T.C.
ERCİYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
MATEMATİK ANABİLİM DALI

KARMA DAĞILIM MODELLERİ KULLANILARAK ÇOK DEĞİŞKENLİ VERİDE GRUP YAPILARININ BELİRLENMESİ, AYRIŞTIRILMASI, KÜMELENMESİ VE SINIFLANDIRILMASI

Hazırlayan
Maruf GÖGEBAKAN

Danışman
Prof. Dr. Hamza EROL

Doktora Tezi

Nisan 2017
KAYSERİ

T.C.
ERCİYES ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
MATEMATİK ANABİLİM DALI

KARMA DAĞILIM MODELLERİ KULLANILARAK ÇOK DEĞİŞKENLİ VERİDE GRUP YAPILARININ BELİRLENMESİ, AYRIŞTIRILMASI, KÜMELENMESİ VE SINIFLANDIRILMASI

(Doktora Tezi)

Hazırlayan
Maruf GÖGEBAKAN

Danışman
Prof. Dr. Hamza EROL

Nisan 2017
KAYSERİ

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Maruf GÖGEBAKAN

YÖNERGEYE UYGUNLUK SAYFASI

"Karma dağılım modelleri kullanılarak çok değişkenli veride grup yapılarının belirlenmesi, ayrıştırılması, kümelenmesi ve sınıflandırılması" adlı Doktora tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi'ne uygun olarak hazırlanmıştır.

Hazırlayan
Maruf GÖGEBAKAN

Danışman
Prof. Dr. Hamza EROL

Matematik Anabilim Dalı Başkanı
Prof. Dr. Fuat GÜRCAN

Prof. Dr. Hamza EROL danışmanlığında **Maruf GÖGEBAKAN** tarafından hazırlanan "**Karma Dağılım Modelleri Kullanılarak Çok Değişkenli Veride Grup Yapılarının Belirlenmesi, Ayrıştırılması, Kümelenmesi ve Sınıflandırılması**" adlı bu çalışma, jürimiz tarafından Erciyes Üniversitesi Fen Bilimleri Enstitüsü Matematik Anabilim Dalında **Doktora** tezi olarak kabul edilmiştir.

21 /04/2017

JÜRİ:

Danışman	: Prof. Dr. Hamza EROL	.....
Üye	:Doç. Dr. Tayfun SERVİ	.....
Üye	: Doç. Dr. Pakize TEMTEK	.....
Üye	:Doç. Dr. Ali DELİCEOĞLU	.....
Üye	:Yrd. Doç.Dr. M. Tarık ATAY	.....

ONAY:

Bu tezin kabulü Enstitü Yönetim Kurulunun 25/04/2017 tarih ve 2017/19-16. sayılı kararı ile onaylanmıştır.

...25.../04.../2017

Prof. Dr. Mehmet Akkurt

Enstitü Müdürü

TEŞEKKÜR

Çalışmalarım boyunca farklı bakış açıları ve bilimsel katkılarıyla beni aydınlatan, yakın ilgi ve yardımlarını esirgemeyen, görüş ve önerileriyle daima yolumu açan değerli hocam Prof. Dr. Hamza EROL ve ikinci danışman hocam Doç. Dr. V. Çağrı GÜNGÖR'e fedakârlıklarından dolayı teşekkür ederim. Eğitim hayatına başladığım günden itibaren üzerimde emeği bulunan tüm hocalarıma, gelişmemizde bizleri etkileyen tüm fikir adamlarına, dünya ve ahiret hayatını öğrenme ve yaşamada bizlere öncülük eden tüm büyüklerimize ve bu yolda beraber yürüdüğümüz tüm arkadaşlarıma teşekkür ederim. Üniversite hayatımda bizlerden emeğini ve bilgisini esirgemeyen, gerek akademik hayatta gerekse sosyal hayatta bizlere her zaman iyi insan olmayı aşılayan, fikri dünyamıza yön veren ve insanlığa faydalı olmayı önceleyen Erciyes Üniversitesi Matematik bölümündeki değerli hocalarıma, bizlere kattıkları bu erdemli değerlerden dolayı ayrıca teşekkür ederim. Akademik çalışma hayatımda aynı odayı paylaştığım araştırma görevlisi arkadaşlarıma, yardımları ve sabırlarından dolayı teşekkür ederim.

Ayrıca; hayatım boyunca bana desteklerini esirgemeyen onca zahmete katlanıp bizlerin eğitimine devam etmesini sağlayan çok değerli annem, babam ve kardeşlerime, evlendikten sonraki hayatımda her zaman beni destekleyen çalışmalarım süresince sabırla fedakârlıkta bulunan çok değerli eşim Fazilet hanıma, kıymetli çocuklarım Ahmet Erdem ve Muhammed Emin'e en içten teşekkürlerimi sunarım.

Maruf GÖĞEBAKAN

Nisan 2017, KAYSERİ

KARMA DAĞILIM MODELLERİ KULLANILARAK ÇOK DEĞİŞKENLİ VERİDE GRUP YAPILARININ BELİRLENMESİ, AYRIŞTIRILMASI, KÜMELENMESİ VE SINIFLANDIRILMASI

Maruf GÖGEBAKAN

Erciyes Üniversitesi, Fen Bilimleri Enstitüsü
Doktora Tezi, Nisan 2017
Danışman: Prof. Dr. Hamza EROL

KISA ÖZET

Karma ayrıştırma analizine dayalı sınıflandırma, nesnelerin ya da gözlemlerin farklı kümelerinin ayrımı ve önceden tanımlanmış grupların yeniden düzenlenmesi ile ilgilenen çok değişkenli yöntemlerdir. Karma kümeleme analizine dayalı sınıflandırma, nesnelerin ya da gözlemlerin farklı kümelerinin ayrımı ve önceden tanımlanmamış grupların yeniden düzenlenmesi ile ilgilenen çok değişkenli yöntemlerdir. Karma ayrıştırma analizi ve karma kümeleme analizi genellikle homojen olmayan veride sıradan ilişkilerin iyi anlaşılmadığı durumlarda gözlenen farklılıkları araştırmak için kullanılır. Ayrıştırma analizinin ve kümeleme analizinin sınıflandırma temelinde iki amacı vardır. Birincisi, bilinen kitledeki, farklı özelliklere sahip gözlemleri grafiksel veya işlemsel olarak tanımlamaktır. İkincisi, gözlemleri iki ya da daha fazla isimlendirilmiş kümelere, gruplara ya da sınıflara ayırmaktır.

Bu çalışmada homojen olmayan çok değişkenli verideki grup sayılarının ve yapılarının modele dayalı olarak belirlenmesinde karma dağılım modelleri kullanılacaktır. Bu tip verilerde homojen olmayan yapının ortaya çıkarılmasında kullanılan yöntemler incelenecektir. Homojen olmayan çok değişkenli verideki grup sayılarının belirlenmesinde kullanılan yöntemler araştırılacaktır. Homojen olmayan çok değişkenli veride belirlenen grupların büyüklükleri, gruplar arasındaki ilişkiler ifade edilecektir. Homojen olmayan çok değişkenli verideki grup yapılarının modellenmesi, fonksiyonlarının oluşturulması ve uygulamaları ele alınacaktır. Homojen olmayan çok değişkenli verinin ayrıştırılması, kümelenmesi ve sınıflandırılması araştırılacaktır.

Anahtar Kelimeler: Kümeleme, Sınıflandırma ve Ayrıştırma analizi; Modele dayalı kümeleme; Bileşen küme sayısı; Sonlu karma dağılım; Çok değişkenli normal karma dağılım modeli; K-ortalamalar algoritması.

DETERMINATION OF GROUP STRUCTURES IN MULTIVARIATE DATA, DISCRIMINATION, CLUSTERING AND CLASSIFICATION USING MIXTURE DISTRIBUTION MODELS

Maruf GÖGEBAKAN

Erciyes University, Graduate School of Natural and Applied Sciences
Ph. D Thesis, April 2017
Supervisor: Prof. Dr. Hamza EROL

ABSTRACT

Mixture discriminant analysis based on classification, separation of different sets of objects or observations and multivariate methods are dealing with the reorganization of pre-defined group. Mixed clustering analysis based on classification, separation of different sets of objects or observations and multivariate methods are dealing with the reorganization of the pre-defined groups. Mixture discriminant analysis and mixture cluster analysis is often used to investigate the differences observed in the case of ordinary relationships in non-homogeneous data are not well understood. Discriminant Analysis and classification on the basis of the cluster analysis has two purposes. First, the known mass is to identify operational observations graphically or with different features. Secondly, the observation of the two or more named sets to separate into groups or classes.

In this study will be mixture distribution models used to determine the number of groups in the non-homogeneous multivariate data and structure-based models. Such data will be analyzed in the methods used to reveal the structure non-homogeneous. Methods will be explored used in the determination of the groups in the number of non-homogeneous multivariate data. The size of the group of non-homogeneous multivariate data set, the relationships between the groups will be expressed. Multivariate modeling of inhomogeneous structures in the data set, the creation of functions and applications will be discussed. Non-homogeneous decomposition of multivariate data clustering and classification will be investigated.

Keywords: Clustering, Classification and Discriminant Analysis; Model Based Clustering; Numbers of Component Cluster; Finite Mixtures Distribution; Multivariate Normal Mixtures Distribution Model; K-means Algorithm.

İÇİNDEKİLER

KARMA DAĞILIM MODELLERİ KULLANILARAK ÇOK DEĞİŞKENLİ VERİDE GRUP YAPILARININ BELİRLENMESİ, AYRIŞTIRILMASI, KÜMELENMESİ VE SINIFLANDIRILMASI

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Karma dağılımlar; iki veya daha fazla bileşenden oluşan dağılımlardır. Karma dağılım modelleri olayların birçok farklı özelliklerine göre elde edilen verilerin istatistiksel modellemeler altında matematiksel olarak incelenmesidir. Bundan dolayı karma dağılım modelleri birçok alanda kullanılmaktadır. Bu alanların başında genetik, biyoloji, mühendislik, tıp, psikiyatri, astronomi, arkeoloji ve tarım uygulamaları gelmektedir. Karma dağılım modelleri yığılım analizi, ayrıştırma analizi ve görüntü analizi gibi istatistiksel sınıflandırma yöntemlerinde kullanılmaktadır. Ayrıştırma ve yığılım analizinde karma dağılım modelleri sınıflandırma yapmak için kullanılmaktadır. Sınıflandırma, kümeleme analizi olarak da adlandırılabilir. Kümeleme analizinde, homojen olmayan verinin küme sayısı ve küme yapısı incelenir. Kümelemede amaç, küme içerisindeki gözlemlerin ya da nesnelerin benzer diğer bir deyişle homojen ve kümelerinde birbirinden farklı yani heterojen olacak şekilde en uygun gruplama yapısını bulmaktır. Karma dağılım modellerindeki bileşen sayısının bilinmesi veya bilinmemesi kullanılacak yöntemi belirlemektedir. Ayrıştırma analizinde karma dağılım modellerindeki bileşen sayısı ön bilgi olarak verilmektedir (eğitimli). Ancak karma dağılım modellerine dayalı kümeleme analizinde modeldeki bileşen sayısı ön bilgi olarak verilmemektedir (eğitimsiz). Homojen olmayan çok değişkenli verideki grup yapılarının modellenmesi, fonksiyonlarının oluşturulması ve uygulamalarında karma dağılım modelleri kullanılmaktadır.

Karma ayrıştırma analizine dayalı sınıflandırma, önceden bilinen bileşen sayısına bağlı olarak gözlemlerin farklı kümelerinin ayrımı ve kümelerin yeniden düzenlenmesi ile ilgilenen çok değişkenli yöntemlerdir. Karma kümeleme analizine dayalı sınıflandırma, bileşen sayılarının ön bilgi olarak verilmediği durumlarda gözlemlerin farklı kümelerinin ayrımı ve kümelerin yeniden düzenlenmesi ile ilgilenen çok değişkenli yöntemlerdir. Karma ayrıştırma analizi ve karma kümeleme analizi çoğunlukla homojen olmayan, iki

veya daha fazla normalin karması, veride sıradan ilişkilerin iyi anlaşılmadığı durumlarda gözlenen farklılıkları araştırmak için kullanılır. Ayrıştırma analizinin ve kümeleme analizinin sınıflandırma temelinde iki amacı vardır. Birincisi, bilinen kitledeki, farklı özelliklere sahip gözlemleri grafiksel veya hesaplamalı yöntemler olarak tanımlamaktır. İkincisi, gözlemleri iki ya da daha fazla isimlendirilmiş kümelere, gruplara ya da sınıflara ayırmaktır.

Verideki kümelenmeler ve grup yapıları gözlemlerden elde edilen grafiklerden faydalanarak da belirlenebilir [1]. Verideki değişken sayısı iki olduğu durumda saçılım (scatter) grafiği küme sayısını belirlemede kullanılabilir. Verideki değişken sayısı ikiden fazla olması durumunda boyut indirgeme [2] ve temel bileşenler analizi kullanılarak iki boyuta indirgenebilir. Verideki değişkenlerin yapısına bağlı olarak homojen yapıdaki değişkenlerin heterojen veriler üzerine izdüşümü (projection) alınarak değişken sayısı azaltılabilir [1]. Çok değişkenli verinin grafiksel gösteriminden verinin kümelenebilir olup olmadığı hakkında ön bilgi elde edilebilir. Tek değişkenli veride tek kümelenmenin bulunması kümelenmenin olmadığı homojen yapıya denk gelmektedir. Diğer yandan iki veya daha fazla tepenin bulunduğu yapı iki veya daha fazla normalin karmasının olduğu yapıya karşılık gelmektedir.

Sonlu karma dağılım modelleri, modele dayalı kümeleme analizinde kullanılmaktadır. Kümeleme analizinde kullanılan yöntemlerin uygulamalarında ortaya çıkan sorunların çözümünde sonlu karma dağılım modelleri kullanılmaktadır [5], [6].

Modele dayalı kümeleme analizi, verideki küme sayısını ve yapısını inceleyen önemli bir kümeleme analizi yöntemidir. Kümelerin sayısı ve yapısının çok değişkenli karma dağılım modelleriyle analiz edilmesi karma model kümeleme analizi veya diğer bir ifadeyle modele dayalı kümeleme analizi olarak adlandırılır [5]. Modele dayalı kümeleme p-boyutlu çok değişkenli heterojen veriyi anlamlı alt gruplara bölmek için kullanılan yöntemlerden biridir [6]. Çok değişkenli normal dağılımların karmasının her bileşeni, çok değişkenli heterojen verideki bir kümeye karşılık gelir [7]. Çok değişkenli heterojen verideki kümelenmenin n tane p-boyutlu $\boldsymbol{x}_1,...,\boldsymbol{x}_n$ gözleminde her biri bilinmeyen $\pi_1,...,\pi_g$ olasılıkları ile sonlu sayıdaki g grup yoğunluklarının karmasından geldiği

varsayılır [8]. $j = 1,...,n$ için j. gözlem değeri $\boldsymbol{x}_j$ de varsayılan normal dağılımların karma modeli,

$$f(\boldsymbol{x}_j;\boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i f_i(\boldsymbol{x}_j;\boldsymbol{\psi}_i) \qquad (1)$$

şeklinde yazılır. Burada $i = 1,...,g$ için π_i, $0 < \pi_i < 1$ ve $\sum_{i=1}^{g} \pi_i = 1$ olacak biçimde i. kümelenme veya grup için karma oranını göstermektedir. $j = 1,...,n$ için grup koşullu yoğunluk $f_i(\boldsymbol{x}_j;\boldsymbol{\psi}_i)$, bilinmeyen parametreler vektörü $\boldsymbol{\psi}_i$'ye bağlıdır. Bu çalışmada $f_i(\boldsymbol{x}_j;\boldsymbol{\psi}_i)$'nin $\boldsymbol{\mu}_i$ ortalamalı $\boldsymbol{\Sigma}_i$ varyans-kovaryans matrisli çok değişkenli normal dağılım olduğu varsayılır. $\boldsymbol{\psi}_i$, bileşenlerin parametrelerinin vektörüdür. Böylece $\boldsymbol{\psi}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ şeklinde gösterilir. $f_i(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\Sigma_i)$ çok değişkenli yoğunluğu,

$$f_i(\boldsymbol{x}_j;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{p}{2}} \left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)\right\} \qquad (2)$$

yapısındadır. (2)'deki eşitlikte T üst indisi, matrisin devriğini (transpoz) göstermektedir. Böylece (1)'deki eşitlikteki $\boldsymbol{\theta} = (\pi_1,...,\pi_g,\boldsymbol{\psi}_1,...,\boldsymbol{\psi}_g)$ vektörü, Ω parametre uzayında çok değişkenli normal dağılımların karmasının bilinmeyen parametrelerinin tümünü temsil eden vektördür [8].

Bu tez çalışmasında modele dayalı kümeleme analizi için çok değişkenli karma dağılım modelindeki bileşenlerin çok değişkenli normal dağılımlardan geldiği varsayılarak verideki bileşen sayısının belirlenmesi için mevcut yöntemler incelenmiş ve yeni bir yöntem ileri sürülmüştür. Çok değişkenli normal karma dağılımlarından gelen verideki bileşen sayısı belirlendikten sonra modele dayalı kümeleme analizi yapmak için model oluşturulmuştur. Oluşturulan modelde varsayım altında uygun aday modellerin sayısı belirlenmiştir. Değişkenlerdeki parametreler, kitlenin ortalama vektörü, varyans-kovaryans matrisi ve kitledeki gözlem sayılarına bağlı olarak olasılık ağırlıkları olarak alınmıştır. Karma kümeleme analizinde modele dayalı kümeleme için uygun aday modeller arasından en iyi modelin seçiminde istatistiksel bilgi kriterleri kullanılmıştır.

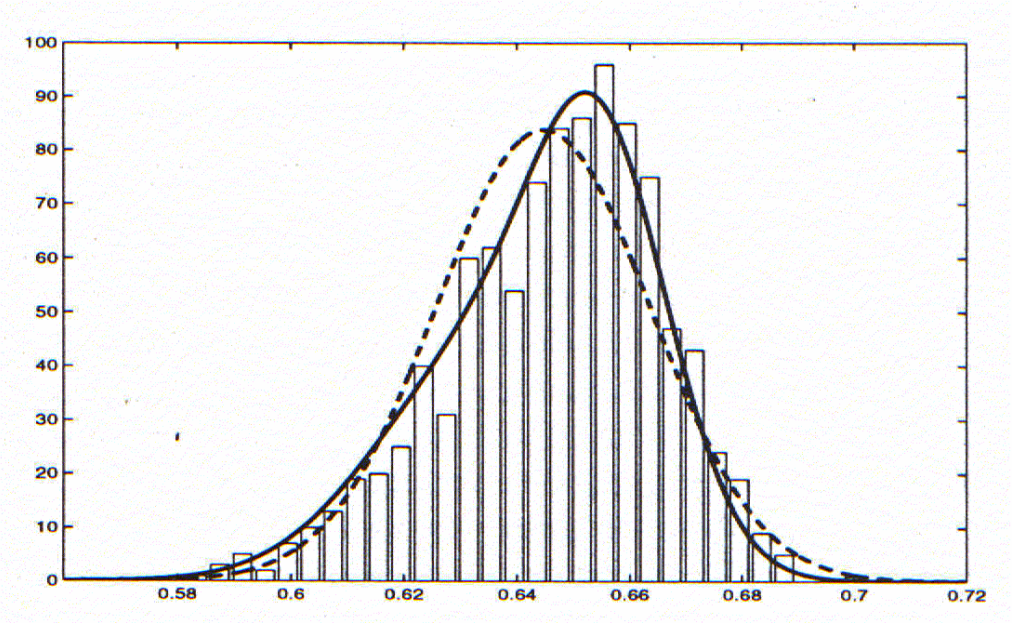
Bu tez çalışması dört bölümden oluşmaktadır. Birinci bölümde çok değişkenli kümeleme analizi ile modele dayalı kümeleme analizi ve en iyi model seçimi ile ilgili yapılan çalışmalar incelenmiştir. Çok değişkenli kümeleme analizinde tanımlar, kavramlar ve gösterimler araştırılmıştır. İkinci bölümde kümeleme analizinde kullanılan yöntemler gösterilmiştir. Çok değişkenli karma normal dağılım modeline dayalı kümeleme analizi, sonlu karma dağılım modelleri ve modele dayalı kümeleme analizi için yöntemler incelenmiştir. Üçüncü bölümde modele dayalı karma normal kümeleme için geliştirilen genetik algoritma ve adımları anlatılmıştır. Geliştirilen genetik algoritma farklı veri setlerine uygulanarak en iyi kümelenmenin seçimi gösterilmiştir. İki değişkenli ve normal karma modellere dayalı kümeleme analizi için iki, üç ve dört bilşenli modeller için sentetik veri seti üzerinde uygulama anlatılmıştır. İki değişkenli her değişkenin üçe bölündüğü veri seti üzerindeki uygulama anlatılmıştır. Geliştirilen genetik algoritma iki değişkenli ve her değişkenin dörde bölündüğü gerçek veri seti üzerine uygulanarak en iyi kümelenme yapısı ortaya çıkarılmıştır. Üç boyutlu ve değişkenlerin bir, iki ve üçe bölündüğü veri seti üzerinde değişken seçimi ve en iyi modelin seçimi için uygulama anlatılmıştır. Değişkenlerdeki heterojenliğin kümelenmeyi belirlediğini göstermek için, değişkenli ve değişkenlerin bir, iki ve iki parçalanmaya sahip olduğu veri setindeki bulgularla iki değişkenli ve değişkenlerin her birisinin ikiye parçalandığı veri setindeki bulguların aynı olduğunu gösteren uygulama gösterilmiştir. Veri madenciliği metodu olarak belirlenen genetik algoritmanın on beş değişkenli veri seti üzerindeki uygulaması anlatılmıştır. Üçüncü bölümde son olarak üç boyutlu uzaktan algılanmış uydu görüntü verisi üzerine yarı eğitimli sınıflandırma metodu ve en iyi kümelenme modelinin belirlendiği gerçek veri seti uygulaması anatılmıştır. Dördüncü bölümde sonuç ve öneriler verilmiştir.

1. BÖLÜM

LİTERATÜR ÇALIŞMASI ve TEMEL KAVRAMLAR

1.1. Literatür Çalışması

Karma dağılım modellerinin parametre tahmini için ilk çalışma 1894 yılında Karl Pearson tarafından yapılan çalışma olarak kabul edilmektedir. Pearson birbirinden farklı μ_1 ve μ_2 ortalamalı, σ_1 ve σ_2 varyanslı ve π ve $(1-\pi)$ karma oranlarına sahip tek değişkenli iki bileşenli normal karma dağılım modelini incelemiştir. Weldon [9],[8] tarafından ölçülen veride 1000 yengeçten alınan veriler modellemiş ve histogramı simetrik olmadığından dolayı yengeç kitlesinin iki farklı türden meydana geldiği iddia edilmiştir. Pearson bu çalışmada tek boyutlu iki normal karma dağılımın parametrelerini momentler yöntemi kullanarak hesaplamıştır.



Şekil 1.1. Yengeç verisi için oluşturulan tek bileşenli normal dağılım modeli (kesikli çizgi) ve iki bileşenli normal karma dağılım modeli (düz çizgi) [11].

Yengeç verisi önce bir tek normal dağılımla modellenmiş ve daha sonra simetrik olmayan yapı nedeniyle Şekil 2.1'de görüldüğü gibi normal dağılımların karması kullanılmıştır. Bu modellemede; modeldeki parametrelerin tahminlerini elde etmek için yüksek dereceden polinomların köklerinin hesaplanması gerekmektedir ve parametrelerin tahmini için momentler yöntemini kullanmıştır. Pearson'un karma dağılım modelleriyle ilgili bu çalışması bir biyoloji probleminin çözümüne yardımcı olduğundan genetik çalışmalarda kullanılan ilk çalışma olarak da kabul edilir

Karma dağılım modellerindeki bileşen sayısının belirlenmesi problemi zor ve tam olarak çözülememiş bir problemdir [11]. Karma dağılımların bileşen sayılarının belirlenmesinde momentler yöntemi, grafiksel yöntemler ve likelihood kestirimleri metotları kullanılmıştır. Karma dağılımlardaki parametre tahminlerinde likelihood kestrimleri metodu ilk defa [12] tarafından kullanılmıştır.

Rao tarafından önerilen yöntemde, iki normal karma dağılımın bileşenlerinin varyanslarını eşit alarak parametreleri ilk defa en çok olabilirlik (Maximum Likelihood Estimation-MLE) yöntemiyle elde etmiştir. Çok değişkenli normal karma dağılımlar kümeleme analizinde istatistiksel bir yöntem olarak [13] ve [14] tarafından kullanılmıştır. Kümeleme analizinde hangi kriterin daha etkili olduğunun kesin olarak belirlenemeyeceğini "çok değişkenli normal karmalarda kümeleme kriterleri" [15] adlı çalışmasında belirtmiştir. Symons bu çalışmasında çok değişkenli normal karma dağılımlarıyla oluşabilecek yeni kümeleme kriterleri üzerinde durmuştur. Bu kriterler çok değişkenli normal karma dağılımların bileşen kovaryans matrislerinin en çok olabilirlik ve Bayes yaklaşımlarından elde edilmesi, grup içi kareler toplamı matrisinin determinantı yöntemleridir. Bu yöntemlerin karşılaştırılması [16] tarafından üç değişkenli veri üzerindeki çalışma ile gösterilmiştir.

Karma yoğunluklar, en çok olabilirlik ve (Expectation and Maximization-EM) algoritması [17] başlıklı çalışmada üstel dağılımlardan oluşan karma modellerin parametrelerinin EM algoritması kullanılarak en çok olabilirlik kestirimlerinin (MLE) nasıl elde edildiği incelenmiştir. EM algoritmasının üstel dağılımların karma modellerindeki özellikleri incelenmiştir. Karma modellerin parametre tahminleri hesaplanırken EM algoritması kullanılmasının bazı olumsuz sonuçları anlatılmıştır. EM algoritmasındaki ardışık hesaplamalar için seçilen başlangıç değerlerinin önemli olduğu

ortaya çıkarılmıştır. EM algoritmasının ardışık hesap adımlarındaki yakınsama, çok yavaş olabileceği ve aynı zamanda algoritmanın parametrelerinin mutlak maksimum değeri yerine bir yerel maksimum değerine yakınsayabileceği saptanmıştır. EM algoritmasındaki bu olumsuzluklar başlangıç parametre seçimi için ayrı bir algoritma (yöntem) kullanılması durumunda algoritma hızı ve yakınsamadaki yerel maksimum noktalarının seçimi gibi problemlerin ortadan kaldırılabileceği karma algoritmaların olabileceği ileri sürülmüştür.

Kümeleme için "sınıflandırma EM algoritması ve iki zamana bağlı (stokastik) versiyon" [18] başlıklı çalışmada EM algoritmasındaki olumsuzluğa karşı sınıflandırma en çok olabilirlik yöntemi altında optimizasyona dayalı kümeleme yöntemleri için genel sınıflandırma EM algoritması tanımlamışlardır. Geliştirilen bu algoritmadan optimizasyona dayalı kümeleme algoritmalarının parametre başlangıç değerlerine olan bağımlılıkların minimize edildiği ve rasgele karmaşıklık algoritmasını da içeren iki stokastik algoritma elde edilmiştir. Bu iki stokastik algoritmayı karşılaştırmak için k-ortalamalar algoritması (k-means) ve varyans kriterini kullanmıştır. Geliştirilen bu genel sınıflandırma EM algoritması, CEM (Classification EM), stokastik EM algoritması SEM (Stokastik EM) ve sınıflandırma tavlayıcısı CAEM (Classification Annealing EM) olarak adlandırılmıştır. Örneklem hacmi küçük olan verilerde CAEM ve örneklem hacmi büyük olan verilerce SEM algoritmasının daha iyi sonuç verdiği belirlenmiştir. SEM ve CAEM algoritmasının her ikisinde de başlangıç parametre değerlerine bağımlılığı minimize etmek için kullanılan algoritmadan dolayı hesaplama adımları oldukça fazladır.

"Modele dayalı normal ve normal olmayan dağılımlarda kümeleme" [19] çalışmasında çok değişkenli normal karma dağılımlarındaki bileşenlerden oluşan varyans-kovaryans matrisinin özdeğer ayrışımı kullanılarak parametrik olarak ifade edilmiştir. Bu parametrik ifade kullanılarak verilerin ortalama vektörü, standart sapma ve olasılık ağırlıkları gibi parametrelerine bağlı kümeler üzerindeki yön, hacim ve şekil gibi bazı geometrik özelliklerinde ele alınması sağlanmıştır. Aynı çalışmada önceden tanımlanmış küme ve küme merkezlerine uygun olmayan gürültü verilerin varlığından bahsedilmiştir. Bu gürültü verilerinin küme merkez sayısını artırmaması için etiket vektörü değeri sıfır alınmıştır.

İstatistikte çalışılan en zor problemlerden birisi karma dağılım modelleri kullanılarak kümeleme analizidir [20]. “Karma model kümeleme analizi” olarak adlandırılan çalışmada bileşenlerden meydana gelen küme sayısının seçimi ele alınmıştır. Küme sayısı genel olarak bileşenlerin parametrelerinden oluşan varyans-kovaryans matris yapıları göz önüne alınarak belirlenmeye çalışılmıştır. Karma dağılım modelindeki bileşen kümelerinin aynı varyans-kovaryans matris yapısına sahip olması, bileşen kümelerinin farklı matris yapısına sahip olması, bileşen kümelerinin matris yapısı aynı fakat köşegen varyans-kovaryans yapısına sahip olması ve bileşen kümelerinin aynı ancak değişkenlerin ikişer ikişer bağımsız olacak şekilde köşegen varyans-kovaryans matris yapısına sahip olması durumları ayrı ayrı incelenerek bileşenlerin kümelenme sayısı belirlenmeye çalışılmıştır.

[21] karma ayrıştırma analizinde karma yoğunluk tahminini sınıflandırmada kullanmışlardır. Bu çalışmada ayrıştırma yapmadan önce k-ortalamalar (k-means) algoritması veya benzer algoritmalar yardımı ile veri için başlangıç sınıfları oluşturulmuştur. Daha sonra EM algoritmasının M adımında her bir bileşen için parametrelerin ağırlıklandırılmış maksimum likelihood tahminleri kullanılarak veriye uygulanmıştır. Bu şekilde veri için parametre tahminleri yapıldıktan sonra ayrıştırma fonksiyonu oluşturulmuş ve gözlemleri aitlik olasılıklarının maksimum olduğu kümelere sınıflandırılarak ayrıştırma analizi yapılmıştır.

Erol ve Akdeniz [22] tarımsal bir bölgedeki farklı alanları sınıflandırmak için tek değişkenli normal dağılımların karmalarına dayalı çok bantlı (multispectral) sınıflandırma algoritması önermişlerdir. Önerilen bu algoritma tek değişkenli durumda farklı iki karma normal dağılım modeline ait olasılık yoğunluk fonksiyonlarının karşılaştırılmasına dayalı bir algoritmadır. Erol ve Akdeniz [22] uzaktan algılanmış uydu görüntü verisin 3, 4 ve 5. bantlarını kullanarak tarımsal bölgedeki her bir kontrol (eğitim) ve test alanı için tek değişkenli ve üç bileşenden oluşan karma normal dağılım modeli oluşturmuşlardır. Erol ve Akdeniz [22] önerdikleri çok bantlı sınıflandırma algoritması için ayrıştırma fonksiyonu olarak Hellinger uzaklığını ve karar kuralı olarak ta Hellinger uzaklığının minimum değerini kullanmışlardır.

Karma dağılım modellerinde incelenen homojen olmayan çok değişkenli veri gruplandırılmamış veridir. Ayrıştırma analizi ve kümeleme analizine dayalı

sınıflandırmada ayrıştırma fonksiyonunun belirlenmesi için öncelikle verinin gruplara doğru bir şekilde ayrılması gerekmektedir [23]. Diğer bir ifade ile veriye en uygun karma dağılım modeli belirlenmelidir. Veri için karma dağılım modelleri oluşturulurken EM algoritması [11] kullanılabilir. Oluşturulan karma dağılım modellerinden en uygun modeli belirlemede AIC ve BIC gibi bilgi kriterleri [11] kullanılır. Veri gruplara ayrıldıktan sonra oluşturulan model için ayrıştırma fonksiyonu ve karar kuralı belirlenir. Karar kuralının belirlenmesi ile hangi sınıfa ait olduğu bilinmeyen yeni verilerin sınıflandırılması yapılacak en son işlemdir.

Bir veri kümesinin içerdiği küme sayısı ve bileşenlerin oluşturduğu kümelenme yapısı hakkında herhangi bir bilgiye sahip olmadan kümelenme yapısı ve sayısının nasıl belirleneceği [8] tarafından önerilmiştir. Heterojen verilerin karma dağılımlar ile temsil edilebileceğini aynı zamanda karma dağılımın her bileşeninin farklı bir kümeye karşılık geldiğini belirtmişlerdir. Karma dağılımın yoğunluk fonksiyonunun çok değişkenli normal karma dağılımlardan gelmesi durumunda bileşenlerin varyans-kovaryans matrislerinin parametrik özelliklerine göre farklı geometrik yapıya sahip modellerin elde edilebileceğini belirtmişlerdir.

Karma normal dağılımda ayrıştırma analizi Hastie ve Tibshirani [21] tarafından ele alınmıştır. Hastie ve Tibshirani, özellikle verideki kümelenmelerin belirgin olduğu durumda normal olmayan veri kümesini etkili bir şekilde sınıflandırmak için karma normal dağılım yaklaşımını incelemişlerdir. Lineer ayrıştırma analizinin uygulanması için veri kümesinin düşük boyutlu olması önemlidir. Alt küme merkezlerinin veri kümeleri içerisindeki yayılımını kontrol etmek kümeler arası yayılımı kontrol etmeye göre daha kolaydır. Hastie ve Tibshirani tarafından önerilen karma ayrıştırma analizi (Mixture Discriminat Analysis- MDA) karma yoğunluk tahminini sınıflandırmaya genelleştirir. Yani ayrıştırma yapmadan önce k-ortalamalar (k-means) algoritması veya benzer algoritmalar yardımı ile veri için başlangıç sınıfları oluşturulur. Daha sonra EM algoritması, M-adımında her bir bileşen için parametrelerin ağırlıklandırılmış maksimum likelihood tahminleri kullanılarak veriye uygulanır. Bu şekilde ayrıştırma analizi yapılmış olur. Karma ayrıştırma analizi lineer ayrıştırma analizinin genelleştirilmiş halidir.

Modele dayalı kümeleme ve ayrıştırma analizinde model seçimi [24] tarafından araştırılmıştır. [25] varyans-kovaryans matrisinin özdeğer ayrışımını kullanarak normal

dağılım modeline dayalı kümeleme ve ayrıştırma analizi için uygun modeller oluşturmuşlardır. Bu çalışmada [24], [25]'in oluşturduğu bu modelleri seçmek için Monte Carlo simülasyon verisini kullanarak bir çok sınıflandırma kriterinin performanslarını karşılaştırmışlardır. Bu kriterler arasında bilgi kriteri olarak AIC ve BIC, ve sınıflandırma kriteri içinde NEC ve çapraz geçerlilik (cross-validation) kriterleri örnek verilebilir. Çapraz geçerlilik kriteri ayrıştırma analizinde iyi sonuçlar vermektedir. Bilgi kriterleri ve BIC hesaplamalarda zaman kazandırdığı için yeterli sonuçlar vermektedir.

Kümeleme analizi veriyi anlamlı gruplara parçalamayla ilgilidir. Kümeleme analizinde çalışılan en önemli problem, verideki kümelenmelerin sayısını ve yapısını bulmaktır. Verideki kümelenmelerin sayısını ve yapısını bulmanın bir yöntemi modele-dayalı kümeleme analizidir. Modele-dayalı kümeleme analizinde veri, çok değişkenli karma dağılım modelleriyle temsil edilir. Bu temsilde her bir bileşen bir kümelenmeye karşılık gelir [26]. Kümeleme analizinde amaç her bir grup içinde homojenliğin maksimum yapılması aynı zamanda gruplar arasındaki farkın da maksimum yapılmasıdır [27].

Amerikan deniz kuvvetlerinde denizde mayınların yerlerinin belirlenmesi amacıyla geliştirilen projede mayın tarama verisi olarak adlandırılan görüntü verisi [28] tarafından oluşturulan COBRA (Coastal Battlefield Reconnaissance and Analysis) programında kullanılmıştır. Fraley ve Raftery [6] aynı veriye COBRA programıyla elde edilen sonuçlarla karşılaştırmak amacıyla modele dayalı kümeleme analizini uyguladılar. Analiz sonucunda BIC (Bayesian Information Creteria) değerlerine göre dört bileşenli karma model elde etmişlerdir. Karşılaştırma sonucunda COBRA programının sonucundan iki kata kadar daha iyi sonuç elde etmişlerdir.

Ju vd. [29], karma normal dağılıma dayalı ayrıştırma analizi ve uzaktan algılamada bir uygulama gerçekleştirmişlerdir. Karma dağılım modelleri için yapılan analizler, uzaktan algılama ile elde edilen görüntülerin incelenmesinde de kullanılmaktadır. Bu analizler, algılanan görüntünün yapısında bulunan farklılıkları belirlemede önemli rol oynar. Uzaktan algılamada karma normal dağılım modellerinin lineer karma dağılım modellerinden lineer olmayan yapay sinir ağı karma dağılım modellerine kadar birçok uygulama alanı vardır. Ju vd. [29], bu çalışmalarında karma ayrıştırma analizi ile yapay sinir ağlarına dayalı ayrıştırma analizini yöntemlerini karşılaştırmış ve uzaktan algılamada karma ayrıştırma analizinin oldukça etkili sonuçlar verdiğini göstermişlerdir.

Model tabanlı sınıflandırma, yoğunluk tahmini ve ayrıştırma analizi yazılımı MCLUST, Fraley ve Raftery [30] tarafından oluşturulmuştur. MCLUST model tabanlı sınıflandırma, yoğunluk tahmini ve ayrıştırma analizi yapmak için oluşturulmuş bir paket programdır. Bu program normal dağılıma dayalı hiyerarşik sınıflandırma algoritmalarına ve EM algoritmasına bağlı olarak çalışmaktadır. Aynı zamanda bu yazılım hiyerarşik sınıflandırma algoritmasını, EM algoritmasını ve Bayesçi bilgi kriterini (BIC) birleştiren fonksiyonları içermektedir. Bu fonksiyonlar sınıflandırma, yoğunluk tahmini ve ayrıştırma analizi için kullanılabilecek çok yönlü fonksiyonlardır. MCLUST programı ile sınıflandırma sonuçlarını iki ve üç boyutlu grafikler üzerinde görmek mümkündür.

Karma normal dağılıma dayalı ayrıştırma analizi ve uzaktan algılamada bir uygulama, Kolaczyk ve ark. [31] tarafından gerçekleştirilmiştir. Karma dağılım modelleri için yapılan analizler, uzaktan algılama ile elde edilen görüntülerin incelenmesinde de kullanılmaktadır. Bu analizler, algılanan görüntünün yapısında bulunan farklılıkları belirlemede önemli rol oynar. Uzaktan algılamada karma normal dağılım modellerinin lineer karma dağılım modellerinden lineer olmayan yapay sinir ağı karma dağılım modellerine kadar birçok uygulama alanı vardır.

Soffritti [32] "bir veri matrisinde çoklu kümelenme yapılarının tanımlanması" çalışmasında kümeleme analizinde değişkenlerin seçimi ve değişken ağırlıkları ile ilgili bir yöntem geliştirmiştir. Çalışmasında kümelenme yapılarının tanımlanabilmesi değişkenlerin alt gruplara bölünmesi ve değişkenlerin farklı ağırlıklandırılmasıyla olabileceğini göstermiştir.

Li [33], çok tabakalı karma modele dayalı sınıflandırma geliştirmiştir. Model tabanlı sınıflandırmada her bir kümenin (sınıfın) yoğunluğu genellikle normal dağılım gibi belirli temel parametrik dağılıma sahip olduğu varsayılır. Özellikle çok değişkenli veri için bir kümenin dağılımı belirlenirken hangi parametrik dağılımın uygun olacağına karar vermek pratikte zordur. Bunun yanında, her bir sınıf kendi içerisinde birden fazla moda sahip olabilir. Dolayısı ile temel parametrik bir dağılımla doğru olarak modellenmeyebilir. Sonuçta ise genel model çok tabakalı karma normal dağılıma sahip olabilir. Modeli tahmin etmek ve sınıflandırma yapabilmek için sınıflandırma maksimum likelihood kriterine (CML) ve karma maksimum likelihood kriterine (MML) dayalı algoritmalar geliştirilmiştir. Bunun yanında her bir kümedeki normal dağılıma sahip bileşenlerin

sayısını belirlemek için BIC ve (integrated completed likelihood-ICL) ICL-BIC gibi bilgi kriterleri incelenmiştir.

Halbe ve Aledjem [34] modele dayalı karma ayrıştırma analizini kullanarak deneysel bir çalışma yapmışlardır. Halbe ve Aladjem [34] bu çalışmalarında modele dayalı ayrıştırma analizini gerçek ve simülasyonla elde edilmiş çeşitli veri kümeleri üzerinde incelemişlerdir. Az sayıda gözlem içeren veri kümelerindeki bileşenler için uygun varyans-kovaryans yapıları belirleyerek doğru ayrıştırma ve/veya sınıflandırma oranını artırmışlardır. Ayrıca Halbe ve Aladjem bu çalışmalarında model seçimi için Bayesci bilgi kriterini kullanan yeni bir yöntem önermişlerdir

Bashir ve Carter [35], rank indirgeme yöntemi ile karma ayrıştırma analizini incelemişlerdir. Bu makalede çok sayıda özellik vektörünün olması durumunda çok değişkenli normal karma dağılım modelleri kullanılarak yapılan ayrıştırma analizinde indirgenmiş alt uzaylar çalışılmıştır. Bu çalışma Hastie ve Tibshirani [21]' nin çalışmalarının genelleştirilmiş halidir. Normal karma dağılım modellerinin olması durumlarında rank indirgeme yöntemi ile yapılan ayrıştırma analizi ağırlıklandırılmış k ranklı lineer ayrıştırma analizine eşittir.

Çok değişkenli normal karma modellerdeki indirgenmiş rank çözümleri tam ranklı dayanıklı (robust) karma dağılım modellerinin çözümlerinden elde edilir. Yeni yönlerdeki sınıflandırmalar robust S-tahmin edicileri kullanarak orijinal koordinatlara dayalı ayrıştırma analizi yaklaşımı ile karşılaştırılmaktadır. Birçok durumda rank indirgemeye dayalı ayrıştırma analizi test verileri için daha iyi sonuç vermektedir. Bununla birlikte kovaryans matrisinin köşegen ve ortak olması durumu için rank indirgemeye dayalı karma ayrıştırma analizi, tam ranklı karma ayrıştırma analizinden sınıflandırmada daha az hata ile daha iyi sonuç vermektedir.

Değişkenlerdeki parçalanma sayısının ön bilgi olarak verilmediği (eğitimsiz) normal karma dağılımlarda kümelenme için bir algoritma Bouman [36] tarafından, örneklem verisinden normal karna dağılımların parametrelerini otomatik olarak tahmin eden bir kümeleme paket programı anlatılmıştır.

Raftery ve Dean [37] "modele dayalı kümelemede değişken seçimi" çalışmalarında modele dayalı kümeleme analizi için değişken seçimi veya değişkenlerin

belirlenmesindeki özelliklerin seçimi için bir yöntem öne sürmüşlerdir. Değişken seçimindeki küme sayısı ve en iyi kümelenme modelinin belirlenmesi için karma dağılım modellerinin bileşen kovaryans matris yapısı ile ilgili bir algoritma geliştirmişlerdir. Algoritma ardışık adımlardan oluşmaktadır. Birinci değişken tek kümelenme yapısının en yüksek olasılıkla elde edildiği değişken olarak belirlenir. Sonraki adımda ilk değişkenle birlikte iki kümelenme yapısının en yüksek olasılıkla elde edildiği değişken seçilir. Aynı işlem önceki değişkenlerle birlikte devam ettirilir ve en yüksek olasılıkla çok değişkenli kümelenme yapısını veren değişken belirlenir. Aradaki adımlarda değişken seçimi için denenen adaylarda algoritmanın ilerlemesine katkı sağlamayan değişkenler ayıklanır. Algoritma daha çok küme sayısını veren değişken bulunamayıncaya kadar devam eder ve bulunamayınca sonlandırılır. Kümelemedeki değişkenlerin seçiminde en yüksek BIC değerini veren değişkenler alınır. Algoritmanın performansının ölçülmesi için iki gerçek veri kümesi üzerinde uygulama yapılmıştır. Elde edilen sonuçlar algoritmanın modele dayalı kümeleme analizinde daha iyi sonuçlar verdiğidir.

Karma modelleme MIXMOD programı ile modele dayalı kümeleme ve ayrıştırma analizi Biernacki ve ark. [38] tarafından oluşturulmuştur. MIXMOD programı yoğunluk tahmini, kümeleme ve ayrıştırma analizi yapmak üzere verilen bir veri kümesi için karma model oluşturmada kullanılmaktadır. Karma modelde parametreleri tahmin etmek için EM, Sınıflandırma EM ve Stokastik EM gibi çeşitli algoritmalar önerilmiştir. Burada amaç likelihood fonksiyonunu maksimum yapacak parametre tahminlerini bulmaktır. MİXMOD programı çok değişkenli normal karmaları ve kırk farklı normal modeli için kullanılmaktadır. MİXMOD'da bileşenlerin varyans-kovaryans matrislerinin özdeğer ayrışımlarına göre ayrıştırma analizi ve modelleme yapılmaktadır.

"Çok değişkenli kalibrasyon yöntemlerinde Mclust paket programı kullanılarak modele dayalı sınıflandırma yöntemleri" başlıklı çalışmalarında Fraley ve Raftery [39], modele dayalı kümeleme analizindeki ilerlemeler sayesinde deney tabanlı yöntemler yerine olasılık tabanlı yöntemlerin daha çok kullanıldığını öne sürmüşlerdir. Çok değişkenli kalibrasyon yöntemlerindeki ölçümlerden elde edilen görüntü verilerinin oluşturduğu kümeler üzerinde kullanılan programın modele dayalı kümeleme analizi, ayrıştırma analizi ve yoğunluk tahmininde nasıl kullanılacağını göstermişlerdir.

“Modele dayalı kümeleme yöntemleri kullanarak verideki birden fazla küme yapılarının belirlenmesi” başlıklı çalışmalarında Galimberti ve Soffritti [40] farklı değişkenlerde parçalanışlardan oluşan alt kümeler üzerinde durulmuştur. Bu yöntem modele dayalı kümeleme yapılarak normal karmalar arasında karşılaştırma yapmak için BIC (Bayesian Information Criteria) kullanılmıştır. Modele dayalı kümeleme yöntemlerinin birleşmesinden oluşan bu genel yöntem için gerçek veri kümesi üzerinde başarılı uygulamalar yapmışlardır.

Servi ve Erol [41] modele dayalı kümelemede aday karma modellerin ve aday bileşen kümelenme merkezlerinin toplam sayısı başlıklı çalışmalarında çok değişkenli normal karmaların toplam kümelenme merkez sayılarının hesaplanabilmesi için bir aralık önermişlerdir. Bu aralık değişkenlerdeki bileşenlerden kaynaklanan olabilecek maksimum ve minimum kümelenme merkez sayısının belirlenmesini sağlamaktadır. Oluşabilecek merkez sayılarına bağlı olarak oluşabilecek toplam karma model sayısının hesaplanması için bir hesaplama yöntemi önermişlerdir.

“Normal karma dağılım modelleri ile konuşmacı tanımada parametre değerlerinin seçimi” çalışmalarında Eskidere ve Ertaş [42], konuşmanın (ses) temiz (TIMIT veri tabanı) ve telefon ortamında (NTIMIT veri tabanı) iletildiği iki veri tabanı için, Normal karmaların bilşen sayısı, eğitim süresi, test süresi ve kişi sayısı değişimlerinin konuşmacı tanımaya etkisini incelemişlerdir. Parametre seçiminin tanıma üzerendeki önemli sonuçları ortaya konmuştur.

Jain [43] “k-ortalamalar algoritmasından 50 yıl sonra veri kümeleme” başlıklı derleme çalışmasında kümeleme analizi ve kümeleme analizinde kullanılan algoritmaların kullanım zorlukları, avantajları ve çıktıları üzerine geniş bir özet sunmuştur. Çalışmada kümeleme analizinde kullanılan algoritmaların yarı eğitimli kümeleme, kümelenme grupları, kümeleme yapılırken eş zamanlı olarak özellik çıkarımı ve büyük verilerde kümelemede karşılaşılan zorlukları ve sonuçları aktarılmıştır.

Li ve Qiao ve [44] makalelerinde, karma ayrıştırma analizi paradigması altında, yüksek boyutlu veri sınıflandırmak için iki yönlü bir Gauss karma modeli önermişlerdir. Bu model gruplar halinde değişkenlerin bölünmesiyle karma bileşenleri düzenler ve daha sonra aynı olan benzer grupların değişkenleri için parametreleri kısıtlar. Burada değişken grupları önceden belirlenmemiş fakat model tahmininin parçası olarak optimize

edilmiştir. Makalede genel bir üstel aileden gelen dağılımların iki yönlü karması için bir boyut indirgeme özelliğini ispatlamışlardır. Sonuç olarak, bazı gerçek veri setlerinde iki yönlü karma grup değişkenleri olmayan karma modelden daha iyi performans gösterdiğini belirtmişlerdir. Ayrıca bir yan ürün olarak, önemli bir boyut indirgeme elde etmişlerdir.

"Modele dayalı kümelemede boyut indirgeme" [2] çalışmasında, normal yoğunlukların sonlu karmalarından elde edilen görsel küme yapıları için boyut indirgeme metodunu tanıttılar. Boyut indirmedeki alt uzaylar hakkındaki bilgiler, karma model tahminine dayalı farklı grup ortalama vektörü ve kovaryans matrislerinden elde edilir. İleri sürülen yöntemin asıl amacı, veridek,i küme yapılarının orijinal özellikleri, kovaryans matrisinin birleştirilmiş öz değerlerinin özellikleri ve bunların lineer kombinasyonlarının kümeleri tarafından tanımlanan boyut indirgemesidir. Gözlem değerlerinin indirgenen alt uzaya iz düşümü alınarak küme yapısının görsel olarak özeti hakkında bilgi sahibi olunabileceğini göstermişlerdir.

McNicholas vd. [45] "çok değişkenli t-faktör analizinin karmaları kullanılarak modele dayalı sınıflandırma" başlıklı çalışmalarında karma dağılımdaki parametre tahmini için AECM (Alternating Expectation Converge and Maximization) algoritması kullanılmış, bu algoritmanın yakınsaklığını belirlemek için Aitken's ivmelenmesi (Aitken's acceleration) kullanılmıştır. Model seçimi için BIC ve ICL (integrated completed likelihood-ICL) teknikleri önerilmiştir.

Seo ve Kim [46] "normal karma dağılımlarda kök seçimi" çalışmalarında bir sonlu karma model için olabilirlik (likelihood) denkleminin çoklu köklerinin varlığında istatistiksel olarak anlamlı köklerin seçimine dayalı yeni bir olabilirlik yaklaşımı önerdiler. Önerilen yaklaşımın metodolojisindeki olabilirlik denkleminin köklerinin tutarlı olarak seçilebileceğini ispatlamışlardır. Yöntemi hem simülasyon hem de gerçek veri üzerinde uygulamış iyi sonuçlar elde etmişlerdir.

Çalış ve Erol [47] "karma ayrıştırma analizleri kullanılarak yeni bir alan bazlı sınıflandırma metodu" çalışmalarında normal karma dağılımların ayrıştırma analizini kullanarak tarım arazilerinin uzaktan algılanmış çok bantlı görüntü verisinin değişkenlerindeki parçalanmanın bilindiği (eğitimli) sınıflandırma için yeni bir alan bazlı sınıflandırma yöntemi geliştirmişlerdir. Bu alan bazlı sınıflandırma yönteminde çok

değişkenli normal karma dağılımların yapısı için test ve kontrol alanlarının sabit veya farklı sayıda bileşen içerebileceği ve her bileşenin farklı veya genel kovaryans matris yapısında olabileceği gösterilmiştir.

Servi ve Erol [48] "dinamik modele dayalı kümeleme kullanarak verideki grupların inceltilmesi için veri madenciliği metodu" başlıklı çalışmalarında dinamik modele dayalı kümeleme kullanarak çok değişkenli heterojen verideki küme yapısına tam veya kısmen uyan, inceltilen grupları kapsayan, kümelenmelerin yapısı ve sayısının belirlenmesini anlatmışlardır.

Erol [49] tarafından "çok değişkenli heterojen veriye uyan normal karma modeller arasından en iyi modelin bulunması" için bir model seçimi algoritması geliştirilmiştir. Algoritmada önce normal karma modellerin toplan sayısı belirlenip ardından normal karmalar arasından uygun aday modeller belirlenir, son olarak uygun aday modeller arasından en iyi modelin seçilmesi için istatistiksel bilgi kriterleri kullanılır. En iyi model seçiminde AIC ve BIC değerleri kullanılmaktadır.

Huang vd. [50] "normal karma dağılımlar için model seçimi" başlıklı çalışmalarında sonlu karma dağılım modellerinde karma bileşenlerin sayısının seçimi üzerinde durmuşlardır. Sonlu çok değişkenli normal karma modellerin model seçimi için yeni bir cezalı olabilirlik (penalized likelihood) metodu önermişlerdir. Önerilen metotta bileşenlerin sayısının belirlenmesinin istatistiksel olarak tutarlı olduğu gösterilmiştir. Uyarlanan EM algoritması ile bileşen sayısı ve karma oranları eşanlı belirlenmekte, ayrıca normal dağılımların bilinmeyen parametreleri ile karma olasılıkları belirlenebilmektedir.

"Yüksek boyutlu verinin modele dayalı kümelenmesi" çalışmalarında Bouveyron ve Brunet-Saumard [51], modele dayalı kümeleme, boyut indirgeme yaklaşımları, düzenlileştirmeye dayalı teknikler, cimri modeller, alt uzay kümeleme metotları ve kümeleme metotlarına dayalı değişken seçimi yöntemlerini incelenmiştir. Yüksek boyutlu verilere dayalı kümeleme metotları R paket programı kullanımı gözden geçirilmiş her bir yöntemin pratik kullanımı gerçek veri kümeleri üzerinde anlatılmıştır.

"Karma modellerdeki parametrelerin bileşenlerine ayrılmış kovaryans matrisi ve özdeğer ile tahmini için Ortogonal Stiefel manifold optimzasyonu" başlıklı çalışmalarında

Browne ve McNicholas [52], değişkenlerine ayrılmış kovaryans matrisi ve özdeğerlere dayalı on dört model arasından dört modeldeki parametre tahminin zorluğu üzerinde durulmuştur. Bu dört durumun yakın incelenmesi sonucunda MCLUST paket programının algoritması kullanılarak iki kolay uygulama için yeni yaklaşımlar ortaya çıkarılmıştır. Bu yeni yaklaşımlar iki veri seti üzerinde uygulanmış ve sonuçlarının oldukça iyi performansa sahip olduğu gösterilmiştir.

Cheballah vd. [53], "paket kare matrisler üzerindeki kombinatoryal hopf cebir yapıları" çalışmalarında bir kare matrisin her bir satırında ve her bir sütununda sıfırdan farklı en az bir elemanın bulunduğu matris yapılarının sayısını incelemişlerdir. nxn tipindeki bir matriste her satır ve her sütunun en az sıfırdan farklı bir eleman bulunması durumunda oluşabilecek farklı matris sayısı için kombinasyonla elde edilmiş sayı dizisi için bir toplam sembolü altında denklem elde etmişlerdir.

Ayrıca "Karma normal dağılımlar kullanılarak veri seti üzerinde kümeleme" çalışmaları ile ilgili Çalış [54] Yüksek lisans tezi ve Servi [55] Doktora tezi önceki çalışmalar kısmının incelenmesinde detaylı çalışmaları ile önemli katkı sağlamışlardır.

1.2. Temel Kavramlar

1.2.1. Matematiksel Tanımlar

Tanım 1.2.1. Sezgisel olarak belli özelliklere sahip nesnelerin bir topluluğu, bir sınıfı veya bir koleksiyonuna küme adı verilir. Kümeyi meydana getiren nesnelere kümenin elemanları adı verilir. Eğer bir a elemanı A kümesine ait ise $a \in A$ denir. Eğer aynı eleman kümeye ait değil ise $a \notin A$ denir.

Tanım 1.2.2. $X, Y \neq \varnothing$ kümeler olmak üzere $XxY = \{(x, y) | x \in X, \mathrm{y} \in Y\}$ kümesine $X \text{ ve } Y$ kümelerinin Kartezyen çarpımı, $X \text{ ve } Y$ kümelerinin herhangi bir alt kümesine $X \to Y$ bir bağıntı denir.

Tanım 1.2.3. $f : X \to Y$ ye bir bağıntı olsun. Eğer aşağıdaki şartlar sağlanıyorsa f bağıntısına bir fonksiyon denir

(a) Her $x, y \in X$ için $(x, y) \in f$ olacak şekilde bir $y \in Y$ vardır.

(b) Eğer $(x,y)\in f$ ve $(x,z)\in f$ ise $y=z$ dir.

Metrik kavramı, Öklid (Euclid) uzayındaki uzunluk kavramından gelmektedir. Bir metrik uzay, kabaca üzerinde bir uzunluk kavramı tanımlı olan bir kümedir.

Tanım 1.2.4. $X\neq\varnothing$ bir küme olmak üzere aşağıdaki şartları sağlayan reel değerli bir $d:XxX\to\mathbb{R}$ fonksiyonuna **metrik,** (X,d) ikilisine bir **metrik uzay** denir.

1-) $d(x,y)=0$ dır ancak ve ancak $x=y$ (pozitif tanımlılık)

2-) $\forall x,y\in X$ için $d(x,y)=d(y,x)$ (simetri özelliği)

3-) $\forall x,y,z\in X$ için $d(x,z)\leq d(x,y)+d(y,z)$ (üçgen eşitsizliği)

Bu aksiyomlardan;

$0=d(x,x)\leq d(x,y)+d(y,x)=d(x,y)+d(x,y)=2d(x,y)$ olduğundan $\forall x,y\in X$ için $d(x,y)\geq 0$ dır.

Tanım 1.2.5. $X\neq\varnothing$ bir küme olmak üzere $\forall x,y,z\in X$ için

$$d(x,y)=\begin{cases}0, & \text{x = y}\\ 1, & x\neq y\end{cases}$$

ile tanımlanan d fonksiyonuna X üzerinde ayrık (discrete) metrik denir.

Tanım 1.2.6. $X\neq\varnothing$ bir küme olmak üzere aşağıdaki şartları sağlayan reel değerli bir $d:XxX\to\mathbb{R}$ fonksiyonuna **yarı metrik,** (X,d) ikilisine bir **yarı metrik uzay** denir.

1-) $\forall x\in X$ için $d(x,x)=0$

2-) $\forall x,y\in X$ için $d(x,y)=d(y,x)$ (simetri özelliği)

3-) $\forall x,y,z\in X$ için $d(x,z)\leq d(x,y)+d(y,z)$ (üçgen eşitsizliği)

Tanımdan da görüleceği gibi her metrik uzay bir yarı metrik uzaydır ancak her yarı metrik uzay bir metrik uzay değildir.

Tanım 1.2.7. Çok değişkenli veri yapısı, bir deneyde bir olayın araştırılması için n adet gözlem değerinin her biri bir özellik belirten $p>1$ sayıdaki değişkenin oluşturduğu ölçüm değerlerine çok değişkenli veri adı verilir. Her değişken bir boyut temsil etmektedir. Çok değişkenli veride değişkenlerin gözlem değerlerinin meydana getirdiği matrise çok değişkenli veri matrisi denir.

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1i} & \cdots & x_{1(n-1)} & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2i} & \cdots & x_{2(n-1)} & x_{2n} \\ \vdots & & \ddots & & & & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{ki} & \cdots & x_{k(n-1)} & x_{kn} \\ \vdots & & & & & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mi} & \cdots & x_{m(n-1)} & x_{mn} \end{bmatrix}_{mxn} \tag{1.2.1}$$

Yukarıdaki mxn tipindeki matriste satır elemanları verideki değişkenleri, sütun elemanları ise her bir değişkendeki gözlem değerlerini ifade etmektedir. X matrisindeki n adet değişkenin her birisinde m adet gözlem bulunduğu gözlenmektedir. x_{ki} elemanı $i.$ değişkenin $k.$ gözlemini göstermektedir.

X çok değişkenli veri matrisi $k=1,2,...,m$ ve $i=1,2,...,n$ olmak üzere her biri $1xn$ tipinde x_i gözlem vektörleri $X=\left(X_1^T, X_2^T, ...X_n^T\right)^T$ şeklinde gösterilir.

Tanım 1.2.8. X bir değişken olmak üzere $[X]=\{1,2,...,n\}$ in bir parçalanışı veya bölütlenmesi $K=\left\{A_1, A_2, ..., A_{|k|},\right\}$ aşağıdaki şartları $\forall i \in \{1,...,n\}$ için sağlıyorsa,

1) $A_i \subset [X]$

2) $A_i \neq \varnothing$

3) $i \neq j$ için $A_i \cap A_i = \varnothing$

4) $\bigcup_i A_i = [X]$

bir parçalanışlar (bölütlenme) kümesidir.

Tanım 1.2.9. Değişken tipleri sürekli ve kesikli değişkenler olmak üzere ikiye ayrılır. Eğer bir rasgele değişkenin alabileceği değerlerin sayısı sonlu veya sayılabilir çoklukta ise kesikli rasgele değişken (discrete random variable), eğer bir aralıkta ve sayılamaz sonsuzlukta ve her değeri alıyorsa sürekli rasgele değişken (continious random variable) denir [56]. Bir zar atıldığındaki üst yüze gelebilecek sayılar kesikli değişkenin, saat 14:15-16:00 arasındaki zaman aralığı sürekli değişkenin elemanları örnek olarak gösterilebilir.

Tanım 1.2.10. Kümeleme analizinde değişken ölçekleri değişkenlerin kullanım yeri ve amacına göre önem kazanmaktadır. Değişkenlerin ölçüm düzeyi matematiksel olarak nasıl ele alınacağına karşılık gelir. Değişkenlerin ölçüm düzeyleri kategorik, sıralı, aralık ve oran olmak üzere dört grupta toplanmıştır [57].

Tanım 1.2.11. Veri seti üzerinde kümeleme yapmak için verideki gözlemlerin benzerlikleri veya gözlem değerleri arasındaki uzaklıklar kullanılarak elde edilir. Değişkenlerin kesikli yada sürekli olmalarına ya da değişkenlerin aralık, oran, nominal veya ordinal ölçekte olmalarına göre hangi uzaklık veya benzerlik ölçüsünün kullanılacağına karar verilir. Veri setindeki gözlem çiftlerinin aralarındaki uzaklıklarına göre küme merkezlerine atanması için algoritmalar kullanılmaktadır. K-ortalamalar algoritması gibi kümeleme algoritmasında her elemanın küme merkezlerine uzaklığını ölçmek için farklı yöntemler kullanılabilir. En yaygın olarak kullanılan Öklid uzaklık ölçütüdür. Öklid uzaklık ölçütüne ek olarak, Manhattan uzaklık ölçütü ve Minkowski uzaklık ölçütü de sık kullanılan yöntemler arasındadır.

Tanım 1.2.12. Genel bir uzaklık ölçüsü olan Minkovski uzaklık ölçüsü değişkenlerdeki gözlem çiftleri arasındaki uzaklıkların bulunması için bir metrik bağıntısıdır. $p>1$ için L_p normu olarak adlandırılır. $x=(x_1,x_2,...,x_n)\in\mathbb{R}^n, y=(y_1,y_2,...,y_n)\in\mathbb{R}^n\ (n\geq 2)$ için,

$$d(x,y)=\left(\sum_{i=1}^{n}(x_i-y_i)^p\right)^{\frac{1}{p}},\ p>1 \tag{1.2.2}$$

ile tanımlanan d fonksiyonuna $\mathbb{R}^n$ üzerinde Minkovski uzaklığı denir.

Tanım 1.2.13. Manhattan uzaklığı Minkovski uzaklığının özel hali olup değişkende veri çiftleri arasındaki farkların ortalamasına eşittir. Manhatten uzaklıklar hesaplanırken

gürültü verilerinin etkisi minimum olmaktadır. Minkovski uzaklık ölçüsünde özel olarak $p=1$ alındığında Manhattan uzaklığı veya L_1 normu elde edilir. $x=(x_1,x_2,...,x_n)\in\mathbb{R}^n, y=(y_1,y_2,...,y_n)\in\mathbb{R}^n\ (n\geq 2)$ için,

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i| \tag{1.2.3}$$

olarak elde edilir. Bu uzaklık ölçüsünde birimler arasındaki mutlak uzaklık kullanılır. Manhattan uzaklık ölçüsüne, "city block uzaklık ölçüsü" adı da verilir.

