ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

Tayfun SERVİ

ÇOK DEĞİŞKENLİ KARMA DAĞILIM MODELİNE DAYALI KÜMELEME ANALİZİ

İSTATİSTİK ANABİLİM DALI

ADANA, 2009

ÇUKUROVA ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

ÇOK DEĞİŞKENLİ KARMA DAĞILIM MODELİNE DAYALI KÜMELEME ANALİZİ

Tayfun SERVİ

DOKTORA TEZİ

İSTATİSTİK ANABİLİM DALI

Bu tez 20 / 04 /2009 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından Oybirliği/ ~~Oyçokluğu~~ İle Kabul Edilmiştir.

İmza.....
Prof.Dr. Hamza EROL
DANIŞMAN

İmza.....
Prof.Dr. Fikri AKDENİZ
ÜYE

İmza.....
Yrd.Doç.Dr. Sami ARICA
ÜYE

İmza.....
Prof.Dr. Selahattin KAÇIRANLAR
ÜYE

İmza.....
Doç.Dr. Aşır GENÇ
ÜYE

Bu tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.
Kod No:

Prof. Dr. Aziz ERTUNÇ
Enstitü Müdürü
İmza ve Mühür

ÖZ

DOKTORA TEZİ

ÇOK DEĞİŞKENLİ KARMA DAĞILIM MODELİNE DAYALI KÜMELEME ANALİZİ

Tayfun SERVİ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

Danışman: Prof.Dr. Hamza EROL
Yıl: 2009, Sayfa: 242
Jüri: Prof.Dr. Hamza EROL
Prof.Dr. Fikri AKDENİZ
Yrd.Doç.Dr. Sami ARICA
Prof.Dr. Selahattin KAÇIRANLAR
Doç.Dr. Aşır GENÇ

Sonlu karma dağılım modeli, modele dayalı kümeleme analizinde kullanılır. Sonlu karma dağılım modeline dayalı kümeleme analizine modele dayalı kümeleme analizi denir. Kümeleme analizinde amaç her bir grup içindeki elemanlar arasındaki uzaklığın minimum ve gruplar arasındaki farkın maksimum yapılmasıdır. Modele dayalı kümeleme analizinde veri, çok değişkenli karma dağılım modeliyle temsil edilir. Bu temsilde modeldeki her bir bileşen verideki bir kümelenmeye karşılık gelir. Modeldeki her bir bileşen için çok değişkenli olasılık yoğunluk fonksiyonu, çok değişkenli veri setindeki ilgili küme yapısını açıklar. Bu tez çalışmasında: i) Çok değişkenli verinin karma dağılım modeline sahip olduğu ve karma dağılım modelindeki her bir bileşenin çok değişkenli normal dağıldığı varsayımıyla çok değişkenli karma dağılım modeline dayalı kümeleme analizinde bileşen küme sayısının belirlenmesi için yeni bir yöntem önerilmiştir. ii) Çok değişkenli verinin karma dağılım modeline dayalı kümeleme analizi için çok değişkenli verideki grupların inceltilerek kümelenmesi amacıyla yeni bir algoritma geliştirilmiştir. iii) Çok değişkenli verideki grupların inceltilerek kümelenmesi için geliştirilen algoritma kullanılarak uzaktan algılanmış çok bantlı uydu görüntü verisi eğitimli olarak alan bazında sınıflandırılmıştır.

Anahtar kelimeler: Kümeleme analizi, Modele dayalı kümeleme, Bileşen küme sayısı, Bhattacharyya uzaklığı, Çok değişkenli karma normal dağılım modeli.

