
GAZİ ÜNİVERSİTESİ
Gazi
1926

# MİNİMUM KOVARYANS DETERMİNANTINA DAYALI SAĞLAM DİSKRİMİNANT ANALİZİ

**Sercan SEZER**

**YÜKSEK LİSANS TEZİ**
**İSTATİSTİK ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**

**EKİM 2019**

Sercan SEZER tarafından hazırlanan "MİNİMUM KOVARYANS DETERMİNANTINA DAYALI SAĞLAM DİSKRİMİNANT ANALİZİ" adlı tez çalışması aşağıdaki jüri tarafından OY BİRLİĞİ ile Gazi Üniversitesi İstatistik Ana Bilim Dalında YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

**Danışman:** Doç. Dr. Necla GÜNDÜZ

İstatistik Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ..……

**Başkan:** Doç. Dr. İhsan KARABULUT

İstatistik Ana Bilim Dalı, Ankara Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

**Üye:** Dr. Öğr. Üyesi Jale BALİBEYOĞLU

İstatistik Ana Bilim Dalı, Gazi Üniversitesi

Bu tezin, kapsam ve kalite olarak Yüksek Lisans Tezi olduğunu onaylıyorum. ……...

Tez Savunma Tarihi: 21/10/2019

Jüri tarafından kabul edilen bu tezin Yüksek Lisans Tezi olması için gerekli şartları yerine getirdiğini onaylıyorum.

…….…….

Prof. Dr. Sena YAŞYERLİ

Fen Bilimleri Enstitüsü Müdürü

## ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Sercan SEZER

21/10/2019

# MİNİMUM KOVARYANS DETERMİNANTINA DAYALI SAĞLAM DİSKRİMİNANT ANALİZİ

(Yüksek Lisans Tezi)

Sercan SEZER

GAZİ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

Ekim 2019

## ÖZET

Sağlıklı tahminleme süreci (öngörü süreci) çeşitli araç ve yöntemlerle derlenmiş verinin, yığını temsil kabiliyetinin nitelikli olduğuna kanaat edilmesiyle başlamaktadır. Gerçek hayatta karşılaştığımız veriler her zaman bu ideali taşımayabilirler. Veride yer alan bazı gözlem ya da gözlem grupları verinin genelinin yoğunlaştığı bölgeden farklı bir noktada konumlanmış olabilir. Bu gözlem ya da gözlem gruplarına aykırı gözlem ya da başka bir dağılımdan karışan gözlem denilmektedir. Mevcut klasik metodlar veride aykırı gözlem ya da başka bir dağılımdan karışan gözlemlerin varlığından çok fazla etkilenebilirler. Bu durum tahmin edicilerin güvenilirliğinin düşmesine sebep olabilir. Sağlam istatistiklerle, (sağlam tahmin edicilerle) başka bir dağılımdan karışan gözlem ya da gözlem grubunu barındıran eldeki mevcut veri setinden hareketle, aykırı gözlemlerin bulunmadığı ya da etkisinin az tutulduğu duruma yakın tahmin ediciler elde edilebilir. Bu çalışmada, çok değişkenli gözlem birimlerinin sınıflandırılmasında kullanılan diskriminant analizinde, kırılma noktası yüksek ve sağlamlık özelliğine sahip tahmin edicilerden Minimum Kovaryans Determinantı (MCD) ve MCD temelli çeşitli algortimalara yer verilmiş olup, bu algoritmalar sonucunda elde edilen tahmin ediciler ile mevcut En Çok Olabilirlik (MLE) tahmin edicileri karşılaştırılmış ve diskriminant analizi sürecinde gözlemlerin hatalı sınıflandırma oranları (ARE) üzerindeki etkileri incelenmiştir.

Bilim Kodu : 20513

Anahtar Kelimeler : Diskriminant analizi, karışan gözlem, yüksek kırılma noktası, sağlamlık, MCD, hatalı sınıflandırma oranı

Sayfa Adedi : 69

Danışman : Doç. Dr. Necla GÜNDÜZ

ROBUST DISCRIMINANT ANALYSIS BASED ON MINIMUM COVARIANCE DETERMINANT

(M. Sc. Thesis)

Sercan SEZER

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

October 2019

## ABSTRACT

The healthy estimation process begins with the conviction that the data compiled by various tools and methods is qualified to represent the mass. The data we encounter in real life may not always carry this ideal. Some observations or observation groups in the data may be located at a different point from the region where the overall data is concentrated. These observations or observation groups are called outlier or contamination with another distribution. Existing classical methods may be greatly influenced by the presence of observations contamination with another distribution or outliers in the data. Hence, this situation may result in reduced reliability of the estimators. With robust statistics, estimators can be obtained for the original data set containing the observation or observation group contamination from another distribution, close to the situation where there are no or low affect for outlier or contamination with another distribution. In this study, Minimum Covariance Determinant (MCD) and based on other types of MCD algorithms are handled with high breakdown point and robustness for discriminant analysis used in the classification of multivariate observations. As a result of this situation, estimators obtained from these algorithms were compared with the existing Maximum Likelihood Estimators (MLE) and their effects on apparent rate error (ARE) for misclassification rates during discriminant analysis process.

Science Code : 20513

Key Words : Discriminant analysis, outlier, contamination, high breakdown point, robustness, MCD, misclassification, apparent rate error

Page Number : 69

Supervisor : Assoc. Prof. Dr. Necla GÜNDÜZ

# TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesinde, değerli bilgilerini benimle paylaşan, zaman mefhumu gözetmeksizin desteğini hiç bir zaman esirgemeyen ve üzerimdeki emeği ödenemeyecek düzeyde olan, kendisine ne zaman danışsam bana kıymetli zamanını ayırıp sabırla ve büyük bir ilgiyle faydalı olabilmek için elinden gelenden fazlasını sunan, her sorun yaşadığımda yanına çekinmeden gidebildiğim, güler yüzünü ve samimiyetini benden esirgemeyen, mevcut ve gelecekteki mesleki hayatımda da bana verdiği değerli bilgilerden faydalanacağımı düşündüğüm çok kıymetli danışman hocam Doç. Dr. Necla GÜNDÜZ'e teşekkürü bir borç biliyor ve şükranlarımı sunuyorum.

Bu günlere gelmemde emeği olan tüm hocalarıma, hayatım boyunca bana kazandırdıkları her şey için ve beni gelecekte söz sahibi yapacak bilgilerle donattıkları için hepsine teker teker teşekkürlerimi sunarım.

Beni bu günlere sevgi ve saygı kelimelerinin anlamlarını bilecek şekilde yetiştirerek getiren ve benden hiçbir zaman desteğini esirgemeyen, bu hayattaki en büyük şansım olan annem, babam ve kardeşime sonsuz teşekkürler.

# İÇİNDEKİLER

**Sayfa**

**Sayfa**

**Sayfa**

**Sayfa**

# ÇİZELGELERİN LİSTESİ

**Çizelge** **Sayfa**

**Çizelge** **Sayfa**

# ŞEKİLLERİN LİSTESİ

**Şekil** **Sayfa**

**Şekil** **Sayfa**

# SİMGELER VE KISALTMALAR

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

| Simgeler | Açıklamalar |
|---|---|
| **p** | Değişken sayısı |
| **g** | Grup sayısı |
| **n** | Gözlem birimi sayısı |
| **ρ** | İki değişken arasındaki Pearson korelasyon |
| **ε** | Başka bir dağılımdan karışma oranı |
| $\nu$ | Lokasyon şişirme faktörü |
| **κ** | Ölçek şişirme föktörü |
| $l$ | Lineer fonksiyonel |
| **h** | Seçilecek alt matrislerin boyu |
| **X** | n x p boyutlu veri matrisi |
| **Y** | p boyutlu bağımlı değişken vektörü |
| $F_{\varepsilon}$ | Başka bir dağılımdan karışmış dağılım |
| $F_X$ | X rassal değişkeninin dağılım fonksiyonu |
| $\mu_j$ | j. gruba ait p boyutlu ortalama vektörü |
| $\bar{\mathbf{x}}_\mathbf{j}$ | j. grubun ortalama istatistiği |
| $\pi_j$ | j. gruba ait n x p boyutlu çok değişkenli yığın |
| $\pi_{j_\varepsilon}$ | ε oranda karışmış çok değişkenli yığın |
| **C** | Grup içi var-cov matrisi |
| $f_i$ | Fisher'in i. grup için diskriminant fonksiyonu |
| $Q_p$ | Konum öteleme faktörü |
| $\mathbf{I_{pxp}}$ | p x p boyutlu birim matris |
| $\mathbf{p_i}$ | i. Sınıfa atanma önsel olasılığı |
| **T(F)** | F üzerinde tanımlı lineer fonksiyonel |

| Simgeler | Açıklamalar |
|---|---|
| $\Delta_{x_0}$ | Basamak fonksiyonu |
| $\varepsilon^*(X,T)$ | Kırılma noktası |
| $z_{ik}$ | k. grup için merkezileştirilen gözlemler |
| $m_k$ | k. grup için nihai grup ortalaması |
| **C(r,s)** | Birleştirilmiş grup içi varyans-kovaryans bileşenleri |
| $\widehat{\Sigma_{MLE}}$ | MLE'ye dayalı varyans-kovaryans matrisi |
| $\widehat{\mu_{MLE}}$ | MLE'ye dayalı ortalama vektörü |
| $\widehat{\Sigma_{MCD}}$ | MCD'ye dayalı varyans-kovaryans matrisi |
| $\widehat{\mu_{MCD}}$ | MCD'ye dayalı ortalama vektörü |

| Kısaltmalar | Açıklamalar |
|---|---|
| **ARE** | Hatalı sınıflandırma oranı |
| **BLUE** | En İyi Doğrusal Sapmasız Tahmin Edici |
| **BP** | Kırılma Noktası |
| **COV** | Kovaryans |
| **HBK** | Hawkins- Bradu- Kass yapay veri seti |
| **IF** | Etki fonksyionu |
| **INF** | En büyüklerin en küçüğü |
| **LR** | Olabilirlik Oranı |
| **MAD** | Medyandan mutlak sapma |
| **MCD** | Minimum Kovaryans Determinant |
| **MD** | Mahalanobis Uzaklığı |
| **MLE** | Maksimum Olabilirlik Tahmin Edicisi |
| **MSE** | Hata Kareler Ortalaması |
| **RD** | Sağlam Uzaklık |
| **SC** | Duyarlılık Eğrisi |
| **SUP** | En küçüklerin en büyüğü |

# 1 . GİRİŞ

Klasik doğrusal diskriminant analizi, mevcut verideki orjinal değişkenlerin doğrusal fonksiyonlarını (diskriminant fonksyionları) kullanarak, gözlem birimlerini sınıflandırmak ve gerçekte hangi gruptan geldiği belli olmayan yeni bir gözlemin, taşımış olduğu değişkenler bakımından hangi gruba atanacağı hususunda öngörüde bulunan çok değişkenli istatistiksel bir yöntemdir.

Klasik doğrusal diskriminant analizinin en önemli kısıtlayıcılarından biri, veride aykırı gözlem ya da gözlem grubunun olmamasıdır. Zira, klasik doğrusal diskriminant analizi sonucu elde edilen tahmin ediciler aykırı gözlem ya da başka dağılımdan karışan gözlem gruplarına karşı oldukça duyarlıdır. Veri setinde aykırı gözlem ya da gözlem grubunun olması durumunda, klasik doğrusal diskriminant analizi sonucu elde edilen tahmin edicilerin değeri gerçek değerinden çok farklı çıkabilmektedir. Bu durum, ilgili tahmin ediciler kullanılarak elde edilen diskriminant fonksiyonları katsayılarını olumsuz etkilemekte ve bunun sonucu olarak gözlem birimlerinin sınıflandırılması veya yeni bir gözlem biriminin gruplara atama öngörüsünde hataya sebep olmaktadır.

Sağlam, sağlıklı, dayanıklı ve gürbüz gibi anlamlar taşıyan robust sözcüğü, ilk kez 1953 yılında Box tarafından istatistik literatüründe kullanılmış olup, varsayımların sağlanmadığı durumda sağlıklı sonuçlar veren istatistiksel yöntemler olarak adlandırılmıştır (Huber, 1964).

Veride aykırı gözlem ya da gözlem grubunun bulunması durumunda, yeni gözlemler için etkili bir grup ataması yapabilmek için aykırı değerlerin varlığından, mevcut klasik tahmin edicilere göre nispeten az ya da hiç etkilenmeyen, (varsayımlardan sapmalara duyarlı olmayan tahmin ediciler) sağlam istatiksel yöntemler sonucu elde edilen sağlam istatistikler tercih edilmektedir. İstatistiksel modellemede ve veri analizinde kullanılan sağlam yöntemlerle, varsayımların sağlandığı ve veri setinde aykırı değerlerin bulunmadığı durumların yanı sıra, varsayımların sağlanmadığı ve ölçüm hatasından ya da verinin doğasından kaynaklanan aykırı değerlerin bulunduğu durumlarda da güvenilir tahmin ediciler üretilebilir ve bu tahmin edicilere bağlı olarak sağlıklı güven aralıkları ve olasılık cümleleri kurulabilir.

Klasik doğrusal diskriminant analizinin uygulanması hususunda klasik tahmin edicilerin sağlam olmama problemi, Todorov (2009), Chorck ve Rousseeuw (1992), Hawkins ve McLachlan (1997), He ve Fung (2000), Croux ve Dehon (2001), Hubert ve Van Drieesen (2004) gibi birçok yazar tarafından ele alınmıştır. Bu araştırmacılar tarafından önerilen istatistiksel yöntemler, gelişmiş bilgisayar teknolojisi sayesinde dünyada birçok yazarın

sanal istatistik kütüphanesi oluşumuna katkı sağladığı, açık kod kaynaklı R paket programının gelişimini desteklemiştir.

Bu tez çalışmasında, çok değişkenli gözlemlerin sınıflandırılması ve yeni bir gözlemin gruplara atanmasında kullanılan diskriminant analizinde kırılma noktası yüksek ve sağlamlık özelliğine sahip MCD ve MCD'ye dayalı çeşitli tahmin ediciler incelenmektedir. Bu tahmin edicilere dayalı algoritmalar sonucu elde edilen tahmin ediciler ile mevcut MLE tahmin edicileri karşılaştırılmış ve diskriminant analizi sürecinde gözlemlerin hatalı sınıflandırma oranları üzerindeki etkileri incelenmiştir.

İkinci bölümde, klasik doğrusal diskriminant analizine ait genel bilgilere yer verilmiş ve yöntemin işleyişi anlatılmıştır. Üçüncü bölümde, çok değişkenli lokasyon ve ölçek parameterleri tanıtılmış, bu parameterlerin tahmininde en çok olabilirlik yöntemi ve bu yöntem sonucu elde edilen tahmin edicilerin taşımış olduğu özelliklere yer verilmiştir. Dördüncü bölümde, MLE'ye dayalı aykırı gözlem tespiti ve aykırı gözlemlerin klasik doğrusal diskriminant analizi üzerindeki etkileri anlatılmıştır. Beşinci bölümde, sağlam tahmin ediciler, taşımış olduğu özellikler ve bazı çok değişkenli sağlam tahmin edicilere ilişkin algoritmalara yer verilmiştir. Altıncı bölümde, çalışma kapsamında gerçek veri setleri üzerinde gerek manuel (elle aşama aşama) gerekse de paket program desteği ile uygulamalar yapılmış ve sonuçlar yorumlanmıştır. Yedinci bölümde, sağlam tahmin ediciler ve klasik tahmin edicilerle, R paket programı desteği ile yapılan diskriminant analizine ilişkin simülasyon çalışmalarına yer verilmiştir. Sekizinci bölümde, elde edilen bulgular açıklanmıştır.

# 2 . KLASİK DOĞRUSAL DİSKRİMİNANT ANALİZİ

Diskriminant analizi; her birinde $p$ tane değişken bulunan $k$ sayıda gruptan ($k > 2$) elde edilecek doğrusal fonksyionlar yardımı ile, gözlemlerin sınıflandırılması ve $p$ tane değişkene sahip yeni bir gözlem birimini, herhangi bir gruba atamak istediğimizde kullanılan bir yöntemdir.

Özellikleri bakımından ölçülen ve birimden birime farklılık gösteren, deney sonuçlarının yer aldığı örnek uzay üzerindeki her bir noktayı reel sayılara bağlayan fonksiyonlara rassal değişken denir (Casella ve Berger, 2002:27). Klasik doğrusal diskriminant analizinde değişken değerleri bakımından ölçülen bir gözlem birimi, o gözlem biriminin atanacağı grubu etkilediğinden ve aralarında böyle bir nedensellik olduğundan ilgili gözlem birimini ölçen değişkenler bağımsız değişken, gözlem biriminin yer aldığı grup ise bağımlı değişken olacaktır.

Bu bağlamda, bağımsız değişkenlerin doğrusal bir bileşeni yada bileşenleri elde edilir ve bu bileşenler kullanılarak gözlem birimleri ait olduğu gruba atanır. Diskriminant analizi; grup ortalama vektörleri arasındaki farklılığı maksimum yapacak şekilde bağımsız değişkenlerin doğrusal fonksiyonlarını bulma işlemidir. Yukarıdaki tanımlar çerçevesinde diskriminant analizinin amaçları ve kullanım yerleri aşağıdaki gibi özetlenebilir:

1) Grupları birbirinden ayırmayı sağlayacak olan doğrusal fonksiyonları bulmak,

2) Bulunan fonksiyonlar yardımıyla, yeni bir gözlemi en az hata ile gruplardan birine atamak,

3) Çalışmaya alınan değişkenlerin hangilerinin grup üyeliğini kestirmekteki katkısının en fazla olduğunu belirlemektir (Alpar, 2011:691).

## 2.1. Genel bilgiler

### 2.1.1. Bağımlı ve bağımsız değişkenler

Yapılan tanımlamalardan anlaşılacağı üzere, diskriminant analizinde bağımlı değişken iki ya da daha fazla kategorili niteliksel değişken iken, bağımsız değişkenler sürekli ya da kesikli niceliksel değişkenlerdir. Yani, bağımlı değişken gruplar iken bağımsız değişkenler gruplara ilişkin özellikleri ifade eden değişkenlerdir.

### 2.1.2. Klasik doğrusal diskriminant analizinin aşamaları

Diskriminant analizi, iki aşamalı bir işlem olarak ele alınabilir;

1) Diskriminant fonksiyonlarının önemlilik testi,

2) Yeni gözlemler için grup üyeliklerinin kestirilmesi.

Birinci aşamada; tüm değişkenler dikkate alınarak, gruplar arasında önemli bir farklılık olup olmadığını ya da incelenen bağımsız değişkenlerin doğrusal fonksiyonunun yazılıp yazılamayacağı belirlenir. Çok değişkenli test sonucunda önemli bir farklılık bulunduğunda yani; ayırıcı fonksiyonlar oluşturulabileceği kanısına varıldığında, hangi değişkenin modele katkısının daha önemli olduğu da belirlenebilir.

İkinci aşamada ise, klasik doğrusal diskriminant fonksiyonları ve gözlem birimlerinin gruplara atanması, aşağıda verilen sıra ile uygulanır :

1) $k.$ grupta yer alan gözlem birimlerinin, genel grup ortalamasından farkı şeklinde tanımlı ($x_i^0$) yeni rassal değişken matrisi oluşturulur.

2) Grup içi varyans-kovaryans matrisi ($c_i$) oluşturulur.

$$c_i = \frac{(x_i^0)^T \cdot x_i^0}{n_i} \tag{2.1}$$

3) Birleştirilmiş grup içi varyans-kovaryans matisi (C(r,s)) bileşenleri hesaplanır.

$$C(r,s) = \frac{1}{n} \sum_{i=1}^{g} n_i \cdot c_i(r,s) \tag{2.2}$$

4) Gözlem biriminin ait olduğu grup hakkında herhangi bir bilgi bulunmadığı durumda, ilgili gözlemin $i$ gruba atanma olasılığı ($p_i$) hesaplanır.

$$p_i = \frac{n_i}{n} \tag{2.3}$$

5) Elde edilen bilgiler Fisher'ın Kanonik Ayırma Fonksiyonunda yerine yazılır.

$$f_i = \mu_i C^{-1} \cdot x_k^T - \frac{1}{2} \mu_i C^{-1} \cdot \mu_i^T + \ln(p_i) \tag{2.4}$$

Grup sayısı kadar doğrusal fonksiyon olacaktır. Doğrusal fonksiyon değerleri herbir gözlem için ayrı ayrı hesaplanır. İlgili gözlem en yüksek fonksiyon skoruna sahip diskriminant fonksiyonunu temsil eden gruba atanır.

## 2.2. Diskriminant analizinin varsayımları ve kısıtlayıcıları

### 2.2.1. Bağımsız değişkenlerin çok değişkenli normal dağılım göstermesi

Grup ortalama vektörlerinin karşılaştırılmasında parametrik bir aileyi temsilen $F$ istatistiğinin kullanılabilmesi için; her bir gruptaki değişkenlerin çok değişkenli normal dağılım göstermesi gerekir.

### 2.2.2. Bağımsız değişkenlerin varyans-kovaryans matrislerinin homojenliği

Gruplardaki örneklem büyüklüğünün yeterli olması, ancak varyans-kovaryans matrislerinin homojen olmaması durumunda, gözlemler daha yüksek kovaryansa sahip olan gruplara fazlasıyla sınıflandırılır. Dolayısıyla bu durum sınıflandırma hatasını arttırır.

### 2.2.3. Bağımsız değişkenler arasında çoklubağlantı sorununun olmaması

Bağımsız değişkenlerin biri (birkaçı), diğeri (diğerleri) ile yüksek derecede ilişkili ise ya da diğer değişkenlerin bir fonksiyonu şeklinde yazılabiliyorsa çoklu bağlantı sorunu ile karşılaşılır. Bu nedenle açıklayıcı değişkenler arasındaki korelasyon matrisi incelenmeli ve varyans şişme faktörleri yani, korelasyon matrisinin tersinin köşegen elemanlarına bakılmalıdır. Bu değerler 10'dan büyükse açıklayıcı değişkenler arasında çoklu bağlantı problemi olabileceği düşünülür.

### 2.2.4. Aykırı gözlem olmaması

Diskriminant analizi, diğer tüm istatistiksel yöntemler de olduğu gibi aykırı değerlere oldukça duyarlıdır. Örneğin, gruplardan biri ortalama istatistiği vektörünün değerini etkileyecek düzeyde aykırı değere sahipse, bu gözlemler aynı zamanda değişkenliğin bir ölçüsü olarak varyans tahminlerini de önemli ölçüde değiştirecektir. Bu ise sonuçları tamamen değiştirecek, hatalı grup üyeliklerinin belirlenmesine sebebiyet verecektir. Bu amaçla, her bir gurup için ayrı ayrı, tek ve çok değişkenli olarak aykırı değerler açısından inceleme yapılabileceği gibi veride aykırı gözlem olup olmadığını araştırmak yerine, aykırı değerlerden etkilenimi az olan sağlam diskriminant fonksiyonlarının kullanılmasının uygun olduğu düşünülmektedir.

## 2.3. Diskriminant analizinin geçerliliği

Diskriminant analizi sonuçlarının başarılı olup olmadığını (sınıflandırma yeterliliğini) farklı yaklaşımlarla inceleyebiliriz.

### 2.3.1. Sınıflandırma tutarlılığı tablosu

Diskriminant analizinin geçerliliği konusunda en basit yaklaşım, gözlemlerin gerçek grup üyelikleri ile diskriminant analizi sonucunda kestirilen grup üyeliklerinin çapraz tablo (Sınıflandırma Tutarlılığı Tablosu) ile incelenmesidir. $A, B, C$ gibi üç grup olduğunda üç yaklaşım için bir çapraz tablo örneği Çizelge 2.1'de verilmiştir. Tabloda $f_{AA}, f_{BB}, f_{CC}$ ile gösterilen köşegen elemanları doğru sınıflama sıklıklarını, köşegen dışındaki gözelerde oluşacak sıklıklar ise hatalı sınıflama sıklıklarını göstermektedir.

Çizelge 2.1. Üç Grup İçin Sınıflandırma Tutarlılığı Tablosu

| | Kestirilen(Atanan)Grup | | | |
|---|---|---|---|---|
| Gerçek grup | A | B | C | Toplam |
| A | $f_{AA}$ | | | |
| B | | $f_{BB}$ | | |
| C | | | $f_{CC}$ | |
| Toplam | | | | n |

Sınıflama tutarlılığı tablosunun irdelenmesinde genellikle iki yaklaşımdan yararlanılır. Birinci yaklaşımda, hazırlanan çapraz tabloda köşegen elemanlarının toplamı genel toplama bölünerek analizin genel tutarlılık/ geçerlilik oranı (doğru sınıflama oranı) belirlenir. (Yani çizelge 2.1'e göre doğru sınıflandırma oranı $\left(\frac{f_{AA}+f_{BB}+f_{CC}}{n}\right)$'dir.) Eğer Diskriminant analizinin kestirimi *mükemmel* ise tüm gözlemler köşegen elemanları üzerinde toplanacak ve genel tutarlılık oranı 1 olacaktır.