Tanım 1.2.14. Öklid uzaklığında kullanılan yöntem ile standartlaştırılmış verilerle değil, verilerin ham hali ile hesaplama yapılır. Öklid uzaklıkları kümeleme analizinde verilere gürültü verisi (noise data) eklenmesinden etkilenmezler. Ancak değişkenler arasındaki ölçek farklılıkları Öklid uzaklıklarının hesaplamasında sapmalara neden olmaktadır. Öklid uzaklık formülü en yaygın olarak kullanılan uzaklık hesaplama formülüdür. $x=(x_1,x_2,...,x_n)\in\mathbb{R}^n, y=(y_1,y_2,...,y_n)\in\mathbb{R}^n\ (n\geq 2)$ için,

$$d(x,y)=\left(\sum_{i=1}^{n}(x_i-y_i)^2\right)^{\frac{1}{2}} \tag{1.2.4}$$

ile tanımlanan d fonksiyonuna $\mathbb{R}^n$ üzerinde Euclid uzaklığı denir.

Burada $p=2$ alındığında iki boyutlu uzayda L_2 normu olarak ta adlandırılan Öklid metriği, değişkenlerdeki gözlem çiftleri arasındaki uzaklıkların toplamlarının bulunmasında simetrik yapıda kullanılmaktadır.

Tanım 1.2.15. Sürekli değişkenler arasındaki yakınlığın bulunmasında karesel Mahalanobis uzaklığı kullanılabilir. Her biri n boyutlu $i.$ gözlem çiftleri arasındaki Mahalanobis uzaklığı,

$$d(x,y)=(x_i-y_i)^T\Sigma^{-1}(x_i-y_i),\quad i=1,2,...,n \tag{1.2.5}$$

ile elde edilir. Burada Σ^{-1} değişkenler veya örneklem varyans-kovaryans matrisidir. Mahalanobis uzaklığı kullanılarak elde edilen küme yapıları eliptik yapıdadır Mahalanobis uzaklığının hesaplanmasında Σ varyans-kovaryans matrisinin tersi her zaman hesaplanamayabilir. Veride değişkenler arasında ilişki bulunmadığı yani Pearson korelasyon katsayısı sıfır olduğu durumlarda matrisin köşegenleri haricindeki diğer elemanları sıfır olacağından birim matris yapısına dönüşür. Bu durumda Mahalanobis uzaklığı karesel Öklid uzaklığı gibi işlem görür [58].

Tanım 1.2.16. Gözlemler arasındaki uzaklık değişkenlerdeki gözlem çiftleri arasındaki ilişkinin yönüne ve büyüklüğüne bağlı olarakta elde edilebilir Cliff vd. [59]. Pearson korelasyon katsayısı değişkenlerin gözlem çiftleri arasındaki ilişkiyi [-1,1] aralığındaki değerlerde oransal olarak saptayabilmektedir. Pearson korelasyon katsayısı,

$$\rho_{ij} = \frac{\left(\sum_{k=1}^{n} \left(x_{ik} - \bar{x}_i \right) \left(x_{jk} - \bar{x}_j \right) \right)}{\left(\sum_{k=1}^{n} \left(x_{ik} - \bar{x}_i \right)^2 \sum_{k=1}^{n} \left(x_{jk} - \bar{x}_j \right)^2 \right)^{\frac{1}{2}}} \tag{1.2.6}$$

şeklinde elde edilebilmektedir. Burada $\bar{x}_i \quad i = 1, 2, ..., n$ gözlemlerinden oluşan n gözlem değerlerinin örneklem ortalaması olup

$$\bar{x}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ik} \tag{1.2.7}$$

şeklinde ifade edilmektedir. Pearson uzaklık ölçüsüne "karesel Pearson uzaklığı" veya "standardize Öklid uzaklığı" adı da verilir.

Tanım 1.2.17. Değişkenlerin standardizasyonu ve değişkenlerin dönüştürülmesi, çok değişkenli veri matrisindeki değişkenlerin ortalama vektörleri ve varyansları birbirinden mutlak değerce çok uzak olduğu durumlarda, gözlem çiftleri arasındaki uzaklıklar hesaplanırken ortalama vektörü veya varyansı daha büyük olan değişkenlerin hesaplaması yapılan uzaklık değerine daha fazla etkide bulunmaktadır. Böylece kümelemede istenmeyen sonuçlarla karşılaşılması söz konusu olmaktadır. Değişkenlerde aşırı değerler (noise data), uzaklık değerine etkisi olan başka bir etkendir. Aşırı değerler kümeleme analizi sonucunda ayrı bir küme olarak da oluşabilir. Aşırı kümelenmenin önüne geçmek için değişkenlerin dönüştürülmesi gerekmektedir. Verilerdeki dönüştürme işlemi

standardizasyon veya belirli aralıklara indirgeme işlemi olarak adlandırılır. Bu dönüştürme teknikleri z puanlarına dönüştürme, standart sapması 1 olacak şekilde indirgeme, en büyük değer 1 olacak şekilde indirgeme olarak ele alınabilir.

1.2.2. Değişkenlerdeki İstatistiksel Tanım ve Gösterimler

Tanım 1.2.18. X kesikli rasgele değişken olmak üzere,

1) Tanım bölgesi dışında $\left(x \notin \mathbb{R} \ \ için\right)\ f\left(x\right)=0$

2) Tanım bölgesi içinde $\left(x \in \mathbb{R} \ \ için\right)\ 0 \le f\left(x\right) \le 1 \quad 0 \le f\left(x\right);\quad x \in \mathbb{R} \ \ için$

3) Tanım bölgesindeki her değer için $\sum_{x} f\left(x\right)=1$

şartlarını sağlayan $f\left(x\right)$ fonksiyonuna kesikli olasılık ve X sürekli rasgele değişken olmak üzere,

1) $f\left(x\right)=0;\quad x \notin \mathbb{R} \ \ için$

2) $0 \le f\left(x\right);\quad x \in \mathbb{R} \ \ için$

3) $\int_{x \in \mathbb{R}} f\left(x\right).dx=1$

şartlarını sağlıyorsa, $f\left(x\right)$ fonksiyonuna sürekli olasılık fonksiyonu veya olasılık yoğunluk fonksiyonu denir. X sürekli rasgele değişkeninin olasılık yoğunluk fonksiyonu $f\left(x\right)$ olmak üzere beklenen değer,

1) Eğer X rasgele değişkeni kesikli ise,

$$E\left(X\right)=\sum_{X} x_i P\left(X=x_i\right)=\sum_{X} x\, f\left(x\right) \tag{1.2.8}$$

2) Eğer X rasgele değişkeni sürekli ise,

$$E\left(X\right)=\int_{-\infty}^{\infty} x\, f\left(x\right) dx \tag{1.2.9}$$

ile ifade edilir. Burada $E\left(X\right)$ in var olabilmesi için $E\left(X\right)$in mutlak yakınsak olması gerekmektedir. Yani $\left|E\left(X\right)\right|<\infty$ olmasıdır [60].

Tanım 1.2.19. X rasgele değişkeninin $E\left(X-E\left(X\right)\right)^2$ beklenen değerine X rasgele değişkeninin varyansı denir.

1) Eğer X rasgele değişkeni kesikli ise, $E\left(X\right)=\mu$ olmak üzere

$$\sigma^2=E\left[\left(X-\mu\right)^2\right]=\sum_x\left(x-\mu\right)^2 f\left(x\right) \tag{1.2.10}$$

2) Eğer X rasgele değişkeni sürekli ise, $E\left(X\right)=\mu$ olmak üzere

$$\sigma^2=E\left[\left(X-\mu\right)^2\right]=\int_{-\infty}^{\infty}\left(x-\mu\right)^2 f\left(x\right)dx=E\left(X^2\right)-\mu^2 \tag{1.2.11}$$

olarak elde edilir [56].

Tanım 1.2.20. X rasgele değişkeninin Moment çıkaran fonksiyonu $M_x\left(t\right)$ ile gösterilip t değişkeninin bütün gerçel değerleri için,

$$M_X\left(t\right)=E\left(e^{tX}\right) \tag{1.2.12}$$

ile tanımlanır.

1) Eğer X rasgele değişkeni kesikli ise,

$$M_X\left(t\right)=E\left(e^{tX}\right)=\sum_X e^{tx} f\left(x\right) \tag{1.2.13}$$

2) Eğer X rasgele değişkeni sürekli ise,

$$M_X\left(t\right)=E\left(e^{tX}\right)=\int_{-\infty}^{\infty} e^{tx} f\left(x\right)dx \tag{1.2.14}$$

olarak elde edilir. Normal dağılımın olasılık yoğunluk fonksiyonunun moment çıkaran fonksiyonu,

$$M_X\left(t\right)=e^{\mu t+\frac{1}{2}\sigma^2 t^2} \tag{1.2.14}$$

şeklinde bulunur.

Teorem1.2.1. $X_1, X_2, ..., X_n$ bağımsız rasgele değişkenleri $N\left(\mu_1,\sigma_1^2\right), N\left(\mu_2,\sigma_2^2\right), ..., N\left(\mu_n,\sigma_n^2\right)$ normal dağılımına sahip olsun. Bu durumda

$Y = k_1X_1 + k_2X_2 + ... + k_nX_n$ rasgele değişkenlerinin $k_1, k_2, ..., k_n$ reel sabitlerle dağılım ortalaması $k_1\mu_1 + k_2\mu_2 + ... + k_n\mu_n$ ve varyansı $k_1\sigma^2{}_1 + k_2\sigma^2{}_2 + ... + k_n\sigma^2{}_n$ olan normal dağılıma sahip olur ve $N\left(\sum_{i=1}^{n} k_i\mu_i, \sum_{i=1}^{n} k_i\sigma_i^2\right)$ olacak şekilde elde edilir.

İspat 1.2.1. $X_1, X_2, ..., X_n$ değişkenleri bağımsız değişkenler oldukları için Y değişkeninin moment çıkaran fonksiyonu,

$$M_y(t) = E\left(\exp\left[t\left(k_1X_1 + k_2X_2 + ... + k_nX_n\right)\right]\right) = E\left(e^{tk_1X_1}\right)E\left(e^{tk_2X_2}\right)...E\left(e^{tk_nX_n}\right) \qquad (1.2.15)$$

elde edilir. Normal dağılımın moment çıkaran fonksiyonu,

$E\left(\exp\left(tX_i\right)\right) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right)$; $\forall t, i = 1, ..., n$ için,

$$E\left(\exp\left(tk_iX_i\right)\right) = \exp\left(\mu_i\left(k_i t\right) + \frac{\sigma_i^2\left(k_i t\right)^2}{2}\right) \qquad (1.2.16)$$

elde edilir. (1.2.16) den elde edilen sonuca göre Y değişkeninin normal dağılımlarının moment çıkaran fonksiyonu,

$$M_{y(t)} = \prod_{i=1}^{n} \exp\left(\left(k_i\mu_i\right)t + \frac{\left(k_i^2\sigma_i^2\right)^2 t^2}{2}\right) = \exp\left(\left(\sum_{i=1}^{n} k_i\mu_i\right)t + \frac{\left(\sum_{i=1}^{n} k_i^2\sigma_i^2\right)^2 t^2}{2}\right) N\left(\sum_{i=1}^{n} k_i\mu_i, \sum_{i=1}^{n} k_i^2\sigma_i^2\right) \qquad (1.2.17)$$

elde edilmiş olur.

Tanım 1.2.21. Tek değişkenli normal dağılımın olasılık yoğunluk fonksiyonu μ ortalama vektörü, σ standart sapması olmak üzere

$$f\left(x; \mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-1}{2}\frac{\left(x-\mu\right)^2}{\sigma^2}\right\} \; x \in \mathbb{R} \qquad (1.2.18)$$

şeklinde ifade edilir.

Tanım 1.2.22. Tek değişkenli normal dağılımın olasılık yoğunluk fonksiyonu matematiksel olarak aşağıdaki gibi elde edilelir.

$f\left(x\right) = e^{\frac{-x^2}{2}}$ eşitliği ele alındığında

$$I = \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx \tag{1.2.19}$$

integralinin çözümü aranır. Bu integral bu hali ile çözülemeyeceğinden yardımcı bir I integrali,

$$I = \int_{-\infty}^{\infty} f(y)dy = \int_{-\infty}^{\infty} e^{\frac{-y^2}{2}} dy \tag{1.2.20}$$

tanımlanır. (1.2.19) ve (1.2.20) eşitliğinden yararlanılarak,

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\frac{-1}{2}\left(x^2+y^2\right)} dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy \tag{1.2.21}$$

eşitliği elde edilir. Bu formattan kutupsal koordinatlara geçilerek çözüm yapılırsa,

$x = r\cos\theta$, $y = r\sin\theta$ dönüşümü uygulanarak yeni integral,

$$\iint_G f(r\cos\theta, r\sin\theta) r dr d\theta \tag{1.2.22}$$

şeklini alır. Jakobiyen determinantı $J = \begin{vmatrix} \frac{dx}{dr} & \frac{dy}{dr} \\ \frac{dx}{d\theta} & \frac{dy}{d\theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & \sin\theta \\ -\mathrm{r}\sin\theta & \mathrm{r}\cos\theta \end{vmatrix} = rdrd\theta$

uygulandığında integral,

$$I^2 = \int_0^{2\pi}\int_0^{\infty} e^{\frac{-r^2}{2}} r dr d\theta = \int_0^{2\pi} d\theta = 2\pi \tag{1.2.23}$$

elde edilir. Ayrıca, $I^2 = 2\pi$ ise , $I = \sqrt{2\pi}$ olur. Dolayısıyla

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-x^2}{2}} dx = 1 \ ve \ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-y^2}{2}} dy = 1 \tag{1.2.24}$$

elde edilmiş olur. İntegralde $x = \frac{(x-\mu)}{\sigma}$, $\sigma > 0$ dönüşümü yapılıp yerine konulduğunda,

$$\int_{-b}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = 1 \tag{1.2.25}$$

olur. Buradan normal dağılım fonksiyonu olan

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],\ -\infty<x<\infty \qquad (1.2.26)$$

elde edilir.

Tanım 1.2.23. İki değişkenli normal dağılımların olasılık yoğunluk fonksiyonları μ ortalama vektör, Σ varyans-kovaryans matrisi olmak üzere

$$f(x;\mu,\Sigma)=\frac{1}{(2\pi)^{p/2}}.\frac{1}{|\Sigma|^{1/2}}.\exp\left(-\frac{1}{2}(x-\mu)|\Sigma|^{-1}(x-\mu)^T\right) \qquad (1.2.27)$$

denkleminde $p=2$ alınarak elde edilir. σ_i değişkenlerin standart sapması, ρ_i pearson korelasyon katsayısı olmak üzere iki değişkenli veride varyans-kovaryans matrisi

$$\Sigma=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_2\sigma_1 & \sigma_2^2\end{bmatrix} \qquad (1.2.28)$$

şeklinde gösterilir. Olasılık yoğunluk fonksiyonunda parametreler yerine konulup işlem yapıldığında iki değişkenli normal dağılımın olasılık yoğunluk fonksiyonu

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}.\frac{1}{\sqrt{1-\rho^2}}.\exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)\right) \qquad (1.2.29)$$

olacak şekilde $\mu, \sigma ve \rho$ parametrelerine bağlı olarak elde edilir.

Tanım 1.2.24. İki değişkenli normal dağılımın moment çıkaran fonksiyonu,

$$M_x(t_1,t_2)=\exp\left(\mu_1t_1+\mu_2t_2+\frac{\sigma_{11}t_1^2+2\rho_{12}\sigma_1\sigma_2t_1t_2+\sigma_{22}t_2^2}{2}\right) \qquad (1.2.30)$$

şeklinde tanımlanır.

Tanım 1.2.25. X_i , $i=1,2,...,p$ için normal dağılıma sahip rasgele değişkenler, ortalama vektörü μ_i varyans-kovaryans matrisi Σ_i olan çok değişkenli normal dağılımın olasılık yoğunluk fonksiyonu

$$f_i\left(x_j;\mu_i,\Sigma_i\right)=(2\pi)^{\frac{-p}{2}}\left|\Sigma_i\right|^{\frac{-1}{2}}\exp\left\{\frac{-1}{2}\left(x_j-\mu_i\right)^T\Sigma_i^{-1}\left(x_j-\mu_i\right)\right\} \tag{1.2.31}$$

şeklinde ifade edilir.

Tanım 1.2.26. Çok değişkenli normal dağılımın moment çıkaran fonksiyonu

$$M_X(t)-E\left(e^{t'x}\right)=\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty}\frac{1}{\sqrt{(2\pi)^n}\sqrt{|\Sigma|^{-1}}}\exp\left(t'x-\frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{2}\right)$$

$$=e^{t'\mu+\frac{t'\Sigma t}{2}} \tag{1.2.32}$$

olarak elde edilir.

Tanım 1.2.27. Çok değişkenli normal dağılımın beklenen değeri,

$$\mu=E(X)=\begin{bmatrix}E(X_1)\\E(X_2)\\\vdots\\E(X_p)\end{bmatrix}\begin{bmatrix}\mu_1\\\mu_2\\\vdots\\\mu_p\end{bmatrix} \tag{1.2.33}$$

şeklinde tanımlanmış olup kovaryansı,

$$\Sigma=E\left((X-\mu)(X-\mu)'\right)=E\left(\begin{bmatrix}(X_1-\mu_1)\\(X_2-\mu2)\\\vdots\\(X_p-\mu p)\end{bmatrix}\begin{bmatrix}(X_1-\mu_1) & (X_2-\mu_2) & \cdots & (X_1-\mu p)\end{bmatrix}\right)$$

$$=E\left(\begin{bmatrix}(X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_p-\mu_p)\\(X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_p-\mu_p)\\\vdots & \vdots & \ddots & \vdots\\(X_p-\mu_p)(X_1-\mu_1) & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_p-\mu_p)^2\end{bmatrix}\right)$$

$$=\begin{bmatrix}E\left((X_1-\mu_1)^2\right) & E((X_1-\mu_1)(X_2-\mu_2)) & \cdots & E\left((X_1-\mu_1)(X_p-\mu_p)\right)\\E((X_2-\mu_2)(X_1-\mu_1)) & E\left((X_2-\mu_2)^2\right) & \cdots & E\left((X_2-\mu_2)(X_p-\mu_p)\right)\\\vdots & \vdots & \ddots & \vdots\\E\left((X_p-\mu_p)(X_1-\mu_1)\right) & E((X_1-\mu_1)(X_2-\mu_2)) & \cdots & E\left((X_p-\mu_p)^2\right)\end{bmatrix}$$

$$= \begin{bmatrix} V(X_1) & Cov(X_1X_2) & \cdots & Cov(X_1X_p) \\ Cov(X_2X_1) & V(X_2) & \cdots & Cov(X_2X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_pX_1) & Cov(X_pX_2) & \cdots & V(X_p) \end{bmatrix} \qquad (1.2.34)$$

şeklinde elde edilir.

2. BÖLÜM

2.1. KÜMELEME ANALİZİNDE KULLANILAN YÖNTEMLER

Kümeleme analizi yöntemleri; nesneleri, desenleri, bireyleri, birimleri, olayları veya gözlemleri belirli sayıdaki kümelere, gruplara, alt kümelere veya kategorilere bölen yöntemlerdir. Başka bir deyişle kümeleme analizi, X veri matrisindeki doğal gruplamaları kesin olarak bilinmeyen nesneleri, değişkenleri, nesne ve değişkenleri birbiri ile benzer alt gruplara ayırmakta kullanılan yöntemlerdir. Küme terimi için genel olarak ifade edilebilecek tek bir tanım olmadığı gibi, tanımlama yapılırken yanlış anlamlara yol açmakta mümkündür [1]. Küme terimi için kısaca:

1) Küme birbirine benzeyen bireylerin kümesi olmasının yanında kümedeki bireylerin aynı zamanda diğer kümedeki bireylerden farklı olması durumudur.

2) Küme örneklem uzayında noktaları kapsayan ve küme içindeki iki nokta çifti arasındaki uzaklığın biri küme içinde diğeri küme dışındaki iki nokta arasındaki uzaklıktan küçük olmasıdır.

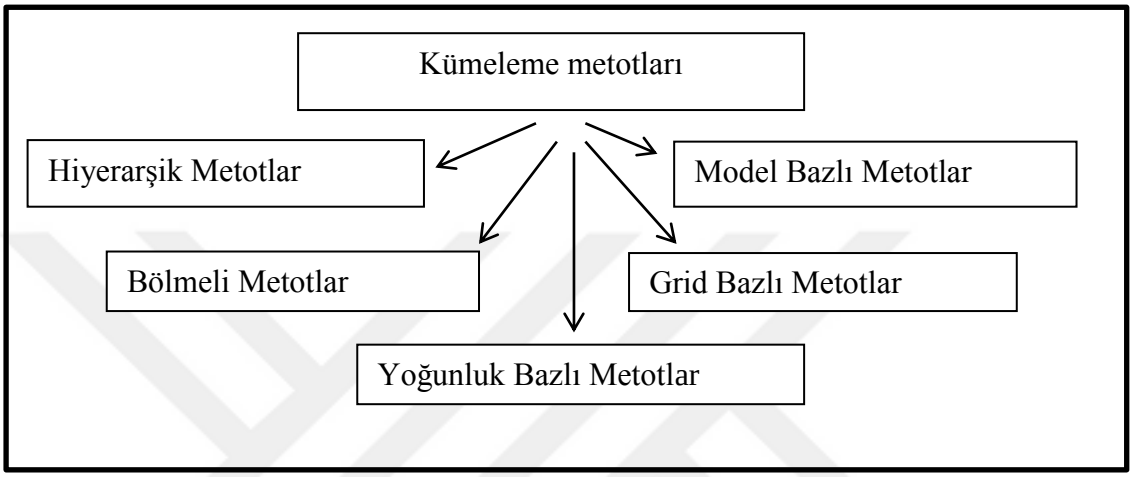
3) Kümeler p boyutlu özellikler uzayında yoğunluklu noktaları içeren ve bunlara kıyasla daha az yoğunluktaki noktaları içeren diğer bölgelerden ayrılan sürekli bölgeler olarak tanımlanabilir [61].

Bu küme tanımları altında kümeleme aynı küme içerisindeki gözlemlerin birbirlerine benzer, diğer kümelerdeki gözlemlerden farklı olacak şekilde yapılmasıdır. Bu amaç için kullanılan benzerlik ve farklık kavramları anlaşılabilir ve mantıklı olmalıdır [62].

Kümeleme analizi için " Kümeleme analizi, temel amacı nesneleri sahip olduğu karakteristik özellikleri temel alarak gruplamak olan çok değişkenli teknikler grubudur. Kümeleme analizi, nesneleri küme içerisinde çok benzer, kümeler arasında farklı olacak biçimde kümeler. Kümeleme işlemi başarılı olursa, bir geometrik çizimle nesneler küme

içinde birbirine çok yakın, kümeler ise, birbirinden çok uzak olacaktır ” [63] şeklinde farklı bir tanımlamada yapılmaktadır.

Değişik kaynaklarda kümeleme metotları farklı şekillerde sınıflandırılır. Çok değişkenli veri analizi kitaplarında kullanılan en genel kümeleme metotları hiyerarşik kümeleme ve hiyerarşik olmayan (bölünmeli) kümeleme metotlarıdır.



Şekil 2.1.1. Kümeleme analizinde kullanılan metotların şematik gösterimi

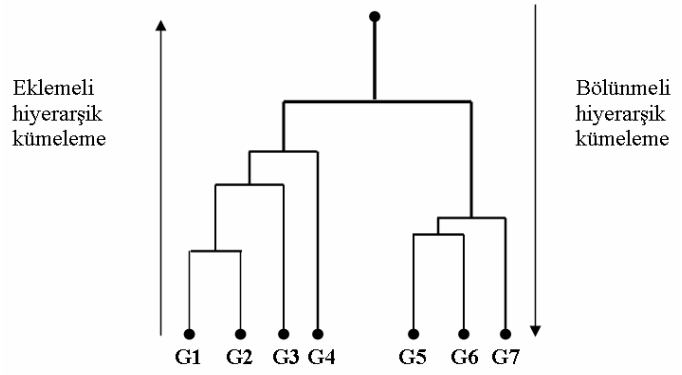
2.1.1. Hiyerarşik Kümeleme Metotları

Hiyerarşik kümeleme metotları verideki gözlemlerin birbirlerine olan uzaklıklarını kullanarak, verideki birimlerin hiyerarşik ayrıştırmasını yapar. Hiyerarşik kümelemede “dendogram“ olarak adlandırılan ağaç veri yapısı kullanılır. Dendogram kümelemedeki grupları ve gruplara düşen elemanları görsel olarak belirtir. Dendogram kök, iç düğüm ve yapraklardan oluşan bir yapıya sahiptir. Dendogram kökü tüm birimlerin bir araya gelmesiyle oluşan ana kümeyi içerir. Dendogramın yaprakları da bir araya getirilmeyen tek bir elemandan oluşan kümeleri içerir. Dendogramın iç düğümleri elemanların bir araya gelerek oluşturdukları kümeyi gösterir.

n gözlemden oluşan bir veri k farklı küme yapısı içeriyorsa, n gözlemin k kümeye parçalanmalarının sayısı,

$$N(n,k)=\frac{1}{k!}\sum_{k=1}^{n}\binom{n}{k}(-1)^{n-k}\,k^{n} \tag{2.1}$$

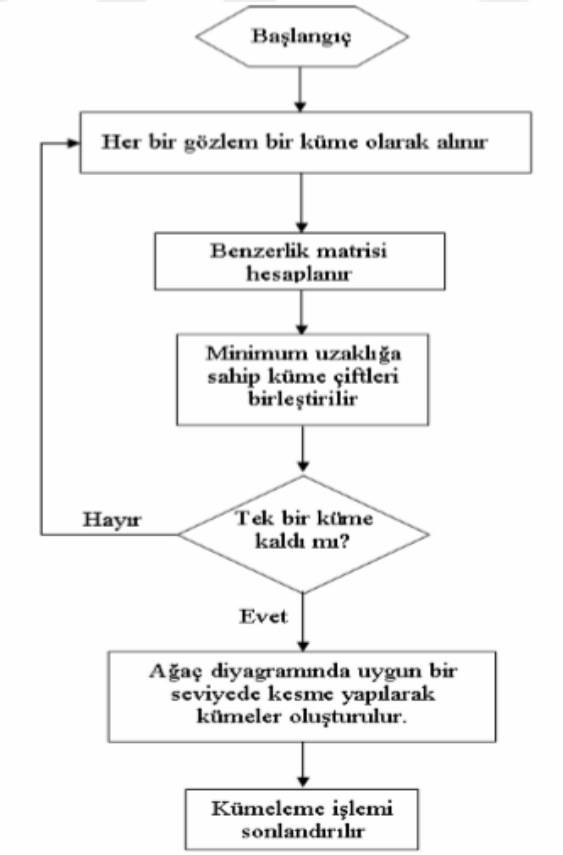
ile elde edilir [64]. Bu parçalanma sayısı $\frac{k^{n}}{k!}$ oranından da yaklaşık olarak hesaplanabilir.



Şekil 2.1.2. Hiyerarşik kümelemede Dendogram (ağaç verisi) yapısı.

2.1.1.1. Yığılmalı Hiyerarşik Kümeleme Yöntemi

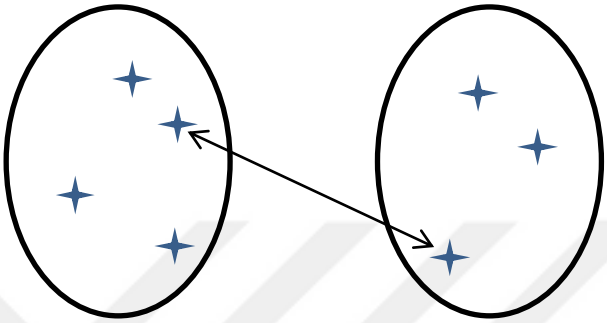
Yığılmalı hiyerarşik kümeleme yönteminde verideki n sayıda gözlemin her biri bir küme olarak düşünülür. Bu kümeler bir seri birleştirme işlemleri ile bu kümeler tek bir küme oluşturuncaya kadar uygulanır. Bu yönteme yığılmalı (agglomerative) ya da kaynaştırmalı hiyerarşik kümeleme denir [65]. Genel bir yığılmalı hiyerarşik kümeleme yönteminin adımları aşağıda verilmiştir



Şekil 2.1.3. Yığılmalı hiyerarşik kümeleme metodunun akış diyagramı.

2.1.1.1.1. Tek Bağlantı Algoritması

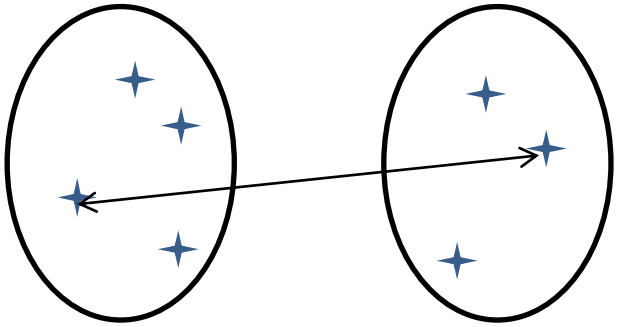
Tek bağlantı algoritması iki farklı küme arasındaki uzaklığın, her bir kümedeki iki eleman arasındaki en kısa mesafedir. Tek bağlantı kümeleme aynı zamanda en yakın komşuluk yöntemi olarak da adlandırılır [66].



Şekil 2.1.4. Tek bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi

2.1.1.1.2. Tam Bağlantı Algoritması

Tam bağlantı yönteminde, iki farklı kümedeki birbirine en uzak gözlem çifti arasındaki uzaklık değeri iki küme arasındaki uzaklık olarak alınır.



Şekil 2.1.5. Tam bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi

2.1.1.1.3. Ortalama Bağlantı Algoritması

Ortalama bağlantı yönteminde ayrı gruplarda yer alan gözlem çiftleri arasındaki ortalama uzaklık iki küme arasındaki uzaklık olarak alınır [1]. Bu yöntem aynı zamanda ağırlıksız grup çiftleri yöntemi olarak da adlandırılır [66].

2.1.1.1.4. Ağırlıklı Ortalama Bağlantı Algoritması

İki küme arasındaki uzaklığın bulunması için ortalama bağlantı yönteminde olduğu gibi bir uzaklık hesaplanır. Bu teknikte ortalama bağlantıdan farklı olarak yeni oluşan küme ile diğer kümeler arasındaki uzaklık her bir kümedeki gözlem sayısı ile değerlendirilir.

2.1.1.1.5. Merkezi Bağlantı Algoritması

Merkezi bağlantı yönteminde iki küme arsasındaki uzaklık kümelerin kendi merkezleri arasındaki uzaklık olarak alınır. G_i ile ifade edilen $i.$ kümenin merkezi ya da ortalaması,

$$m_i = \frac{1}{n_i} \sum_{x \in G_i} X \tag{2.1.1}$$

ile bulunur. Burada n_i , G_i kümesindeki gözlem sayısını ifade eder.

2.1.1.1.6. Medyan Bağlantı Algoritması

Merkezi bağlantı yönteminden farklı olarak, medyan bağlantı yönteminde iki küme arasındaki uzaklık, iki kümenin merkezleri arasındaki uzaklığın eşit ağırlıklı olarak hesaplanmasıyla elde edilir [67].

2.1.1.1.7. Ward Algoritması

Artırımlı (incremental) hata kareler toplamı yöntemi olarak da adlandırılan Ward'ın hiyerarşik kümeleme yönteminde amaç artan sınıf içi hata kareler toplamının minimum yapılmasıdır [1], [68]. Sınıf içi hata kareleri toplamı,

$$E = \sum_{k=1}^{K} \sum_{x_i \in G_i} \left\| x_i - m_k \right\|^2 \tag{2.1.2}$$

ile ifade edilir. Burada K , küme sayısını ve $k = 1,...,K$ olmak üzere m_k, (2.1.1) de gösterilen $k.$ küme ortalamasını gösterir.

2.1.2. Bölmeli Kümeleme Metotları

Bölmeli metotlar hiyerarşik olmayan kümeleme metotlarıdır. Bu metotlar n adet elemandan (gözlem) oluşan veri setini başlangıçta belirlenen $k < n$ olmak üzere k adet kümeye ayırmak için kullanılır. Bölmeli metotların hiyerarşik metotlardan en belirgin farkı budur. Bölmeli metotlar hiyerarşik metotlara göre daha büyük veri setlerine uygulanabilir. Bölmeli metotlarda oluşturulacak k adet kümede her bir küme en az bir

elemen içerir ve her elemen yalnız bir grupta bulunur. Bölmeli metotlarda kullanılan işlem sırasını şöyle özetleyebiliriz,

1) Başlangıç küme merkezleri rasgele seçilir.

2) Elemanların seçilen kümelerin merkezlerine olan uzaklıklarına göre yeni küme merkezleri belirlenir.

3) Bu işlemler birbirinden farklı kendi içinde homojen ve birbirleri arasında benzerlik bulunmayan k adet küme oluşturuluncaya kadar devam eder.

Bölmeli metotlar arasındaki en önemli metot k-ortalamalar (k-means) algoritmasıdır.

2.1.2.1. K-ortalamalar Algoritması

K-ortalamalar algoritması [69], günümüzde kümeleme için kullanılan en yaygın denetimsiz öğrenme yöntemlerinden biridir [43]. k –ortalamalar algoritması, hata kareler toplamının en aza indirgenmesine dayalı olarak veri yapısına en uygun parçalanmayı belirlemek için kullanılan ve adım adım işlem yapan bir algoritma kullanan optimizasyon yöntemidir.

K-ortalamalar algoritması değişkendeki n adet veriyi k adet farklı küme oluşturacak şekilde gruplara ayırmak prensibine göre çalışmaktadır. Algoritmanın çalışma prensibi her gözlemin sadece bir kümenin elemanı olmasına dayalı olarak işlemektedir. Birbirine yakın veriler aynı kümede yer alırken birbirinden uzak veriler farklı kümelerde yer alırlar. Algoritmanın amacı; gerçekleştirilen gruplama işleminin sonunda küme içi benzerlikleri maksimum (homojenlik), kümeler arası benzerlikleri ise minimum (heterojenlik) hale getirmektir. Küme içi benzerlik, kümenin ağırlık merkezi olarak seçilen gözlemin kümedeki diğer gözlemler arasındaki uzaklıkların ortalama değeri ile ölçülmektedir. Bulunan ortalama değerlerine göre kümeleme, verilerin en yakın oldukları küme merkezlerine atanması ile elde edilir.

Kümelemede önceden bilinen k sayısı, kümeleme işlemi bitene kadar değeri değişmeyen sabit bir pozitif tam sayıdır. Algoritmanın adı bu k sabit sayısından gelmektedir.

K-algoritmalar algoritmasının performansını başlangıçta belirlenen k sayısı, başlangıç olarak seçilen merkez değerleri ve seçilen benzerlik ölçüsü metotları etkilemektedir. Sadece sayısal verilerde kullanılabilir olması yani kategorik verilerde kullanılamaması K-ortalamalar algoritmasının olumsuz yönlerindendir.

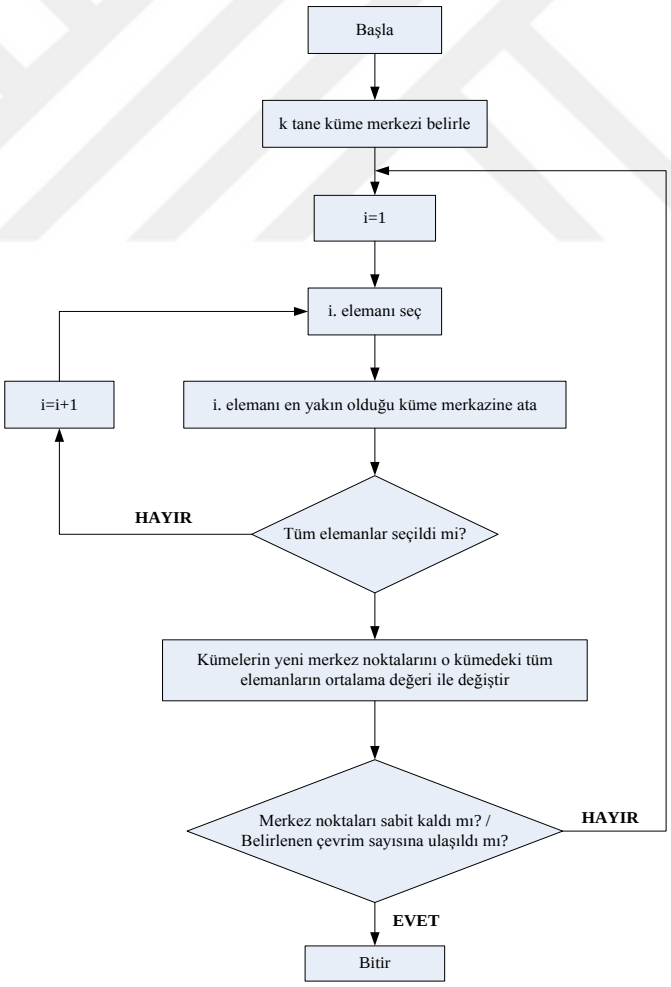
K-ortalamalar algoritmasının çalışma sistemi aşağıdaki adımlarda şu şekilde gösterilebilir,

1. Adım: Başlangıç küme merkezleri rasgele belirlenebilir veya bu işlem için farklı bir algoritmalar kullanılabilir.

2. Adım: Her gözlem değerinin belirlenen merkezlere olan uzaklıkları hesaplanır. Elde edilen sonuçlara göre tüm gözlemler k satıdaki kümeden en yakın kümeye atanır.

3. Adım: Oluşan kümelerin yeni merkez noktaları o kümedeki tüm elemanların ortalama değeri ile değiştirilir.

4. Adım: Merkez noktalar değişmeyene kadar algoritmanın seçme adımları yani 2. ve 3. adımlar tekrarlanır.



Şekil 2.1.6. K-ortalamalar algoritmasının akış diyagramı.

2.1.3. Paylaştırmalı (Partitional) Kümeleme Yöntemleri

Paylaştırmalı, ayırmalı veya bölmeli kümeleme metotlarında gözlemler benzerlik ya da uzaklık matrislerine dayalı olan bir hiyerarşik yaklaşım kullanılmadan K kümeye ayrılır. Küme sayısı K daha önceden belirlenmiş olabileceği gibi kümeleme yöntemi süresinde de hesaplanabilir. Paylaştırmalı kümeleme yöntemlerinde amaç n sayıda gözlemin K kümeye parçalanması için bütün ihtimallerin incelenmesi ve belli kriterlere göre en uygun kümelenmenin belirlenmesidir. Yalnız burada sorun parçalanma sayılarının oldukça fazla olmasıdır. Bu nedenden dolayı veride gözlem sayısı, küme sayısı veya her ikisinin de fazla olması durumunda, kümeleme için basit yöntemlere ihtiyaç duyulur.

2.2. Sonlu Karma Dağılımlar

Tanım 2.2.1 $\boldsymbol{x}$, $\boldsymbol{X}$'in gözlenen değerini göstersin. $i = 1,2,...,g$ için 0 ve 1 gösterge değişkenlerinin g-boyutlu $\boldsymbol{W}$ etiket vektöründe $x \in G_i$ ise $w_i = 1$ ve $x \notin G_i$ ise $w_i = 0$ olur. $i = 1,2,...,g$ için G_i grubundaki $\boldsymbol{X}$ değişkeninin olasılık yoğunluk fonksiyonu $f_i(\boldsymbol{x})$ ile gösterilsin. Ayrıştırma analizine ve kümeleme analizine karma dağılım modeli yaklaşımı altında $G_1, G_2, ..., G_g$ şeklindeki g grubun bir G karmasından çekilen elemanlar sırasıyla $\pi_1, \pi_2, ..., \pi_g$ olasılık oranlarına sahiptir. Burada $\sum_{i=1}^{g} \pi_i = 1$ ve $\pi_i \geq 0$ dır. G'deki $\boldsymbol{X}$'in olasılık yoğunluk fonksiyonu, $f_X(\boldsymbol{x}) = \sum_{i=1}^{g} \pi_i \, f_i(\boldsymbol{x})$ biçiminde bir sonlu karma dağılım modeliyle gösterilebilir [11].

Tanım 2.2.2 Çok değişkenli normal karma dağılımların parametrik gösterimi, g bileşenli karma dağılım modeli altında X_j rasgele vektörünün çok değişkenli karma dağılım modeli genel olarak,

$$f\left(x_j; \Psi\right) = \sum_{i=1}^{g} \pi_i \, f_i\left(x_j; \Theta_i\right) \tag{2.2.1}$$

formuna sahiptir [11]. Burada $\boldsymbol{\Psi}$ parametre vektörü karma dağılım modelindeki tüm parametreleri içerir. $\boldsymbol{\Psi}$ parametre vektörü $\Psi = \left(\pi_1, \pi_2, ..., \pi_{g-1}, \xi^T\right)^T$ şeklinde

yazılabilir. ξ parametre vektörü birbirinden farklı olduğu önbilgisine sahip olunan $\Theta_1, \Theta_2, ..., \Theta_g$'deki parametreleri içerir. π_i'ler $0 < \pi_i < 1$ olmak üzere $\sum_{i=1}^{g} \pi_i = 1$ dir. $\pi = \left(\pi_1, \pi_2, ..., \pi_g\right)^T$ şeklinde yazılabilir. (1)'deki çok değişkenli karma dağılımın olasılık yoğunluk fonksiyonunda: f, çok değişkenli karma dağılımın olasılık yoğunluk fonksiyonunu; x_j, rasgele değişkenleri; $\Psi = (\pi, \Theta)$ olmak üzere parametre vektörünü; g, bileşen sayısını; π_i, bileşenlerin karma oranlarını: f_i, bileşen olasılık yoğunluk fonksiyonlarını ve $\Theta_i = \left(\mu_i, \Sigma_i\right)$, bileşen olasılık yoğunluk fonksiyonlarının parametrelerini göstermektedir.

2.2.1. Çok Değişkenli Normal Karma Dağılım Fonksiyonu

Çok değişkenli Normal dağılımların karma yoğunluk fonksiyonu $i = 1, 2, ..., g \quad ve \quad j = 1, 2, ..., n$ olmak üzere,

$$f\left(x_j; \Psi\right) = \sum_{i=1}^{g} \pi_i f_i\left(x_j; \Theta_i\right) \tag{2.2.2}$$

ile verilir. $f_i\left(x_j; \Theta_i\right) = f_i\left(x_j; \mu_i, \Sigma_i\right)$ fonksiyonu ortalama vektörü μ_i varyans-kovaryans matrisi Σ_i olan,

$$f_i\left(x_j; \mu_i, \Sigma_i\right) = (2\pi)^{\frac{-p}{2}} \left|\Sigma_i\right|^{\frac{-1}{2}} \exp\left\{\frac{-1}{2}\left(x_j - \mu_i\right)^T \Sigma_i^{-1}\left(x_j - \mu_i\right)\right\} \tag{2.2.3}$$

şeklinde ifade edilen çok değişkenli normal dağılım fonksiyonudur. Bu durumda karma modelin tüm bilinmeyen parametreler vektörü $\Psi = \left(\pi_1, \pi_2, ..., \pi_{g-1}, \xi^T\right)^T$ şeklinde yazılır. Burada ξ karma dağılım modelindeki bileşen olasılık yoğunluk fonksiyonlarının parametreleri olan bileşen ortalama vektörleri $\mu = \left(\mu_1, \mu_2, ..., \mu_g\right)$ ve bileşen varyans-kovaryans matrisleri $\Sigma = \left(\Sigma_1, \Sigma_2, ..., \Sigma_g\right)$ dan oluşur. $i = 1, 2, ..., g \quad ve \quad j = 1, 2, ..., n$ olmak üzere x_j gözlem vektörünün sonraki *i.* bileşenine ait olma olasılığı $\tau_i = \left(x_j; \Psi\right)$,

$$\tau_i = \left(x_j; \Psi\right) = \frac{\pi_i f_i\left(x_j; \mu_i, \Sigma_i\right)}{\sum_{m=1}^{g} \pi_m f_i\left(x_j; \mu_i, \Sigma_i\right)} \tag{2.2.4}$$

Olarak elde edilir. EM algoritmasının $(k+1)$. iterasyonunundaki M-adımında güncellenmiş karma oranları π_i ve ortalama vektörü μ_i nin en çok olabilirlik kestiricileri (maximum likelihood) sırasıyla,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_i\left(x_j; \Psi^{(k)}\right)}{n} \tag{2.2.5}$$

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i\left(x_j; \Psi^{(k)}\right) x_j}{\sum_{j=1}^{n} \tau_i\left(x_j; \Psi^{(k)}\right)} \tag{2.2.6}$$

olarak elde edilir. Bileşen olasılık yoğunluklarının varyans-kovaryans matrisi Σ_i nin güncel tahmini,

$$\Sigma_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i\left(x_j; \Psi^{(k)}\right)\left(x_j - \mu_i^{(k+1)}\right)\left(x_j - \mu_i^{(k+1)}\right)^T}{\sum_{j=1}^{n} \tau_i\left(x_j; \Psi^{(k)}\right)} \tag{2.2.7}$$

şeklinde hesaplanır.

2.2.2. Çok Değişkenli Karma Normal Modeller

Çok değişkenli normal karma dağılımlar kullanılarak modele dayalı kümeleme analizini Celeux ve Govaert [71] önermişlerdir. Burada karma normal dağılımın bileşen olasılık yoğunluk fonksiyonlarındaki varyans-kovaryans matrisinin bazı kısıtlamaları ve varyans-kovaryans matrisinin geometrik özellikleri model kelimesi için kullanılmıştır [70]. Celeux ve Govaert [71], çok değişkenli normal dağılımların bileşen dağılımlarındaki varyans-kovaryans matrisi Σ_i'yi,

$$\Sigma_i = \lambda_i D_i A_i D_i^T \tag{2.2.8}$$

şeklinde özdeğer ayrışımı ile ifade etmiştir. Burada p değişken sayısını göstermek üzere $\lambda_i = |\Sigma_i|^{\frac{1}{p}}$, D_i Σ_i'nin öz vektörlerinin matrisini, A_i Σ_i'nin bileşen varyans-kovaryans matrisinin normalleştirilmiş öz değerlerini azalan sırada içeren bir köşegen matris olup determinantı $|A_i| = 1$ dir. λ_i $i.$kümenin hacmini (büyüklüğünü), D_i yönünü ve A_i şeklini belirler.

Celeux ve Govaert [71] modelleri daha kolay elde edebilmek için karakteristiklerin hepsi eşanlı olmamak üzere λ_i, D_i ve A_i parametrelerinin kabul edilen çeşitleri ile farklı 14 varyans-kovaryans matris yapısı ileri sürmüşlerdir. Farklı 14 varyans-kovaryans matris yapısına göre karma dağılımdaki bileşen olasılıklarının eşit olması veya olmaması durumunda yirmi sekiz çok değişkenli normal karma modeli elde edilir. Bu modeller üç kategori altında toplanır. Bu modeller genel, köşegen ve küresel varyans-kovaryans model aileleri olarak farklı üç gruba ayrılırlar. Genel varyans-kovaryans yapı ailesi içerisinde 8, köşegen varyans-kovaryans yapı ailesi içerisinde 4 ve küresel varyans-kovaryans yapı ailesinde 2 farklı karma model vardır. Bu modeller Tablo 2.2.1 - Tablo 2.2.3 ile verilmiştir.

Tablo 2.2.1. Karma normal modellerin genel bileşen varyans-kovaryans yapıları.

MODEL	HACİM	ŞEKİL	YÖN
$\Sigma_i = \lambda DAD^T$	Sabit	Sabit	Sabit
$\Sigma_i = \lambda_i DAD^T$	Farklı	Sabit	Sabit
$\Sigma_i = \lambda DA_iD^T$	Sabit	Farklı	Sabit
$\Sigma_i = \lambda D_iAD_i^T$	Sabit	Sabit	Farklı
$\Sigma_i = \lambda_i DA_iD^T$	Farklı	Farklı	Sabit
$\Sigma_i = \lambda_i D_iAD_i^T$	Farklı	Sabit	Farklı
$\Sigma_i = \lambda D_iA_iD_i^T$	Sabit	Farklı	Farklı
$\Sigma_i = \lambda_i D_iA_iD_i^T$	Farklı	Farklı	Farklı

Tablo 2.2.2. B köşegen bir matris olmak üzere karma normal modellerin köşegen bileşen varyans-kovaryans yapıları.

MODEL	HACİM	ŞEKİL	YÖN
$\Sigma_i = \lambda B$	Sabit	Sabit	Eksen
$\Sigma_i = \lambda_i B$	Farklı	Sabit	Eksen
$\Sigma_i = \lambda B_i$	Sabit	Farklı	Eksen
$\Sigma_i = \lambda_i B_i$	Farklı	Farklı	Eksen

Tablo 2.2.3. I birim matris olmak üzere karma normal modellerin küresel bileşen varyans-kovaryans yapıları.

MODEL	HACİM	ŞEKİL	YÖN
$\Sigma_i = \lambda I$	Sabit	Sabit	-
$\Sigma_i = \lambda_i I$	Farklı	Sabit	-

2.2.3. EM Algoritması

Tanım 2.2.3.1. $i = 1,...,g$ olmak üzere π_i, $\boldsymbol{\mu}_i$ ve $\boldsymbol{\Sigma}_i$ parametreleri, oluşturulan her bir sentetik veri için likelihood fonksiyonları kullanılarak EM algoritmasıyla aşağıdaki eşitlikler ardışık hesaplanarak tahmin edilebilir [11].

$$\mathbf{z}_j^{i(k)} = \frac{\boldsymbol{\pi}_i^{(k)} f_i(\mathbf{x}_j;\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)})}{\sum_{i=1}^{g} \boldsymbol{\pi}_i^{(k)} f_i(\mathbf{x}_j;\boldsymbol{\mu}_i^{(k)},\boldsymbol{\Sigma}_i^{(k)})} \tag{2.2.9}$$

$$\boldsymbol{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \mathbf{z}_j^{i(k)}}{n} \tag{2.2.10}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{j=1}^{n} (\mathbf{z}_j^{i(k)} \mathbf{x}_j)}{\sum_{j=1}^{n} \mathbf{z}_j^{i(k)}} \tag{2.2.11}$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \mathbf{z}_j^{i(k)} (\mathbf{x}_j - \boldsymbol{\mu}_i^{(k+1)})(\mathbf{x}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \mathbf{z}_j^{i(k)}} \tag{2.2.12}$$

Burada $\mathbf{z}_j^{i(k)}$, k. iterasyonda i. gruptaki j. gözlem değerinin sonraki olasılığını gösterir. $\boldsymbol{\pi}_i^{(k+1)}$ değeri, $(k+1)$ iterasyonda i. bileşenin karma oranlarını gösterir. $\boldsymbol{\mu}_i^{(k+1)}$ değeri, $(k+1)$ iterasyonda i. bileşenin ortalama vektörünü ve $\boldsymbol{\Sigma}_i^{(k+1)}$ değeri de $(k+1)$ iterasyonda i. bileşenin varyans-kovaryans matrisini gösterir.

2.2.4. Çok Değişkenli Sonlu Karma Dağılımlarda Model Oluşturma

Çok değişkenli karma modellerdeki her bir değişkenin anlamlı alt gruplara ayrılmasına değişkenlerin parçalanması denir. Değişkenlerdeki parçalanmalara karar vermek bazen çok kolay olmayabilir. Bir değişkende parçalanmanın olmaması verinin homojen olması anlamına gelmektedir. Eğer değişkende birden fazla alt grup veya parçalanma bulunuyorsa veri homojen olmayan yani heterojen yapıya sahiptir denir. Modele dayalı kümeleme p-boyutlu çok değişkenli heterojen veriyi anlamlı alt gruplara bölmek için birçok kümeleme metodundan birisidir [6]. Çok değişkenli normal dağılımların karmasının her bileşeni, çok değişkenli heterojen verideki bir kümeye karşılık gelir [7]. Çok değişkenli heterojen verinin değişkenlerindeki parçalanmaları diğer bir ifadeyle karma dağılımdaki g bileşen sayısını belirlemek için histogram veya P-P grafiksel yöntemlere başvurulabilir [41]. Çok değişkenli heterojen verideki normal karma dağılım modelindeki bileşen sayısını belirlemek için verideki her bir değişken tek değişkenli normal karma gibi alınıp değişkendeki loglikelihood, AIC ve BIC değerlerine bakılarak karar verilebilir. Elde edilen değerlere göre verideki en uygun parçalanma sayısının loglikelihood değerinin maksimum ve AIC ile BIC değerlerinin minimum olduğu sayı olarak karar verilir. Her bir değişkendeki k_s parçalanma sayısı belirlendikten sonra, bu parçalanmalara karşılık gelen modeldeki kümelenme sayısı veya çok değişkenli normal karma dağılım modelindeki g bileşen sayısı hesaplanabilir [41].

2.2.4.1. Çok Değişkenli Sonlu Karma Dağılımlarda Model Sayısının Belirlenmesi

Çok değişkenli heterojen verideki değişkenlerin yapısı çok değişkenli heterojen verideki kümelenmeyi belirlemektedir [41]. p-boyutlu çok değişkenli heterojen verideki $\boldsymbol{X}_S$ değişkeninin yapısı ya da bileşenlerinin sayısının $k_S \geq 1$ olmak üzere k_S olduğunu varsayalım. Bu durumda her bir değişken için k_S değeri k-means algoritması gibi bir

ayrıştırma-kümeleme algoritması kullanılarak elde edilebilir. p-boyutlu çok değişkenli heterojen verideki s. değişken ve j. gözlem arasındaki ilişki x_{sj} olarak gösterilebilir. Burada x_{sj}, p-boyutlu çok değişkenli heterojen verideki s. değişkeninin j. gözlemini ifade eder. Çok değişkenli heterojen verideki kümelenmeyi verideki her bir değişkenin yapısı ve çok değişkenli normal dağılımların karmalarını kullanarak modele dayalı biçimde elde etmek veya belirlemek için öncelikle veride her bir değişkendeki bölünmeye bakılır. Veride her bir değişkendeki bölünmelere göre verinin kümelenme merkezleri ve bu merkezlerin oluşturabileceği kümelenme yapıları belirlenir. $C_{\max}$ ile gösterilen değişkenlerdeki bölünmelerin oluşturduğu kümelenme merkezlerinin maksimum sayısı, Servi ve Erol [41] tarafından önerilen

$$C_{\max} = \prod_{s=1}^{p} k_s \tag{2.2.13}$$

eşitliğiyle hesaplanır. $C_{\min}$ ile gösterilen değişkenlerdeki bölünmelerden oluşan kümelenme merkez sayısının minimum sayısı,

$$C_{\min} = \max\{k_s\} \tag{2.2.14}$$

eşitliğiyle hesaplanır. Burada p, değişken sayısını ve k_s, her bir değişkendeki bölünme sayısını göstermektedir.

M_{Toplam} ile gösterilen normal dağılımların karmalarıyla oluşturulabilecek toplam model sayısı,

$$M_{Toplam} = 2^{C_{\max}} - 1 \tag{2.2.15}$$

eşitliğinden elde edilebilir.

2.2.4.2. Sonlu Karma Dağılım Modellerinde Uygun Aday Model Sayısının Belirlenmesi

Sonlu karma dağılım modellerinde değişkenlerdeki parçalanmalardan kaynaklanan kümelenme merkezlerinin sayısı, kümelenme merkezlerinin oluşturduğu oluşabilecek toplam model sayısı M_{Toplam}, Servi ve Erol [41] tarafından önerildi. Çok değişkenli

heterojen verideki değişkenler ve bu değişkenlerdeki parçalanmalardan kaynaklanan kümelenme merkezlerinin modeldeki yeri ve sayısı varsayıma uyan modeller olarak hesaplanabilir. Karma dağılım modelindeki değişken sayısı ve her bir değişkendeki parçalanmaların oluşturduğu maksimum kümelenme sayısı $C_{\max} = \prod_{s=1}^{p} k_S$ ve minimum kümelenme sayısı $C_{\min} = \max\{k_S\}$ olarak belirlenir. Modeldeki geçerli olabilecek kümelenme sayısı için aralık,

$$\max\{k_S\} \le k \le \prod_{s=1}^{p} k_S \tag{2.2.16}$$

şeklinde elde edilir. Burada k modeldeki kümelenme merkez sayısını göstermektedir.

Sonlu karma modellerdeki $M_{Toplam} = 2^{C_{\max}} - 1$ model arasından her bir parçalanmaya en az bir kümelenmenin karşılık geldiği modellere uygun aday model denir. İki değişkenli modeldeki uygun aday model sayısı $i = 1,...,n \;\; j = 1,...,m$ olmak üzere,

$$f(n,m,k) = \sum_{i=0}^{n} (-1)^{i} \binom{n}{i} \sum_{j=0}^{m} (-1)^{j} \binom{m}{j} \binom{(n-i)(m-j)}{k} \tag{2.2.17}$$

denklemi ile elde edilir [53]. Burada $n \; ve \; m$ değişkenlerdeki parçalanma sayısı, k değişkenlerdeki parçalanmalara karşılık gelen modelin kümelenme merkez sayısını göstermektedir. Ayrıca değişkenlerdeki parçalanmalar eşit yani $n = m$ alındığında uygun aday model sayısını veren denklem, $i = 1,...,n$ olmak üzere,

$$f(n,k) = \sum_{i,j=0}^{n} (-1)^{i+j} \binom{n}{i} \binom{n}{j} \binom{(n-i)(n-j)}{k} \tag{2.2.18}$$

şeklinde elde edilir.

2.2.5. Modele Dayalı Kümeleme

Modele dayalı kümeleme sonlu karma modeller gibi verinin olasılık dağılımların bir karmasından elde edildiği düşünülür. Başka bir şekilde ifade edilecek olursa, her bir bileşen dağılımı verideki bir kümelenmeye karşılık gelen karma olasılık dağılımlarından üretildiği varsayılır. Modele dayalı kümeleme verideki karma kümelemenin ve küme

yapılarının modellenmesinde iki yaklaşım kullanılır. Bunlar sınıflandırma olabilirlik yaklaşımı ve karma olabilirlik yaklaşımıdır[72]. Sınıflandırma olabilirlik fonksiyonu,

$$L_C = \left(\theta_1, \theta_2, ..., \theta_g; \gamma_1, \gamma_2, ..., \gamma_n \middle| x\right) = \prod_{j=1}^{n} f_{\gamma j}\left(x_j \middle| \theta_{\gamma j}\right) \tag{2.2.19}$$

şeklinde ifade edilir. Burada $i = 1, 2, ..., g \ ve \ \ j = 1, 2, ..., n$ dir. Eğer x_j rastgele vektörü karma dağılımın *i.* bileşeninden elde edilmişse $\gamma_j = i$ olarak etiketlenir.

Karma olabilirlik (likelihood) yaklaşımında bir olasılık fonksiyonu olasılık ağırlıklarının bileşen fonksiyon toplamı olduğu kabul edilir. Kümelemede karma olabilirlik yaklaşımı kullanılırsa problem karma dağılım modelindeki parametrelerin tahminine dönüşür. En çok olabilirlik fonksiyonu,

$$L_M = \left(\theta_1, \theta_2, ..., \theta_g; \pi_1, \pi_2, ..., \pi_n \middle| x\right) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_i f_i\left(x_j \middle| \theta_i\right) \tag{2.2.20}$$

olarak elde edilir. π_i bir gözlem değerinin *i.* bileşende bulunma olasılığını göstermektedir.

Tanım 2.2.5.1. Olabilirlik Kestirim Fonksiyonu (Likelihood);

$X_1 = x_1, X_2 = x_2, ...X_n = x_n$ gözlemleri bir rasgele örneklem uzayını göstersin. X_i'nın olasılık yoğunluk fonksiyonu $f\left(x_i; \Psi\right) = \sum_{i=1}^{g} \pi_j f\left(x_i; \theta_j\right) \ \ i = 1, 2, ...n \ \ ve \ \ j = 1, 2, ...g$ olsun. $f\left(x_i; \theta_j\right)$ karma olasılık yoğunluk fonksiyonunun olabilirlik kestirim fonksiyonu,

$$L\left(\Psi\right) = \prod_{i=1}^{n} f\left(x_i; \Psi\right) = \prod_{i=1}^{n}\left[\sum_{j=1}^{g} f\left(x_i; \theta_j\right)\right] \ \ i = 1, 2, ...n \ \ ve \ \ j = 1, 2, ...g$$

şeklinde elde edilir. Logaritması alınmış olabilirlik kestirim fonksiyonu,

$$\log L\left(\Psi\right) = \sum_{j=1}^{n} \log\left(f\left(x_i; \Psi\right)\right) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \pi_i f_i\left(x_j; \theta_i\right)\right) \tag{2.2.21}$$

olarak elde edilir.

Tanım 2.2.5.2. Sınıflandırma en çok olabilirlik yaklaşımında; $i = 1, 2, ...g \ \ ve \ \ j = 1, 2, ...n$ olmak üzere z_{ij} , x_j gözlemi, karmanın *i.* bileşeninden elde ediliyorsa 1 diğer

bileşenlerinden elde ediliyorsa sıfır değerini alan elemanlardan oluşan g boyutlu z_j bileşen etiket vektörleri bilinmeyen parametre vektörleri olarak düşünülür. Örnekleme yapısına göre iki farklı sınıflandırma en çok olabilirlik kriteri bulunur. Bu örnekleme yapıları ayrık örnekleme ve karma örnekleme olarak adlandırılır.

Ayrık örneklemede, $x_1, x_2, ..., x_n$ örneklemi bağlantısız olarak n_i sayıda gözlemin *i.* bileşenden alınmasıyla oluşturulur. Burada n_i örneklemeye başlanmadan önce alınan sabit bir sayıdır. Bu gösterimde bileşen oranları π_i'ler belirgin şekilde görülmezler. Bu nedenden dolayı karma oranları π_i'nin eşit olduğu kabul edilir. Bu durumda kısıtlı sınıflandırma en çok olabilirlik kriteri,

$$L(\Psi) = \prod_{i=1}^{n} f\left(x_i; \Psi\right) \tag{2.2.22}$$

olarak elde edilir. Burada $i = 1, 2, ..., g \; ve \; j = 1, 2, ..., \mathrm{n}$ olmak üzere $P = (P_1, P_2, ..., P_n)$, z_j etiket vektörleri ile belirlenmiş $x_1, x_2, ..., x_n$ gözlemlerinin alt gruplara parçalanışlarıdır. $P_i = \left\{ x_j \middle| z_{ij} = 1 \right\}$ parçalanması *i.* parçalanışın gözlemlerinden elde edilir.

Karma örnekleme yapısı içerisinde $x_1, x_2, ..., x_n$ örneklemi, çok değişkenli karma dağılım fonksiyonu dağılımından, bileşenlerinden alınan gözlemlerin sayıları örneklem hacmi n ve olasılık parametreleri $\pi_1, \pi_2, ..., \pi_g$ olan çok terimli bir dağılıma sahip olacak şekilde rastgele elde edilirler. Bu durumda sınıflandırma en çok olabilirlik kriteri [11].

$$\log L(\Psi) = \sum_{j=1}^{n} \log\left(f\left(x_i; \Psi\right)\right) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \pi_i f_i\left(x_j; \theta_i\right)\right) \tag{2.2.23}$$

şeklinde elde edilir.

2.2.5.3. Karma Kümeleme Analizinde Model Seçim Kriterleri

Çok değişkenli karma dağılım modeline dayalı kümeleme analizinde model seçimi olabilirlik fonksiyonuna (likelihood) dayalı olarak olabilirlik fonksiyonunun logaritması alınmış hali olan log-likelihood fonksiyonu kullanılır. Logaritması alınmış olabilirlik,

$$\log L(\Psi) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \pi_i f_i\left(x_j; \theta_i\right)\right) \tag{2.2.24}$$

Şeklinde elde edilir. Logaritmik olabilirlik fonksiyonuna dayalı olarak yapılan model seçimi bilgi kriterleri olarak adlandırılır [11].

Tanım 2.2.5.4. Bayesci Bilgi Kriteri (BIC);

İntegrallenebilir olabilirlik fonksiyonu modellerin giriş olasılık değerlerine bağlı olduğundan, modellerde logaritmik olabilirlik fonksiyonunun değeri hesaplanabilir. Bu değere kısaca Bayesci bilgi kriteri (BIC) veya Schwarz bilgi kriteri denir. Sonlu karma dağılım modeli g sayıdaki bileşenin bilinmeyen parametrelerini $\Psi = \left\{ \pi_1, \pi_2, ..., \pi_g, \theta_1, \theta_2, ..., \theta_g \right\}$ vektörü ile belirler. Veri yapısına uyan en iyi modelin belirlenmesi için sonlu karma dağılım modelinde model seçimi metodu kullanılır. $i = 1, ..., g$ *ve* $j = 1, ..., n$ olmak üzere $f\left(x_j; \theta_i\right)$ karma modelin olasılık yoğunluk fonksiyonu, $L\left(\Psi\right)$ karma modelin yoğunluk fonksiyonundan elde edilmiş olabilirlik fonksiyonu olsun. Bayesci bilgi kriteri,

$$BIC = -2 \ln L\left(\Psi\right) + d \log n \tag{2.2.25}$$

olarak elde edilir. Burada n yoğunluk fonksiyonundaki gözlem sayısını, d modeldeki bağımsız parametre sayısını K bileşen sayısı ve p değişken sayısına bağlı olarak

$$d = \left(K - 1\right) + \left(Kp\right) + \left(Kp \frac{\left(p+1\right)}{2} \right) \tag{2.2.26}$$

eşitliğinden hesaplanır [5]. Buradaki $d \log n$ terimi bilgi kriterindeki yanlılık düzeltmesinde kullanılan ceza terimi olarak kullanılmıştır. Ceza terimi gürültü verisinden kaynaklanan fazla kümelenmeyi önlemek amacıyla olabilirlik fonksiyonundan çıkarılır [73].