ABSTRACT

PhD. THESIS

MULTIVARIATE MIXTURE DISTRIBUTION MODEL BASED CLUSTER ANALYSIS

Tayfun SERVİ

DEPARTMENT OF STATISTICS
INSTITUTE OF NATURAL AND APPLIED SCIENCES
UNIVERSITY OF ÇUKUROVA

Supervisor: Prof.Dr. Hamza EROL
Year: 2009, Pages: 242
Jury: Prof.Dr. Hamza EROL
Prof.Dr. Fikri AKDENİZ
Assist.Prof.Dr. Sami ARICA
Prof.Dr. Selahattin KAÇIRANLAR
Assoc.Prof.Dr. Aşır GENÇ

Mixture distribution model is used for model based cluster analysis. Cluster analysis based on mixture distribution models, is called as model based cluster analysis. The aim of the model based cluster analysis is minimizing the distances between each elements in a group and maximizing the differences between groups in multivariate data set. In model based cluster analysis approach, it is assumed that the multivariate data is generated by a mixture distribution model in which each component corresponds to a different cluster in multivariate data set. The multivariate probability density function for each component in mixture distribution model explains the structure of the corresponding cluster in multivariate data set. In this thesis: i) A new method is proposed for determining the number of component clusters in the multivariate normal mixture model based cluster analysis by assuming that the multivariate data having mixture distribution model and each component in mixture distribution model is multivariate normal distribution. ii) A new algorithm is developed for multivariate normal mixture distribution model based cluster analysis by refining groups in multivariate data set. iii) The new algorithm developed for multivariate normal mixture distribution model based cluster analysis by refining groups in multivariate data set is applied for supervised per field classification of remotely sensed multispectral image data of an an agricultural region in Adana.

Key words: Cluster analysis, Model based clustering, Mixture Component cluster number, Bhattacharyya distance, Multivariate normal mixture distribution model.

TEŞEKKÜR

Bu tez çalışması sürecinde değerli bilgi ve görüşlerini benimle paylaşan danışmanım sayın Prof.Dr. Hamza EROL'a; tez izleme komitesi üyeleri sayın Prof.Dr. Fikri AKDENİZ ve sayın Yrd.Doç.Dr. Sami ARICA'ya; yardımlarını esirgemeyen İstatistik Bölümü öğretim elemanlarına teşekkürlerimi sunarım.

Ayrıca eğitim ve öğrenim hayatım boyunca maddi ve manevi katkılarını benden hiçbir zaman esirgemeyen aileme; en sıkıntılı günlerimde bana destek olan eşim Filiz DOĞANÇAY SERVİ ve hayatımın neşe kaynağı oğlum Yusuf SERVİ'ye teşekkürü bir borç bilirim.