Gözlemlerin yanlış sınıflandırılmasına ait hatalı sınıflandırma oranı ARE, kullanılan tahmin edici ile yapılan diskriminant analizinin test hatasını göstermekte olup; ARE = (1-doğru sınıflandırma oranı) şeklinde hesaplanır.

### 2.3.2. Press $\theta$ istatistiği

$H_0$: Diskriminant analizinde; bağımsız değişkenlerin doğrusal fonksiyonları istatistiksel açıdan yazılamaz.

$H_1$: Diskriminant analizinde; bağımsız değişkenlerin doğrusal fonksiyonları istatistiksel açıdan yazılabilir.

şeklinde $H_0$ hipotezini $\theta$ istatistiği yardımıyla test edebiliriz. Press $\theta$ istatistiği de denilen bu istatistik "1" serbestlik dereceli bir $\chi^2$ dağılımıdır (Alpar, 2011:708:711).

$$\theta = \frac{(n - k\sum f_{ii})^2}{n(k-1)} \tag{2.5}$$

# 3. ÇOK DEĞİŞKENLİ LOKASYON (KONUM) VE ÖLÇEK PARAMETRELERİNİN TAHMİNİNDE EN ÇOK OLABİLİRLİK YÖNTEMİ

## 3.1. Çok değişkenli lokasyon ve ölçek parametrelerinin tanımlanması

Bir rassal değişkene ait lokasyon parametresi; o rassal değişkene ait gözlem birimlerinin yoğun olarak konumlandığı noktadır. Ölçek parametresi ise; ilgili rassal değişkene ait gözlem birimlerinin lokasyon parametresi etrafındaki yayılımının bir ölçüsüdür. Örneğin; tek değişken ile temsil edilen bir rassal değişkende ortalama ve medyan değerleri lokasyon parametreleri iken;varyans ve medyandan mutlak sapmaların medyanı (Median Absolute Deviation (MAD)) ölçek parametreleridir.

Çok değişkenli yapı içerisinde $p$ boyutlu $n$ defa tekrardan oluşan $X$ veri matrisi Eş. 3.1'teki gibi tanımlanmış olsun:

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{bmatrix}. \tag{3.1}$$

Örneğin, $i.$ nci satır $i.$ nci denemedeki $p$ değiskene ilişkin sonuçlar $x_{i1}, x_{i2}, ..., x_{ip}$ ile gösterilmektedir.

## 3.2. En çok olabilirlik yöntemi

Çok değişkenli normal dağılım ailesi, karakteristik özelliklerinin ortaya konulması bakımından ortalama vektörü $\mu$ ve varyans-kovaryans matrisi $\Sigma$ parametreleri ile ifade edilmektedir. Bu ifade kısaca; $N(\mu, \Sigma)$ şeklinde gösterilmektedir. Rassal değişken, mümkün çözüm uzayındaki (örnek uzaydaki) herbir gözlem birimine rassal örneğin bir fonksiyonu olan yeni değerleri atamak şeklinde çalışmaktadır. Örneğin, bu amaçla çalışan $X$ rassal değişkeni, çok değişkenli normal dağılıma sahiptir ifadesi kısaca; $X \sim N(\mu, \Sigma)$ şeklinde gösterilir.

İstatistiğin en önemli konusunu, $X$ rassal değişkenine ait yığından alınan rassal bir örneğin $(X_1, X_2...., X_n)$ sağlamış olduğu veri bilgisinden istifade ederek, $\mu$ ve $\Sigma$ parametrelerinin maksimum olasılıkla ne olabileceği hususunda tahminleme yapma süreci oluşturmaktadır. Bu amaçla, kullanılan yöntemlerden biri MLE yöntemi iken, bu yöntem ile bulunan parametre tahmin edicilerine de En Çok Olabilirlik Tahmin Edicileri (Maksimum Likelihood Estimator-MLE) denir (Anderson, 2003:67).

$N$ gözleme sahip $X$ rassal değişkeni, çok değişkenli normal dağılıma sahip olsun. Bu dağılımdan $n$ çaplı birbirinden bağımsız ve aynı dağılımlı ($X_1$, ...., $X_n$) rassal örneği alınsın ve yeni bir rassal değişken, alınan rassal örneğin bir fonksiyonunu yeni bir değer kümesine atasın. Bu yeni rassal değişkenin bir ismi vardır. İstatistik literatüründe adını sıklıkla duymuş olduğumuz bu rassal değişkenin adı Olabilirlik Fonksiyonudur $(L)$ veya Olabilirlik Oranı (LR)'dir.

$$L = \prod_{i,k}^{n} n(x_{i,k}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}pn}|\Sigma|^{\frac{1}{2}n}} exp\left[-\frac{1}{2}\sum_{i,k}^{n}(x_{i,k}-\mu)'\Sigma^{-1}(x_{i,k}-\mu)\right]. \tag{3.2}$$

şeklinde ifade edilir. Eş.3.2'deki ifadeyi maksimum yapan $\mu^*$ ve $\Sigma^*$'yı bulmak için sürekli ve doğrusal bir fonksiyon niteliğinin kazandırılması gerekmektedir. Bu bağlamda logaritması alınan olabilirlik fonksiyonu pozitif tanımlı $\Sigma^*$ varyans- kovaryans matrisi ve $\mu^*$ vektörüne göre artan bir fonsiyon olduğundan, bu parametrelere göre türevinin değişmediği noktada, $(X_1, X_2...., X_n)$ rassal örneğinin mümkün çözüm uzayı üzerinde rast gelmesi sıklık bilgisinin maksimum olasılıkla olacağı sonucunu doğurur:

$$\log L = -\frac{1}{2}pn\log 2\pi - \frac{1}{2}n\log|\Sigma^*| - \frac{1}{2}\sum_{i,k}^{n}(x_{i,k}-\mu^*)'\Sigma^{*-1}(x_{i,k}-\mu^*). \tag{3.3}$$

Dolayısıyla, $\mu$ ve $\Sigma$ parametrelerinin tahmin edicileri olan $\Sigma^*$ varyans- kovaryans matrisi ve $\mu^*$ vektörü örneklem ortalaması ve varyans kovaryans matrisi tahmin edicileri olarak kaşımıza çıkmaktadır:

$$\widehat{\Sigma^*} = \Sigma_{MLE} = \frac{\left(\sum_{k=1}^{n}\begin{bmatrix} x_{k1}-\overline{x_1} \\ x_{k2}-\overline{x_2} \\ \cdots \\ x_{kp}-\overline{x_p}\end{bmatrix}\begin{bmatrix} x_{k1}-\overline{x_1} & x_{k2}-\overline{x_2} & \cdots & x_{kp}-\overline{x_p}\end{bmatrix}\right)}{n-1}, \tag{3.4}$$

$$\widehat{\Sigma^*} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{pmatrix}, \tag{3.5}$$

$$\widehat{\mu^*} = \mu_{MLE} = \frac{1}{n}\left(\begin{bmatrix} x_{11} \\ x_{12} \\ \cdots \\ x_{1p}\end{bmatrix} + \begin{bmatrix} x_{21} \\ x_{22} \\ \cdots \\ x_{2p}\end{bmatrix} + \ldots + \begin{bmatrix} x_{n1} \\ x_{n2} \\ \cdots \\ x_{np}\end{bmatrix}\right), \tag{3.6}$$

burada $\overline{x}_j = \sum_{i=1}^{n} x_{i,j}$ $(j = 1, ..., p)$, ve $\overline{x} = [\overline{x}_1, \overline{x}_2, ...\overline{x}_p]^T$'dır.

## 3.3. En çok olabilirlik yöntemine dayalı klasik tahmin ediciler için değerlendirme kriterleri

### 3.3.1. Hata kareler ortalaması (Mean Square Error (MSE)

Tahmin edicinin parametreden farkı olarak tanımlanan yeni rassal değişkeninin karesinin beklenen değeridir. Bir diğer anlamda tahmin edicinin parametreye olan ortalama uzaklığıdır. Elimizde birden fazla tahmin edici olduğu durumda paremetreyi yüksek bir olasılıkla öngörmek hususunda hangisinin daha iyi bir iş olacağı konusunda MSE önemli bir değerlendirme kriteri olarak karşımıza çıkmaktadır. (Casella ve Berger, 2002:330:334)

Parametresi $\theta$ olan $X$ rassal değişkeninin dağılımından, birbirinden bağımsız ve aynı dağılımlı $X_1, ...., X_n$ rassal örnek alınsın. $\widehat{\theta}$ tahmin edicisi rassal örnek üzerinde tanımlı bir istatistik olmak üzere, $\widehat{\theta}$ için MSE,

$$\begin{aligned} MSE(\widehat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2) \\ &= E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2 \\ &= V(\hat{\theta}^2) + \Big[E(\hat{\theta}) - \theta)\Big]^2 \\ &= V(\hat{\theta}) + \Big[Bias(\hat{\theta})\Big]^2 \end{aligned} \tag{3.7}$$

şeklinde tanımlanmaktadır ve burada; $Sapma = Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$'dır.

Eğer sapma sıfır ise; tahmin edicinin MSE değeri doğrudan varyansına eşit olmaktadır. Ayrıca bu durum, tahmin edicinin parametre için sapmasız bir istatistik olduğu sonucunu doğurmaktadır.

### 3.3.2. Yansızlık (Sapmasızlık)

Yansızlık, tahmin edicilerde en çok aranan özelliklerden biridir. Parametresi $\theta$ olan X rassal değişkeninin dağılımından birbirinden bağımsız ve aynı dağılımlı $X_1$,$X_2$,...,$X_n$ rassal örneği alınsın. $\widehat{\theta}$ tahmin edicisi rassal örnek üzerinde tanımlı bir istatistik olsun. $E(\widehat{\theta}) = \theta$ ise, $\widehat{\theta}$ istatistiği $\theta$ parametresi için yansız bir tahmin edicidir (Akdi, 2005:254). Tanımdan anlaşılacağı üzere, yansız bir tahmin edicinin var olması demek, istatistiğin (tahmin edicinin) beklenen değerinin bizi parametreye götürmesi demektir.

### 3.3.3. Tutarlılık

Büyük sayılar yasasının güçlü tarafı gereği eğer gözlem birimleri birbirinden bağımsız ve aynı dağılıma sahip iseler örnek çapının arttığı durumda tahmin edicinin varyansının hızlı bir şekilde sıfıra yaklaşması o tahmin edicinin parametre için tutarlı bir istatistik olacağı sonucunu doğurur (Casella ve Berger, 2002:467)

### 3.3.4. Yeterlilik

Bilinmeyen $\theta$ parametresi için yeterli istatistik, veri seti içinde bulunan parametre hakkındaki bilgileri tek olarak (veya birkaç tane) özetleyen istatistiktir. Yani, veri içinde parametre hakkında ne kadar bilgi varsa, bu bilgileri bize özet olarak (hiç bilgi kaybı olmadan) veren tahmin edici parametre için yeterli bir tahmin edicidir. Başka bir ifade ile, tahmin edicinin verdiği bilgi dışında, veri içinde başka bilgi yoktur. Yani, tanımlı bir istatistik koşulu altında yığının dağılımından alınan bir rassal örneğin önümüze gelmesi, sıklık bilgisi $\theta$ parametresinden bağımsız ise tanımlı istatistiğin kendisi bir yeterli istatistiktir. Tanımın kendisinden de anlaşılacağı üzere yeterli istatistik tek değildir ve bir parametre için birçok yeterli tahmin edici elde edilebilir (Casella ve Berger, 2002:348).

### 3.3.5. Tamlık

$X_1$,$X_2$,...,$X_n$; $X$ rassal değişkeninin dağılımdan alınan bir rastgele örneklem olmak üzere, eğer rastgele örneklemin bir fonksiyonu şeklinde tanımlı $\hat{\theta}$'nın $\theta$ altındaki tüm beklenen değerleri sıfır oluyorsa $\hat{\theta}$ tahmin edicisi $X_1$,$X_2$,...,$X_n$ rassal örneğinin geldiği dağılım ailesi için tamlık vasfına haizdir denir (Hogg ve Craig, 1978:389-395).

### 3.3.6. En iyi doğrusal sapmasız tahmin edici (Best Linear Unbiased Estimator (BLUE))

Bu özellik, tahmin edicide sapmasızlık ve etkinliğin yanında doğrusallık özelliğininde aranması gerektiğini ifade etmektedir.

1) $\theta$ parametresinin tahmin edicisi olan $\hat{\theta}$ istatistiği örnek birimlerinin doğrusal fonksiyonu ise,

2) $\hat{\theta}$, $\theta$ parametersinin sapmasız tahmincisi ise,

3) Var($\hat{\theta}$) $\leq$ Var($\bar{\theta}$) ise

$\hat{\theta}$, $\theta$ parametersinin en iyi doğrusal sapmasız tahmincisidir. Burada $\bar{\theta}$; $\theta$ parametresinin tahmin edicisi olan diğer doğrusal sapmasız tahmin edicileri ifade etmektedir. Tahmin edicinin örnek gözlemlerinin doğrusal fonksiyonu olma şartı tahmin için kullanılacak formüle bakılarak belirlenir. $X_1$,$X_2$,...,$X_n$ birbirinden bağımsız ve aynı dağılımlı rassal örnek ve $a_1$,$a_2$,...$a_n$ sabit sayılar olmak üzere, tahmin edicinin $a_1X_1$+$a_2X_2$+...+$a_nX_n$ yapısında olması durumunda doğrusallıktan söz edilecektir. Doğrusallık şartının aranma sebebi, minimum varyansa sahip sapmasız tahmin edicilerin belirlenmesinde karşılaşılan problemdir. Rao-Cramer teoremi ile minimum varyansın alt sınırı yığın dağılımın bilinmesi durumunda belirlenebilir. $\hat{\theta}$, $\theta$'nın sapmasız tahmin edicisi ise, $\hat{\theta}$'nın varyansı,

$$Var(\hat{\theta}) \geq \frac{1}{n \cdot E\left[\left(\frac{\partial Lnf(X)}{\partial \theta}\right)^2\right]} \tag{3.8}$$

eşitsizliğini sağlar. Burada $n$ örnek birim sayısı, $f(X)$ ise yığına ilişkin olasılık yoğunluk fonksiyonudur. Buna göre $\hat{\theta}$, $\theta$'nın sapmasız tahmin edicisi ise ve varyansı,

$$Var(\hat{\theta}) = \frac{1}{n \cdot E\left[\left(\frac{\partial Lnf(X)}{\partial \theta}\right)^2\right]} \tag{3.9}$$

ise, $\hat{\theta}$, $\theta$'nın en etkin sapmasız tahmin edicisi olacaktır. Fakat yığın dağılımı bilinmiyorsa doğrusallık şartının aranması gerekmektedir. Bu şart ile yığın dağılımının bilinmemesi nedeni ile ortaya çıkan problemin çözümü kolaylaşmaktadır. Doğrusallık şartı ile ilgili üç farklı durumdan söz edilebilir:

1) Etkin tahmin edici, örnek gözlemlerinin doğrusal fonksiyondur. Bu durumda en iyi doğrusal sapmasız tahmin edici aynı zamanda etkin tahmin edicidir.

2) Etkin tahmin edici yaklaşık olarak doğrusaldır. Bu durumda en iyi doğrusal sapmasız tahmin edici etkin değildir; fakat en iyi doğrusal sapmasız tahmin edici ile etkin tahmin edicilerin varyansları birbirine yakın olacaktır.

3) Etkin tahmin edici doğrusal ise etkin tahmin edicinin varyansı, en iyi doğrusal sapmasız tahmin edicinin varyansından oldukça küçük olacaktır. Bu durum problem yaratacaktır.

### 3.3.7. Lineer fonksiyonel

Lineer fonksyionel $(l)$, $X_1$,$X_2$,...,$X_n$ birbirinden bağımsız ve aynı dağılımlı rassal örneğin bir fonksiyonudur. n boyutlu mümkün çözüm uzayı ($\Omega$) üzerindeki her bir noktayı reel sayılara bağlayan istatistiklerdir. Dolayısıyla bir rassal değişkendir. Örneğin; rassal örneğin geldiği yığının dağılımının parametresi ya da parametreleri belli, tek değişkenli dağılıma sahip olduğu durumda lokasyon ve ölçek tahmin edicileri üzerinde tanımlı lineer

fonksiyoneller; beklenen değer (bir veri setinde gözlem birimlerinin görülmesi sıklığı ile ağırlıklandırılmış ortalaması), medyan, varyans, MAD, 1. çeyreklik, 3. çeyreklik ve çeyreklikler arası açıklık istatistikleri olabilir. Rassal örneğin geldiği yığının dağılımının parametresi ya da parametreleri belli, çok değişkenli dağılıma sahip olduğu durumda lokasyon ve ölçek tahmin edicileri üzerinde tanımlı lineer fonksiyoneller en çok olabilirlik yöntemi ile elde edilmiş örnek ortalaması istatistiği vektörü, varyans-kovaryans matrisinin tahmin edicisi ve bu tahmin edicilerin sağlam versiyonları olabilir (Maronna ve diğerleri, 2006:71:75).

Konu ile ilgisi açısından ve daha sonra kullanacağımızdan, lineer fonksiyonelin temel özelliklerini kısaca açıklayalım. Beklenen değer ve varyans operatörünün lineer fonksiyonel olduğunu gösterelim:

$$l = a_1X_1 + a_2X_2 + ... + a_nX_n = \sum_{i=1}^{n} a_iX_i \tag{3.10}$$

olarak tanımlanmış olsun. Burada $a_i$ sabit sayıları, $X_1$,$X_2$,...,$X_n$ birbirinden bağımsız ve aynı dağılımlı rassal örneği ifade etmektedir. Fonksiyonun beklenen değeri,

$$E(l) = \sum_{i=1}^{n} a_iE(X_i) = \sum_{i=1}^{n} a_i\mu_i \tag{3.11}$$

ve varyansı,

$$\sigma_l^2 = E[l - E(l)]^2 = E\left[\sum_{i=1}^{n} a_i(X_i - \mu_i)\right]^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i<}\sum_{j} a_i\sigma_i\sigma_j \tag{3.12}$$

olacaktır. Ele alınan örneklem; rassal, aynı dağılımlı ve birbirinden bağımsız bir örneklem olduğundan $\sigma_{ij} = 0$ kabul edilerek,

$$\sigma_i^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2 \tag{3.13}$$

olacaktır. Ortalama ve varyans sabit ise, yani $\mu_i$=$\mu$ ve $\sigma_i^2$=$\sigma^2$ olduğu durumda,

$$E(l) = \mu\sum_{i=1}^{n} a_i \tag{3.14}$$

$$\sigma_i^2 = \sigma^2\sum_{i=1}^{n} a_i^2 \tag{3.15}$$

olacaktır (Güriş, 2010).

# 4. EN ÇOK OLABİLİRLİK TAHMİN EDİCİLERİ AÇISINDAN KLASİK DOĞRUSAL DİSKRİMİNANT ANALİZİ YAKLAŞIMINDA AYKIRI GÖZLEMLER

## 4.1. Kavram olarak "Aykırılık"

Olay ya da olguların doğal gelişimi evresi içerisinde bir gurubun, teşkilatın, kümenin ya da topluluğun tipikliğini reddeden, tavır, tutum ve davranışları bakımından diğerlerinden farklılık gösteren, muhalif olma husssiyeti taşıyan şey anlamına gelmektedir.

## 4.2. İstatistiksel açıdan "Aykırı Değer"

Bir veri setindeki gözlem birimlerinin geneli göz önünde bulundurulduğunda, verinin yoğunlaşma bölgesinden uzakta davranış seyreden (diğer gözlemlerden farklı türde hareket eden) gözlem ya da gözlem gurubu istatistiksel açıdan aykırı ya da başka bir dağılımdan karışan gözlem olarak nitelendirilir.

## 4.3. Klasik doğrusal diskriminant analizinde aykırı gözlemlerin etkisi

Çalışmanın 2. Bölümünde bahsi geçen ve üzerinde durulan klasik doğrusal diskriminant analizi yaklaşımının en önemli varsayımlarından biri de çok değişkenli yapıya sahip grupların aykırı gözlem barındırmamasıdır. Çünkü gruplarda yer alan aykırı gözlemler diskriminant fonksiyonlarının doğru sonuç vermesini engelleyebilir hale getirebilir. Veride aykırı gözlemlerin varlığında; en çok olabilik yöntemi ile hesaplanan diskriminant fonksiyonu katsayılarının parametre tahminleri, gerçek değerlerinden çok uzakta gerçekleşebilir ve bu durum en çok olabilirlik tahmin edicileri ile hesaplanan diskriminant fonksiyonlarını gerçekte olması gerekenden çok farklı bir noktaya taşıyabilir. En nihayetinde de diskriminant fonksiyonlarının gözlemleri doğru sınıflandırma oranı azalır ve gözlemlerin gerçekte olması gerekenden farklı gruplara atanması oranı ARE artar.

## 4.4. Çok değişkenli dağılımlarda aykırı gözlem tespiti

Diskriminant analizinde, grupların herhangi birine ait olmayan ya da olmadığı düşünülen bir gözlem diskriminant analizi için bir aykırı ya da dağılıma karışan gözlem olarak nitelendirilir. Çok değişkenli dağılımlarda aykırı gözlemleri tespit etmek için her bir $x_i$ gözleminin, yoğunlaşmanın merkezinden ne kadar uzakta olduğunu gösteren Mahalanobis uzaklığından (MD) yararlanılabilir.

1) MD; çok değişkenli normal dağılıma sahip $X$ veri matrisindeki her bir gözlemin, verinin merkezinden uzaklığının bir ölçüsüdür. Mahalanobis uzakığı $\alpha = 0,005$ gibi küçük bir anlamlılık düzeyi için $(\chi^2_{(p;0.005)})$ değerini bulmak ve bu değerden büyük $MD(x_i)$ değerine

sahip gözlemleri aykırı gözlem olarak belirleme prensibiyle çalışmakatadır. Klasik yaklaşımda her $x_i$ ölçümünün Mahalanobis uzaklığı aşağıdaki şekilde hesaplanmaktadır.

$$MD(x_i) = \sqrt{(x_i - \hat{\mu}_{MLE})' \hat{\Sigma}^{-1}_{MLE} (x_i - \hat{\mu}_{MLE})} \tag{4.1}$$

2) Bu konuda diğer bir yaklaşım ise; herbir grup için standartlaştırılmamış kanonik skorlara ilişkin "Mahalanobis uzaklıkları"nın gruplara ilişkin varyans-kovaryans matrislerinin homojen olduğu varsayımı altında hesaplanmasıdır. Bu durumda, hesaplamalarda kullanılacak $\Sigma$ matrisi $min(k-1, p)$ kadar kanonik fonksiyon için herbir gruptan elde edilen "Mahalanobis uzaklıkları", $min(k-1, p)$ serbestlik dereceli bir $\chi^2$ dağılımını gösterir. Ancak her iki yaklaşımda da yanılma düzeyinin $\alpha = 0.01$ gibi küçük seçilmesi önerilir.

Ancak birden fazla aykırı gözlem olması durumunda, aykırı gözlemlerin düzenli gözlem gibi görünmesi (maskeleme) ve düzenli gözlemlerin de aykrı gözlem gibi görünmesi (süpürme etkileri) söz konusu olabilir. ve Mahalanobis uzaklığı bu durumda doğru sonuç vermeyebilir. Örneğin; bir gözlem biriminin gruplara atanma sonsal olasılıklarının hesaplanması neticesinde elde edilen skora göre atandığı grup, bu gözlemin grup merkezine olan MD'nin büyüklüğüne göre potansiyel olarak aykırı bir gözlem olup olmadığı söylenebilir. Dolayısıyla, bu gözlem gerçekte kendi bulunduğu gruba atanmış olsa bile, MD'nin büyük çıkması bu gözlemin tipik bir biçimde kendi grup gözlemi olmadığı kuşkusunu uyandırır (Hubert ve Debruyne, 2010).

# 5 . SAĞLAM TAHMİN EDİCİLER

## 5.1. Kavram olarak ”Sağlamlık”

[1]Bir ölçme aracıyla elde edilen ölçümlerin yüksek güvenirlik ve geçerlik düzeyinde kazandığı dayanıklılığa ilişkin süreğenlik [2] anlamına gelmektedir.

## 5.2. İstatistiksel açıdan ”Sağlamlık”

En çok olabilirlik yöntemine dayalı olarak elde edilen MLE tahmin edicileri, aykırı gözlem ya da gözlem gruplarına karşı oldukça duyarlıdır. Aykırı gözlemler içeren veri setleri ile çalışırken karar vericinin, elde edilen bulguların politika amacıyla kullanılabilmesini mümkün kılan yüksek güvenilirlikte bulgulara ulaştığı sonucuna varması pek mümkün değildir. Örneğin, tek değişkenli bir yapıda verinin lokasyon parametresi hakkında öngörüde bulunan örnek ortalaması istatistiği, veri seti içerisinde kabaca büyük bir gözlemin varlığından önemli derecede etkilenir. Aynı şekilde varyans istatistiğinin değeri de bu durumdan önemli derecede etkilenecektir. Bu sebeple karar vericinin başka türde tahmin edicilerle çalışmayı yeğlemesi ancak, aykırı gözlemlerin etkisini tamamen tahminleme sürecinin dışında tutulduğu ya da aykırı gözlemlerin tahminleme sürecinde işin içerisinde olduğu, fakat etkisinin az tutulduğu başka türde tahmin edicilerle olacaktır.