Tanım 2.2.5.5. Akaike Bilgi Kriteri (AIC); Sonlu karma dağılım modellerinde $L\left(\Psi\right)$ olabilirlik fonksiyonuna dayalı bilgi kriteri Akaike [74] tarafından ileri sürülmüştür. Karma modelin parametrelerinin tahmini için olabilirlik fonksiyonu kullanılarak hesaplanan bilgi kriteri,

$$AIC = -2 \ln L\left(\Psi\right) + 2d \tag{2.2.27}$$

Akaike bilgi kriteri (AIC) olarak elde edilir. Burada d modeldeki bağımsız parametre sayısını göstermektedir. Model seçiminde AIC modeldeki bileşen sayısını olduğundan fazla göstermeye meyilli olmasına rağmen en yaygın kullanılan bilgi kriteridir.

3. BÖLÜM

NORMAL KARMA MODELLERİN MODELE DAYALI KÜMELENMESİ İÇİN ADAY MODELLER ARASINDA EN İYİ MODEL SEÇİMİNE DAYALI YENİ BİR METOD

3.1. Normal Karma Dağılımlarda Modele Dayalı Kümeleme İçin Genetik Algoritma

Çok değişkenli verinin modele dayalı kümelenmesi için verideki değişkenler ve bu değişkenlerin alt gruplarının oluşturduğu kümelenme merkezlerinin yapısı ve sayısı istatiksel ve matematiksel metotlarla incelenmiştir. Veri setindeki değişkenlerin parçalanması (segmentasyonu) ve bu parçalanmaların karşılık geldiği kümelenmelerin ortaya çıkarılması için bir algoritma geliştirilmiştir. Normal karma modellerin modele dayalı kümelemesi için geliştirilen genetik algoritmanın adımları kısaca şu şekilde gösterilebilir,

1) Çok değişkenli heterojen verideki her bir değişkenin verinin yapısına en uygun alt gruplara ayrılması veya başka bir deyişle değişkenlerin parçalanması

2) Çok değişkenli veride k-ortalamalar algoritması ile değişkenlerdeki parçalanmalara düşen gözlemlerin belirlenmesi

3) Çok değişkenli normal karma modellerde değişkenlerdeki parçalanmalara dayalı kümelenme merkezi sayısının ve yapısının belirlenmesi

4) Çok değişkenli normal karma modellerde toplam model sayısı ve modellerin yapısının belirlenmesi

5) Toplam modeller arasından varsayımlara uyan uygun aday model sayısının belirlenmesi ve her bir modelin temsili gösteriminin elde edilmesi

6) Uygun aday modellerdeki parametrelerin normal karma dağılımlardan elde edilmiş verinin gözlem değerlerine dayalı olarak tahmin edilmesi (EM algoritması kullanmadan parametrelerin giriş değerleri kullanılır)

7) Her bir Uygun aday modeli meydana getiren küme merkezlerini oluşturan parametreler kullanılarak modelin log-likelihood fonksiyonu, AIC ve BIC değerlerinin elde edilmesi

8) Log-likelihood, AIC ve BIC değerlerine göre optimizasyonla uygun aday modeller arasından en iyi model seçilmesi

3.2. Çok Değişkenli Üç Farklı Veri Setinde Heterojen Değişkenlerin Normal Karma Modeller ile Modele Dayalı Kümelenmesi

Bu bölümde çok değişkenli normal karma modellerin modele dayalı kümelenmesi için iki değişkenli, gözlemleri normal dağılımlardan gelen üç farklı sentetik veri seti Minitab17 Paket Programının Deneme Sürümü (http://it.minitab.com/en-us/products/minitab/free-trial.aspx) kullanılarak elde edilmiştir. Gözlem değerleri her bir değişken için ayrı ayrı ortalama vektörleri ve standart sapmaları girilerek elde edilmiştir. Değişkenlerin her birisindeki gözlem sayısı eşit olup, her bir değişkendeki bölünmelere düşen gözlem sayıları farklıdır. Simülasyonla üretilen veri seti üzerindeki prototip çalışma için üretilen farklı üç veri setinde iki değişken ve her bir değişkende eşit gözlem sayısı bulunmaktadır. Yani, X_1 ve X_2 değişkenlerinin gözlem sayıları sırasıyla $n_1 = 300$ *ve* $n_2 = 300$ olarak elde edilmiştir. Üç farklı sentetik veri setindeki değişkenlerde bulunan heterojen yapılar araştırılmış, değişkenlerdeki parçalanmalara dayalı kümelenme merkezlerinin terleri ve sayısı belirlenmiştir. Karma modeller oluşturulmuş ve aralarından varsayıma uyan aday modellerin sayıları ve karma modelleri oluşturulmuştur. Aday modeller arasından veri yapısına en uygun karma modelin seçiminde istatistiksel bilgi kriterleri kullanılmıştır. Simülasyon ile oluşturulan birinci, ikinci ve üçüncü veri setlerinde değişkenlerdeki parçalanmalar ve buna dayalı karma modeller arasından en iyi model sırasıyla iki, üç ve dört kümelenme merkezli modeller olarak elde edilmiştir.

3.2.1. Heterojen Değişken İçeren Çok Değişkenli Veride Değişkenlerin Grafiksel Yöntemlere ve Tek Değişkenli Normal Dağılımlara Dayalı Parçalanması

Çok değişkenli veride değişkenlerdeki parçalanmaların belirlenmesi için çeşitli yöntemlerden faydalanılabilir. Her bir değişkendeki parçalanmalar değişkenlerin alt

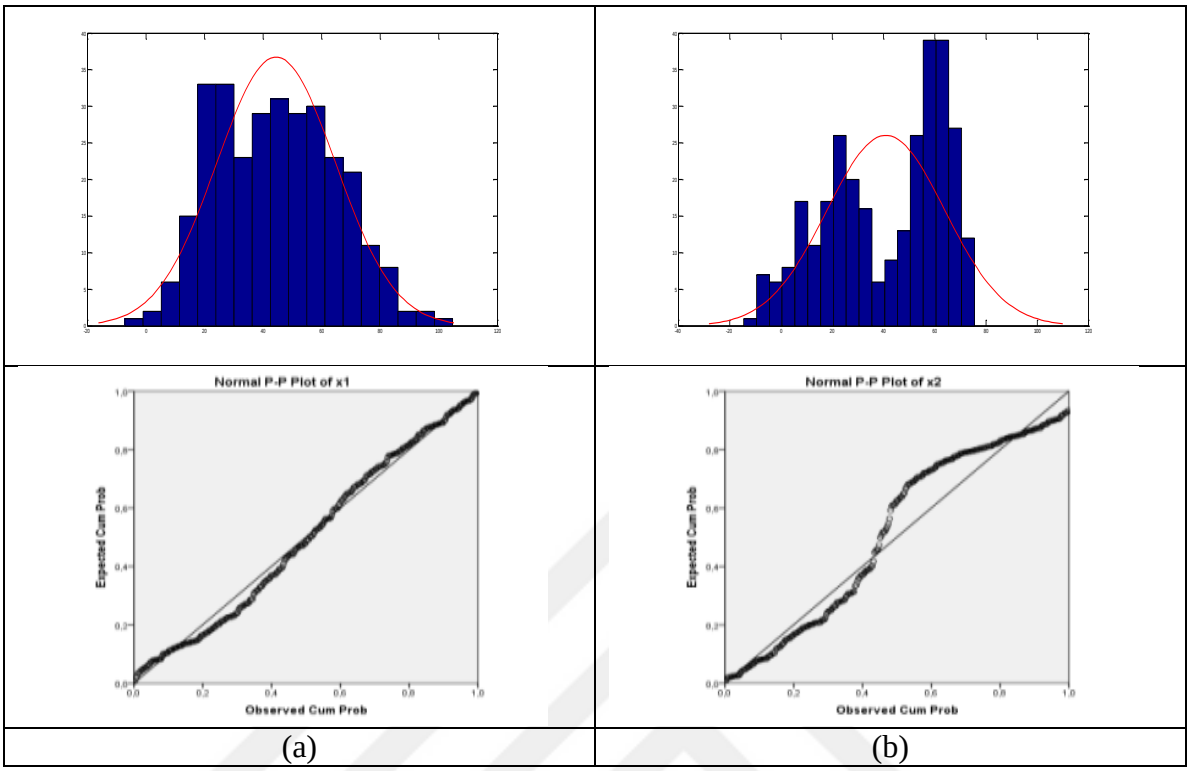
grupları olarak adlandırılır. Değişkenlerdeki parçalanmaları belirlemek ve bu parçalanmaları anlamlı alt gruplara dönüştürmek için istatistiksel grafiksel yöntemler ve normal karma modeller kullanılır. Her bir değişkendeki anlamlı parçalanma sayısı histogram grafiğindeki tepelenme sayısı ve P-P grafiğinde gözlemlerin normal doğrusu ile kesişme sayısına göre tahmin edilebilir.

Her bir değişkendeki anlamlı alt grup sayısını belirlemek amacıyla tek değişkenli normal karma modelin log-likelihood, AIC ve BIC değerlerine bakılır.

3.2.1.1. Değişkenlerdeki Parçalanmaların Grafiksel Yöntemlere Dayalı Belirlenmesi

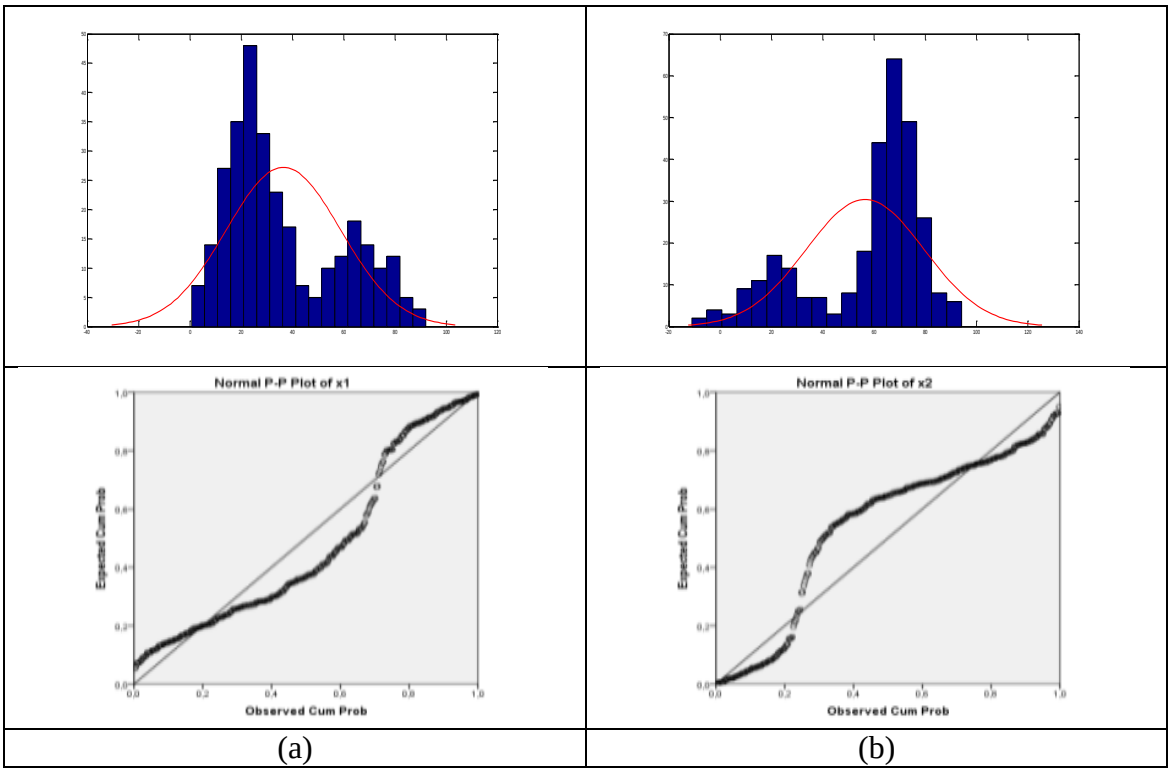
Bu çalışmada literatürdekilerden farklı olarak verideki değişkenlerin heterojenliği test edilir. Değişkenlerde heterojenlik varsa değişken içerdiği alt grup sayısı kadar parçalanır. Verideki değişkenlerdeki parçalanmalar, verideki kümelenme merkez sayılarını belirler. $s = 1,..., p$ olmak üzere verideki her bir X_s değişkeninin heterojenliği incelensin. Verideki değişkenler homojen ya da heterojen yapıda olabilir. Heterojen verideki değişkenlerin yapısı verideki kümelenmeyi belirlemektedir [41]. p-boyutlu veride X_s değişkeninin yapısına göre parçalanmalarının sayısı $k_s \geq 1$ olsun. $k_s = 1$ durumunun olması değişkenin homojen yapıda ve $k_s > 1$ olması değişkenin heterojen yapıda olması anlamına gelir. p-boyutlu heterojen verideki $s.$ değişken ve $j.$ gözlem arasındaki ilişki x_{sj} olarak gösterilebilir. Burada x_{sj}, p -boyutlu heterojen verideki $s.$ değişkeninin $j.$ gözlemini ifade eder. Veri setinde bulunan her bir değişkendeki alt gruplarını (parçalanma sayıları) belirlemede en uygun parçalanma sayısını bulmak için histogram ve P-P grafiklerinden tahmini parçalanma sayısı hakkında bilgi edinilebilir. Üç farklı sentetik veri setinde iki değişkenli birinci veri seti, iki değişkenli ikinci veri seti ve iki değişkenli üçüncü veri setindeki her bir değişkenin homojenlik veya heterojenlik yapısının ortaya çıkarılmasında istatistiksel grafiksel yöntemler kullanılır [41].

İki değişkenli birinci, ikinci ve üçüncü veri setlerinde X_1 ve X_2 değişkenlerinde meydana gelen en uygun parçalanmayı belirlemek amacıyla her bir değişken ayrı ayrı ele alınır. Değişkenlerin histogram ve P-P grafikleri elde edilerek parçalanma sayısı hakkında tahmini bir sayı elde belirlenebilir.



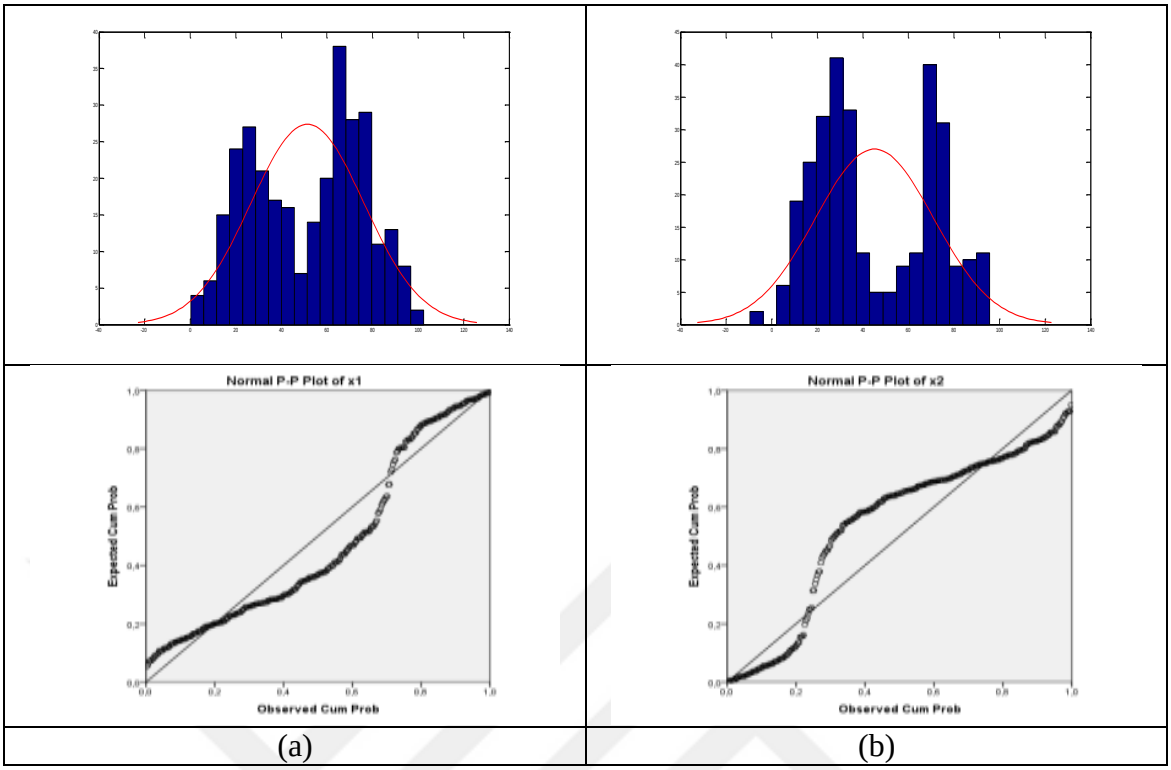
(a) (b)

Şekil 3.2.1. İki değişkenli birinci veri setindeki (a) X_1 değişkeni (b) X_2 değişkeni için histogram ve P-P grafikleri.



(a) (b)

Şekil 3.2.2. İki değişkenli ikinci veri setindeki (a) X_1 değişkeni (b) X_2 değişkeni için histogram ve P-P grafikleri.



Şekil 3.2.3. İki değişkenli üçüncü veri setinde (a) X_1 değişkeni (b) X_2 değişkeni için histogram grafikleri.

İki değişkenli üç farklı veri setinde, her bir veri setindeki X_1 ve X_2 değişkenleri için uygun parçalanmanın belirlenmesinde histogram grafiğindeki tepelenme (mod) sayısı ve P-P grafiğinin normal doğrusu veya $y = x$ doğrusu ile verideki gözlemlerin oluşturduğu eğrinin kesişim sayısına bakılarak değişkenlerdeki parçalanma sayıları belirlenir. **Şekil 3.2.1 - Şekil 3.2.3** teki grafiklere bakılarak birinci veri setinde $k_1 = 2$ ve $k_2 = 2$, ikinci veri setinde $k_1 = 2$ ve $k_2 = 2$ ve üçüncü veri setinde $k_1 = 2$ ve $k_2 = 2$ olarak tahmin edilmiştir.

3.2.1.2. Değişkenlerdeki Parçalanmaların Tek Değişkenli Normal Karmalara Dayalı Belirlenmesi

İki değişkenli veride X_1 ve X_2 değişkenlerinin her birinde meydana gelen optimum parçalanmayı belirlemek amacıyla her bir değişken ayrı ayrı ele alınır. İki değişkenli veride her bir heterojen değişkendeki bölünme ya da parçalanma sayısının belirlenmesinde grafiksel yöntemlerle birlikte tek değişkenli karma modeller kullanılabilir.

Sentetik veri setlerindeki her bir değişkene tek değişkenli karma normal model oluşturarak her bir değişken için oluşturulan tek değişkenli karma normal modelde bileşen sayısını belirlenir. Tek değişkenli normal dağılımların karması

$$f(x;\theta)=\sum_{i=1}^{g}\pi_i f_i(x;\mu_i,\sigma_i) \tag{3.2.1}$$

olarak gösterilir. Burada $f(x)$ tek değişkenli normal dağılımın karmalarının olasılık yoğunluk fonksiyonu, g karma dağılımdaki bileşen sayısı, π_i karma olasılık ağırlıklarını göstermektedir. Tek değişkenli normal karma dağılımların olasılık yoğunluk fonksiyonları $f_i(x;\mu_i,\sigma_i)$, ortalama vektörü μ ve standart sapması σ olmak üzere,

$$f\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{\frac{-1}{2}\frac{\left(x-\mu\right)^2}{\sigma^2}\right\}\ x\in\mathbb{R} \tag{3.2.2}$$

şeklinde ifade edilir. Değişkenlerdeki parçalanma sayısını elde etmek için normal karma dağılımların log-likelihhod, AIC ve BIC gibi istatistiksel bilgi kriterleri kullanılmaktadır. Oluşturulan her bir model için log-likelihood, AIC ve BIC değerleri değişkenlerdeki olasılık ağırlıkları π, ortalama vektörü μ ve varyansı σ^2 değerlerinden hesaplanır. Her bir heterojen değişkendeki parçalanmayı model tabanlı belirlemek amacıyla her bir k değerleri için hesaplanan log-likelihood, AIC ve BIC değeri karşılaştırılır ve optimum k değeri belirlenir. Log-likelihood fonksiyon değerinin maksimum, AIC ve BIC değerlerinin minimum olduğu modelde parçalanma sayısı optimum olarak bulunur. Hesaplamalarda Matlab programının istatistik paketi kütüphanesinde yer alan *gmdistribution.fit(data,k)* hazır fonksiyonu kullanılmıştır. İki değişkenli veride her bir heterojen değişken için hangi parçalanmanın optimum olduğunu belirlemek amacıyla $k=1,2,3,4$ bileşen sayıları girilerek oluşturulan karma normal modeller için log-likelihood, AIC ve BIC değerleri hesaplanmış ve hesaplama sonuçlarına göre en uygun parçalanma sayısı belirlenmiştir.

İki değişkenli üç farklı veri setinde her bir değişkendeki uygun parçalanma sayısını belirlemek için tek değişkenli normal karma dağılım modeli uygulanmıştır.

Değişkenlerdeki k bileşen sayısı aralıktaki en küçük değerden başlayıp üst sınıra kadar verilerek en uygun sayı elde edilmiştir.

Tablo 3.2.1. Birinci veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için tek değişkenli karma normal modellerdeki π, μ ve σ^2 parametre değerlerinden elde edilen log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-1327.4	2658.9	2666.3
k=2	**-1318.0**	**2646.0**	**2664.5**
k=3	-1317.1	2650.2	2679.8

Tablo 3.2.2. Birinci veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için tek değişkenli karma normal modellerdeki π, μ ve σ^2 parametre değerlerinden elde edilen log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-1365.8	2735.6	2743.0
k=2	**-1297.1**	**2604.2**	**2622.8**
k=3	-1297.1	2610.2	2639.9

Tablo 3.2.3. İkinci veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için tek değişkenli karma normal modellerdeki π, μ ve σ^2 parametre değerlerinden elde edilen log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-1327.4	2658.9	2666.3
k=2	**-1318.0**	**2646.0**	**2664.5**
k=3	-1317.1	2650.2	2679.8

Tablo 3.2.4. İkinci veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için tek değişkenli karma normal modellerdeki π, μ ve σ^2 parametre değerlerinden elde edilen log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-1365.8	2735.6	2743.0
k=2	**-1297.1**	**2604.2**	**2622.8**
k=3	-1297.1	2610.2	2639.9

Tablo 3.2.5. Üçüncü veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için tek değişkenli karma normal modellerdeki π, μ ve σ^2 parametre değerlerinden elde edilen log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-1327.4	2658.9	2666.3
k=2	**-1318.0**	**2646.0**	**2664.5**
k=3	-1317.1	2650.2	2679.8

Tablo 3.2.6. Üçüncü veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için tek değişkenli karma normal modellerdeki π , μ ve σ^2 parametre değerlerinden elde edilen log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-1365.8	2735.6	2743.0
k=2	**-1297.1**	**2604.2**	**2622.8**
k=3	-1297.1	2610.2	2639.9

İki değişkenli birinci, ikinci ve üçüncü sentetik veri setlerine tek değişkenli normal karma dağılım modeli uygulanarak elde edilen bilgi kriterleri tablolarına göre üç farklı veri setindeki her bir değişkende heterojen yapıda uygun parçalanma sayısı $k = 2$ olarak elde edilmiştir. Böylece üç farklı sentetik veri setindeki her bir değişkende uygun parçalanma sayısı grafiksel yöntemler ve tek değişkenli karma modeller ile kümelemedeki algoritmalar kullanılarak elde edilmiştir.

3.2.2. K-Ortalamalar Algoritması İle Parçalanmalara Düşen Gözlemlerin Belirlenmesi

İki değişkenli ve her değişkenin ikiye parçalandığı veri setlerindeki değişkenlerin parçalanmalarına düşen gözlemlerin belirlenmesi için k-ortalamalar algoritması kullanılmaktadır. Başlangıçta sentetik veri setlerindeki her bir değişkendeki anlamlı alt grup sayısı veya uygun parçalanma sayısı $k = 2$ belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık hesaplaması

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 \qquad (3.2.3)$$

şeklinde hesaplanmaktadır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir.

İki değişkenli üç farklı sentetik veri setindeki her değişkenin ikiye bölündüğü veride k-ortalamalar algoritması uygulandığında iki ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri

atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum en uygun parçalanma sayısını vermektedir. K-ortalamalar algoritması kullanılarak verideki her bir değişken ikiye parçalanmış ve her bir gözlemin düştüğü veri grubu (alt grup) belirlenmiştir. X_1 *ve* X_2 değişkenlerindeki parçalanmalara karşılık gelen alt gruplar, ait oldukları değişkene göre isimlendirilmiştir. X_1 değişkenindeki iki parçalanmaya karşılık gelen alt gruplar X_{11} *ve* X_{12}, X_2 değişkenindeki iki parçalanmaya karşılık gelen alt gruplar ise X_{21} *ve* X_{22} olarak adlandırılmıştır.

İki değişkenli normal karma modellerdeki üç farklı veri setinde bulunan her bir değişkenin parçalanmaları ve bu parçalanmalara düşen gözlem sayıları **Tablo 3.2.7-9** da verilmiştir.

Tablo 3.2.7. İki değişkenli birinci sentetik veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişkenler	X_1		X_2	
Değişkendeki parçalanmalar	X_{11}	X_{12}	X_{21}	X_{22}
Gözlem sayısı	$n_{11}=157$	$n_{12}=143$	$n_{21}=168$	$n_{22}=132$
Toplam	$n_1=300$		$n_2=300$	

Tablo 3.2.8. İki değişkenli ikinci sentetik veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişkenler	X_1		X_2	
Değişkendeki parçalanmalar	X_{11}	X_{12}	X_{21}	X_{22}
Gözlem sayısı	$n_{11}=89$	$n_{12}=211$	$n_{21}=74$	$n_{22}=226$
Toplam	$n_1=300$		$n_2=300$	

Tablo 3.2.9. İki değişkenli üçüncü sentetik veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişkenler	X_1		X_2	
Değişkendeki parçalanmalar	X_{11}	X_{12}	X_{21}	X_{22}
Gözlem sayısı	$n_{11}=133$	$n_{12}=167$	$n_{21}=126$	$n_{22}=174$
Toplam	$n_1=300$		$n_2=300$	

3.2.3. Normal Karma Modellerde Değişkenlerdeki Parçalanmalara Dayalı Kümelenme Merkezi Sayısı ve Yapısının Belirlenmesi

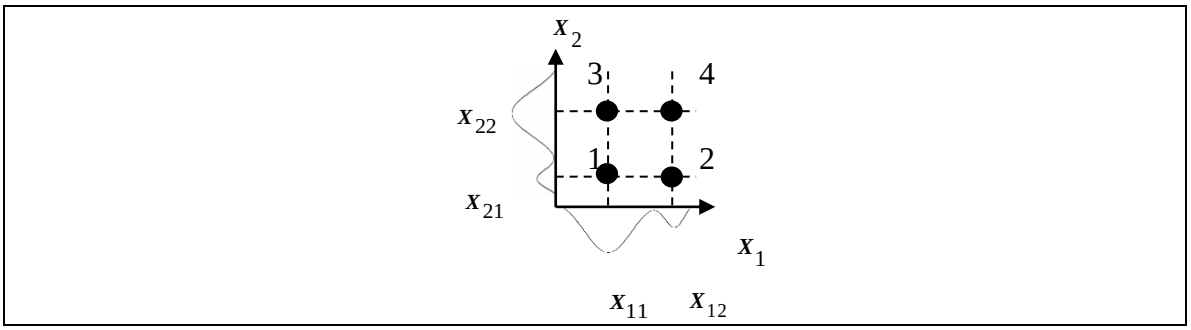
İki değişkenli ve her değişkenin ikiye bölündüğü üç farklı veri setinde modeldeki kümelenme merkez sayıları değişkenlerdeki parçalanmalara bağlı olarak hesaplanır. Değişkenlerdeki parçalanmaların sayısı kümelenme merkez sayısını, parçalanmalara düşen gözlem değerleri sırasıyla, kümelenme merkezlerinin yerini, şeklini ve büyüklüğünü belirlemektedir [49]. Modeldeki değişkenlerin parçalanmalarının oluşturduğu maksimum ve minimum kümelenme merkezlerinin sayısı C_{max} ve C_{min} $s = 1,..., p$ olmak üzere k_s değerlerine bağlı olarak

$$C_{max} = \prod_{s=1}^{p} k_s \tag{3.2.4}$$

ve

$$C_{min} = \max\{k_s\} \tag{3.2.5}$$

olarak elde edilir. (3.2.4) ve (3.2.5) eşitliklerinde verideki parçalanmalara karşılık gelen en çok kümelenme merkezi sayısı $C_{max} = k_1.k_2 = 2.2 = 4$ ve en az kümelenme merkez sayısı $C_{min} = \max\{k_1, k_2\} = \max\{2,2\} = 2$ olarak elde edilir.



Şekil 3.2.4. İki değişkenli ve her değişkenin ikiye parçalandığı normal karma modelde X_1 değişkeninin X_{11}, X_{12} ve X_2 değişkeninin X_{21}, X_{22} parçalanmalarının oluşturduğu kümelenme merkezleri ve bu merkezlerin numaraları.

Bir numaralı kümelenme merkezi: X_1 değişkeninin X_{11} ve X_2 değişkeninin X_{21} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_1

ve varyans-kovaryans matrisi Σ_1, sırasıyla $\mu_1 = \begin{bmatrix} \mu_{11} \\ \mu_{21} \end{bmatrix}$ ve $\Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{21} \\ \rho_1\sigma_{21}\sigma_{11} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_1 = Corr(X_{11}, X_{21})$ olmak üzere X_{11} ve X_{21} parçalanmaları arasındaki korelasyon katsayısını göstermektedir. $\rho_1 = \frac{\sigma_{1121}}{\sigma_{11}\sigma_{21}}$ olarak tanımlanır. Bu kümelenme merkezindeki veriler, $N(x; \mu_1, \Sigma_1)$ normal dağılımına sahip olsun.

İki numaralı kümelenme merkezi: X_1 değişkeninin X_{12} ve X_2 değişkeninin X_{21} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_2 ve varyans-kovaryans matrisi Σ_2, sırasıyla $\mu_2 = \begin{bmatrix} \mu_{12} \\ \mu_{21} \end{bmatrix}$ ve $\Sigma_2 = \begin{bmatrix} \sigma_{12}^2 & \rho_2\sigma_{12}\sigma_{21} \\ \rho_2\sigma_{21}\sigma_{12} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_2 = Corr(X_{12}, X_{21})$ olmak üzere X_{12} ve X_{21} parçalanmaları arasındaki korelasyon katsayısını göstermektedir. $\rho_2 = \frac{\sigma_{1221}}{\sigma_{12}\sigma_{21}}$ olarak tanımlanır. Bu kümelenme merkezindeki veriler, $N(x; \mu_2, \Sigma_2)$ normal dağılımına sahip olsun.

Üç numaralı kümelenme merkezi: X_1 değişkeninin X_{11} ve X_2 değişkeninin X_{22} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_3 ve varyans-kovaryans matrisi Σ_3, sırasıyla $\mu_3 = \begin{bmatrix} \mu_{11} \\ \mu_{22} \end{bmatrix}$ ve $\Sigma_3 = \begin{bmatrix} \sigma_{11}^2 & \rho_3\sigma_{11}\sigma_{22} \\ \rho_3\sigma_{22}\sigma_{11} & \sigma_{22}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_3 = Corr(X_{11}, X_{22})$ olmak üzere X_{11} ve X_{22} parçalanmaları arasındaki korelasyon katsayısını göstermektedir. $\rho_3 = \frac{\sigma_{1122}}{\sigma_{11}\sigma_{22}}$ olarak tanımlanır. Bu kümelenme merkezindeki veriler, $N(x; \mu_3, \Sigma_3)$ normal dağılımına sahip olsun.

Dört numaralı kümelenme merkezi: X_1 değişkeninin X_{12} ve X_2 değişkeninin X_{22} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_4 ve varyans-kovaryans matrisi Σ_4, sırasıyla $\mu_4 = \begin{bmatrix} \mu_{12} \\ \mu_{22} \end{bmatrix}$ ve $\Sigma_4 = \begin{bmatrix} \sigma_{12}^2 & \rho_4\sigma_{12}\sigma_{22} \\ \rho_4\sigma_{22}\sigma_{12} & \sigma_{22}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_4 = Corr(X_{12}, X_{22})$ olmak üzere X_{12} ve X_{22}

parçalanmaları arasındaki korelasyon katsayısını göstermektedir. $\rho_4 = \frac{\sigma_{1222}}{\sigma_{12}\sigma_{22}}$ olarak tanımlanır. Bu kümelenme merkezindeki veriler, $N(x;\mu_4,\Sigma_4)$ normal dağılımına sahip olsun.

3.2.4. Normal Karma Modellerde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

İki değişkenli ve her değişkenin ikiye parçalanması durumunda kümelenme merkezleri için M_{toplam} ile gösterilen oluşabilecek tüm modellerin sayısı değişkenlerdeki parçalanmalara bağlı olarak elde edilir. İki değişkenli normal karma modellerin toplam sayısı

$$M_{toplam} = 2^{C\max} - 1 = 2^4 - 1 = 15 \tag{3.2.6}$$

olarak elde edilebilir. Burada çıkarılan 1 model kümelenmenin bulunmadığı boş modeldir. Oluşabilecek toplam modeller içerisinde C_{max} eşitliğinden elde edilen maksimum kümelenme sayısına göre tek kümelenme merkezli, iki kümelenme merkezli, üç kümelenme merkezli ve dört kümelenme merkezli modeller bulunmaktadır. Toplam model sayısı, $\binom{4}{0} = 1$ boş model, $\binom{4}{1} = 4$ tek kümelenme merkezli model, $\binom{4}{2} = 6$ iki kümelenme merkezli model, $\binom{4}{3} = 4$ üç kümelenme merkezli model ve $\binom{4}{4} = 1$ dört kümelenme merkezli dolu model olarak elde edilip boş model sayısı çıkarılaraktan hesaplanabilir.

3.2.5. Normal Karma Modellerde Uygun Aday Model Sayısının Hesaplanması

İki değişkenli normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin sayısı ve yine parçalanmalara bağlı olarak toplam model sayısı elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli model veya uygun modellerin sayısı araştırılmıştır. Bu geçerli modeller Şekil 3.2.3.1 de gösterilen merkezler üzerinden kısaca her satır ve her sütunda en az bir kümelenme merkezi bulunacak

varsayımı ile anlatılabilir. Varsayıma uyan modellere Uygun Aday Model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanabilir.

Teorem 3.2.1. Değişkendeki g parçalanmaya karşılık gelen k adet kümelenme veya grup merkezlerinin sayısı

$$S(g,k)=\sum_{i=0}^{k}(-1)^{i}\binom{k}{i}(k-i)^{g} \tag{3.2.7}$$

denkleminden elde edilir. Burada kümelenme merkez sayısının parçalanma sayısından az veya eşit olması yani $k \le g$ durumlarının dışındaki durumlar hesaplanır.

Her satır ve sütunda en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı her bir merkezin bulunduğu satır ve sütundaki konumuna göre kombinasyon hesabı ile elde edilebilir. Hesaplama sonucunda iki değişkenli veride her değişkenin ikiye bölündüğü durumda veride olabilecek minimum kümelenme merkez sayısı 2, maksimum kümelenme merkez sayısı 4 için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları **Tablo 3.2.10** da verilmiştir.

Tablo 3.2.10. İki değişkenli ve her değişkenin ikiye bölündüğü normal karma modellerde uygun aday modeller için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları.

Merkez Sayısı	**Merkezlerin Konumu**	**Modellerin Sayısı İçin Bağıntı**	**Model Sayısı**
İki merkezli	1 1 şeklinde sütunlara dağılan	$\binom{2}{1}\binom{1}{1}=2!=2$	2
Üç merkezli	2 1 şeklinde sütunlara dağılan	$\binom{2}{1}2!=2.2=4$	4
Dört merkezli	2 2 şeklinde sütunlara dağılan	$\binom{4}{4}=1$	1
Toplam model sayısı		$2^{2.2}-1=15$	
Toplam Uygun model sayısı		$0+0+2+4+1=7$	

Tek merkezli toplam model sayısı $\binom{4}{1}=\frac{4!}{(4-1)!.1!}=4$ adettir. Ancak modellerin hiçbiri varsayıma uymadığından uygun aday model yoktur.

İki merkezli toplam model sayısı $\binom{4}{2}=\frac{4!}{(4-2)!.2!}=6$ adettir. Ancak varsayıma uyan uygun aday model sayısı yukarıda hesaplandığı gibi 2 adettir.

Üç merkezli toplam model sayısı $\binom{4}{3}=\frac{4!}{(4-3)!.3!}=4$ adettir. Üç kümelenme merkezli modellerin hepsi varsayıma uyan uygun aday modeldir.

Dört merkezli toplam model sayısı $\binom{4}{4}=\frac{4!}{(4-4)!.4!}=1$ adettir. Dört kümelenme merkezli model varsayıma uyan uygun aday modeldir. Toplamda 2+4+1=7 uygun aday model bulunmaktadır.

Varsayım altındaki uygun aday model sayısının hesaplanmasındaki kombinatorik problem bir matris yapısına karşılık gelmektedir. Varsayımdaki değişkenler ve değişkenlerin parçalanmaları matrisin satır ve sütunlarına karşılık gelmektedir. **Tablo 3.2.10** da kombinasyonla elde edilen uygun aday modellerin sayısı başka bir şekilde, n_{ij} elemanı *i.* satır ve *j.* sütundaki kümelenme merkezini göstermek üzere, *nxm* tipindeki sıfırlardan oluşan bir matriste her satır ve her sütunda en az bir kümelenmenin olduğu başka bir ifade ile sıfırdan farklı en az bir elemanın bulunduğu matrislerin sayısı Cheballah vd. [53] tarafından önerilen

$$f(n,k)=\sum_{i=1}^{n}(-1)^{i}\binom{n}{i}\sum_{j=1}^{m}(-1)^{j}\binom{m}{j}\binom{(n-i)(m-j)}{k} \tag{3.2.8}$$

eşitliği ile elde edilebilir. Burada n ve m satır ve sütun sayısını veya değişkenlerdeki parçalanma sayısını göstermek üzere $n=m$ alındığında

$$f(n,k)=\sum_{i,j=1}^{n}(-1)^{i+j}\binom{n}{i}\binom{n}{j}\binom{(n-i)(n-j)}{k} \tag{3.2.9}$$

eşitliği elde edilebilir. Burada $i, j=1,...,4$ değişkenlerin parçalanmalarındaki kümelenme sayısı ve k modeldeki kümelenme sayısını göstermektedir.

Tablo 3.2.10 da ve (3.2.8) eşitliğinden elde edilen yedi uygun duruma karşılık gelen kümelenme yapılarının modelleri **Tablo 3.2.11** de verilmiştir.

Tablo 3.2.11. İki değişkenli ve her değişkenin ikiye bölündüğü normal karma modeller arasından varsayıma uyan iki, üç ve dört kümelenme merkezli toplam model sayısı, uygun aday modellerin sayısı ve karşılık gelen kümelenme yapılarının dizi gösterimi.

Kümelenme Merkez Sayısı	Durum Sayısı	Uygun Durum Sayısı	Uygun Durumlara Karşılık Gelen Modellerin Dizi Gösterimleri
			1234
Bir Kümelenme Merkezli Modeller	4	-	-
İki Kümelenme Merkezli Modeller	6	2	1001
			0110
Üç Kümelenme Merkezli Modeller	4	4	1110
			1101
			1011
			0111
Dört Kümelenme Merkezli Modeller	1	1	1111

3.2.6. Normal Karma Modellerde Uygun Aday Modellerin Oluşturulması ve Parametre Tahminleri

İki değişkenli normal karma dağılım modelleri için bileşen ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir. EM algoritması gibi adımsal hesaplama yapan ve hesaplama karmaşıklığı yüksek algoritmaların kullanılmaması, işlem karmaşıklığını en aza indirmekte ve hızını yükseltmektedir. Karma modeldeki her bir merkezin olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisi modeldeki olasılık yoğunluk fonksiyonunu elde etmek için kullanılır. **Tablo 3.2.7** de uygun aday modellerin temsili gösterimleri modeldeki merkezlerin numaralarına göre dizi gösterimi şeklinde verilmiştir. Modelin temsili gösteriminde kullanılan "0" ve "1" elemanları (digit) modeldeki karşılık gelen merkezde yığılmanın (kümelenmenin) olup olmadığını göstermektedir [75]. Eğer uygun aday modelde herhangi bir merkezde kümelenme varsa "1" yoksa "0" ile gösterilmiştir. Veri setindeki değişkenlerin her birisi matris gösteriminde satır ve sütunlara karşılık geldiğinden, iki boyutlu düzlemde uygun durumlar ve bu durumların merkezlerinin yerlerinin anlatıldığı modeller elde edilmiştir. İki değişkenli ve her değişkenin ikiye parçalandığı veri setinde oluşabilecek uygun aday modellerin dizi gösteriminden elde edilen grafiksel gösterimler **Şekil 3.2.8** de verilmiştir.

Bir numaralı kümelenme merkezinin ortalama vektörü μ_1 ve varyans-kovaryans matrisi Σ_1 sırasıyla $\boldsymbol{\mu}_1 = \begin{bmatrix} \mu 11 \\ \mu 21 \end{bmatrix}$ ve $\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_1 \sigma_{11} \sigma_{21} \\ \rho_1 \sigma_{21} \sigma_{11} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_1 = \boldsymbol{Corr(X}_{11}, \boldsymbol{X}_{21})$ olmak üzere $\rho_1 = \dfrac{\sigma_{1121}}{\sigma_{11}\sigma_{21}}$ pearson korelasyon katsayısını göstermektedir. Bu durumda birinci kümelenme merkezindeki veriler, $N(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ çok değişkenli normal dağılıma sahip olsun. **Şekil 3.2.7**'deki iki numaralı kümelenme merkezi X_1 değişkeninin X_{12} ve X_2 değişkeninin X_{21} parçalarının oluşturduğu kümelenme merkezdir.

İki numaralı kümelenme merkezinin ortalama vektörü μ_2 ve varyans-kovaryans matrisi Σ_2 sırasıyla $\boldsymbol{\mu}_2 = \begin{bmatrix} \mu 12 \\ \mu 21 \end{bmatrix}$ ve $\boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma_{12}^2 & \rho_2 \sigma_{12} \sigma_{21} \\ \rho_2 \sigma_{21} \sigma_{12} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın.

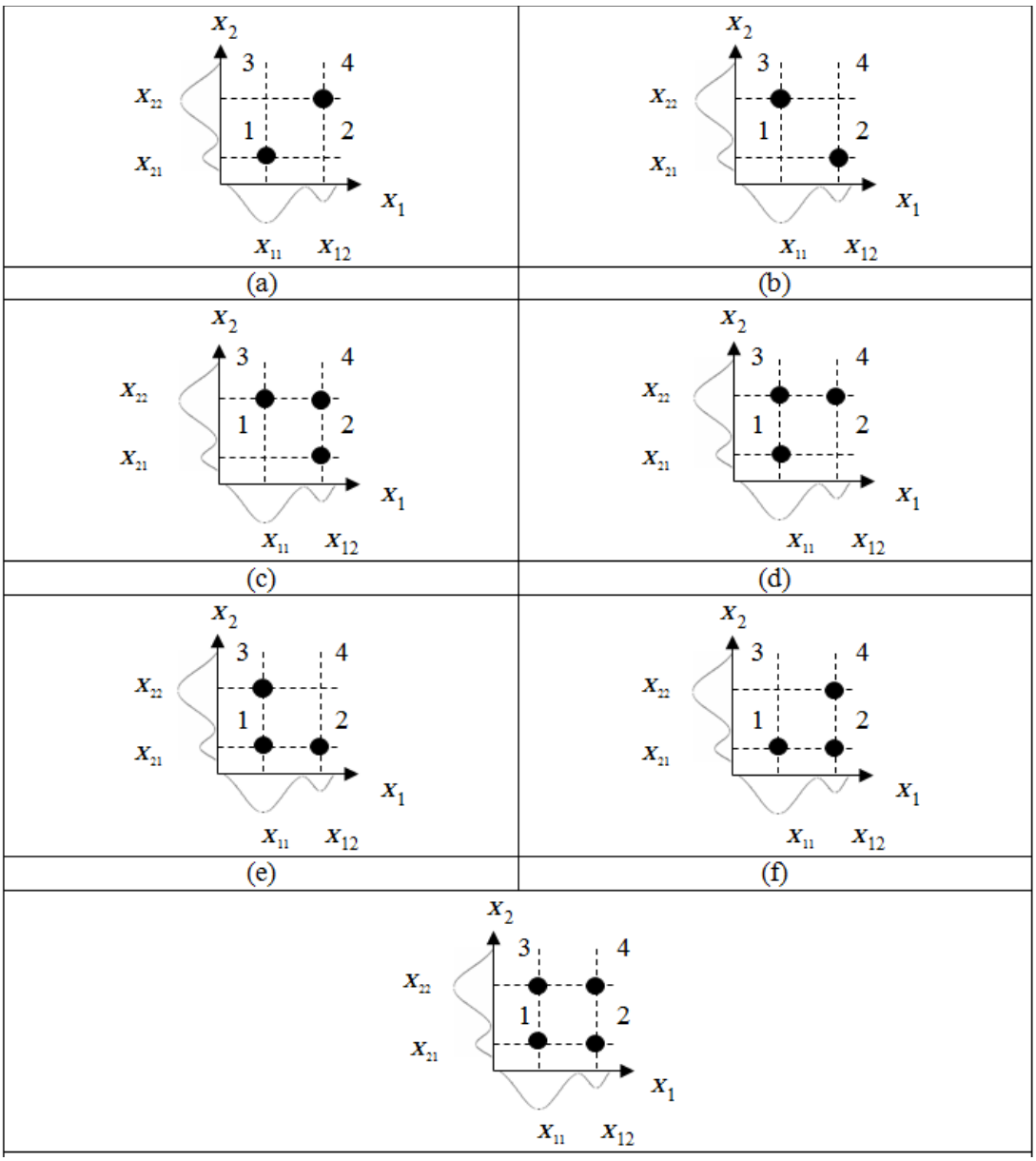
Burada $\rho_2 = \boldsymbol{Corr(X}_{12}, \boldsymbol{X}_{21})$ olmak üzere $\rho_2 = \dfrac{\sigma_{1221}}{\sigma_{12}\sigma_{21}}$ pearson korelasyon katsayısını göstermektedir. Bu durumda ikinci kümelenme merkezindeki veriler, $N(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ çok değişkenli normal dağılıma sahip olsun. **Şekil 3.2.7**'deki üç numaralı kümelenme merkezi X_1 değişkeninin X_{11} ve X_2 değişkeninin X_{22} parçalarının oluşturduğu kümelenme merkezdir.

Üç numaralı kümelenme merkezinin ortalama vektörü μ_3 ve varyans-kovaryans matrisi Σ_3 sırasıyla $\boldsymbol{\mu}_3 = \begin{bmatrix} \mu 11 \\ \mu 22 \end{bmatrix}$ ve $\boldsymbol{\Sigma}_3 = \begin{bmatrix} \sigma_{11}^2 & \rho_3 \sigma_{11} \sigma_{22} \\ \rho_3 \sigma_{22} \sigma_{11} & \sigma_{22}^2 \end{bmatrix}$ şeklinde tanımlansın.

Burada $\rho_3 = \boldsymbol{Corr(X}_{11}, \boldsymbol{X}_{22})$ olmak üzere $\rho_3 = \dfrac{\sigma_{1122}}{\sigma_{11}\sigma_{22}}$ pearson korelasyon katsayısını göstermektedir. Bu durumda üçüncü kümelenme merkezindeki veriler, $N(\mathbf{x}; \boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ çok değişkenli normal dağılıma sahip olsun. **Şekil 3.2.7**'deki dört numaralı kümelenme merkezi X_1 değişkeninin X_{12} ve X_2 değişkeninin X_{22} parçalarının oluşturduğu kümelenme merkezdir.

Dört numaralı kümelenme merkezinin ortalama vektörü μ_4 ve varyans-kovaryans matrisi Σ_4 sırasıyla $\boldsymbol{\mu}_4 = \begin{bmatrix} \mu 12 \\ \mu 22 \end{bmatrix}$ ve $\boldsymbol{\Sigma}_4 = \begin{bmatrix} \sigma_{12}^2 & \rho_4 \sigma_{12} \sigma_{22} \\ \rho_4 \sigma_{22} \sigma_{12} & \sigma_{22}^2 \end{bmatrix}$ şeklinde tanımlansın.

Burada $\rho_4 = \boldsymbol{Corr}(\boldsymbol{X}_{12}, \boldsymbol{X}_{22})$ olmak üzere $\rho_4 = \dfrac{\sigma_{1222}}{\sigma_{12}\sigma_{22}}$ pearson korelasyon katsayısını göstermektedir. Bu durumda dördüncü kümelenme merkezindeki veriler, $N(\mathbf{x}; \boldsymbol{\mu}_4, \boldsymbol{\Sigma}_4)$ çok değişkenli normal dağılıma sahip olsun.



Şekil 3.2.5. İki değişkenli ve her değişkenin ikiye parçalandığı veri setinde (a) ve (b) iki merkezli, (c), (d),(e) ve (f) üç merkezli ve (g) dört kümelenme merkezli uygun aday modellerin düzlemsel grafiklerini göstermektedir.

3.2.6.1. Normal Karma Modeller ve Parametre Tahminleri

Simülasyon ile oluşturulan üç sentetik veri setindeki iki değişkenin her birinin ikiye bölünmesi durumunda oluşacak kümelenme merkezleri **Şekil 3.2.7** de gösterilmiştir. **Tablo 3.2.11** de her bir veri setindeki parçalanmalara karşılık gelen ve varsayımlara uyan uygun aday modellerin dizi gösterimi verilmiştir.

Toplam 7 uygun aday model içerisinden iki bileşenli normal karma modele karşılık gelen modeller $u = 1,2$ ve $i = 1,2$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{2} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.2.10}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{2} \pi i}$, $x = 1,2$ ve $y = 1,2$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir [76]. İki kümeleme merkezli çok değişkenli normal dağılımların karma dağılım modelleri: $0 < \pi_i < 1$ ve $\sum_{i=1}^{2} \pi_i = 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$$

altı karma model arasından iki uygun aday model vardır.

Toplam 7 uygun aday model içerisinden üç bileşenli normal karma modele karşılık gelen modeller $u = 3,...,6$ ve $i = 1,2,3$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{3} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.2.11}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{3} \pi i}$, $x = 1,2$ ve $y = 1,2$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir. Üç kümeleme merkezli çok değişkenli normal dağılımların karma dağılım modelleri: $0 < \pi_i < 1$ ve $\sum_{i=1}^{3} \pi_i = 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

dört uygun aday model vardır.

Toplam 7 uygun aday model içerisinden üç bileşenli normal karma modele karşılık gelen modeller $u = 7$ ve $i = 1,...,4$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{4} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \qquad (3.2.12)$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{4} \pi i}$, $x = 1,2$ ve $y = 1,2$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi

$$\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)}\sigma_{1x}^{(u)}\sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)}\sigma_{2y}^{(u)}\sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$$ olarak ifade edilir. Dört kümeleme merkezli çok değişkenli normal dağılımların karma dağılım modeli: $0 < \pi_i < 1$ ve $\sum_{i=1}^{4} \pi_i = 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

bir uygun aday model vardır.

Böylece uygun aday modellerdeki $i = 1,...,g$ olmak üzere π_i, μ_i ve Σ_i parametreleri tahmin edilmelidir.

Tablo 3.2.11 de dizi gösterimi verilen: İki bileşenli iki karma model ve (3.2.10) denklemindeki $u = 1,2$ için karma modellerin yapısına uyan modeller **Şekil 3.2.8** de gösterilmiştir. İki bileşenli modellerin olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{2} n_i}$, $x = 1,2$ ve $y = 1,2$ olmak üzere, ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)}s_{1x}^{(u)}s_{2y}^{(u)} \\ r_{2y,1x}^{(u)}s_{2y}^{(u)}s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}$, $i = 1,2$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir.

Üç bileşenli dört karma model ve (3.2.11) denklemindeki $u = 3,...,6$ için karma modellerin yapısıma uyan modeller **Şekil 3.2.8** de gösterilmiştir. Üç bileşenli modellerin olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{3} n_i}$, $x = 1,2$ ve $y = 1,2$ olmak üzere, ortalama

vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri

$\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}$, $i = 1,2,3$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir.

Dört bileşenli karma model ve (3.2.12) denklemindeki $u = 7$ için karma modellerin yapısıma uyan modeller **Şekil 3.2.8** de gösterilmiştir. Dört bileşenli modelin olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{4} n_i}$, $x = 1,2$ ve $y = 1,2$ olmak üzere, ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}$,

$i = 1,...,4$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir.

İki değişkenli birinci, ikinci ve üçüncü veri setlerindeki normal karma modellerin kümelenmelerinden oluşan uygun aday modellerin parametre değerleri **Tablo 3.2.11-Tablo 3.2.13** de gösterilmiştir.

Tablo 3.2.12. İki değişkenli birinci veri setindeki yedi uygun aday normal karma model için parametre tahminleri.

Model No	π_i	μ_i	Σ_i	ρ
1. Model	$\pi_1 = 0,5416$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 28,2664 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 109,277 & 23,113 \\ 23,113 & 67,624 \end{bmatrix}$	$\rho_1 = 0,2704$
	$\pi_4 = 0,4583$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 62,2378 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 134,152 & -13,605 \\ -13,605 & 146,631 \end{bmatrix}$	$\rho_4 = -0,0954$
2. Model	$\pi_2 = 0,5183$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 62,2378 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 134,152 & 18,965 \\ 18,965 & 67,624 \end{bmatrix}$	$\rho_2 = 0,1958$
	$\pi_3 = 0,4816$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 28,2664 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 109,277 & 1,616 \\ 1,616 & 146,631 \end{bmatrix}$	$\rho_3 = 0,0129$
3. Model	$\pi_1 = 0,3514$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 28,2664 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 109,277 & 23,113 \\ 23,113 & 67,624 \end{bmatrix}$	$\rho_1 = 0,2704$
	$\pi_2 = 0,3362$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 62,2378 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 134,152 & 18,965 \\ 18,965 & 67,624 \end{bmatrix}$	$\rho_2 = 0,1958$
	$\pi_3 = 0,3124$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 28,2664 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 109,277 & 1,616 \\ 1,616 & 146,631 \end{bmatrix}$	$\rho_3 = 0,0129$
4. Model	$\pi_1 = 0,3567$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 28,2664 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 109,277 & 23,113 \\ 23,113 & 67,624 \end{bmatrix}$	$\rho_1 = 0,2704$
	$\pi_2 = 0,3414$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 62,2378 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 134,152 & 18,965 \\ 18,965 & 67,624 \end{bmatrix}$	$\rho_2 = 0,1958$
	$\pi_4 = 0,3019$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 62,2378 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 134,152 & -13,605 \\ -13,605 & 146,631 \end{bmatrix}$	$\rho_4 = -0,0954$
5. Model	$\pi_1 = 0,3656$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 28,2664 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 109,277 & 23,113 \\ 23,113 & 67,624 \end{bmatrix}$	$\rho_1 = 0,2704$
	$\pi_3 = 0,3251$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 28,2664 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 109,277 & 1,616 \\ 1,616 & 146,631 \end{bmatrix}$	$\rho_3 = 0,0129$
	$\pi_4 = 0,3093$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 62,2378 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 134,152 & -13,605 \\ -13,605 & 146,631 \end{bmatrix}$	$\rho_4 = -0,0954$
6. Model	$\pi_2 = 0,3554$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 62,2378 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 134,152 & 18,965 \\ 18,965 & 67,624 \end{bmatrix}$	$\rho_2 = 0,1958$
	$\pi_3 = 0,3303$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 28,2664 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 109,277 & 1,616 \\ 1,616 & 146,631 \end{bmatrix}$	$\rho_3 = 0,0129$
	$\pi_4 = 0,3143$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 62,2378 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 134,152 & -13,605 \\ -13,605 & 146,631 \end{bmatrix}$	$\rho_4 = -0,0954$
7. Model	$\pi_1 = 0,2708$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 28,2664 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 109,277 & 23,113 \\ 23,113 & 67,624 \end{bmatrix}$	$\rho_1 = 0,2704$
	$\pi_2 = 0,2592$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 62,2378 \\ 59,2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 134,152 & 18,965 \\ 18,965 & 67,624 \end{bmatrix}$	$\rho_2 = 0,1958$
	$\pi_3 = 0,2408$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 28,2664 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 109,277 & 1,616 \\ 1,616 & 146,631 \end{bmatrix}$	$\rho_3 = 0,0129$
	$\pi_4 = 0,2292$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 62,2378 \\ 17,7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 134,152 & -13,605 \\ -13,605 & 146,631 \end{bmatrix}$	$\rho_4 = -0,0954$

Tablo 3.2.13. İki değişkenli ikinci veri setindeki yedi uygun aday normal karma model için parametre tahminleri.

Model No	π_i	μ_i	Σ_i	ρ
1. Model	$\pi_1 = 0,2717$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67,3698 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108,335 & -2,386 \\ -2,386 & 136,873 \end{bmatrix}$	$\rho_1 = -0,0201$
	$\pi_4 = 0,7283$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23,5331 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92,963 & 6,761 \\ 6,761 & 80,099 \end{bmatrix}$	$\rho_4 = 0,0779$
2. Model	$\pi_2 = 0,4750$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23,5331 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92,963 & 18,051 \\ 18,051 & 136,873 \end{bmatrix}$	$\rho_2 = 0,1752$
	$\pi_3 = 0,5250$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67,3698 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108,335 & 15,393 \\ 15,393 & 80,099 \end{bmatrix}$	$\rho_3 = 0,1806$
3. Model	$\pi_1 = 0,2717$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67,3698 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108,335 & -2,386 \\ -2,386 & 136,873 \end{bmatrix}$	$\rho_1 = -0,0201$
	$\pi_2 = 0,4750$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23,5331 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92,963 & 18,051 \\ 18,051 & 136,873 \end{bmatrix}$	$\rho_2 = 0,1752$
	$\pi_3 = 0,5250$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67,3698 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108,335 & 15,393 \\ 15,393 & 80,099 \end{bmatrix}$	$\rho_3 = 0,1806$
4. Model	$\pi_1 = 0,2717$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67,3698 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108,335 & -2,386 \\ -2,386 & 136,873 \end{bmatrix}$	$\rho_1 = -0,0201$
	$\pi_2 = 0,4750$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23,5331 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92,963 & 18,051 \\ 18,051 & 136,873 \end{bmatrix}$	$\rho_2 = 0,1752$
	$\pi_4 = 0,7283$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23,5331 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92,963 & 6,761 \\ 6,761 & 80,099 \end{bmatrix}$	$\rho_4 = 0,0779$
5. Model	$\pi_1 = 0,2717$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67,3698 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108,335 & -2,386 \\ -2,386 & 136,873 \end{bmatrix}$	$\rho_1 = -0,0201$
	$\pi_3 = 0,5250$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67,3698 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108,335 & 15,393 \\ 15,393 & 80,099 \end{bmatrix}$	$\rho_3 = 0,1806$
	$\pi_4 = 0,7283$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23,5331 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92,963 & 6,761 \\ 6,761 & 80,099 \end{bmatrix}$	$\rho_4 = 0,0779$
6. Model	$\pi_2 = 0,4750$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23,5331 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92,963 & 18,051 \\ 18,051 & 136,873 \end{bmatrix}$	$\rho_2 = 0,1752$
	$\pi_3 = 0,5250$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67,3698 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108,335 & 15,393 \\ 15,393 & 80,099 \end{bmatrix}$	$\rho_3 = 0,1806$
	$\pi_4 = 0,7283$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23,5331 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92,963 & 6,761 \\ 6,761 & 80,099 \end{bmatrix}$	$\rho_4 = 0,0779$
7. Model	$\pi_1 = 0,2717$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67,3698 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108,335 & -2,386 \\ -2,386 & 136,873 \end{bmatrix}$	$\rho_1 = -0,0201$
	$\pi_2 = 0,4750$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23,5331 \\ 20,1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92,963 & 18,051 \\ 18,051 & 136,873 \end{bmatrix}$	$\rho_2 = 0,1752$
	$\pi_3 = 0,5250$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67,3698 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108,335 & 15,393 \\ 15,393 & 80,099 \end{bmatrix}$	$\rho_3 = 0,1806$
	$\pi_4 = 0,7283$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23,5331 \\ 68,5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92,963 & 6,761 \\ 6,761 & 80,099 \end{bmatrix}$	$\rho_4 = 0,0779$

Tablo 3.2.14. İki değişkenli üçüncü veri setindeki yedi uygun aday normal karma model için parametre tahminleri.

Model No	π_i	μ_i	Σ_i	ρ
1. Model	$\pi_1 = 0,5416$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 26,6768 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 114,677 & -14,283 \\ -14,283 & 100,252 \end{bmatrix}$	$\rho_1 = -0,1319$
	$\pi_4 = 0,4583$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 71,2124 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 132,427 & 25,461 \\ 25,461 & 104,498 \end{bmatrix}$	$\rho_4 = 0,2146$
2. Model	$\pi_2 = 0,5183$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 71,2124 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 132,427 & 1,616 \\ 1,616 & 100,252 \end{bmatrix}$	$\rho_2 = 0,0142$
	$\pi_3 = 0,4816$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 26,6768 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 114,677 & 22,735 \\ 22,735 & 104,498 \end{bmatrix}$	$\rho_3 = 0,2041$
3. Model	$\pi_1 = 0,5416$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 26,6768 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 114,677 & -14,283 \\ -14,283 & 100,252 \end{bmatrix}$	$\rho_1 = -0,1319$
	$\pi_2 = 0,5183$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 71,2124 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 132,427 & 1,616 \\ 1,616 & 100,252 \end{bmatrix}$	$\rho_2 = 0,0142$
	$\pi_3 = 0,4816$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 26,6768 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 114,677 & 22,735 \\ 22,735 & 104,498 \end{bmatrix}$	$\rho_3 = 0,2041$
4. Model	$\pi_1 = 0,5416$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 26,6768 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 114,677 & -14,283 \\ -14,283 & 100,252 \end{bmatrix}$	$\rho_1 = -0,1319$
	$\pi_2 = 0,5183$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 71,2124 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 132,427 & 1,616 \\ 1,616 & 100,252 \end{bmatrix}$	$\rho_2 = 0,0142$
	$\pi_4 = 0,4583$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 71,2124 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 132,427 & 25,461 \\ 25,461 & 104,498 \end{bmatrix}$	$\rho_4 = 0,2146$
5. Model	$\pi_1 = 0,5416$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 26,6768 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 114,677 & -14,283 \\ -14,283 & 100,252 \end{bmatrix}$	$\rho_1 = -0,1319$
	$\pi_3 = 0,4816$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 26,6768 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 114,677 & 22,735 \\ 22,735 & 104,498 \end{bmatrix}$	$\rho_3 = 0,2041$
	$\pi_4 = 0,4583$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 71,2124 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 132,427 & 25,461 \\ 25,461 & 104,498 \end{bmatrix}$	$\rho_4 = 0,2146$
6. Model	$\pi_2 = 0,5183$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 71,2124 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 132,427 & 1,616 \\ 1,616 & 100,252 \end{bmatrix}$	$\rho_2 = 0,0142$
	$\pi_3 = 0,4816$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 26,6768 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 114,677 & 22,735 \\ 22,735 & 104,498 \end{bmatrix}$	$\rho_3 = 0,2041$
	$\pi_4 = 0,4583$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 71,2124 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 132,427 & 25,461 \\ 25,461 & 104,498 \end{bmatrix}$	$\rho_4 = 0,2146$
7. Model	$\pi_1 = 0,5416$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 26,6768 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 114,677 & -14,283 \\ -14,283 & 100,252 \end{bmatrix}$	$\rho_1 = -0,1319$
	$\pi_2 = 0,5183$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 71,2124 \\ 73,0996 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 132,427 & 1,616 \\ 1,616 & 100,252 \end{bmatrix}$	$\rho_2 = 0,0142$
	$\pi_3 = 0,4816$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 26,6768 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 114,677 & 22,735 \\ 22,735 & 104,498 \end{bmatrix}$	$\rho_3 = 0,2041$
	$\pi_4 = 0,4583$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 71,2124 \\ 24,8488 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 132,427 & 25,461 \\ 25,461 & 104,498 \end{bmatrix}$	$\rho_4 = 0,2146$

3.2.7. Normal Karma Modellerin Log-likelihood Fonksiyonu, AIC ve BIC Değerleri

Modele dayalı kümeleme yapmak için iki değişkenli ve her değişkenin ikiye parçalandığı birinci, ikinci ve üçüncü veri setlerindeki uygun aday modellerin normal karma modelleri belirlenmiştir. Normal karma modeller arasından en iyi modelin seçimi için her bir uygun aday modelin çok değişkenli normal karma dağılımlardan log-likelihood, AIC ve BIC değerleri elde edilmiştir. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu, $i = 1, 2, ...n$ *ve* $j = 1, 2, ...g$ olmak üzere $f\left(x_i; \theta_j\right)$ karma olasılık yoğunluk fonksiyonunun likelihood fonksiyonu

$$L(\Psi) = \prod_{i=1}^{n} f\left(x_i; \Psi\right) = \prod_{i=1}^{n} \left[\sum_{j=1}^{g} f\left(x_i; \theta_j\right) \right] \tag{3.2.13}$$

şeklinde elde edilir. Logaritması alınmış likelihood fonksiyonu

$$\log L(\Psi) = \sum_{j=1}^{n} \log\left(f\left(x_i; \Psi\right)\right) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{g} \pi_i f_i\left(x_j; \theta_i\right)\right) \tag{3.2.14}$$

olarak elde edilir [73].