İÇİNDEKİLER **SAYFA**

TABLOLAR DİZİNİ **SAYFA**

ŞEKİLLER DİZİNİ **SAYFA**

1. GİRİŞ

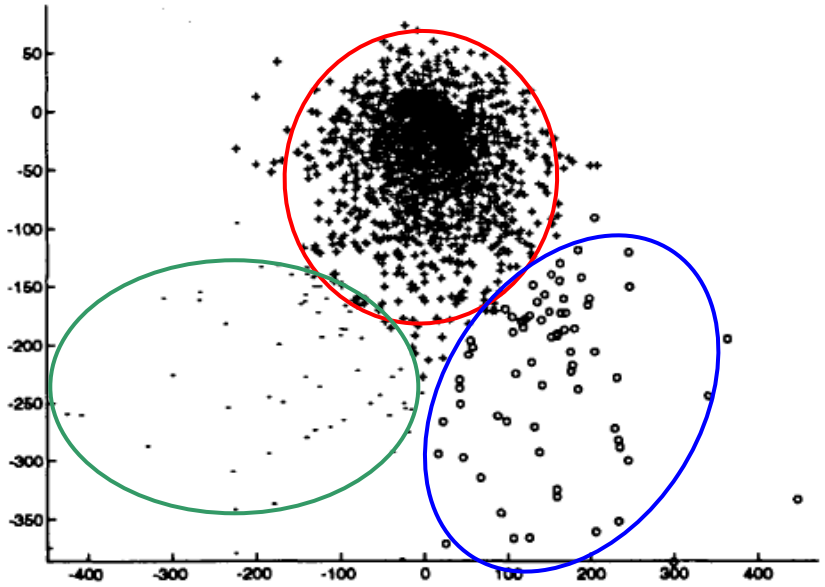
Kümeleme analizinde, heterojen yapıdaki verinin küme sayısı ve küme yapısı araştırılır. Kümelemede amaç, her bir küme içerisindeki gözlemlerin ya da nesnelerin birbirlerine benzer ve kümelerin de birbirinden farklı olacak şekilde en uygun gruplama yapısını bulmaktır. Kümeleme analizi temel olarak ayrıştırma analizinden farklıdır. Ayrıştırma analizinde gözlemler daha önceden tanımlanmış ve sayısı bilinen gruplara parçalanırken, kümeleme analizinde ne grup sayısı hakkında ne de grupların yapısı hakkında bir ön bilgi bulunmaz. Gözlemlerin kümelere gruplanması için geliştirilen bazı yöntemlerde kümeleme, tüm gözlem çiftleri arasındaki benzerliklerin bulunmasıyla başlar. Bazı durumlarda benzerlikler, uzaklık ölçümlerine dayalı olarak bulunur. Diğer kümeleme yöntemlerinde, küme merkezlerinin seçimi veya küme içi ve kümeler arası değişimin karşılaştırılması yapılır. Değişkenlerin de kümelenmesi mümkündür. Bu durumda benzerlik için korelasyon kullanılır (Soffritti 2003, Galimberti ve Soffritti 2007).

Kümeleme analizi, sınıflandırma olarak da adlandırılabilir. Birçok alandaki veri kümeleri üzerinde kümeleme analizi uygulanmaktadır. Bu alanlar arasında tıp, psikiyatri, sosyoloji, kriminoloji, antropoloji, arkeoloji, jeoloji, coğrafya, uzaktan algılama, pazar araştırması, ekonomi ve mühendislik sayılabilir. Kümeleme analizinin uygulama alanlarına örnekler olarak aşağıdaki alanlar verilebilir:

Astronomi: Faundez-Abans ve ark.(1996), Ward yöntemi (Everitt ve ark., 2001) olarak bilinen kümeleme yöntemini, uzayda bulut biçiminde görünen 192 gezgin gaz ve toz yığınının kimyasal bileşimleri ile ilgili veriye uygulamışlardır. Çalışmalarında, 6 farklı grup olduğunu tespit etmişlerdir. Astronomi alanında bir başka çalışma da Celeux ve Govaert (1995) tarafından ele alınmıştır. Bu çalışmada karma normal modeli, 2370 yıldıza ait iki değişkenli verinin kümelenmesi için uygulamışlardır. Sonuçta üç farklı küme yapısı ile veriyi modellemişlerdir. Bu veriye ait kümelenme yapısı Şekil 1.1'de verilmiştir.

Arkeoloji: Houdson (1971), k-ortalamalar (k-means) algoritmasını kullanarak Britanya adalarında bulunan ve taş devrinden kalan baltaların tasnifini yapmıştır. Bu çalışmada örnekler için kullanılan değişkenler: uzunluk, genişlik ve

keskinlik olarak alınmıştır. Çalışmanın sonucunda ince küçük ve kalın büyük olmak üzere iki balta sınıfı oluşturmuşlardır. Benzer bir çalışma da Gordon (1999) tarafından yapılmıştır. Dellaportas (1998), cilalı taş devrinden kalan aletlerin sınıflandırılması için bir Bayesci (Bayesian) yaklaşım kullanmıştır. Mallory-Greenough ve Greenough (1998), eski Mısır çömleklerini tek bağlantı (single linkage) ve ortalama bağlantı (average linkage) kümeleme yöntemlerini kullanarak üç farklı gruba sınıflandırmışlardır.