Tahminleme sürecinde aykırı gözlemlerin dışlandığı durumda elde edilen klasik tahmin edicilerle parametre tahmini yapmanın dezavantajları oldukça fazladır. Örneğin aykırı gözlem niteliği kazandığı düşünülen bir gözlem biriminin veri setinden çıkarılması, elde edilecek test istatistiği değerini ve varsa bu istatistiğe ilişkin serbestlik derecesini değiştirecektir. Bu durum gerçekte kabul ya da red edilmesi beklenen yokluk hipotezinin zıttı yönde üstelik çok yüksek bir güvenilirlikle karara bağlanmasına sebep olacak ve parametre öngörülerimiz sağlıklı olmayacaktır.

Tek değişkenli yapı içerisinde lokasyon parametresinin tahmininde örnek ortalaması istatistiği yerine; medyan istatistiği, varyans istatistiği yerine ise gözlemlerin medyanından mutlak farkı şeklinde tanımlı yeni rassal değişkenin medyanı olarak tanımlanan MAD istatistiğinin kullanılması sağlıklı olasılık cümleleri ve güven aralıkları kurulması bakımından önemlidir. Bu türde tahmin edicilerin en önemli vasfı klasik tahmin edicilerde aranan özelliklere haiz olmasının yanı sıra, etkinlik özelliği bakımından klasik tahmin edicilere yakın, aykırı gözlem ya da başka bir dağılımdan karışan gözlem olması durumundan çok fazla etkilenmeyen ve yüksek kırılma noktasına sahip tahmin ediciler olmasıdır. Bu türde tahmin edici elde etmek için kullanılan yöntemlere sağlam istatistiksel

[1] *Yöntembilimleri Sözlüğü*,Türk Dil Kurumu, 1981, Ankara

[2]Süreğenlik: Ne kadar süreceği belli olmaksızın sürüp giden, müzmin, kronik

yöntem, elde edilen tahmin edicilere ise sağlam istatistik denilmektedir (Maronna ve diğerleri, 2006:1:5).

### 5.3. Sağlam tahmin ediciler için değerlendirme kriterleri

MSE'nin yanı sıra, bu tahmin edicilerin sağlamlık açısından değerlendirilmesine olanak sağlayan ve tek bir noktanın tahmin ediciye olan etkisini ölçen duyarlılık eğrisi (sensivity curve(SC)), duyarlılık eğrisinin limit durumu olarak düşünülen etki fonksyionu (influence function(IF)) ve tahmin edicinin maksimum ne kadar karışma oranına dayanabileceğinin ölçüsü olarak kullanılan kırılma noktası (breakdown point(BP)) sağlam istatistiksel yöntemler sonucu elde edilen tahmin edicinin genel olarak değerlendirilmesinde kullanılan önemli ölçütlerdir (Maronna ve diğerleri, 2006:55).

#### 5.3.1. Etki fonksiyonu (Influence Function (IF))

IF olarak adlandırıdığımız fonksiyon, sağlamlık karakteristiğinin bir ölçüsü olarak gösterilebilir. Her tahmin edici için bir etki fonksiyonu vardır. Gerçek veri setine aykırı gözlem veya başka bir dağılımdan karışan bir gözlemin tahmin edici üzerindeki etkilerinin, neler olduğunu ortaya koymada önemli bir rol oynayan fonksiyon olarak karşımıza çıkar.

IF, örnek çarpının büyük olduğu durumda, dışardan dağılıma karışan bir gözlem ya da gözlem grubunun tahmin ediciye olan etkisinin ölçüsü olarak tanımlanmaktadır. IF, dağılım fonksiyonu ve bu dağılım fonksiyonu üzerinde tanımlı lineer fonksiyonel ve dışardan dağılıma karışan $x_0$ gözlemine ait etkinin bir fonksiyonudur. Bir başka ifadeyle, dışardan karışan $x_0$ gözleminin, dağılım fonksiyonu üzerinden hesaplanan lineer fonksiyonel üzerinde ne gibi değişiklikler gösterdiğini ifade eden fonksiyon IF'dir. Bu sebeple IF, bir tahmin ediciye ilişkin etkileri gözlemlerken aynı zamanda aykırı gözlem veya başka bir dağılımdan karışan bir gözlem olması durumunu da hesaba katar.

IF'nin bazı özelliklere sahip olması istenir. Veri setine aykırı gözlem veya başka bir dağılmdan gözlem karışması durumunda IF'nin değerinin çok değişmediği tahmin ediciler tercih edilir. Bir tahmin ediciye ilişkin etki fonksiyonu sınırlı (bounded) kalıyorsa, o tahmin edicinin IF'nu; aykırı gözlem ya da başka bir dağılımdan karışan gözlem ne olursa olsun sınırlı kalır.

$X$, rassal değişkeni üzerinde tanımlanan dağılım fonksiyonu $F$ ve $\varepsilon$ küçük bir sayı olmak üzere ($0<\varepsilon<1$); $F_\varepsilon$, yani $\varepsilon$ oranda karışmış dağılım fonksyionu,

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{x_0}, \tag{5.1}$$

biçiminde ifade edilir (Maronna ve diğerleri, 2006:55). Burada, $\Delta_{x_0} = \begin{cases} 1, & x > x_0 \\ 0, & x \leq .x_0 \end{cases}$ şeklinde tanımlı basamak fonksiyonudur. $T$ lineer fonsiyonel olmak üzere, $T$ için etki fonksiyonu aşağıdaki gibi tanımlanır:

$$IF(x_0; T, F) = \lim_{\varepsilon \to 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = T'_\varepsilon(F). \tag{5.2}$$

*Örnek:* $Y \sim N(0,1)$ olsun. $x_0$ ise $Y$ rassal değişkeninin dağılımına başka bir dağılımdan karışmış gözlem birimi olmak üzere, karışmanın olduğu dağılım fonksiyonunun beklenen değeri Eş.5.1 (1. momenti) aşağıdaki gibi ifade edilir:

$$F_\varepsilon = (1 - \varepsilon)F + \varepsilon\Delta_{x_0} \tag{5.3}$$

$$E[F_\varepsilon] = E[(1 - \varepsilon)F + \varepsilon\Delta_{x_0} \tag{5.4}$$

$$(1 - \varepsilon)E[F] + \varepsilon = E[\Delta_{x_0}] = (1 - \varepsilon) \cdot 1 + \varepsilon x \tag{5.5}$$

Bu durum için Eş. 5.2'den etki fonksiyonu,

$$\begin{aligned} IF(x_0; E, F) &= \frac{\partial((1 - \varepsilon) \cdot 1 + \varepsilon x)}{\partial \varepsilon} \\ &= -1 + x_0 \end{aligned} \tag{5.6}$$

şeklinde olmaktadır. $IF(x_0; E, F) = \infty$ olduğunda, beklenen değer lineer fonksiyonelinin (1. momente ilişkin) etki fonksyionu yazılamaz (Arık, 2014). Yani dağılım fonksiyonu üzerinde tanımlı linner fonksiyonelin sınırlı kalmadığını gösterir.

### 5.3.2. Duyarlılık eğrisi (Sensivity Curve (SC))

SC, veriye aykırı gözlem ya da başka bir dağılımdan gözlem ya da gözlem grubu karıştığı durumda, tahmin edicinin bundan nasıl etkilendiğini görmek bakımından önemlidir. Aykırı gözlem dahil edilerek hesaplanan tahmin edicinin değeri ile aykırı gözlem hariç tutularak hesaplanan tahmin edicinin değeri arasındaki fark, duyarlılık eğrisinin oluşumuna zemin hazırlar. Örneğin; $T_n$ ; $(x_1, ..., x_n)$ gözlemleri üzerinde, $T_{n+1}$ bu gözlemlere $x_0$ gözlemi eklendiğinde tanımlı lineer fonksiyoneller olmak üzere, karışma oranı $\frac{1}{n+1}$ için duyarlılık eğrisi,

$$\begin{aligned} SC_n(x_0) &= \frac{T_{n+1}(x_1, ..., x_n, x_0) - T_n(x_1, ..., x_n)}{\frac{1}{n+1}} \\ &= (n + 1)[T_{n+1}(x_1, ..., x_n, x_0) - T_n(x_1, ..., x_n)], \end{aligned} \tag{5.7}$$

şeklinde ifade edilir (Maronna ve diğerleri, 2006:55:57). Örnek ortalaması ve medyan istatistiklerinin duyarlılıkları kolayca hesaplanabilir. Örneğin, belli bir dağılımdan alınan $(X_1, X_2, ..., X_n)$ rassal örneğine farklı bir $Y$ gözlemi eklensin ve ortalamanın duyarlılık eğrisine olan etkisi inceleniyor olsun. O halde;

$$T_n = \overline{X}_n$$
$$T_{n+1} = \frac{1}{n+1}(n\overline{X}_n + Y)$$
$$SC_n(Y) = (n+1)\left[\left(\frac{n}{n+1} - 1\right)\bar{X}_n + \frac{1}{n+1}Y\right]$$
$$SC_n(Y) = (n+1)\left[\frac{1}{n+1}(Y - \bar{X}_n)\right]$$
$$SC_n(Y) = Y - \bar{X}_n$$
$$SC_n(Y) = sup|Y - \bar{X}_n|$$

$SC_n(Y) = sup|Y - \bar{X}_n| = \infty$, ifadesi elde edilen bu duyarlılığın en küçük üst sınırının sonsuz olacağını söyler. Bu yüzden, örnekleme farklı bir değerin eklenmesi sonucunda ortalama istatistiğinin sınırsız duyarlılığa sahip olabileceği görülmektedir.

Örneğin, $n = 2m + 1$ ve $X_{(1)} \leq X_{(2)} \leq, ..., \leq X_{(n)}$ sıralı gözlem değerleri için medyan istatistiğini ele alalım. $T_n = T_n(X_{(1)}), X_{(2)}, ..., X_{(n)}) = X_{(m+1)}$ olur. $Y$ gözlemi eklendiğinde, $T_n = T_n(X_{(1)}, X_{(2)}, ..., X_{(n)}, Y)$ tahmin edicisinin değeri, $Y$'nın örneklem içindeki konumuna göre aşağıdaki değerleri alır:

$$T_{n+1} = \begin{cases} \frac{X_{(m)}+X_{(m+1)}}{2}, & Y < X_{(m)} \\ \frac{X_{(m+1)}+X_{(m+2)}}{2}, & Y > X_{(m+2)} \\ \frac{Y+X_{(m+1)}}{2}, & X_{(m)} \leq Y \leq X_{(m+2)} \end{cases}$$

$Y$ gözleminin eklenmesi durumunda,

$$SC_n(Y) = \begin{cases} \frac{X_{(m)}-X_{(m+1)}}{2}, & Y < X_{(m)} \\ \frac{X_{(m+2)}-X_{(m+1)}}{2}, & Y > X_{(m+2)} \\ \frac{Y-X_{(m+1)}}{2}, & X_{(m)} \leq Y \leq X_{(m+2)} \end{cases}$$

şeklinde olur. Görüldüğü gibi duyarlılık eğrisinin değerlerinin hepsi sonlu değerlerdir. Medyan tahmin edicisi sınırlı duyarlılık eğrisine sahiptir (Jureckova ve Picek, 2006).

### 5.3.3. Kırılma noktası (Breakdown Point (BP))

Bir tahmin edicinin kırılma noktası, tahmin edicinin gerçek parametre hakkında bilgi vermesini engellemeyecek maksimum karışma oranıdır (Maronna ve diğerleri, 2006:58). $T_n$, $X = (x_1, x_2, \cdots x_n)$ gözlemleri üzerinde tanımlı bir tahmin edici olsun. Gözlemlerin

$m$ $(m < n)$ tanesi, aykırı gözlemler ile değiştirilsin. Yeni oluşan gözlemlerin üzerinde hesaplanan yeni bir tahmin edici $T_n(Y)$ olarak elde edilmiş olsun. $T_n$ tahmin edicisi için kırılma noktası,

$$\varepsilon^*(X,T) = \inf\{\varepsilon : \sup || (T_n(Y) - T_n(X)|| = \infty\} \tag{5.8}$$

biçiminde ifade edilmektedir (Huber ve Ronchetti, 2009:281). Örneğin, medyan istatistiğinin kırılma noktası % 50'dir.

### 5.4. Çok değişkenli lokasyon ve ölçek parametrelerinin sağlam tahmin edicileri

Tek değişken durumunda olduğu gibi Eş.3.4 ve Eş. 3.6'da tanımlı $\overline{x}_{MLE}$ ve $\Sigma_{MLE}$ tahmin edicilerinin kırılma noktası %'0 dır Yani, veri grubu içinde aykırı gözlemlerin payı çok küçük olsa bile bu tahmin edicilerin değeri aşırı derecede etkilenmektedirler. Veri setine karışmış aykırı gözlemler, bu istatistiklerin aldığı değerin çok büyük ya da çok küçük olmasına yol açabilmektedir. Bu bağlamda aykırı değerlerin etkisinin düşük tutulduğu ve kırılma noktası bakımından yüksek başka türde tahmin edicilere gereksinim vardır (Hubert ve Debruyne, 2010). Bu çalışmada ilgili özellikleri taşıyan ve çok değişkenli sağlam tahmin ediciler elde etme sürecinde, $n$ gözlem biriminden h çaplı alt kümeler seçilerek, varyans-kovaryans matrisinin determinantının minimum olmasına dayalı algoritma sonucu elde edilen, lokasyon ve ölçek tahmin edicilerinin sağlamlık özelliğine sahip tahmin edicilerine yer verilmiştr.

#### 5.4.1. Minimum kovaryans determinantı algoritması (MCD yöntemi)

Lokasyon ve ölçek parametrelerinin sağlamlık özelliğine sahip tahmin edicileri, kovaryans matrisinin determinantının minimum olmasına dayalı Rousseeuw (1999)'in MCD algoritması kullanılarak elde edilmektedir. MCD algortimasının temel çalışma prensibi; gözlemlerden hangi $h$ tanesi için hesaplanacak determinant değerinin en küçük olacağını tespit etmek üzerine kuruludur (Hubert ve Debruyne, 2010)

MCD algortimasının adımları:

1) En küçük kovaryans determinanatı metodu $p$ boyutlu $n$ çaplı gözlem grubu içinden $h$ tane gözlem rassal olarak seçilmektedir.

2) Dışardan karışmanın büyük oranda olduğu beklentisi var ise bu durumlarda $h$, $0,5n$'e yakın alınmalıdır. Dışardan karışmanın yüksek oranlı olmadığı durumlarda, küçük çaplı örnek için daha yüksek etkinlik elde etmek üzere; $h$ için ara değerler (Örneğin; $0,75n$) alınması önerilir.

3) $\binom{n}{h}$ tane kovaryans matrisleri ve determinat değerleri elde edilir.

4) Elde edilen $\binom{n}{h}$ tane determinant içinde en küçüğü belirlenir, böylece en küçük determinant değerini veren $h$ tane gözlem tespit edilmiş olur.

5) Başlangıç tahminleri, en küçük determinant değerini veren $h$ gözlem için yapılır: $\widehat{\mu_0},\ \widehat{\Sigma_0}$.

6) Bütün gözlemler ($n$ sayıda gözlem) üzerinden sağlamlık özelliğine sahip uzaklık (Sağlam Uzaklık (RD)) Eş.5.9'deki gibi tanımlanır:

$$RD_{x_i}(x_i, \widehat{\mu}_0, \widehat{\Sigma}_0) = \sqrt{(x_i - \widehat{\mu}_0)^T \widehat{\Sigma}_0^{-1} (x_i - \widehat{\mu}_0)} \leq \sqrt{\chi^2_{p,0.975}}$$
$$RD_{x_i}(x_i, \widehat{\mu}_0, \widehat{\Sigma}_0) = \sqrt{(x_i - \widehat{\mu}_0)^T \widehat{\Sigma}_0^{-1} (x_i - \widehat{\mu}_0)} > \sqrt{\chi^2_{p,0.975}} \tag{5.9}$$

Bu uzaklık, $\widehat{\mu_0}$ ve $\widehat{\Sigma_0}$ metriğine göre her gözlem birimi için hesaplanmaktadır. Gözlem noktaları, $\widehat{\mu_0}$ merkezine uzaklıkları bakımından ikiye ayrılır. $\chi^2$ kontürünün içine düşen gözlemlere 1 ataması, dışına düşen gözlemlere de 0 ataması yapılarak bütün gözlemler ağırlıklandırılır. Bir başka ifade ile, veri merkezine uzaklığı $\sqrt{\chi^2_{p,0.975}}$'den küçük olan gözlemler için $w_i = 1$ ağırlığı atanmakta, diğerlerine $w_i = 0$ atanmaktadır.

7) Eş.5.9'de tanımlanan sağlam uzaklıkla oluşan elips kontürünün içine düşen gözlem değerleri için yani, $w_i = 1$ ağırlığına sahip gözlemler için tekrar örnek ortalaması ve varyansı istatistikleri aşağıdaki gibi elde edilir:

$$\widehat{\mu_w} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{5.10}$$
$$\widehat{\Sigma_w} = \frac{\sum_{i=1}^{n} w_i \left(x_i - \widehat{\mu_w}\right)\left(x_i - \widehat{\mu_w}\right)^T}{\sum_{i=1}^{n} w_i - 1} \tag{5.11}$$

8) Bütün gözlem değerleri için sağlamlık özelliğine sahip uzaklık tekrar yeni tahmin değerlerine göre belirlenir:

$$RD_{x_i}(x_i, \widehat{\mu}_w, \widehat{\Sigma}_w) = \sqrt{(x_i - \widehat{\mu}_w)^T \widehat{\Sigma}_w^{-1} (x_i - \widehat{\mu}_w)} \leq \sqrt{\chi^2_{p,0.975}}$$
$$RD_{x_i}(x_i, \widehat{\mu}_w, \widehat{\Sigma}_w) = \sqrt{(x_i - \widehat{\mu}_w)^T \widehat{\Sigma}_w^{-1} (x_i - \widehat{\mu}_w)} > \sqrt{\chi^2_{p,0.975}} \tag{5.12}$$

Bu uzaklık, $\widehat{\mu_w}$ ve $\widehat{\Sigma_w}$ metriğine göre her gözlem birimi için hesaplanmaktadır. Gözlem noktaları, $\widehat{\mu_w}$ merkezine uzaklıkları bakımından ikiye ayrılır. $\chi^2$ kontürünün içine düşen gözlemlere yani, veri merkezine uzaklığı $\sqrt{\chi^2_{p,0.975}}$'den küçük olan gözlemler için $w_i = 1$ ağırlığı atanmakta, diğerlerine $w_i = 0$ atanmaktadır.

9) Elpisin merkezi sabit kalıncaya kadar bu işlem devam eder ve tahmin edilen ortalama ve varyans aynı kalmaya devam ettiğinde işlem durdurulur.

Böylece MCD algoritmasının sonucunda ağırlıklı ortalama ve ağırlıklı varyans-kovaryans matrisi aşağıdaki gibi tahmin edilir:

$$\hat{\mu}_{MCD} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{5.13}$$

$$\hat{\Sigma}_{MCD} = \frac{\sum_{i=1}^{n} w_i (x_i - \hat{\mu}_w)(x_i - \hat{\mu}_w)^T}{\sum_{i=1}^{n} w_i - 1}. \tag{5.14}$$

En küçük kovaryans determinant tahmin edicisi ve onun ağırlıklandırılmış versiyonunda etki fonksiyonu sınırlıdır. Bu tahmin edicinin kırılma noktası değeri $(\frac{n-h+1}{n})$'dir. Tahmin edicinin sağlamlık karakterini bu söz konusu $h$ sayısı belirler. En küçük kovaryans determinant tahmin edicisi $h = \frac{n+p+1}{2}$ için kırılma noktası değeri mümkün en yüksek değerine erişir.

## 5.5. Çok değişkenli tahmin edici elde etme sürecinde Minimum Kovaryans Determinantına dayalı bazı algoritmalar

MCD yoluyla elde edilen tahmin ediciler sağlamlık karakteristiğini yansıtıyor olmasına rağmen, özellikle $h$'ın maksimum kırılma noktası değerine ulaştığı normal modeller için etkin değildir. Fakat bu yolla elde edilen tahmin ediciler düşük etkinlik sağlamasına rağmen, uygulamada en çok tercih gören tahmin edicilerdir. Bunun da sebebi; bazı istatistik tabanlı programlarda MCD yoluyla elde edilen tahmin edicilere ulaşmanın sağladığı kolaylıktır. Bu nedenle MCD metoduna alternatif olarak, MCD'ye dayalı çeşitli algoritmalar geliştirilmiştir. Bu algoritmalarla elde edilen tahmin ediciler sağlamlık karakteristiği açısından farklı nitelikler taşımaktadırlar.

### 5.5.1. MCD-A algortiması

MCD-A algoritmasında temel mantık, gruplarda aykırı değer tespiti yapıldıktan sonra, herbir grup için aynı budama şekli ve oranını koruyarak klasik MCD algortimasının adımlarını uygulamaktır.

MCD-A Algoritmasının Adımları :

1) Başlangıç değerleri için MCD'den farklı olarak grup ortalamaları $m_k^0$ ve ortak varyans-kovaryans matrisi $C_0$ belirlenir ve algoritmanın geri kalan kısmı MCD metodundaki gibi gözlem birimlerine yeniden ağırlıklandırma yapılarak yürütülür.

2) Ancak burada MCD algoritmasından farklı olarak ortak varyans kovaryans matrisinin elde edilmesi gerekmektedir. Yüksek kırılma noktasına sahip ortak varyans-kovaryans matrisinin tahmini için en kolay yaklaşımlardan birisi bireysel gruplardan hareketle grup ortalamalarını ve grup varyans-kovaryans matrisini $(m_k, C_k)$ $(k = 1, 2, ..., g)$ tahmin etmek, daha sonra birleştirilmiş ortak varyans-kovaryans matrisini oluşturmaktır:

$$C_{MCD-A} = \frac{\sum_{k=1}^{g} n_k C_k}{\sum_{k=1}^{g} n_k - g}. \tag{5.15}$$

MCD-A metodunu bu şekilde uygulamanın dezavantajı tüm gruplara aynı budama oranının uygulanmasıdır. Bazı gruplarda aykırı gözlem olmamasına rağmen, tüm gruplara budama yapılırsa bu işlem sonucu elde edilen tahmin edicilerde etkinlik kaybı meydana gelecektir (Todorov ve Pires, 2007).

### 5.5.2. MCD-B algoritması

Diğer bir method, He ve Fung (2000) tarafından önerilen $S$ metodunun, Hubert ve Van Driessen (2004) tarafından MCD metoduna uyarlanmış şeklidir ve MCD-B olarak adlandırılmaktadır. Bu metotta, birleştirilmiş varyans-kovaryans matrisi yerine, tek bir gruptan tahmin edilen varyans-kovaryans matrisini merkezileştirilen ve birleştiren gözlemler kullanılır.

MCD-B Algoritmasının Adımları

1) Bu metoda, her bir grubun yeniden ağırlıklandırılmış MCD lokasyon tahmin edicisi olarak bireysel grup lokasyon tahmin edicileri $t_k$, $k = 1, 2, \cdots, g$ elde edilerek başlanır.

2) Bu grup ortalamaları merkezileştirilen gözlemleri elde etmek için orjinal gözlemlerle yer değiştirir.

$$Z = z_{ik}, \qquad z_{ik} = x_{ik} - t_k \tag{5.16}$$

3) Ortak varyans-kovaryans matrisi $C$, merkezileştirilen $Z$ gözlemlerinin yeniden ağırlıklandırılmış MCD varyans-kovaryans matrisi olarak tahmin edilir.

4) $Z$'nin lokasyon tahmin edicisi $\delta$, grup ortalamaları $m_k$'yı düzeltmek için kullanılır ve böylece nihahi grup ortalamaları;

$$m_k = t_k + \delta \tag{5.17}$$

olacak şekilde elde edilir. Bu süreç yakınsama olana kadar tekrar edilir (Todorov ve Pires, 2007).

### 5.5.3. MCD-C algoritması

Hawkins ve McLachlan (1997) tarafından gruplar içi minimum varyans-kovaryans determinantına dayalı olarak tanımlanan yöntemin çok yoğun iş hacmi gerektirmesi, Hubert ve Van Driessen (2004)'in bu algoritmayı modifiye ederek (FAST-MCD metodunun avantajlarını kullanarak) yeni bir metot geliştirmesine yol açmıştır, ancak bu metotta da her bir grup için MCD hesaplaması gerekmektedir.