Çok değişkenli normal karma modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kriteri (BIC)

$$BIC = -2 \ln L(\Psi) + d \log n \tag{3.2.15}$$

olarak elde edilir. Burada n olasılık yoğunluk fonksiyonundaki gözlem sayısını, d modeldeki bağımsız parametre sayısını K bileşen veya grup sayısı ve p değişken sayısına bağlı olarak

$$d = (K - 1) + (Kp) + \left(Kp \frac{(p+1)}{2} \right) \tag{3.2.16}$$

elde edilir. Aynı şekilde Akaike bilgi kriteri de (AIC) log-likelihood fonksiyonuna bağlı olarak

$$AIC = -2 \ln L(\Psi) + 2d \tag{3.2.17}$$

şeklinde elde edilir [74].

Tablo 3.2.15. İki değişkenli normal karma dağılımların modele dayalı kümelenmesi için iki değişkenli sentetik veri setlerinden log-l, AIC, BIC değerlerinin elde edildiği matlab kodu.

```
a=load('p12x11.txt');b=load('p12x12.txt');c=load('p12x21.txt');d=load('p12x22.txt');
x(1,:)=load('p12x1.txt');x(2,:)=load('p12x2.txt');
pi1=(size(a)+size(c))/(size(a) +size(c) +size(b) +size(d));
pi4=(size(b)+size(d))/(size(a) +size(c) +size(b) +size(d));
mu(1,:)=mean(a);mu(2,:)=mean(c);
sigma =[ 109.257 1.616;1.616 146.631];

for i=1:length(x(1,:))
pdf1(i)=(1/(2*pi))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-mu)'...
*inv(sigma)*(x(:,i)-mu));
end
mu(1,:)=mean(b);mu(2,:)=mean(d);
sigma =[ 134.152 18.965;18.965 67.624];


for i=1:length(x(1,:))
pdf4(i)=(1/(2*pi))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-mu)'
*inv(sigma)*(x(:,i)-mu));
end
oyf1=pi1*pdf1+pi4*pdf4;
% Log-likelihood fonksiyonu
lnl= -11*log(length(oyf1))+ sum(log(pi1*pdf1(:)+pi4*pdf4(:)));
lnl
%AIC bilgi kriteri
aic=-2*lnl+2*11;
aic
%BIC bilgi kriteri
bic=-2*lnl+11*log(length(lnl));
bic
```

Tablo 3.2.16. İki değişkenli normal karma dağılımların modele dayalı kümelenmesi için iki değişkenli birinci sentetik veri setinde log-l, AIC, BIC değerleri ile karma modeldeki bağımsız değişken sayısı.

Model no:	$\ln L(\Psi)$	$AIC = -2\ln L(\Psi) + 2d$	$BIC = -2\ln L(\Psi) + d\log n$	d
1.	**-2726.3**	**5474.5**	**5452.5**	**11**
2.	-3481	6984.1	6962.1	11
3.	-3118.5	6270.9	6236.3	17
4.	-2746.9	5527.9	5493.9	17
5.	-2772.9	5579.8	5545.8	17
6.	-3154.6	6343.2	6309.2	17
7.	-2758.8	5563.6	5517.6	23

Tablo 3.2.17. İki değişkenli normal karma dağılımların modele dayalı kümelenmesi için iki değişkenli ikinci sentetik veri setinde log-l, AIC, BIC değerleri ile karma modeldeki bağımsız değişken sayısı.

Model no:	$\ln L(\Psi)$	$AIC = -2\ln L(\Psi) + 2d$	$BIC = -2\ln L(\Psi) + d\log n$	d
1.	-3675.3	7372.7	7350.7	11
2.	-3460.4	6942.8	6920.8	11
3.	-3146.3	6326.6	6292.6	17
4.	-3767	7568	7534	17
5.	**-2641.8**	**5317.7**	**5283.7**	**17**
6.	-3054.3	6142.5	6108.5	17
7.	-2705.2	5456.4	5410.4	23

Tablo 3.2.18. İki değişkenli normal karma dağılımların modele dayalı kümelenmesi için iki değişkenli üçüncü sentetik veri setinde log-l, AIC, BIC değerleri ile karma modeldeki bağımsız değişken sayısı.

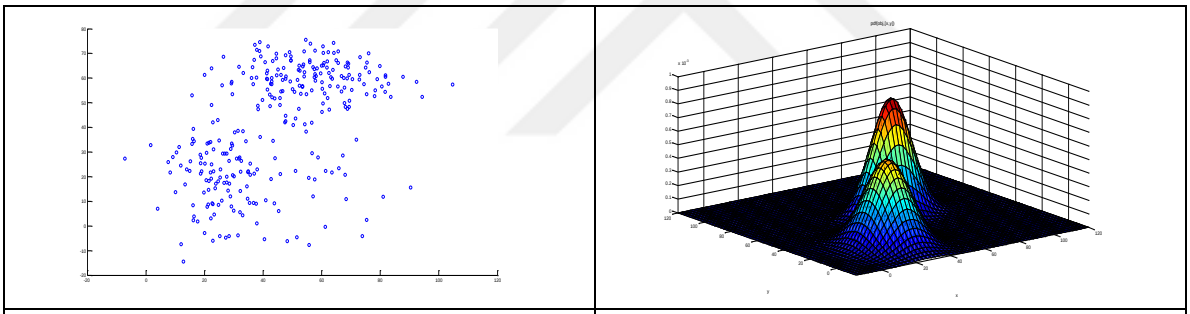
Model no:	$\ln L(\Psi)$	$AIC = -2\ln L(\Psi) + 2d$	$BIC = -2\ln L(\Psi) + d\log n$	d
1.	-3364.5	6751	6729	11
2.	-3826.7	7675.3	7653.3	11
3.	-3351.4	6736.8	6702.8	17
4.	-2957	5948	5914	17
5.	-3246	6526	6492	17
6.	-3314.2	6662.3	6628.3	17
7.	**-2803.9**	**5653.7**	**5607.7**	**23**

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin log-l, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. İki değişkenli birinci, ikinci ve üçüncü sentetik veri setlerinde normal karma modellerden uygun aday modellerin log-l, AIC ve BIC değerleri hesaplanmış ve **Tablo 3.2.15-17** te verilmiştir. Uygun aday modeller arasından modele dayalı kümelemede en iyi model likelihood değeri en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir.

Tablo 3.2.19. İki değişkenli iki kümelenme merkezli normal karma model için sentetik veri setlerinden elde edilen en iyi modelin MATLAB yüzey grafiği kodu.

```
% Merkezleri oluşturan veriler
a=load('p12x11.txt');
b=load('p12x12.txt');
c=load('p12x21.txt');
d=load('p12x22.txt');
%Merkezlerin yerlerini (konumlarını) veren parametre değerleri
MU = [mean(a) mean(c);mean(b) mean(d)];
% Merkezlerin şekillerini veren parametre değerleri
SIGMA = cat(3,[ 109.277 23.113;23.113 67.624],[ 134.152 -13.605;-13.605 146.631]);
%Merkezlerin normal dağılımdan yüzey grafikleri
p = ones(1,2)/2;
obj = gmdistribution(MU,SIGMA,p);
figure;
ezsurf(@(x,y)pdf(obj,[x y]),[-15 120],[-15 120]);
```

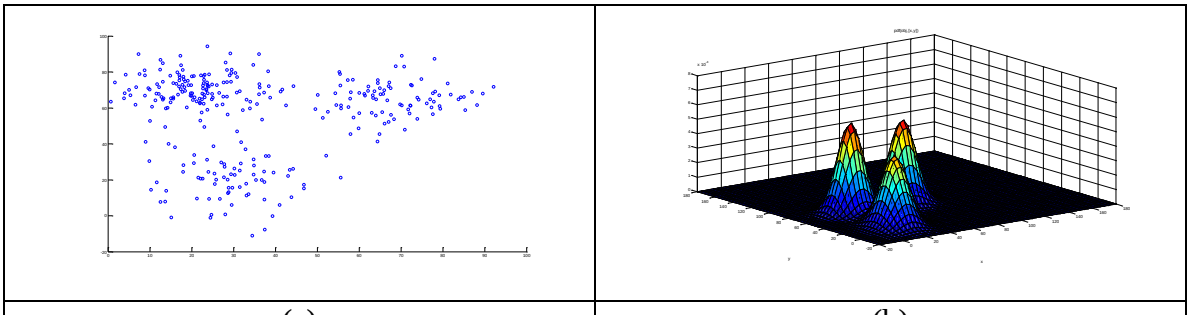
İki değişkenli ve her değişkenin ikiye parçalandığı birinci veri setinde en iyi model iki bileşenli birinci model olarak belirlenmiştir.



(a) (b)

Şekil 3.2.6. İki değişkenli ve her değişkenin ikiye bölündüğü birinci veri setindeki modeller arasından en iyi modelin (a) saçılım grafiği (b) yüzey grafiği.

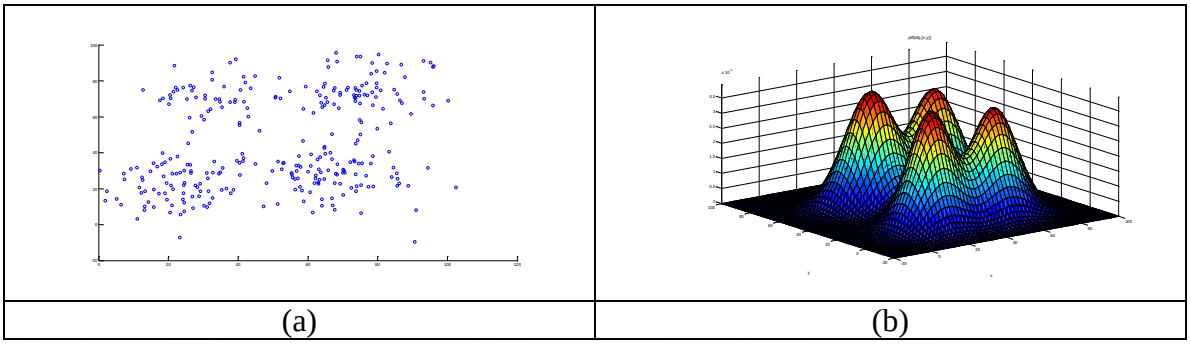
İki değişkenli ve her değişkenin ikiye parçalandığı ikinci veri setinde en iyi model üç bileşenli binci model olarak belirlenmiştir.



(a) (b)

Şekil 3.2.7. İki değişkenli ve her değişkenin ikiye bölündüğü ikinci veri setindeki modeller arasından en iyi modelin (a) saçılım grafiği (b) yüzey grafiği.

İki değişkenli ve her değişkenin ikiye parçalandığı üçüncü veri setinde en iyi model dört bileşenli yedinci model olarak belirlenmiştir.



(a) (b)

Şekil 3.2.8. İki değişkenli ve her değişkenin ikiye bölündüğü üçüncü veri setindeki modeller arasından en iyi modelin (a) saçılım grafiği (b) yüzey grafiği.

3.3. İki Değişkenli Normal Karma Modellerde Heterojen Değişkenlerdeki Bölütlenmelerden (Segmentasyondan) Kaynaklanan Verinin Modele Dayalı Kümelenmesi

Bu kısımda çok değişkenli normal karma modellerde değişkenlerdeki bölütlenmeden (segmentasyondan) kaynaklanan modele dayalı kümeleme için yöntem, yeni bir veri seti üzerinde anlatılmıştır. Çok değişkenli heterojen verideki kümelenmeyi normal dağılımların karmalarının bir kümesini kullanarak modele dayalı biçimde belirlemek amacıyla geliştirilen en iyi model seçimi yöntemi açıklanmıştır. Çok değişkenli heterojen verinin her bir değişkenindeki bölünme sayısı grafiksel yöntemler ve tek değişkenli normal karma modellerin bilgi kriterleri yardımıyla saptanıp, verideki kümelenme yapısı ve kümelenme merkezlerinin sayısı hesaplanmıştır.

Değişkenlerdeki parçalanmalara bağlı olarak kümelenme merkez sayıları ve bu merkezlerin oryantasyonundan oluşan modellerin sayısı belirlenmiştir. Elde edilen toplam model sayıları içerisinde bazı modeller varsayıma uymadığından çıkarılmıştır. Böylece uygun aday modellerin sayısı belirlenmiş, modellerin temsili (dizi) gösterimleri elde edilmiştir. Temsili gösterimlerden yararlanarak normal karma modeller oluşturulmuş ve parametreleri tahmin edilmiştir. Modellerin parametre tahminlerinden normal dağılımların olasılık yoğunluk fonksiyonu elde edilmiş, yoğunluk fonksiyonuna bağlı

olarak her bir modelin log-likelihood, AIC ve BIC değerleri hesaplanmıştır. En iyi modelin belirlenmesi aşamasında elde edilen bilgi kriterlerine göre seçim yapılmıştır.

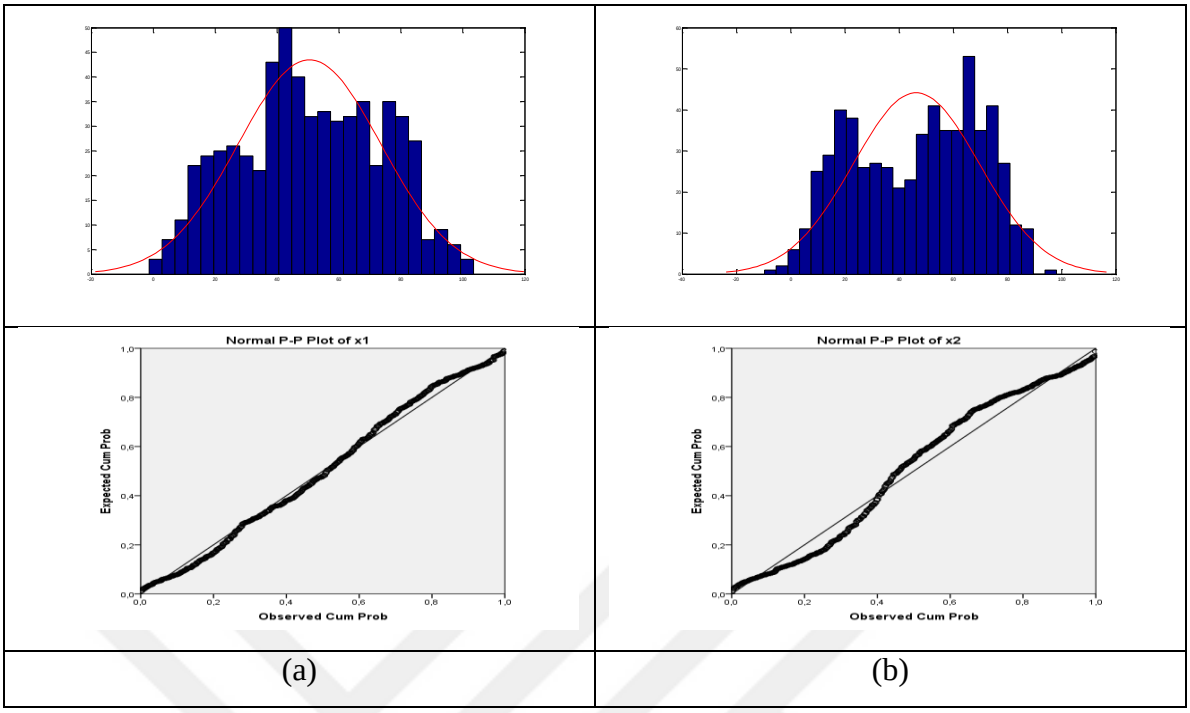
3.3.1. Heterojen Veride Normal Karma Modellerin Modele Dayalı Kümelenmesi İçin Değişken Veri Segmentasyonu

Çok değişkenli veride değişkenlerdeki parçalanmaların belirlenmesi için çeşitli yöntemlerden faydalanılabilir. Her bir değişkendeki parçalanmaların sayısı değişkenlerin alt grupları olarak adlandırılır. Değişkenlerdeki parçalanmaları belirlemek ve bu parçalanmaları anlamlı alt gruplara dönüştürmek için istatistiksel grafiksel yöntemler ve normal karma modeller kullanılır. Her bir değişkendeki anlamlı parçalanma sayısı histogram grafiğindeki tepelenme sayısı ve P-P grafiğinde gözlemlerin normal doğrusunu ile kesişme sayısına göre tahmin edilebilir.

Her bir değişkendeki anlamlı alt grup sayısını belirlemek amacıyla tek değişkenli normal karma modelin log-likelihood, AIC ve BIC değerlerine bakılır.

3.3.1.1. Normal Karma Modellerin Modele Dayalı Kümelenmesi İçin Değişkenlerdeki Bölütlenmenin Grafiksel Yöntemlere Dayalı Tahmini

Homojen veya homojen olmayan değişkenlerdeki parçalanmanın olup olmadığını belirlemek amacıyla grafiksel yöntemler kullanılabilir. Veri setindeki her bir değişkenin parçalanmasındaki en uygun parçalanma sayısını belirlemek için verinin parçalanma sayısı ile ilgili bir tahmini aralık Histogram ve P-P grafiklerinden elde edilebilir. İki değişkenli veri setinde X_1 *ve* X_2 değişkenlerinin her birinde meydana gelen en uygun parçalanmayı belirlemek amacıyla her bir değişken ayrı ayrı ele alınır. Her bir değişkendeki Histogram ve P-P grafikleri elde edilerek değişkenlerdeki parçalanma sayısı hakkında tahmini bir sayı elde edilir.



Şekil 3.3.1. İki değişkenli veride (a) X_1 değişkenindeki (b) X_2 değişkenindeki parçalanmayı gösteren histogram grafikleri ve P-P grafikleri.

X_1 *ve* X_2 değişkenleri için uygun parçalanmanın belirlenmesinde histogram grafiğindeki tepelenme sayısı ve P-P grafiğindeki grafiğin normal doğrusu veya $y = x$ doğrusu ile verideki gözlem eğrisinin kesişim sayısına göre değişkenlerdeki alt grup veya segment sayısı belirlenebilir. **Şekil 3.3.1** teki grafiklere bakılarak heterojen veri setinde X_1 *ve* X_2 değişkenlerinde parçalanma sayıları sırasıyla $k_1 = 3$ ve $k_2 = 3$ olarak tahmin edilmiştir.

3.3.1.2. Normal Karma Modellerin Modele Dayalı Kümelenmesi İçin Değişkenlerdeki Bölütlenmenin Tek Değişkenli Normal Karmalara Dayalı Belirlenmesi

İki değişkenli veride X_1 ve X_2 değişkenlerinin her birinde meydana gelen optimum parçalanmayı belirlemek amacıyla her bir değişken ayrı ayrı ele alınır. İki değişkenli veride her bir heterojen değişkendeki bölünme ya da parçalanma sayısının belirlenmesinde grafiksel yöntemlerle birlikte tek değişkenli karma modeller kullanılabilir.

Sentetik veri setindeki her bir için oluşturulan tek değişkenli karma normal modelde bileşen sayısı (3.2.1) ve (3.2.2) denklemlerindeki gibi belirlenir. Değişkenlerdeki

parçalanma sayısını elde etmek için normal karma dağılımların log-likelihhod değeri ile AIC ve BIC gibi istatistiksel bilgi kriterleri kullanılmaktadır. Oluşturulan her bir model için log-likelihood, AIC ve BIC değerleri değişkenlerdeki olasılık ağırlıkları π, ortalama vektörü μ ve varyansı σ^2 değerlerinden hesaplanır. Her bir heterojen değişkendeki parçalanmayı model tabanlı belirlemek amacıyla her bir k değerleri için hesaplanan log-likelihood, AIC ve BIC değeri karşılaştırılır ve optimum k değeri belirlenir. Log-likelihood fonksiyon değerinin maksimum, AIC ve BIC değerlerinin minimum olduğu modelde parçalanma sayısı optimum olarak bulunur. Hesaplamalarda Matlab programının istatistik paketi kütüphanesinde yer alan *gmdistribution.fit(data,k)* hazır fonksiyonu kullanılmıştır. İki değişkenli veride her bir heterojen değişken için hangi parçalanmanın optimum olduğunu belirlemek amacıyla grafiksel yöntemlerden tahmin edilen sayıda $k = 1,...\mathrm{n}$ değerleri girilerek oluşturulan karma normal modeller için log-likelihood, AIC ve BIC değerleri hesaplanmış ve hesaplama sonuçlarına göre en uygun parçalanma sayısı belirlenmiştir.

İki değişkenli heterojen veri setinde her bir değişkendeki uygun parçalanma sayısını belirlemek için tek değişkenli normal karma dağılım modeli uygulanmıştır. Değişkenlerdeki k bileşen sayısı aralıktaki en küçük değerden başlayıp üst sınıra kadar verilerek en uygun sayı elde edilmiştir.

Tablo 3.3.1. İki değişkenli veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametreleri ile log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-2734.0	5472.1	5480.8
k=2	-2716.8	5443.7	5465.7
k=3	**-2705.7**	**5427.4**	**5462.6**
k=4	-2705.5	5433.1	5481.4

Tablo 3.3.2. İki değişkenli veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametreleri ile log-likelihood, AIC ve BIC değerleri.

k	Log-l	AIC	BIC
k=1	-2742.1	5488.1	5496.9
k=2	-2681.5	5373.1	5395.1
k=3	**-2676.7**	**5369.3**	**5404.5**
k=4	-2676.1	5376.2	5422.6

İki değişkenli veri setinde tek değişkenli normal karma dağılım modeli uygulanarak elde edilen **Tablo 3.3.1 -2** deki bilgi kriterlerine göre X_1 ve X_2 değişkenleri için en iyi modelin olduğu uygun parçalanma sayısı $k_1 = 3$ ve $k_2 = 3$ olarak elde edilmiştir.

3.3.2. K-Ortalamalar Algoritması ile Veride Gözlem Değerlerinin Değişkenlerdeki Bölütlenmelere (Segmentlere) Atanması

İki değişkenli ve her değişkenin üçe parçalandığı veri setindeki değişkenlerin parçalanmalarına düşen gözlemlerin belirlenmesi için k-ortalamalar algoritması kullanılır. Başlangıçta $k = 3$ merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık (3.2.3) denlemindeki gibi hesaplanmaktadır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir.

İki değişkenli her değişkenin üçe bölündüğü veride k-ortalamalar algoritması uygulandığında üç ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum en uygun parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki her bir değişken üçe parçalanmış ve her bir gözlemin düştüğü gözlemlerin grubu belirlenmiştir. $X_1 \; ve \; X_2$ değişkenlerindeki parçalanmalara karşılık gelen alt gruplar ait oldukları değişkene göre isimlendirilmiştir. X_1 değişkenindeki üç parçalanmaya karşılık gelen alt gruplar $X_{11}, X_{12} \, ve \, X_{13}$, X_2 değişkenindeki üç parçalanmaya karşılık gelen alt gruplar ise $X_{21}, X_{22} \, ve \, X_{23}$ dür. $nx2$ tipindeki veri matrisi $X = \left[X_1 \; X_2 \right]$ şeklinde gösterilebilir. n elemanlı X_1 değişkenindeki parçalanma $X_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ X_{13} \end{bmatrix}$ şeklinde olur. Burada $X_{11}, X_{12} \, ve \, X_{13}$ sırasıyla $n_{11}, n_{12} \, ve \, n_{13}$ elemanlıdır. Yani $n = n_{11} + n_{12} + n_{13}$ olarak elde edilir. n elemanlı X_2 değişkenindeki

parçalanma $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ X_{23} \end{bmatrix}$ şeklinde olur. Burada $X_{21}, X_{22} ve X_{23}$ sırasıyla $n_{21}, n_{22} ve n_{23}$ elemanlıdır. Yani $n = n_{21} + n_{22} + n_{23}$ olarak elde edilir.

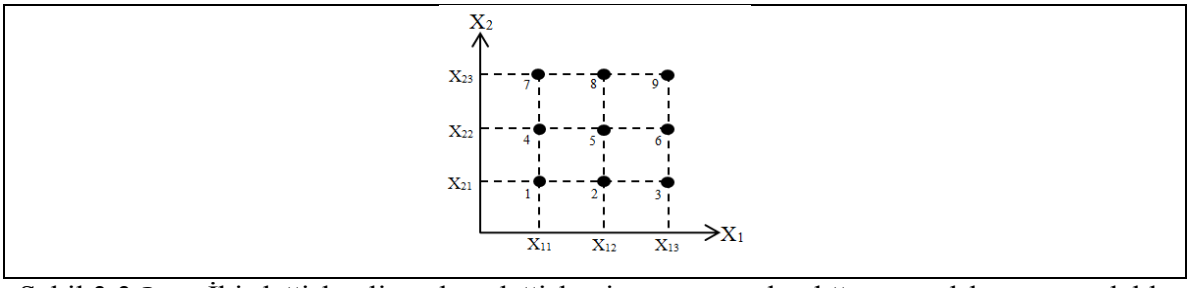
İki değişkenli heterojen veri setinde bulunan her bir değişkenin parçalanmaları ve bu parçalanmalara düşen gözlem sayıları **Tablo 3.3.3** te verilmiştir.

Tablo 3.3.3. İki değişkenli veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişken	X_1			X_2		
Segmentler	X_{11}	X_{12}	X_{13}	X_{21}	X_{22}	X_{23}
Gözlem sayısı	$n_{11} = 154$	$n_{12} = 241$	$n_{13} = 245$	$n_{21} = 205$	$n_{22} = 180$	$n_{23} = 215$
Toplam	$n_1 = 600$			$n_2 = 600$		

3.3.3. Normal Karma Modellerin Değişkenlerdeki Segmentlere Dayalı Kümelenme Merkez Sayısı ve Yapısının Belirlenmesi

İki değişkenli ve her değişkenin üçe bölündüğü veri setinde modeldeki kümelenme merkez sayıları değişkenlerdeki parçalanmalara bağlı olarak hesaplanır. Değişkenlerdeki parçalanmaların sayısı kümelenme merkez sayısını, parçalanmalara düşen gözlem değerleri sırasıyla, kümelenme merkezlerinin yerini, şeklini ve büyüklüğünü belirlemektedir [49]. Modeldeki değişkenlerin parçalanmalarının oluşturduğu maksimum ve minimum kümelenme merkezlerinin sayısı C_{max} ve $C_{min}, s = 1,..., p$ olmak üzere k_s değerlerine bağlı olarak (3.2.4)-(3.2.5) denklemleri ile hesaplanır. Burada, verideki parçalanmalara karşılık gelen en çok kümelenme merkezi sayısı $C_{max} = k_1.k_2 = 3.3 = 9$ ve en az kümelenme merkez sayısı $C_{min} = \max\{k_1, k_2\} = \max\{3,3\} = 3$ olarak elde edilir.



Şekil 3.3.2. İki değişkenli ve her değişkenin üçe parçalandığı normal karma modelde X_1 değişkeninin X_{11}, X_{12}, X_{13} ve X_2 değişkeninin X_{21}, X_{22}, X_{23} parçalanmalarının oluşturduğu kümelenme merkezleri ve bu merkezlerin numaraları.

Bir numaralı kümelenme merkezi: X_1 değişkeninin X_{11} ve X_2 değişkeninin X_{21} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_1 ve varyans-kovaryans matrisi Σ_1, sırasıyla $\mu_1 = \begin{bmatrix} \mu_{11} \\ \mu_{21} \end{bmatrix}$ ve $\Sigma_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{21} \\ \rho_1\sigma_{21}\sigma_{11} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_1 = Corr(X_{11}, X_{21})$ olmak üzere korelasyon katsayısını göstermektedir. Bu kümelenme merkezindeki veriler, $N(x; \mu_1, \Sigma_1)$ normal dağılımına sahiptir.

İki numaralı kümelenme merkezi: X_1 değişkeninin X_{12} ve X_2 değişkeninin X_{21} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_2 ve varyans-kovaryans matrisi Σ_2, sırasıyla $\mu_2 = \begin{bmatrix} \mu_{12} \\ \mu_{21} \end{bmatrix}$ ve $\Sigma_2 = \begin{bmatrix} \sigma_{12}^2 & \rho_2\sigma_{12}\sigma_{21} \\ \rho_2\sigma_{21}\sigma_{12} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_2 = Corr(X_{12}, X_{21})$ olmak korelasyon katsayısını göstermektedir. $N(x; \mu_2, \Sigma_2)$ normal dağılımına sahiptir.

Üç numaralı kümelenme merkezi: X_1 değişkeninin X_{13} ve X_2 değişkeninin X_{21} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_3 ve varyans-kovaryans matrisi Σ_3, sırasıyla $\mu_3 = \begin{bmatrix} \mu_{13} \\ \mu_{21} \end{bmatrix}$ ve $\Sigma_3 = \begin{bmatrix} \sigma_{13}^2 & \rho_3\sigma_{13}\sigma_{21} \\ \rho_3\sigma_{21}\sigma_{13} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_3 = Corr(X_{13}, X_{21})$ olmak üzere korelasyon katsayısını

göstermektedir. Bu kümelenme merkezindeki veriler, $N(x;\mu_3,\Sigma_3)$ normal dağılımına sahiptir.

Dört numaralı kümelenme merkezi: X_1 değişkeninin X_{11} ve X_2 değişkeninin X_{22} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_4 ve varyans-kovaryans matrisi Σ_4, sırasıyla $\mu_4=\begin{bmatrix}\mu_{11}\\ \mu_{22}\end{bmatrix}$ ve $\Sigma_4=\begin{bmatrix}\sigma_{11}^2 & \rho_4\sigma_{11}\sigma_{22}\\ \rho_4\sigma_{22}\sigma_{11} & \sigma_{22}^2\end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_4=Corr(X_{11},X_{22})$ olmak üzere korelasyon katsayısını göstermektedir. Bu kümelenme merkezindeki veriler, $N(x;\mu_4,\Sigma_4)$ normal dağılımına sahiptir.

Beş numaralı kümelenme merkezi: X_1 değişkeninin X_{12} ve X_2 değişkeninin X_{22} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_5 ve varyans-kovaryans matrisi Σ_5, sırasıyla $\mu_5=\begin{bmatrix}\mu_{12}\\ \mu_{22}\end{bmatrix}$ ve $\Sigma_5=\begin{bmatrix}\sigma_{12}^2 & \rho_5\sigma_{12}\sigma_{22}\\ \rho_5\sigma_{22}\sigma_{12} & \sigma_{22}^2\end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_5=Corr(X_{12},X_{22})$ olmak üzere korelasyon katsayısını göstermektedir. Bu kümelenme merkezindeki veriler, $N(x;\mu_5,\Sigma_5)$ normal dağılımına sahiptir.

Altı numaralı kümelenme merkezi: X_1 değişkeninin X_{13} ve X_2 değişkeninin X_{22} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_6 ve varyans-kovaryans matrisi Σ_6, sırasıyla $\mu_6=\begin{bmatrix}\mu_{13}\\ \mu_{22}\end{bmatrix}$ ve $\Sigma_6=\begin{bmatrix}\sigma_{13}^2 & \rho_6\sigma_{13}\sigma_{22}\\ \rho_6\sigma_{22}\sigma_{13} & \sigma_{22}^2\end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_6=Corr(X_{13},X_{22})$ olmak üzere korelasyon katsayısını göstermektedir. Bu kümelenme merkezindeki veriler, $N(x;\mu_6,\Sigma_6)$ normal dağılımına sahiptir.

Yedi numaralı kümelenme merkezi: X_1 değişkeninin X_{11} ve X_2 değişkeninin X_{23} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_7 ve varyans-kovaryans matrisi Σ_7, sırasıyla $\mu_7=\begin{bmatrix}\mu_{11}\\ \mu_{23}\end{bmatrix}$ ve $\Sigma_7=\begin{bmatrix}\sigma_{11}^2 & \rho_7\sigma_{11}\sigma_{23}\\ \rho_7\sigma_{23}\sigma_{11} & \sigma_{23}^2\end{bmatrix}$

şeklinde tanımlansın. Burada $\rho_7 = Corr(X_{11}, X_{23})$ olmak üzere korelasyon katsayısını göstermektedir. Bu kümelenme merkezindeki veriler, $N(x; \mu_7, \Sigma_7)$ normal dağılımına sahiptir.

Sekiz numaralı kümelenme merkezi: X_1 değişkeninin X_{12} ve X_2 değişkeninin X_{23} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_8 ve varyans-kovaryans matrisi Σ_8, sırasıyla $\mu_8 = \begin{bmatrix} \mu_{12} \\ \mu_{23} \end{bmatrix}$ ve $\Sigma_8 = \begin{bmatrix} \sigma_{12}^2 & \rho_8\sigma_{12}\sigma_{23} \\ \rho_8\sigma_{23}\sigma_{12} & \sigma_{23}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_8 = Corr(X_{12}, X_{23})$ olmak üzere korelasyon katsayısını göstermektedir. Bu kümelenme merkezindeki veriler, $N(x; \mu_8, \Sigma_8)$ normal dağılımına sahiptir.

Dokuz numaralı kümelenme merkezi: X_1 değişkeninin X_{13} ve X_2 değişkeninin X_{23} parçalarının oluşturduğu kümelenme merkezidir. Buradaki merkezin ortalama vektörü μ_9 ve varyans-kovaryans matrisi Σ_9, sırasıyla $\mu_9 = \begin{bmatrix} \mu_{13} \\ \mu_{23} \end{bmatrix}$ ve $\Sigma_9 = \begin{bmatrix} \sigma_{13}^2 & \rho_9\sigma_{13}\sigma_{23} \\ \rho_9\sigma_{23}\sigma_{13} & \sigma_{23}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_9 = Corr(X_{13}, X_{23})$ olmak üzere korelasyon katsayısını göstermektedir. Bu kümelenme merkezindeki veriler, $N(x; \mu_9, \Sigma_9)$ normal dağılımına sahiptir.

3.3.4. Normal Karma Modellerde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

İki değişkenli ve her değişkenin üçe parçalanması durumunda kümelenme merkezleri için M_{toplam} ile gösterilen oluşabilecek tüm modellerin sayısı değişkenlerdeki parçalanmalara bağlı olarak elde edilir. İki değişkenli normal karma modellerin toplam sayısı (3.2.6) denkleminden $M_{toplam} = 2^{C\max} - 1 = 2^{3.3} - 1 = 511$ olarak elde edilebilir. Burada çıkarılan 1 model kümelenmenin bulunmadığı boş modeldir. Oluşabilecek toplam modeller içerisinde $r = 1,...,9$ kümelenme merkezli modeller bulunmaktadır. $\binom{n}{r}$ ifadesi n elemanlı kümedeki r elemanlı kombinasyonları göstermek üzere, tek kümelenme

merkezli oluşturulabilecek model sayısı $\binom{9}{1}=9$, iki kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{2}=36$, üç kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{3}=84$, dört kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{4}=126$, beş kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{5}=126$, altı kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{6}=84$, yedi kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{7}=36$, sekiz kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{8}=9$ ve dokuz kümelenme merkezli normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\binom{9}{9}=1$ dir. Toplamda normal dağılımların karmalarıyla oluşturulabilecek model sayısı $\sum_{k=1}^{9}\binom{9}{k}-1=511$ şeklinde merkez sayılarına göre bulunur.

3.3.5. Normal Karma Modellerde Uygun Aday Model Sayısının Hesaplanması

İki değişkenli normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ve yine parçalanmalara bağlı olarak toplam model sayısı (3.3.3) denkleminden elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli veya uygun modellerin sayısı araştırılmıştır. Bu geçerli modeller **Şekil 3.3.4.** de gösterilen merkezler üzerinden kısaca değişkenlerdeki parçalanmaların karşılık geldiği her satır ve her sütunda en az bir kümelenme merkezi bulunacak varsayımı ile anlatılabilir. Varsayıma uyan

modellere Uygun Aday Model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanabilir.

Her satır ve sütunda en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı her bir merkezin bulunduğu satır ve sütundaki konumuna göre kombinasyon ile elde edilebilir [75]. Hesaplama sonucunda iki değişkenli veride her değişkenin üçe bölündüğü durumda, veride olabilecek minimum kümelenme merkez sayısı 3, maksimum kümelenme merkez sayısı 9 için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları **Tablo 3.3.4** te verilmiştir.

Tablo 3.3.4. İki değişkenli ve her değişkenin üçe bölündüğü normal karma modellerde uygun aday modeller için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları.

Merkez Sayısı	**Merkezlerin Konumu**	**Modellerin Sayısı İçin Bağıntı**	**Model Sayısı**
Üç merkezli	1 1 1 şeklinde sütunlara dağılan	$\binom{3}{1}\binom{2}{1}\binom{1}{1}=3!$	6
Dört merkezli	2 1 1 şeklinde sütunlara dağılan	$\binom{3}{1}[1+1+1]\frac{3!}{2!}=3.5.3$	45
Beş merkezli	3 1 1 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{1}\binom{3}{1}\frac{3!}{2!}=1.3.3.3$	90
	2 2 1 şeklinde sütunlara dağılan	$\binom{3}{2}[1+3+3]\frac{3!}{2!}=3.7.3$	
Altı merkezli	3 2 1 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{2}\binom{3}{1}3!=1.3.3.6$	78
	2 2 2 şeklinde sütunlara dağılan	$\binom{3}{2}[2+3+3]\frac{3!}{3!}=3.8.1$	
Yedi merkezli	3 3 1 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{3}\binom{3}{1}\frac{3!}{2!}=1.1.3.3$	36
	3 2 2 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{2}\binom{3}{2}\frac{3!}{2!}=1.3.3.3$	
Sekiz merkezli	3 3 2 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{3}\binom{3}{2}\frac{3!}{2!}=1.1.3.3$	9
Dokuz merkezli	3 3 3 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{3}\binom{3}{3}\frac{3!}{3!}=1.1.1.1$	1
Toplam model sayısı		$2^{3.3}-1=511$	
Toplam Uygun model sayısı		0+0+6+45+90+78+36+9+1=265	

Varsayım altındaki uygun aday model sayısının hesaplanmasındaki kombinatorik problem bir matris yapısına karşılık gelmektedir. Varsayımdaki değişkenler ve değişkenlerin parçalanmaları matrisin satır ve sütunlarına karşılık gelmektedir. n_{ij} *i.* satır ve *j.* sütundaki kümelenme merkezini göstermek üzere, *nxm* tipindeki sıfırlardan oluşan bir kare matriste her satır ve her sütunda en az bir kümelenmenin olduğu başka bir ifade ile sıfırdan farklı en az bir elemanın bulunduğu matrislerin sayısı (3.2.8) denkleminden hesaplanır.

Bu yedi uygun duruma karşılık gelen kümelenme yapılarının modelleri **Tablo 3.3.4** te verilmiştir. İki değişkenli ve her değişkenin üçe bölündüğü veri setinde toplam 511 model arasından varsayıma uyan 265 uygun aday model belirlenmiş, varsayıma uymayan modeller çıkarılmıştır. Çalışmada üç kümelenme merkezli altı uygun aday modelin temsili gösterimi ve bu modellerden elde edilen normal karma modeller için hesaplama yapılmıştır.

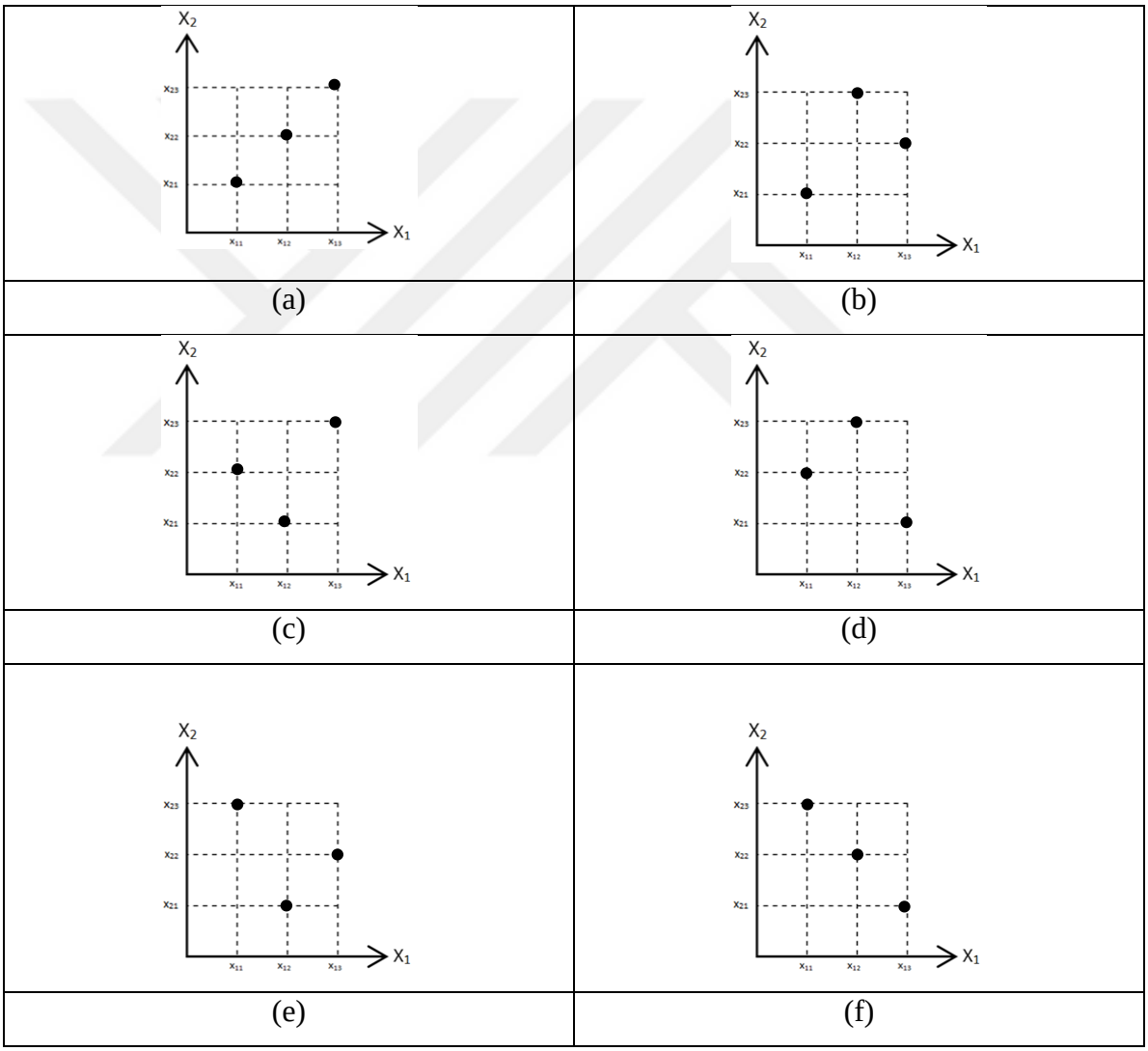
Tablo 3.3.5. İki değişkenli ve her değişkenin üçe bölündüğü normal karma modeller arasından varsayıma uyan üç kümelenme merkezli uygun aday modellerin sayısı ve karşılık gelen kümelenme yapılarının temsili modelleri.

	Modelin Dizi Gösterimi	Uygun Model sayısı
Üç kümelenme merkezli uygun modeller	100010001	6
	100001010	
	010100001	
	001100010	
	010001100	
	001010100	

3.3.6. Normal Karma Modellerde Uygun Modellerin Bölütlenmeye Dayalı Oluşturulması ve Parametre Tahminleri

İki değişkenli normal karma dağılım modelleri için bileşen ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir. Karma modeldeki her bir merkezin olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisi modeldeki olasılık yoğunluk fonksiyonunu elde etmek için kullanılır. **Tablo 3.3.5** te

uygun aday modellerin temsili gösterimleri modeldeki merkezlerin numaralarına göre dizi gösterimi şeklinde verilmiştir. Modelin temsili gösteriminde kullanılan “0” ve “1” elemanları modeldeki karşılık gelen merkezde yığılmanın olup olmadığını göstermektedir. Eğer uygun aday modelde herhangi bir merkezde kümelenme varsa “1” yoksa “0” ile gösterilmiştir. İki değişkenli ve her değişkenin üçe parçalandığı veri setinde oluşabilecek üç kümelenme merkezli uygun aday modellerin dizi gösteriminden elde edilen grafiksel gösterimler **Şekil 3.3.4** te verilmiştir.



Şekil 3.3.3. İki değişkenli ve her değişkenin üçe parçalandığı veri setinde (a) - (f) üç kümelenme merkezli uygun aday modellerin düzlemsel grafiklerini göstermektedir.

3.3.6.1. Normal Karma Modeller ve Parametre Tahminleri

Simülasyon ile oluşturulan sentetik veri setindeki iki değişkenin her birinin üçe bölünmesi durumunda oluşacak kümelenme merkezleri **Şekil 3.3.4** de gösterilmiştir. **Tablo 3.3.4** te veri setindeki parçalanmalara karşılık gelen ve varsayımlara uyan uygun aday modeller verilmiştir. Toplam 265 uygun aday model içerisinden üç bileşenli normal karma modele karşılık gelen modeller $u = 1,...,6$ ve $i = 1,2,3$ için normal bileşen yoğunluk fonksiyonları,

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{3} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.3.1}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{3} \pi i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir. Dört bileşenli normal karma modele karşılık gelen modeller $u = 7,...,51$ ve $i = 1,...,4$ için normal bileşen yoğunluk fonksiyonları,

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{4} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.3.2}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{4} \pi i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir. Beş bileşenli normal

karma modele karşılık gelen modeller $u = 52,...,141$ ve $i = 1,...,5$ için normal bileşen yoğunluk fonksiyonları,

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{5} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.3.3}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{4} \pi i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir. Altı bileşenli normal karma modele karşılık gelen modeller $u = 142,...,219$ ve $i = 1,...,6$ için normal bileşen yoğunluk fonksiyonları,

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{6} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.3.4}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{6} \pi i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir. Yedi bileşenli normal karma modele karşılık gelen modeller $u = 220,...,255$ ve $i = 1,...,7$ için normal bileşen yoğunluk fonksiyonları,

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{7} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.3.5}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{7} \pi i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir. Sekiz bileşenli normal karma modele karşılık gelen modeller $u = 256,...,264$ ve $i = 1,...,8$ için normal bileşen yoğunluk fonksiyonları,

$$f^{(u)}(x; \mu^{(u)}, \Sigma^{(u)}) = \sum_{i=1}^{8} \pi_i^{(u)} f_i\left(x; \mu_i^{(u)}, \Sigma_i^{(u)}\right) \tag{3.3.6}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{8} \pi i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir. Dokuz bileşenli normal karma modele karşılık gelen modeller $u = 265$ ve $i = 1,...,9$ için normal bileşen yoğunluk fonksiyonları,

$$f^{(u)}(x; \mu^{(u)}, \Sigma^{(u)}) = \sum_{i=1}^{9} \pi_i^{(u)} f_i\left(x; \mu_i^{(u)}, \Sigma_i^{(u)}\right) \tag{3.3.7}$$

olmak üzere, olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{9} \pi i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix}$ ve varyans-kovaryans matrisi $\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)} \sigma_{1x}^{(u)} \sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)} \sigma_{2y}^{(u)} \sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix}$ olarak ifade edilir.

$i = 1,...,g$ olmak üzere normal karma modellerdeki yoğunluk fonksiyonlarının π_i, μ_i ve Σ_i parametreleri tahmin edilmelidir.

Tablo 3.3.4 te elde edilen üç bileşenli uygun aday modellerden altı karma model ve (3.3.1) denklemindeki $u = 1,...,6$ için karma modellerin yapısına uyan modellerin, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{3} n_i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}, i = 1,2,3$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. **Tablo 3.3.4** te elde edilen dört bileşenli uygun aday modellerden kırk beş karma model ve (3.3.2) denklemindeki $u = 7,...,51$ için karma modellerin yapısına uyan modeller için, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{4} n_i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$,

varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}, i = 1,...,4$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. **Tablo 3.3.4** te elde edilen beş bileşenli uygun aday modellerden doksan karma model ve (3.3.3) denklemindeki $u = 52,...,141$ için karma modellerin yapısına uyan modeller için, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{5} n_i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}, i = 1,...,5$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. **Tablo 3.3.4** te elde edilen altı bileşenli uygun aday modellerden yetmiş sekiz karma model ve (3.3.4) denklemindeki $u = 142,...,219$ için karma modellerin yapısına uyan modeller için, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{6} n_i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}$, $i = 1,...,6$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. **Tablo 3.3.4** te elde edilen yedi bileşenli uygun aday modellerden otuz altı karma model ve (3.3.5) denklemindeki $u = 220,...,255$ için karma modellerin yapısına uyan modeller için, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{7} n_i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere

ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri

$\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}, i = 1,...,7$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. **Tablo 3.3.4** te elde edilen sekiz bileşenli uygun aday modellerden dokuz karma model ve (3.3.6) denklemindeki $u = 256,...,264$ için karma modellerin yapısına uyan modeller için, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{8} n_i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}, i = 1,...,8$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. **Tablo 3.3.4** te elde edilen dokuz bileşenli uygun aday modelden bir karma model ve (3.3.7) denklemindeki $u = 265$ için karma modellerin yapısına uyan modeller için, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{9} n_i}$, $x = 1,2,3$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri

$\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}, i = 1,...,9$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. İki değişkenli heterojen veri setindeki normal karma modellerin kümelenmelerinden oluşan uygun aday modellerin parametre değerleri **Tablo 3.3.6** da gösterilmiştir.

Tablo 3.3.6. İki değişkenli veride uygun aday modeller için parametre tahminleri.

Model No	π	μ	Σ
Üç Merkezli Birinci Model	$\pi_1 = 0.2992$ $\pi_5 = 0.3508$ $\pi_9 = 0.3500$	$\mu_1 = \begin{bmatrix} 20.6545 \\ 18.8291 \end{bmatrix}$ $\mu_5 = \begin{bmatrix} 47.5032 \\ 47.8592 \end{bmatrix}$ $\mu_9 = \begin{bmatrix} 76.3423 \\ 71.3897 \end{bmatrix}$	$\Sigma_1 = \begin{bmatrix} 71.353 & 9.662 \\ 9.662 & 74.302 \end{bmatrix}$ $\Sigma_5 = \begin{bmatrix} 55.425 & -0.680 \\ -0.680 & 55.535 \end{bmatrix}$ $\Sigma_9 = \begin{bmatrix} 90.141 & 4.046 \\ 4.046 & 55.737 \end{bmatrix}$
Üç Merkezli İkinci Model	$\pi_1 = 0.2992$ $\pi_6 = 0.3208$ $\pi_8 = 0.3800$	$\mu_1 = \begin{bmatrix} 20.6545 \\ 18.8291 \end{bmatrix}$ $\mu_6 = \begin{bmatrix} 76.3423 \\ 47.8592 \end{bmatrix}$ $\mu_8 = \begin{bmatrix} 47.5032 \\ 71.3897 \end{bmatrix}$	$\Sigma_1 = \begin{bmatrix} 71.353 & 9.662 \\ 9.662 & 74.302 \end{bmatrix}$ $\Sigma_6 = \begin{bmatrix} 90.141 & 5.35 \\ 5.35 & 55.535 \end{bmatrix}$ $\Sigma_8 = \begin{bmatrix} 55.425 & -0.609 \\ -0.609 & 55.737 \end{bmatrix}$
Üç Merkezli Üçüncü Model	$\pi_2 = 0.3717$ $\pi_4 = 0.2783$ $\pi_9 = 0.3500$	$\mu_2 = \begin{bmatrix} 47.5032 \\ 18.8291 \end{bmatrix}$ $\mu_4 = \begin{bmatrix} 20.6545 \\ 47.8592 \end{bmatrix}$ $\mu_9 = \begin{bmatrix} 76.3423 \\ 71.3897 \end{bmatrix}$	$\Sigma_2 = \begin{bmatrix} 55.425 & -6.117 \\ -6.117 & 74.302 \end{bmatrix}$ $\Sigma_4 = \begin{bmatrix} 71.353 & -4.260 \\ -4.260 & 55.535 \end{bmatrix}$ $\Sigma_9 = \begin{bmatrix} 90.141 & 4.046 \\ 4.046 & 55.737 \end{bmatrix}$
Üç Merkezli Dördüncü Model	$\pi_3 = 0.3417$ $\pi_4 = 0.2783$ $\pi_8 = 0.3800$	$\mu_3 = \begin{bmatrix} 76.3423 \\ 18.8291 \end{bmatrix}$ $\mu_4 = \begin{bmatrix} 20.6545 \\ 47.8592 \end{bmatrix}$ $\mu_8 = \begin{bmatrix} 47.5032 \\ 71.3897 \end{bmatrix}$	$\Sigma_3 = \begin{bmatrix} 90.141 & -11.249 \\ -11.249 & 74.302 \end{bmatrix}$ $\Sigma_4 = \begin{bmatrix} 71.353 & -4.260 \\ -4.260 & 55.535 \end{bmatrix}$ $\Sigma_8 = \begin{bmatrix} 55.425 & -0.609 \\ -0.609 & 55.737 \end{bmatrix}$
Üç Merkezli Beşinci Model	$\pi_2 = 0.3717$ $\pi_6 = 0.3208$ $\pi_7 = 0.3075$	$\mu_2 = \begin{bmatrix} 47.5032 \\ 18.8291 \end{bmatrix}$ $\mu_6 = \begin{bmatrix} 76.3423 \\ 47.8592 \end{bmatrix}$ $\mu_7 = \begin{bmatrix} 20.6545 \\ 71.3897 \end{bmatrix}$	$\Sigma_2 = \begin{bmatrix} 55.425 & -6.117 \\ -6.117 & 74.302 \end{bmatrix}$ $\Sigma_6 = \begin{bmatrix} 90.141 & 5.35 \\ 5.35 & 55.535 \end{bmatrix}$ $\Sigma_7 = \begin{bmatrix} 71.353 & -0.291 \\ -0.291 & 55.737 \end{bmatrix}$
Üç Merkezli Altıncı Model	$\pi_3 = 0.3417$ $\pi_5 = 0.3508$ $\pi_7 = 0.3075$	$\mu_3 = \begin{bmatrix} 76.3423 \\ 18.8291 \end{bmatrix}$ $\mu_5 = \begin{bmatrix} 47.5032 \\ 47.8592 \end{bmatrix}$ $\mu_7 = \begin{bmatrix} 20.6545 \\ 71.3897 \end{bmatrix}$	$\Sigma_3 = \begin{bmatrix} 90.141 & -11.249 \\ -11.249 & 74.302 \end{bmatrix}$ $\Sigma_5 = \begin{bmatrix} 55.425 & -0.680 \\ -0.680 & 55.535 \end{bmatrix}$ $\Sigma_7 = \begin{bmatrix} 71.353 & -0.291 \\ -0.291 & 55.737 \end{bmatrix}$

3.3.7. Normal Karma Modellerin Log-Likelihood Fonksiyonu, AIC ve BIC Değerlerinin Hesaplanması.

Modele dayalı kümeleme yapmak için iki değişkenli ve her değişkenin ikiye parçalandığı birinci, ikinci ve üçüncü veri setlerindeki uygun aday modellerin normal karma modelleri belirlenmiştir. Normal karma modeller arasından en iyi modelin seçimi için her bir uygun aday modelin çok değişkenli normal karma dağılımlardan log-likelihood fonksiyonu, AIC ve BIC değerleri elde edilmiştir. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu, $i=1,2,...n$ ve $j=1,2,...g$ olmak üzere $f\left(x_i;\theta_j\right)$ karma olasılık yoğunluk fonksiyonunun likelihood fonksiyonu ve logaritması alınmış likelihood fonksiyonu (3.2.13) ve (3.2.14) denklemleri ile hesaplanır.

Çok değişkenli normal karma modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kriteri (BIC) ve Akaike bilgi kriteri (AIC) (3.2.15) ve (3.2.17) denklemlerinden hesaplanır.

Tablo 3.3.7. İki değişkenli normal karma dağılımların modele dayalı kümelenmesi için veri setinden log-l, AIC ve BIC değerlerinin elde edilmesi için matlab kodu.

```
a=load('p13x11.txt');
b=load('p13x12.txt');
c=load('p13x21.txt');
d=load('p13x22.txt');
pi1=(size(a)+size(c))/(size(a) +size(c) +size(b) +size(c)+size(b)+size(d));
pi2=(size(b)+size(c))/(size(a) +size(c) +size(b) +size(c)+size(b)+size(d));
pi4=(size(b)+size(d))/(size(a) +size(c) +size(b) +size(c)+size(b)+size(d));
x(1,:)=load('p13x1.txt');
x(2,:)=load('p13x2.txt');
mu(1,:)=mean(a);
mu(2,:)=mean(c);
sigma=[92.963 18.051  ;18.051 136.873];
for i=1:length(x(1,:))
pdf1(i)=(1/(2*pi))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-mu)'*inv(sigma)*(x(:,i)-mu));
end
mu(1,:)=mean(b);
mu(2,:)=mean(c);
sigma=[108.335 -2.386;-2.386 136.873];
for i=1:length(x(1,:))
pdf2(i)=(1/(2*pi))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-mu)'...
   *inv(sigma)*(x(:,i)-mu));
end
mu(1,:)=mean(b);
mu(2,:)=mean(d);
sigma=[108.335 15.393;15.393 80.099];
for i=1:length(x(1,:))
pdf4(i)=(1/(2*pi))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-mu)'...
   *inv(sigma)*(x(:,i)-mu));
end
 oyf1=pi1*pdf1+pi2*pdf2+pi4*pdf4;
```

```
% Log-likelihood fonksiyonu
lnl= -17*log(length(oyf1))+ sum(log(pi1*pdf1(:)+pi2*pdf2(:)+pi4*pdf4(:)));
%AIC bilgi kriteri ve BIC bilgi kriteri
aic=-2*lnl+2*17;
bic=-2*lnl+17*log(length(lnl));
```

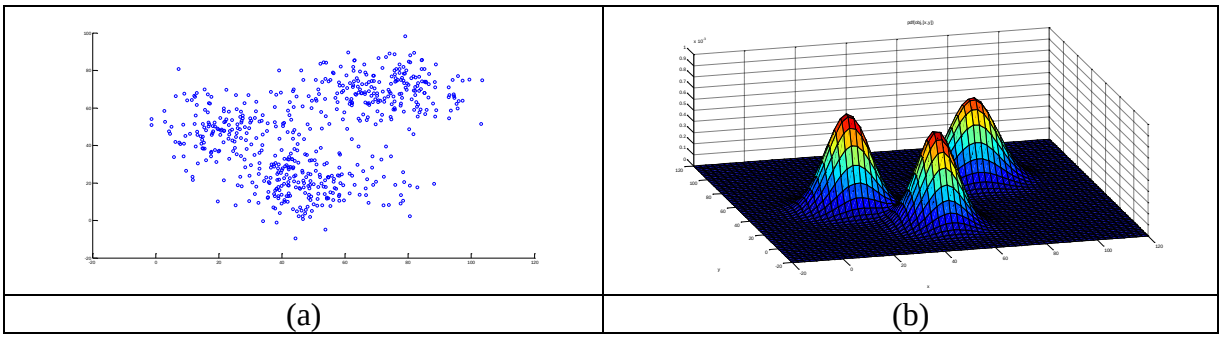
Tablo 3.3.8. İki değişkenli normal karma dağılımların modele dayalı kümelenmesi için log-l, AIC ve BIC değerleri.

Model no:	$\log L(\Psi)$	$AIC = -2\ln L(\Psi) + 2d$	$BIC = -2\ln L(\Psi) + d\log n$	d
1. Model	-11259	22521	22528	17
2. Model	-12193	24390	24397	17
3. Model	**-9229.1**	**1846.2**	**1846.9**	17
4. Model	-11424	22851	22859	17
5. Model	-12300	24603	24610	17
6. Model	-14376	28755	28763	17

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin log-l, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. İki değişkenli veri setinde normal karma modellerden uygun aday modellerin log-l, AIC ve BIC değerleri hesaplanmış ve **Tablo 3.3.7** de verilmiştir. Uygun aday modeller arasından modele dayalı kümelemede en iyi model log-likelihood değeri en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir.

```
a=load('p13x11.txt');b=load('p13x12.txt');c=load('p13x21.txt');
d=load('p13x22.txt');
MU = [mean(b) mean(c);mean(a) mean(d)];
SIGMA = cat(3,[ 92.963 18.051;18.051 136.873],[ 108.335 15.393;15.393
80.099]);
p = ones(1,2)/2;
obj = gmdistribution(MU,SIGMA,p);
figure;ezsurf(@(x,y)pdf(obj,[x y]),[-15 120],[-15 120]);
```

İki değişkenli ve her değişkenin üçe parçalandığı veri setinde en iyi model üç bileşenli birinci model olarak belirlenmiştir.



(a) (b)

Şekil 3.3.4. İki değişkenli ve her değişkenin üçe bölündüğü modeller arasından en iyi modelin (a) saçılım grafiği (b) yüzey grafiği.

3.4. Çok Değişkenli Veride Normal Karma Modellerin Modele Dayalı Kümelenmesi için Yeni Bir Kümeleme Algoritması: Gerçek Veri Seti (Ruspini) Üzerine Bir Uygulama

Çok değişkenli veride modele dayalı kümeleme yapmak için verideki her bir değişkenin anlamlı alt gruplara ayrılması gerekmektedir. Normal karma modeldeki bileşenler değişkenlerdeki alt gruplardan veya değişkenlerdeki segmentlerden meydana gelmektedir. Normal karma modellerdeki her bileşene bir kümelenme karşılık gelir, dolayısıyla modeldeki bileşen sayısının belirlenmesi kümeleme analizindeki en zor problemlerdendir [5]. Bu bölümdeki çalışmada geliştirilen modele dayalı kümeleme algoritması gerçek veri üzerinde uygulanarak kümeleme analizi yapılmıştır. Ruspini [77] verisi olarak adlandırılan iki değişkenli veri seti üzerinde kümeleme analizi uygulanarak verideki bileşen sayısı ve buna bağlı olarak en iyi normal karma model ortaya çıkarılmıştır. Değişkenler üzerindeki parçalanmalar grafiksel ve tek değişkenli normal karmalara dayalı belirlenmiştir. Değişkenler üzerindeki parçalanmalara karşılık gelen kümelenme merkezlerinin yeri şekli ve sayısı belirlenmiş ve aday modeller oluşturulmuştur. Aday modeller arasından en iyi karma model bilgi kriterlerine göre belirlenmiştir.

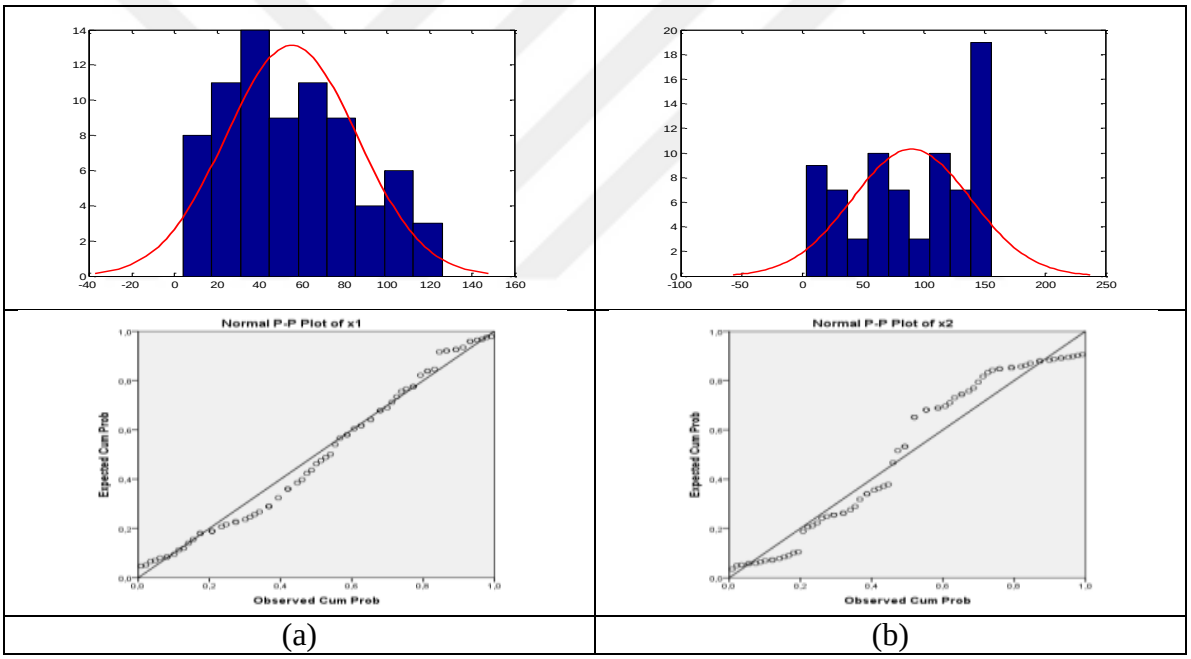
3.4.1. Ruspini Verisinde Normal Karma Modellerin Modele Dayalı Kümelenmesi için Değişkenlerdeki Grup (Segment) Sayısının Belirlenmesi

Çok değişkenli veride değişkenlerdeki parçalanmaların belirlenmesi için çeşitli yöntemlerden faydalanılabilir. Her bir değişkendeki parçalanmalar değişkenlerin alt grupları (segmentleri) olarak adlandırılır. Değişkenlerdeki parçalanmaları belirlemek ve bu parçalanmaları anlamlı alt gruplara dönüştürmek için istatistiksel grafiksel yöntemler ve normal karma modeller kullanılır. Her bir değişkendeki anlamlı parçalanma sayısı histogram grafiğindeki tepelenme sayısı ve P-P grafiğinde gözlemlerin normal doğrusu ile kesişme sayısına göre tahmin edilebilir.

Her bir değişkendeki anlamlı alt grup sayısını belirlemek amacıyla tek değişkenli normal karma modelin log-likelihood, AIC ve BIC değerlerine bakılır.

3.4.1.1. Ruspini Verisinde Normal Karma Dağılımların Modele Dayalı Kümelenmesi İçin Değişkenlerdeki Parçalanmanın Grafiksel Yöntemlerle Tahmini

Homojen veya homojen olmayan (heterojen) değişkenlerdeki parçalanmanın olup olmadığını belirlemek amacıyla grafiksel yöntemler kullanılabilir. Veri setindeki her bir değişkenin parçalanmasındaki en uygun parçalanma sayısını belirlemek için verinin parçalanma sayısı ile ilgili bir tahmini aralık Histogram ve P-P grafiklerinden elde edilebilir. İki değişkenli veri setinde X_1 *ve* X_2 değişkenlerinin her birinde meydana gelen en uygun parçalanmayı belirlemek amacıyla her bir değişken ayrı ayrı ele alınır. Her bir değişkendeki Histogram ve P-P grafikleri elde edilerek değişkenlerdeki parçalanma sayısı hakkında tahmini bir sayı elde edilir.



(a) (b)

Şekil 3.4.1. İki değişkenli veride (a) X_1 değişkenindeki (b) X_2 değişkenindeki parçalanmayı gösteren histogram ve P-P grafikleri.

X_1 *ve* X_2 değişkenleri için uygun parçalanmanın belirlenmesinde histogram grafiğindeki tepelenme sayısı ve P-P grafiğindeki normal doğrusu veya $y = x$ doğrusu ile verideki gözlem eğrisinin kesişim sayısına göre her bir değişkendeki parçalanma diğer bir ifade ile anlamlı alt grup sayısı belirlenebilir. **Şekil 3.4.1** deki grafiklere göre heterojen Ruspini verisindeki X_1 *ve* X_2 değişkenlerindeki parçalanma sayısı sırasıyla, $k_1 = 4$ ve $k_2 = 4$ olarak tahmin edilmiştir.

3.4.1.2. Ruspini Verisinde Normal Karma Modellerin Modele Dayalı Kümelenmesi İçin Değişkenlerdeki Parçalanmanın Tek Değişkenli Normal Karmalara Dayalı Belirlenmesi

Ruspini veri setindeki her bir değişkene tek değişkenli karma normal model oluşturarak her bir değişken için oluşturulan tek değişkenli karma normal modelde bileşen sayısı (3.2.1) ve (3.2.2) denklemlerinden elde edilebilir. Değişkenlerdeki parçalanma sayısını elde etmek için normal karma dağılımların log-likelihhod, AIC ve BIC gibi istatistiksel bilgi kriterleri kullanılmaktadır. Oluşturulan her bir model için log-likelihood, AIC ve BIC değerleri değişkenlerdeki olasılık ağırlıkları π, ortalama vektörü μ ve varyansı σ^2 değerlerinden hesaplanır. Her bir heterojen değişkendeki parçalanmayı model tabanlı belirlemek amacıyla her bir k değerleri için hesaplanan log-likelihood, AIC ve BIC değeri karşılaştırılır ve optimum k değeri belirlenir. Log-likelihood fonksiyon değerinin maksimum, AIC ve BIC değerlerinin minimum olduğu modelde parçalanma sayısı optimum olarak bulunur. Hesaplamalarda Matlab programının istatistik paketi kütüphanesinde yer alan *gmdistribution.fit(data,k)* hazır fonksiyonu kullanılmıştır. İki değişkenli veride her bir heterojen değişken için hangi parçalanmanın optimum olduğunu belirlemek amacıyla grafiksel yöntemlerden tahmin edilen sayıda $k = 1,...$ değerleri girilerek oluşturulan karma normal modeller için log-likelihood, AIC ve BIC değerleri hesaplanmış ve hesaplama sonuçlarına göre en uygun parçalanma sayısı belirlenmiştir.

Heterojen Ruspini veri setinde her bir değişkendeki uygun parçalanma sayısını belirlemek için tek değişkenli normal karma dağılım modeli uygulanmıştır. Değişkenlerdeki k bileşen sayısı aralıktaki en küçük değerden başlayıp üst sınıra kadar verilerek en uygun sayı elde edilmiştir.

Tablo 3.4.1. Ruspini veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-363.27	730.55	735.18
2	-359.06	728.12	739.71
3	-359.05	734.11	752.65
4	**-357.44**	**736.88**	**762.37**
5	-357.03	742.88	774.18

Tablo 3.4.2. İki değişkenli veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-397.89	799.78	804.42
2	-382.25	774.51	786.09
3	-370.93	757.86	776.40
4	**-362.50**	**747.01**	**772.50**
5	-360.03	748.24	780.51

X_1 *ve* X_2 değişkenleri için uygun parçalanmanın belirlenmesinde tek değişkenli normal karma modellerden elde edilen log-likelihood, AIC ve BIC değerleri kullanılmıştır. Her iki değişkende log-likelihood değerinin en büyük aynı zamanda AIC ve BIC değerinin en küçük olduğu parçalanma sayısı en uygun parçalanma sayısı olarak belirlenmiştir. Veri setindeki değişkenler için elde edilen parçalanma sayısı grafiksel yöntemlere dayalı tahmin ile örtüşmüş ve $k_1 = 4$, $k_2 = 4$ olarak elde edilmiştir. İki değişken için elde edilen parçalanma sayısı $k = 4$ için, log-likelihood, AIC ve BIC değerlerinde kırılma gözlenmektedir. Bu kırılmada log-likelihood fonksiyon değerinin artmadan azalışa geçişi aynı zamanda AIC ve BIC değerlerinde de azalmadan artmaya geçiş olduğu gözlemlenmiştir.

3.4.2. Ruspini Verisinde K-Ortalamalar Algoritması ile Gözlem Değerlerinin Değişkenlerdeki Parçalanmalara Atanması

İki değişkenli ve her değişkenin dörde parçalandığı veri setindeki değişkenlerin parçalanmalarına düşen gözlemlerin belirlenmesi için k-ortalamalar algoritması kullanılır. Başlangıçta $k = 4$ merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık (3.2.3) denkleminden hesaplanmaktadır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir.

İki değişkenli her değişkenin dörde bölündüğü veride k-ortalamalar algoritması uygulandığında dört ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum en uygun parçalanma sayısını verir. K-ortalamalar

algoritması kullanılarak verideki her bir değişken dörde parçalanmış ve her bir gözlemin düştüğü veri grubu belirlenmiştir. X_1 *ve* X_2 değişkenlerindeki parçalanmalara karşılık gelen alt gruplar ait oldukları değişkene göre isimlendirilmiştir. X_1 değişkenindeki dört parçalanmaya karşılık gelen alt gruplar X_{11}, X_{12}, X_{13} *ve* X_{14} , X_2 değişkenindeki dört parçalanmaya karşılık gelen alt gruplar ise X_{21}, X_{22}, X_{23} *ve* X_{24} olarak belirlenmiştir. $nx2$ tipindeki X veri matrisi $X = [X_1 X_2]$ şeklinde matris formunda gösterilebilir. n elemanlı X_1 değişkenindeki parçalanma $X_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ X_{13} \\ X_{14} \end{bmatrix}$ şeklinde olur. Burada X_{11}, X_{12}, X_{13} *ve* X_{14} sırasıyla n_{11}, n_{12}, n_{13} *ve* n_{14} elemanlıdır. Yani $n = n_{11} + n_{12} + n_{13} + n_{14}$ olarak elde edilir. n elemanlı X_2 değişkenindeki parçalanma $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ X_{23} \\ X_{24} \end{bmatrix}$ şeklinde olur. Burada X_{21}, X_{22}, X_{23} *ve* X_{24} sırasıyla n_{21}, n_{22}, n_{23} *ve* n_{24} elemanlıdır. Yani $n = n_{21} + n_{22} + n_{23} + n_{24}$ olarak elde edilir.

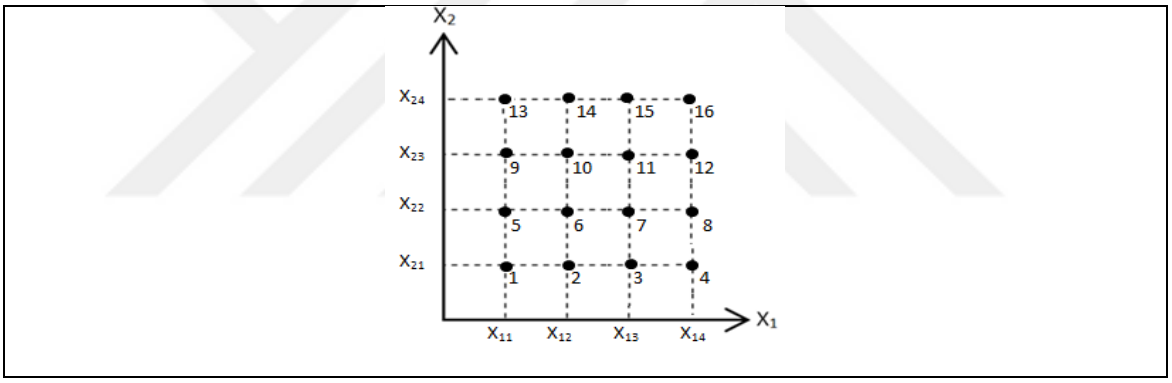
İki değişkenli heterojen veri setinde bulunan her bir değişkenin parçalanmaları ve bu parçalanmalara düşen gözlem sayıları **Tablo 3.4.3** te verilmiştir.

Tablo 3.4.3. İki değişkenli Ruspini veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişken	X_1				X_2			
Değişkenin parçalanmaları	X_{11}	X_{12}	X_{13}	X_{14}	X_{21}	X_{22}	X_{23}	X_{24}
Gözlem sayısı	$n_{11}=12$	$n_{12}=28$	$n_{13}=23$	$n_{14}=12$	$n_{21}=15$	$n_{22}=20$	$n_{23}=17$	$n_{24}=23$
Toplam	$n_1 = 75$				$n_2 = 75$			

3.4.3. Normal Karma Modellerde Değişkenlerdeki Parçalanmalara Dayalı Kümelenme Merkez Sayısının ve Yapısının Belirlenmesi

İki değişkenli ve her değişkenin dörde bölündüğü veri setinde modeldeki kümelenme merkez sayıları değişkenlerdeki parçalanmalara bağlı olarak hesaplanır. Değişkenlerdeki parçalanmaların sayısı kümelenme merkez sayısını, parçalanmalara düşen gözlem değerleri sırasıyla, kümelenme merkezlerinin yerini, şeklini ve büyüklüğünü belirlemektedir [49]. Modeldeki değişkenlerin parçalanmalarının oluşturduğu maksimum ve minimum kümelenme merkezlerinin sayısı C_{max} ve C_{min}; $s=1,...,p$ olmak üzere k_s değerlerine bağlı olarak (3.2.4) ve (3.2.5) denklemleri ile verideki parçalanmalara karşılık gelen en çok kümelenme merkezi sayısı $C_{max}=k_1.k_2=4.4=16$ ve en az kümelenme merkez sayısı $C_{min}=\max\left\{k_1,k_2\right\}=\max\left\{4,4\right\}=4$ olarak elde edilir.



Şekil 3.4.2. İki değişkenli ve her değişkenin dörde parçalandığı normal karma modelde X_1 değişkeninin $X_{11}, X_{12}, X_{13}\ ve\ X_{14}$ ve X_2 değişkeninin $X_{21}, X_{22}, X_{23}\ ve\ X_{24}$ parçalanmalarının oluşturduğu kümelenme merkezleri ve bu merkezlerin numaraları.

$s=1,...,16$ olmak üzere, $X_1\ ve\ X_2$ değişkenlerindeki parçalanmalara karşılık gelen k_s küme merkezleri için parametreler: $i=1,...,16$ olmak üzere merkezin ortalama vektörü μ_i ve varyans-kovaryans matrisi Σ_i, sırasıyla $\mu_i=\begin{bmatrix}\mu_{1x}\\ \mu_{2y}\end{bmatrix}$, $\Sigma_i=\begin{bmatrix}\sigma_{1x}^2 & \rho_i\sigma_{1x}\sigma_{2y}\\ \rho_i\sigma_{2y}\sigma_{1x} & \sigma_{2y}^2\end{bmatrix}$, $x,y=1,...,4$ şeklinde tanımlanır. Burada $\rho_i=Corr\left(X_{1x},X_{2y}\right)$ korelasyon katsayısını göstermek üzere kümelenme merkezindeki veriler, $N\left(x;\mu_i,\Sigma_i\right)$ normal dağılımına sahiptir.

3.4.4. Normal Karma Modellerde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

İki değişkenli ve her değişkenin dörde parçalanması durumunda kümelenme merkezleri için M_{toplam} ile gösterilen oluşabilecek tüm modellerin sayısı değişkenlerdeki parçalanmalara bağlı olarak elde edilir. İki değişkenli normal karma modellerin toplam sayısı (3.2.6) denkleminden $M_{Toplam} = 2^{k_S} - 1 = 2^{k_1 . k_2} - 1 = 2^{4.4} - 1 = 65535$ elde edilir. Oluşabilecek modeller içerisinde $s = 1,...,16$ kümelenme merkezli modeller $\binom{s}{r}$ ifadesi s toplam küme merkezli modeller arasından r alt küme sayılı modellerin sayısını $\sum_{k=1}^{16} \binom{16}{k} - 1 = 65535$ olarak merkez sayılarına göre bulunur.

3.4.5. Normal Karma Modellerde Uygun Aday Model Sayısının Hesaplanması

İki değişkenli normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ve yine parçalanmalara bağlı olarak toplam model sayısı (3.3.3) denkleminden elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli veya uygun modellerin sayısı araştırılmıştır. Bu geçerli modeller **Şekil 3.4.2** de gösterilen merkezler üzerinden kısaca değişkenlerdeki parçalanmaların karşılık geldiği her satır ve her sütunda en az bir kümelenme merkezi bulunacak varsayımı ile anlatılabilir. Varsayıma uyan modellere Uygun Aday Model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanabilir.

Her satır ve sütunda en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı her bir merkezin bulunduğu satır ve sütundaki konumuna göre kombinasyon hesabı ile elde edilebilir. Hesaplama sonucunda iki değişkenli veride her değişkenin dörde bölündüğü durumda, veride olabilecek minimum kümelenme merkez sayısı 4, maksimum kümelenme merkez sayısı 16 için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları **Tablo 3.4.4** te verilmiştir.

Tablo 3.4.4. Normal karma modellerde uygun aday modeller için merkez sayıları, toplam model sayısı, uygun aday model sayısı ve bağımsız parametrelerin sayısı.

Küme merkez sayısı	Toplam model sayısı	Uygun aday model sayısı	Bağımsız parametre sayısı
1	$\binom{16}{1}=16$	-	6
2	$\binom{16}{2}=120$	-	11
3	$\binom{16}{3}=560$	-	17
4	$\binom{16}{4}=1820$	24	23
5	$\binom{16}{5}=4368$	432	29
6	$\binom{16}{6}=8008$	2248	35
7	$\binom{16}{7}=11440$	5776	41
8	$\binom{16}{8}=12870$	9066	47
9	$\binom{16}{9}=11440$	9696	53
10	$\binom{16}{10}=8008$	7480	59
11	$\binom{16}{11}=4368$	4272	65
12	$\binom{16}{12}=1820$	1812	71
13	$\binom{16}{13}=560$	560	77
14	$\binom{16}{14}=120$	120	83
15	$\binom{16}{15}=16$	16	89
16	$\binom{16}{16}=1$	1	95
Toplam	65535	41503	-

Varsayım altındaki uygun aday model sayısının hesaplanmasındaki kombinatorik problem bir matris yapısına karşılık gelmektedir. Varsayımdaki değişkenler ve değişkenlerin parçalanmaları matrisin satır ve sütunlarına karşılık gelmektedir. n_{ij} $i.$ satır ve $j.$ sütundaki kümelenme merkezini göstermek üzere, nxm tipindeki sıfırlardan oluşan bir kare matriste her satır ve her sütunda en az bir kümelenmenin olduğu başka bir ifade ile sıfırdan farklı en az bir elemanın bulunduğu matrislerin sayısı (3.2.8) denkleminden hesaplanır.

İki değişkenli ve her değişkenin dörde bölündüğü veri setinde toplam 65535 model arasından varsayıma uyan 41503 uygun aday model belirlenmiş, varsayıma uymayan modeller çıkarılmıştır. Çalışmada dört kümelenme merkezli yirmi dört uygun aday modelin temsili gösterimi **Tablo 3.4.4** te dört merkezli aday modellerin dizi gösterimi verilmiş ve bu modellerden elde edilen normal karma modeller için hesaplama yapılmıştır. Ayrıca dizi gösterimi ile gösterilen aday modellerden en iyi modelin belirlenmesi için yazılan Matlab kodu **Tablo 3.4.5** te verilmiştir.

Tablo 3.4.5. Ruspini verisindeki değişkenler ve değişkenlerdeki parçalanmalara göre oluşabilecek aday modelleri belirleyip bu modellerin matris dizilimini veren ve bu matris dizilimindeki merkezleri okuyup her bir modelin log-likelihood, AIC ve BIC değerlerini hesaplayan ve en iyi modeli veren matlab kodu.

```
DELIMITER = ' ';
HEADERLINES = 0;
fileToRead1 = 'C:\Users\y1.txt';

% Import the file
rawData1 = importdata(fileToRead1, DELIMITER, HEADERLINES);
% matches that from the Import Wizard.

[~,name] = fileparts(fileToRead1);
newData1.(genvarname(name)) = rawData1;

% Create new variables in the base workspace from those fields.
vars = fieldnames(newData1);

for i = 1:length(vars)
    assignin('base', vars{i}, newData1.(vars{i}));
end

% parcalanma degiskenleri
a=load('rx11.txt');b=load('rx12.txt');c=load('rx13.txt');d=load('rx14.tx
t');e=load('rx21.txt');f=load('rx22.txt');g=load('rx23.txt');
h=load('rx24.txt');

size_p(1) = length(a) + length(e);size_p(2) = length(b) + length(e);
size_p(3) = length(c) + length(e);size_p(4) = length(d) + length(e);
size_p(5) = length(a) + length(f);size_p(6) = length(b) + length(f);
size_p(7) = length(c) + length(f);size_p(8) = length(d) + length(f);
size_p(9) = length(a) + length(g);size_p(10) = length(b) + length(g);
size_p(11) = length(c) + length(g);size_p(12) = length(d) + length(g);
size_p(13) = length(a) + length(h);size_p(14) = length(b) + length(h);
size_p(15) = length(c) + length(h);size_p(16) = length(d) + length(h);
x(1,:)=load('rx1.txt');x(2,:)=load('rx2.txt');size_x = size(x,1);
% ortalama vektörleri

mu(:,1) = [mean(a) mean(e)];mu(:,2) = [mean(b) mean(e)];
mu(:,3) = [mean(c) mean(e)];mu(:,4) = [mean(d) mean(e)];
mu(:,5) = [mean(a) mean(f)];mu(:,6) = [mean(b) mean(f)];
mu(:,7) = [mean(c) mean(f)];mu(:,8) = [mean(d) mean(f)];
mu(:,9) = [mean(a) mean(g)];mu(:,10) = [mean(b) mean(g)];
mu(:,11) = [mean(c) mean(g)];mu(:,12) = [mean(d) mean(g)];
mu(:,13) = [mean(a) mean(h)];mu(:,14) = [mean(b) mean(h)];
mu(:,15) = [mean(c) mean(h)];mu(:,16) = [mean(d) mean(h)];
%Varyans-kuvaryans matrisleri

sigma_p(1,:,:) = [ 39.33 21.909;21.909 51.129];
sigma_p(2,:,:) = [ 74.784 26.625;26.625 51.129];
sigma_p(3,:,:) = [ 85.893 -27.338;-27.338 51.129];
sigma_p(4,:,:) = [ 86.970 4.818; 4.818 51.129];
sigma_p(5,:,:) = [ 39.33 19.303;19.303 101.524];
sigma_p(6,:,:) = [ 74.787 5.726;5.726 101.524];
sigma_p(7,:,:) = [ 85.893 -2.466;-2.466 101.524];
sigma_p(8,:,:) = [ 86.970 12.788;12.788 101.524];
```

```
sigma_p(9,:,:) = [ 39.33 -6.121;-6.121 107.007];
sigma_p(10,:,:) = [ 74.787 -26.257;-26.257 107.007];
sigma_p(11,:,:) = [ 85.893 -12.257;-12.257 107.007];
sigma_p(12,:,:) = [ 86.970 -6.333;-6.333 107.007];
sigma_p(13,:,:) = [ 39.33 -10.697;-10.697 39.732];
sigma_p(14,:,:) = [ 74.787 -19.139;-19.139 39.732];
sigma_p(15,:,:) = [ 85.893 -4.459;-4.459 39.732];
sigma_p(16,:,:) = [ 86.970 -14.758;-14.758 39.732];

Bütün merkezler kullanılarak oluşabilecek aday modellerin dizi
gösterimini veren döngü
% sadece gecerli olan modeller

ihtimal = y1(y1(:,17)>0,:);
for i = 1:1:length(ihtimal(:,1))
    % pdf ve top size hesapla
    top_size = 0;
    clear pdf;
    m = 1;
    for j = 1:1:16
        if(1 == ihtimal(i,j))
            for k=1:length(x(1,:))
                clear sigma_p_temp;
                sigma_p_temp(:,:) = sigma_p(j,:,:);
                pdf(m,k)=(1/(2*pi))...
                    *(1/((det(sigma_p_temp))^(1/2)))...
                    *exp((-0.5)*(x(:,k)-mu(:,j))'...
                    *inv(sigma_p_temp)...
                    *(x(:,k)-mu(:,j)));
            end
            m = m + 1;
            top_size = top_size + size_p(j);
        end
    end
    size_pdf = size(pdf,1);
    % p hesapla
    clear p;
    m = 1;
    for j = 1:1:16
        if(1 == ihtimal(i,j))
            p(m) = size_p(j)/top_size;
            m = m + 1;
        end
    end
    % loglikelihood AIC ve BIC hesapla
    for j = 1:1:length(p(1,:))
        lnl_mul(j,:) = p(j).*pdf(j,:);
    end
    lnl(i) = -(log(size(pdf,2)) * ((size_pdf - 1) + (size_pdf * size_x)
+ (size_pdf * (size_x * ((size_x + 1) / 2)))))+...
        ((size_pdf - 1) + (size_pdf * size_x) + (size_pdf * (size_x *
((size_x + 1) / 2))))*sum(log(sum(lnl_mul)));
    aic(i) = (-2 * lnl(i)) + (2 * ((size_pdf - 1) + (size_pdf * size_x)
+ (size_pdf * (size_x * ((size_x + 1) / 2)))));
    bic(i) = (-2 * lnl(i)) + (log(size(pdf,2)) * ((size_pdf - 1) +
(size_pdf * size_x) + (size_pdf * (size_x * ((size_x + 1) / 2)))));
end
```

Tablo 3.4.6. İki değişkenli ve her değişkenin dörde bölündüğü normal karma modellerde dört kümelenme merkezli uygun aday modeller ve dizi gösterimi.

Kümelenme Merkez Sayısı	Uygun Aday Model Sayısı	Modellerin Dizi Gösterimi 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6
4	24	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
		1 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0
		1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1
		1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0
		1 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0
		1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0
		0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 1
		0 1 0 0 1 0 0 0 0 0 0 1 0 0 1 0
		0 1 0 0 0 0 1 0 1 0 0 0 0 0 0 1
		0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0
		0 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0
		0 1 0 0 0 0 0 1 0 0 1 0 1 0 0 0
		0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 1
		0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 0
		0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 1
		0 0 1 0 0 1 0 0 0 0 0 1 1 0 0 0
		0 0 1 0 0 0 0 1 1 0 0 0 0 1 0 0
		0 0 1 0 0 0 0 1 0 1 0 0 1 0 0 0
		0 0 0 1 1 0 0 0 0 1 0 0 0 0 1 0
		0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0
		0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 0
		0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0
		0 0 0 1 0 0 1 0 1 0 0 0 0 1 0 0
		0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0

3.4.6. Normal Karma Modellerde Uygun Aday Modellerin Oluşturulması ve Parametre Tahminleri

İki değişkenli normal karma dağılım modelleri için bileşen ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir. Karma modeldeki her bir merkezin olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisi modeldeki olasılık yoğunluk fonksiyonunu elde etmek için kullanılır. **Tablo 3.4.6** da dört kümelenme merkezli uygun aday modellerin temsili gösterimleri modeldeki merkezlerin numaralarına göre dizi gösterimi şeklinde verilmiştir [75]. Modelin temsili gösteriminde kullanılan “0” ve “1” elemanları modeldeki karşılık gelen merkezde yığılmanın olup olmadığını göstermektedir. Eğer uygun aday modelde herhangi bir merkezde kümelenme varsa “1” yoksa “0” ile gösterilmiştir.

3.4.6.1. Normal Karma Modeller ve Parametre Tahminleri

Ruspini veri setindeki iki değişkenin her birinin dörde bölünmesi durumunda oluşacak kümelenme merkezleri **Şekil 3.4.2** de gösterilmiştir. **Tablo 3.4.4** te veri setindeki parçalanmalara karşılık gelen ve varsayımlara uyan uygun aday modellerin sayısı verilmiştir. Toplam 65535 model arasından 41503 uygun aday model içerisinden dört bileşenli normal karma modele karşılık gelen modellerin normal bileşen yoğunluk fonksiyonları, $u = 1,...,41503$ ve $i = 1,...,16$ için

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{k} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.4.1}$$

olarak elde edilir. Normal dağılımların bileşen yoğunluğunun olasılık ağırlıkları

$$\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{k} \pi i} \tag{3.4.2}$$

eşitliğinden elde edilir. Ortalama vektörleri $x, y = 1,...,4$ olmak üzere

$$\mu_i^{(u)} = \begin{bmatrix} \mu_{1x}^{(u)} \\ \mu_{2y}^{(u)} \end{bmatrix} \tag{3.4.3}$$

ve varyans-kovaryans matrisi

$$\Sigma_i^{(u)} = \begin{bmatrix} \left(\sigma_{1x}^{(u)}\right)^2 & \rho_{1x,2y}^{(u)}\sigma_{1x}^{(u)}\sigma_{2y}^{(u)} \\ \rho_{2y,1x}^{(u)}\sigma_{2y}^{(u)}\sigma_{1x}^{(u)} & \left(\sigma_{2y}^{(u)}\right)^2 \end{bmatrix} \tag{3.4.4}$$

olarak ifade edilir.

Dört bileşenli 24 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 1,...,24$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. Beş bileşenli 432 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 25,...,456$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. Altı bileşenli 2248 uygun aday model (3.4.1) denklemindeki normal karma modellerin

yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 457,...,2704$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. Yedi bileşenli 5776 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 2705,...,8480$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. Sekiz bileşenli 9066 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 8481,...,17546$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. Dokuz bileşenli 9696 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 17547,...,27242$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. On bileşenli 7480 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 27243,...,34722$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. On bir bileşenli 4272 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 34723,...,38994$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. On iki bileşenli 1812 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 38995,...,40806$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. On üç bileşenli 560 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 40807,...,41366$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. On dört bileşenli 120 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 41367,...,41486$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır.

On beş bileşenli 16 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modellerdir. Karma modellerdeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 41487,...,41502$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır. On altı bileşenli 1 uygun aday model (3.4.1) denklemindeki normal karma modellerin yapısına uyan modeldir. Karma modeldeki $\pi_i^{(u)}$, $\mu_i^{(u)}$ ve $\Sigma_i^{(u)}$ parametreleri $u = 41503$ olmak üzere (3.4.2), (3.4.3) ve (3.4.4) denklemlerindeki yapıdadır.

$i = 1,...,g$ olmak üzere normal karma modellerdeki yoğunluk fonksiyonlarının π_i, μ_i ve Σ_i parametreleri tahmin edilmelidir.

Tablo 3.4.4 te sayıları verilen uygun aday modellerin (3.4.1) denklemindeki $u = 1,...,41503$ için karma modellerin yapısına uyan modellerin, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \dfrac{n_i}{\sum_{i=1}^{k} n_i}$, $x, y = 1,...,4$ olmak üzere, ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}$, normal dağılımların bileşen yoğunluk fonksiyonlarının bilinmeyen parametrelerinin tahmini olarak ifade edilir.

Tablo 3.4.7. Ruspini verisinin değişkenlerdeki parçalanmalara karşılık gelen alt gruplarının oluşturduğu 16 merkezin veri setinden elde edilen μ_i, Σ_i ve ρ_i parametre tahminleri.

Merkezler	**Ortalama vektörü**	**Varyans-kovaryans matrisi**	**Korelasyon katsayısı**
1	$\boldsymbol{\mu}_1 = \begin{bmatrix} 13{,}6667 \\ 19{,}0625 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 39{,}333 & 21{,}909 \\ 21{,}909 & 51{,}129 \end{bmatrix}$	$\rho_1 = 0{,}676$
2	$\boldsymbol{\mu}_2 = \begin{bmatrix} 38{,}25 \\ 19{,}0625 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 74{,}787 & 26{,}625 \\ 26{,}625 & 51{,}129 \end{bmatrix}$	$\rho_2 = 0{,}817$
3	$\boldsymbol{\mu}_3 = \begin{bmatrix} 70{,}43 \\ 19{,}0625 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 85{,}893 & -27{,}338 \\ -27{,}338 & 51{,}129 \end{bmatrix}$	$\rho_3 = -0{,}494$
4	$\boldsymbol{\mu}_4 = \begin{bmatrix} 106{,}333 \\ 19{,}0625 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 86{,}970 & 4{,}818 \\ 4{,}818 & 51{,}129 \end{bmatrix}$	$\rho_4 = 0{,}100$
5	$\boldsymbol{\mu}_5 = \begin{bmatrix} 13{,}6667 \\ 64{,}9500 \end{bmatrix}$	$\boldsymbol{\Sigma}_5 = \begin{bmatrix} 39{,}333 & 19{,}303 \\ 19{,}303 & 101{,}524 \end{bmatrix}$	$\rho_5 = 0{,}281$
6	$\boldsymbol{\mu}_6 = \begin{bmatrix} 38{,}2500 \\ 64{,}9500 \end{bmatrix}$	$\boldsymbol{\Sigma}_6 = \begin{bmatrix} 74{,}787 & 5{,}726 \\ 5{,}726 & 101{,}524 \end{bmatrix}$	$\rho_6 = 0{,}106$
7	$\boldsymbol{\mu}_7 = \begin{bmatrix} 70{,}4348 \\ 64{,}9500 \end{bmatrix}$	$\boldsymbol{\Sigma}_7 = \begin{bmatrix} 85{,}893 & -2{,}466 \\ -2{,}466 & 101{,}524 \end{bmatrix}$	$\rho_7 = -0{,}027$
8	$\boldsymbol{\mu}_8 = \begin{bmatrix} 106{,}3333 \\ 64{,}9500 \end{bmatrix}$	$\boldsymbol{\Sigma}_8 = \begin{bmatrix} 86{,}970 & 12{,}788 \\ 12{,}788 & 101{,}524 \end{bmatrix}$	$\rho_8 = 0{,}125$
9	$\boldsymbol{\mu}_9 = \begin{bmatrix} 13{,}6667 \\ 114{,}4118 \end{bmatrix}$	$\boldsymbol{\Sigma}_9 = \begin{bmatrix} 39{,}333 & -6{,}121 \\ -6{,}121 & 107{,}007 \end{bmatrix}$	$\rho_9 = -0{,}169$
10	$\boldsymbol{\mu}_{10} = \begin{bmatrix} 38{,}250 \\ 114{,}4118 \end{bmatrix}$	$\boldsymbol{\Sigma}_{10} = \begin{bmatrix} 74{,}787 & -26{,}257 \\ -26{,}257 & 107{,}007 \end{bmatrix}$	$\rho_{10} = -0{,}575$
11	$\boldsymbol{\mu}_{11} = \begin{bmatrix} 70{,}4348 \\ 114{,}4118 \end{bmatrix}$	$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} 85{,}893 & -12{,}257 \\ -12{,}257 & 107{,}007 \end{bmatrix}$	$\rho_{11} = -0{,}153$
12	$\boldsymbol{\mu}_{12} = \begin{bmatrix} 106{,}3333 \\ 114{,}4118 \end{bmatrix}$	$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} 86{,}970 & -6{,}333 \\ -6{,}333 & 107{,}007 \end{bmatrix}$	$\rho_{12} = -0{,}118$
13	$\boldsymbol{\mu}_{13} = \begin{bmatrix} 13{,}6667 \\ 146{,}2727 \end{bmatrix}$	$\boldsymbol{\Sigma}_{13} = \begin{bmatrix} 39{,}333 & -10{,}697 \\ -10{,}697 & 39{,}732 \end{bmatrix}$	$\rho_{13} = -0{,}342$
14	$\boldsymbol{\mu}_{14} = \begin{bmatrix} 38{,}2500 \\ 146{,}2727 \end{bmatrix}$	$\boldsymbol{\Sigma}_{14} = \begin{bmatrix} 74{,}787 & -19{,}139 \\ -19{,}139 & 39{,}732 \end{bmatrix}$	$\rho_{14} = -0{,}500$
15	$\boldsymbol{\mu}_{15} = \begin{bmatrix} 70{,}4348 \\ 146{,}2727 \end{bmatrix}$	$\boldsymbol{\Sigma}_{15} = \begin{bmatrix} 85{,}893 & -4{,}459 \\ -4{,}459 & 39{,}732 \end{bmatrix}$	$\rho_{15} = -0{,}075$
16	$\boldsymbol{\mu}_{16} = \begin{bmatrix} 106{,}3333 \\ 146{,}2727 \end{bmatrix}$	$\boldsymbol{\Sigma}_{16} = \begin{bmatrix} 86{,}970 & -14{,}758 \\ -14{,}758 & 39{,}732 \end{bmatrix}$	$\rho_{16} = -0{,}318$

3.4.7. Normal Karma Modellerin Log-Likelihood Fonksiyonu, AIC ve BIC Değerlerinin Hesaplanması

Modele dayalı kümeleme yapmak için iki değişkenli ve her değişkenin dörde parçalandığı veri setindeki uygun aday modellerin normal karma modelleri belirlenmiştir. Normal karma modeller arasından en iyi modelin seçimi için her bir uygun aday modelin çok değişkenli normal karma dağılımlardan log-likelihood fonksiyonu, AIC ve BIC değerleri elde edilmiştir. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu ve logaritması alınmış likelihood fonksiyonu (3.2.13) ve (3.2.14) denkleminden elde edilir. Çok değişkenli normal karma modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kriteri (BIC) ve Akaike bilgi kriteri (AIC) (3.2.15) ve (3.2.17) denklemlerinden elde edilir [74].

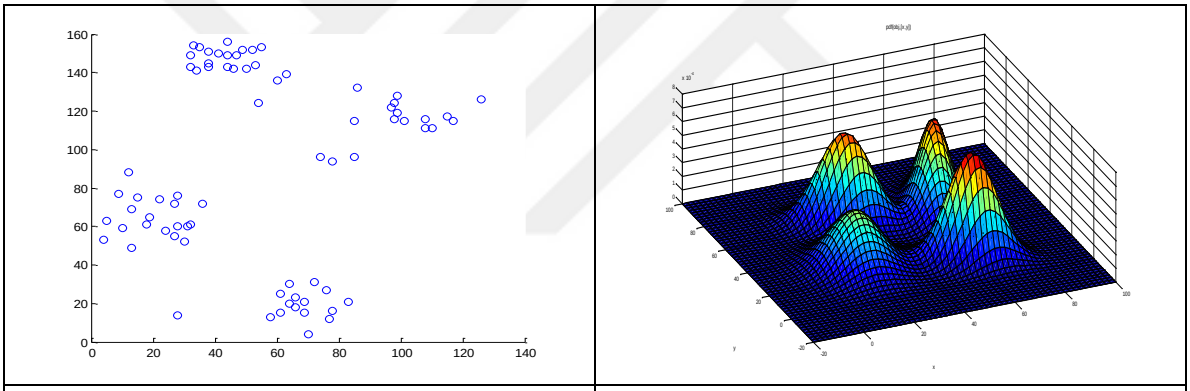
Ruspini veri setinde genetik algoritmanın adımları uygulanarak elde edilen 41503 aday modelden elde edilen bilgi kriterlerine göre dört merkezli modeller en iyi model olarak belirlenmiştir. **Tablo 3.4.8** de veri setinden elde edilen bilgi kriterlerine göre en iyi modelin belirlenmesinde kullanılan dört merkezli aday modellerin değerleri verilmiştir.

Tablo 3.4.8. Ruspini verisinde normal karma dağılımların modele dayalı kümelenmesi için log-l, AIC ve BIC değerleri.

Model no:	$\log L(\Psi)$	$AIC = -2\ln L(\Psi) + 2d$	$BIC = -2\ln L(\Psi) + d\log n$	d
1. Model	-27614,3	55274,53	55327,83	23
2. Model	-24208,5	48462,93	48516,23	23
3. Model	-30565,8	61177,55	61230,85	23
4. Model	-24858,6	49763,29	49816,59	23
5. Model	-32082,2	64210,5	64263,8	23
6. Model	-31089,6	62225,3	62278,6	23
7. Model	-23826,4	47698,72	47752,03	23
8. Model	-20420,9	40887,78	40941,08	23
9. Model	-32660,7	65367,37	65420,67	23
10. Model	-29857,6	59761,19	59814,49	23
11. Model	-28672,4	57390,78	57444,08	23
12. Model	-33702,6	67451,12	67504,43	23
13. Model	-22301,1	44648,13	44701,43	23
14. Model	**-15699,7**	**31445,31**	**31498,61**	23
15. Model	-27386,3	54818,51	54871,81	23
16. Model	-21865	43776,08	43829,38	23
17. Model	-25258,4	50562,75	50616,05	23
18 Model	-30361,5	60769,07	60822,37	23
19. Model	-24523,2	49092,35	49145,65	23
20. Model	-19849,1	39744,23	39797,54	23
21. Model	-25614,1	51274,26	51327,57	23
22. Model	-23282,1	46610,25	46663,55	23
23. Model	-28275,8	56597,61	56650,91	23
24. Model	-33105,5	66257,05	66310,35	23

3.4.8. Normal Karma Modellerin Modele Dayalı Kümelenmesi için En İyi Model Seçimi

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin log-l, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. Ruspini veri setinde normal karma modellerden uygun aday modellerin log-l, AIC ve BIC değerleri hesaplanmış ve **Tablo 3.4.8** de verilmiştir. Uygun aday modeller arasından modele dayalı kümelemede en iyi model log-likelihood değeri en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir. Matlab programında yazılan kod ile 41503 uygun aday model modelin tamamının log-likelihood, AIC ve BIC değerleri hesaplanmış ve aralarından en iyi model dört kümelenme merkezli on dördüncü "0010100000010100" dizi gösterimine sahip olan model olarak belirlenmiştir.



(a) (b)

Şekil 3.4.3. İki değişkenli Ruspini verisinde 41503 aday normal karma model arasından en iyi modelin (a) saçılım grafiği (b) yüzey grafiği.

3.5. Değişken Veri Segmentasyonu Kullanarak Model Seçimine Dayalı Karma Model Kümeleme

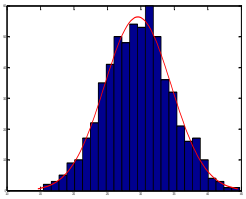
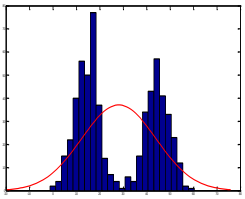
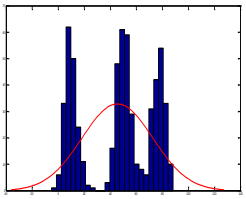
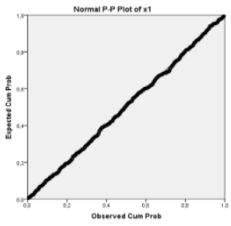
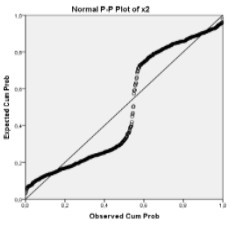
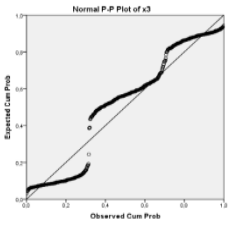
Çok değişkenli normal dağılımların karmalarından elde edilen verilerde modele dayalı kümeleme ve oluşabilecek modeller arasından en iyi modelin seçilmesine dayalı genetik algoritma verilerin parçalanması, parçalanan veriden oluşabilecek kümelenme sayılarının ve yapısının belirlenmesi, verinin yapısına göre aday model sayısının belirlenmesi ve aday modeller arasından en iyi karma modelin bilgi kriterlerine göre belirlenmesini sağlamaktadır. Bu kısımdaki çalışmada üç değişkenli sentetik veri setinde her bir değişkendeki farklı parçalanma sayısına göre, verinin yapısı ve aday modellerin

belirlenmesi ve aday modeller arasından veri yapısına uyan en iyi modelin varlığı araştırılmıştır. Çok değişkenli veri setinde homojen değişkenlerin elenmesi boyut indirgeme olarak adlandırılır. Homojen değişkenlerin elenip heterojen değişkenlere dayalı kümeleme analizinde değişken seçimi yapılmıştır. Değişkenlerdeki heterojenliğin test edildiği değişken veri segmentasyonu metodunda homojen yapıdaki değişkenlerin verideki küme yapılarını ve sayısını etkilemediği belirlenmiştir.

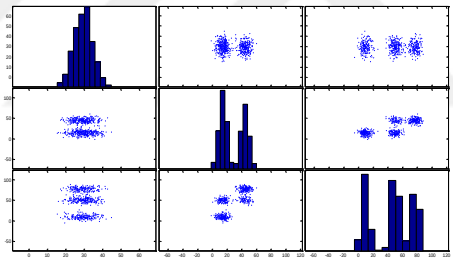
3.5.1. Heterojen Değişkenlerin Parçalanmasına Dayalı Kümelenme Merkez Sayılarının ve Yapılarının Belirlenmesi

Bu çalışmada daha önceki yapılan çalışmalardan farklı olarak verideki değişkenlerin heterojenliği test edilir. Değişkenlerde heterojenlik varsa değişken içerdiği alt grup sayısı kadar parçalanır. Verideki değişkenlerdeki parçalanmalar verideki kümelenme merkez sayılarını verir. $s = 1,...,p$ olmak üzere verideki her bir $\boldsymbol{X}_S$ değişkeninin heterojenliği incelensin. Verideki değişkenler homojen ya da heterojen yapıda olabilir. Heterojen verideki değişkenlerin yapısı verideki kümelenmeyi belirlemektedir [41]. p-boyutlu veride $\boldsymbol{X}_S$ değişkeninin yapısına göre parçalanmalarının sayısı $k_S \geq 1$ olsun. $k_S = 1$ durumu değişkenin homojen yapıda ve $k_S > 1$ durumu değişkenin heterojen yapıda olması anlamına gelir. Verideki her bir değişken için k_S değeri k-ortalamalar algoritması gibi bir ayrıştırma algoritması kullanılarak elde edilebilir. p-boyutlu heterojen verideki s. değişken ve j. gözlem arasındaki ilişki x_{sj} olarak gösterilebilir. Burada x_{sj}, p-boyutlu heterojen verideki s. değişkeninin j. gözlemini ifade eder.

Simülasyon ile oluşturulan sentetik veri setinde X_2 ve X_3 değişkenlerindeki heterojen ve X_1 değişkenindeki homojen yapıyı gösteren histogram grafikleri **Şekil 3.5.1-2** de gösterilmiştir.

X_1 Değişkeni	X_2 Değişkeni	X_3 Değişkeni
		
(a)	(b)	(c)
		
(d)	(e)	(f)

Şekil 3.5.1. a, b ve c grafikleri sırasıyla X_1 değişkeninde homojen X_2 ve X_3 değişkenindeki heterojen yapıyı gösteren histogram ve d, e ve f P-P grafiklerini göstermektedir



Şekil 3.5.2. X_1, X_2 ve X_3 değişkenlerinin saçılım ve histogram grafiklerinin matris gösterimi.

Şekil 3.5.1-2 de çok değişkenli verideki X_1, X_2 ve X_3 değişkenlerindeki parçalanmalar sırasıyla, $k_1 = 1$, $k_2 = 2$ ve $k_3 = 3$ olarak tahmin edilmiştir. Değişkenlerdeki parçalanmalara göre X_1 değişkeni homojen, X_2 ve X_3 değişkenlerinin heterojen yapıda olduğu belirlenmiştir.

3.5.1.1. Değişken Veri Parçalanmalarının Tek Değişkenli Normal Karma Modellere Dayalı Olarak Belirlenmesi

Çok değişkenli veride değişkenlerdeki parçalanmanın belirlenmesi tek değişkenli normal karma dağılım modeli (3.2.1) ve (3.2.2) denklemlerindeki gibi hesaplanır. Değişkendeki uygun parçalanma sayısına göre tek değişkenli normal karma dağılımın olasılık ağırlıkları π , ortalama vektörü μ ve varyansı σ^2 parametreleri hesaplanır. Her bir değişkendeki parçalanmanın ortaya çıkarılması için tek değişkenli normal karma modelin log-likelihood, AIC ve BIC değerleri modelin parametrelerinden parçalanma sayısına bağlı olarak hesaplanır. Tek değişkenli normal karma dağılımlardan elde edilen log-likelihood değerinin en büyük, AIC ve BIC değerlerinin en küçük olduğu parçalanma sayısı en uygun parçalanma sayısı olarak belirlenir. Hesaplamalarda MATLAB programının istatistik paketi kütüphanesinde yer alan *gmdistribution.fit(data,k)* fonksiyonu kullanılmıştır. Çok değişkenli veride heterojen yapının ortaya çıkarılması için her bir değişkende optimum segment (parçalanma) sayısını belirlemek amacıyla grafiksel yöntemlerden tahmin edilen sayıda $k = 1,...$ değerleri algoritmaya girilerek oluşturulan karma normal modeller için log-likelihood, AIC ve BIC değerleri hesaplanmış ve hesaplama sonuçlarına göre en uygun parçalanma sayısı belirlenmiştir.

Tablo 3.5.1. İki değişkenli veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-l	AIC	BIC
k=1	**-1813.6**	**3631.3**	**3640.1**
k=2	-1813.6	3637.2	3659.2
k=3	-1813.4	3642.8	3678.0
k=4	-1813.4	3648.8	3697.1

Tablo 3.5.2. İki değişkenli veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-l	AIC	BIC
k=1	-2510.7	5025.4	5034.2
k=2	**-2236.5**	**4483.0**	**4505.0**
k=3	-2236.3	4488.7	4523.9
k=4	-2234.6	4491.0	4539.7

Tablo 3.5.3. İki değişkenli veri setinde X_3 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-l	AIC	BIC
k=1	-2833.1	5670.2	5679.0
k=2	-2607.9	5225.7	5247.7
k=3	**-2454.0**	**4924.0**	**4959.1**
k=4	-2453.1	4929.1	4977.5

Hesaplama sonuçlarına göre X_1, X_2 ve X_3 değişkenleri için log-likelihood değerinin maksimum aynı zamanda AIC ve BIC değerlerinin minimum olduğu en uygun parçalanma sayısını veren karma normal dağılım modeli X_1 homojen değişkeni için $k_1 = 1$, X_2 ve X_3 heterojen değişkenleri için parçalanma sayıları sırasıyla $k_2 = 2$ ve $k_3 = 3$ olarak belirlenmiştir.

3.5.2. K-Ortalamalar Algoritması ile Verideki Gözlem Değerlerinin Değişkenlerdeki Segmentlere Atanması

Üç değişkenli ve değişkenlerin sırasıyla bire, ikiye ve üçe parçalandığı veri setindeki değişkenlerin parçalanmalarına düşen gözlemlerin belirlenmesi için k-ortalamalar algoritması kullanılır. Başlangıçta her bir değişkenin parçalanma sayısı kadar k merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık (3.2.3) denkleminden hesaplanır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir.

Üç değişkenli ve değişkenlerin sırasıyla bir, iki ve üçe parçalandığı veride k-ortalamalar algoritması uygulandığında her bir değişken için sırasıyla bir, iki ve üç ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum en uygun parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki her bir

değişken bir, iki ve üçe parçalanmış ve her bir gözlemin düştüğü veri grubu belirlenmiştir.

Üç değişkenli veride k-ortalamalar algoritması uygulandığında X_1 değişkeni için tek küme, X_2 değişkeni için iki ayrı küme, X_3 değişkeni için üç ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum optimum parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki değişkenler sırasıyla bir, iki ve üçe parçalanmış ve her bir gözlemin düştüğü veri grubu belirlenmiştir. Simülasyon ile oluşturulan üç değişkenli veride her değişkendeki parçalanmalar k-ortalamalar algoritması uygulanarak X_1 , X_2 ve X_3 değişkenleri için sırasıyla $k_1 = 1$, $k_2 = 2$ ve $k_3 = 3$ olarak elde edilir. X_1 değişkeninde parçalanma olmadığından verinin kendisi alınır, X_2 değişkenindeki parçalanmalar X_{21}, X_{22} ve X_3 değişkeninde X_{31} , X_{32} ve X_{33} olarak gösterilsin. $n \times 3$ tipindeki veri matrisi $X = \lfloor X_1 \quad X_2 \quad X_3 \rfloor$ şeklinde gösterilebilir. n_1 elemanlı X_1 değişkeni $X_1 = \left[X_1 \right]$ şeklinde gösterilir, burada $n_1 = n_1$ olarak değişkendeki eleman sayısını gösterir. n_2 elemanlı X_2 değişkeni $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}$ şeklinde gösterilir, burada X_{21} ve X_{22} sırasıyla n_{21} ve n_{22} elemanlı olup $n_2 = n_{21} + n_{22}$ olarak elde edilir. n_3 elemanlı X_3 değişkeni $X_3 = \begin{bmatrix} X_{31} \\ X_{32} \\ X_{33} \end{bmatrix}$ şeklinde gösterilir, burada X_{31}, X_{32} ve X_{33} sırasıyla n_{31}, n_{32} ve n_{33} elemanlı olup $n = n_{31} + n_{32} + n_{33}$ olarak elde edilir.

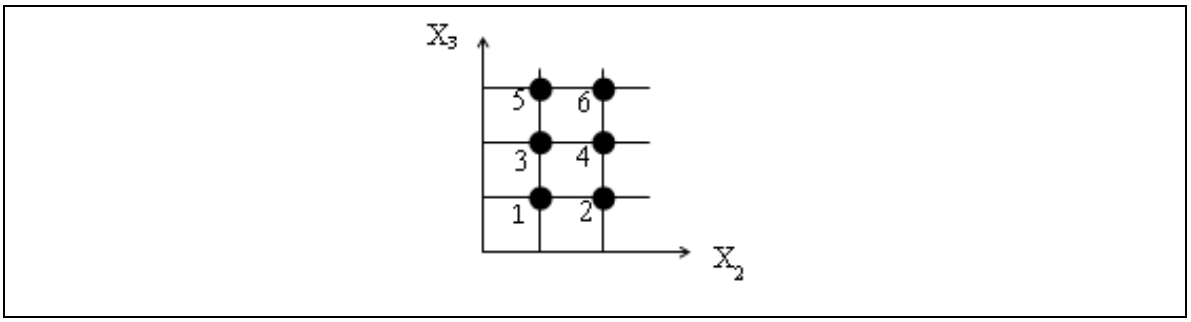
Çok değişkenli veri setindeki değişkenler ve değişkenlere k-means algoritması uygulandıktan sonra değişkenlerdeki segmentlere düşen gözlem sayıları **Tablo 3.5.4** te verilmiştir.

Tablo 3.5.4. Çok değişkenli veri setindeki değişkenler, değişkenlerin segmentleri ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişkenler	X_1	X_2		X_3		
Değişkendeki seğmenler	X_1	X_{21}	X_{22}	X_{31}	X_{32}	X_{33}
Gözlem sayısı	$n_1 = 600$	$n_{21} = 329$	$n_{22} = 271$	$n_{31} = 179$	$n_{32} = 231$	$n_{33} = 190$
Toplam	600	600		600		

3.5.3. Normal Karma Modellerde Değişkenlerdeki Parçalanmalara Dayalı Kümelenme Merkez Sayısının Ve Yapısının Belirlenmesi

Çok değişkenli veride değişkenlerdeki segmentler sırasıyla, $k_1 = 1$, $k_2 = 2$ ve $k_3 = 3$ olarak **Tablo 3.5.1-3** te elde edilmiştir. Normal karma modeldeki kümelenme merkez sayıları değişkenlerdeki segmentlere dayalı olarak hesaplanır. Değişkenlerdeki segmentlerin sayısı kümelenme merkez sayısını, segmentlere düşen gözlem değerleri sırasıyla, kümelenme merkezlerinin yerini, şeklini ve büyüklüğünü belirlemektedir [49]. Modeldeki değişkenlerin parçalanmalarının oluşturduğu maksimum ve minimum kümelenme merkezlerinin sayısı C_{max} ve $C_{min}, s = 1,..., p$ olmak üzere k_s değerlerine bağlı olarak (3.2.4) ve (3.2.5) denklemlerinden verideki parçalanmalara karşılık gelen en çok kümelenme merkezi sayısı $C_{max} = k_1.k_2.k_3 = 1.2.3 = 6$ ve en az kümelenme merkez sayısı $C_{min} = \max\{k_1, k_2, k_3\} = \max\{1,2,3\} = 3$ olarak elde edilir.



Şekil 3.5.3. X_1 değişkeninin homojen yapısı, X_2 ve X_3 değişkenleri üzerine izdüşüm alınarak elde edilen iki değişkenli grafik ve oluşabilecek kümelenme merkezlerinin numaraları.

Şekil 3.5.3'de birinci kümelenme merkezi; n_1, n_{21} ve n_{31} elemanlı sırasıyla X_1, X_{21} ve X_{31} parçalanmalarının yer aldığı $\left[X_1 X_{21} X_{31}\right]$ veri matrisi; ikinci kümelenme merkezi; n_1, n_{22} ve n_{31} elemanlı sırasıyla X_1, X_{22} ve X_{31} parçalanmalarının yer aldığı $\left[X_1 X_{22} X_{31}\right]$ veri matrisi; üçüncü kümelenme merkezi; n_1, n_{21} ve n_{32} elemanlı sırasıyla X_1, X_{21} ve X_{32} parçalanmalarının yer aldığı $\left[X_1 X_{21} X_{32}\right]$ veri matrisi; dördüncü kümelenme merkezi; n_1, n_{22} ve n_{32} elemanlı sırasıyla X_1, X_{22} ve X_{32} parçalanmalarının yer aldığı $\left[X_1 X_{22} X_{32}\right]$ veri matrisi; beşinci kümelenme merkezi; n_1, n_{21} ve n_{33} elemanlı sırasıyla X_1, X_{21} ve X_{33} parçalanmalarının yer aldığı $\left[X_1 X_{21} X_{33}\right]$ veri matrisi; altıncı kümelenme merkezi; n_1, n_{22} ve n_{33} elemanlı sırasıyla X_1, X_{22} ve X_{33} parçalanmalarının yer aldığı $\left[X_1 X_{22} X_{33}\right]$ veri matrisi tarafından oluşturulur.

3.5.4. Normal Karma Modellerde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

Çok değişkenli veride X_1 değişkeninde parçalanmanın bulunmadığı yani $k_1 = 1$ ve X_2 ve X_3 değişkenlerinin sırasıyla $k_2 = 2, k_3 = 3$ parçalanmalarına dayalı olarak kümelenme merkezleri için M_{Toplam} ile gösterilen normal dağılımların karma modelleriyle oluşturulabilecek toplam model sayısı (3.2.6) denklemi kullanılarak

$M_{Toplam} = 2^{1.2.3} - 1 = 2^6 - 1 = 63$ olarak elde edilir. Burada çıkarılan 1 model varsayıma uymayan kümelenme merkezi bulunmayan boş modeldir. Bu veri seti için oluşturulabilecek model sayısı $\binom{n}{r}$ ifadesi n toplam kümelenme merkez sayısı ve r modeldeki kümelenme merkez sayısını göstermek üzere $r = 3,...6$ kümelenme merkezli,

oluşturulabilecek normal dağılımların karma modellerinin sayısı sırasıyla $\binom{6}{1}=6$, $\binom{6}{2}=15$, $\binom{6}{3}=20$, $\binom{6}{4}=15$, $\binom{6}{5}=6$ ve $\binom{6}{6}=1$ olarak elde edilir.

Toplam normal dağılımların karma modelleriyle oluşturulabilecek model sayısı $\sum_{k=1}^{6}\binom{6}{1}-1=63$ olacak şekilde merkez sayılarına bağlı olarak elde edilir[41].

3.5.5. Normal Karma Modellerde Uygun Aday Model Sayısının Hesaplanması

Üç değişkenli normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ve yine parçalanmalara bağlı olarak toplam model sayısı elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli veya uygun modellerin sayısı araştırılmıştır. Bu geçerli modeller **Şekil 3.5.3** te gösterilen merkezler üzerinden kısaca her boyutta en az bir kümelenme merkezi bulunacak varsayımı ile anlatılabilir. Varsayıma uyan modellere Uygun Aday Model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanabilir.

Her boyutta en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı her bir merkezin bulunduğu boyuttaki konumuna göre kombinasyon hesabı ile elde edilebilir. Hesaplama sonucunda üç değişkenli veride olabilecek minimum kümelenme merkez sayısı 3, maksimum kümelenme merkez sayısı 6 için merkez sayıları, merkezlerin konumu, modellerin sayısı için bağıntı ve uygun aday model sayıları **Tablo 3.5.5** te verilmiştir.

Tablo 3.5.5. Çok değişkenli veride değişkenlerde $k_1 = 1$, $k_2 = 2$ ve $k_3 = 3$ parçalanmalarına karşılık gelen merkezlerin konumu, model sayısı için bağıntı ve uygun aday model sayısı

Merkez Sayısı	**Merkezlerin Konumu**	**Modellerin Sayısı İçin Bağıntı**	**Uygun Aday Model Sayısı**
Üç merkezli	2 1 şeklinde sütunlara dağılan	$\binom{3}{2}\binom{3}{1}2! = 6$	6
Dört merkezli	3 1 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{1}2! = 6$	12
	2 2 şeklinde sütunlara dağılan	$\binom{3}{2}\binom{2}{1}1! = 6$	
Beş merkezli	3 2 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{2}2! = 6$	6
Altı merkezli	3 3 şeklinde sütunlara dağılan	$\binom{3}{3}\binom{3}{3}1! = 1$	1
Toplam model sayısı		$2^{1.2.3} - 1 = 63$	
Toplam Uygun model sayısı		0+0+6+12+6+1=25	

Varsayım altındaki uygun aday model sayısının hesaplanmasındaki problem bir matris yapısına karşılık gelmektedir. Varsayımdaki değişkenler ve değişkenlerin parçalanmaları matrisin satır ve sütunlarına karşılık gelmektedir. n_{ij} $i.$ satır ve $j.$ sütundaki kümelenme merkezini göstermek üzere, nxm tipindeki sıfırlardan oluşan bir kare matriste her satır ve her sütunda en az bir kümelenmenin olduğu başka bir ifade ile sıfırdan farklı en az bir elemanın bulunduğu

$$f(n,m,s,k) = \sum_{i=0}^{n} (-1)^i \binom{n}{i} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \sum_{t=0}^{s} (-1)^t \binom{s}{t} \binom{(n-i)(m-j)(s-t)}{k} \quad (3.5.1)$$

eşitliği ile elde edilebilir. Burada $n, m \; ve \; s$ sırasıyla $X_1, X_2 \; ve \; X_3$ değişkenlerindeki parçalanma sayısını göstermek üzere $i, j, t = 1, ..., 6$ parçalanmalardaki kümelenme merkez sayısını ve k modeldeki istenilen kümelenme merkez sayısını göstermektedir. Birinci değişkende veri yapısı homojenlik gösterdiğinden parçalanma sayısı 1 olarak elde edildiğinden (3.5.1) numaralı denklemdeki hesaplama iki boyutlu denklemlerle yapılan hesaplamalar ile aynıdır. Homojen yapıdaki değişkenlerin modellerin oluşturulmasında ve hesaplamalarda en iyi modelin belirlenmesinde etkisi olmadığı gösterilmiştir. Modellerin yapısı ve aday modellerin sayısı veri setindeki heterojen değişkenlerden kaynaklandığı için değişken veri parçalanmadaki bileşenlerin heterojenliğinin test

edilmesi kümelemede en iyi modelin bulunmasında önemli yer tutmaktadır. Değişkenlerdeki segmentasyona dayalı aday model sayısı ve modellerin dizi gösterimleri **Tablo 3.5.6** da verilmiştir.

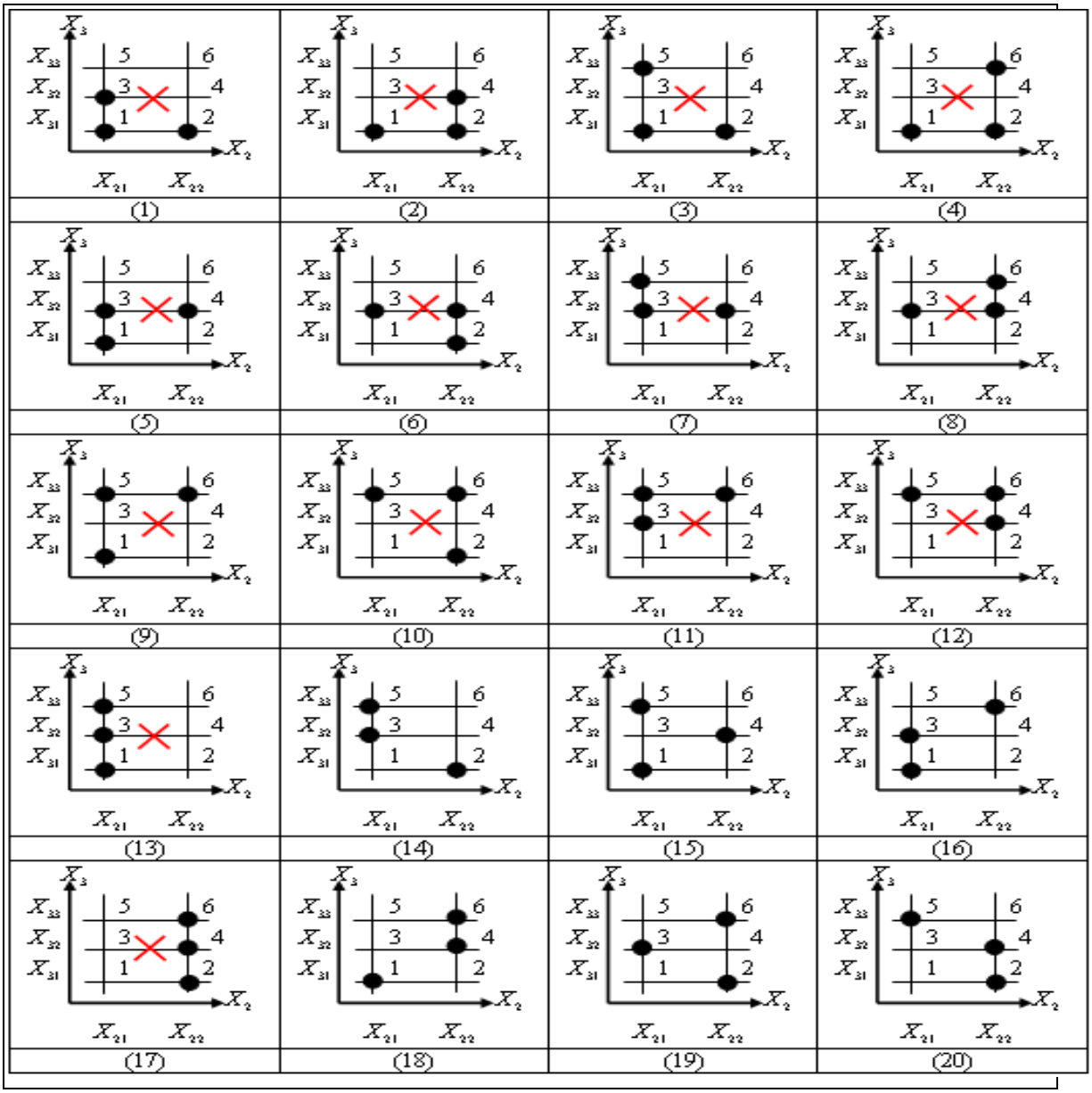
Tablo 3.5.6. Aday modellerin dizi gösterimleri, kümelenmelere göre model sayıları, modellerin uygun olup olmadıkları ve uygun model sayıları.