MCD-C Algortimasının Adımları :

1) MCD metodu ile elde edilen $m_k^0$ ve $C_0$ tahmin edicileri kullanılarak, başlangıç sağlam uzaklıklar hesaplanır:

$$RD_{ik}^0 = \sqrt{(x_{ik} - m_k^0)^t C_0^{-1} (x_{ik} - m_k^0)} \tag{5.18}$$

2) Hesaplanan başlangıç sağlam uzaklıklar ile her bir gözlem için (yani, $x_{ik},\ i = 1, ..., n_k,\quad k = 1, ..., g$ için) bir ağırlık tanımlanır:

$$w_{ik} = \begin{cases} 1 & RD_{ik}^0 \leq \sqrt{\chi^2_{p,0.975}} \\ 0 & d.h \end{cases} \tag{5.19}$$

3) Bu ağırlıklar ile nihai grup ortalamaları $m_k$'nın yeniden ağırlandırılmış tahmin edicileri ve sağlam sınıflama kuralları için gerekli olan ortak gruplar içi varyans-kovaryans matrisi $C$ tahmin edilir.

$$m_{k_{MCD-C}} = \frac{\sum_{i=1}^{n_k} w_{ik} x_{ik}}{v_k}$$

$$C_{MCD-C} = \frac{1}{v-g} \sum_{k=1}^{g} \sum_{i=1}^{n_k} w_{ik} (x_{ik} - m_k)(x_{ik} - m_k)^t \tag{5.20}$$

Burada, $v_k$ yani, $v_k = \sum_{i=1}^{n_k} w_{ik}$, $(k = 1, ..., g)$; $k$. grup içi ağırlıklar toplamı ve $v$, yani $v = \sum_{k=1}^{g} v_k$ ise toplam ağırlık toplamını ifade etmektedir (Todorov ve Pires, 2007).

# 6 . UYGULAMA

Bu bölümde, öncelikle klasik ve sağlam doğrusal diskriminant analizi uygulamasının detaylı olarak açıklanması; daha sonra da gerçek veri seti üzerinde, aykırı gözlem ya da başka bir dağılımdan gözlem karışmış olması durumunda, sağlam tahmin edicilerle en çok olabilirlik tahmin edicilerinin karşılaştırılması amaçlamakta olup, üç alt kısımdan oluşmaktadır.

İlk olarak, küçük bir veri seti üzerinde, klasik doğrusal diskriminant analizi adım-adım uygulanmıştır. Sonra aynı veri seti manipüle edilerek, bazı gözlemlerin aykırı değer niteliği kazanması sağlanmış ve tekrar klasik doğrusal diskriminant analizi uygulanmıştır. Bu iki durum arasında hatalı sınıflandırma oranları bakımından, bir farklılık olup olmadığı incelenmiştir. Ayrıca, bu iki farklı küçük veri seti için klasik ve sağlam uzaklıklar manuel (elle) olarak hesaplanmış ve elde edilen sonuçlara ilişkin karşılaştırmalar yapılmıştır (Syed ve diğerleri, 2017).

İkinci olarak, 75 gözlem içinde 14 aykırı gözlem barındıran ve maskeleme etkisinin örneklendiği Hawkins, Bradu ve Kass (1984) veri seti üzerinde klasik ve sağlam tahmin edicilerle gözlemlerin, uzaklıklar bakımından karşılaştırmalı uygulaması yapılmıştır.

Üçüncü olarak, Alcohol veri seti [1] üzerinde klasik ve sağlam tahmin ediciler kullanılarak klasik doğrusal diskriminant üzerinde gözlem birimlerinin hatalı sınıflandırılma oranları elde edilmektedir (Syed ve diğerleri, 2016).

Özellikle, ikinci ve üçüncü uygulama R paket programı ve bu program içinde yer alan rrcov ve robustbase kütüphaneleri kullanılarak yapılmıştır. (Alrawashdeh ve diğerleri, 2018)

## 6.1. Kalite kontrol veri seti

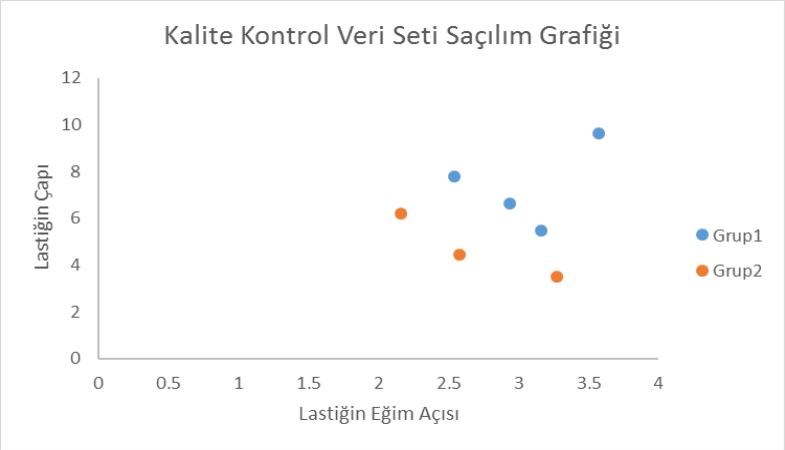
Bir fabrikada çalışan mühendisler ideal lastik üretiminde iki değişkenin (lastiklerin eğiminin ve çapının) lastiklerin kalitesinin üzerinde anlamlı bir etkiye sahip olduğunu tespit etmiş, bu kapsamda gün içerisinde üretilen lastiklerden 7 tanesini (gözlem birimini) rastgele seçmiş ve iki değişken bakımından incelemesini yapmışlardır [2]. İncelemesi yapılan 7 lastikten 4 lastik ölçüm değerleri açısından kalite kontrolü aşamasından geçerken, sadece 3 adet lastiğin bu süreçte başarılı olmadığı gözlemlenmiştir. İlgili veri seti Çizelge 6.1'de gösterilmiştir.

[1] http://data-mining-tutorials.blogspot.com/search?q =alcohol

[2] Bu veri seti ve senaryo uygulama amacıyla keyfi olarak üretilmiştir.

Çizelge 6.1. Kalite Kontrol Veri Seti

| Testi Geçenler (Grup 1) | | Testi Geçemeyenler (Grup 2) | |
|---|---|---|---|
| Lastiğin Eğim Açısı ($x_1$) | Lastiğin Çapı ($x_2$) | Lastiğin Eğim Açısı ($x_1$) | Lastiğin Çapı ($x_2$) |
| 2.94 | 6.62 | 2.58 | 4.46 |
| 2.54 | 7.80 | 2.16 | 6.22 |
| 3.57 | 9.65 | 3.27 | 3.52 |
| 3.16 | 5.47 | | |




Şekil 6.1. Kalite Kontrol Veri Seti Saçılım Grafiği

Bir mühendis ilgili veri seti üzerinde bir gözlem biriminin gerçekte olması gereken gruptan bir başka gruba atanıp atanmayacağı hususunda kestirimde bulunmak istemekte ve bu amaçla ilgili veri seti üzerinde klasik doğrusal diskriminant analizi tekniğini uygulamayı uygun bulmaktadır. Aşağıda kalite Kontrol Veri Seti üzerinde klasik doğrusal diskriminanat analizinin manuel (elle) uygulaması yapılmıştır.

### 6.1.1. Kalite kontrol veri seti üzerinde klasik doğrusal diskriminant analizinin manuel (elle) uygulanması

1) Kalite kontrol veri setinde, $X$ bağımsız değişken matrisini $(x_i, i = 1, 2, \cdots, 7)$ (lastiğin eğim açısı ve lastiğin çapı) ve $Y$ bağımlı değişken vektörünü (Grup1, Grup2) göstermektedir:

$$X = \begin{bmatrix} 2.94 & 6.62 \\ 2.54 & 7.80 \\ 3.57 & 9.65 \\ 3.16 & 5.47 \\ 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix}, \qquad Y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}. \tag{6.1}$$

2) Örneğin, $x_3$ gösterimi bağımsız değişken matrisindeki 3. gözlemi ($x_3 = [3.57,\ 9.65]$); $x_7$ gösterimi bağımsız değişken matrisindeki 7. gözlemi ($x_7 = [3,27,\ 3.52]$) ifade etmektedir.

3) $X_1$ bağımsız değişken matrisindeki 1. grubu, $X_2$ bağımsız değişken matrisindeki 2. grubu temsil etmektedir:

$$X_1 = \begin{bmatrix} 2.94 & 6.62 \\ 2.54 & 7.80 \\ 3.57 & 9.65 \\ 3.16 & 5.47 \end{bmatrix}, \qquad X_2 = \begin{bmatrix} 2.58 & 4.46 \\ 2.16 & 6.22 \\ 3.27 & 3.52 \end{bmatrix}. \tag{6.2}$$

4) $\hat{\mu}$, gruplara ilişkin genel ortalama vektörü tahmin edicisini; $\hat{\mu_1}$ ve $\hat{\mu_2}$ sırasıyla, bağımsız değişken matrisindeki 1. gruba ilişkin ortalama vektörü ve bağımsız değişken matrisindeki 2. gruba ilişkin ortalama vektörü tahmin edicisini ifade etmektedir:

$$\hat{\mu_1} = [\ 3.05 \quad 7.38\ ], \qquad \hat{\mu_2} = [\ 2.67 \quad 4.73\ ], \qquad \hat{\mu} = [\ 2.88 \quad 6.24\ ]. \tag{6.3}$$

5) $x_1^0$, 1. grupta yer alan gözlem birimlerinin, genel grup ortalamasından farkı şeklinde tanımlı yeni rassal değişken matrisini ve $x_2^0$, 2. grupta yer alan gözlem birimlerinin genel grup ortalamasından farkı şeklinde tanımlı yeni rassal değişken matrisini ifade etmektedir:

$$x_1^0 = \begin{bmatrix} 2.94 - 2.88 & 6.62 - 6.24 \\ 2.34 - 2.88 & 7.80 - 6.24 \\ 3.57 - 2.88 & 9.65 - 6.24 \\ 3.16 - 2.88 & 5.47 - 6.24 \end{bmatrix}, \qquad x_2^0 = \begin{bmatrix} 2.58 - 2.88 & 4.46 - 6.24 \\ 2.16 - 2.88 & 6.22 - 6.24 \\ 3.27 - 2.88 & 3.52 - 6.24 \end{bmatrix}. \tag{6.4}$$

6) $c_i$, grup içi varyans-kovaryans matrislerini ifade etmektedir:

$$c_i = \frac{(x_i^0)^T \cdot x_i^0}{n_i} \tag{6.5}$$

$$c_1 = \begin{bmatrix} 0.185 & 0.280 \\ 0.280 & 3.185 \end{bmatrix}, \qquad c_2 = \begin{bmatrix} 0.314 & -0.731 \\ -0.731 & 1.879 \end{bmatrix}, \tag{6.6}$$

7) C(r,s) ise birleştirilmiş grup içi varyans-kovaryans matrisini ifade etmektedir:

$$C(r,s) = \frac{1}{n} \sum_{i=1}^{g} n_i \cdot c_i(r,s), \tag{6.7}$$

$$\begin{aligned} C(1,1) &= \frac{4}{7} \cdot 0.185 + \frac{3}{7} \cdot 0.314 = 0.237, \\ C(2,2) &= \frac{4}{7} \cdot 0.280 + \frac{3}{7} \cdot -0.731 = -0.124, \\ C(1,2) &= \frac{4}{7} \cdot 3.185 + \frac{3}{7} \cdot 1.879 = 2.662, \end{aligned} \tag{6.8}$$

$$C = \begin{bmatrix} 0.237 & -0124 \\ -0.124 & 2.662 \end{bmatrix}, \qquad C^{-1} = \begin{bmatrix} 4.11 & -0.19 \\ -0.19 & -0.36 \end{bmatrix}. \tag{6.9}$$

8) $p_i$, herbir gözlemin dahil olduğu grupta (i. grupta) bulunması olasılığıdır:

$$p_i = \frac{n_i}{n}, \tag{6.10}$$

$$p = \begin{bmatrix} \frac{4}{7} \\ \frac{3}{7} \end{bmatrix} = \begin{bmatrix} 0.571 \\ 0.429 \end{bmatrix}. \tag{6.11}$$

9) Elde edilen bilgiler Fisher'in kanonik diskriminant fonksiyonunda yerine yazılır:

$$f_i = \mu_i C^{-1} \cdot x_k^T - \frac{1}{2}\mu_i C^{-1} \cdot \mu_i^T + \ln(p_i). \tag{6.12}$$

10) Grup Sayısı kadar diskriminant (ayrıcı) fonksiyonu olacaktır:

$$f_1 = -35.946 + 14.723\,\text{Lastiğin Eğim Açısı} + 3.462\,\text{Lastiğin Çapı}, \tag{6.13}$$

$$f_2 = -23.012 + 12.529\,\text{Lastiğin Eğim Açısı} + 2.363\,\text{Lastiğin Çapı}. \tag{6.14}$$

11) Diskriminant fonksiyon değerleri herbir gözlem için ayrı ayrı hesaplanır. İlgili gözlem en yüksek ayırıcı skora sahip diskriminant fonksiyonunu temsil eden gruba atanır. Örneğin, 1. grupta yer alan 1. gözlem için diskriminant fonksiyonu değeri, en yüksek değerini 1. grubu temsil eden diskriminant fonksiyonunda aldığından, ilgili gözlem birimi 1. gruba atanır. Böylece gerçekte 1. grupta yer alan gözlem biriminin klasik doğrusal diskriminant analizi sonucu yine 1. grupta olması öngörülür.

$$X_1^1 = \begin{bmatrix} 2.94 & 6.62 \end{bmatrix} \tag{6.15}$$

$$f_1^1 = -35.946 + 14.723^{*}2.94 + 3.462^{*}6.62 = 30.25 \tag{6.16}$$

$$f_2^2 = -23.012 + 12.529^{*}2.94 + 2.363^{*}6.62 = 29.46 \tag{6.17}$$

11) Diğer gözlem birimleri içinde 6.16 ve 6.17 Eş.leri kullanarak gözlemlerin yüksek olasılıkla atanacağı gruplar öngörülür. Çizelge 6.2'de Kalite Kontrol Veri Setine ait gözlemlerin gerçekte bulunduğu ve klasik doğrusal diskriminant analizi sonucu atama öngörüsünün yapıldığı gruplar gösterilmiştir.

Çizelge 6.2. Kalite Kontol Veri Setine Ait Gözlemlerin Atandığı Nihai Grupları Gösterir Tablo

| **Gerçekte Yer Aldığı Grup** | $x_1$ | $x_2$ | $f_1$ | $f_2$ | **Kestirilen (Atanan) Grup** |
|---|---|---|---|---|---|
| Grup 1 | 2.94 | 6.62 | 30.25* | 29.46 | Grup 1 |
| Grup 1 | 2.54 | 7.80 | 28.45* | 27.24 | Grup 1 |
| Grup 1 | 3.57 | 9.65 | 50.02* | 44.51 | Grup 1 |
| Grup 1 | 3.16 | 5.47 | 29.51* | 29.50 | Grup 1 |
| Grup 2 | 2.58 | 4.46 | 17.47 | 19.85* | Grup 2 |
| Grup 2 | 2.16 | 6.22 | 17.38 | 18.74* | Grup 2 |
| Grup 2 | 3.27 | 3.52 | 24.38 | 26.27* | Grup 2 |

Çizelge 6.3. Kalite Kontrol Veri Setine ait Sınıflandırma Tutarlılığı Tablosu

| | Kestirilen (Atanan) Grup | | |
|---|---|---|---|
| Gerçek grup | Grup 1 | Grup 2 | Toplam |
| Grup 1 | 4 | 0 | 4 |
| Grup 2 | 0 | 3 | 3 |
| Toplam | 4 | 3 | 7 |

Çizelge 6.3'te Kalite Kontrol Veri Seti üzerinde tanımlı 7 gözlem biriminden, gerçekte kendi grubunda olupta klasik doğrusal diskiriminant analizi sonucu yine kendi grubuna atanan gözlem sayısı 7 dir. Dolayısıyla gözlem birimlerinin % 100 'ü doğru sınıflanırılmış olup, hatalı sınıflandırılma (ARE) oranı % 0'dır.

### 6.1.2. Manipüle edilmiş kalite kontrol veri seti

Doğrusal diskriminant analizinin geçerliliği hususunda önemli varsayımlardan biri, gruplarda çok değişkenli yapıya ilişkin aykırı gözlem olmamasıdır. Bu durum bir gözlemin gerçekte var olduğu gruptan başka bir gruba atanmasına sebep olarak, hatalı sınıflandırma oranında bir artışa neden olacaktır. Aşağıda yer alan ve manipüle edilmiş veri seti üzerinde bu durumun tahmin ediciler ve hatalı sınıflandırma oranı üzerindeki yansımaları incelenmiştir. 2. Grupta lastiğin eğim açısı 3.27 iken 200, lastik çapına ilişkin gözlem değeri 3.52 iken 100 olarak alınmıştır.

Çizelge 6.4. Manipüle Edilmiş Kalite Kontrol Veri Seti

| Testi Geçenler (Grup 1) | | Testi Geçemeyenler (Grup 2) | |
|---|---|---|---|
| Lastiğin Eğim Açısı ($x_1$) | Lastiğin Çapı ($x_2$) | Lastiğin Eğim Açısı ($x_1$) | Lastiğin Çapı ($x_2$) |
| 2.94 | 6.62 | 2.58 | 4.46 |
| 2.54 | 7.80 | 2.16 | 6.22 |
| 3.57 | 9.65 | 200 | 100 |
| 3.16 | 5.47 | | |

1) Manipüle edilmiş kalite kontrol veri seti grup içi varyans-kovaryans matrisleri:

$$c = \begin{bmatrix} 5207.811 & 2494.448 \\ 2494.448 & 1196.956 \end{bmatrix}, \qquad c^{-1} = \begin{bmatrix} 0.106 & -0.221 \\ -0.221 & 0.462 \end{bmatrix}, \tag{6.18}$$

şeklinde elde edilmiştir.

2) Her bir gözlemin dahil olduğu grupta (i. grupta) bulunması olasılığı hesaplanmıştır:

$$p = \begin{bmatrix} \frac{4}{7} \\ \frac{3}{7} \end{bmatrix} = \begin{bmatrix} 0.571 \\ 0.429 \end{bmatrix}. \tag{6.19}$$

3) Grup ortalama vektörleri aşağıdaki gibi elde edilmiştir:

$$\mu_1 = [\ 3.05 \quad 7.38\ ], \mu_2 = [\ 68.24 \quad 36.89\ ], \mu = [\ 30.99 \quad 20.013\ ] \tag{6.20}$$

4) Fisher'in kanonik diskriminant fonksyionları;

$$f_i = \mu_i C^{-1} \cdot x_k^T - \frac{1}{2}\mu_i C^{-1} \cdot \mu_i^T + \ln(p_i) \tag{6.21}$$

$$f_1 = -8.814 + -1.313 \text{Lastiğin Eğim Açısı} + 2.742\,\text{Lastiğin Çapı} \tag{6.22}$$

$$f_2 = -5.232 - 0.919\,\text{Lastiğin Eğim Açısı} + 1.946\,\text{Lastiğin Çapı} \tag{6.23}$$

olarak elde edilmiştir. Örneğin, 1. grupta yer alan 4. gözlem değeri $X_4^1 = [3.16, \ \ 5.47]$ için; 1. ve 2. diskriminant fonksiyonlarından diskriminant skorları aşağıdaki gibi elde edilmiştir:

$$f_1^1 = -8.814 - 1.313{}^*3.16 + 2.742{}^*5.47 = 2.03, \tag{6.24}$$

$$f_2^2 = -5.232 - 0.919{*}3.16 + 1.946{}^*5.47 = 2.50. \tag{6.25}$$

Çizelge 6.5 ve 6.6 incelendiğinde; Kalite Kontrol Veri Setinin 2. grup 3. gözlem biriminin diğer gözlem birimlerine göre anormal büyüklüğe çıkarılarak oluşturulmuş, yani manipüle edilmiş Kalite Kontrol Veri setinde yer alan 7 gözlem biriminden, gerçekte kendi grubunda olupta klasik doğrusal diskiriminant analizi sonucu yine kendi grubuna atanan gözlem sayısının 5 olduğu görülmektedir. Dolayısıyla gözlem birimlerinin $\frac{5}{7}$ = %71'i doğru sınıflanırılmış olup, hatalı sınıflandırılma oranı (ARE) % 29'dur. 2. grup 3. gözlem

biriminin diğer gözlem birimlerine göre anormal büyüklüğe çıkarılması, yani aykırılaştırılması gözlem birimlerinin gerçekte bulundukları gruptan başka bir gruba atanması oranını önemli derecede etkilemiştir.

Çizelge 6.5. Manipüle Edilmiş Kalite Kontrol Veri Setine Ait Gözlemlerin Atandığı Nihai Grupları Gösterir Tablo

| Gerçekte Yer Aldığı Grup | $x_1$ | $x_2$ | $f_1$ | $f_2$ | Kestirilen (Atanan) Grup |
|---|---|---|---|---|---|
| Grup 1 | 2.94 | 6.62 | 5.47* | 4.94 | Grup 1 |
| Grup 1 | 2.54 | 7.80 | 9.23* | 7.61 | Grup 1 |
| Grup 1 | 3.57 | 9.65 | 12.95* | 10.26 | Grup 1 |
| Grup 1 | 3.16 | 5.47 | 2.03 | 2.50* | Grup 2 |
| Grup 2 | 2.58 | 4.46 | 0.02 | 1.07* | Grup 2 |
| Grup 2 | 2.16 | 6.22 | 5.40* | 4.88 | Grup 1 |
| Grup 2 | 200 | 100 | 2.78 | 5.56* | Grup 2 |

Çizelge 6.6. Manipüle Edilmiş Kalite Kontrol Veri Setine ait Sınıflandırma Tutarlılığı Tablosu

| | Kestirilen (Atanan) Grup | | |
|---|---|---|---|
| Gerçek grup | Grup 1 | Grup 2 | Toplam |
| Grup 1 | 3 | 1 | 4 |
| Grup 2 | 1 | 2 | 3 |
| Toplam | 4 | 3 | 7 |

### 6.1.3. Kalite kontrol veri seti üzerinde tanımlı yeni veri seti

Bu kısımda, diskriminant analizinde MCD metodunun manuel olarak uygulanması amacıyla, Kalite Kontrol Veri Seti revize edilmiş ve $X^{yeni}$ isimli yeni bir veri seti oluşturulmuştur:

$$X^{yeni} = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \\ 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix}. \tag{6.26}$$

Örneğin, $x_1^{yeni} = [2.95,\ 6.63]$ gösterimi bağımsız değişken matrisindeki 1. gözlemi; $x_2^{yeni} = [2.53,\ 7.79]$ gösterimi bağımsız değişken matrisindeki 2. gözlemi; $x_3^{yeni} = [3.57,\ 5.65]$ gösterimi bağımsız değişken matrisindeki 3. gözlemi; $x_4^{yeni} = [3.16,\ 5.47]$ gösterimi bağımsız değişken matrisindeki 4. gözlemi ifade etmektedir.

### 6.1.4. MCD algoritmasının adımlarının manuel uygulanması

MCD algortimasının adımları:

1) MCD metodu, gözlem grubundan $h$ tanesini seçerek kovaryans matrisinin determinantını hesaplamaktadır. Dolayısıyla, öncelikle alt matrislerin sayısı, $h$'ın belirlenmesi ile başlanır. $X^{yeni}$ veri matrisi $p = 2$ değişken ve $n = 4$ gözlem biriminden oluşmaktadır. Değişken sayısı 2 olduğundan seçilecek mümkün alt matrislerin sayısının, yani $h$ değerinin 2 olması beklenir.

2) Bu durumda, $\binom{n}{h}=\binom{4}{2}$=6 tane kovaryans matrisi ve determinant değeri elde edilir:

$$x_1 = \begin{bmatrix} 2.95 & 6.63 \\ 2.53 & 7.79 \end{bmatrix}, x_2 = \begin{bmatrix} 2.95 & 6.63 \\ 3.57 & 5.65 \end{bmatrix}, x_3 = \begin{bmatrix} 2.95 & 6.63 \\ 3.16 & 5.47 \end{bmatrix}, \tag{6.27}$$

$$x_4 = \begin{bmatrix} 2.53 & 7.79 \\ 3.57 & 5.65 \end{bmatrix}, x_5 = \begin{bmatrix} 2.53 & 7.79 \\ 3.16 & 5.47 \end{bmatrix}, x_6 = \begin{bmatrix} 3.57 & 5.65 \\ 3.16 & 5.47 \end{bmatrix}, \tag{6.28}$$

$$\begin{aligned} &det(x_1) = 6.20 && det(x_2) = -7.00 && det(x_3) = -4.81 \\ &det(x_4) = -13.51 && det(x_5) = -10.77 && det(x_6) = 1.67 \end{aligned} \tag{6.29}$$

3) Eş. 6.29'de hesaplanan determinat değerlerinden en küçük olan alt matris tespit edilir. İlgili örnekte $x_4$ matrisinin determinantı en küçük çıkmıştır $(det(x_4) = -13.51)$. Böylece minimum determinanta sebep olan alt matrise ait gözlem grubu aşağıdaki şekliyle belirlenmiş olur:

$$x_4 = \begin{bmatrix} 2.53 & 7.79 \\ 3.57 & 5.65 \end{bmatrix}. \tag{6.30}$$

4) Bu belirlemeden sonra, yani kovaryans matrisinin determinantını minimum yapan alt matrisin dahil olduğu gözlem grubu belirlendikten sonra, başlangıç için konum ve ölçek parametrelerinin tahmin edicileri, söz konusu alt matrisin gözlem grubu üzerinden hesaplanır.