Kümelenme Merkez Sayısı	Durum sayısı	Uygun Durum Sayısı	Aday Modellerin Dizi Gösterimi
			123456
Bir Kümelenme Merkezli Modeller	6	0	-
İki Kümelenme Merkezli Modeller	15	0	-
Üç Kümelenme Merkezli Modeller	20	6	101001
			100110
			011010
			010110
			011001
			100101
Dört Kümelenme Merkezli Modeller	15	12	111010
			101110
			101011
			110101
			011101
			011110
			011011
			110110
			100111
			101101
			111001
Beş Kümelenme Merkezli Modeller	6	6	011111
			101111
			110111
			111011
			111101
			111110
Altı Kümelenme Merkezli Model	1	1	111111

Üç değişkenli veride X_1 değişkeninde parçalanma olmadığından X_2 ve X_3 değişkenlerindeki segmentlere dayalı aday normal karma modeller oluşturulmuştur. **Şekil 3.5.3** te X_2 ve X_3 değişkenlerinin düzlem üzerine izdüşümü ile oluşan iki değişkenli veride her bir değişkenin sırasıyla ikiye ve üçe parçalanması durumunda veride normal

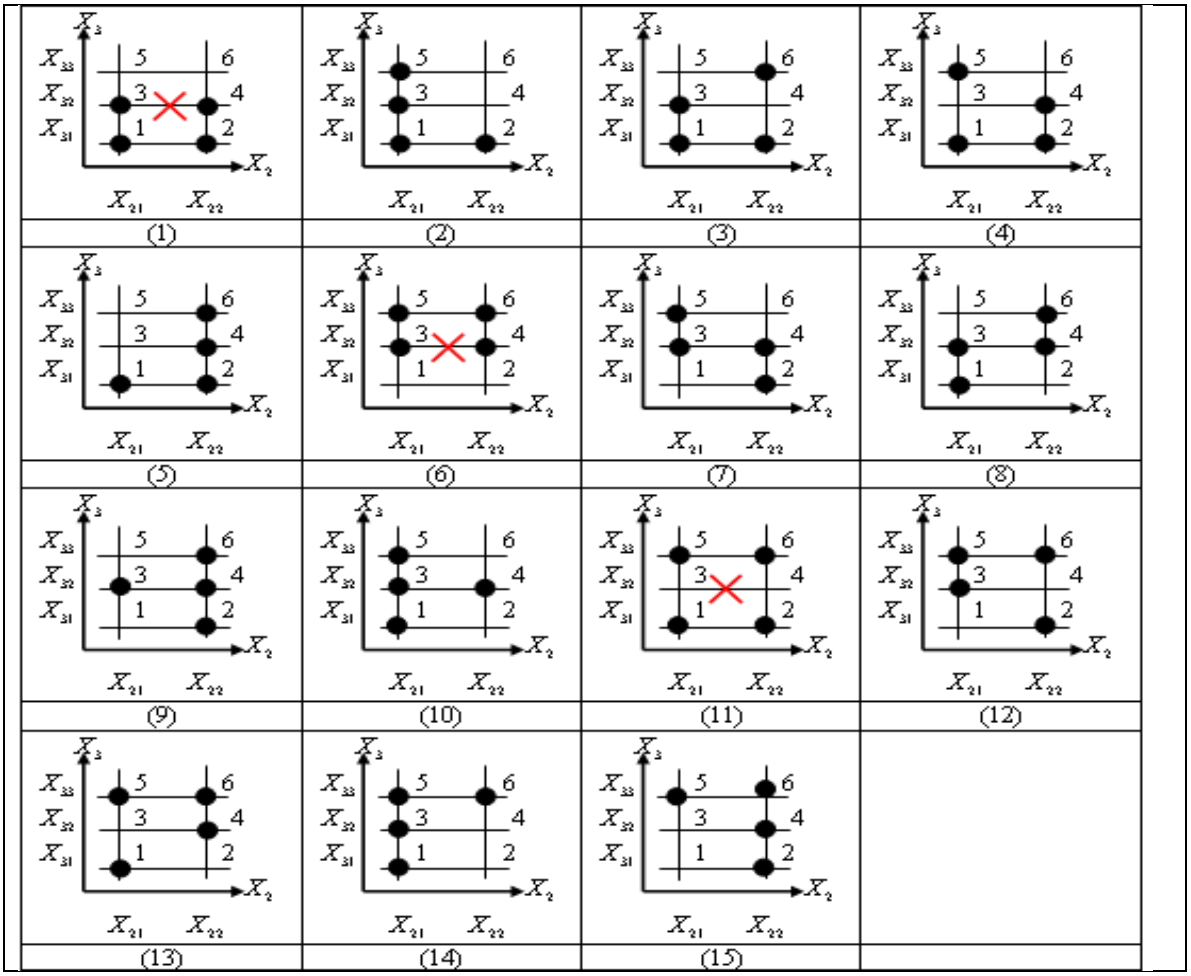
dağılımların karma modelleriyle oluşturulabilecek ve $M_{Muhtemel}$ ile gösterilen aday modellerin grafikleri **Şekil 3.5.4 -9** da gösterilmiştir.

Üç kümelenme merkezli modellerden varsayıma uymayan modeller **Şekil 3.5.4** te X işareti ile belirtilen modellerdir.

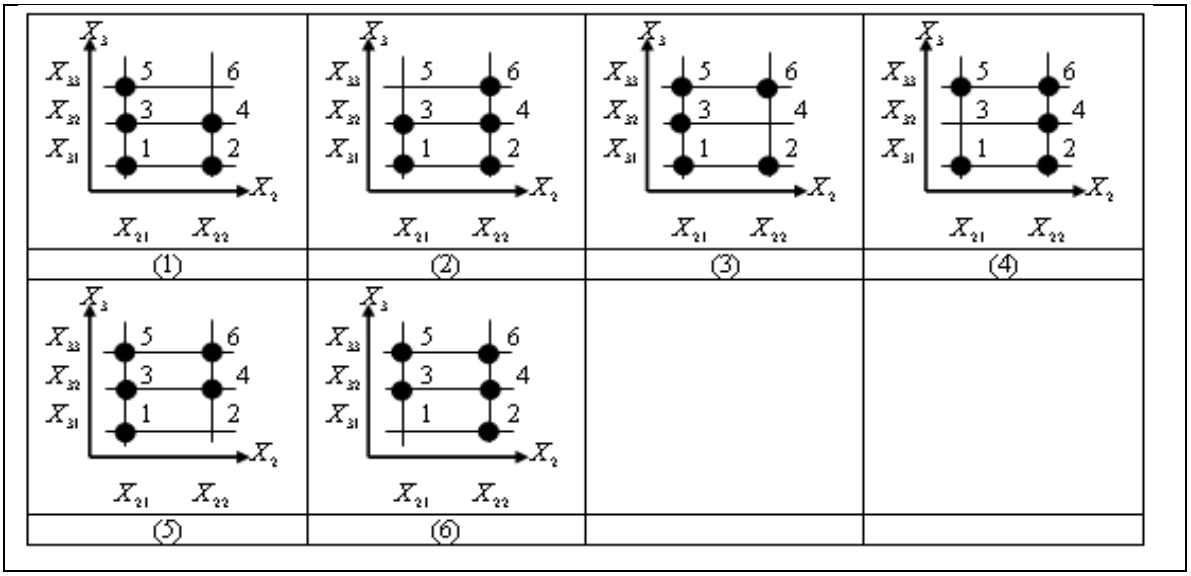


Şekil 3.5.4. X_1 homojen, X_2 *ve* X_3 değişkenlerinde sırasıyla iki ve üç parçalanmanın bulunduğu karma normal modeller arasından üç kümelenme merkezli modeller.

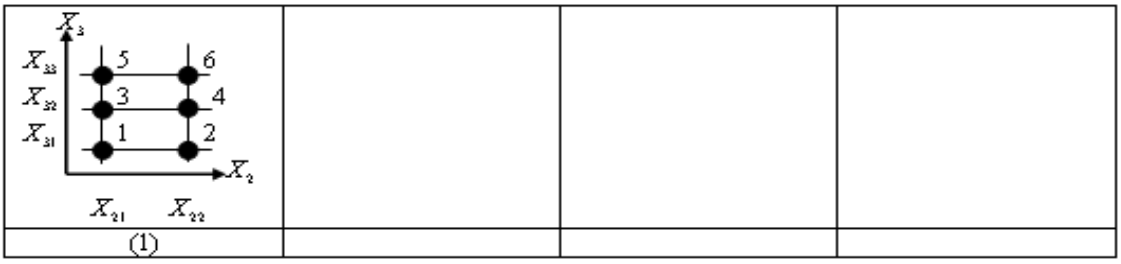
Dört kümelenme merkezli modellerden varsayıma uymayan modeller **Şekil 3.5.5** te X işareti ile belirtilen modellerdir.



Şekil 3.5.5. X_1 homojen, X_2 *ve* X_3 değişkenlerinde sırasıyla iki ve üç parçalanmanın bulunduğu karma normal modeller arasından dört kümelenme merkezli modeller.



Şekil 3.5.6. X_1 homojen, X_2 *ve* X_3 değişkenlerinde sırasıyla iki ve üç parçalanmanın bulunduğu karma normal modeller arasından beş kümelenme merkezli modeller.



Şekil 3.5.7. X_1 homojen, X_2 *ve* X_3 değişkenlerinde sırasıyla iki ve üç parçalanmanın bulunduğu karma normal modeller arasından altı kümelenme merkezli model.

Böylece üç değişkenli veride değişkenlerin bir, iki ve üçe parçalanması durumunda simülasyon ile oluşturulan sentetik veri seti için toplamda normal dağılımların karma modelleriyle oluşturulabilecek uygun aday model sayısı **Şekil 3.5.4-7** deki grafiklere göre 25 model elde edilmiştir.

3.5.6. Normal Karma Modellerde Değişken Veri Segmentasyonuna Dayalı Aday Modellerin Oluşturulması ve Parametre Tahminleri

Çok değişkenli veride normal karma dağılım modelleri için bileşen ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir. Karma modeldeki her bir merkezin olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisi modeldeki olasılık yoğunluk fonksiyonunu elde etmek için kullanılır. **Tablo 3.5.6** da uygun aday modellerin temsili gösterimleri modeldeki merkezlerin numaralarına göre dizi gösterimi şeklinde verilmiştir. Modelin dizi gösteriminde kullanılan "0" ve "1" elemanları modeldeki karşılık gelen merkezde kümelenme merkezinin olup olmadığını göstermektedir. Eğer uygun aday modelde herhangi bir merkezde kümelenme varsa "1" yoksa "0" ile gösterilmiştir.

3.5.6.1. Normal Karma Modeller ve Parametre Tahminleri

Çok değişkenli veride değişkenlerdeki parçalanmalara karşılım gelen kümelenme merkezlerini temsil eden ortalama vektörleri, varyans-kovaryans matrisleri ve korelasyon katsayıları verinin parçalanmalarının ortalamaları ve standart sapmalarından elde edilmiştir. Üç değişkenli verinin ortalama vektörü,

$\mu_i = \begin{bmatrix} \mu_1 \\ \mu_{2x} \\ \mu_{3y} \end{bmatrix}$ ile temsil edilir. $i = 1,2,...,6$ olmak üzere, aday modellerin varyans-kovaryans matrisleri $\Sigma_i = \begin{bmatrix} {\sigma_1}^2 & \rho_i\sigma_1\sigma_{2x} & \rho_j\sigma_1\sigma_{3y} \\ \rho_i\sigma_{2x}\sigma_1 & {\sigma_{2x}}^2 & \rho_k\sigma_{2x}\sigma_{3y} \\ \rho_j\sigma_{3y}\sigma_1 & \rho_k\sigma_{3y}\sigma_{2x} & {\sigma_{3y}}^2 \end{bmatrix}$ ile temsil edilir.

Burada $x = 1,2$ ve $y = 1,2,3$ olacak şekilde sırasıyla $X_1 ve X_2$ değişkenlerindeki parçalanmayı, $\rho_i = \boldsymbol{Corr(X}_1, \boldsymbol{X}_{2x}, \boldsymbol{X}_{3y})$ Pearson korelasyon katsayısını göstermektedir.

Bu merkezindeki veriler, $N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ normal dağılımına sahip olsun. Üç değişkenli veride değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ile normal karma modeller oluşturulurken olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri kullanılmıştır. Değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin varyans-kovaryans matrislerindeki bileşenler arasındaki ilişkinin yönünü ve derecesini veren Pearson korelasyon katsayısı üç bileşenli merkez yapısı için $3x3$ tipinde matris içindeki bileşenlerde altı adet bulunmaktadır. Varyans-kovaryans matrisi simetrik matris yapısında olduğundan her bir varyans-kovaryans matrisinde üç farklı korelasyon katsayısı bulunur. Modellerdeki farklı korelasyon katsayılarının sayısı değişkenlerdeki parçalanmalar kullanılarak

$$k_1k_2 + k_1k_3 + k_2k_3 = 1.2 + 1.3 + 2.3 = 11 \tag{3.5.2}$$

şeklinde elde edilir [76].

Toplam 25 uygun aday model içerisinden üç bileşenli normal karma modele karşılık gelen modeller $u = 1,...,6$ ve $i = 1,2,3$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{3} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.5.3}$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{3} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

altı normal karma model bulunur.

Dört bileşenli normal karma modele karşılık gelen modeller $u = 7,...,18$ ve $i = 1,...,4$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{4} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.5.4}$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \dfrac{\pi_i}{\sum_{i=1}^{4} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

on iki normal karma model bulunur.

Beş bileşenli normal karma modele karşılık gelen modeller $u = 19,...,24$ ve $i = 1,...,5$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{5} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.5.5}$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{5} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5)$$
$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$
$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$
$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$
$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$
$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

altı normal karma model bulunur.

Altı bileşenli normal karma modele karşılık gelen modeller $u = 25$ ve $i = 1,...,6$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{6} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.5.6}$$

Şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{6} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4) + \pi_5 N(\mathbf{x};\boldsymbol{\mu}_5,\boldsymbol{\Sigma}_5) + \pi_6 N(\mathbf{x};\boldsymbol{\mu}_6,\boldsymbol{\Sigma}_6)$$

bir normal karma model bulunur.

(3.5.3)-(3.5.6) denklemlerindeki $u = 1,...,25$ için karma modellerin yapısına uyan modellerin olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{k} n_i}$, $x = 1,2$ ve $y = 1,2,3$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_1^{(u)} \\ \bar{x}_{2x}^{(u)} \\ \bar{x}_{3y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri

$$\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_1^{(u)}\right)^2 & r_{1,2x}^{(u)} s_1^{(u)} s_{2x}^{(u)} & r_{1,3y}^{(u)} s_1^{(u)} s_{3y}^{(u)} \\ r_{2x,1}^{(u)} s_{2x}^{(u)} s_1^{(u)} & \left(s_{2x}^{(u)}\right)^2 & r_{2x,3y}^{(u)} s_{2x}^{(u)} s_{3y}^{(u)} \\ r_{3y,1}^{(u)} s_{3y}^{(u)} s_1^{(u)} & r_{3y,2x}^{(u)} s_{3y}^{(u)} s_{2x}^{(u)} & \left(s_{3y}^{(u)}\right)^2 \end{bmatrix}, i = 1,2,3 \text{ için normal bileşen}$$

yoğunluk fonksiyonlarının bilinmeyen parametreleridir. Veri setindeki değişkenlerin gözlem değerlerinden elde edilen parametre tahminleri **Tablo 3.5.7-8** de verilmiştir.

Tablo 3.5.7. Değişkenlerdeki parçalanmalar arasındaki ilişkinin yönü ve derecesini gösteren Pearson korelasyon katsayısı

$\rho_1 = -0,077$	$\rho_2 = 0,011$	$\rho_3 = -0,059$	$\rho_4 = 0,012$
$\rho_5 = -0,005$	$\rho_6 = -0,050$	$\rho_7 = 0,209$	$\rho_8 = 0,014$
$\rho_9 = -0,016$	$\rho_{10} = 0,033$	$\rho_{11} = -0,076$	

Tablo 3.5.8. Değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin ortalama vektörü ve varyans-kovaryans parametreleri tahmini.

$\mu_1 = \begin{bmatrix} 29.5538 \\ 14.4367 \\ 77.1946 \end{bmatrix}, \Sigma_1 = \begin{bmatrix} 27.762 & -1.962 & -0.393 \\ -1.962 & 25.120 & 0.824 \\ -0.393 & 0.824 & 23.227 \end{bmatrix}$	$\mu_2 = \begin{bmatrix} 29.5538 \\ 44.6978 \\ 77.1946 \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 27.762 & 0.305 & -0.393 \\ 0.305 & 26.431 & -1.897 \\ -0.393 & -1.897 & 23.227 \end{bmatrix}$
$\mu_3 = \begin{bmatrix} 29.5538 \\ 14.4367 \\ 50.2912 \end{bmatrix}, \Sigma_3 = \begin{bmatrix} 27.762 & -1.962 & -1.291 \\ -1.962 & 25.120 & 5.221 \\ -1.291 & 5.221 & 24.940 \end{bmatrix}$	$\mu_4 = \begin{bmatrix} 29.5538 \\ 44.6978 \\ 50.2912 \end{bmatrix}, \Sigma_4 = \begin{bmatrix} 27.762 & 0.305 & -1.291 \\ 0.305 & 26.431 & 0.362 \\ -1.291 & 0.362 & 24.940 \end{bmatrix}$
$\mu_5 = \begin{bmatrix} 29.5538 \\ 14.4367 \\ 9.6469 \end{bmatrix}, \Sigma_5 = \begin{bmatrix} 27.762 & -1.962 & 0.289 \\ -1.962 & 25.120 & -1.444 \\ 0.289 & -1.444 & 23.294 \end{bmatrix}$	$\mu_6 = \begin{bmatrix} 29.5538 \\ 44.6978 \\ 9.6469 \end{bmatrix}, \Sigma_6 = \begin{bmatrix} 27.762 & 0.305 & 0.289 \\ 0.305 & 26.431 & -0.116 \\ 0.289 & -0.116 & 23.294 \end{bmatrix}$

3.5.7. En İyi Kümelenme Yapısını Veren Normal Dağılımların Karma Modelleri İçin Bilgi Kriterlerinin Hesaplanması.

Modele dayalı kümeleme yapmak için iki değişkenli ve değişkenlerin bir, iki ve üçe parçalandığı veri setindeki uygun aday modellerin normal karma modelleri belirlenmiştir. Normal karma modeller arasından en iyi modelin seçimi için her bir uygun aday modelin çok değişkenli normal karma dağılımlardan log-likelihood, AIC ve BIC değerleri elde edilmiştir. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu,

$i = 1, 2, ...n \;\; ve \;\; j = 1, 2, ...g$ olmak üzere $f\left(x_i; \theta_j\right)$ karma olasılık yoğunluk fonksiyonunun likelihood fonksiyonu ve logaritması alınmış likelihood fonksiyonu (3.2.13) ve (3.2.14) denklemleri kullanılarak elde edilir.

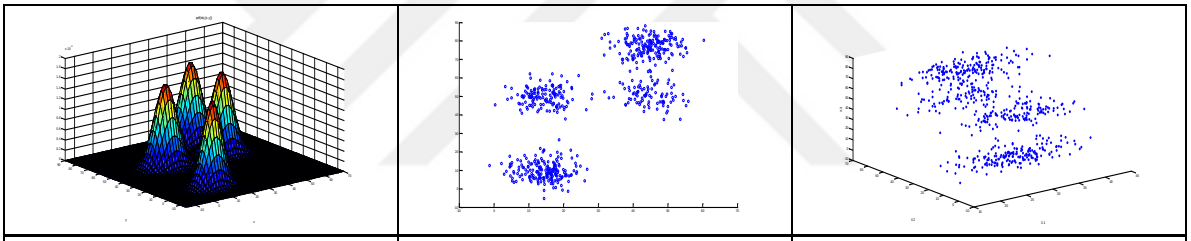
Çok değişkenli normal karma modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kriteri (BIC) ve Akaike bilgi kriteri (AIC) (3.2.15) ve (3.2.17) denklemleri kullanılarak elde edilir.

Tablo 3.5.9. Ruspini verisinde normal karma dağılımların modele dayalı kümelenmesi için log-l, AIC ve BIC değerleri.

No	Merkez Model no	log-l	AIC	BIC	Matris gösterimi
-	Tek merkezli	-	-	-	-
-	İki merkezli	-	-	-	-
1	Üç Merkezli 1.model	-7548.5	15155	15097	101001
2	Üç Merkezli 2.model	-10401	20860	20802	100110
3	Üç Merkezli 3.model	-14581	29220	29162	011010
4	Üç Merkezli 4.model	-13765	27588	27530	010110
5	Üç Merkezli 5.model	-10732	21521	21463	011001
6	Üç Merkezli 6.model	-8671.6	17401	17343	100101
7	Dört Merkezli 1.model	-11460	22999	22921	111010
8	Dört Merkezli 2.model	-8792.4	17663	17585	101110
9	Dört Merkezli 3.model	-7784.2	15646	15568	101011
10	Dört Merkezli 4.model	-8904	17886	17808	110101
11	Dört Merkezli 5.model	-9749.5	19577	19499	011101
12	Dört Merkezli 6.model	-11675	23428	13350	011110
13	Dört Merkezli 7.model	**-6523.9**	**13126**	**13048**	**011011**
14	Dört Merkezli 8.model	-7737.4	15553	15475	110110
15	Dört Merkezli 9.model	-8368.2	16814	16736	100111
16	Dört Merkezli 10.model	-10631	21340	21262	101101
17	Dört Merkezli 11.model	-12019	24116	24038	111001
18	Dört Merkezli 12.model	-10970	22018	21940	011111
19	Beş Merkezli 1.model	-8977.9	18054	17956	101111
20	Beş Merkezli 2.model	-6710.9	13520	13422	110111
21	Beş Merkezli 3.model	-7935.7	15969	15871	111011
22	Beş Merkezli 4.model	-8556.4	17211	17113	111101
23	Beş Merkezli 5.model	-6725.5	13529	13431	111110
24	Beş Merkezli 6.model	-9943.6	19985	19887	111111
25	Altı Merkezli model	-6885.4	13889	13771	101001

3.5.8. Üç Değişkenli Veride Normal Karma Modellerin Modele Dayalı Kümelenmesi için En İyi Modelin Seçimi

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin log-likelihood, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. Çok değişkenli veri setlerinde normal karma modellerden uygun aday modellerin log-likelihood, AIC ve BIC değerleri hesaplanmış ve **Tablo 3.5.9** da verilmiştir. Uygun aday modeller arasından modele dayalı kümelemede en iyi model log-likelihood değeri en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir. Matlab programında yazılan kod ile 25 uygun aday model modelin tamamının log-likelihood, AIC ve BIC değerleri hesaplanmış aralarından en iyi model dört kümelenme merkezli on üçüncü model olarak belirlenmiştir. Model sayısı çok fazla olduğundan sadece dört kümelenme merkezli modeller tablo halinde gösterilmiştir.



(a) (b) (c)

Şekil 3.5.8. Üç değişkenli ve değişkenlerin sırasıyla bir, iki ve üçe bölündüğü normal karma dağılımlar arasından bilgi kriterlerine göre belirlenen en iyi modelin (a) yoğunluk fonksiyonunun yüzey grafiği (b) iki boyutlu saçılım grafiği (c) üç boyutlu saçılım grafiği

3.6. Çok değişkenli Veride Modele Dayalı Karma Kümeleme ile En İyi Model Seçimi

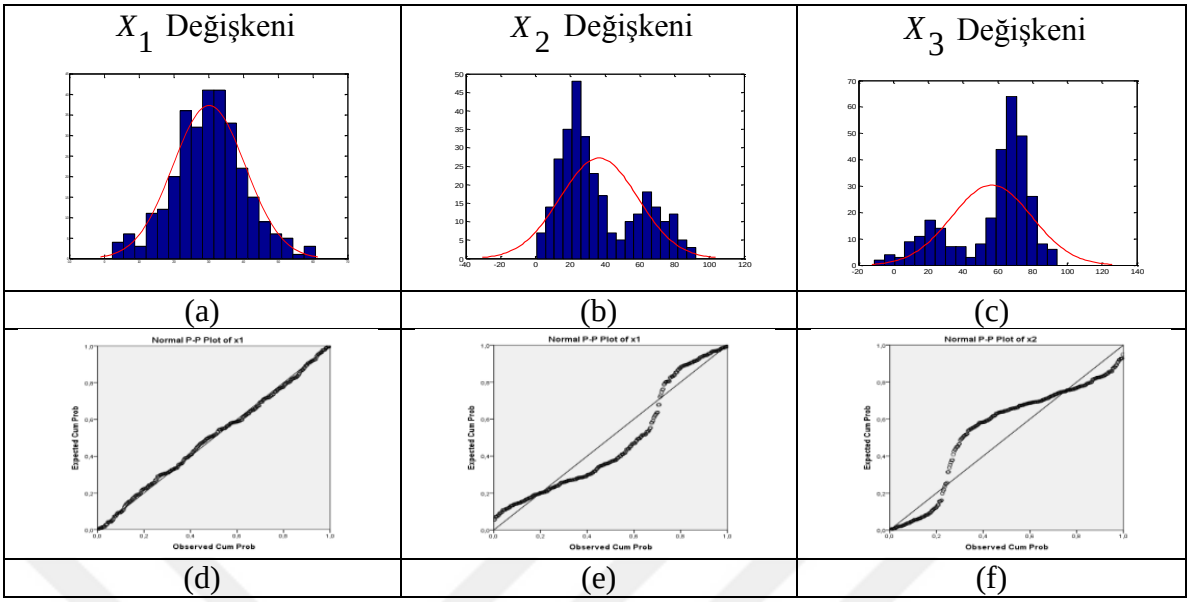
Çok değişkenli veride verinin yapısına göre meydana gelebilecek küme yapıları modele dayalı olarak belirlenebilir [11]. Bu kısımdaki çalışmada üç değişkenli sentetik veri setinde değişkenlerdeki farklı sayıdaki parçalanmalardan kaynaklanan veri yapısına göre oluşabilecek kümelenme sayısı ve bu kümelenme sayılarından oluşan aday modeller arasından en iyi karma modelin belirlenmesi araştırılmıştır.

3.6.1. Çok Değişkenli Veride En İyi Modelin Seçimi için Değişkenlerdeki Parçalanmaların Grafiksel ve Normal Karma Modellerin Kullanılması

Çok değişkenli veride her bir değişkendeki anlamlı alt grupların ortaya çıkarılması oluşturulacak karma modellerin belirlenmesi için kullanılmaktadır [41]. Çok değişkenli veride her bir değişkende homojen veya heterojen yapıdaki değişkenlerin belirlenmesi ve buna bağlı olarak toplam model sayılarının elde edilmesi için parçalanma sayıları belirlenmiştir. Değişkenlerdeki parçalanmalara bağlı oluşabilecek model sayıları ve parçalanmalardan elde edilen parametre değerleri ile modellerin yapıları belirlenmektedir. Çok değişkenli veride değişkenlerdeki parçalanma sayıları grafiksel yöntemlerden ve tek değişkenli normal karma modellerden faydalanarak tahmin edilebilir.

3.6.1.1. Çok Değişkenli Veride Değişkenlerdeki Parçalanmalarla En İyi Modelin Seçimi İçin Aday Modellerin Ortaya Çıkarılmasında Grafiksel Yöntemler

Bu çalışmada daha önceki yapılan çalışmalardan farklı olarak verideki değişkenlerin heterojenliği test edilir. Değişkenlerde heterojenlik varsa değişken içerdiği alt grup sayısı kadar parçalanır. Verideki değişkenlerdeki parçalanmalar verideki kümelenme merkez sayılarını verir. $s = 1,\ldots,p$ olmak üzere verideki her bir $\boldsymbol{X}_S$ değişkeninin heterojenliği incelensin. Verideki değişkenler homojen ya da heterojen yapıda olabilir. Heterojen verideki değişkenlerin yapısı verideki kümelenmeyi belirlemektedir [41]. p-boyutlu veride $\boldsymbol{X}_S$ değişkeninin yapısına göre parçalanmalarının sayısı $k_S \geq 1$ olsun. $k_S = 1$ olması değişkenin homojen yapıda ve $k_S > 1$ olması değişkenin heterojen yapıda olması anlamına gelir. Verideki her bir değişken için k_S değeri k-ortalamalar algoritması gibi bir ayrıştırma algoritması kullanılarak elde edilebilir. p-boyutlu heterojen verideki s. değişken ve j. gözlem arasındaki ilişki x_{Sj} olarak gösterilebilir. Burada x_{Sj}, p-boyutlu heterojen verideki s. değişkeninin j. gözlemini ifade eder.



Şekil 3.6.1. a,b,c grafikleri sırasıyla X_1 değişkeninde homojen yapıyı X_2 ve X_3 değişkenindeki heterojen yapıyı gösteren histogram grafikleri. d,e ve f grafikleri sırasıyla X_1 değişkeninde homojen yapıyı, X_2 ve X_3 değişkenindeki heterojen yapıyı gösteren P-P grafikleri.

X_1, X_2 *ve* X_3 değişkenleri için uygun parçalanmanın belirlenmesinde histogram grafiğindeki tepelenme sayısı ve P-P grafiğinin normal doğrusu veya $y = x$ doğrusu ile verideki gözlem eğrisinin kesiştikleri nokta sayısına bakılarak değişkendeki parçalanma sayısı tahmin edilebilir. **Şekil 3.6.1** de çok değişkenli verideki X_1, X_2 ve X_3 değişkenlerindeki parçalanmalar sırasıyla, $k_1 = 1$, $k_2 = 2$ ve $k_3 = 2$ olarak tahmin edilmiştir.

3.6.1.2. Çok Değişkenli Veride Değişkenlerdeki Parçalanmalarının Tek Değişkenli Normal Karma Modellere Dayalı Belirlenmesi

İki değişkenli veride değişkenlerdeki parçalanmanın belirlenmesi tek değişkenli normal karma dağılım modeli (3.2.1) ve (3.2.2) denklemelerindeki gibi hesaplanır. Değişkendeki uygun parçalanma sayısınıa göre tek değişkenli normal karma dağılımın olasılık ağırlıkları π , ortalama vektörü μ ve varyans-kovaryans matrisi σ^2 parametreleri hesaplanır. Her bir değişkendeki parçalnmanın ortaya çıkarılması için tek değişkenli normal karma dağılımın log-likelihood, AIC ve BIC değerleri modelin parametrelerinden parçalanma sayısına bağlı olarak hesaplanır. Tek değişkenli normal karma dağılımlardan

elde edilen log-likelihood değerinin en büyük, AIC ve BIC değerlerinin en küçük olduğu parçalanma sayısı en uygun sayı olarak belirlenir Hesaplamalarda MATLAB programının istatistik paketi kütüphanesinde yer alan *gmdistribution.fit(data,k)* fonksiyonu kullanılmıştır. Çok değişkenli veride heterojen yapının ortaya çıkarılması için her bir değişkende optimum parçalanma sayısını belirlemek amacıyla grafiksel yöntemlerden tahmin edilen sayıda $k=1,...$ değerleri algoritmaya girilerek oluşturulan karma normal modeller için log-likelihood, AIC ve BIC değerleri hesaplanmış ve hesaplama sonuçlarına göre en uygun parçalanma sayısı belirlenmiştir.

Tablo 3.6.1. İki değişkenli veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

k	$Log-L$	AIC	BIC
$k=1$	**1128,4**	**2260,8**	**2268,2**
$k=2$	1128,1	2266,2	2284,8
$k=3$	1126,7	2269,3	2298,9

Tablo 3.6.2. İki değişkenli veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

k	$Log-L$	AIC	BIC
$k=1$	-1327,4	2658,9	2666,3
$k=2$	**-1318,0**	**2646,0**	**2664,5**
$k=3$	-1317,1	2650,2	2679,8

Tablo 3.6.3. İki değişkenli veri setinde X_3 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

k	$Log-L$	AIC	BIC
$k=1$	-1365.8	2735.6	2743.0
$k=2$	**-1297.1**	**2604.2**	**2622.8**
$k=3$	-1297.1	2610.2	2639.9

Hesaplama sonuçlarına göre X_1, X_2 ve X_3 heterojen değişkenleri için log-likelihood değerinin maksimum aynı zamanda AIC ve BIC değerlerinin minimum olduğu en uygun

parçalanma sayısını veren karma normal dağılım modeli X_1 değişkeni için $k_1 = 1$, X_2 değişkeni için $k_2 = 2$ ve X_3 değişkeni için $k_3 = 2$ olarak belirlenmiştir.

3.6.2. Çok Değişkenli Veride K-Ortalamalar Algoritması ile Gözlem Değerlerinin Değişkenlerdeki Parçalanmalara Atanması

Üç değişkenli ve değişkenlerin sırasıyla bire, ikiye ve ikiye parçalandığı veri setindeki değişkenlerin parçalanmalarına düşen gözlemlerin belirlenmesi için k-ortalamalar algoritması kullanılır. Başlangıçta her bir değişkenin parçalanma sayısı kadar k merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık (3.2.3) denkleminden hesaplanır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir.

Üç değişkenli ve değişkenlerin sırasıyla bire, ikiye ve ikiye parçalandığı veride k-ortalamalar algoritması uygulandığında her bir değişken için sırasıyla bir, iki ve iki ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum en uygun parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki her bir değişken bir, iki ve ikiye parçalanmış ve her bir gözlemin düştüğü veri grubu **Tablo 3.6.4** te verilmiştir.

Üç değişkenli veride k-ortalamalar algoritması uygulandığında X_1 değişkeni için tek küme, X_2 değişkeni için iki ayrı küme, X_3 değişkeni için iki ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum optimum parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki değişkenler sırasıyla bir, iki ve ikiye parçalanmış ve her bir gözlemin düştüğü veri grubu

belirlenmiştir. Aşağıda X_1 , X_2 ve X_3 için sırasıyla $k_1 = 1$, $k_2 = 2$ ve $k_3 = 2$ olarak değişkenlerdeki parçalanmalar elde edilmiştir.

X_1 değişkeninde parçalanma olmadığından değişken tek küme olarak alınır, X_2 değişkenindeki parçalanmalar X_{21}, X_{22} ve X_3 değişkeninde X_{31} ve X_{32} olarak gösterilsin. $n \times 3$ tipindeki veri matrisi $X = [X_1 X_2 X_3]$ şeklinde gösterilebilir. n_1 elemanlı X_1 değişkeni $X_1 = [X_1]$ şeklinde gösterilir, burada $n_1 = n_1$ olarak değişkendeki eleman sayısını gösterir. n_2 elemanlı X_2 değişkeni $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}$ şeklinde gösterilir, burada X_{21} ve X_{22} sırasıyla n_{21} ve n_{22} elemanlı olup $n_2 = n_{21} + n_{22}$ olarak elde edilir. n_3 elemanlı X_3 değişkeni $X_3 = \begin{bmatrix} X_{31} \\ X_{32} \end{bmatrix}$ şeklinde gösterilir, burada X_{31} ve X_{32} sırasıyla n_{31} ve n_{32} elemanlı olup $n = n_{31} + n_{32}$ olarak elde edilir.

Çok değişkenli veri setindeki değişkenler ve değişkenlere k-means algoritması uygulandıktan sonra değişkenlerdeki parçalanmalara düşen gözlem sayıları **Tablo 3.6.4** te verilmiştir.

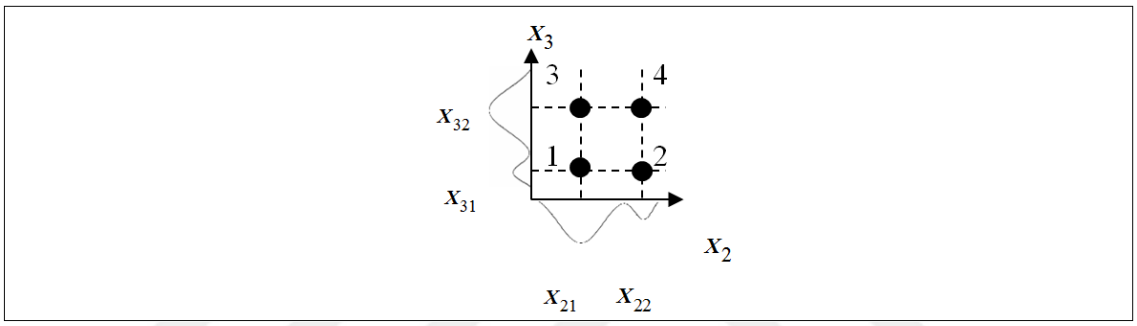
Tablo 3.6.4. İki değişkenli veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişkenler	X_1	X_2		X_3	
Değişkendeki parçalanmalar	X_{11}	X_{21}	X_{22}	X_{31}	X_{32}
Gözlem sayısı	$n_{11} = 300$	$n_{21} = 211$	$n_{22} = 89$	$n_{31} = 74$	$n_{32} = 226$
Toplam	300	300		300	

3.6.3. Çok Değişkenli Veride Aday Modeller Arasından En İyi Modelin Seçimi İçin Normal Karma Modellerde Değişkenlerdeki Parçalanmalara Dayalı Kümelenme Merkez Sayısının Ve Yapısının Belirlenmesi

Çok değişkenli veride değişkenlerdeki parçalanmalar sırasıyla, $k_1 = 1$, $k_2 = 2$ ve $k_3 = 2$ olarak **Tablo 3.6.1-3** te elde edilmiştir. Normal karma modeldeki kümelenme merkez

sayıları değişkenlerdeki parçalanmalara dayalı olarak hesaplanır. Değişkenlerdeki parçalanmaların sayısı kümelenme merkez sayısını, parçalanmalara düşen gözlem değerleri sırasıyla, kümelenme merkezlerinin yerini, şeklini ve büyüklüğünü belirlemektedir [49]. Modeldeki değişkenlerin parçalanmalarının oluşturduğu maksimum ve minimum kümelenme merkezlerinin sayısı C_{max} ve $C_{min}, s=1,...,p$ olmak üzere k_s değerlerine bağlı olarak (3.2.4) ve (3.2.5) denklemlerinden verideki parçalanmalara karşılık gelen en çok kümelenme merkezi sayısı $C_{max}=k_1.k_2.k_3=1.2.2=4$ ve en az kümelenme merkez sayısı $C_{min}=\max\left\{k_1,k_2,k_3\right\}=\max\left\{1,2,2\right\}=2$ olarak elde edilir.



Şekil 3.6.2. X_1 değişkeninde parçalanmanın olmadığı, X_2 ve X_3 değişkenleri üzerine izdüşüm alınarak elde edilen iki değişkenli model grafiği ve oluşabilecek merkez numaraları.

Şekil 3.6.2 de birinci kümelenme merkezi; n_1, n_{21} ve n_{31} elemanlı sırasıyla X_1, X_{21} ve X_{31} parçalanmalarının yer aldığı $\left[X_1X_{21}X_{31}\right]$ veri matrisi ikinci kümelenme merkezi; n_1, n_{22} ve n_{31} elemanlı sırasıyla X_1, X_{22} ve X_{31} parçalanmalarının yer aldığı $\left[X_1X_{22}X_{31}\right]$ veri matrisi üçüncü kümelenme merkezi; n_1, n_{21} ve n_{32} elemanlı sırasıyla X_1, X_{21} ve X_{32} parçalanmalarının yer aldığı $\left[X_1X_{21}X_{32}\right]$ veri matrisi dördüncü kümelenme merkezi; n_1, n_{22} ve n_{32} elemanlı sırasıyla X_1, X_{22} ve X_{32} parçalanmalarının yer aldığı $\left[X_1X_{22}X_{32}\right]$ veri matrisi tarafından oluşturulur.

3.6.4. Çok Değişkenli Veride Aday Modeller Arasından En İyi Modelin Seçimi için Normal Karma Modellerde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

Çok değişkenli veride X_1 değişkeninde parçalanmanın bulunmadığı yani $k_1 = 1$ ve X_2 ve X_3 değişkenlerinin sırasıyla $k_2 = 2, k_3 = 2$ parçalanmalarına dayalı olarak kümelenme merkezleri için M_{Toplam} ile gösterilen normal dağılımların karma modelleriyle oluşturulabilecek toplam model sayısı (3.2.6) denkleminden $M_{Toplam} = 2^{1.2.2} - 1 = 2^4 - 1 = 15$ olarak elde edilir. Burada çıkarılan 1 model varsayıma uymayan kümelenme merkezi bulunmayan boş modeldir. Bu veri seti için oluşturulabilecek model sayısı $\binom{n}{r}$ ifadesi *n* toplam kümelenme merkez sayısını ve *r* modeldeki kümelenme merkez sayısını göstermek üzere $r = 1,...4$ kümelenme merkezli, oluşturulabilecek normal dağılımların karma modellerinin sayısı sırasıyla $\binom{4}{1} = 4$, $\binom{4}{2} = 6$, $\binom{4}{3} = 4$ ve $\binom{4}{4} = 1$ olarak elde edilir. Toplamda normal dağılımların karma modelleriyle oluşturulabilecek model sayısı $\sum_{k=1}^{4} \binom{4}{k} - 1 = 15$ olacak şekilde merkez sayılarına bağlı olarak elde edilir.

3.6.5. Çok Değişkenli Veride Aday Modeller Arasından En İyi Modelin Seçimi için Normal Karma Modellerde Uygun Aday Model Sayısının Hesaplanması

Üç değişkenli normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ve yine parçalanmalara bağlı olarak toplam model sayısı elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli veya uygun modellerin sayısı araştırılmıştır. Bu geçerli modeller **Şekil 3.6.2** de gösterilen merkezler üzerinden kısaca her boyutta en az bir kümelenme merkezi bulunacak varsayımı ile anlatılabilir. Varsayıma uyan modellere uygun aday model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanabilir.

Her boyutta en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı her bir merkezin bulunduğu boyuttaki konumuna göre kombinasyon ile elde edilebilir. Hesaplama sonucunda üç değişkenli veride olabilecek minimum kümelenme merkez sayısı 2, maksimum kümelenme merkez sayısı 4 için merkez sayıları, merkezlerin konumu, modellerin sayısı için bağıntı ve uygun aday model sayıları **Tablo 3.6.5** te verilmiştir.

Tablo 3.6.5. Üç değişkenli veride değişkenlerde bir, iki ve iki parçalanma durumunda oluşan toplam model sayısı ve uygun aday model sayısı

Merkez Sayısı	**Merkezlerin Konumu**	**Modellerin Sayısı İçin Bağıntı**	**Uygun Aday Model Sayısı**
Tek Merkezli	1 şeklinde sütunlara dağılan	-	-
İki merkezli	1 1 şeklinde sütunlara dağılan	$\binom{2}{1}\binom{1}{1}1! = 2$	2
Üç merkezli	2 2 şeklinde sütunlara dağılan	$\binom{2}{2}\binom{2}{1}2! = 4$	4
Dört merkezli		$\binom{2}{2}\binom{2}{2}1! = 1$	1
Toplam model sayısı		$2^{1.2.2} - 1 = 15$	
Toplam Uygun model sayısı		0+2+4+1=7	

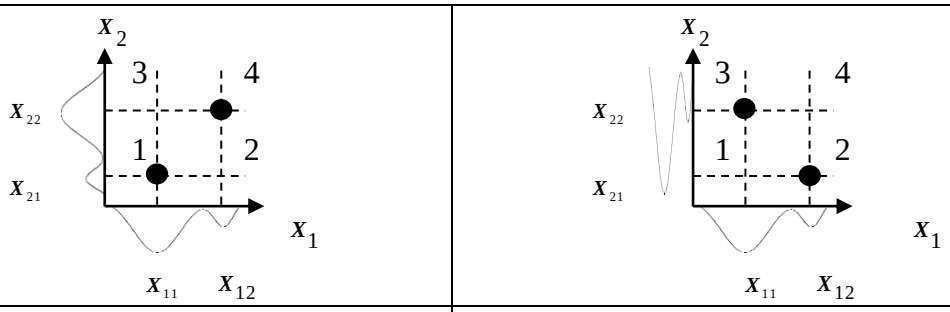
Varsayım altındaki uygun aday model sayısının hesaplanmasındaki problem bir matris yapısına karşılık gelmektedir. Varsayımdaki değişkenler ve değişkenlerin parçalanmaları matrisin satır ve sütunlarına karşılık gelmektedir. n_{ij} $i.$ satır ve $j.$ sütundaki kümelenme merkezini göstermek üzere, nxm tipindeki sıfırlardan oluşan bir kare matriste her satır ve her sütunda en az bir kümelenmenin olduğu başka bir ifade ile sıfırdan farklı en az bir elemanın bulunduğu matrislerin sayısı (3.5.1) denkleminden hesaplanır. Birinci değişkende veri yapısı homojenlik gösterdiğinden parçalanma sayısı 1 olarak elde edildiğinden (3.5.1) numaralı denklemdeki hesaplama iki boyutlu denklemlerle yapılan hesaplamalar ile aynıdır. Homojen yapıdaki değişkenlerin modellerin oluşturulmasında ve hesaplamalarda en iyi modelin belirlenmesinde etkisi olmadığı gösterilmiştir. Modellerin yapısı ve aday modellerin sayısı veri setindeki heterojen değişkenlerden kaynaklandığı için değişken veri parçalanmadaki bileşenlerin heterojenliğinin test edilmesi kümelemede en iyi modelin bulunmasında önemli yer tutmaktadır.

Değişkenlerdeki parçalanmalara dayalı aday model sayısı ve modellerin dizi gösterimleri **Tablo 3.6.6** da verilmiştir.

Tablo 3.6.6. Modellerin temsili gösterimleri, kümelenmelere göre model sayıları, modellerin uygun olup olmadıkları ve uygun model sayıları.

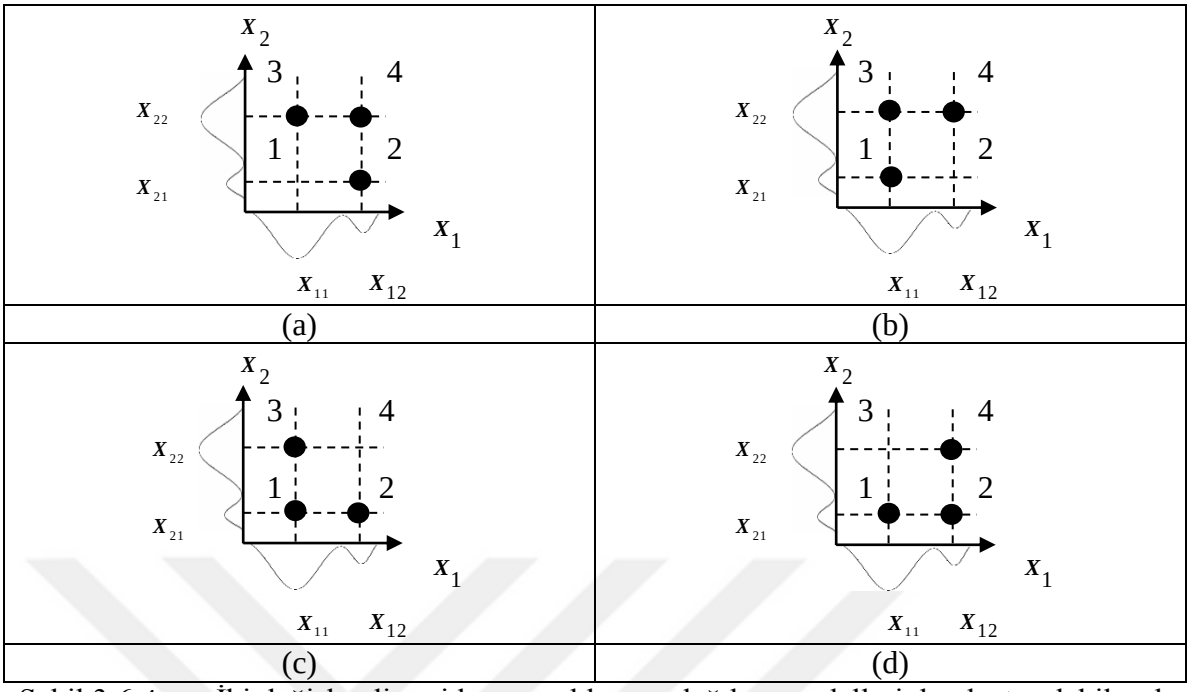
Kümelenme Merkez Sayısı	Durum sayısı	Uygun Durum Sayısı	Uygun Durumlara Karşılık Gelen Modellerin Temsili Gösterimleri
			1234
Bir Kümelenme Merkezli Modeller	4	-	-
İki Kümelenme Merkezli Modeller	6	2	1001
			0110
Üç Kümelenme Merkezli Modeller	4	4	1110
			1101
			1011
			0111
Dört Kümelenme Merkezli Modeller	1	1	1111

Üç değişkenli veride X_1 değişkeninde parçalanma olmadığından X_2 ve X_3 değişkenlerindeki parçalanmalara dayalı aday normal karma modeller oluşturulmuştur. **Şekil 3.6.2** de X_2 ve X_3 değişkenlerinin düzlem üzerine izdüşümü ile oluşan iki değişkenli veride her bir değişkenin sırasıyla ikiye parçalanması durumunda veride normal dağılımların karma modelleriyle oluşturulabilecek ve $M_{Muhtemel}$ ile gösterilen aday modellerin grafikleri **Şekil 3.6.3 -5**'te gösterilmiştir.

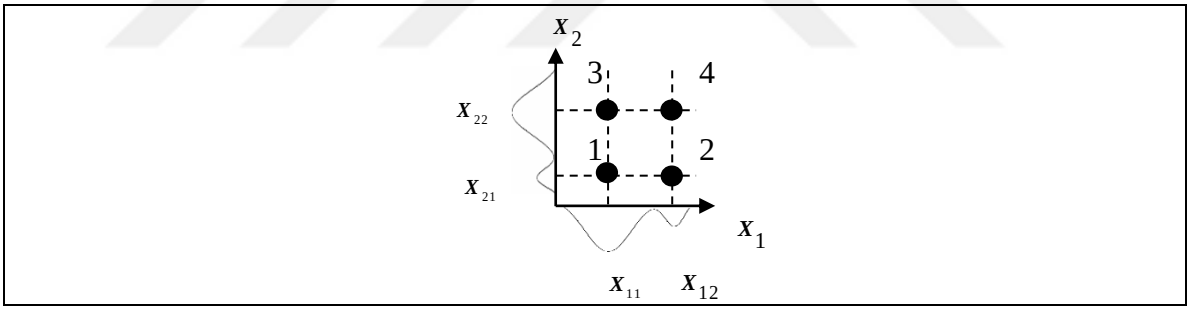


(a) (b)

Şekil 3.6.3. İki değişkenli veride normal karma dağılım modelleriyle oluşturulabilecek iki kümelenme merkezli (a) birinci uygun aday model (b) ikinci uygun aday model grafiği.



Şekil 3.6.4. İki değişkenli veride normal karma dağılım modelleriyle oluşturulabilecek üç kümelenme merkezli (a) üçüncü uygun aday model (b) dördüncü uygun aday model (c) beşinci uygun aday model (d) altıncı uygun aday model grafiği.



Şekil 3.6.5. Dört kümelenme merkezli. Normal dağılımların karma modelleriyle oluşturulabilecek dört kümelenme merkezli bir muhtemel modelin kümelenme merkezleri.

Böylece üç değişkenli veride değişkenlerin bir, iki ve ikiye parçalanması durumunda simülasyon ile oluşturulan sentetik veri seti için toplamda normal dağılımların karma modelleriyle oluşturulabilecek uygun aday model sayısı **Şekil 3.6.3-5** teki grafiklerde gösterilen 7 aday modeldir.

3.6.6. Çok Değişkenli Veride Aday Modeller Arasından En İyi Modelin Seçimi için Normal Karma Modellerde Uygun Aday Modellerin Oluşturulması ve Parametre Tahminleri

Üç değişkenli normal karma dağılım modelleri için bileşen ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir. Karma modeldeki her bir merkezin olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisi modeldeki olasılık yoğunluk fonksiyonunu elde etmek için kullanılır. **Tablo 3.6.6** da uygun aday modellerin temsili gösterimleri modeldeki merkezlerin numaralarına göre dizi gösterimi şeklinde verilmiştir. Modelin dizi gösteriminde kullanılan "0" ve "1" elemanları modeldeki karşılık gelen merkezde yığılmanın olup olmadığını göstermektedir. Eğer uygun aday modelde herhangi bir merkezde kümlenme varsa "1" yoksa "0" ile gösterilmiştir.

3.6.6.1. Çok Değişkenli Veride Aday Modeller Arasından En İyi Modelin Seçimi için Normal Karma Modeller ve Parametre Tahminleri

Üç değişkenli veride değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerini temsil eden ortalama vektörleri, varyans-kovaryans matrisleri ve korelasyon katsayıları verinin parçalanmalarının ortalamaları ve standart sapmalarından elde edilmiştir. Üç değişkenli verinin ortalama vektörü,

$\mu_i = \begin{bmatrix} \mu_1 \\ \mu_{2\bullet} \\ \mu_{3*} \end{bmatrix}$ ile temsil edilir. $i = 1, 2$ olmak üzere, aday modellerin varyans-kovaryans matrisleri $\Sigma_i = \begin{bmatrix} {\sigma_1}^2 & \rho_i\sigma_1\sigma_{2\bullet} & \rho_j\sigma_1\sigma_{3*} \\ \rho_i\sigma_{2\bullet}\sigma_1 & {\sigma_{2\bullet}}^2 & \rho_k\sigma_{2\bullet}\sigma_{3*} \\ \rho_j\sigma_{3*}\sigma_1 & \rho_k\sigma_{3*}\sigma_{2\bullet} & {\sigma_{3*}}^2 \end{bmatrix}$, ile temsil edilir. Burada $\bullet: X_2$ değişkenindeki 1,2 parçalanmayı, $*: X_3$ değişkenindeki 1,2 parçalanmayı ve $\rho = \boldsymbol{Corr}(\boldsymbol{X}_1, \boldsymbol{X}_{2\bullet}, \boldsymbol{X}_{3*})$ Pearson korelasyon katsayısını göstermektedir.

Bu merkezindeki veriler, $N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ normal dağılımına sahip olsun. Üç değişkenli veride değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ile normal karma modeller oluşturulurken olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri kullanılmıştır. Değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin varyans-kovaryans matrislerindeki bileşenler arasındaki ilişkinin yönünü ve derecesini veren Pearson korelasyon katsayısı üç bileşenli merkez yapısı için $3x3$ tipinde matris içindeki bileşenlerde altı adet bulunmaktadır. Varyans-kovaryans matrisi simetrik matris yapısında olduğundan her bir varyans-kovaryans matrisinde üç farklı Pearson korelasyon katsayısı bulunur. Modellerdeki farklı korelasyon katsayılarının sayısı; değişkenlerdeki parçalanmalar kullanılarak (3.5.2) denkleminden elde edilir.

Toplam 7 uygun aday model içerisinden:

İki bileşenli normal karma modele karşılık gelen modeller $u=1,2$ ve $i=1,2$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)})=\sum_{i=1}^{2}\pi_i^{(u)}f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \qquad (3.6.1)$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)}=\frac{\pi_i}{\sum_{i=1}^{2}\pi i}$, $0<\pi_i<1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta})=\pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+\pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta})=\pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)+\pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$$

iki normal karma model bulunur.

Üç bileşenli normal karma modele karşılık gelen modeller $u=3,...,6$ ve $i=1,...,3$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)})=\sum_{i=1}^{3}\pi_i^{(u)}f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \qquad (3.6.2)$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)}=\frac{\pi_i}{\sum_{i=1}^{3}\pi i}$, $0<\pi_i<1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

dört normal karma model bulunur.

Dört bileşenli normal karma modele karşılık gelen model $u = 7$ ve $i = 1,...,4$ için normal bileşen yoğunluk fonksiyonu

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{4} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.6.3}$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{4} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

bir normal karma model bulunur.

(3.6.1)-(3.6.3) denklemlerindeki $u = 1,...,7$ için karma modellerin yapısına uyan modellerin, olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{k} n_i}$, $2\bullet = 1,2$ ve $3* = 1,2$ olmak üzere ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_1^{(u)} \\ \bar{x}_{2\bullet}^{(u)} \\ \bar{x}_{3*}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri

$$\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_1^{(u)}\right)^2 & r_{1,2\bullet}^{(u)} s_1^{(u)} s_{2\bullet}^{(u)} & r_{1,3*}^{(u)} s_1^{(u)} s_{3*}^{(u)} \\ r_{2\bullet,1}^{(u)} s_{2\bullet}^{(u)} s_1^{(u)} & \left(s_{2\bullet}^{(u)}\right)^2 & r_{2\bullet,3y}^{(u)} s_{2\bullet}^{(u)} s_{3*}^{(u)} \\ r_{3*,1}^{(u)} s_{3*}^{(u)} s_1^{(u)} & r_{3y,2\bullet}^{(u)} s_{3*}^{(u)} s_{2\bullet}^{(u)} & \left(s_{3*}^{(u)}\right)^2 \end{bmatrix}, i = 1,2 \text{ için normal bileşen}$$

yoğunluk fonksiyonlarının bilinmeyen parametreleridir. Veri setindeki değişkenlerin gözlem değerlerinden elde edilen parametre tahminleri **Tablo 3.6.7** de verilmiştir.

Tablo 3.6.7. Değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin ortalama vektörü ve varyans-kovaryans parametreleri tahmini.

μ	Σ
$\mu_1 = \begin{bmatrix} 30.1168 \\ 23.5331 \\ 20.1245 \end{bmatrix}$, $\Sigma_1 = \begin{bmatrix} 108.641 & -5.744 & -11.069 \\ -5.744 & 92.963 & 18.051 \\ -11.069 & 18.051 & 136.873 \end{bmatrix}$	$\mu_2 = \begin{bmatrix} 30.1168 \\ 67.3698 \\ 20.1245 \end{bmatrix}$, $\Sigma_2 = \begin{bmatrix} 108.641 & 16.330 & -11.069 \\ 16.330 & 108.335 & -2.386 \\ -11.069 & -2.386 & 136.873 \end{bmatrix}$
$\mu_3 = \begin{bmatrix} 30.1168 \\ 23.5331 \\ 68.5173 \end{bmatrix}$, $\Sigma_3 = \begin{bmatrix} 108.641 & -5.744 & 0.382 \\ -5.744 & 92.963 & 6.761 \\ 0.382 & 6.761 & 80.099 \end{bmatrix}$	$\mu_4 = \begin{bmatrix} 30.1168 \\ 67.3698 \\ 68.5173 \end{bmatrix}$, $\Sigma_4 = \begin{bmatrix} 108.641 & 16.330 & 0.382 \\ 16.330 & 108.335 & 15.393 \\ 0.382 & 15.393 & 80.099 \end{bmatrix}$

3.6.7. Çok Değişkenli Veride Aday Modeller Arasından En İyi Modelin Seçimi için Normal Dağılımların Karma Modelleri ile Bilgi Kriterlerinin Hesaplanması.

Modele dayalı kümeleme yapmak için iki değişkenli ve değişkenlerin bir, iki ve ikiye parçalandığı veri setindeki uygun aday modellerin normal karma modelleri belirlenmiştir. Normal karma modeller arasından en iyi modelin seçimi için her bir uygun aday modelin çok değişkenli normal karma dağılımlardan log-likelihood, AIC ve BIC değerleri elde edilmiştir. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu, $i = 1, 2, ...n \ \ ve \ \ j = 1, 2, ...g$ olmak üzere $f\left(x_i; \theta_j\right)$ karma olasılık yoğunluk fonksiyonunun likelihood fonksiyonu ve logaritması alınmış likelihood fonksiyonu (3.2.13) ve (3.2.14) denklemeleri kullanılarak elde edilir.

çok değişkenli normal karma modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kriteri (BIC) ve

Akaike bilgi kriteri (AIC) (3.2.15) ve (3.2.17) denklemeleri kullanılarak elde edilir.

Tablo 3.6.8. Üç değişkenli sentetik veri setinde en iyi modelin elde edilmesi için kullanılan log-likelihood, AIC ve BIC değerlerinin hesaplanması için Matlab kodu.

```
a=load('urgx1.txt') ;
b=load('urgx21.txt');
c=load('urgx22.txt');
d=load('urgx31.txt');
e=load('urgx32.txt');
pi1=(size(a)+size(b)+size(d))/(size(a)+size(b)+size(d)+size(a)+size(b)+s
ize(e)+size(a)+size(c)+size(e));
pi3=(size(a)+size(b)+size(e))/(size(a)+size(b)+size(d)+size(a)+size(b)+s
ize(e)+size(a)+size(c)+size(e));
pi4=(size(a)+size(c)+size(e))/(size(a)+size(b)+size(d)+size(a)+size(b)+s
ize(e)+size(a)+size(c)+size(e));
x(1,:)=load('urgx1.txt');
x(2,:)=load('urgx2.txt');
x(3,:)=load('urgx3.txt');
mu(1,:)=mean(a);
mu(2,:)=mean(b);
mu(3,:)=mean(d);
sigma=[108.641 -5.744 -11.069 ; -5.744 92.963 18.051;-11.069 18.051
136.873];
for i=1:length(x(1,:))
pdf1(i)=(1/(2*pi)^(3/2))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-
mu)'*inv(sigma)*(x(:,i)-mu));
end
mu(1,:)=mean(a);
mu(2,:)=mean(b);
mu(3,:)=mean(e);
sigma=[108.641 -5.744 0.382;-5.744 92.963 6.761;0.382 6.761 80.099];
for i=1:length(x(1,:))
pdf3(i)=(1/(2*pi)^(3/2))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-
mu)'*inv(sigma)*(x(:,i)-mu));
end
mu(1,:)=mean(a);
mu(2,:)=mean(c);
mu(3,:)=mean(e);
sigma=[108.641 16.330 0.382; 16.330 108.335 15.393; 0.382 15.393
80.099];
for i=1:length(x(1,:))
pdf4(i)=(1/(2*pi)^(3/2))*(1/((det(sigma))^(1/2)))*exp((-0.5)*(x(:,i)-
mu)'*inv(sigma)*(x(:,i)-mu));
end
 oyf1=pi1*pdf1+pi3*pdf3+pi4*pdf4;
 lnl= -29*log(length(oyf1))+
sum(log(pi1*pdf1(:)+pi3*pdf3(:)+pi4*pdf4(:)));
lnl
%AIC bilgi kriteri
aic=-2*lnl+2*29;
aic
%BIC bilgi kriteri
bic=-2*lnl+29*log(length(lnl));
bic
end
```

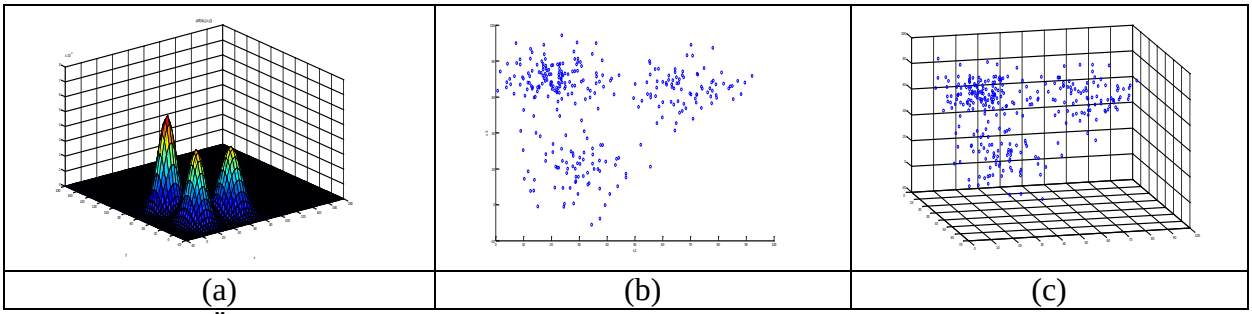
Tablo 3.6.9. Çok değişkenli veride normal karma dağılımların modele dayalı kümelenmesi için log-likelihood, AIC ve BIC değerleri.

Model no	log-l	AIC	BIC	Matris gösterimi
1	-4863.6	9765.3	9727.3	1001
2	-4636.1	9310.1	9272.1	0110
3	-4357.5	8772.9	8714.9	1110
4	-5001.9	10062	10004	1101
5	**-3842.9**	**7743.8**	**7685.8**	**1011**
6	-4261.3	8580.6	8522.6	0111
7	-3942.7	7973.3	7895.3	1111

Model seçiminde Log-likelihood değeri en büyük ve AIC ile BIC değerlerinin en küçük olduğu 5. model elde edilen değerlere göre tüm uygun aday modeller arasındaki en iyi model olarak seçilmiştir.

3.6.8. Çok Değişkenli Veride Aday Modeller Arasından En İyi Modelin Seçimi için Normal Karma Modellerin Modele Dayalı Kümelenmesi ile En İyi Modelin Seçimi

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin log-l, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. Üç değişkenli veri setinde normal karma modellerden uygun aday modellerin log-l, AIC ve BIC değerleri hesaplanmış ve **Tablo 3.6.9** da verilmiştir. Uygun aday modeller arasından modele dayalı kümelemede en iyi model log-likelihood değer en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir. Bu değerlere göre üç kümelenme merkezli beşinci uygun aday model en iyi model olark belirlenmiştir.