5) En düşük determinant değerine sahip olan $x_4$ alt matrisi üzerinden hesaplanan konum ve ölçek parametreleri başlangıç için MLE tahmin değerleri aşağıdaki gibi elde edilir ve elde edilen tahmin edicilerin bu aşamada dahi sağlamlık vasfı vardır.

$$\widehat{\mu}_0 = \begin{bmatrix} 3.05 & 6.72 \end{bmatrix}, \tag{6.31}$$

$$\widehat{\Sigma}_0 = \begin{bmatrix} 0.54 & -1.11 \\ -1.11 & 2.28 \end{bmatrix}. \tag{6.32}$$

6) Her gözlem birimi için sağlamlık özelliği olan bir uzaklık (RD) tanımlanmaktadır. Bu uzaklık, $\widehat{\mu}_0$ merkezine ve $\widehat{\Sigma}_0$ metriğine göre her nokta için hesaplanmaktadır. Gözlem noktaları $\widehat{\mu}_0$ merkezine uzaklıkları bakımından ikiye ayrılmaktadır:

$$RD_{x_i^{yeni}}(x_i^{yeni}, \widehat{\mu}_0, \widehat{\Sigma}_0) = \sqrt{(x_i^{yeni} - \widehat{\mu}_0)^T \widehat{\Sigma}_0^{-1} (x_i^{yeni} - \widehat{\mu}_0)} \leq \sqrt{\chi^2_{p,0.975}}, \tag{6.33}$$

$$RD_{x_i^{yeni}}(x_i^{yeni}, \widehat{\mu}_0, \widehat{\Sigma}_0) = \sqrt{(x_i^{yeni} - \widehat{\mu}_0)^T \widehat{\Sigma}_0^{-1} (x_i^{yeni} - \widehat{\mu}_0)} > \sqrt{\chi^2_{p,0.975}}. \tag{6.34}$$

Eş. 6.33 ve 6.34'den veri merkezine uzaklığı $\sqrt{\chi^2_{p,0.975}}$'den küçük olan gözlem birimlerine $w_i = 1$ ağırlığı atanmakta, diğer gözlem birimlerine ise $0$ ağırlığı atanmaktadır. $\widehat{\Sigma}_0$ matrisinin tersi,

$$\widehat{\Sigma}_0^{-1} = \begin{bmatrix} -2533.33 & -1233.33 \\ -1233.33 & -599.99 \end{bmatrix} \tag{6.35}$$

olarak elde edilmektedir. Bunun nedeni $\widehat{\Sigma}_0$ matrisinin determinantının sıfıra yakın olmasıdır. Eş. 6.34'daki $\chi^2$ değeri için ilgili integralin üst sınırı, Eş. 6.36'da gösterildiği gibi 7.37 olarak hesaplanır:

$$\chi^2_{2,0.975} = \int_0^{\chi^2=7.37} f(\chi_2^2)dx = 0.975 \tag{6.36}$$

ve $\sqrt{\chi^2_{2,0.975}} = \sqrt{7.37} = 2.71$ olarak elde edilir.

Örnek olarak, $x_1^{yeni} = [2.95,\ 6.63]$ gözlemine yapılacak ağırlık ataması $w_1$ ile bu gözlemin $\widehat{\mu_w}$ ve $\hat{\Sigma}_w$ sağlam tahmin edicilerine katkısını hesaplayalım:

$$RD_{x_1^{yeni}}(x_1^{yeni}, \widehat{\mu}_0, \widehat{\Sigma}_0) = \sqrt{(x_1^{yeni} - \widehat{\mu}_0)^T \widehat{\Sigma}_0^{-1} (x_1^{yeni} - \widehat{\mu}_0)} > \sqrt{\chi^2_{p,0.975}} \tag{6.37}$$

$$= \sqrt{\begin{bmatrix} 2.95 - 3.05 & 6.63 - 6.72 \end{bmatrix}' \begin{bmatrix} -2533.33 & -1233.33 \\ -1233.33 & -599.99 \end{bmatrix} \begin{bmatrix} 2.95 - 3.05 & 6,63 - 6,72 \end{bmatrix}}$$

$$= \sqrt{19.63} = 4.43. \tag{6.38}$$

olarak elde edilir.

$RD_{x_1^{yeni}}(x_1^{yeni}, \widehat{\mu}_0, \widehat{\Sigma}_0) = 4.43 > 2.71$ olduğundan; $RD_{x_1^{yeni}}(x_1^{yeni}, \widehat{\mu}_0, \widehat{\Sigma}_0)$ fonksiyonu; $x_1^{yeni} = [2.95,\ 6.63]$ ağırlık fonksiyonuna $w_1 = 0$ değerini atar.

7) Tahmin edilen ortalama ve varyans aynı kalmaya devam ettiğinde işlem durdurulur.

8) Sonuç olarak ağırlıklı ortalama ve ağırlıklı varyans-kovaryans matrisi Eş 5.10 ve 5.11 gibi tahmin edilir.

Bu şekilde hesaplanmasından, yani sadece $\widehat{\mu}_0$'nın yakın komşuluğunda yer alan gözlemlerin hesaba katılmasından dolayı, bu tahmin edicilerin küçük çaplı örnekler dahi söz konusu iken etkinliğinin artmış olması beklenir. Kalite Kontrol yeni veri seti üzerinde tanımlı verisi için sağlam lokasyon ve varyans-kovaryans matrisi tahmin edicileri aşağıdaki gibi elde edilir:

$$\hat{\mu}_{MCD} = \begin{bmatrix} 2.88 & 6.63 \end{bmatrix}, \tag{6.39}$$

$$\hat{\Sigma}_{MCD} = \begin{bmatrix} -0.42 & -1.50 \\ -1.50 & 5.54 \end{bmatrix}. \tag{6.40}$$

### 6.1.5. En çok olabilirlik tahmin edicilerinin bulunması

Çalışmanın 3. bölümünde kalite kontrol veri seti üzerinde tanımlı yeni veri seti için, MLE tahmin edicileri (lokasyon tahmin edicisi için örnek ortalaması istatistiği ve ölçek tahmin edicisi için varyans-kovaryans istatistiği) Eş. 6.41, 6.42'de verildiği gibi elde edilmiştir:

$$\hat{\mu}^* = \begin{bmatrix} 3.05 & 6.38 \end{bmatrix} \tag{6.41}$$

$$\hat{\Sigma}^* = \begin{bmatrix} 0.18 & -0.41 \\ -0.41 & -1.13 \end{bmatrix} \tag{6.42}$$

### 6.1.6. Sağlam ve klasik uzaklıklar açısından aykırı gözlem tespiti

Sağlam Uzaklıklar Açısından Aykırı Gözlem Tespiti

Eş. 6.39 ve 6.40'da sağlam lokasyon ve varyans-kovaryans matrisi tahmin edicileri, kalite kontrol veri seti üzerinde tanımlı yeni veri seti için elde edilmiştir. Sağlam uzaklık metriği, her bir gözlem birimi için ilgili gözlem birimlerinin yoğunlaşma elipsoidinin merkezi olarak tahmin edilen sağlam lokasyon ve varyans-kovaryans matrisi ölçülerinden uzaklığına bakarak çalışır ve aşağıdaki gibi hesaplanmıştır:

$$RD_{x_i^{yeni}}(x_i^{yeni}, \widehat{\mu}_{MCD}, \widehat{\Sigma}_{MCD}) = \sqrt{(x_i^{yeni} - \widehat{\mu}_{MCD})^T \widehat{\Sigma}_{MCD}^{-1} (x_i^{yeni} - \widehat{\mu}_{MCD})}$$

$$RD_{x_1^{yeni}} = \sqrt{\begin{bmatrix} 2.95 - 2.88 & 6.63 - 6.63 \end{bmatrix}' \begin{bmatrix} -0.42 & -1.50 \\ -1.50 & 5.54 \end{bmatrix} \begin{bmatrix} 2.95 - 2.88 & 6.63 - 6.63 \end{bmatrix}}$$

$$RD_{x_2^{yeni}} = \sqrt{\begin{bmatrix} 2.53 - 2.88 & 7.79 - 6.63 \end{bmatrix}' \begin{bmatrix} -0.42 & -1.50 \\ -1.50 & 5.54 \end{bmatrix} \begin{bmatrix} 2.53 - 7.88 & 7.79 - 6.63 \end{bmatrix}}$$

$$RD_{x_3^{yeni}} = \sqrt{\begin{bmatrix} 3.57 - 2.88 & 5.65 - 6.63 \end{bmatrix}' \begin{bmatrix} -0.42 & -1.50 \\ -1.50 & 5.54 \end{bmatrix} \begin{bmatrix} 3.57 - 2.88 & 5.65 - 6.63 \end{bmatrix}}$$

$$RD_{x_4^{yeni}} = \sqrt{\begin{bmatrix} 3.16 - 2.88 & 5.47 - 6.63 \end{bmatrix}' \begin{bmatrix} -0.42 & -1.50 \\ -1.50 & 5.54 \end{bmatrix} \begin{bmatrix} 3.16 - 2.88 & 5.47 - 6.63 \end{bmatrix}}$$

Klasik Uzaklıklar Açısından Aykırı Gözlem Tespiti

Eş. 6.41 ve 6.42'de yer alan tahmin ediciler, kalite kontrol veri seti üzerinde tanımlı yeni veri seti için oluşturulan en çok olabilirlik tahmin edicileri ile elde edilen Mahalanobis uzaklıklarıdır ve her bir gözlem birimi için ilgili gözlem birimlerinin yoğunlaşma elipsoidinin merkezi olarak tahmin edilen en çok olabilirlik tahmin edicilerine ait uzaklığa bakarak çalışır ve aşağıdaki gibi hesaplanmıştır:

$$MD(x_i^{yeni}) = \sqrt{(x_i^{yeni} - \hat{\mu}^*)' \hat{\Sigma^*}^{-1} (x_i^{yeni} - \hat{\mu}^*)}. \tag{6.43}$$

$$MD(x_1^{yeni}) = \sqrt{\begin{bmatrix} 2.95 - 3.05 & 6.63 - 6.38 \end{bmatrix}' \begin{bmatrix} 0.18 & -0.41 \\ -0.41 & -1.13 \end{bmatrix} \begin{bmatrix} 2.95 - 3.05 & 6.63 - 6.38 \end{bmatrix}}$$

$$MD(x_2^{yeni}) = \sqrt{\begin{bmatrix} 2.53 - 3.05 & 7.79 - 6.38 \end{bmatrix}' \begin{bmatrix} 0.18 & -0.41 \\ -0.41 & -1.13 \end{bmatrix} \begin{bmatrix} 2.53 - 3.05 & 7.79 - 6.38 \end{bmatrix}}$$

$$MD(x_3^{yeni}) = \sqrt{\begin{bmatrix} 3.57 - 3.05 & 5.65 - 6.38 \end{bmatrix}' \begin{bmatrix} 0.18 & -0.41 \\ -0.41 & -1.13 \end{bmatrix} \begin{bmatrix} 3.57 - 3.05 & 5.65 - 6.38 \end{bmatrix}}$$

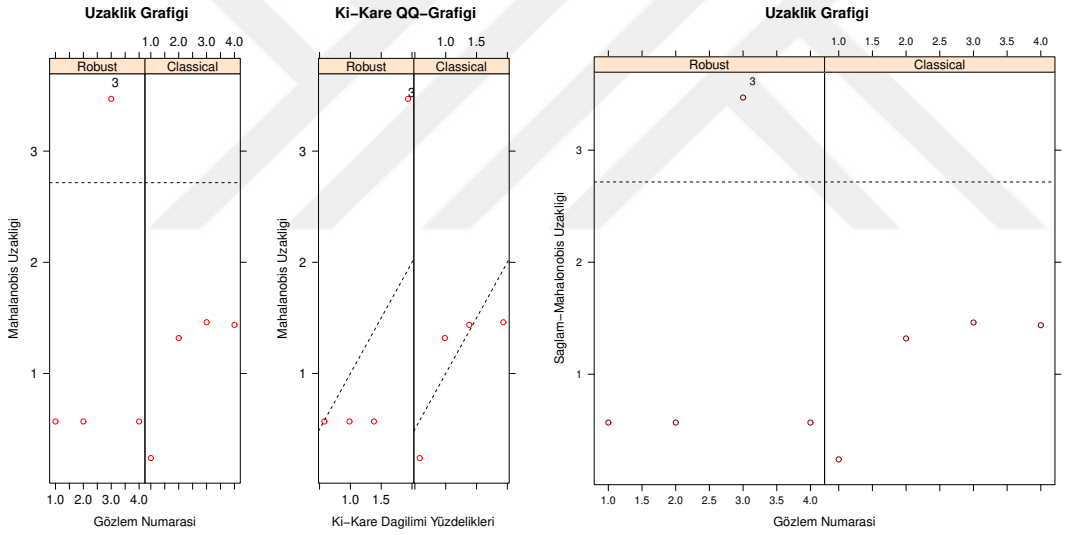
$$MD(x_4^{yeni}) = \sqrt{\begin{bmatrix} 3.16 - 3.05 & 5.47 - 6.38 \end{bmatrix}' \begin{bmatrix} 0.18 & -0.41 \\ -0.41 & -1.13 \end{bmatrix} \begin{bmatrix} 3.16 - 3.05 & 5.47 - 6.38 \end{bmatrix}}$$

Sağlam ve Klasik Uzaklıların Karşılaştırmalı Gösterimi :

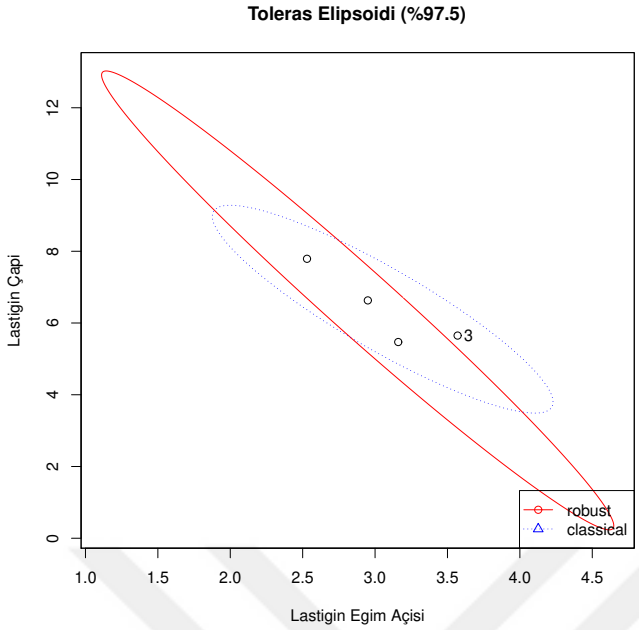
Çizelge 6.7. Kalite Kontrol Veri Seti Üzerinde Tanımlı Kontrol Grubu için Sağlam ve Klasik Uzaklıklar.

| Gözlem No | Sağlam Uzaklık (RD) | Klasik Uzaklık (MD) |
|---|---|---|
| 1 | 0.5688614 | 0.2401326 |
| 2 | 0.5688614 | 1.3191459 |
| 3 | 3.4697463 | 1.4612372 |
| 4 | 0.5688614 | 1.4376983 |

Kalite Kontrol Veri Seti üzerinde tanımlı yeni veri seti için elde edilen uzaklıklar Eş. 6.2; Sağlam tahmin ediciler kullanıldığında 3 numaralı gözlem aykırı gözlem olarak tespit edilirken, en çok olabilirlik tahmin edicileri kullanılarak hesaplanan uzaklıklarda 3 numaralı gözlem aykırı gözlem olarak tespit edilmemiştir.




Şekil 6.2. Kalite Kontrol Veri Seti Üzerinde Tanımlı Yeni Veri Seti için Sağlam ve Klasik Uzaklıklara ilişkin sırasıyla Uzaklık, Normallik Sınaması Ki-Kare Q-Q ve Uzaklık grafikleri.




Şekil 6.3. Kalite Kontrol Veri Seti Üzerinde Tanımlı Yeni Veri Seti için Sağlam ve Klasik Tahmin Edicilerin Tolerans Elipsoidi

## 6.2. Hawkins, Bradu ve Kass yapay verisi (HBK)

Bu kısımda, Hawkins, Bradu ve Kass (1984) tarafından üretilen yapay veri seti dikkate alınmıştır. 4 boyutlu (bir bağımlı ve ve üç bağımsız değişken) ve 75 gözleme sahip olan bu veri setinden, sadece X1 X2 X3 olarak tanımlanan bağımsız değişkenler uygulamada kullanılmaktadır. Bu veri seti oluşturulurken iki gruptaki ilk 14 gözlem aykırı gözlem olarak üretilmiştir.

### 6.2.1. Klasik en çok olabilirlik ve sağlam tahmin edicileri

Çizelge 6.8. HBK veri seti için Klasik En Çok Olabilirlik Tahmin Edicileri

| X1 | X2 | X3 |
|---|---|---|
| 3.206667 | 5.597333 | 7.230667 |

Çizelge 6.9. HBK veri seti için Varyans-Kovaryans Matrisinin Klasik En Çok Olabilirlik Tahmin Edicisi

| | X1 | X2 | X3 |
|---|---|---|---|
| X1 | 13.34 | 28.46 | 41.24 |
| X2 | 28.46 | 67.88 | 94.66 |
| X3 | 41.24 | 94.66 | 137.83 |

Çizelge 6.10. HBK Veri Seti için Sağlam (MCD algoritması) Lokasyon Tahmin Edicileri

| X1 | X2 | X3 |
|---|---|---|
| 1.538 | 1.780 | 1.687 |

Çizelge 6.11. HBK Veri Seti İçin Sağlam (MCD algoritması) Varyans-Kovaryans Tahmin Edicisi

| | X1 | X2 | X3 |
|---|---|---|---|
| X1 | 1.6528 | 0.0741 | 0.1713 |
| X2 | 0.0741 | 1.6823 | 0.2055 |
| X3 | 0.1713 | 0.2055 | 1.5624 |

### 6.2.2. Maskeleme etkisinin izlenilebilirliği

İlgili veri seti oluşum amacı dahilinde Maskeleme Etkisi'nin iyi bir örneğini sunmaktadır (Hawkins ve diğerleri, 1984). Bilindiği üzere Maskeleme Etkisi aykırı gözlem olması durumunda, aykırı gözlemlerin düzenli gözlem gibi görünmesi durumudur. Ele alınan veri setinde, her iki yöntem için de kritik değer $\sqrt{\chi^2_{3,0.975}}$=3.05 olarak elde edilmiştir ve bu değerden büyük uzaklıklar aykırı gözlem olarak değerlendirilmektedir. MCD yöntemine dayalı sağlam tahmin ediciler kullanarak yapılan ölçümlerde [1:14] no'lu gözlemler, aykırı gözlem niteliği kazanırken; klasik en çok olabilirlik tahmin edicilerine dayalı olarak elde edilen uzaklıklar baz alınarak yapılan ölçümlerde sadece [12,14] no'lu gözlemler aykırı gözlem niteliği kazanmıştır.

Çizelge 6.12. HBK Veri Seti İçin Sağlam Uzaklıklar

| | Gözlem No: | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | [1] | [10] | [19] | [28] | [37] | [46] | [55] | [64] | [73] |
| | 593.7527 | 655.81 | 1.13 | 0.73 | 2.77 | 2.39 | 1.2934 | 2.35 | 1.20 |
| | 624.9270 | 919.45 | 2.92 | 0.88 | 1.51 | 3.56 | 1.8535 | 1.63 | 1.57 |
| | 696.6151 | 986.73 | 0.76 | 3.05 | 2.20 | 2.43 | 1.2673 | 1.46 | 2.91 |
| Sağlam Uzaklık Değerleri | 739.6908 | 933.52 | 2.08 | 2.01 | 0.80 | 1.72 | 2.0793 | 0.20 | |
| | 713.6212 | 1156.5 | 0.92 | 2.12 | 2.86 | 1.48 | 1.1212 | 3.04 | |
| | 640.9069 | 2.74 | 1.18 | 1.11 | 2.22 | 1.57 | 3.0114 | 2.08 | |
| | 644.7492 | 3.21 | 2.70 | 2.85 | 3.19 | 3.00 | 3.4469 | 1.26 | |
| | 608.2405 | 2.57 | 1.99 | 2.43 | 2.79 | 4.33 | 2.7047 | 0.69 | |
| | 699.3602 | 0.42 | 2.72 | 0.89 | 2.35 | 2.48 | 2.2251 | 0.58 | |

Çizelge 6.13. HBK Veri Seti İçin Klasik Uzaklıklar

| | Gözlem No : | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | [1] | [11] | [21] | [31] | [41] | [51] | [61] | [71] |
| Klasik Uzaklık Değerleri | 1.91 | 2.44 | 1.089 | 1.8384961 | 1.6997 | 1.3026508 | 1.6749449 | 0.64 |
| | 1.85 | 3.10 | 1.54 | 1.3072295 | 1.7650 | 2.0760547 | 0.7595327 | 1.05 |
| | 2.31 | 2.66 | 1.08 | 0.9819878 | 1.8700 | 2.2104432 | 1.2922585 | 1.47 |
| | 2.22 | 6.38 | 0.97 | 1.1750140 | 1.4204 | 1.4142877 | 0.9738683 | 1.64 |
| | 2.10 | 1.81 | 0.79 | 1.2436358 | 1.0759 | 1.2304551 | 1.1482076 | 1.89 |
| | 2.14 | 2.15 | 1.16 | 0.8508035 | 1.3441 | 1.3311013 | 1.2967463 | |
| | 2.01 | 1.38 | 1.44 | 1.8323781 | 1.9663 | 0.8327441 | 0.6298270 | |
| | 1.91 | 0.84 | 0.86 | 0.7520607 | 1.4242 | 1.4044015 | 1.5495480 | |
| | 2.22 | 1.14 | 0.57 | 1.2650409 | 1.5697 | 0.5912349 | 1.0705111 | |
| | 2.33 | 1.59 | 1.56 | 1.1120381 | 0.4239 | 1.8897366 | 0.9977606 | |

## 6.3. Alcohol veri seti

Bu kısımda, 77 gözlem birimi ve 3 gruptan oluşan Alcohol veri seti ele alınmaktadır. Bu gruplar ”KIRSCH”, ”MIRAB” ve ”POIRE” isimleriyle nitelendirilmiştir ve Gruplar ”MEOH, ACET, BU1, MPR, ACAL, LNPRO1” isimli 6 değişken bakımından oluşturulmuştur. Birinci grup 18, ikinci grup 29 ve üçüncü grup 30 gözlemden oluşmaktadır.

Bu veri setinin incelenmesindeki amaç, diskriminant analizi uygulamasında, en önemli varsayımlardan aykırı gözlem olmaması, grupların çok değişkenli normal dağılıma sahip olması ve grupta yer alan değişkenler arasında doğrusal bir ilişkinin olmaması varsayımlarının sağlanıp sağlanmadığıdır. Bu amaçla, her gruptaki altı değişken tek tek aykırı gözlem içerip içermediği bakımından kutu grafikleri ile incelenmiştir. Öte yandan, her gruptaki 6 değişken için ikili klasik tolerans elipsoidi ve sağlam tolerans elipsoidi aynı düzlem üzerinde gösterilmiştir ve her bir grupta yer alan değişken ikilileri için klasik ve sağlam korelasyon grafikleri değerleri ile birlikte verilmiştir. Uygulama, klasik MLE, MCD tahmin edicilerine dayalı metotlar için ayrı ayrı yapılmıştır. (Hubert ve diğerleri, 2018)

Alcohol veri setinin genel görünümü Çizelge 6.14’de verilmektedir.

Çizelge 6.14. Alcohol Veri Seti

| Gözlem No | KIRSCH | | | | | | MIRAB | | | | | | POIRE | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
| 1 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 2 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| 3 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| n | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

### 6.3.1. En çok olabilirlik ve sağlam tahmin edicileri

Aşağıdaki Çizelge'lerde sırasıyla, MLE ve başlangıç için MCD metoduna dayalı lokasyon ve varyans-kovaryans tahmin edicileri verilmektedir.