(a) (b) (c)

Şekil 3.6.6. Üç değişkenli ve değişkenlerin sırasıyla bir, iki ve ikiye bölündüğü normal karma dağılımlar arasından bilgi kriterlerine göre belirlenen en iyi modelin (a) yoğunluk fonksiyonunun yüzey grafiği (b) iki boyutlu saçılım grafiği (c) üç boyutlu saçılım grafiği

3.7. Çok Değişkenli Büyük Veride Modele Dayalı Karma Kümeleme ile En İyi Modelin Seçimi için Yeni Bir Veri Madenciliği Yaklaşımı

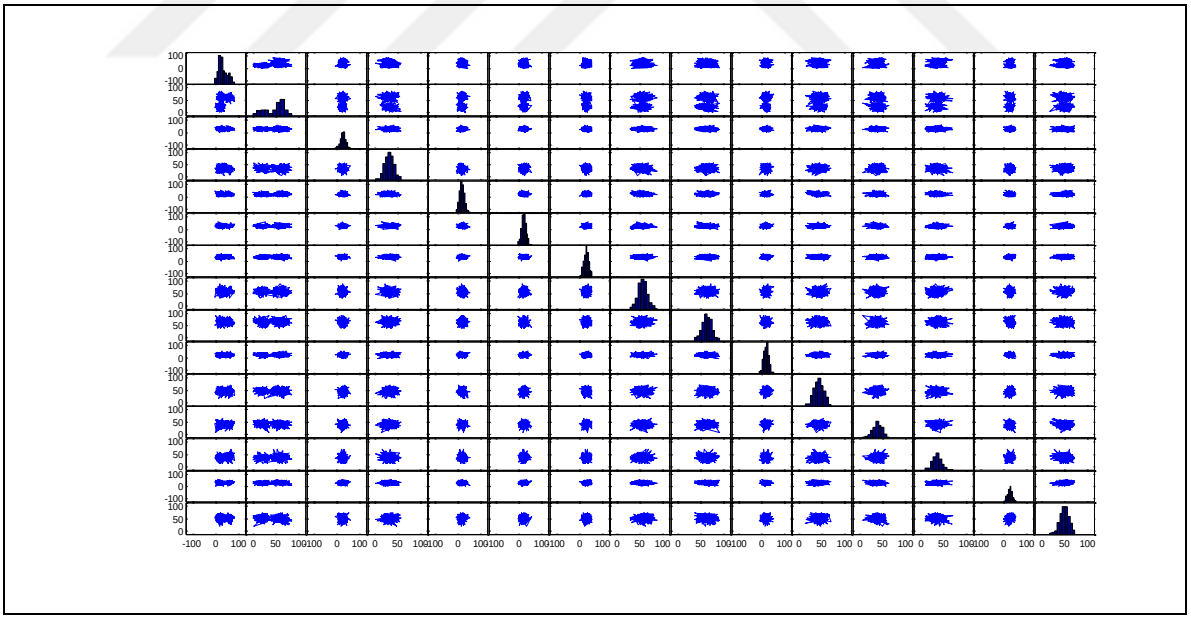
Çok değişkenli veride modele dayalı kümeleme yapmak için verideki her bir değişkenin anlamlı alt gruplara ayrılması gerekmektedir. Normal karma modellerdeki bileşenler değişkenlerdeki alt gruplardan oluşmaktadır. Normal karma modellerdeki her bileşene bir kümelenme karşılık gelir. Dolayısıyla modeldeki bileşen sayısının belirlenmesi kümeleme analizindeki en zor problemlerdendir [5]. Bu kısımdaki çalışmada büyük veri üzerinde kümelenme araştırılmıştır. Büyük veri ister çok fazla değişken (yatay büyük veri), ister her bir değişkende çok fazla gözlem ihtiva etsin (dikey büyük veri) geliştirilen genetik algoritma ile değişkenlerdeki dağılımlara göre en iyi kümelenme belirlenmektedir. Verinin çok fazla değişken içermesi iterasyonla parametre belirleyen algoritmalarda işlem karmaşıklığına sebep olmaktadır. Geliştirilen genetik algoritma parametre tahmini ve model seçiminde iterasyonla (EM algoritması gibi) işlem yapmadığından işlem karmaşıklığı meydana gelmemektedir. Geliştirilen veri madenciliği genetik algoritması ile on beş değişkenli veri setinde değişkenlerdeki parçalanma sayıları ile parçalanma sayılarına bağlı oluşabilecek küme yapıları ve en iyi kümelenmenin olduğu normal karma model ortaya çıkarılmıştır. Veri setindeki homojen değişkenler elenip heterojen yapıdaki değişkenler ile küme yapısı ve sayısı belirlenmiştir.

3.7.1. Çok Değişkenli Büyük Veride Normal Karma Modellerin Modele Dayalı Kümelenmesi için Değişkenlerdeki Parçalanmanın Grafiksel ve Tek Değişkenli Normal Karma Modeller ile Tahmini

Simülasyon ile üretilen çok değişkenli sentetik veride değişkenlerdeki parçalanmaların belirlenmesi için çeşitli yöntemlerden faydalanılabilir. Her bir değişkendeki parçalanmalar değişkenlerin alt grupları olarak adlandırılır. Değişkenlerdeki parçalanmaları belirlemek ve bu parçalanmaları anlamlı alt gruplara dönüştürmek için istatistiksel grafiksel yöntemler ve normal karma modeller kullanılır. Her bir değişkendeki anlamlı parçalanma sayısı histogram grafiğindeki tepelenme sayısı ve P-P grafiğinde gözlemlerin normal doğrusunu kesme sayısına göre tahmin edilebilir.

3.7.1.1. Çok Değişkenli Büyük Veride Normal Karma Modellerin Modele Dayalı Kümelenmesi için Değişkenlerdeki Parçalanmanın Grafiksel Yöntemlerle Tahmini.

Çok değişkenli veri setinde her bir değişkendeki parçalanmalar değişkenin heterojen olup olmadığını test etmek için önemlidir. Veri setindeki değişkenler tek tek ele alınıp parçalanmaların varlığı araştırılır. $s = 1,\ldots,p$ olmak üzere verideki her bir $\boldsymbol{X}_S$ değişkeninin heterojenliği incelensin. Verideki değişkenler homojen ya da heterojen yapıda olabilir. Heterojen verideki değişkenlerin yapısı verideki kümelenmeyi belirlemektedir [41]. p -boyutlu veride $\boldsymbol{X}_S$ değişkeninin yapısına göre parçalanmalarının sayısı $k_S \geq 1$ olsun. $k_S = 1$ durumu değişkenin homojen yapıda ve $k_S > 1$ durumu değişkenin heterojen yapıda olması anlamına gelir. Histogram ve P-P grafikleri gibi istatistiksel grafiksel yöntemlerle değişkenlerde parçalanmaların varlığı belirlenebilir. **Şekil 3.7.1** de veri setindeki her bir değişkenin diğer değişkenler ile oluşturduğu saçılım grafiği ve Histogram grafikleri verilmiştir.



Şekil 3.7.1. On beş değişkenli büyük veride değişkenindeki parçalanmayı gösteren Histogram ve saçılım grafiği.

Şekil 3.7.1 de on beş değişkenin birbiri ile karşılaştırmalı histogram ve saçılım grafikleri verilmiştir. Grafikte köşegen üzerinde sırasıyla $p = 1,\ldots,15$ değişkenin histogram grafiği ve satırlarda değişkenlerin karşılıklı saçılım grafiği gösterilmiştir. Elde edilen grafikten $X_1 ve X_2$ değişkenlerinin her birinde iki parçalanma yani $k_1 = 2$ ve $k_2 = 2$ bulunduğu

tahmin edilmiştir. Diğer değişkenlerde parçalanma gözlenememektedir. Yani $X_3, X_4, ..., X_{15}$ değişkenlerinin her birinde homojen yapı olduğu tahmin edilmiştir.

3.7.1.2. Çok Değişkenli Heterojen Büyük Veride Normal Karma Modellerin Modele Dayalı Kümelenmesi için Değişkenlerdeki Parçalanmanın Tek Değişkenli Normal Karmalara Dayalı Belirlenmesi.

Çok değişkenli veride değişkenlerdeki parçalanmanın belirlenmesi tek değişkenli normal karma dağılım modeli (3.2.1) ve (3.2.2) denklemleri kullanılarak elde edilebilir. Değişkendeki uygun parçalanma sayısına göre tek değişkenli normal karma dağılımın olasılık ağırlıkları π , ortalama vektörü μ ve varyansı σ^2 parametreleri hesaplanır. Her bir değişkendeki parçalanmanın ortaya çıkarılması için tek değişkenli normal karma modelin log-likelihood, AIC ve BIC değerleri modelin parametrelerinden parçalanma sayısına bağlı olarak hesaplanır. Tek değişkenli normal karma dağılımlardan elde edilen log-likelihood değerinin en büyük, AIC ve BIC değerlerinin en küçük olduğu parçalanma sayısı en uygun parçalanma sayısı olarak belirlenir. Hesaplamalarda MATLAB programının istatistik paketi kütüphanesinde yer alan *gmdistribution.fit(data,k)* fonksiyonu kullanılmıştır. Çok değişkenli veride heterojen yapının ortaya çıkarılması için her bir değişkende optimum parçalanma sayısını belirlemek amacıyla grafiksel yöntemlerden tahmin edilen sayıda $k = 1, ... n$ değerleri algoritmaya girilerek oluşturulan karma normal modeller için log-likelihood, AIC ve BIC değerleri hesaplanmış ve hesaplama sonuçlarına göre en uygun parçalanma sayısı belirlenmiştir.

Tablo 3.7.1. Çok değişkenli veri setinde X_1 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1753.4	3510.7	3510.8
2	**-1704.7**	**3419.4**	**3439.4**
3	-1702.3	3420.6	3452.5

Tablo 3.7.2. Çok değişkenli veri setinde X_2 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1775.8	3555.7	3563.7
2	**-1693**	**3369**	**3416**
3	-1693	3402	3433.9

Tablo 3.7.3. Çok değişkenli veri setinde X_3 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1485.9	2975.9	2983.8
2	-1485.8	2981.5	3001.5
3	-1485.4	2986.7	3018.6

Tablo 3.7.4. Çok değişkenli veri setinde X_4 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1471.2	2948.3	2956.3
2	-1471.7	2953.4	2973.4
3	-1471.8	2959.6	2991.5

Tablo 3.7.5. Çok değişkenli veri setinde X_5 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1482.6	2969.3	2977.3
2	-1482.6	2975.3	2995.3
3	-1480	2977	3009

Tablo 3.7.6. Çok değişkenli veri setinde X_6 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1489.4	2982.8	2990.7
2	-1488.5	2986.9	3006.9
3	-1486.6	2989.2	3021.2

Tablo 3.7.7. Çok değişkenli veri setinde X_7 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1495	2993.9	3001.9
2	-1495	2999.9	3019.9
3	-1494	3003.9	3035.9

Tablo 3.7.8. Çok değişkenli veri setinde X_8 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1491.4	2986.8	2994.8
2	-1490.7	2991.3	3011.3
3	-1490.4	2996.7	3028.7

Tablo 3.7.9. Çok değişkenli veri setinde X_9 değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1500.6	3005.2	3013.1
2	-1500.5	3010.9	3030.9
3	-1500.5	3017	3048.9

Tablo 3.7.10. Çok değişkenli veri setinde X_{10} değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1499.8	3003.7	3011.7
2	-1499.9	3009.8	3029.7
3	-1499.8	3015.6	3047.6

Tablo 3.7.11. Çok değişkenli veri setinde X_{11} değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1490.6	2985.2	2993.2
2	-1490.6	2991.2	3011.2
3	-1490.3	2996.6	3028.6

Tablo 3.7.12. Çok değişkenli veri setinde X_{12} değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1468.8	2941.5	2949.5
2	-1466.3	2942.6	2962.6
3	-1460	2948.1	2980

Tablo 3.7.13. Çok değişkenli veri setinde X_{13} değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1488.2	2980.4	2988.4
2	-1487.1	2984.2	3004.2
3	-1487	2989.9	3021.9

Tablo 3.7.14. Çok değişkenli veri setinde X_{14} değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1478.6	2961.2	2969.2
2	-1478.6	2967.3	2987.2
3	-1475.4	2965.1	2997

Tablo 3.7.15. Çok değişkenli veri setinde X_{15} değişkenindeki parçalanma sayısını belirlemek için modellerdeki $\pi, \mu ve \sigma^2$ parametrelerine bağlı log-likelihood, AIC ve BIC değerleri.

K	Log-likelihood	AIC	BIC
1	-1475.3	2954.5	2962.5
2	-1476.6	2959.3	2979.2
3	-1471.5	2959	2991.

Çok değişkenli büyük veri setindeki her bir değişkendeki parçalanma sayısı tek değişkenli normal karma modellerin Log-likelihood, AIC ve BIC değerlerine bakılarak elde edilir. On beş değişkenin her birisinde log-likelihood değerinin en büyük aynı zamanda AIC ve BIC değerinin en küçük olduğu parçalanma sayısı en uygun parçalanma sayısı olarak belirlenmiştir. Her bir değişken için elde edilen değerlere bakılarak $X_1 \; ve \; X_2$ değişkenleri için uygun parçalanma sayısı $k_1 = 2$ ve $k_2 = 2$ olarak elde edilmiş, $X_3, X_4, ..., X_{15}$ değişkenleri için $i = 3,...,15$ olmak üzere $k_i = 1$ yani parçalanmanın bulunmadığı homojen yapı gözlenmiştir.

3.7.2. Çok Değişkenli Büyük Veride K-Ortalamalar Algoritması İle Gözlem Değerlerinin Değişkenlerdeki Parçalanmalara Atanması

Çok değişkenli $X_1 \; ve \; X_2$ değişkenlerinin her birinde 2 geriye kalan on üç değişkende parçalanmanın bulunmadığı veri setindeki değişkenlerin parçalanmalarına düşen gözlemlerin belirlenmesi için k-ortalamalar algoritması kullanılır. Başlangıçta her bir

değişkenin parçalanma sayısı k kadar merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık (3.2.3) denkleminden hesaplanır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir.

Çok değişkenli ve X_1 *ve* X_2 değişkenlerinin her birinde 2 parçalanmanın olduğu veride k-ortalamalar algoritması uygulandığında her bir değişken için iki ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum en uygun parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki gözlemlerin parçalanmanın bulunduğu değişkenlerdeki alt gruplara düştüğü veri grubu belirlenmiştir.

Çok değişkenli veride k-ortalamalar algoritması uygulandığında X_1 ve X_2 değişkenleri için iki ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojenlik) aynı zamanda küme içi mesafenin minimum (homojenlik) olduğu durum optimum parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki her bir değişkende iki parçalana ve her bir gözlemin düştüğü veri grubu belirlenmiştir. Simülasyon ile oluşturulan çok değişkenli veride her değişkendeki parçalanmalar k-ortalamalar algoritması uygulanarak X_1 ve X_2 değişkenleri için sırasıyla $k_1 = 2$ ve $k_2 = 2$ olarak elde edilir. X_1 değişkenindeki parçalanmalar X_{11}, X_{12} ve X_2 değişkenindeki parçalanmalar X_{21}, X_{22} olarak gösterilsin. $n \times 15$ tipindeki veri setinde değişken veri parçalama ve parçalanmaların bulunduğu değişkenlere k-ortalamalar algoritması uygulanarak değişkenlerdeki heterojenlik incelenmiştir. Çok değişkenli veri setinde X_1 ve X_2 değişkenlerinde parçalanma olduğundan veri iki değişkenli $n \times 2$ formundadır. İki değişkenli veri matrisi $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ şeklinde gösterilebilir. n_1 elemanlı X_1 değişkeni $X_1 = \begin{bmatrix} X_{11} \\ X_{12} \end{bmatrix}$ şeklinde

gösterilir, burada X_{11} ve X_{12} sırasıyla n_{11} ve n_{12} elemanlı olup $n_1 = n_{11} + n_{12}$ olarak elde edilir. n_2 elemanlı X_2 değişkeni $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \end{bmatrix}$ şeklinde gösterilir, burada X_{21} ve X_{22} sırasıyla n_{21} ve n_{22} elemanlı olup $n_2 = n_{21} + n_{22}$ olarak elde edilir. Çok değişkenli veri setindeki değişkenler ve değişkenlere k-means algoritması uygulandıktan sonra değişkenlerdeki parçalamalara düşen gözlem sayıları **Tablo 3.7.16** da verilmiştir.

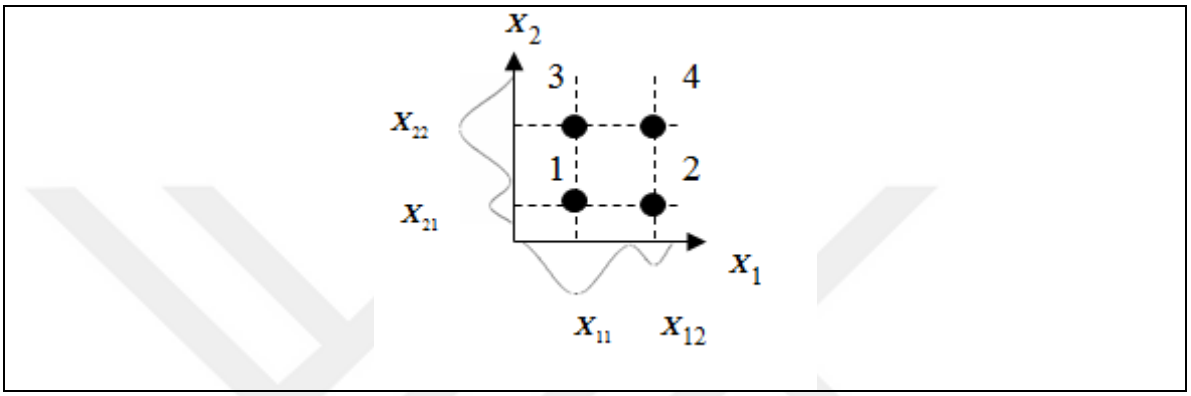
Tablo 3.7.16. On beş değişkenli veri setindeki değişkenler ve değişkenlerdeki parçalanmalara düşen gözlem sayıları.

Değişken	X_1		X_2		X_3	X_4	X_5
Değişken parçaları	$X_{1.1}$	$X_{1.2}$	$X_{2.1}$	$X_{2.2}$	X_3	X_4	X_5
Gözlem sayısı	$n_{1.1} = 120$	$n_{1.2} = 280$	$n_{2.1} = 150$	$n_{2.2} = 250$	$n_3 = 400$	$n_4 = 400$	$n_5 = 400$
Toplam	400		400		400	400	400
Değişken	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Değişken parçaları	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
Gözlem sayısı	$n_6 = 400$	$n_7 = 400$	$n_8 = 400$	$n_9 = 400$	$n_{10} = 400$	$n_{11} = 400$	$n_{12} = 400$
Toplam	400	400	400	400	400	400	400
Değişken	X_{13}			X_{14}		X_{15}	
Değişken parçaları	X_{13}			X_{14}		X_{15}	
Gözlem sayısı	$n_{13} = 400$			$n_{14} = 400$		$n_{15} = 400$	
Toplam	400			400		400	

3.7.3. Çok Değişkenli Büyük Veride Normal Karma Modellerde Değişkenlerdeki Parçalanmalara Dayalı Kümelenme Merkez Sayısının ve Yapısının Belirlenmesi

On beş değişkenli X_1, X_2 değişkenin ikiye bölündüğü ve diğer değişkenlerde parçalanmanın olmadığı veri setinde modeldeki kümelenme merkez sayıları değişkenlerdeki parçalanmalara bağlı olarak hesaplanır. Değişkenlerdeki parçalanmaların sayısı kümelenme merkez sayısını, parçalanmalara düşen gözlem değerleri ve sırasıyla, kümelenme merkezlerinin yerini, şeklini ve büyüklüğünü belirlemektedir [49].

Modeldeki değişkenlerin parçalanmalarının oluşturduğu maksimum ve minimum kümelenme merkezlerinin sayısı C_{max} ve $C_{min}, s=1,...,p$ olmak üzere k_s değerlerine bağlı olarak (3.2.4) ve (3.2.5) denklemlerinden verideki parçalanmalara karşılık gelen en çok kümelenme merkezi sayısı $C_{max}=k_1.k_2...k_{15}=2.2.1...1=4$ ve en az kümelenme merkez sayısı $C_{min}=\max\left\{k_1,k_2,...,k_{15}\right\}=\max\left\{2,2,1,...,1\right\}=2$ olarak elde edilir.



Şekil 3.7.2. On beş değişkenli büyük veride X_1 *ve* X_2 değişkeninin ikiye parçalandığı diğer değişkenlerde parçalanmanın olmadığı normal karma modelde X_1 *ve* X_2 değişkenleri üzerine iz düşüm grafiği, değişkenlerdeki parçalanmalar ve merkezlerin numaraları.

Şekil 3.7.2 de birinci kümelenme merkezi; n_{11} ve n_{21} elemanlı sırasıyla X_{11} ve X_{21} parçalanmalarının yer aldığı $\left[X_{11}\ \ X_{21}\right]$ veri matrisi; ikinci kümelenme merkezi; n_{12} ve n_{22} elemanlı X_{12} ve X_{22} parçalanmalarının yer aldığı $\left[X_{12}\ \ X_{21}\right]$ veri matrisi; üçüncü kümelenme merkezi; n_{11} ve n_{21} elemanlı X_{11} ve X_{22} parçalanmalarının yer aldığı $\left[X_{11}\ \ X_{22}\right]$ veri matrisi; dördüncü kümelenme merkezi; n_{12} ve n_{22} elemanlı X_{12} ve X_{22} parçalanmalarının yer aldığı $\left[X_{12}\ \ X_{22}\right]$ veri matrisi tarafından oluşturulur.

3.7.4. Çok Değişkenli Büyük Veride Normal Karma Modellerde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

On beş değişkenli X_1, X_2 değişkenin ikiye bölündüğü ve diğer değişkenlerde parçalanmanın olmadığı durumda kümelenme merkezleri için M_{toplam} ile gösterilen

oluşabilecek tüm modellerin sayısı değişkenlerdeki parçalanmalara bağlı olarak elde edilir. On beş değişkenli normal karma modellerin toplam sayısı (3.2.6) denkleminden $M_{Toplam} = 2^{2.2.1...1} - 1 = 2^4 - 1 = 15$ olarak elde edilir. Burada çıkarılan 1 model varsayıma uymayan kümelenme merkezi bulunmayan boş modeldir. Bu veri seti için oluşturulabilecek model sayısı $\binom{n}{r}$ ifadesi n toplam kümelenme merkez sayısı ve r modeldeki kümelenme merkez sayısını göstermek üzere $r = 1,...,4$ kümelenme merkezli, oluşturulabilecek normal dağılımların karma modellerinin sayısı sırasıyla; $\binom{4}{1} = 4$, $\binom{4}{2} = 6$, $\binom{4}{3} = 4$ ve $\binom{4}{4} = 1$ olarak elde edilir.

Toplam normal dağılımların karma modelleriyle oluşturulabilecek model sayısı $\sum_{k=1}^{4}\binom{4}{k} - 1 = 15$ olacak şekilde merkez sayılarına bağlı olarak elde edilir.

3.7.5. Çok Değişkenli Büyük Veride Değişkenlerdeki Parçalanmaya Dayalı Normal Karma Modellerde Aday Model Sayısının Hesaplanması

Çok değişkenli normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ve yine parçalanmalara bağlı olarak toplam model sayısı elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli veya uygun modellerin sayısı araştırılmıştır. Bu geçerli modeller **Şekil 3.7.2** de gösterilen merkezler üzerinden kısaca her satır ve her sütunda en az bir kümelenme merkezi bulunacak varsayımı ile anlatılabilir. Varsayıma uyan modellere uygun aday model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanabilir.

Her satır ve sütunda en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı her bir merkezin bulunduğu satır ve sütundaki konumuna göre kombinasyon hesabı ile elde edilebilir. Hesaplama sonucunda on beş değişkenli veride her değişkenin ikiye bölündüğü durumda veride olabilecek minimum kümelenme merkez sayısı 2, maksimum

kümelenme merkez sayısı 4 için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları **Tablo 3.7.17.** de verilmiştir.

Tablo 3.7.17. On beş değişkenli X_1, X_2 değişkenin ikiye parçalandığı ve diğer değişkenlerde parçalanmanın olmadığı normal karma modellerde uygun aday modeller için merkez sayıları, merkezlerin konumları, model sayısı için bağıntılar ve model sayıları.

Merkez Sayısı	**Merkezlerin Konumu**	**Modellerin Sayısı İçin Bağıntı**	**Model Sayısı**
İki merkezli	1 1 şeklinde sütunlara dağılan	$\binom{2}{1}\binom{1}{1}=2!=2$	2
Üç merkezli	2 1 şeklinde sütunlara dağılan	$\binom{2}{1}2!=2.2=4$	4
Dört merkezli	2 2 şeklinde sütunlara dağılan	$\binom{4}{4}=1$	1
Toplam model sayısı		$2^{2.2}-1=15$	
Toplam Uygun model sayısı		$0+0+2+4+1=7$	

Tek merkezli toplam model sayısı $\binom{4}{1}=\frac{4!}{(4-1)!.1!}=4$ adettir. Ancak modellerin hiçbiri varsayıma uymadığından uygun aday model yoktur.

İki merkezli toplam model sayısı $\binom{4}{2}=\frac{4!}{(4-2)!.2!}=6$ adettir. Ancak varsayıma uyan uygun aday model sayısı yukarıda hesaplandığı gibi 2 adettir.

Üç merkezli toplam model sayısı $\binom{4}{3}=\frac{4!}{(4-3)!.3!}=4$ adettir. Üç kümelenme merkezli modellerin hepsi varsayıma uyan uygun aday modeldir.

Dört merkezli toplam model sayısı $\binom{4}{4}=\frac{4!}{(4-4)!.4!}=1$ adettir. Dört kümelenme merkezli model varsayıma uyan uygun aday modeldir. Toplamda 2+4+1=7 uygun aday model bulunmaktadır.

Varsayım altındaki uygun aday model sayısının hesaplanmasındaki problem bir matris yapısına karşılık gelmektedir. Varsayımdaki değişkenler ve değişkenlerin parçalanmaları matrisin satır ve sütunlarına karşılık gelmektedir. n_{ij} $i.$ satır ve $j.$ sütundaki kümelenme merkezini göstermek üzere, nxm tipindeki sıfırlardan oluşan bir kare matriste her satır ve

her sütunda en az bir kümelenmenin olduğu başka bir ifade ile sıfırdan farklı en az bir elemanın bulunduğu $n \geq 3$ alındığında aday modellerin temsil eden matrislerin sayısı

$$f\left(n_1,...,n_r;k\right)=\sum_{i_1=1}^{n_1}(-1)^{i_1}\binom{n_1}{i_1}...\sum_{i_r=1}^{n_r}(-1)^{i_r}\binom{n_r}{i_r}\binom{\left(n_1-i_1\right)...\left(n_r-i_r\right)}{k} \tag{3.7.5}$$

şeklinde hesaplanır [75].

Burada n_r değişkenlerdeki veya matrisin boyutlarına karşılık gelen parçalanmaları, i_r parçalanmalardai küme sayısını ve k karma modeldeki küme sayısı göstermektedir.

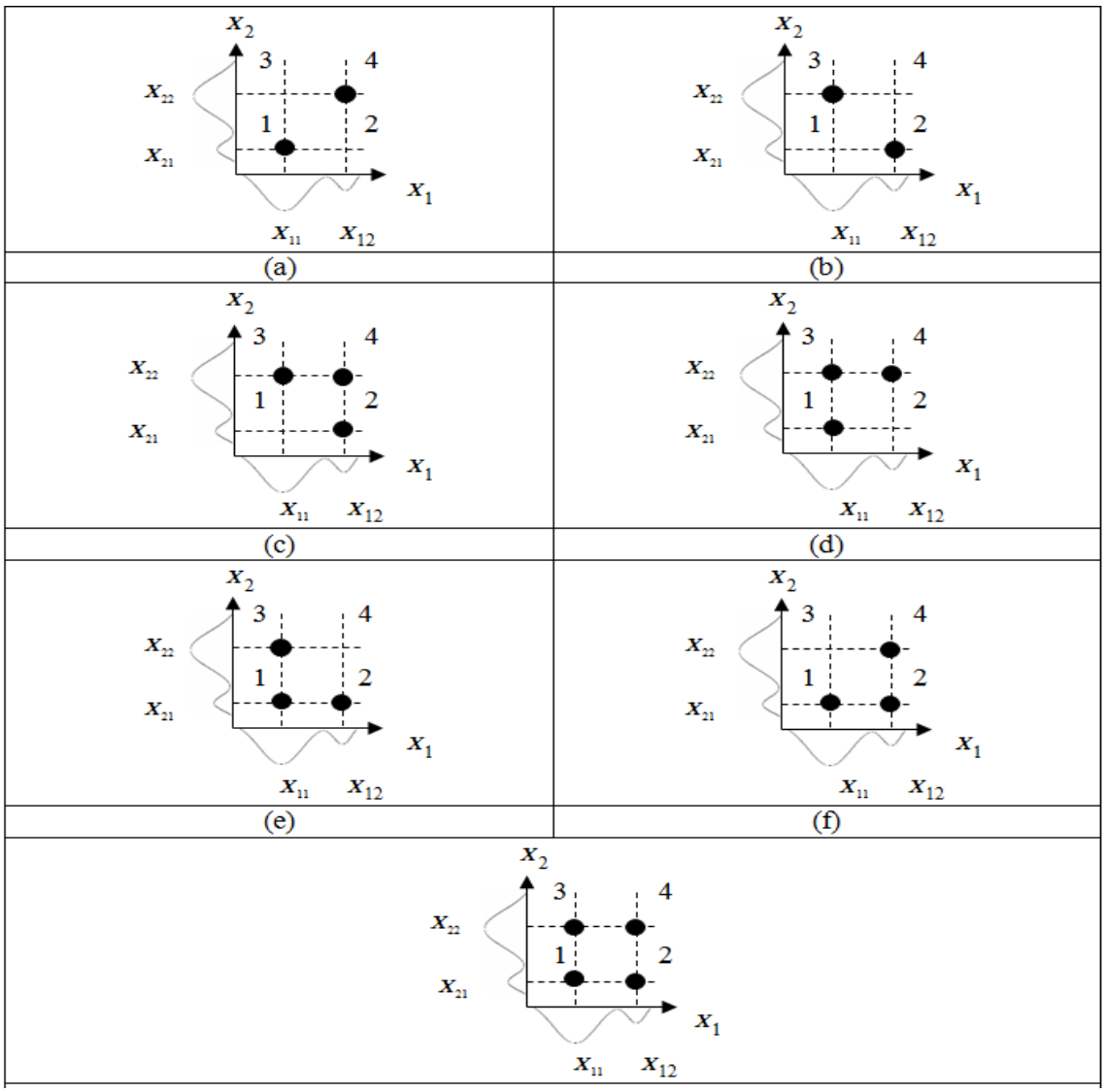
Bu yedi uygun duruma karşılık gelen kümelenme yapılarının modelleri **Tablo 3.7.18** de verilmiştir.

Tablo 3.7.18. On beş değişkenli X_1 ve X_2 değişkenin ikiye bölündüğü ve diğer değişkenlerde parçalanmanın olmadığı normal karma modeller arasından varsayıma uyan iki, üç ve dört kümelenme merkezli toplam model sayısı, uygun aday modellerin sayısı, karşılık gelen kümelenme yapılarının temsili dizi gösterimleri.

Kümelenme Merkez Sayısı	Durum Sayısı	Uygun Durum Sayısı	Uygun Durumlara Karşılık Gelen Modellerin Temsili Gösterimleri
			1234
Bir Kümelenme Merkezli Modeller	4	-	-
İki Kümelenme Merkezli Modeller	6	2	1001
			0110
Üç Kümelenme Merkezli Modeller	4	4	1110
			1101
			1011
			0111
Dört Kümelenme Merkezli Modeller	1	1	1111

Çok değişkenli büyük veride değişkenlerdeki heterojenlik, her bir değişkene tek değişkenli normal karma model uygulanarak değişken veri parçalama metodu ile ortaya çıkarılmıştır. On beş değişkenli veri setinde **Tablo 3.7.1-15** deki değerlere bakılarak $X_1 ve X_2$ değişkenlerinde $k_1=2$ ve $k_2=2$ parçalanma belirlenmiş diğer $X_3,...,X_{15}$

değişkenlerinde parçalanmanın bulunmadığı yani homojenlik tespit edilmiştir. Veri setindeki homojen değişkenler model oluşturmada etkisi olmadığından elenmiştir. Veri setindeki model oluşturmada etkisi bulunmayan değişkenlerin elenmesi değişken seçimi (variable selection) olarak adlandırılır. Çok değişkenli veri setinde değişken seçimi yapılınca boyut indirgeme (dimension reduction) işlemi yapılmıştır [76]. Boyut indirgenince iki değişkenli veri setinde her bir değişkenin ikiye parçalandığı yapı oluşmuştur. İki değişkenli ve her bir değişkenin ikiye parçalandığı veri setinde iki, üç ve dört kümelenme merkezli aday modellerin grafikleri **Şekil 3.7.3** te gösterilmiştir.



Şekil 3.7.3. İki değişkenli ve her değişkenin ikiye parçalandığı veri setinde (a) ve (b) iki merkezli, (c), (d),(e) ve (f) üç merkezli ve (g) dört kümelenme merkezli uygun aday modellerin düzlemsel grafiklerini göstermektedir.

3.7.6. Çok Değişkenli Büyük Veride Değişkenlerdeki Parçalanmalara Dayalı Normal Karma Modellerde Uygun Aday Modellerin Oluşturulması ve Parametre Tahminleri

On beş değişkenli veri seti, heterojenliğe dayalı değişken seçimi ve boyut indirgeme işlemlerinden sonra iki boyutlu yani $p=2$ olacak şekilde iki değişkenli X_1, X_2 değişkenlerin ikiye bölündüğü yapıya dönüşmüştür. İki değişkenli veride değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerini temsil eden ortalama vektörleri, varyans-kovaryans matrisleri ve korelasyon katsayıları verinin parçalanmalarının ortalamaları ve standart sapmalarından elde edilmiştir. On beş değişkenli verinin ortalama vektörü,

$\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{1\bullet} \\ \mu_{2*} \\ \mu_3 \\ \vdots \\ \mu_{15} \end{bmatrix}$ olarak elde edilmiş ve boyut indirgeme işleminden sonra iki değişkenli veri setinin ortalama vektörü $\mu_i = \begin{bmatrix} \mu_{1x} \\ \mu_{2y} \end{bmatrix}$ olarak elde edilmiştir. $i=1,2$ olmak üzere, on beş değişkenli veri setinin aday modellerin varyans-kovaryans matrisleri

$\Sigma_i = \begin{bmatrix} {\sigma_{1\bullet}}^2 & \rho_k\sigma_{1\bullet}\sigma_{2*} & \cdots & \rho_k\sigma_{1\bullet}\sigma_{15} \\ \rho_k\sigma_{2*}\sigma_{1\bullet} & {\sigma_{2*}}^2 & \cdots & \rho_k\sigma_{2*}\sigma_{15} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k\sigma_{15}\sigma_{1\bullet} & \rho_k\sigma_{15}\sigma_{2*} & \cdots & {\sigma_{15}}^2 \end{bmatrix}$ olarak elde edilmiş ve boyut indirgeme işleminden sonra iki değişkenli veri setinin varyans-kovaryans matrisi

$\Sigma_i = \begin{bmatrix} {\sigma_{1x}}^2 & \rho_i\sigma_{1x}\sigma_{2y} \\ \rho_i\sigma_{2y}\sigma_{1x} & {\sigma_{2y}}^2 \end{bmatrix}$ olarak elde edilmiştir. Burada $x: X_1$ değişkenindeki 2 parçalanmayı, $y: X_2$ değişkenindeki 2 parçalanmayı ve $\rho = \boldsymbol{Corr}\left(\boldsymbol{X}_{1x}, \boldsymbol{X}_{2y}\right)$ Pearson korelasyon katsayısını göstermektedir.

Bu merkezindeki veriler, $N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ normal dağılımına sahip olsun. İki değişkenli veride değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ile normal

karma modeller oluşturulurken olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri kullanılmıştır. Değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin varyans-kovaryans matrislerindeki bileşenler arasındaki ilişkinin yönünü ve derecesini veren Pearson korelasyon katsayısı iki bileşenli merkez yapısı için $2x2$ tipinde matris içindeki bileşenlerde 4 adet bulunmaktadır. Varyans-kovaryans matrisi simetrik matris yapısında olduğundan her bir varyans-kovaryans matrisinde üç farklı Pearson korelasyon katsayısı bulunur. Toplam 7 uygun aday model içerisinden iki bileşenli normal karma modele karşılık gelen modeller $u = 1,2$ ve $i = 1,2$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{2} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.7.2}$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{2} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$$

iki normal karma model bulunur. Üç bileşenli normal karma modele karşılık gelen modeller $u = 3,...,6$ ve $i = 1,...,3$ için normal bileşen yoğunluk fonksiyonları

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{3} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.7.3}$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{3} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

dört normal karma model bulunur. Dört bileşenli normal karma modele karşılık gelen model $u = 7$ ve $i = 1,...,4$ için normal bileşen yoğunluk fonksiyonu

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)}) = \sum_{i=1}^{4} \pi_i^{(u)} f_i\left(x;\mu_i^{(u)},\Sigma_i^{(u)}\right) \tag{3.7.4}$$

şeklinde ifade edilir. Olasılık ağırlıkları $\pi_i^{(u)} = \frac{\pi_i}{\sum_{i=1}^{4} \pi i}$, $0 < \pi_i < 1$ olmak üzere

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2 N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_3 N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3) + \pi_4 N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$$

bir normal karma model bulunur. (3.7.2)-(2.7.4) denklemlerindeki $u = 1,...,7$ için karma modellerin yapısına uyan modellerin olasılık ağırlıklarının tahmini $\hat{\pi}_i^{(u)} = \frac{n_i}{\sum_{i=1}^{k} n_i}$, $1x = 1,2$ ve $2y = 1,2$ olmak üzere, ortalama vektörleri $\hat{\mu}_i^{(u)} = \begin{bmatrix} \bar{x}_{1x}^{(u)} \\ \bar{x}_{2y}^{(u)} \end{bmatrix}$, varyans-kovaryans matrisleri $\hat{\Sigma}_i^{(u)} = \begin{bmatrix} \left(s_{1x}^{(u)}\right)^2 & r_{1x,2y}^{(u)} s_{1x}^{(u)} s_{2y}^{(u)} \\ r_{2y,1x}^{(u)} s_{2y}^{(u)} s_{1x}^{(u)} & \left(s_{2y}^{(u)}\right)^2 \end{bmatrix}, i = 1,...,4$ için normal bileşen yoğunluk fonksiyonlarının bilinmeyen parametreleridir. Veri setindeki değişkenlerin gözlem değerlerinden elde edilen ortalama vektörünün parametre tahminleri **Tablo 3.7.7** de verilmiştir.

Tablo 3.7.19. On beş değişkenli büyük veri setindeki kümelenme merkezlerine ait ortalama vektörleri.

Parametre	μ_i			
No	$\boldsymbol{\mu}_1 = \begin{bmatrix} 21,3411 \\ 21,4425 \\ 27,5768 \\ 32,3862 \\ 18,8737 \\ 23,5194 \\ 30,3363 \\ 56,8775 \\ 64,3351 \\ 20,7174 \\ 44,3828 \\ 39,1720 \\ 36,2523 \\ 28,1805 \\ 50,1940 \end{bmatrix}$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 57,1664 \\ 21,4425 \\ 27,5768 \\ 32,3862 \\ 18,8737 \\ 23,5194 \\ 30,3363 \\ 56,8775 \\ 64,3351 \\ 20,7174 \\ 44,3828 \\ 39,1720 \\ 36,2523 \\ 28,1805 \\ 50,1940 \end{bmatrix}$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 21,3411 \\ 61,1944 \\ 27,5768 \\ 32,3862 \\ 18,8737 \\ 23,5194 \\ 30,3363 \\ 56,8775 \\ 64,3351 \\ 20,7174 \\ 44,3828 \\ 39,1720 \\ 36,2523 \\ 28,1805 \\ 50,1940 \end{bmatrix}$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 57,1664 \\ 61,1944 \\ 27,5768 \\ 32,3862 \\ 18,8737 \\ 23,5194 \\ 30,3363 \\ 56,8775 \\ 64,3351 \\ 20,7174 \\ 44,3828 \\ 39,1720 \\ 36,2523 \\ 28,1805 \\ 50,1940 \end{bmatrix}$

3.7.7. Çok Değişkenli Büyük Veri Setinde Normal Karma Modellerin Log-likelihood, AIC ve BIC değerleri

Modele dayalı kümeleme yapmak için on beş değişkenli veride homojen değişkenler değişken seçimi ile elenip veride boyut indirgeme işleminden sonra iki değişkenli ve her bir değişkenin ikiye parçalandığı veri setindeki uygun aday modellerin normal karma modelleri belirlenmiştir. Normal karma modeller arasından en iyi modelin seçimi için her bir uygun aday modelin çok değişkenli normal karma dağılımlardan log-likelihood, AIC ve BIC değerleri elde edilmiştir. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu, $i=1,2,...n$ *ve* $j=1,2,...g$ olmak üzere $f\left(x_i;\theta_j\right)$ karma olasılık yoğunluk fonksiyonunun likelihood fonksiyonu ve logaritması alınmış likelihood fonksiyonu (3.2.13) ve (3.2.14) denklemeleri kullanılarak elde edilir 73].

Çok değişkenli normal karma modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kriteri (BIC) ve Akaike bilgi kriteri (AIC) (3.2.15) ve (3.2.17) denklemeleri kullanılarak elde edilir.

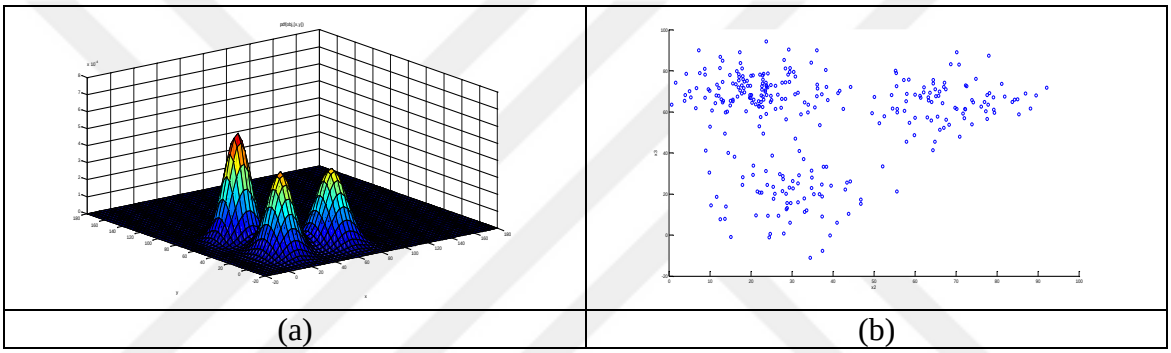
On beş değişkenli büyük veride değişken seçimi (variable selection) yapılarak homojen yapıdaki değişkenler çıkarıldıktan sonra ilk iki değişkendeki ikişer parçalanmadan kaynaklanan kümelenme merkezlerine göre uygun aday modellerin log-likelihood, AIC ve BIC değerlerinin hesaplandığı değerler **Tablo 3.7.20** de verilmiştir.

Tablo 3.7.20. İki değişkenli normal karma dağılımların modele dayalı kümelenmesi için birinci veri setinden log-l, AIC ve BIC değerleri.

Model no:	$\log L\left(\Psi\right)$	$AIC=-2\ln L\left(\Psi\right)+2d$	$BIC=-2\ln L\left(\Psi\right)+d\log n$	d
1.	-3979	7980	7958	11
2.	-4544.6	9111.1	9089.1	11
3.	-4091.8	8217.6	8183.6	17
4.	-4103.8	8241.5	8207.5	17
5.	**-3466.5**	**6967**	**6933**	**17**
6.	-4050.2	8134.5	8100.5	17
7.	-3550.9	7147.9	7101.9	23

3.7.8. İki Değişkenli Her Değişkenin İkiye Bölündüğü Normal Karma Modellerin Modele Dayalı Kümelenmesi için En İyi Modelin Seçimi

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin log-l, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. Çok değişkenli veri setinde normal karma modellerden uygun aday modellerin log-l, AIC ve BIC değerleri hesaplanmış ve **Tablo 3.7.20.** de verilmiştir. Uygun aday modeller arasından modele dayalı kümelemede en iyi model log-likelihood değer en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir. Bu değerlere göre üç kümelenme merkezli beşinci uygun aday model en iyi model olarak belirlenmiştir.



(a) (b)

Şekil 3.7.4. On beş değişkenli büyük verideki X_1, X_2 değişkenlerindeki parçalanmalardan oluşan modeller arasından en iyi modelin (a) yoğunluk fonksiyonunun yüzey grafiği (b) saçılım grafiği.

3.8. Uzaktan Algılanmış Çok Bantlı Uydu Görüntü Verisi için Karma Model Kümelemeye Dayalı Yeni bir Yarı-Denetimli Sınıflandırma Metodu

Bu çalışmada uzaktan algılanmış çok bantlı uydu görüntü verisi için yeni bir yarı-denetimli sınıflandırma (semi-supervised classification) metodu geliştirilmiştir [78]. Geliştirilen bu yarı-denetimli sınıflandırma metodunda uydu görüntü verisin alındığı banlardaki dalga boyundaki veriye uyan eğri (spectral response curves) veya dalga boyu izleri (spectral signature) kullanılarak verideki kümeler, eğitimsiz kümeleme (unsupervised clustering) ve eğitimli sınıflandırmalardan (supervised classification) meydana gelmektedir. Model seçimine dayalı karma model kümelemede, uzaktan algılanmış çok bantlı uydu görüntü verisinde spectral bantlardaki (değişkenlerdeki) heterojen yapıların, parçalanmalarının verideki kümelenme sayısını ve yapısını belirlediği ve şekillendirdiği varsayılmış ve gösterilmiştir. Uzaktan algılanmış çok bantlı uydu

görüntü verisinde değişkenlerde parçalanma olup olmadığı durum incelenmiştir. Değişkenlerdeki parçalanmalara göre veride olabilecek tüm kümelenme merkez sayıları belirlenmiştir. Tüm kümelenme merkez sayıları arasında bazıları varsayımları sağlamadığı için çıkarılmıştır. Geriye kalan kümelenme merkez sayıları mümkün kümelenme merkezlerini verir ve bunlar varsayımları sağlar. Uzaktan algılanmış çok bantlı uydu görüntü verisinde mümkün kümelenme merkezleri için kümelenme sayısını ve yapısını belirlemek amacıyla normal dağılımların karma modelleri kullanılarak aday modeller oluşturulmuştur. Oluşturulan normal dağılımların karma modelleri arasından en iyisi bilgi kriterleri kullanılarak seçilmiştir.

Verideki kümeler, dalga boyu izleri (spectral signature) kullanılarak sınıflandırılır. Eğitimli sınıflandırma metodu için ayırt edici fonksiyon olarak Öklid uzaklığı (Euclidean distance) kullanılır. Öklid uzaklığı değerleri eğitimli sınıflandırma metodu için karar kuralı olarak kullanılır. En iyi olarak seçilen normal dağılımların karma modeli verideki kümelenme sayısını ve yapısını belirler. En iyi normal dağılımların karma modelindeki bileşen sayısı uydu görüntü verisindeki kümelenme sayısına karşılık gelir. En iyi normal dağılımların karma modelindeki bileşenlerin oranları, ortalama vektörleri ve varyans – kovaryans matrisleri uzaktan algılanmış çok bantlı uydu görüntü verisindeki kümelenmelerin sırasıyla büyüklüğüne, konumuna ve yapısına karşılık gelir.

3.8.1. Uzaktan Algılanmış Uydu görüntü Verisi

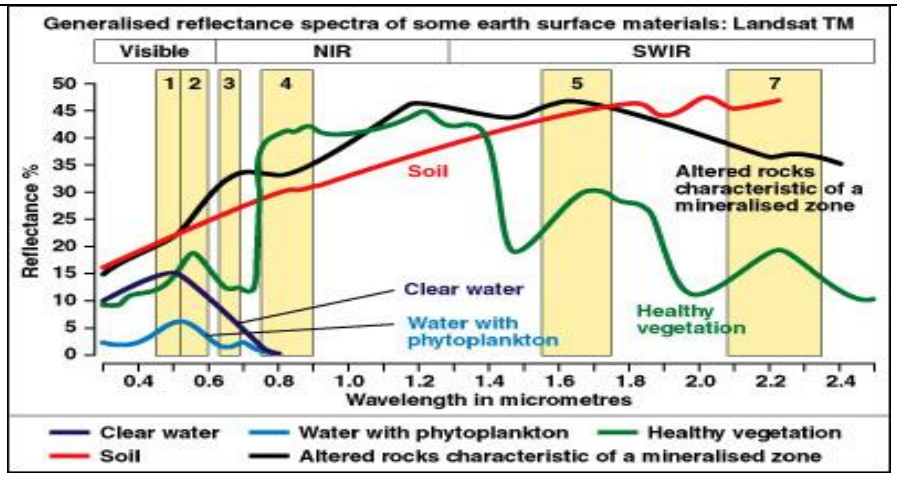
Önerilen modele dayalı karma model kümeleme metodu, bir tarımsal bölgesinin uzaktan algılanmış Landsat Thematic Mapper (Landsat TM) görüntü verisine uygulanmıştır. Çok bantlı görüntü verisi olarak Adana ili Seyhan ilçesinde ($\approx 37^{\circ}$ kuzey ve 36° doğu) bir bölgeye ait 27.03.1992 tarihli, yörünge 175- satır 34 numarası olan Landsat TM uydusunun 3., 4. ve 5. bantlarındaki $198x200$ (toplamda 39600) piksel değerleri kullanılmıştır [79]. Landsat TM uydusunun 7 bandından yukarıda belirtilen üç bandın kullanılmasının nedeni bitki örtüsü, su ve toprağın bu bantlardan alınan dalga boyuna göre daha iyi belirlenmesidir. Landsat TM uydusunun 7 bandının veri topladıkları dalga boyları ve çözünürlükleri **Tablo 3.8.1** de verilmiştir.

Tablo 3.8.1. Landsat 4 ve 5 uydusunun Thematic Mapper (TM) algılayıcısının bantları, veri topladıkları dalga boyları ve çözünürlükleri.

Uydu	Bant	Veri Topladığı Dalga Boyu	Çözünürlük(metre)
Landsat TM	Bant 1	0.45-0.52	30
	Bant 2	0.52-0.60	30
	Bant 3	0.63-0.69	30
	Bant 4	0.75-0.90	30
	Bant 5	1.55-1.75	30
	Bant 6	10.40-12.5	120
	Bant 7	2.08-2.35	30

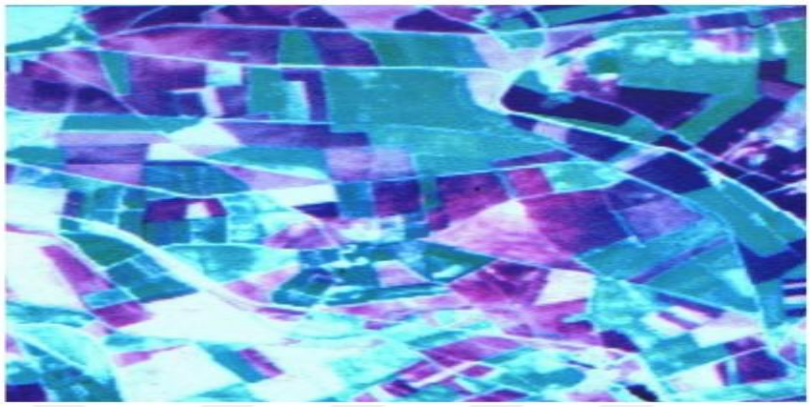
Landsat TM bantları farklı türde yeryüzü kaynaklarını belirlemede kullanılır. Kullanılan bantların algılayıcılarının temsil ettiği renklerin özelliklerine göre mesela, 4. bant kırmızı rengi temsil eder ve parlak kırmızı renk yetişen bitki örtüsünü gösterir. Toprak ya da seyrek bitki örtülü alanlar nem ve organik madde ihtiva oranına bağlı olarak beyazdan yeşile ya da kahverengiye kadar renk birleşimleri ile temsil edilir.

Ölçüm prensipleri algılayıcı tipine göre değişir. Ön işlemden geçen ölçüm türleri arazide bir bölgeye karşılık gelir ve görüntüde bir piksel ile temsil edilir. Bölgeler arası uzaklık bitişik görüntü pikselleri arasındaki uzaklık ile belirlenir. Bu uzaklık bir görüntü algılayıcısının uzaysal çözünürlüğünü belirler. Landsat TM Uydusunun bantları ve veri topladığı dalga boyları **Şekil 3.8.1** de verilmiştir.



Şekil 3.8.1. Uzaktan algılanmış çok bantlı uydu görüntü verisinin alındığı LANDSAT TM Uydusunun bantları (değişkenleri) ve veri topladığı dalga boyları.

Landsat TM 3., 4., ve 5. bantları modele dayalı kümeleme için kullanılır. Tarımsal bölge üzerinde çalışılan uydu görüntü verisinin belirttiği alanın uydu görüntüsü **Şekil 3.8.2** de gösterilmiştir.



Şekil 3.8.2. Seyhan Ovası Adana - Türkiye (≈37^{o}Kuzey, 36^{o}Doğu) bölgesinde yer alan tarımsal bölgenin 27 Mart 1992 tarihli (Yörünge 175 - Satır 34) 198 satır ve 200 sütundan oluşan Landsat TM uydu görüntüsü

3.8.2. Çok Bantlı Uydu Görüntü Verisindeki Heterojen Değişkenler için Uygun Parçalanma Sayısının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntüsünün sayısallaştırılmış verisindeki 3.bant, 4.bant ve 5.bant değerleri bu çalışmada değişken değerleri olarak alınmıştır. Veride X_1 değişkeni 3.banttaki 198×200=39600, X_2 değişkeni 4.banttaki 198×200=39600 ve X_3 değişkeni 5.banttaki 198×200=39600 gözlem değerinden oluşmaktadır. Veride toplam 39600×3=118800 gözlem değeri bulunmaktadır. Öncelikle bu üç değişkendeki verilerin homojen mi? yani dağılımının normal mi? ya da heterojen mi? yani dağılımının normal dağılımların karması mı? olduğu araştırılır. Veri setindeki her bir bant bir değişkeni temsil ettiğinden modeller oluşturulurken bu bantlardan elde edilen değişkenlerin yapıları ya da parçalanmaları kullanılmıştır [78]. Değişkenlerdeki homojen ve heterojen yapılar, yani parçalanmalar belirlenirken hem hesaplama yöntemi hem de grafiksel yöntemler kullanılmıştır. Verideki her bir değişkenin histogram ve P-P grafiklerine bakılarak da değişkendeki parçalanma sayısı kontrol edilerek belirlenmiştir. Çok bantlı uydu görüntü verisinin (3.bant) X_1 değişkenindeki, (4.bant) X_2 değişkenindeki ve (5.bant) X_3 değişkenindeki parçalanmalar önce histogram ve P-P grafikleri kullanılarak grafiksel yöntemlerle incelenmiştir. Bu durum **Şekil 3.8.3** te gösterilmiştir. Çok bantlı uydu

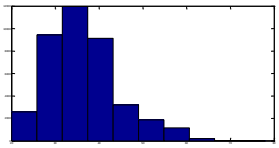
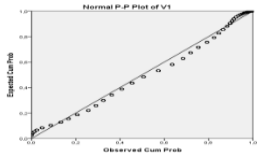
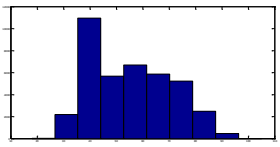
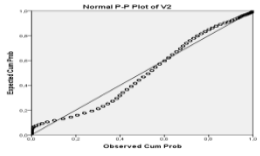
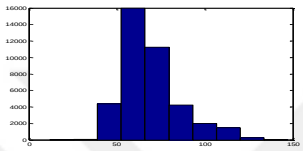
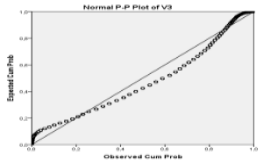
görüntü verisindeki (3.bant) X_1 değişkeni, (4.bant) X_2 değişkeni ve (5.bant) X_3 değişkeni histogram ve P-P grafikleri kullanılarak grafiksel yöntemlerle inceleme sonucunda heterojen yapıda oldukları ve bu değişkenlerdeki parçalanmaların üç olduğu belirlenmiştir. Üç değişkenli uydu görüntü verisinde değişkenlerdeki parçalanma sayılarının belirlenmesi için hesaplama yönteminde her bir değişken için tek değişkenli normal karma model kullanılarak belirlenmiştir [79]. Tek değişkenli normal dağılımların karmaları,

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i f_i(x;\mu_i,\sigma_i) \tag{3.8.1}$$

formundadır. Burada $f(x)$ normal karma modelin olasılık yoğunluk fonksiyonunu, k karma normal modeldeki bileşen (parçalanma) sayısını, π_i bileşen ağırlığını gösterir. $f_i(x;\mu_i,\sigma_i)$ bileşenin olasılık yoğunluk fonksiyonunu göstermek üzere

$$f_i(x;\boldsymbol{\mu}_i,\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{1}{2}\left(\frac{x-\boldsymbol{\mu}_i}{\sigma_i}\right)^2\right\} \tag{3.8.2}$$

şeklinde ifade edilir. Burada μ_i bileşen ortalamasını, σ_i bileşen standart sapmasını göstermektedir. Tek değişkenli normal karma model değişkendeki uygun parçalanma sayısına göre tek değişkenli normal karma modeldeki karma ağırlıkları π, ortalamalar μ ve varyanslar σ^2 tahmin edilir. Her bir değişkendeki parçalanmanın ortaya çıkarılması için tek değişkenli normal karma modelin log-likelihood, AIC ve BIC değerleri modelin parametrelerinden parçalanma sayısına bağlı olarak hesaplanır. Tek değişkenli normal karma modelden elde edilen log-likelihood değerinin en büyük, AIC ve BIC değerlerinin en küçük olduğu parçalanma sayısı en uygun parçalanma sayısı olarak belirlenir.

X_1 	X_1 
(a)	(b)
X_2 	X_2 
(a)	(b)
X_3 	X_3 
(a)	(b)

Şekil 3.8.3. Çok bantlı uydu görüntü verisinin (3.bant) X_1 değişkeni, (4.bant) X_2 değişkeni ve (5.bant) X_3 değişkeni için sırasıyla (a) histogram ve (b) P-P grafikleri.

Çok bantlı uydu görüntü verisindeki X_1 değişkeni için veri değerleri kullanılarak uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri **Tablo 3.8.2** de verilmiştir.

Tablo 3.8.2. Çok bantlı uydu görüntü verisinde X_1 değişkeni için uygun parçalanma sayısını belirlemek amacıyla, modeldeki karma ağırlıklar π, ortalama vektörler μ ve varyanslar σ^2 için parametrelerin tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri.

K	Log-L	AIC	BIC
k=1	-13947	27895	27897
k=2	-13767	27534	27538
k=3*	**-13567**	**27235**	**27242**
k=4	-13617	27236	27246
k=5	-13617	27237	27249

Tablo 3.8.2 deki sonuçlara göre X_1 değişkeni için uygun parçalanma sayısı üçtür. Çok bantlı uydu görüntü verisindeki X_2 değişkeni için veri değerleri kullanılarak uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri **Tablo 3.8.3** de verilmiştir.

Tablo 3.8.3. Çok bantlı uydu görüntü verisinde X_2 değişkeni için uygun parçalanma sayısını belirlemek amacıyla, modeldeki karma ağırlıklar π, ortalama vektörler μ ve varyanslar σ^2 için parametrelerin tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri.

K	Log-L	AIC	BIC
k=1	-16418	32837	32838
k=2	-15912	31825	31829
k=3*	**-15858**	**31717**	**31724**
k=4	-15857	31715	31725
k=5	-15857	31716	31728

Tablo 3.8.3 teki sonuçlara göre X_2 değişkeni için uygun parçalanma sayısı üçtür. Çok bantlı uydu görüntü verisindeki X_3 değişkeni için veri değerleri kullanılarak uygun parçalanma sayısını belirlemek amacıyla modeldeki karma ağırlıklar π, ortalamalar μ ve varyanslar σ^2 için tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri **Tablo 3.8.1** de verilmiştir.

Tablo 3.8.4. Çok bantlı uydu görüntü verisinde X_3 değişkeni için uygun parçalanma sayısını belirlemek amacıyla, modeldeki karma ağırlıklar π, ortalama vektörler μ ve varyanslar σ^2 için parametrelerin tahmin değerlerine dayalı olarak hesaplanan log-likelihood, AIC ve BIC değerleri.

K	Log-L	AIC	BIC
k=1	-16664	33329	33331
k=2	-16169	32340	32344
k=3*	**-16126**	**32254**	**32261**
k=4	-16134	32330	32339
k=5	-16138	32340	32352

Tablo 3.8.4 teki sonuçlara göre X_3 değişkeni için uygun parçalanma sayısı üçtür. Uzaktan algılanmış çok bantlı uydu görüntü verisindeki X_1, X_2 ve X_3 değişkenleri için log-likelihood fonksiyon değerinin maksimumu, aynı zamanda AIC ve BIC

değerlerinin minimumu kullanılarak uygun parçalanma sayısını veren karma normal model: X_1 değişkeni için $k=3$, X_2 değişkeni için $k=3$ ve X_3 değişkeni için $k=3$ olarak belirlenmiştir.

3.8.3. Çok Bantlı Uydu Görüntü Verisinde Değişkendeki Uygun Parçalanmaların Belirlenmesi

Üç değişkenli ve değişkenlerin her birinin üçe parçalandığı veri setindeki değişkenlerin anlamlı alt gruplarına düşen gözlem değerlerinin belirlenmesi için k-ortalamalar algoritması kullanılmıştır. Başlangıçta her bir değişkenin parçalanma sayısı kadar k merkez sayısı belirlenerek adımsal işlemlerle gözlemler arasındaki uzaklıklara göre merkez etrafındaki en yakın gözlemler parçalanmalara atanmaktadır. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık (3.2.3) denklemi kullanılarak hesaplanır. Burada her bir grup ya da parçalanma için gözlem değeri ile grup merkezi arasındaki uzaklıkların toplamı alınarak her bir gözlem değeri için en uygun grup merkezi seçilir. Üç değişkenli ve değişkenlerin her birinde üç anlamlı alt grup bulunan veride k-ortalamalar algoritması uygulandığında her bir değişken için üç ayrı küme merkezi belirlenmiş ve her bir küme merkezinin etrafına aralarındaki mesafe minimum olacak şekilde gözlem değerleri atanmıştır. Kümeler arası mesafenin maksimum (heterojen) aynı zamanda küme içi mesafenin minimum (homojen) olduğu durum en uygun parçalanma sayısını verir. K-ortalamalar algoritması kullanılarak verideki her bir değişken üçe parçalanmış ve her bir gözlemin düştüğü anlamlı alt grup belirlenmiştir. Üç değişkenli veride k-ortalamalar algoritması uygulandığında X_1, X_2 ve X_3 değişkenleri için parçalanma sayıları sırasıyla $k_1=3$, $k_2=3$ ve $k_3=3$ olarak elde edilmiştir. X_1 değişkenindeki parçalanmalar: X_{11}, X_{12} ve X_{13}; X_2 değişkenindeki parçalanmalar: X_{21}, X_{22} ve X_{23}; ve X_3 değişkeninde parçalanmalar: X_{31}, X_{32} ve X_{33} olarak tanımlanır ve alınır. Değişkenlerdeki parçalanmaların gözlem sayıları **Tablo 3.8.5** te verilmiştir.

Tablo 3.8.5. Üç değişkenli uydu görüntü verisinde değişkenlerdeki parçalanmalara düşen gözlem sayıları.

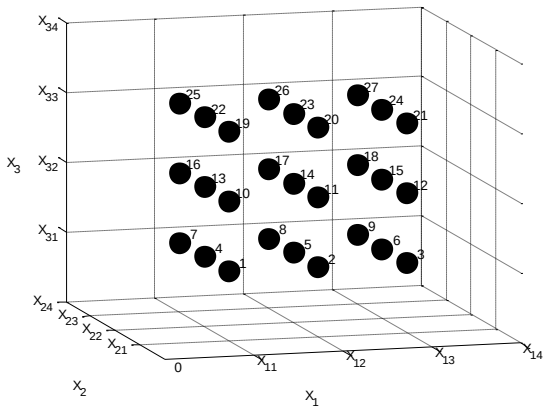
Değişken	X_1			X_2			X_3		
Değişken parçalanmaları	X_{11}	X_{12}	X_{13}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{33}
Parçalanmadaki Gözlem sayısı	4291	17241	18068	10279	12697	16624	17929	4999	16672
Toplam	39600			39600			39600		

3.8.4. Çok Bantlı Uydu Görüntü Verisinde Değişkenlerdeki Uygun Parçalanmalara Dayalı Kümelenme Merkez Sayılarının Ve Yapılarının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntü verisindeki X_1, X_2 ve X_3 değişkenlerindeki parçalanmalar sırasıyla $k_1 = 3$, $k_2 = 3$ ve $k_3 = 3$ olarak elde edilmiştir. Çok değişkenli verideki değişken (bant) sayısı $p = 3$ alınmıştır. Üç değişkenli veri seti için $C_{\min}$ ve C_{mak} ile gösterilen değişkenlerdeki parçalanmaların oluşturduğu kümelenme merkezlerinin sırasıyla minimum ve maksimum sayısı, [41] tarafından önerilen (3.2.4) ve (3.2.5) denklemeleri kullanılarak hesaplanır. Burada p, verideki değişken sayısını ve k_S, X_S değişkenindeki parçalanma sayısını göstermektedir. $n \times 3$ tipindeki X veri matrisi $X = [X_1 \quad X_2 \quad X_3]$ şeklinde matris formunda gösterilebilir. n_1 elemanlı X_1 değişkenindeki parçalanma $X_1 = \begin{bmatrix} X_{11} \\ X_{12} \\ X_{13} \end{bmatrix}$ şeklinde olur. Burada X_{11}, X_{12} ve X_{13} sırasıyla n_{11}, n_{21} ve n_{13} elemanlıdır. Yani n_1, X_1 değişkenindeki eleman sayısı olmak üzere, $n_1 = n_{11} + n_{12} + n_{13}$ dir. n_2 elemanlı X_2 değişkenindeki parçalanma $X_2 = \begin{bmatrix} X_{21} \\ X_{22} \\ X_{23} \end{bmatrix}$ şeklinde olur. Burada X_{21}, X_{22} ve X_{23} sırasıyla n_{21}, n_{22} ve n_{23} elemanlıdır. Yani n_2, X_2 değişkenindeki eleman sayısı olmak üzere,

$n_2 = n_{21} + n_{22} + n_{23}$ dir. n_3 elemanlı X_3 değişkenindeki parçalanma $X_3 = \begin{bmatrix} X_{31} \\ X_{32} \\ X_{33} \end{bmatrix}$ şeklinde olur. Burada X_{31}, X_{32} ve X_{33} sırasıyla n_{31}, n_{32} ve n_{33} elemanlıdır. Yani n_3, X_3 değişkenindeki eleman sayısı olmak üzere $n_3 = n_{31} + n_{32} + n_{33}$ dir.