Çizelge 6.15. Birleştirilmiş Alcohol Veri Setinde Lokasyon parametrelerine ait En Çok Olabilirlik Tahmin Edicileri

| MEOH | ACET | BU1 | MEPR | ACAL | LNPRO1 |
|---|---|---|---|---|---|
| 378.7 | 243.7 | 1.622 | 32.07 | 13.94 | 6.231 |

Çizelge 6.16. Birleştirilmiş Alcohol Veri Setine Ait Kovaryans matrisinin En Çok Olabilirlik Tahmin Edicisi

| | MEOH | ACET | BU1 | MEPR | ACAL | LNPRO1 |
|---|---|---|---|---|---|---|
| MEOH | 33550.68 | 16796.45 | 133.48 | 1862.59 | 1102.77 | 68.27 |
| ACET | 16796.45 | 38549.03 | 90.43 | 1290.53 | 969.59 | 37.29 |
| BU1 | 133.48 | 90.43 | 1.30 | 6.15 | 16.09 | 0.22 |
| MEPR | 1862.59 | 1290.53 | 6.15 | 140.78 | 34.62 | 3.01 |
| ACAL | 1102.77 | 969.59 | 16.09 | 34.62 | 236.11 | 1.46 |
| LNPRO1 | 68.27 | 37.29 | 0.22 | 3.01 | 1.46 | 0.72 |

Çizelge 6.17. Birleştirilmiş Alcohol Veri Setine Ait Sağlam Lokasyon Tahmin Edicileri (MCD)

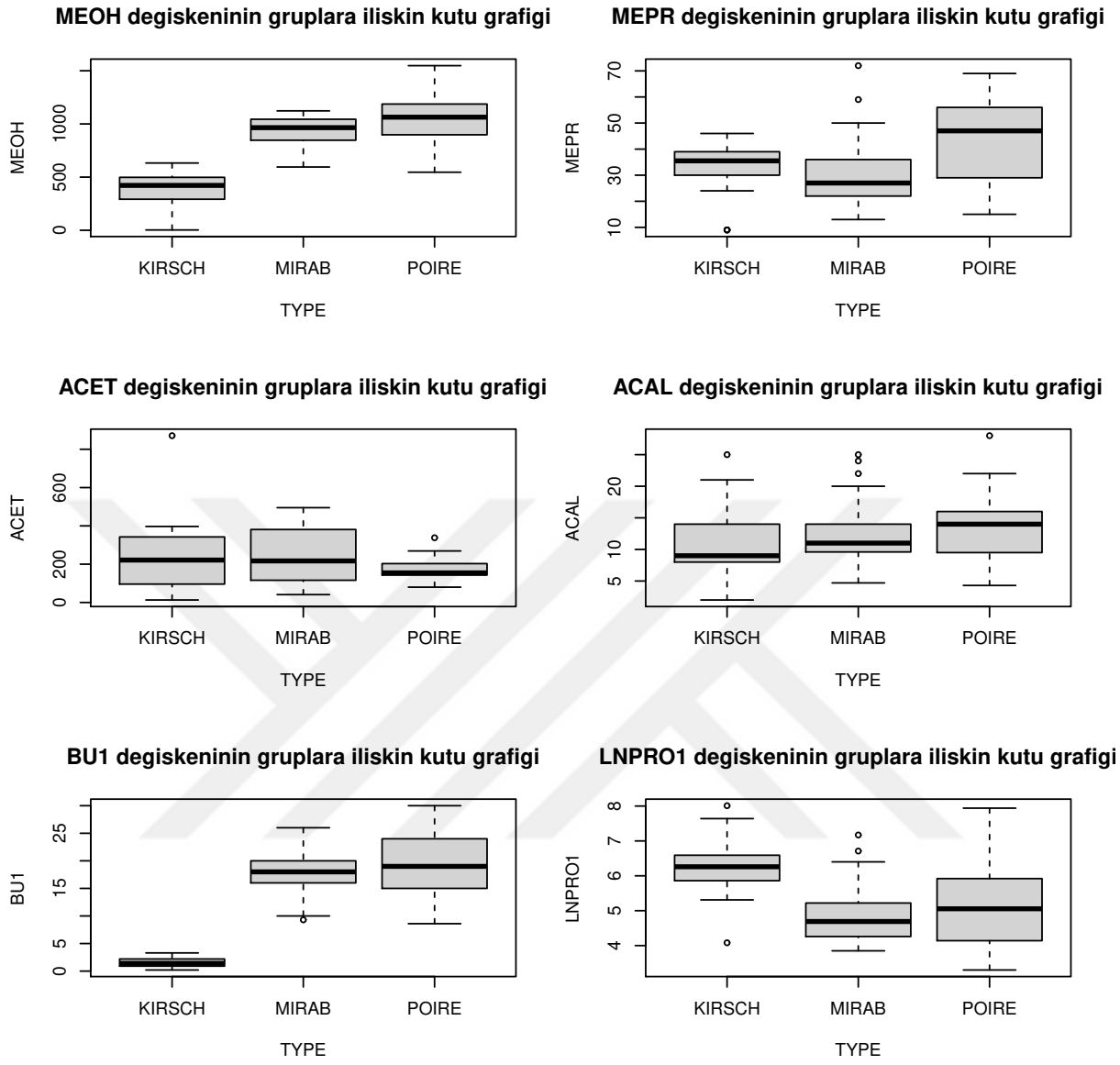
| MEOH | ACET | BU1 | MEPR | ACAL | LNPRO1 |
|---|---|---|---|---|---|
| 438.708 | 230.275 | 1.583 | 36.267 | 12.083 | 6.302 |

Çizelge 6.18. Birleştirilmiş Alcohol Veri Setine Ait Sağlam Varyans-Kovaryans tahmin edicisi (MCD)

| | MEOH | ACET | BU1 | MEPR | ACAL | LNPRO1 |
|---|---|---|---|---|---|---|
| MEOH | 23440 | 15890 | 108 | 457.1 | 930.6 | 186.1 |
| ACET | 15890 | 24740 | 125.1 | 1100 | 1170 | 131.3 |
| BU1 | 108 | 125.1 | 0.8703 | 6.310 | 7.715 | 0.7717 |
| MEPR | 457.1 | 1100 | 6.310 | 77.3 | 58.53 | 5.445 |
| ACAL | 930.6 | 1170 | 7.715 | 58.53 | 95.28 | 8.158 |
| LNPRO1 | 186.1 | 131.3 | 0.7717 | 5.445 | 8.158 | 2.486 |

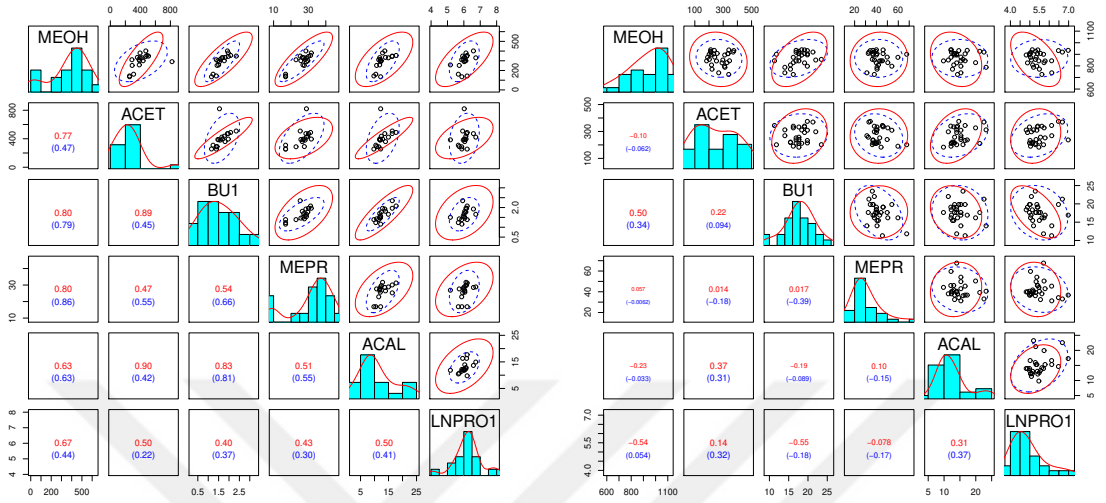
### 6.3.2. Alcohol veri setindeki her bir değişken için kutu grafikleri ve gruplara ilişkin aykırı gözlemler

Alcohol veri setine ait gruplarda yer alan her bir değişken için tek değişkenli kutu grafikleri (box-plot) Şekil 6.4'de verilmektedir ve kutu grafiklerinden görüldüğü gibi MIRAB grubunda yer alan ACAL, LNPRO1 ve MPR değişkenlerinin nispeten fazla aykırı değer içerdiği, dolayısıyla ilgili grubun çok değişkenli aykırı değer içerme potansiyelinin yüksek olduğu düşünülmektedir.

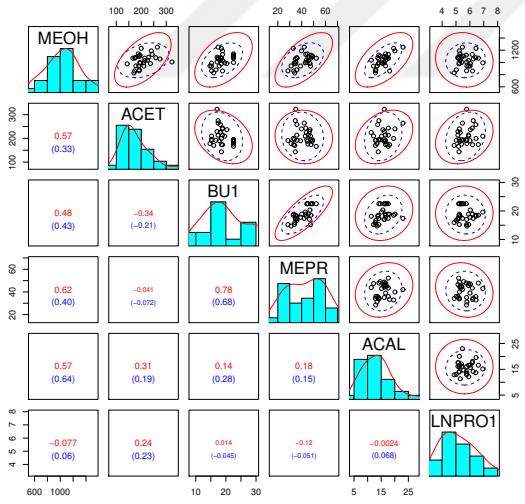



Şekil 6.4. Alcohol Veri Setindeki Her bir Değişken İçin Kutu Grafikleri ve Gruplara ilişkin Aykırı Gözlemler

### 6.3.3. Alcohol veri seti için klasik ve sağlam tahmin ediciler bakımından Tolerans Elipsoidi grafikleri ve Korelasyon değerleri



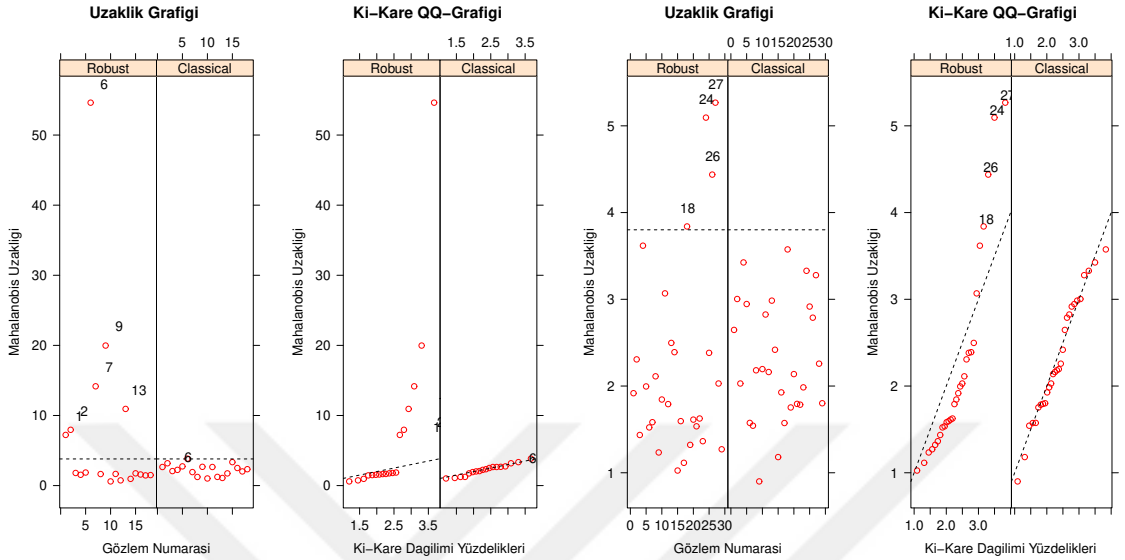

Şekil 6.5. KIRSCH ve MIRAB Grubu için Tolerans Elipsoidi Grafiği



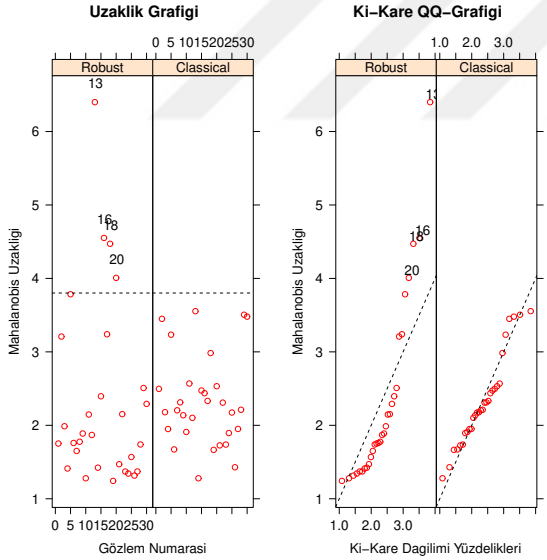

Şekil 6.6. POIRE Grubu için Tolerans Elipsoidi Grafiği

Şekil 6.5 ve Şekil 6.6 incelendiğinde; KIRSCH, MIRAB ve POIRE grubu içerisinde yer alan değişken ikilileri için MCD ve MLE tahmin edicisine dayalı robust tolerans elipsoidi ve klasik tolerans elipsoidi grafiği, robust korelasyon değerleri ve iki değişken arasındaki bilinen Pearson korelasyon katsayısının, ayrıca değişkenlerin histogramlarının aynı grafik üzerinde çizdirildiği görülmektedir.

### 6.3.4. Alcohol veri seti için klasik ve sağlam tahmin ediciler bakımından uzaklık grafikleri

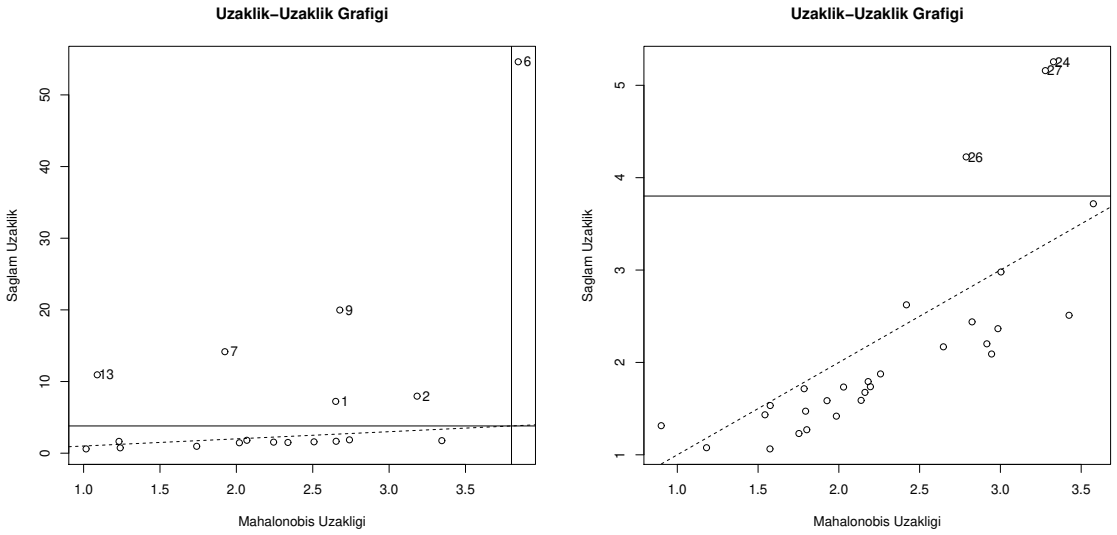



Şekil 6.7. KIRSCH ve MIRAB Grubu için Uzaklık Grafikleri

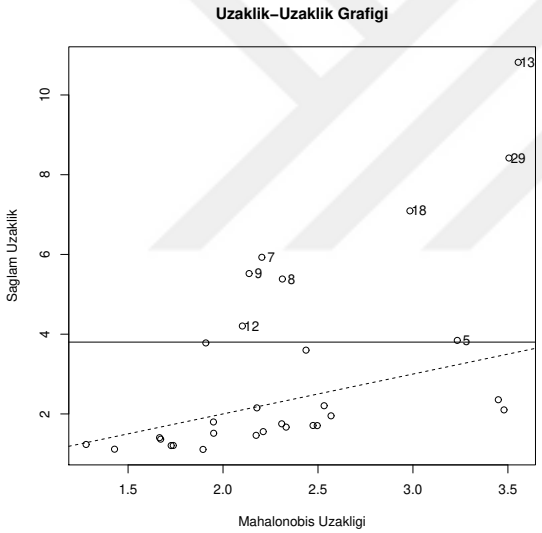



Şekil 6.8. POIRE Grubu için Uzaklık Grafikleri

Şekil 6.7 ve 6.8 incelendiğinde; MCD algoritması ile elde edilen çok değişkenli sağlam istatistikler ve MLE tahmin ediciler kullanılarak aykırı gözlem incelemesi yapılmış, $\sqrt{\chi^2_{p,0.975}}$'den $(p = 6)$ büyük gözlemler aykırı gözlem niteliği kazanmıştır.




Şekil 6.9. KIRSCH ve MIRAB Grubu için Uzaklık Grafikleri




Şekil 6.10. POIRE Grubu için Uzaklık Grafikleri

Şekil 6.9 ve Şekil 6.10 incelendiğinde; MCD algoritması ile elde edilen çok değişkenli sağlam istatistikler ve MLE tahmin ediciler kullanılarak aykırı gözlem incelemesi yapılmış ve aykırı gözlemler bakımından 2 boyulu grafiği çizdirilmiştir. $\sqrt{\chi^2_{p,0.975}}$'den $(p = 6)$ büyük gözlemler aykırı gözlem niteliği kazanmıştır.

### 6.3.5. Alcohol veri seti için klasik ve sağlam tahmin ediciler bakımından hatalı sınıflandırma oranları

Çizelge 6.19. Alcohol Veri Setinde MLE, MCD, MCD-A, MCD-B, MCD-C Tahmin Edicileri kullanarak Diskriminant Analizi Sonucu Oluşturulmuş Sınıflandırma Tutarlılığı Tablosu

| Tahmin ediciler | Sınıflandırma Tutarlılığı Tablosu | | | | Hatalı Sınıflandırma Oranı (ARE) |
|---|---|---|---|---|---|
| MLE | | | Tahmin | | 0.2078 |
| | | KIRSCH | MIRAB | POIRE | |
| | Gerçek KIRSCH | 18 | 0 | 0 | |
| | Gerçek MIRAB | 0 | 22 | 7 | |
| | Gerçek POIRE | 0 | 9 | 21 | |
| Klasik MCD | | | Tahmin | | 0.2078 |
| | | KIRSCH | MIRAB | POIRE | |
| | Gerçek KIRSCH | 18 | 0 | 0 | |
| | Gerçek MIRAB | 1 | 23 | 5 | |
| | Gerçek POIRE | 1 | 9 | 20 | |
| MCD-A | | | Tahmin | | 0.2078 |
| | | KIRSCH | MIRAB | POIRE | |
| | Gerçek KIRSCH | 18 | 0 | 0 | |
| | Gerçek MIRAB | 1 | 23 | 5 | |
| | Gerçek POIRE | 1 | 9 | 20 | |
| MCD-B | | | Tahmin | | 0.1818 |
| | | KIRSCH | MIRAB | POIRE | |
| | Gerçek KIRSCH | 18 | 0 | 0 | |
| | Gerçek MIRAB | 1 | 22 | 6 | |
| | Gerçek POIRE | 1 | 6 | 23 | |
| MCD-C | | | Tahmin | | 0.1818 |
| | | KIRSCH | MIRAB | POIRE | |
| | Gerçek KIRSCH | 17 | 1 | 0 | |
| | Gerçek MIRAB | 1 | 22 | 6 | |
| | Gerçek POIRE | 1 | 5 | 24 | |

Klasik doğrusal diskriminant analizinin yukarıda bahsi geçen varsayımlarının sağlanmadığı Alcohol veri seti için yapılan diskriminant analizi uygulamasında, Klasik MLE, MCD ve MCD-A metodu ile elde edilen sınıflandırma tutarlılığı tabloları farklı olmasına rağmen, gözlemlerin hatalı sınıflandırılma oranları aynı çıkmıştır. Öte yandan, MCD-B ve MCD-C gerek sınıflandırma tutarlılığı tabloları gerekse gözlemlerin hatalı sınıflandırılma oranları daha düşük olduğu için daha güvenilir sonuçlar vermektedir.

# 7 . SİMÜLASYON ÇALIŞMASI

Çalışmanın bu kısmında, parametreleri belli tek değişkenli ve çok değişkenli normal dağılımlar için özellikle başka dağılımdan karışan gözlem ya da gözlem grubu olması durumunda karışma dağılımları üzerinde tanımlı lineer fonksyioneller klasik en çok olabilirlik tahmin edicileri ile sağlamlık özelliğine sahip tahmin edicilerin kıyaslaması yapılmıştır. Çok değişkenli normal dağılıma sahip modeller için açıklayıcı değişken sayısı $p = \{2, 6\}$, bağımlı değişken sayısı yani grup sayısı (sınıf sayısı) $g = \{2, 3\}$ herbir gruptaki gözlem sayısı eşit ve $n = \{20, 100\}$ olarak belirlenmiş, Klasik MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicilerinin ve bu tahmin edicilerle uygulanan diskriminant analizi sonucu elde edilen hatalı sınıflandırma oranlarının nasıl bir farklılık gösterdiğinin ortaya konulması amaçlanmıştır (Todorov ve Pires, 2007), (Gündüz ve Fokoue, 2017).

Bu simülasyon çalışması $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ karışma oranı için ayrı ayrı uygulanmış olup, $\varepsilon = \{0\}$ olması öngörülen dağılıma karışmanın olmadığı anlamına gelmektedir.

## 7.1. Tek değişkenli yapıda Lokasyon ve Ölçek karışması durumu

Tek değişkenli bir rassal değişkenin dağılımına, yine tek değişkenli başka rassal değişkenin dağılımdan belirli oranda veri karışması olduğunda, gerçek dağılımın yapısı kirlenir. Bu kirlenmenin tahmin edici üzerindeki etkileri, birleştirilmiş yeni rassal değişken üzerinde tanımlı lineer fonsiyoneller ile izlenebilir. Eş.7.1'deki $F_\varepsilon$ dağılımı, $X$ rassal değişkeni ile temsil edilen dağılım fonksiyonuna, $Y$ rassal değişkeninin dağılım fonksiyonundan $\varepsilon$ oranında karışma olduğunu göstermektedir:

$$F_\varepsilon = (1 - \varepsilon)F_X + (\varepsilon)F_Y. \tag{7.1}$$

$X$ rassal değişkeninin dağılım fonksiyonuna, $Y$ rassal değişkeninin dağılım fonksiyonundan $\varepsilon$ oranında karışma olması durumunda, yeni dağılım fonksiyonunun elde edilmesine ilişkin algoritmanın adımları aşağıdaki gibidir:

1) $F_X$ ve $F_Y$ sırasıyla herhangi bir dağılım ailesi içinde tanımlı $X$ ve $Y$ rassal değişkenleri için birer dağılım fonksiyonu olsun.

2) $[0 - 1]$ aralığında tanımlı Tekdüze (Uniform) dağılımdan tesadüfi olarak bir reel sayı seçilsin.

3) $[0 - 1]$ aralığında tanımlı Tekdüze (Uniform) dağılımdan tesadüfi olarak seçilen reel sayının $(1 - \varepsilon)$ kadarı $X$ rassal değişkeninin dağılımından, $(\varepsilon)$ kadarı $Y$ rassal değişkeninin dağılımından meydana gelinceye kadar 2. ve 3. adımlar tekrarlanır.

4) $\varepsilon$ oranında karışmış, yeni dağılım fonksiyonu ($F_\varepsilon = (1-\varepsilon)F_X + (\varepsilon)F_Y$) üzerinde tanımlı lineer fonksiyoneller araclığıyla karışmanın tahmin edici üzerindeki etkileri incelenebilir.

Tek değişkeli normal dağılıma sahip $X$ rassal değişkeninin dağılımına, tek değişkenli normal dağılıma sahip $Y$ rassal değişkenin dağılımından $\varepsilon$ oranında karışma olsun. Bu karışma için iki farklı senaryo dikkate alınmıştır:

Gerçek dağılıma, ortalaması (lokasyon parametresi) farklı, varyansı (ölçek parametresi) aynı ve ortalaması aynı, varyansı farklı dağılımlardan karışma olması durumları sırasıyla Eş. 7.2 ve Eş. 7.3 kullanılarak ifade edilebir:

$$F_\varepsilon = (1-\varepsilon).N(\mu_X, \sigma_X^2) + (\varepsilon).N(\mu_Y, \sigma_Y^2), \quad \sigma_X^2 = \sigma_Y^2 = \sigma^2 \tag{7.2}$$

$$F_\varepsilon = (1-\varepsilon).N(\mu_X, \sigma_X^2) + (\varepsilon).N(\mu_Y, \sigma_Y^2), \quad \mu_X = \mu_Y = \mu \tag{7.3}$$

Aşağıda Eş. 7.2 ve Eş. 7.3 durumlarına ilişkin ayrı ayrı örnekler verilmektedir.