Bu durumda üç değişkendeki parçalanmaların oluşturduğu kümelenme merkezlerinin minimum ve maksimum sayısı (3.2.4) ve (3.2.5) denklemlerii kullanılarak $C_{\min} = \max\left\{k_1, k_2, k_3\right\} = \max\left\{3,3,3\right\} = 3$ ve $C_{mak} = k_1 k_2 k_3 = 3.3.3 = 27$ olarak elde edilir. Üç değişkenli uzaktan algılanmış çok bantlı uydu görüntü verisinde X_1, X_2 ve X_3 değişkenlerindeki sırasıyla X_{11}, X_{12} ve X_{13}; X_{21}, X_{22} ve X_{23}; X_{31}, X_{32} ve X_{33} parçalanmalara karşılık gelen C_{mak} ile hesaplanan 27 kümelenme merkezi ve bu merkezleri meydana getiren değişkenlerdeki bileşenler **Şekil 3.8.4** ve **Tablo 3.8.6** da verilmiştir.



Şekil 3.8.4. Çok bantlı uydu görüntü verisinin X_1, X_2 ve X_3 değişkenlerindeki parçalanmalar ve bu parçalanmalara karşılık gelen kümelenme merkezleri

Tablo 3.8.6. Üç değişkenli uzaktan algılanmış çok bantlı uydu görüntü verisinde X_1, X_2 ve X_3 değişkenlerindeki sırasıyla X_{11}, X_{12} ve X_{13}; X_{21}, X_{22} ve X_{23}; X_{31}, X_{32} ve X_{33} parçalanmalara karşılık gelen C_{mak} ile hesaplanan 27 kümelenme merkezi ve bu merkezleri meydana getiren değişkenlerdeki bileşenler.

Merkez No	**Merkez Bileşenleri**	**Merkez No**	**Merkez Bileşenleri**	**Merkez No**	**Merkez Bileşenleri**
1.	(X_{11}, X_{21}, X_{31})	10.	(X_{11}, X_{21}, X_{32})	19.	(X_{11}, X_{21}, X_{33})
2.	(X_{12}, X_{21}, X_{31})	11.	(X_{12}, X_{21}, X_{32})	20.	(X_{12}, X_{21}, X_{33})
3.	(X_{13}, X_{21}, X_{31})	12.	(X_{13}, X_{21}, X_{32})	21.	(X_{13}, X_{21}, X_{33})
4.	(X_{11}, X_{22}, X_{31})	13.	(X_{11}, X_{22}, X_{32})	22.	(X_{11}, X_{22}, X_{33})
5.	(X_{12}, X_{22}, X_{31})	14.	(X_{12}, X_{22}, X_{32})	23.	(X_{12}, X_{22}, X_{33})
6.	(X_{13}, X_{22}, X_{31})	15.	(X_{13}, X_{22}, X_{32})	24.	(X_{13}, X_{22}, X_{33})
7.	(X_{11}, X_{23}, X_{31})	16.	(X_{11}, X_{23}, X_{32})	25.	(X_{11}, X_{23}, X_{33})
8.	(X_{12}, X_{23}, X_{31})	17.	(X_{12}, X_{23}, X_{32})	26.	(X_{12}, X_{23}, X_{33})
9.	(X_{13}, X_{23}, X_{31})	18.	(X_{13}, X_{23}, X_{32})	27.	(X_{13}, X_{23}, X_{33})

Üç değişkenli normal dağılımların karma modelleri kullanılarak elde edilen modele dayalı kümelemede bu merkezler ve merkezlerden elde edilen modeller araştırılacaktır. Her bir merkezi meydana getiren değişkenlerin parçalanmaları, modellerin hesaplamalarında kullanılacak parametrelerin elde edilmesinde kullanılmıştır.

3.8.5. Çok Bantlı Uydu Görüntü Verisinde Toplam Model Sayısı ve Modellerin Yapısının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntü verisindeki X_1, X_2 ve X_3 değişkenlerindeki parçalanmaların oluşturduğu kümelenme merkezleri için M_{Toplam} ile gösterilen üç değişkenli normal dağılımların karma modelleriyle oluşturulabilecek toplam model sayısı (3.2.6) denklemi kullanılarak elde edilir [41]. Uzaktan algılanmış çok bantlı

uydu görüntü verisi için $M_{Toplam} = 2^{3.3.3} - 1 = 2^{27} - 1 = 134.217.727$ olarak elde edilir. Burada çıkarılan 1 model, sabit modeldir.

3.8.6. Çok Bantlı Uydu Görüntü Verisinde Değişkenlerdeki Parçalanmalara Dayalı Uygun Aday Model Sayısının Hesaplanması

Üç değişkenli uzaktan algılanmış çok bantlı uydu görüntü verisi için normal karma modellerde değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ve parçalanmalara bağlı olarak toplam model sayısı elde edilmiştir. Normal karma modeller oluşturulurken değişkenlerdeki her parçalanmaya en az bir kümelenme merkezi karşılık gelecek şekildeki geçerli veya uygun modellerin sayısı araştırılmıştır. Bu aday modeller **Tablo 3.8.6** da belirtilen 27 merkez üzerinden her boyutta en az bir kümelenme bulunacak varsayımı ile elde edilmiştir. **Tablo 3.8.6** daki değişkenlerin parçalanmalarının oluşturduğu kümelenme merkezleri $C_{\min}$ ve C_{mak} eşitliklerinden elde edilen en az 3 ve en fazla 27 kümelenmenin olduğu varsayımı yapılmıştır. Varsayıma uyan modellere uygun aday model denilmektedir. Uygun aday model sayısı değişken sayısı ve değişkenlerdeki parçalanma sayısına bağlı olarak hesaplanmıştır. Her boyutta (bantta) en az bir merkezin bulunduğu varsayımına uyan modellerin sayısı, her bir merkezin bulunduğu boyuttaki konumuna göre kombinasyon ile elde edilebilir. Uzaktan algılanmış çok bantlı uydu görüntü verisindeki değişken sayısı ve değişkendeki gözlem sayısı büyük veri olduğundan hesaplama geliştirilen algoritma ile elde edilmiştir. Uzaktan algılanmış çok bantlı uydu görüntü verisindeki her bir değişken ve değişkenlerdeki parçalanmalara dayalı aday modellerin sayısı hesaplanmıştır. Değişkenlerin parçalanmasından sonra oluşan merkez sayısı, toplam model sayısı ve uygun aday model sayısını veren MATLAB kodu **Tablo 3.8.7** de ve modeller **Tablo 3.8.8** de verilmiştir.

Tablo 3.8.7. Çok bantlı uzaktan algılanmış uydu görüntü verisinde uygun aday modellerin sayısını veren Matlab kodu.

```
% cozum uzayi boyutları ve degisken sayisi
cozum_uzayi = [3 3 3]; degisken_sayisi = 3; nokta_sayisi = 1;
for i = 1:1:size(cozum_uzayi,2)
   nokta_sayisi = nokta_sayisi * cozum_uzayi(i);
end
%% tum ihtimalleri olustur
baslangic = power(2, degisken_sayisi) - 1;
bitis = 0;
for i = 1:1:degisken_sayisi
   bitis = bitis + power(2, nokta_sayisi + 1 - i);
end
bitis = power(2, nokta_sayisi);
for i = baslangic:1:bitis
   temp = dec2base(i,2);    temp_length = length(temp);
   for temp_zero = 1:1:nokta_sayisi - temp_length
      temp = strcat('0', temp);
   end
   if(gecerli_kontrol(temp))
dlmwrite(strcat('gecerli_333_',int2str(length(strfind(temp,'1'))),'.csv'),temp,'delimiter',' ','-
append');
   end
end
```

Tablo 3.8.8. Çok bantlı uydu görüntü verisindeki değişkenlerin parçalanmalarına dayalı oluşan merkez sayısı, toplam model sayısı ve uygun aday model sayısı.

Merkez Sayısı	**Toplam Model Sayısı**	**Uygun Aday Model Sayısı**
1	27	0
2	351	0
3	2925	36
4	17550	1890
5	80730	24300
6	296010	153828
7	888030	623106
8	2220075	1839672
9	4686825	4255194
10	8436285	8044245
11	13037895	12751803
12	17383860	17216811
13	20058300	19981143
14	20058300	20030760
15	17383860	17376516
16	13037895	13036518
18	8436285	8436123
18	4686825	4686816
19	2220075	2220075
20	888030	888030
21	296010	296010
22	80730	80730
23	17550	17550
24	2925	2925
25	351	351
26	27	27
27	1	1
Toplam	134217728	**131964460**

Uygun aday modellerin sayısı **Tablo 3.8.8** de bilgisayar bilimleri yöntemiyle hesaplandığı

$$f\left(n,m,s,k\right)=\sum_{i,\,j,t=0}^{n,m,s}(-1)^{i+j+t}\binom{n}{i}\binom{m}{j}\binom{s}{t}\binom{(n-i)\quad(m-j)\quad s-t}{\qquad\quad k} \tag{3.8.3}$$

Burada n, m ve s sırasıyla değişkenlerdeki parçalanma sayılarını göstermektedir. i, j ve t sırasıyla değişkenlerdeki parçalanmalardan kaynaklanan kümelenme merkez sayılarını ve k modeldeki kümelenme merkez sayısını göstermektedir.

3.8.7. Çok Bantlı Uydu Görüntü Verisindeki Değişkenlerin Parçalanmalarına Dayalı Uygun Durumlar için Aday Normal Karma Modellerin Oluşturulması

Çok bantlı uydu görüntü verisindeki değişkenlerde, parçalanmalardan oluşan kümelenmeler için uygun durumlarda ortaya çıkan aday normal karma modellerin bileşen ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri verideki heterojen değişkenlerdeki parçalanmalar kullanılarak örnekleme dayalı tahmin edilmektedir [49]. Heterojen değişkenlerin parçalanmalarıyla oluşan her bir kümelenme merkezi için aday normal karma modellerin bileşenlerinin karma ağırlıklarının, ortalama vektörlerinin ve varyans-kovaryans matrislerinin tahminleri **Tablo 3.8.8** deki uygun aday modellerin karma olasılık yoğunluk fonksiyonlarını elde etmek için kullanılır.

Tablo 3.8.8 de uygun aday modeller için merkez sayıları kullanılarak, 0 ve/veya 1 lerden oluşan ve 27 karakterden oluşan modellerin dizi (string) gösterimleri yapılır. Bu dizi gösterimlerinde model oluşturulurken kümelenmenin olduğu merkeze karşılık olarak "1", kümelenmenin olmadığı merkeze karşılık olarak "0" yazılır. Uygun modeller için temsili dizi gösterimler, uygun modeller için hesaplamalarda kullanılmıştır.

Aday karma modellerin her bir dizi gösterimi 131964460 uygun karma modelden birine karşılık gelmektedir. k bileşenli normal karma model $u=1,...,131964460$ için

$$f^{(u)}(x;\mu^{(u)},\Sigma^{(u)})=\sum_{i=1}^{k}\pi_i^{(u)}f_i(x;\mu_i^{(u)},\Sigma_i^{(u)}) \tag{3.8.4}$$

şeklinde elde edilir. Normal yoğunluk fonksiyonlarının bileşenleri için olasılık ağırlıkları $u=1,...,131964460$ ve $l=1,2...,\mathrm{k}$ olacak şekilde

$$\pi_i^{(u)} = \frac{\pi_i}{\sum_{l=1}^{k} \pi_l} \tag{3.8.5}$$

elde edilir. Normal yoğunluk fonksiyonlarının bileşenleri için ortalama vektörü $u = 1,...,131964460$ ve $\mathrm{p,q,r} = 1,2,3$ olmak üzere,

$$\mu_i^{(u)} = \begin{bmatrix} \mu_{1p}^{(u)} \\ \mu_{2q}^{(u)} \\ \mu_{3r}^{(u)} \end{bmatrix} \tag{3.8.6}$$

elde edilir. Normal yoğunluk fonksiyonlarının bileşenleri için varyans-kovaryans matrisi $u = 1,...,131964460$ ve $\mathrm{p,q,r} = 1,2,3$ olmak üzere,

$$\Sigma_i^{(u)} = \begin{bmatrix} (\sigma_{1p}^{(u)})^2 & \rho_{1p,2q}^{(u)} \sigma_{1p}^{(u)} \sigma_{2q}^{(u)} & \rho_{1p,3r}^{(u)} \sigma_{1p}^{(u)} \sigma_{3r}^{(u)} \\ \rho_{2q,1p}^{(u)} \sigma_{2q}^{(u)} \sigma_{1p}^{(u)} & (\sigma_{2q}^{(u)})^2 & \rho_{2q,3r}^{(u)} \sigma_{2q}^{(u)} \sigma_{3r}^{(u)} \\ \rho_{3r,1p}^{(u)} \sigma_{3r}^{(u)} \sigma_{1p}^{(u)} & \rho_{3r,2q}^{(u)} \sigma_{3r}^{(u)} \sigma_{2q}^{(u)} & (\sigma_{3r}^{(u)})^2 \end{bmatrix} \tag{3.8.7}$$

şeklinde ifade edilir.

Tablo 3.8.9. Uygun aday modeller için merkez sayıları kullanılarak, 0 ve/veya 1 elemanlarından ve 27 karakterden oluşan tam modelin dizi (string) gösterimi

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	1	1	1	1	1	1	1	1
10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	1	1	1	1	1	1	1	1
19	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
1	1	1	1	1	1	1	1	1

3.8.8. Çok Bantlı Uydu Görüntü Verisindeki Değişkenlerin Parçalanmalarına Dayalı Uygun Durumlar İçin Aday Normal Karma Modellerdeki Parametrelerin Tahmini

Veride değişkenlerdeki parçalanmalara karşılık gelen merkezlerdeki kümelenmeler için karma oranları, ortalama vektörleri ve varyans-kovaryans matrisleri, örneklemden tahmin edilmiştir. Üç değişkenli veride $i = 1,2,...,27$ için olasılık ağırlıkları, ortalama vektörleri

ve varyans-kovaryans matrisleri $u = 1,...,131964460$ ve $l = 1,2...,\mathrm{k}$ olmak üzere sırasıyla,

$$\bar{\pi}_i^{(u)} = \frac{n_i}{\sum_{l=1}^{k} n_l} \quad , \quad u = 1,...,131964460 \text{ ve } \quad p,q,r = 1,2,3 \text{ olmak üzere } \quad \bar{\mu}_l^{(u)} = \begin{bmatrix} \bar{x}_{1p}^{(u)} \\ \bar{x}_{2q}^{(u)} \\ \bar{x}_{3r}^{(u)} \end{bmatrix} \text{ ve}$$

$$\hat{\Sigma}_i^{(u)} = \begin{bmatrix} (S_{1p}^{(u)})^2 & r_{1p,2q}^{(u)} S_{1p}^{(u)} S_{2q}^{(u)} & r_{1p,3r}^{(u)} S_{1p}^{(u)} S_{3r}^{(u)} \\ r_{2q,1p}^{(u)} S_{2q}^{(u)} S_{1p}^{(u)} & (S_{2q}^{(u)})^2 & r_{2q,3r}^{(u)} S_{2q}^{(u)} S_{3r}^{(u)} \\ r_{3r,1p}^{(u)} S_{3r}^{(u)} S_{1p}^{(u)} & r_{3r,2q}^{(u)} S_{3r}^{(u)} S_{2q}^{(u)} & (S_{3r}^{(u)})^2 \end{bmatrix}$$

ile temsil edilir. Burada $p: X_1$ değişkenindeki $1,2,3$ parçalanmayı, $q: X_2$ değişkenindeki $1,2,3$ parçalanmayı ve $r: X_3$ değişkenindeki $1,2,3$ parçalanmayı göstermektedir. $\rho_i = \boldsymbol{Corr}(\boldsymbol{X}_{1\otimes}, \boldsymbol{X}_{2\bullet}, \boldsymbol{X}_{3*})$ Pearson korelasyon katsayılarını göstermektedir. Her bir kümelenme merkezindeki veriler, $N(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ normal dağılımına sahip olsun. Üç değişkenli veride değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezleri ile normal karma modeller oluşturulurken karma ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri kullanılır. Değişkenlerdeki parçalanmalara karşılık gelen kümelenme merkezlerinin varyans-kovaryans matrislerindeki bileşenler arasındaki ilişkinin yönünü ve derecesini veren Pearson korelasyon katsayısı üç bileşenli merkez yapısı için $3x3x3$ tipinde matris içindeki bileşenlerde dokuz adet bulunmaktadır. Varyans-kovaryans matrisi simetrik matris yapısında olduğundan her bir varyans-kovaryans matrisinde üç farklı Pearson korelasyon katsayısı bulunur. Modellerdeki farklı korelasyon katsayılarının sayısı değişkenlerdeki parçalanmalar kullanılarak (3.5.2) denkleminden elde edilir.

Çok bantlı uydu görüntü verisindeki $i = 1,...,27$ olmak üzere $\hat{\pi}_i^{(u)}, \hat{\mu}_i^{(u)} ve \hat{\Sigma}_i^{(u)}$ parametrelerinin örnekleme dayalı tahminleri **Tablo 3.8.10** da verilmiştir.

Tablo 3.8.10. Çok bantlı uydu görüntü verisindeki değişkenlerde parçalanmalara karşılık gelen 27 kümenin ortalama vektörü ve varyans-kovaryans matrislerinin parametre tahmini.

$\mu_1 = \begin{bmatrix} 52{,}90 \\ 75{,}96 \\ 56{,}23 \end{bmatrix}$ $\Sigma_1 = \begin{bmatrix} 19{,}156 & -{,}594 & -{,}790 \\ -{,}594 & 34{,}965 & {,}437 \\ -{,}790 & {,}437 & 29{,}263 \end{bmatrix}$	$\mu_2 = \begin{bmatrix} 29{,}17 \\ 75{,}96 \\ 56{,}23 \end{bmatrix}$ $\Sigma_2 = \begin{bmatrix} 10{,}692 & {,}158 & {,}719 \\ {,}158 & 34{,}965 & {,}437 \\ {,}719 & {,}437 & 29{,}263 \end{bmatrix}$	$\mu_3 = \begin{bmatrix} 39{,}25 \\ 75{,}96 \\ 56{,}23 \end{bmatrix}$ $\Sigma_3 = \begin{bmatrix} 9{,}691 & -{,}282 & -{,}393 \\ -{,}282 & 34{,}965 & {,}437 \\ -{,}393 & {,}437 & 29{,}263 \end{bmatrix}$
$\mu_4 = \begin{bmatrix} 52{,}90 \\ 40{,}31 \\ 56{,}23 \end{bmatrix}$ $\Sigma_4 = \begin{bmatrix} 19{,}156 & {,}114 & -{,}790 \\ {,}114 & 20{,}095 & -{,}197 \\ -{,}790 & -{,}197 & 29{,}263 \end{bmatrix}$	$\mu_5 = \begin{bmatrix} 29{,}17 \\ 40{,}31 \\ 56{,}23 \end{bmatrix}$ $\Sigma_5 = \begin{bmatrix} 10{,}692 & -{,}080 & {,}719 \\ -{,}080 & 20{,}095 & -{,}197 \\ {,}719 & -{,}197 & 29{,}263 \end{bmatrix}$	$\mu_6 = \begin{bmatrix} 39{,}25 \\ 40{,}31 \\ 56{,}23 \end{bmatrix}$ $\Sigma_6 = \begin{bmatrix} 9{,}691 & {,}906 & -{,}393 \\ {,}906 & 20{,}095 & -{,}197 \\ -{,}393 & -{,}197 & 29{,}263 \end{bmatrix}$
$\mu_7 = \begin{bmatrix} 52{,}90 \\ 58{,}13 \\ 56{,}23 \end{bmatrix}$ $\Sigma_7 = \begin{bmatrix} 19{,}156 & {,}373 & -{,}790 \\ {,}373 & 26{,}399 & {,}672 \\ -{,}790 & {,}672 & 29{,}263 \end{bmatrix}$	$\mu_8 = \begin{bmatrix} 29{,}17 \\ 58{,}13 \\ 56{,}23 \end{bmatrix}$ $\Sigma_8 = \begin{bmatrix} 10{,}692 & {,}545 & {,}719 \\ {,}545 & 26{,}399 & {,}672 \\ {,}719 & {,}672 & 29{,}263 \end{bmatrix}$	$\mu_9 = \begin{bmatrix} 39{,}25 \\ 58{,}13 \\ 56{,}23 \end{bmatrix}$ $\Sigma_9 = \begin{bmatrix} 9{,}691 & -{,}083 & -{,}393 \\ -{,}083 & 26{,}399 & {,}672 \\ -{,}393 & {,}672 & 29{,}263 \end{bmatrix}$
$\mu_{10} = \begin{bmatrix} 52{,}90 \\ 75{,}96 \\ 102{,}42 \end{bmatrix}$ $\Sigma_{10} = \begin{bmatrix} 19{,}156 & -{,}594 & 1{,}164 \\ -{,}594 & 34{,}965 & 3{,}192 \\ 1{,}164 & 3{,}192 & 109{,}734 \end{bmatrix}$	$\mu_{11} = \begin{bmatrix} 29{,}17 \\ 75{,}96 \\ 102{,}42 \end{bmatrix}$ $\Sigma_{11} = \begin{bmatrix} 10{,}692 & {,}158 & -2{,}117 \\ {,}158 & 34{,}965 & 3{,}192 \\ -2{,}117 & 3{,}192 & 109{,}734 \end{bmatrix}$	$\mu_{12} = \begin{bmatrix} 39{,}25 \\ 75{,}96 \\ 102{,}42 \end{bmatrix}$ $\Sigma_{12} = \begin{bmatrix} 9{,}691 & -{,}282 & -{,}216 \\ -{,}282 & 34{,}965 & 3{,}192 \\ -{,}216 & 3{,}192 & 109{,}734 \end{bmatrix}$
$\mu_{13} = \begin{bmatrix} 52{,}90 \\ 40{,}31 \\ 102{,}42 \end{bmatrix}$	$\mu_{14} = \begin{bmatrix} 29{,}17 \\ 40{,}31 \\ 102{,}42 \end{bmatrix}$	$\mu_{15} = \begin{bmatrix} 39{,}25 \\ 40{,}31 \\ 102{,}42 \end{bmatrix}$

$\Sigma_{13} = \begin{bmatrix} 19,156 & ,114 & 1,164 \\ ,114 & 20,095 & ,740 \\ 1,164 & ,740 & 109,734 \end{bmatrix}$	$\Sigma_{14} = \begin{bmatrix} 10,692 & -,080 & -2,117 \\ -,080 & 20,095 & ,740 \\ -2,117 & ,740 & 109,734 \end{bmatrix}$	$\Sigma_{15} = \begin{bmatrix} 9,691 & ,906 & -,216 \\ ,906 & 20,095 & ,740 \\ -,216 & ,740 & 109,734 \end{bmatrix}$
$\mu_{16} = \begin{bmatrix} 52,90 \\ 58,13 \\ 102,42 \end{bmatrix}$ $\Sigma_{16} = \begin{bmatrix} 19,156 & ,373 & 1,164 \\ ,373 & 26,399 & -2,438 \\ 1,164 & -2,438 & 109,734 \end{bmatrix}$	$\mu_{17} = \begin{bmatrix} 29,17 \\ 58,13 \\ 102,42 \end{bmatrix}$ $\Sigma_{17} = \begin{bmatrix} 10,692 & ,545 & -2,117 \\ ,545 & 26,399 & -2,438 \\ -2,117 & -2,438 & 109,734 \end{bmatrix}$	$\mu_{18} = \begin{bmatrix} 39,25 \\ 58,13 \\ 102,42 \end{bmatrix}$ $\Sigma_{18} = \begin{bmatrix} 9,691 & -,083 & -,216 \\ -,083 & 26,399 & -2,438 \\ -,216 & -2,438 & 109,734 \end{bmatrix}$
$\mu_{19} = \begin{bmatrix} 52,90 \\ 75,96 \\ 73,05 \end{bmatrix}$ $\Sigma_{19} = \begin{bmatrix} 19,156 & -,594 & -,475 \\ -,594 & 34,965 & 1,046 \\ -,475 & 1,046 & 37,488 \end{bmatrix}$	$\mu_{20} = \begin{bmatrix} 29,17 \\ 75,96 \\ 73,05 \end{bmatrix}$ $\Sigma_{20} = \begin{bmatrix} 10,692 & ,158 & ,541 \\ ,158 & 34,965 & 1,046 \\ ,541 & 1,046 & 37,488 \end{bmatrix}$	$\mu_{21} = \begin{bmatrix} 39,25 \\ 75,96 \\ 73,05 \end{bmatrix}$ $\Sigma_{21} = \begin{bmatrix} 9,691 & -,282 & -,129 \\ -,282 & 34,965 & 1,046 \\ -,129 & 1,046 & 37,488 \end{bmatrix}$
$\mu_{22} = \begin{bmatrix} 52,90 \\ 40,31 \\ 73,05 \end{bmatrix}$ $\Sigma_{22} = \begin{bmatrix} 19,156 & ,114 & -,475 \\ ,114 & 20,095 & 1,193 \\ -,475 & 1,193 & 37,488 \end{bmatrix}$	$\mu_{23} = \begin{bmatrix} 29,17 \\ 40,31 \\ 73,05 \end{bmatrix}$ $\Sigma_{23} = \begin{bmatrix} 10,692 & -,080 & ,541 \\ -,080 & 20,095 & 1,193 \\ ,541 & 1,193 & 37,488 \end{bmatrix}$	$\mu_{24} = \begin{bmatrix} 39,25 \\ 40,31 \\ 73,05 \end{bmatrix}$ $\Sigma_{24} = \begin{bmatrix} 9,691 & ,906 & -,129 \\ ,906 & 20,095 & 1,193 \\ -,129 & 1,193 & 37,488 \end{bmatrix}$
$\mu_{25} = \begin{bmatrix} 52,90 \\ 58,13 \\ 73,05 \end{bmatrix}$ $\Sigma_{25} = \begin{bmatrix} 19,156 & ,373 & -,475 \\ ,373 & 26,399 & ,376 \\ -,475 & ,376 & 37,488 \end{bmatrix}$	$\mu_{26} = \begin{bmatrix} 29,17 \\ 58,13 \\ 73,05 \end{bmatrix}$ $\Sigma_{26} = \begin{bmatrix} 10,692 & ,545 & ,541 \\ ,545 & 26,399 & ,376 \\ ,541 & ,376 & 37,488 \end{bmatrix}$	$\mu_{27} = \begin{bmatrix} 39,25 \\ 58,13 \\ 73,05 \end{bmatrix}$ $\Sigma_{27} = \begin{bmatrix} 9,691 & -,083 & -,129 \\ -,083 & 26,399 & ,376 \\ -,129 & ,376 & 37,488 \end{bmatrix}$

3.8.9. Çok Bantlı Uydu Görüntü Verisindeki Değişkenlerin Parçalanmalarına Dayalı Uygun Durumlar için Aday Normal Karma Modellerin Log-Likelihood, AIC ve BIC Değerleri

Heterojen veride en iyi kümelenme yapısını normal dağılımların karma modellerini kullanarak modele dayalı belirlemek amacıyla her uygun model için birinci kriter olarak log-likelihood fonksiyonu değeri hesaplanır. Uygun modeller için log-likelihood fonksiyonların değerleri en iyi modeli seçmek için bir kriter olarak kullanılmıştır. Çok değişkenli normal karma modellerin log-likelihood fonksiyonu, $i = 1, 2, ...n \ \ ve \ \ j = 1, 2, ...g$ olmak üzere $f\left(x_i; \theta_j\right)$ karma olasılık yoğunluk fonksiyonu ve logaritması alınmış likelihood fonksiyonu (3.2.13) ve (3.2.14) denklmeleri kullanılarak elde edilir.

Çok değişkenli normal karma modellerin log-likelihood fonksiyonuna bağlı olarak Bayesci bilgi kritri (BIC) ve Akaike bilgi kriteride (AIC) (3.2.15) ve (3.2.17) denklemeleri kullanılarak elde edilir.

Tablo 3.8.11. Çok bantlı uydu görüntü verisinde normal karma dağılımların modele dayalı kümelenmesi için log-l, AIC ve BIC değerleri.

Log-l	**AIC**	**BIC**	**Matris gösterimi**
-105752585,4	211505668,8433	211507806,9028	1101111111111101111111111111

Çok bantlı uydu görüntü verisindeki değişkenlerdeki parçalanmalardaki karşılık kümelenme merkezlerinin sayısı ve yerlerine göre belirlenene ve **Tablo 3.8.8** de sayısı verilen aday modellerin **Tablo 3.8.9** daki dizi gösterimi olarak kaydedilmiştir. 131964460 aday modelin her birinin dizi gösterimini modeldeki kümelenmenin bulunduğu bileşenlere göre okuyan, hesaplayan Matlab kodu yazılmış ve **Tablo 3.8.12** de verilmiştir.

Tablo 3.8.12. Çok bantlı uydu görüntü verisinde normal karma dağılımların modele dayalı kümelenmesi için log-l, AIC ve BIC değerlerini veriden okutup hesaplatan Matlab kodu.

```
% Verinin dosyadan okutulması
DELIMITER = ' ';HEADERLINES = 0;
fileToRead1 = 'D:\geçerli 3_boyut uydu\gecerli_333_3_cift.csv';
% Verinin Matlab dosyasına yazdırılması
rawData1 = importdata(fileToRead1, DELIMITER, HEADERLINES);
 [~,name] = fileparts(fileToRead1);newData1.(genvarname(name)) = rawData1;
% Create new variables in the base workspace from those fields.
vars = fieldnames(newData1);for i = 1:length(vars)
   assignin('base', vars{i}, newData1.(vars{i}));
end
% parcalanma degiskenleri
a=load('tamv11.txt');b=load('tamv12.txt');c=load('tamv13.txt');d=load('tamv21.txt');e=load('tamv22.txt');f=l
oad('tamv23.txt');g=load('tamv31.txt');h=load('tamv32.txt');l=load('tamv33.txt');
size_p(1) = length(a) + length(d)+ length(g);size_p(2) = length(b) + length(d)+ length(g);size_p(3) =
length(c) + length(d)+ length(g);size_p(4) = length(a) + length(e)+ length(g);size_p(5) = length(b) +
length(e)+ length(g);size_p(6) = length(c) + length(e)+ length(g);size_p(7) = length(a) + length(f)+
length(g);size_p(8) = length(b) + length(f)+ length(g);size_p(9) = length(c) + length(f)+
length(g);size_p(10) = length(a) + length(d)+ length(h);size_p(11) = length(b) + length(d)+
length(h);size_p(12) = length(c) + length(d)+ length(h);size_p(13) = length(a) + length(e)+
length(h);size_p(14) = length(b) + length(e)+ length(h);size_p(15) = length(c) + length(e)+
length(h);size_p(16) = length(a) + length(f)+ length(h);size_p(17) = length(b) + length(f)+
length(h);size_p(18) = length(c) + length(f)+ length(h);size_p(19) = length(a) + length(d)+
length(l);size_p(20) = length(b) + length(d)+ length(l);size_p(21) = length(c) + length(d)+
length(l);size_p(22) = length(a) + length(e)+ length(l);size_p(23) = length(b) + length(e)+
length(l);size_p(24) = length(c) + length(e)+ length(l);size_p(25) = length(a) + length(f)+
length(l);size_p(26) = length(b) + length(f)+ length(l);size_p(27) = length(c) + length(f)+ length(l);
x(1,:)=load('tamv1.txt');x(2,:)=load('tamv2.txt');x(3,:)=load('tamv3.txt');size_x = size(x,1);
% ortalama vektörleri
mu(:,1) = [mean(a) mean(d) mean(g)];mu(:,2) = [mean(b) mean(d) mean(g)];mu(:,3) = [mean(c) mean(d)
mean(g)];mu(:,4) = [mean(a) mean(e) mean(g)];mu(:,5) = [mean(b) mean(e) mean(g)];mu(:,6) = [mean(c)
mean(e) mean(g)];mu(:,7) = [mean(a) mean(f) mean(g)];mu(:,8) = [mean(b) mean(f) mean(g)];mu(:,9) =
[mean(c) mean(f) mean(g)];mu(:,10) = [mean(a) mean(d) mean(h)];mu(:,11) = [mean(b) mean(d)
mean(h)];mu(:,12) = [mean(c) mean(d) mean(h)];mu(:,13) = [mean(a) mean(e) mean(h)];mu(:,14) =
[mean(b) mean(e) mean(h)];mu(:,15) = [mean(c) mean(e) mean(h)];mu(:,16) = [mean(a) mean(f)
mean(h)];mu(:,17) = [mean(b) mean(f) mean(h)];mu(:,18) = [mean(c) mean(f) mean(h)];mu(:,19) =
[mean(a) mean(d) mean(l)];mu(:,20) = [mean(b) mean(d) mean(l)];mu(:,21) = [mean(c) mean(d)
mean(l)];mu(:,22) = [mean(a) mean(e) mean(l)];mu(:,23) = [mean(b) mean(e) mean(l)];mu(:,24) = [mean(c)
mean(e) mean(l)];mu(:,25) = [mean(a) mean(f) mean(l)];mu(:,26) = [mean(b) mean(f) mean(l)];mu(:,27) =
[mean(c) mean(f) mean(l)];
%Varyans-kuvaryans matrisleri
sigma_p(1,:,:) = [ 19.156 -0.594 -0.790;-0.594 34.965 0.437;-0.790 0.437 29.263 ];sigma_p(2,:,:) = [
10.962 0.158 0.719;0.158 34.965 0.437;0.719 0.437 29.263 ];sigma_p(3,:,:) = [ 9.691 -0.282 -0.393;-0.282
34.965 0.437;-0.393 0.437 29.263 ];sigma_p(4,:,:) = [ 19.156 0.114 -0.790; 0.114 20.095 -0.197; 0.719 -
0.197 29.263 ];sigma_p(5,:,:) = [ 10.962 -0.080 0.719;-0.080 20.095 -0.197; 0.719 -0.197 29.263 ];
sigma_p(6,:,:) = [ 9.691  0.906 -0.393; 0.906 20.095 -0.197;-0.393 -0.197 29.263 ];sigma_p(7,:,:) = [
19.156  0.373 -0.790; 0.373 26.399 0.672;-0.790 0.672 29.263 ];sigma_p(8,:,:) = [ 10.962  0.545  0.719;
0.545 26.399 0.672; 0.719 0.672 29.263 ];sigma_p(9,:,:) = [ 9.691 -0.083 -0.393;-0.083 26.399 0.672;-
0.393 0.672 29.263 ];sigma_p(10,:,:) = [ 19.156 -0.594 1.164;-0.594 34.965 3.192;1.164 3.192 109.734 ];
sigma_p(11,:,:) = [ 10.962 0.158 -2.117;0.158 34.965 3.192;-2.117 3.192 109.734 ];sigma_p(12,:,:) = [
9.691 -0.282 -0.216;-0.282 34.965 3.192;-0.216 3.192 109.734 ];sigma_p(13,:,:) = [ 19.156 0.114 1.164;
0.114 20.095  0.740; 1.164  0.740 109.734 ];sigma_p(14,:,:) = [ 10.962 -0.080 -2.117;-0.080 20.095  0.740;
-2.117  0.740 109.734 ];sigma_p(15,:,:) = [ 9.691  0.906 -0.216; 0.906 20.095  0.740;-0.216  0.740 109.734
];sigma_p(16,:,:) = [ 19.156  0.373 1.164; 0.373 26.399 -2.438;1.164 -2.438 109.734 ];sigma_p(17,:,:) = [
10.962  0.545  -2.117; 0.545 26.399 -2.438;  -2.117 -2.438 109.734 ];sigma_p(18,:,:) = [ 9.691 -0.083 -
0.216;-0.083 26.399 -2.438;-0.216 -2.438 109.734 ];sigma_p(19,:,:) = [ 19.156 -0.594 -0.475;-0.594 34.965
```

```
1.046;-0.475 1.046 37.488 ];sigma_p(20,:,:) = [ 10.962 0.158 0.541 ;0.158 34.965 1.046;0.541 1.046
37.488 ];sigma_p(21,:,:) = [ 9.691 -0.282 -0.129;-0.282 34.965 1.046;-0.129 1.046 37.488 ];sigma_p(22,:,:)
= [ 19.156 0.114 -0.475; 0.114 20.095 1.193; -0.475 1.193 37.488 ];sigma_p(23,:,:) = [ 10.962 -0.080
0.541;-0.080 20.095 1.193; 0.541 1.193 37.488 ];sigma_p(24,:,:) = [ 9.691 0.906 -0.129; 0.906 20.095
1.193;-0.129 1.193 37.488 ];sigma_p(25,:,:) = [ 19.156 0.373 -0.475; 0.373 26.399 0.376;-0.475 0.376
37.488 ];sigma_p(26,:,:) = [ 10.962 0.545 0.541; 0.545 26.399 0.376; 0.541 0.376 37.488
];sigma_p(27,:,:) = [ 9.691 -0.083 -0.129;-0.083 26.399 0.376;-0.129 0.376 37.488 ];
% sadece gecerli olan modeller
ihtimal = rawData1;
for i = 1:1:length(ihtimal(:,1))
   % pdf ve top size hesapla
   top_size = 0;
   clear pdf;   m = 1;
   for j = 1:1:length(ihtimal(1,:))
      if(1 == ihtimal(i,j))
         for k=1:length(x(1,:))
            clear sigma_p_temp;
            sigma_p_temp(:,:)  =  sigma_p(j,:,:);                    pdf(m,k)=(1/(2*pi))...
*(1/((det(sigma_p_temp))^(1/2)))...                    *exp((-0.5)*(x(:,k)-mu(:,j))'...
*inv(sigma_p_temp)...                *(x(:,k)-mu(:,j)));
         end
         m = m + 1;          top_size = top_size + size_p(j);
      end
   end
   size_pdf = size(pdf,1);
   % p hesapla
   clear p;   m = 1;
   for j = 1:1:27
      if(1 == ihtimal(i,j))          p(m) = size_p(j)/top_size;          m = m + 1;
      end
   end
   % loglikelihood AIC ve BIC hesapla
   clear lnl_mul;   for j = 1:1:length(p(1,:))       lnl_mul(j,:) = p(j).*pdf(j,:);
   end
   lnl(i) = -(log(size(pdf,2)) * ((size_pdf - 1) + (size_pdf * size_x) + (size_pdf * (size_x * ((size_x + 1) /
2)))))+...
      ((size_pdf - 1) + (size_pdf * size_x) + (size_pdf * (size_x * ((size_x + 1) /
2))))*sum(log(sum(lnl_mul)));
   aic(i) = (-2 * lnl(i)) + (2 * ((size_pdf - 1) + (size_pdf * size_x) + (size_pdf * (size_x * ((size_x + 1) /
2)))));
   bic(i) = (-2 * lnl(i)) + (log(size(pdf,2)) * ((size_pdf - 1) + (size_pdf * size_x) + (size_pdf * (size_x *
((size_x + 1) / 2)))));
end
```

3.8.10. Çok Bantlı Uydu Görüntü Verisinin Normal Karma Modellerin Modele Dayalı Kümelenmesi için En İyi Modelin Seçimi

Modele dayalı kümeleme için çok değişkenli normal karma modeller arasından en iyi modelin seçimi modellerin log-l, AIC ve BIC değerlerine dayalı olarak belirlenmektedir. Üç değişkenli veride normal karma modellerden uygun aday modellerin log-l, AIC ve BIC değerleri büyük veriden algoritma yarımı ile 131.964.460 uygun model arasından hesaplanmış ve **Tablo 3.8.11** de verilmiştir.

Uygun aday modeller arasından modele dayalı kümelemede en iyi model log-likelihood değer en büyük aynı zamanda AIC ve BIC değerleri en küçük olan modeldir. Uygun aday model modelin tamamının log-likelihood, AIC ve BIC değerleri hesaplanmış aralarından en iyi model yirmi beş merkezli modeller arasından 60. Sıradaki "1101111111111011111111111111" temsili dizi gösterime sahip model olarak belirlenmiştir.

3.8.11. Spectral Signatures Kullanarak Eğitimli Sınıflandırma

Çok bantlı uydu görüntü verisindeki kümeler spektral signatures kullanılarak sınıflandırılır [81]. Öklid uzaklığı eğitimli sınıflandırma metodunda ayrıştırma fonksiyonu olarak kullanılır [82]. Öklid uzaklığının değerleri eğitimli sınıflandırmada karar kuralı olarak kullanılır. En iyi modeldeki kümelerin sınıflandırma hata matrisi veya olasılık değerleri **Tablo 3.8.13** de verilmiştir.

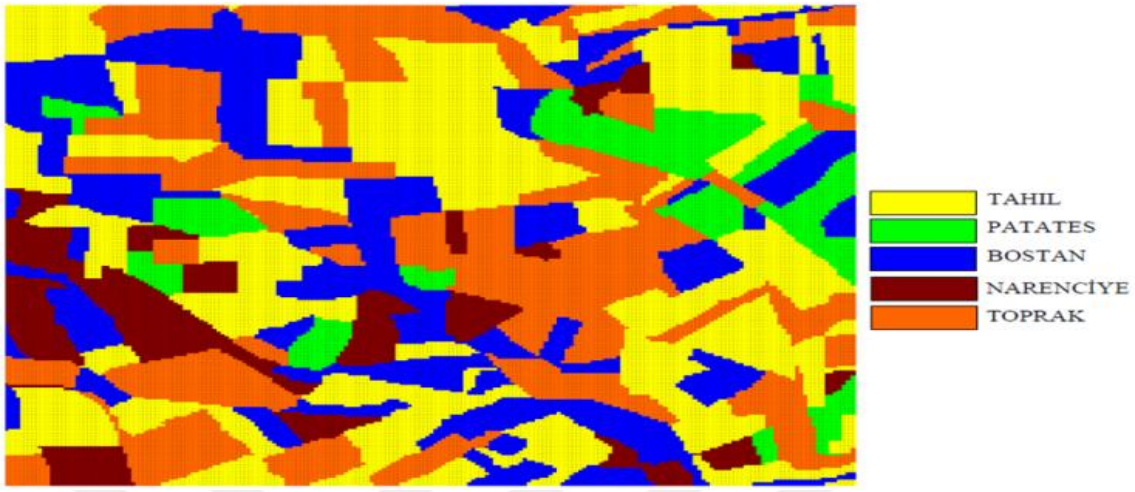
Tablo 3.8.13. Sınıflandırma hata matrisi veya olasılık değerleri

	Tahıl	Patates	Bostan	Narenciye	Toprak	Toplam	Kullanıcı Doğruluğu
Tahıl	12989	235	33	0	273	13530	0,96
Patates	502	3724	79	24	312	4641	0,80
Bostan	0	27	11194	0	86	11307	0,99
Narenciye	188	0	0	1379	0	1567	0,88
Toprak	191	0	151	0	8213	8555	0,96
Toplam	13870	3986	11457	1403	8884	39600	
Üretici Doğruluğu	0,94	0,93	0,98	0,98	0,92		0,95

Toplam doğruluk düzeyi, **Tablo 3.8.13** den 0.95 olarak elde edilir. Çok bantlı uydu görüntü verisinin sınıflandırılması için karma model kümelemeye dayalı eğitimli sınıflandırma için **0.95** olarak elde edilmiştir.

Verideki heterojen değişkenlerin parçalanmaları kullanılarak uzaktan algılanmış çok bantlı uydu görüntü verisinin karma normal modele dayalı kümelenmesinden sonra oluşan herbir kümenin içeriği eğitimli sınıflandırma yöntemiyle gerçekleştirilmiştir. Eğitimli sınıflandırmada [81] tarafından çalışılan kontrol sınıfları kullanılmıştır. Sınıflandırma sonucunda %95 lik bir sınıflandırma doğruluk yüzdesi elde edilmiştir. Uzaktan algılanmış çok bantlı uydu görüntü verisinin kümelenmesinden sonra

sınıflandırma sonucu renklendirilmiş haritası ve her rengin temsil ettiği kategoriler **Şekil 3.8.5** te gösterilmiştir.



Şekil 3.8.5. Uzaktan algılanmış çok bantlı uydu görüntü verisinin kümelenmesinden sonra sınıflandırma sonucu renklendirilmiş haritası ve her rengin temsil ettiği kategoriler.

4. BÖLÜM

SONUÇ VE ÖNERİLER

Çok değişkenli normal karma modellerin modele dayalı olarak ayrıştırılması, sınıflandırılması ve kümelenmesi üzerine yazılan bu tez çalışması dört bölümden meydana gelmektedir. Birinci bölümde karma dağılım modelinin uygulama alanları, çok değişkenli kümeleme analizi ve karma kümeleme analizi ile ilgili literatürde yer alan önceki çalışmalar ve çok değişkenli kümeleme analizinde temel kavramlar, tanımlar, gösterimler ve yöntemler anlatılmıştır.

İkinci bölümde çok değişkenli karma normal dağılım modeline dayalı kümeleme analizinde kullanılan yöntemler verilmiştir. Ayrıca sonlu karma dağılımlar ve modele dayalı kümeleme yöntemleri anlatılmış bu yöntemlerle ilgili genel tanımlar, kavramlar, teoremler ve ispatlarına yer verilmiştir. Tezin ilk iki bölümünde modele dayalı karma normal kümeleme analizi için önemli olan çalışmalara ve genel kavramlara yer verilmiş ve bu verilen kavramlar tezdeki çalışmalarda kullanılmıştır.

Tezin üçüncü bölümünde yapılan çalışmalar sekiz alt bölümde anlatılmış ve bulgular verilmiştir. Kesim 3.1 de modele dayalı karma normal kümeleme analizi için geliştirilen genetik algoritma ve genetik algoritmanın adımları tanıtılmıştır. Kesim 3.2 de geliştirilen modele dayalı kümeleme analizinde en iyi modelin seçimi için temel sayılan baz çalışma anlatılmıştır. Bu çalışmada üç farklı sentetik veri seti elde edilmiş ve algoritma bu üç veri seti üzerinde uygulanmıştır. İlk veri seti iki değişkenli ve her değişkenin ikiye bölündüğü durumdur. Bu veri setinde en iyi model iki bileşenli model olarak elde edilmiştir. İkinci veri seti iki değişkenli ve her değişkenin ikiye bölündüğü durumda en iyi modelin üç bileşenli model olarak belirlendiği veri setidir. Üçüncü veri seti ilk iki veri seti gibi iki değişkenli ve her değişkenin ikiye bölündüğü ve en iyi modelin dört bileşenli model olarak belirlendiği veri setidir. Geliştirilen kümeleme analizi yöntemi için yapılan ve

farklı veri setlerinde uygulanan ilk çalışma veri setlerindeki değişken yapısını ve kümelenmeleri en iyi şekilde ortaya çıkarmıştır.

Kesim 3.3 de farklı ve öncekinden biraz daha kompleks bir veri seti simülasyon ile üretilmiştir. İki değişkenli ve her değişkenin üçe parçalandığı veri setinde en iyi kümelenmenin olduğu model üç bileşenli model olarak belirlenmiştir. Kesim 3.4 de gerçek veri seti üzerinde verideki parçalanmalara dayalı karma model kümeleme yöntemi uygulanmıştır. Literatürde Ruspini verisi olarak bilinen iki değişkenli ve her değişkenin dörde parçalandığı veri setinde en iyi kümelenmenin olduğu model dört bileşenli model olarak belirlenmiştir. Literatürde Ruspini verisi üzerine yapılan kümeleme çalışmaları ile aynı kümelenme yapısını elde eden yöntem diğer çalışmalara göre daha hızlı ve daha az işlem karmaşıklığı ile sonuç almaktadır. Kesim 3.5'de ilk üç veri setinden farklı olarak üç değişkenli veri seti üzerinde yöntem uygulanmıştır. Değişkenlerde sırasıyla bir, iki ve üç parçalanmanın belirlendiği veri setinde çok değişkenli normal karma modeller içerisinden en iyi modelin bilgi kriterlerine göre belirlenmesi sağlanmıştır. Üç değişkenli veri setinde değişken seçimi ve boyut indirgeme işlemleri yapılmış ve en iyi model üç bileşenli model olarak elde edilmiştir. Kesim 3.6 da üç değişkenli ve değişkenlerde bir, iki ve iki parçalanmanın bulunduğu veri seti kullanılmıştır. Bu çalışmada değişken seçimi yapıldıktan sonra metodun uygulanması ile veri setinin Kesim 3.2 deki veri seti için uygulanan yöntem ile aynı olduğu belirlenmiştir. Yani değişken seçimi ve boyut indirgeme ile değişkenlerin heterojen yapısının kümelenmeyi belirlediği gösterilmiştir.

Kesim 3.7 de geliştirilen kümeleme yönteminin veri madenciliğinde büyük veri için uygulaması gösterilmiştir. On beş değişkenli veri setinde heterojen değişkenlere dayalı modeller arasından en iyi küme yapısı ve model belirlenmiştir. Üçüncü bölümün son uygulaması olarak gerçek veri seti üzerinde genetik algoritma uygulanmıştır. Uzaktan algılanmış uydu görüntü verisinin yarı eğitimli sınıflandırılması normal karma modellere göre model seçimi yöntemi ile anlatılmış ve daha önceki çalışmalara göre yüksek başarı elde edilmiştir. Karma kümeleme analizine dayalı olarak elde edilen genel sınıflandırma doğruluk yüzdesi %95 olarak elde edilmiştir.

Çok değişkenli normal karma modellerde modele dayalı kümeleme analizinde geliştirilen yöntem daha farklı ve büyük boyutlardaki veri setleri için genelleştirilebilir.

Dördüncü bölüm olan son bölümde sonuç ve öneriler verilmiştir.

KAYNAKLAR

1. Everıtt, B.S., Landau, S. and Leese, M. (2001). Cluster Analysis. Oxford University Press Inc., New York, NY.

2. Scrucca, L. (2010). Dimension reduction for model-based clustering. **Statistics and Computing, 20**(4), 471-484.

3. Mclachlan, G.J. and Basford, K.E. (1988). Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

4. Banfield, J. D., and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. **Biometrics,** 803-821.

5. Bozdogan, H. (1994b). Mixture-Model Cluster Analysis Using Model Selection Criteria And A New Informational Measure Of Complexity. In Multivariate Statistical Modeling, Vol. 2, H. Bozdogan (ed.), Kluwer Academic Publishers, Dordrecht, the Netherlands, 1994, pp. 69-113.

6. Fraley, C. and Raftery, A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. **Journal of the American Statistical Association, 97**, 611-631.

7. McLachlan, G. J. and Chang, S. U. (2004). Mixture Modelling for Cluster Analysis. **Statistical Methods in Medical Research 13,** 347-361.

8. Fraley, C. and Raftery, A.E. (1998). How many clusters? Which clustering method?-Answers via model-based cluster analysis. **Computer Journal, 41**, 578-588.

9. Weldon, W.F.R., 1892. Certain Correlated Variations in Crangon Vulgaris. Proceedings of the Royal Society of London, 51:2-21.

10. Weldon, W.F.R., 1893. On Certain Correlated Variations in Carcinus Moenas. **Proceedings of the Royal Society of London, 54**:318-329.

11. McLachlan, G. J. and Peel, D. (2000). Finite Mixture Models. Wiley, New York.

12. Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. **Series B (Methodological), 10**(2), 159-203.

13. Wolfe, J.H. (1965). A computer Program for the Maximum-Likelihood Analysis of Types, USNPRA Technical Bulletin 65-15, U.S. Naval Personal Research Activity, San Diego.

14. Binder, D.A. (1978). Bayesian Cluster Analysis, **Biometrika, 65**, 31-38.

15. Symons, M.J. (1981). Clustering Criteria and Multivariate Normal Mixtures. **Biometrics, 37**, 35-43.

16. Reaven, G.M. and Miller, R.G. (1979). An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis. **Diabetologia, 16**, 17– 24.

17. Render, R.A and Walker, H. F. (1984) Mixture Densities, Maximum Likelihood and The EM Algorithm. **Society for Industrial and Applied Mathematics 26** (2), 195-239.

18. Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm for Clustering and Two Stochastic versions. **Computational Statistics and Data Analysis, 14**, 315-332.

19. Banfield, J. D., and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. **Biometrics,** 803-821.

20. Bozdogan, H. (1994a). Choosing the number of clusters, subset selection of variables, and outlier detection in the standart mixture model cluster anlysis. Invited paper in New Approaches in Classification and Data Ana lysis, E. Diday et al. (Eds.), Springer-Verlang, New York, pp. 169-177.

21. Hastıe, T And Tıbshıranı, R. (1996) Discriminant Analysis by Gaussian Mixtures. **J. R. Statist. Soc. 58**(1), 155-176

22. Erol, H., & Akdeniz, F. (1996). A multispectral classification algorithm for classifying parcels in an agricultural region. **International Journal of Remote Sensing, 17**(17), 3357-3371.

23. Srivastava, M. S. (2002). Methods of Multivariate Statistics. John Wiley & Sons, Inc. New York.

24. Biernacki, C. And Govaert, G. (1999) Choosing models in Model-Based Clustering and Discriminant Analysis. **J. Statist. Comput. Simul. 64,** 49-71.

25. Celeux, G. and Govaert, G. (1993). Comparison of the Mixture and the Classification Maximum likelihood in Cluster Analysis. **Journal of Statistical Computation and Simulation, 47**, 127-146.

26. Fraley, C., and Raftery, A. E. (2000). Model-Based Clustering, Discriminant Analysis, Anad Density Estimation. Technical Report No. 380 Department of Statistics University of Washington Box 354322 Seattle, WA 98195-4322, USA.

27. Arabie, P. And Hubert, L. J. (1999). An Overview of Combinatorial Data Analysis (Chapter 1), in Clustering and Clasification, Arabie, P. And Hubert, L. J. And DeSoete, G. Eds, World Scientific.

28. Witherspoon, N.H., Holloway, J.H., Davis, K.S., Miller, R.W. and Dubey, A.C. (1995). Coastal battlefield reconnaissance and analysis program for minefield detection. Proc. SPIE Vol. 2496, p. 500-508, Detection Technologies for Mines and Minelike Targets, Abinash C. Dubey; Ivan Cindrich; James M. Ralston; Kelly A. Rigano; Eds.

29. Ju, J., Kolaczyk, E. D. And Gopal, S. (2003) Gaussian Mixture Discriminant Analysis and Sub-pixel Land Cover Characterization in Remote Sensing. **Remote Sensing and Environment 84,** 550-560.

30. Fraley, C. And Raftery, A. E. (2003) Enhanced software for model-based clustering, discriminant analysis, and density estimation: MCLUST. **Journal of Classification 20**, 263-286.

31. Kolaczyk, E. D., Ju, J., & Gopal, S. (2003). Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing. **Remote Sensing of Environment, 84**(4), 550-560.

32. Soffritti, G. (2003). Identifying multiple cluster structures in a data matrix. Communications in Statistics, **Simulation & Computation, 32** (4), 1151-1181.

33. Li, J. (2005) Clustering Based on a Multi-layer Mixture Model. **Journal of Computational and Graphical Statistics, 14**(3), 547-568.

34. Halbe, Z., & Aladjem, M. (2005). Model-based mixture discriminant analysis—an experimental study. **Pattern Recognition, 38**(3), 437-440.

35. Bashir, S. And Carter, E. M.(2005) High breakdown mixture discriminant analysis. **Journal of Multivariate Anal. 93**(1), 102-111.

36. Bouman, C. A., Shapiro, M., Cook, G. W., Atkins, C. B., & Cheng, H. (1997). Cluster: An unsupervised algorithm for modeling Gaussian mixtures.

37. Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. **Journal of the American Statistical Association, 101**(473), 168-178.

38. Biernacki, C., Celeux, G., Govaert, G. And Langrognet, F. (2006) Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. **Computational Statistics and Data Analysis, 51**, 587-600.

39. Fraley, C. and Raftery, A.E. (2007). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. **Journal of Classification, 24**, 155-181.

40. Galimberti, G., and Soffritti, G. (2007). Model-based methods to identify multiple cluster structures in a data set. **Computational Statistics and Data Analysis, 52**(1), 520-536.

41. Servi, T. and Erol, H. (2007). On Total Number Of Candidate Component Cluster Centers And Total Number of Candidate Mixture Models In Model Based Clustering. **Selçuk Journal of Applied Mathematics, 8**(2), 57–69.

42. Eskidere, O., & Ertas, F. (2007, June). Impact of Pitch Frequency on Speaker Identification. In Signal Processing and Communications Applications, 2007. SIU 2007. IEEE 15th (pp. 1-4). IEEE.

43. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. **Pattern recognition letters, 31**(8), 651-666.

44. Li, F., Qiao, W., Sun, H., Wan, H., Wang, J., Xia, Y., ... & Zhang, P. (2010). Smart transmission grid: Vision and framework. **IEEE transactions on Smart Grid, 1**(2), 168-177.

45. Steane, M. A., McNicholas, P. D., & Yada, R. Y. (2012). Model-based classification via mixtures of multivariate t-factor analyzers. **Communications in Statistics-Simulation and Computation, 41**(4), 510-523.

46. Seo, B. and Kim, D. (2012). Root selection in normal mixture models. **Computational Statistics and Data Analysis. 56**, 2454-2470.

47. Çalış, N. and Erol, H., 2013. A new per-field classification method using mixture discriminant analysis. **Journal of Applied Statistics 39**(10):1-12. DOI: 10.1080/02664763.2012.702263.

48. Servi, T., and Erol, H. (2013). A data mining method for refining groups in data using dynamic model based clustering. In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on(pp. 1-6). IEEE.

49. Erol, H. (2013). A model selection algorithm for mixture model clustering of heterogeneous multivariate data. In Innovations in Intelligent Systems and Applications (INISTA), *2013 IEEE International Symposium*. pp. 1-7.

50. Huang, T., Peng, H., and Zhang, K. (2013). Model Selection for Gaussian Mixture Models. arXiv preprint arXiv:1301.3558.

51. Bouveyron, C., and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. **Computational Statistics and Data Analysis, 71**, 52-78.

52. Browne, R. P., & McNicholas, P. D. (2014). Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. **Statistics and Computing, 24**(2), 203-210.

53. Cheballah, H., Giraudo, S., and Maurice, R. (2015). Hopf algebra structure on packed square matrices. **Journal of Combinatorial Theory, Series A, 133**, 139-182.

54. Çalış, N. (2005), Karma Dağılım Modellerinde Bileşen Sayısını Tahmin Etmek İçin Yöntemler, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Yayımlanmamış Yüksek Lisans Tezi, Adana

55. Servi, T. (2009), Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Yayımlanmamış Doktora Tezi, Adana

56. Akdeniz, F. (2004). Olasılık istatistik. Nobel Kitapevi. Adana.

57. Stevens, S.S. (1946).On the theory of scales of measurement. **Science 103,** pp. 677–680.

58. Mao, J. ve Jain, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (hec). **IEEE Transactions on Neural Networks, 7,** 16-29.

59. Cliff, D., & Miller, G. F. (1995, June). Tracking the red queen: Measurements of adaptive progress in co-evolutionary simulations. In European Conference on Artificial Life (pp. 200-218). Springer Berlin Heidelberg.

60. Meyer, P. L. (1970). Introductury Probability and Statistical Applications. Addison-Wesley Publishing Company, Inc, 183 p.

61. Everıtt, B. S. (1980). Cluster analysis. Halsted Press, New York, p. 25-27

62. Gordon, A.D. (1999). Classification, (2nd edn). Chapman & Hall, New York, NY.

63. Hair Jr, J. F., Anderson, R. E., Tatham, R. L., & William, C. (1995). Black (1995), Multivariate data analysis with readings. New Jersy: Prentice Hall.

64. Duran, B.S. and Odell, P.L. (1974). Cluster Analysis, New York: Springer-Verlag.

65. Hansen, P., Hertz, A. and Quinodoz, N. (1997). Splitting trees. **Discrete Mathematics 165-166**: 403-419.

66. Jaın, A.K. and Dubes, R.C. (1988). Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

67. Gower, J.C. (1971). A general coefficient of similarity and some of its properties. **Biometrics, 27,** 857-872

68. Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. **Journal of the American statistical association, 58**(301), 236-244.

69. Macqueen. J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, eds.

70. Martinez, W.L. and Martinez, A.R. (2005). Explanatory Data Analysis with MatLab, Computer Science and Data Analysis Series, Chapman & Hall/CRC. NewYork.

71. Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. **Pattern recognition, 28**(5), 781-793.

72. Fraley, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. **SIAM Journal on Scientific Computing, 20**(1), 270-281.

73. Schwarz, G. (1978). Estimating the dimension of a model, **Ann. Statist. 6,**461–464.

74. Akaike, H. (1974). A new look at the statistical model identification**. IEEE Transactions on Automatic Control 19** (6), 716–723.

75. Gögebakan, M. and Erol, H., (2016a). Mixture model clustering using variable data segmentation. Conference: 12th German Probability and Statistics Days, At Bochum, Germany.

76. Gögebakan, M. and Erol, H., (2016b). A New Approach For Mixture Model Clustering Based On Selecting The Best Mixture Model Among Candidate Mixture Models. International Conference on Information Complexity and Statistical Modeling in High Dimensions with Applications, At Nevşehir Turkey, Volume: 1.

77. Ruspini E. H. (1970) Numerical methods for fuzzy clustering. **Inform. Sci. 2**, 319–350.

78. Gögebakan, M. ve Erol, H. (2016c) Değişken Veri Parçalanmaları Kullanılarak Model Tabanlı Kümelemeye Dayalı Yeni Bir Sınıflandırma Yöntemi: Uzaktan Algılanmış Çok Bantlı Görüntü Sınıflandırılması için Veri Madenciliği Yaklaşımı. 6. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu (UZAL-CBS 2016) Adana, Türkiye.

79. Erol, H. and Akdeniz, F. (2005). A per-field classification method based onmixture distribution models and an application to Landsat Thematic Mapperdata. **International Journal of Remote Sensing,** 26 (6), 1229–1244.

80. Erol, H. and Erol, R., 2016. Logical Circuit Design Using Orientations Of Clusters In Multivariate Data For Decision Making Predictions: A Data Mining And Artificial Intelligence Algorithm Approach. *International Symposium on INnovations in Intelligent SysTems and Applications* 2-5 August 2016, At Sinaia, Romania, Volume: 1.

81. Çalış, N., & Erol, H. (2013). A new per-field classification method using mixture discriminant analysis. **Journal of Applied Statistics, 39**(10), 2129-2140.

82. David W. Messinger ; Amanda Ziemann ; Bill Basener and Ariel Schlamm (2012). "Metrics of spectral image complexity with application to large area search", **Opt. Eng. 51**(3), 036201 (Mar 29, 2012).

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı, Soyadı **:** Maruf GÖGEBAKAN
Uyruğu **:** T.C.
Doğum Tarihi ve Yeri: 03.05.1981-Gaziantep
Medeni Durumu: Evli
Tel **:** 0 505 859 24 65
Email **:** maruf.gogebakan@agu.edu.tr, marufgogebakan@gmail.com
Yazışma Adresi : Abdullah Gül Üniversitesi Bilgisayar Bilimleri Fakültesi Uygulamalı Matematik Bölümü –Kocasinan/Kayseri

EĞİTİM

Derece	Kurum	Mezuniyet Tarihi
Lise	Şahinbey Cumhuriyet Lisesi/Gaziantep	1999
Lisans	Erciyes Üniversitesi Fen Fakültesi, Matematik	2005
Yüksek Lisans	Niğde Üniversitesi Fen Bilimleri Enstitüsü, Matematik	2011
Doktora	Erciyes Üniversitesi Fen Bilimleri Enstitüsü, Matematik	2017

İŞ DENEYİMLERİ

Yıl	Kurum	Görev
2005-2009	Kayseri Uğur Dershanesi	Matematik Öğretmeni
2009-20011	Kayseri BİL Dershanesi	Matematik Öğretmeni
2011-	Abdullah GÜL Üniversitesi Bilgisayar Bilimleri Fakültesi Uygulamalı Matematik Bölümü	Araştırma Görevlisi

YABANCI DİL

İngilizce

YAYINLAR

ULUSLARARASI KONFERANS BİLDİRİLERİ

- Gogebakan, M. and Erol, H. ,(2016) Mixture Model Clustering Using Variable Data Segmentation. Conference: 12th German Probability and Statistics Days, Bochum, Germany.
- Gogebakan, M. and EROL, H. (May 2016) A New Approach For Mixture Model Clustering Based On Selecting The Best Mixture Model Among Candidate Mixture Models Conference: International Conference on Information Complexity and Statistical Modeling in High Dimensions with Applications, Nevşehir, Turkey.

- Gogebakan, M. and Erol, H. (2016) Değişken Veri Parçalanmaları Kullanılarak Model Tabanlı Kümelemeye Dayalı Yeni Bir Sınıflandırma Yöntemi: Uzaktan Algılanmış Çok Bantlı Görüntü Sınıflandırılması için Veri Madenciliği Yaklaşımı. 6. Uzaktan Algılama ve Coğrafi Bilgi Sistemleri Sempozyumu (UZAL-CBS) Adana, Türkiye.