### 7.1.1. Örnek: Lokasyon parametresinin farklı olduğu karışma durumu

$X \sim N(0,1)$ şeklinde tanımlı standart normal dağılımına, $Y \sim N(10,1)$ şeklinde tanımlı normal dağılımın $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ oranlarında karışması sonucu $\varepsilon$ oranında karışma olan dağılım fonksiyonu Eş. 7.4'deki gibi ifade edilebilir:

$$F_\varepsilon = (1-\varepsilon) \cdot F_X + \varepsilon \cdot F_Y \tag{7.4}$$

Örneğin, manuel(elle) olarak hesaplanması nispeten diğer lineer fonksiyonellere göre daha kolay olduğu için, beklenen değer ve varyans operatörünün, $n = 1000$ gözlem birimi, $\varepsilon = 0.25$ olması durumunda karışma dağılımına ait lokasyon tahmin edicisi üzerindeki etkisini inceleyelim: $\varepsilon = 0.25$ için gözlemlerin yaklaşık 0.75'i $X$ rassal değişkeninin, 0.25'i $Y$ rassal değişkeninin dağılımından meydana gelmektedir. Eş. 7.4'nin beklenen değeri alındığında, Eş. 7.5 elde edilir:

$$E(F_\varepsilon) = E((1-\varepsilon) \cdot F_X) + E(\varepsilon \cdot F_Y) \tag{7.5}$$

Beklenen değer operatörü özelliği gereği Eş. 7.5, Eş. 7.6'teki gibi yazılabilir:

$$E(F_\varepsilon) = (1-\varepsilon) \cdot E(X) + \varepsilon \cdot E(Y) \tag{7.6}$$

$\varepsilon$= {0.25} için, $E(X) = 0, E(Y) = 10$ değerleri Eş. 7.6'te yerine yazılırsa; $\varepsilon$ oranda karışmış dağılım fonksyionunun beklenen değeri, aşağıdaki gibi elde edilir:

$$E(F_\varepsilon) = 0.75 \cdot E(X) + 0.25 \cdot E(Y), \tag{7.7}$$

$$E(F_\varepsilon) = 0.75 \cdot 0 + 0.25 \cdot 10 = 2.5. \tag{7.8}$$

Eş. 7.4 için varyans ise, Eş. 7.9'daki gibi ifade edilir:

$$V(F_\varepsilon) = V((1-\varepsilon)\cdot F_X) + V(\varepsilon \cdot F_Y) + 2\cdot cov((1-\varepsilon)\cdot F_X), (\varepsilon \cdot F_Y)). \tag{7.9}$$

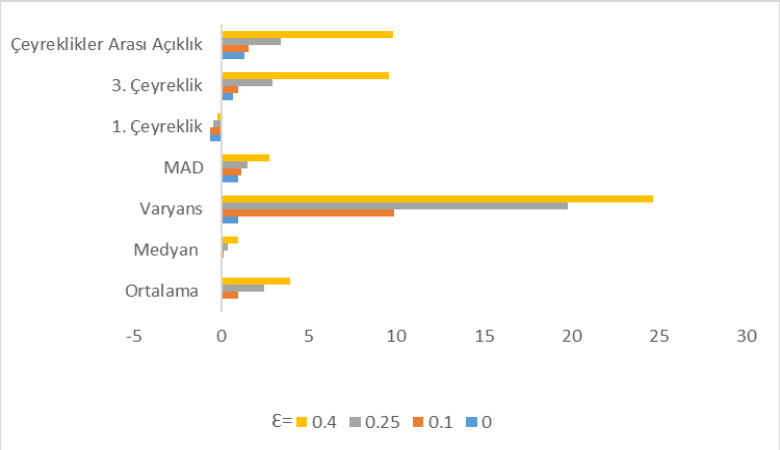
Varyans operatörü gereği Eş. 7.9'dan,

$$V(F_\varepsilon) = (1-\varepsilon)^2 \cdot V(X) + \varepsilon^2 \cdot V(Y) + 2(1-\varepsilon)\cdot(\varepsilon).cov(X,Y) \tag{7.10}$$

şeklinde ifade edilir. $\varepsilon = \{0.25\}$, $V(X) = 1$, $V(Y) = 1$ ve $cov(X,Y)$ değerleri Eş. 7.10'de yerine yazılırsa; $\varepsilon$ oranında karışmış dağılım fonksiyonunun varyans istatistiği değeri $V(F_\varepsilon) \approx 19.80$ olarak bulunur. Aşağıda yer alan Çizelge 7.1'de, $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ karışma oranları esas alınarak hesaplanan lineer fonksiyonellere yer verilmiştir.

Çizelge 7.1. Tek Değişkenli Yapıda Lokasyonu farklı Modelden Karışma Durumunda Tahmin Edicilerin değerleri ($n = 1000$, $X \sim N(0,1)$ ve $Y \sim N(10,1)$)

| $\varepsilon$ | 0.00 | 0.10 | 0.25 | 0.40 |
|---|---|---|---|---|
| Ortalama | 0.00 | 0.97 | 2.47 | 3.95 |
| Medyan | 0.00 | 0.12 | 0.41 | 0.96 |
| Varyans | 1.00 | 9.88 | 19.80 | 24.67 |
| MAD | 1.00 | 1.14 | 1.51 | 2.73 |
| 1.Çeyreklik | -0.65 | -0.60 | -0.44 | -0.19 |
| 3.Çeyreklik | 0.69 | 0.95 | 2.96 | 9.61 |
| Çeyreklikler Arası Açıklık | 1.35 | 1.55 | 3.41 | 9.81 |




Şekil 7.1. Tek Değişkenli Yapıda Lokasyonu farklı Modelden Karışma Durumunda Tahmin Edicilerin değerleri ($n = 1000$, $X \sim N(0,1)$ ve $Y \sim N(10,1)$)

### 7.1.2. Örnek : Ölçek parametresinin farklı olduğu karışma durumu

$X \sim N(0,1)$ şeklinde tanımlı $X$ rassal değişkeninin dağılımına, $Y \sim N(0,10)$ tanımlı $Y$ rassal değişkeninin dağılımının $\varepsilon = \{0.10, 0.25, 0.40\}$ oranlarında karışması halinde $\varepsilon$ oranda karışmış dağılım fonksiyonu Eş. 7.4'te tanımlanmıştır. Lineer fonksiyonel olarak tanımlanan beklenen değer ve varyans operatörünü kullanarak, ölçek karışması halinde $n = 1000$ gözlem birimi için karışma oranı $\varepsilon = \{0.40\}$ olması durumunda, karışma dağılımına ait ortalama ve varyans tahmin edicilerini hesaplayalım. $\varepsilon = \{0.40\}$ olması durumunda gözlemlerin yaklaşık 0.60'ı $X$ rassal değişkeninin, 0.40'ı $Y$ rassal değişkeninin dağılımından meydana gelmektedir. Eş. 7.4'nin beklenen değeri alındığında;

$$E(F_\varepsilon) = E((1-\varepsilon)\cdot F_X) + E(\varepsilon \cdot F_Y) \tag{7.11}$$

Eş. 7.11 elde edilir. Beklenen değer operatörü özelliği gereği Eş. 7.11;

$$E(F_\varepsilon) = (1-\varepsilon)\cdot E(X) + \varepsilon \cdot E(Y) \tag{7.12}$$

Eş. 7.12'teki gibi yazılabilir. $\varepsilon$= {0.40}, $E(X) = 0$, $E(Y) = 0$ değerleri Eş. 7.12'te yerine yazılırsa $\varepsilon$ oranda karışmış dağılım fonksiyonuna ilişkin beklenen değer;

$$E(F_\varepsilon) = 0.60 \cdot E(X) + 0.40 \cdot E(Y), \tag{7.13}$$

$$E(F_\varepsilon) = 0.60 \cdot 0 + 0.40 \cdot 10 = 0 \tag{7.14}$$

olarak elde edilir. Eş. 7.4'nin varyansı alındığında;

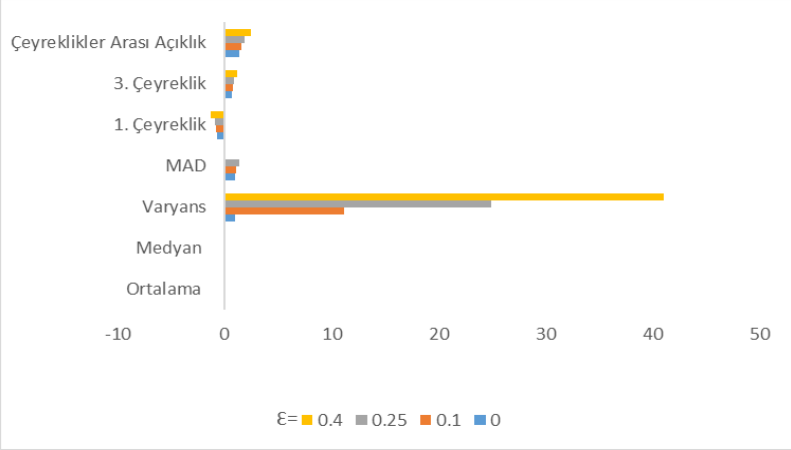
$$V(F_\varepsilon) = V((1-\varepsilon)\cdot F_X) + V(\varepsilon \cdot F_Y) + 2 \cdot cov((1-\varepsilon)\cdot F_X), (\varepsilon \cdot F_Y)) \tag{7.15}$$

$$V(F_\varepsilon) = (1-\varepsilon)^2 \cdot V(X) + \varepsilon^2 \cdot V(Y) + 2(1-\varepsilon)\cdot(\varepsilon).cov(X,Y) \tag{7.16}$$

Varyans operatörü gereği Eş. 7.15, Eş. 7.16 şeklinde yazılır. $\varepsilon$= {0.40}, $V(X) = 1$, $V(Y) = 10$ ve $cov(X,Y)$ değerleri Eş. 7.16'de yerine yazılırsa; $\varepsilon$ oranında karışmış dağılım fonksiyonunun varyans istatistiği değeri $V(F_\varepsilon) \approx 41.08$ olarak bulunur. Aşağıda yer alan Çizelge 7.2'de, $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ karışma oranları esas alınarak hesaplanan lineer fonksiyonellere yer verilmiştir.

Çizelge 7.2. Tek Değişkenli Yapıda Ölçek Parametresi Farklı Modelden Karışma Durumunda Tahmin Edicilerin değerleri ($n = 1000$, $X \sim N(0,1)$ ve $Y \sim N(0,10)$)

| $\varepsilon$ | 0.00 | 0.10 | 0.25 | 0.40 |
|---|---|---|---|---|
| Ortalama | 0.00 | 0.00 | 0.03 | -0.02 |
| Medyan | 0.00 | -0.01 | 0.02 | 0.00 |
| Varyans | 1.00 | 11.13 | 24.92 | 41.08 |
| MAD | 1.00 | 1.13 | 1.36 | 0.00 |
| 1.Çeyreklik | -0.68 | -0.77 | -0.91 | -1.25 |
| 3.Çeyreklik | 0.68 | 0.76 | 0.92 | 1.21 |
| Çeyreklikler Arası Açıklık | 1.37 | 1.53 | 1.84 | 2.47 |




Şekil 7.2. Tek Değişkenli Yapıda Ölçek Parametresi Farklı Modelden Karışma Durumunda Tahmin Edicilerin değerleri ($n = 1000$, $X \sim N(0,1)$ ve $Y \sim N(0,10)$)

Çizelge 7.1 ve 7.2'den görüldüğü gibi, başka dağılımdan karışan gözlem ya da gözlem grubu arttıkça lokasyon ve ölçek tahmin edicileri açısından örnek medyan istatistiğinin örnek ortalaması istatistiğine, örnek MAD istatistiğinin örnek varyans istatistiğine göre karışan gözlem ya da gözlem gurubundan fazla etkilenmediği gözlemlenmiştir. Bu durum başka dağılımdan gözlem ya da gözlem grubu karışması durumunda kırılma noktası yüksek ve sağlam bir tahmin edici olan medyan ve MAD istatistiğinin sahip olduğu özelliğin bir sonucudur.

## 7.2. Diskriminant analizinde sağlam ve klasik tahmin ediciler açısından gözlem birimlerinin hatalı sınıflandırılma oranları

Veri seti $p$ boyutlu, her grup ortalaması farklı, $g$ sayıda grup için çok değişkenli Normal dağılıma sahip yığından üretilmektedir (Eş.7.17):

$$\pi_j \sim N_p(\mu_j, \Sigma_j), \qquad j = 1, \cdots, g, \tag{7.17}$$

ve bu dağılımlar için sırasıyla, ölçek ve lokasyon karışması durumları Eş. 7.18 ve Eş. 7.19 senaryoları ile uygulanacaktır (Todorov ve Pires, 2007):

$$\pi_{j_\varepsilon} \sim (1-\varepsilon)\, N_p(\mu_j, I_p) + \varepsilon N_p(\mu_j, \kappa I_p) \qquad j = 1, \cdots, g \tag{7.18}$$

$$\pi_{j_\varepsilon} \sim (1-\varepsilon)\, N_p(\mu_j, I_p) + \varepsilon N_p(\hat{\mu}_j, \kappa^2 I_p) \qquad j = 1, \cdots, g$$

$$\hat{\mu}_j = \ \mu_j + (\nu Q_p, ..., \ \nu Q_p\,), \tag{7.19}$$

$$Q_p = \sqrt{\chi^2_{p;0.001}/p}$$

Burada $\varepsilon$, karışma oranı; $\pi_{j_\varepsilon}$, $j$. gruba $\varepsilon$ oranda karışmış çok değişkenli dağılımları; $\mu_j$, $j$. grubun ortalama vektörünü; $\kappa$, ölçek şişirme faktörünü; $\nu$, lokasyon şişirme faktörünü; $Q_p$, konum öteleme faktörünü; $\hat{\mu}_j$, j. gruba ilişkin $\varepsilon$ oranda karışmış çok değişkenli normal dağılım ortalama vektörünün parametre tahminini ifade etmektedir.

### 7.2.1. Örnek: Lokasyon ve Ölçek karışması durumunda klasik doğrusal diskriminant analizinde gözlem birimlerinin hatalı sınıflandırılma oranları

Çalışmanın bu kısmında; Eş.7.18 ve 7.19 esas alınarak $g = \{2, 3\}$ gruplu veri setleri için bir senaryo üretilmiştir: Bu senaryoya göre, Eş.7.20'de tanımlanan $p = \{2, 6\}$ değişkene sahip çok değişkenli normal dağılımdan $n = \{20, 100\}$ çaplı örnekler rassal olarak seçilmiş ve $\varepsilon = \{0, 0.10, 0.25, 0.40\}$, $\kappa = \{9, 100\}$, $\nu = \{5, 10\}$ değerleri için MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicileri kullanılarak klasik doğrusal diskriminant analizi yapılmıştır. Ele alınan örneklerde; gruplara ilişkin gözlem birimlerinin gelmiş olduğu yığınların varyans-kovaryans matrislerinin eşit ve $pxp$ boyutlu birim matris olduğu düşünülmektedir.

R paket programı kullanılarak 500 kez tekrar eden deneyde gözlem birimlerinin hatalı sınıflandırılma oranları için 8 ayrı durum; lokasyon ve ölçek karışması durumu için ayrı ayrı 3 farklı senaryo üzerinde incelenmiştir. İncelenen durumlarla ilgili bulgular paylaşılmış ve bu hususta grafiklere yer verilmiştir.

$$\mu_1 = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix}_p$$

$$\mu_2 = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}_p \tag{7.20}$$

$$\mu_3 = \begin{bmatrix} 2 & 2 & \dots & 2 \end{bmatrix}_p$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = I_p$$

Çizelge 7.3. $p = 2, n = n_1 = n_2 = 20$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| **Grup sayısı** | $\kappa$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 9 | 0 | 0.175 | 0.175 | 0.175 | 0.175 | 0.175 |
| | | 0.10 | 0.150 | 0.175 | 0.150 | 0.175 | 0.175 |
| | | 0.25 | 0.175 | 0.175 | 0.150 | 0.175 | 0.125 |
| | | 0.40 | 0.250 | 0.225 | 0.225 | 0.225 | 0.225 |
| | 100 | 0 | 0.175 | 0.175 | 0.175 | 0.175 | 0.175 |
| | | 0.10 | 0.200 | 0.175 | 0.175 | 0.175 | 0.175 |
| | | 0.25 | 0.350 | 0.300 | 0.300 | 0.300 | 0.325 |
| | | 0.40 | 0.375 | 0.350 | 0.400 | 0.350 | 0.375 |
| 3 | 9 | 0 | 0.216 | 0.233 | 0.233 | 0.233 | 0.266 |
| | | 0.10 | 0.266 | 0.316 | 0.300 | 0.316 | 0.350 |
| | | 0.25 | 0.333 | 0.316 | 0.300 | 0.316 | 0.350 |
| | | 0.40 | 0.316 | 0.333 | 0.316 | 0.333 | 0.316 |
| | 100 | 0 | 0.266 | 0.233 | 0.233 | 0.233 | 0.233 |
| | | 0.10 | 0.283 | 0.266 | 0.300 | 0.300 | 0.266 |
| | | 0.25 | 0.416 | 0.416 | 0.400 | 0.416 | 0.433 |
| | | 0.40 | 0.583 | 0.650 | 0.666 | 0.650 | 0.566 |

Çizelge 7.4. $p = 2, n = n_1 = n_2 = 100$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| Grup sayısı | $\kappa$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 9 | 0 | 0.280 | 0.270 | 0.270 | 0.270 | 0.280 |
| | | 0.10 | 0.270 | 0.265 | 0.270 | 0.265 | 0.260 |
| | | 0.25 | 0.285 | 0.285 | 0.280 | 0.285 | 0.285 |
| | | 0.40 | 0.285 | 0.300 | 0.275 | 0.305 | 0.305 |
| | 100 | 0 | 0.280 | 0.270 | 0.270 | 0.270 | 0.270 |
| | | 0.10 | 0.310 | 0.320 | 0.310 | 0.320 | 0.315 |
| | | 0.25 | 0.405 | 0.390 | 0.400 | 0.390 | 0.405 |
| | | 0.40 | 0.430 | 0.420 | 0.425 | 0.420 | 0.425 |
| 3 | 9 | 0 | 0.327 | 0.327 | 0.327 | 0.318 | 0.322 |
| | | 0.10 | 0.204 | 0.209 | 0.204 | 0.209 | 0.195 |
| | | 0.25 | 0.313 | 0.327 | 0.313 | 0.322 | 0.313 |
| | | 0.40 | 0.395 | 0.400 | 0.400 | 0.400 | 0.400 |
| | 100 | 0 | 0.318 | 0.318 | 0.331 | 0.322 | 0.327 |
| | | 0.10 | 0.350 | 0.359 | 0.372 | 0.359 | 0.372 |
| | | 0.25 | 0.422 | 0.445 | 0.427 | 0.445 | 0.436 |
| | | 0.40 | 0.486 | 0.454 | 0.468 | 0.454 | 0.459 |

Çizelge 7.5. $p = 6$, $n = n_1 = n_2 = 20$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| Grup sayısı | $\kappa$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 9 | 0 | 0.186 | 0.186 | 0.186 | 0.186 | 0.181 |
| | | 0.10 | 0.162 | 0.178 | 0.153 | 0.178 | 0.178 |
| | | 0.25 | 0.177 | 0.177 | 0.152 | 0.175 | 0.135 |
| | | 0.40 | 0.300 | 0.250 | 0.261 | 0.261 | 0.261 |
| | 100 | 0 | 0.200 | 0.200 | 0.200 | 0.210 | 0.195 |
| | | 0.10 | 0.250 | 0.200 | 0.200 | 0.200 | 0.199 |
| | | 0.25 | 0.400 | 0.400 | 0.400 | 0.323 | 0.395 |
| | | 0.40 | 0.420 | 0.400 | 0.405 | 0.400 | 0.395 |
| 3 | 9 | 0 | 0.230 | 0.240 | 0.240 | 0.240 | 0.288 |
| | | 0.10 | 0.288 | 0.321 | 0.315 | 0.330 | 0.364 |
| | | 0.25 | 0.400 | 0.350 | 0.392 | 0.356 | 0.356 |
| | | 0.40 | 0.440 | 0.400 | 0.400 | 0.390 | 0.390 |
| | 100 | 0 | 0.286 | 0.240 | 0.242 | 0.242 | 0.242 |
| | | 0.10 | 0.291 | 0.288 | 0.321 | 0.321 | 0.276 |
| | | 0.25 | 0.502 | 0.464 | 0.485 | 0.485 | 0.453 |
| | | 0.40 | 0.685 | 0.640 | 0.650 | 0.664 | 0.612 |

Çizelge 7.6. $p = 6$, $n = n_1 = n_2 = 100$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| Grup sayısı | $\kappa$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 9 | 0 | 0.102 | 0.102 | 0.102 | 0.081 | 0.090 |
| | | 0.10 | 0.116 | 0.140 | 0.133 | 0.144 | 0.144 |
| | | 0.25 | 0.134 | 0.134 | 0.121 | 0.141 | 0.103 |
| | | 0.40 | 0.254 | 0.220 | 0.221 | 0.243 | 0.243 |
| | 100 | 0 | 0.185 | 0.185 | 0.175 | 0.175 | 0.175 |
| | | 0.10 | 0.154 | 0.164 | 0.164 | 0.164 | 0.162 |
| | | 0.25 | 0.261 | 0.258 | 0.258 | 0.185 | 0.185 |
| | | 0.40 | 0.220 | 0.195 | 0.218 | 0.218 | 0.218 |
| 3 | 9 | 0 | 0.195 | 0.195 | 0.215 | 0.215 | 0.215 |
| | | 0.10 | 0.240 | 0.220 | 0.225 | 0.300 | 0.225 |
| | | 0.25 | 0.220 | 0.220 | 0.220 | 0.215 | 0.220 |
| | | 0.40 | 0.320 | 0.315 | 0.315 | 0.333 | 0.333 |
| | 100 | 0 | 0.200 | 0.220 | 0.185 | 0.185 | 0.185 |
| | | 0.10 | 0.200 | 0.215 | 0.300 | 0.321 | 0.276 |
| | | 0.25 | 0.200 | 0.215 | 0.244 | 0.200 | 0.215 |
| | | 0.40 | 0.324 | 0.320 | 0.325 | 0.320 | 0.300 |

Çizelge 7.7. $p = 2, n = n_1 = n_2 = 20$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| Grup sayısı | $\nu$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 5 | 0 | 0.081 | 0.071 | 0.080 | 0.073 | 0.073 |
| | | 0.10 | 0.140 | 0.071 | 0.077 | 0.070 | 0.070 |
| | | 0.25 | 0.145 | 0.071 | 0.076 | 0.070 | 0.070 |
| | | 0.40 | 0.156 | 0.065 | 0.074 | 0.081 | 0.091 |
| | 10 | 0 | 0.155 | 0.068 | 0.074 | 0.074 | 0.071 |
| | | 0.10 | 0.150 | 0.068 | 0.065 | 0.059 | 0.071 |
| | | 0.25 | 0.145 | 0.074 | 0.067 | 0.064 | 0.064 |
| | | 0.40 | 0.150 | 0.072 | 0.067 | 0.067 | 0.067 |
| 3 | 5 | 0 | 0.220 | 0.112 | 0.126 | 0.126 | 0.120 |
| | | 0.10 | 0.224 | 0.115 | 0.100 | 0.124 | 0.120 |
| | | 0.25 | 0.224 | 0.115 | 0.115 | 0.124 | 0.120 |
| | | 0.40 | 0.320 | 0.100 | 0.115 | 0.124 | 0.120 |
| | 10 | 0 | 0.340 | 0.321 | 0.320 | 0.315 | 0.315 |
| | | 0.10 | 0.320 | 0.321 | 0.320 | 0.310 | 0.340 |
| | | 0.25 | 0.342 | 0.314 | 0.320 | 0.310 | 0.340 |
| | | 0.40 | 0.342 | 0.314 | 0.300 | 0.300 | 0.300 |

Çizelge 7.8. $p = 2, n = n_1 = n_2 = 100$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| Grup sayısı | $\nu$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 5 | 0 | 0.075 | 0.071 | 0.074 | 0.070 | 0.070 |
| | | 0.10 | 0.135 | 0.071 | 0.077 | 0.070 | 0.062 |
| | | 0.25 | 0.135 | 0.071 | 0.076 | 0.070 | 0.062 |
| | | 0.40 | 0.146 | 0.065 | 0.074 | 0.081 | 0.081 |
| | 10 | 0 | 0.140 | 0.068 | 0.074 | 0.074 | 0.070 |
| | | 0.10 | 0.150 | 0.068 | 0.065 | 0.059 | 0.070 |
| | | 0.25 | 0.142 | 0.074 | 0.067 | 0.064 | 0.062 |
| | | 0.40 | 0.150 | 0.072 | 0.067 | 0.067 | 0.060 |
| 3 | 5 | 0 | 0.215 | 0.112 | 0.126 | 0.122 | 0.115 |
| | | 0.10 | 0.210 | 0.110 | 0.100 | 0.120 | 0.115 |
| | | 0.25 | 0.210 | 0.110 | 0.115 | 0.120 | 0.115 |
| | | 0.40 | 0.315 | 0.100 | 0.115 | 0.120 | 0.100 |
| | 10 | 0 | 0.340 | 0.300 | 0.320 | 0.310 | 0.300 |
| | | 0.10 | 0.320 | 0.320 | 0.320 | 0.305 | 0.320 |
| | | 0.25 | 0.342 | 0.310 | 0.320 | 0.305 | 0.320 |
| | | 0.40 | 0.342 | 0.300 | 0.300 | 0.301 | 0.300 |

Çizelge 7.9. $p = 6$, $n = n_1 = n_2 = 20$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| Grup sayısı | $\nu$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 5 | 0 | 0.075 | 0.092 | 0.074 | 0.072 | 0.071 |
| | | 0.10 | 0.135 | 0.092 | 0.079 | 0.072 | 0.071 |
| | | 0.25 | 0.140 | 0.100 | 0.079 | 0.081 | 0.071 |
| | | 0.40 | 0.140 | 0.102 | 0.100 | 0.081 | 0.071 |
| | 10 | 0 | 0.150 | 0.070 | 0.075 | 0.070 | 0.082 |
| | | 0.10 | 0.150 | 0.070 | 0.065 | 0.071 | 0.071 |
| | | 0.25 | 0.140 | 0.075 | 0.067 | 0.071 | 0.096 |
| | | 0.40 | 0.125 | 0.096 | 0.074 | 0.061 | 0.096 |
| 3 | 5 | 0 | 0.220 | 0.112 | 0.128 | 0.120 | 0.134 |
| | | 0.10 | 0.200 | 0.115 | 0.102 | 0.120 | 0.134 |
| | | 0.25 | 0.200 | 0.115 | 0.115 | 0.120 | 0.134 |
| | | 0.40 | 0.200 | 0.112 | 0.115 | 0.120 | 0.134 |
| | 10 | 0 | 0.300 | 0.300 | 0.320 | 0.285 | 0.312 |
| | | 0.10 | 0.310 | 0.300 | 0.300 | 0.285 | 0.312 |
| | | 0.25 | 0.310 | 0.320 | 0.300 | 0.300 | 0.300 |
| | | 0.40 | 0.242 | 0.320 | 0.323 | 0.300 | 0.300 |

Çizelge 7.10. $p = 6$, $n = n_1 = n_2 = 100$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

| Grup sayısı | $\nu$ | $\varepsilon$ | MLE | MCD | MCD-A | MCD-B | MCD-C |
|---|---|---|---|---|---|---|---|
| 2 | 5 | 0 | 0.070 | 0.085 | 0.072 | 0.070 | 0.074 |
| | | 0.10 | 0.130 | 0.085 | 0.074 | 0.070 | 0.074 |
| | | 0.25 | 0.135 | 0.092 | 0.070 | 0.074 | 0.072 |
| | | 0.40 | 0.135 | 0.092 | 0.092 | 0.074 | 0.061 |
| | 10 | 0 | 0.140 | 0.065 | 0.071 | 0.062 | 0.081 |
| | | 0.10 | 0.140 | 0.065 | 0.060 | 0.071 | 0.070 |
| | | 0.25 | 0.130 | 0.070 | 0.062 | 0.074 | 0.092 |
| | | 0.40 | 0.120 | 0.090 | 0.072 | 0.043 | 0.094 |
| 3 | 5 | 0 | 0.200 | 0.110 | 0.096 | 0.100 | 0.130 |
| | | 0.10 | 0.190 | 0.110 | 0.090 | 0.100 | 0.130 |
| | | 0.25 | 0.200 | 0.110 | 0.096 | 0.100 | 0.132 |
| | | 0.40 | 0.190 | 0.100 | 0.092 | 0.112 | 0.130 |
| | 10 | 0 | 0.200 | 0.285 | 0.300 | 0.200 | 0.300 |
| | | 0.10 | 0.220 | 0.240 | 0.285 | 0.220 | 0.300 |
| | | 0.25 | 0.220 | 0.300 | 0.285 | 0.200 | 0.280 |
| | | 0.40 | 0.220 | 0.300 | 0.300 | 0.200 | 0.280 |

### 7.2.2. Ölçek karışması durumu:

Senaryo 1: değişken sayısı aynı fakat gruplara ilişkin gözlem sayıları farklı ise:
(Çizelge 7.3 ve Çizelge 7.4)

1) İki ve üç gruplu çok değişkenli normal dağılıma $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ oranında karışmış veri seti için MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicileri kullanıldığında, klasik doğrusal diskriminant analizi sonucu gözlemlerin hatalı sınıflandırma oranları incelendiğinde; değişken sayısı sabit kalıp gruplardaki gözlem sayısı arttığında dağılıma dışardan karışan gözlem ya da gözlemler olmaması halinde; ($\varepsilon = 0$) MLE tahmin edicileri ile klasik doğrusal diskriminant analizi yapmanın hatalı sınıflandırma oranları bakımından MCD, MCD-A, MCD-B, MCD-C tahmin edicilerine göre daha iyi sonuç verdiği gözlemlenmiştir.

2) Aynı düzey grup ve değişken sayısı söz konusu iken, gruplardaki gözlem sayısı ve $\varepsilon$ karışma oranı arttığında; MCD tahmin edicisine dayalı algoritmaların klasik MLE tahmin edicisine göre gözlemlerin hatalı sınıflandırma oranı daha düşük sonuç verdiği gözlemlenmiştir.

3) Aynı düzey grup ve değişken sayısı söz konusu iken, gözlem sayısı ile birlikte $\kappa$ ölçek şişirme faktöründeki artış; MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicilerine dayalı algoritmalar ile yapılan hatalı sınıflandırma oranını arttırmaktadır.

4) Aynı düzey grup, $\varepsilon$, ölçek şişirme faktörü, değişken sayısı söz konusu iken, gruplardaki gözlem birimi sayısındaki artış MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicileri için benzer sonuçlar vermektedir.

Senaryo 2: değişken sayısı farklı, fakat gruplara ilişkin gözlem sayıları aynı ise:
(Çizelge 7.3 ve Çizelge 7.5)

1) İki ve üç gruplu çok değişkenli normal dağılıma $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ oranında karışmış veri seti için MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicileri kullanıldığında, klasik doğrusal diskriminant analizi sonucu gözlemlerin hatalı sınıflanırma oranları incelendiğinde; değişken sayısı artıp, gruplardaki gözlem sayısı sabit kaldığında dağılıma dışardan karışan gözlem ya da gözlemler olmaması halinde; ($\varepsilon = 0$) MLE tahmin edicileri ile klasik doğrusal diskriminant analizi yapmanın hatalı sınıflandırma oranları bakımından MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicileri birbirine benzer sonuçlar vermektedir.

2) Aynı düzey grup ve gruplardaki gözlem sayısı eşit ise; değişken sayısı ile birlikte $\kappa$ ölçek şişirme faktöründeki artış; MLE tahmin edicilerine dayalı algoritmalar ile yapılan hatalı sınıflandırma oranını , MCD, MCD-A, MCD-B, MCD-C tahmnin edicilerine göre daha fazla arttırmaktadır.

3) Gruplardaki karışma oranı arttıkça; karışma oranından en olumsuz etklinen MLE tahmin edicisine dayalı algoritme sonucu elde edilen sınıflandırma oranları olmaktadır.

Senaryo 3: gruplardaki değişken sayısı sabit ve gözlem sayısı bütün gruplar için aynı fakat grup sayıları farklı ise: (Çizelge 7.3)

1) Gruplardaki gözlem sayısı aynı ve grup değişkeni sabit ise $\kappa$ ölçek şişirme faktörü ve grup sayısındaki artış; MLE ve MCD temelli tahmin edicilerle yapılan hatalı sınıflandırma oranlarını birbirine benzer kılmaktadır.

2) Gruplardaki karışma oranı arttıkça; MLE tahmin edicileri MCD temelli tahmin edicilere göre sınıflandırma oranları bakımından daha olumsuz etkilenmektedir.

### 7.2.3. Lokasyon karışması durumu:

Senaryo 1: değişken sayısı aynı fakat gruplara ilişkin gözlem sayıları farklı ise : (Çizelge 7.7 ve Çizelge 7.8)

1) İki ve üç gruplu çok değişkenli normal dağılıma $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ oranında karışmış veri seti için MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicileri kullanıldığında, klasik doğrusal diskriminant analizi sonucu gözlemlerin hatalı sınıflanırma oranları incelendiğinde; değişken sayısı aynı kalıp, gruplardaki gözlem sayısı arttığında dağılıma dışardan karışam gözlem ya da gözlemler olmaması halinde; ($\varepsilon = 0$) MCD temelli tahmin ediciler ile klasik doğrusal diskriminant analizi yapmanın hatalı sınıflandırma oranları bakımından MLE tahmin edicisine göre benzer sonuçlar verdiği gözlemlenmiştir.

2) Aynı düzey grup ve değişken sayısı söz konusu iken; gruplardaki gözlem sayısı ve $\varepsilon$ karışma oranı arttığında; MCD tahmin edicisine dayalı algoritmaların klasik MLE tahmin edicisine göre çok daha iyi sonuç verdiği, yani hatalı sınıflandırma oranlarının oldukça düşük olduğu gözlemlenmiştir.

3) Aynı düzey grup ve değişken sayısı söz konusu iken, gözlem sayısı ile birlikte $\nu$ konum şişirme faktöründeki artış; MCD temelli tahmin edicilerle klasik doğrusal diskriminant analizi yapmanın hatalı sınıflandırma oranları bakımından, MLE tahmin edicisine göre daha iyi sonuçlar verdiği gözlemlenmiştir.

Senaryo 2: değişken sayısı farklı fakat gruplara ilişkin gözlem sayıları aynı ise:
(Çizelge 7.7 ve Çizelge 7.9)

1) İki ve üç gruplu çok değişkenli normal dağılıma $\varepsilon = \{0, 0.10, 0.25, 0.40\}$ oranında karışmış veri seti için MLE, MCD, MCD-A, MCD-B, MCD-C tahmin edicileri kullanıldığında, klasik doğrusal diskriminant analizi sonucu gözlemlerin hatalı sınıflanırma oranları incelendiğinde; değişken sayısı artıp, gruplardaki gözlem sayısı sabit kaldığında dağılıma dışardan karışan gözlem ya da gözlemler olmaması halinde; ($\varepsilon = 0$) MCD temelli tahmin ediciler ile klasik doğrusal diskriminant analizi yapmanın hatalı sınıflandırma oranları bakımından MLE tahmin edicisine göre daha iyi sonuçlar verdiği gözlemlenmiştir.

2) Aynı düzey grup ve gruplardaki gözlem sayısı eşit ise; değişken sayısı ile birlikte $\nu$ lokasyon şişirme faktöründeki artış; MLE tahmin edicilerine dayalı algoritmalar ile yapılan hatalı sınıflandırma oranını, MCD, MCD-A, MCD-B, MCD-C tahmnin edicilerine göre çok daha hızlı arttırmaktadır.

3) Gruplardaki karışma oranı arttıkça, karışma oranından en olumsuz etklinen MLE tahmin edicisine dayalı algoritma sonucu elde edilen sınıflandırma oranları olmaktadır.

Durum 3: gruplardaki değişken sayısı sabit ve gözlem sayısı bütün gruplar için aynı fakat grup sayıları farklı ise: (Çizelge 7.7)

1) Gruplardaki gözlem sayısı aynı ve grup değişkeni sabit ise $\nu$ lokasyon şişirme faktörü ve grup sayısındaki artış; MCD temelli tahmin edicilerle yapılan klasik doğrusal diskriminant analizinde hatalı sınıflandırma oranlarının MLE tahmin edicilerine göre daha düşük çıktığını göstermiştir.

2) Gruplardaki karışma oranı arttıkça; karışma oranından MLE tahmin edicileri MCD temelli tahmin edicilere göre çok daha olumsuz etkilenmektedir.

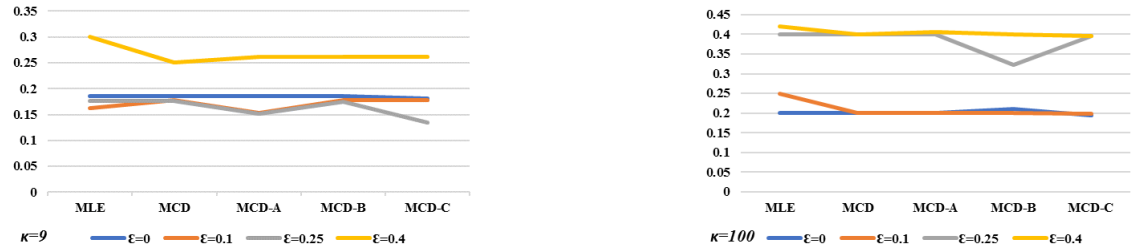
Aşağıda yer alan grafiklerde MLE, MCD, MCD-A, MCD-B MCD-C tahmin edicileri kullanılarak, çok değişkenli normal dağılama sahip dağılımlarda, dağılımın lokasyon ve ölçek parametresine karışma durumunda klasik doğrusal diskriminant analizi için elde edilmiş hatalı sınıflandırma oranları Eş.7.20 için farklı lineer fonksyioneller, $\varepsilon$, $\nu$, $\kappa$, $p$ ve $n$ için ayrı ayrı gösterilmiştir.


Şekil 7.3. İki gruplu veride $p = 2$, $n = n_1 = n_2 = 20$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.4. Üç gruplu veride $p = 2$, $n = n_1 = n_2 = 20$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

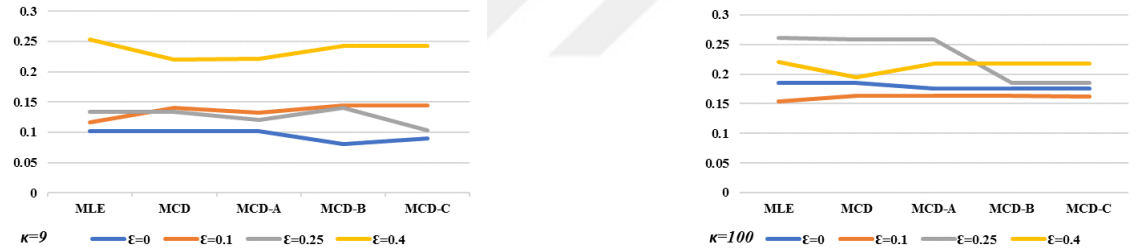

Şekil 7.5. İki gruplu veride $p = 2$, $n = n_1 = n_2 = 100$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.6. Üç gruplu veride $p = 2$, $n = n_1 = n_2 = 100$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları




Şekil 7.7. İki gruplu veride $p = 6$, $n = n_1 = n_2 = 20$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.8. Üç gruplu veride $p = 6$, $n = n_1 = n_2 = 20$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

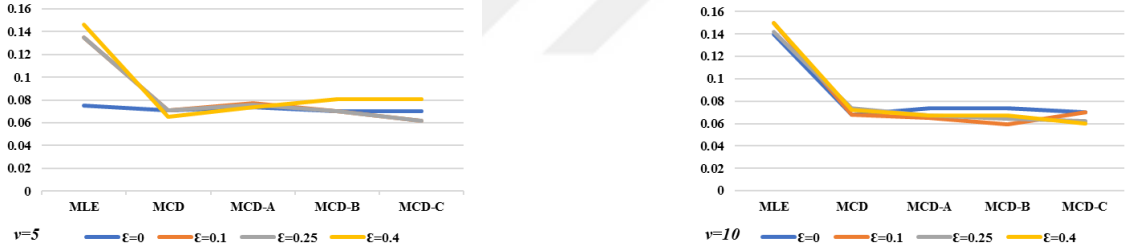



Şekil 7.9. İki gruplu veride $p = 6$, $n = n_1 = n_2 = 100$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.10. Üç gruplu veride $p = 6$, $n = n_1 = n_2 = 100$ iken ölçek tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.11. İki gruplu veride $p = 2$, $n = n_1 = n_2 = 20$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.12. Üç gruplu veride $p = 2$, $n = n_1 = n_2 = 20$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları




Şekil 7.13. İki gruplu veride $p = 2$, $n = n_1 = n_2 = 100$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.14. Üç gruplu veride $p = 2$, $n = n_1 = n_2 = 100$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.15. İki gruplu veride $p = 6$, $n = n_1 = n_2 = 20$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.16. Üç gruplu veride $p = 6$, $n = n_1 = n_2 = 20$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.17. İki gruplu veride $p = 6$, $n = n_1 = n_2 = 100$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları


Şekil 7.18. Üç gruplu veride $p = 6$, $n = n_1 = n_2 = 100$ iken lokasyon tahmin edicisine karışma durumunda hatalı sınıflandırma oranları

# 8 . SONUÇ VE ÖNERİLER

Bu çalışmada, aykırı gözlem ya da başka dağılımdan karışmış gözlemin olması durumunda klasik doğrusal diskriminant analizi için MLE ve sağlamlık vasfına sahip çeşitli MCD metotları ve bu metotların gözlem birimlerinin hatalı sınıflandırma oranları incelenmeye çalışılmıştır. Özellikle, aykırı gözlem ya da başka dağılımdan karışmış gözlemler olduğu zaman klasik doğrusal diskriminant analizinin nasıl etkilendiği, gözlemlerin hatalı sınıflandırma oranlarının gerçekte olması beklenen durumdan oldukça farklılık gösterdiği ve böyle durumlarda sağlam istatistiklerle çalışmanın, MLE tahmin edicilerine göre daha doğru sonuçlar verdiği gösterilmeye çalışılmıştır ve aşağıda bu durumlara ilişkin sonuçlara yer verilmiştir.

Uygulama verisi olarak seçilen veri setlerinden, tek gruplu HBK veri setinde dikkate alınan 3 değişken için klasik tahmin edicilerle Mahalanobis uzaklık ölçüleri uygulandığında gözlemlerin yaklaşık olarak %2.67'si (2/75 =0.0267) aykırı gözlem olarak tespit edilirken; sağlam uzaklık ölçüleri uygulandığında gözlemlerin yaklaşık olarak %18.67'si (14/75 =0.1867) aykırı gözlem olarak tespit edilmiştir. HBK verisi, kendi içinde aykırı gözlem barındıran bir veri seti olarak hazırlanmıştır ( Hawkins ve diğerleri, 1984). Klasik tahmin ediciler bu aykırı gözlemleri tespit etmede oldukça başarısız iken, sağlam tahmin ediciler aykırı gözlemleri tespit etmede daha başarılı olmaktadırlar.

Alcohol veri seti 6 değişken ve 3 grup içermektedir. Klasik, MCD ve MCD-A metodu ile elde edilen sınıflandırma tabloları farklı olmasına rağmen gözlemlerin hatalı sınıflandırılma oranları aynı çıkmıştır. Öte yandan MCD-B ve MCD-C gerek sınıflandırma tabloları gerekse de gözlemlerin hatalı sınıflandırılma oranları bakımından daha güvenilir sonuçlar vermektedir.

Bu çalışama kapsamında ele alınan MLE, MCD, MCD-A, MCD-B ve MCD-C tahmin edicileri ile farklı karışma oranı, gözlem sayısı, değişken sayısı ve grup sayısı baz alınarak çok değişkenli normal dağılıma sahip veri setleri üzerinde yapılan diskriminant analizi için yapılan simülasyonlarda karışma oranı arttıkça, sağlamlık özelliğine sahip MCD'ye dayalı tahmin edicilerin, klasik MLE'ne göre gözlem birimlerinin hatalı sınıflandırma oranları bakımından daha iyi sonuçlar verdiği gözlemlenmiştir. Dağılım yapısına dışardan bir gözlem ya da gözlem grubunun karışmadığı durumda klasik MLE tahmin edicileri en iyi sonucu vermektedir.

# KAYNAKLAR


Akdi, Y. (2005). *Matematiksel İstatistiğe Giriş* (Birinci Baskı). Ankara: Bıçaklar Kitabevi.

Alpar, C. R. (2011). *Uygulamalı Çok Değişkenli İstatistiksel Yöntemler.* (Üçünücü Baskı). Ankara: Detay Anatolia Akademik Yayıncılık Ltd. Şti.

Alrawashdeh, M., Radwan, T., Abunawas, K. (2018). Performance of Linear Discriminant Analysis Using Different Robust Methods. *European Journal of Pure and Applied Mathematics.* 11(1), 284.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. (Second Edition). New York: John Wiley Sons Incorporated Company.

Arık, İ. (2014). *Dağılım Parametrelerinin Tahmininde Kullanılan Robust Yöntemler.* Yüksek Lisans Tezi, Anadolu Üniversitesi, Eskişehir.

Casella, G. ve Berger R. L. (2002). *Statistical Inference* (Second Edition). United States: Duxbery Thompson Learning Incorporated Company.

Chork, C. ve Rousseeuw, P.J. (1992). İntegrating a High Breakdown Option İnto Discriminant Analysis in Exploration Geochemistry, *Journal of Geochemical Exploration*, 43, 191-203.

Croux, C. ve Dehon, C. (2001). Robust Linear Discriminant Analysis using S-Estimators, *The Canadian Journal of Statististics,* 29, 473-492.

Güriş, S. ve Çağlayan, E. (2010). *Ekonometri-Temel Kavramlar* (İkinci Baskı). İstanbul: Der yayınları.

Gündüz, N. ve Fokoue, E. (2017). Predictive Performances of Implicity and Explicity Robust Classifiers On High Dimensional Data. *Communications Faculty of Science University of Ankara Series A1 Mathematics and Statistics,* 66(2), 14-36.

Hawkins, D.M . ve McLachlan, G. (1997). High-breakdown Linear Discriminant Analysis, *Journal of the Ameraican Statistical Association*, 92, 136-143.

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984) Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26, 197--208.

He, X. ve Fung, W. (2000). High-Breakdown Estimation for Multiple Populations with Applications to Discriminant Analysis, *Journal of Multivariate Analysis*, 72, 151-162.

Hogg, R. V. ve Craig, A. T. (1978). *İntroduction to Mathematical Statistics* (Fourth Edition) . NewYork: Macmillan Publishing Incorporated Company.

Huber P. J. Ve Ronchetti, E. M. (2009). *Robust Statistics* (Second Edition), İngiltere: A John Wiley & Sons Incorporated Company Publication.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35 (1), 73-101.

Hubert, M. ve Debruyne M. (2010). Minimum Covariance Determinant. *John Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 36-43.

Hubert, M. ve Van Driessen, K. (2004). Fast and Robust Discriminant Analysis, *Computational Statistics and Data Analysis*, 45 301-320.

Hubert, M. Debruyne M. ve Rousseeuw, P. (2018). Minimum Covariance Determinant and Extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*. 10(3), 1-11.

Jureckova, J. ve Picek, J. (2006). *Robust Statistical Methods with* R (Second Editinon), London : Chapman and Hall.

Maronna, R. A., Martin, R. D. ve Yohaai, V. J. (2006). *Robust Statistics: Theory and Methods.* (Sixth edition). New York: John Wiley Sons Incorporated Company.

Rousseeuw, P. J. ve Hubert, M. (2011). Robust Statistics For Outlier Detection. *John Wiley Sons Reviews Incorporated Company*, 1 (1), 73-79.

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

Syed, Y., Sharipah, S., Lim, Yai F., Ali, H. ve Zurni, O. (2016). Robust Linear Discriminant Analysis. *Journal of Mathematics and Statistics.* 12(4) 312-316.

Todorov, V. ve Filzmoser, P. (2009). An Object-Oriented Framework for robust Multivariate Analysis. *Journal of Statistical Software*, 32 (3), 585-604.

Todorov, V. ve Filzmoser, P. (2013). Robust Tools for the İmperfect World. *Elsevier Jounal*, 245(1), 4-20.

Todorov, V. ve Pires, A.M. (2007). Comparative Performance of Several Robust Linear Disciriminant Analysis. *REVSTAT-Statistical Journal*, 5(1), 63-83.

# ÖZGEÇMİŞ

## Kişisel Bilgiler


| | |
|---|---|
| Soyadı, adı | : SEZER, Sercan |
| Uyruğu | : T.C. |
| Doğum tarihi ve yeri | : 01.08.1990, Ankara |
| Medeni hali | : Bekar |
| Telefon | :+90 (222) 231 54 80 |
| E-mail | : sercan.sezer24@gmail.com |


## Eğitim

| Derece | Eğitim Birimi | Mezuniyet Tarihi |
|---|---|---|
| Yüksek lisans | Gazi Üniversitesi / İstatistik | Devam ediyor. |
| Lisans | Gazi Üniversitesi/ Ekonometri Çiftanadal | 2013 |
| Lisans | Gazi Üniversitesi / İstatistik Bölümü | 2013 |
| Lise | Yıldırım Beyazit Lisesi (Y.D.A) | 2007 |

## İş Deneyimi

| Yıl | Yer | Görev |
|---|---|---|
| 2016-Halen | Sosyal Güvenlik Kurumu (Eskişehir) | Sosyal Güvenlik Denetmeni |

## Yabancı Dil

İngilizce

## Yayınlar

- Sezer, S. ve Gündüz, N. (2015). *Sağlamlık Özelliğine Sahip En Küçük Kovaryans Determinantına İlişkin Tahmin Ediciler*, Antalya: Türk İstatistik Derneği ve Eskişehir Osmangazi Üniversitesi 9. Uluslararası İstatistik Kongresi Bildiri Özeti Kitabı (28 Ekim- 1 Kasım 2015), 81.

## Hobiler

Türk halk müziği dinlemek, bağlama çalmak, tiyatroya gitmek, kitap okumak

*GAZİ GELECEKTİR..*