Şekil 1.1. 2370 yıldıza ait iki değişkenli verinin saçılım grafiği (Celeux ve Govaert 1995).

Psikiyatri: Zihinsel rahatsızlıkların teşhisi ve tedavisi, bedensel rahatsızlıkların teşhisinden ve tedavisinden oldukça zordur. Psikiyatride zihinsel rahatsızlıkların teşhisinde veya uygulanan tedavi yöntemlerinin gruplandırılmasında kümeleme analizi kullanılmaktadır. Pilowsky ve ark. (1969) ve, Wallace ve Boulton (1968) çalışmalarında kümeleme yöntemini kullanarak cinsiyet, yaş ve hastalık süresi bilgilerini de kapsayan depresyon anket verilerini incelemişlerdir. Ankette yer alan 200 hastayı kümelemişlerdir. Benzer bir çalışmayı Paykel (1971), Friedman ve Rubin (1967) tarafından yapılan çalışmadaki yöntemini kullanarak, 165 hasta üzerinde yapmıştır. Bu çalışmanın sonucunda veride dört grubun varlığı

belirlenmiştir. Depresyon sınıflandırmasına ilişkin genel bir çalışma Farmer ve ark. (1983) tarafından verilmiştir. Kurtz ve ark. (1987) ve, Ellis ve ark. (1996) intihara teşebbüs eden hastaları, ortalama bağlantı kümeleme (average linkage clustering) yöntemini kullanarak kümelemişlerdir. Bu çalışmalarının sonucunda da dört farklı grup elde etmişlerdir.

Pazar (piyasa) Araştırması: Dünya'da pazar araştırması için kullanılabilecek oldukça fazla şehir bulunmaktadır. Buna karşın ekonomik nedenlerden dolayı araştırmalar az sayıda şehir üzerinde yapılabilmektedir. Aralarında benzer özellikler gösteren şehirler aynı gruba sınıflandırıldıktan sonra her gruptan seçilen bir şehrin araştırmada kullanılması mümkündür. Bu yaklaşımı Green ve ark. (1967), 88 şehri 14 değişkene göre sınıflandırmışlardır.

Ekin (Ürün) Sınıflandırması: Conway ve ark. (1991), tarımsal bir alanda yetiştirilen ve içeriği bilinen beş ürüne ilişkin görüntüleri parçalamak için k-ortalamalar yöntemini kullanmışlardır. Sonuçta elde edilen parçalanmış görüntüde ürünlerin, geniş yapraklı ve dar yapraklı ekinler olarak iki belirgin kümeye ayrılabileceğini belirtmişlerdir.

Kümeler, gözlemlere ait çizimlerinden faydalanılarak, grafiksel olarak da belirlenebilir (Everitt ve ark. 2001). İki değişken varsa gözlemlere ait saçılım grafiği kullanılabilir. Verinin iki değişkenden fazla olması durumunda temel bileşenler analizi kullanılarak veri iki değişkene indirgenebilir ve grafiği çizilebilir. Çizime dayalı başka bir yaklaşımda izdüşüm grafikleridir. Burada kümelere ilişkin iki boyutlu izdüşümler incelenir (Everitt ve ark. 2001).

Çok değişkenli verinin grafiksel gösterimi bu verilerin kullanıldığı yöntemlerin tüm aşamalarında önemlidir. Genel anlamda çok değişkenli verinin grafiksel gösterimi verinin yapısı hakkında bilgi vermektedir. Bu sayede verinin kümelenebilir olup olmadığı hakkında bir ön bilgi elde edilebilir. Kümelerin belirlenmesinde grafiksel yöntemler de kullanılabilir (Everitt ve ark. 2001). Grafiksel yöntemler, tek veya iki boyutlu veriler için uygundur. Çok değişkenli veri, genel olarak ikiden fazla değişken içerdiğinden grafiksel yöntemler ham veriye doğrudan uygulanamaz. Bu nedenle çok değişkenli verinin daha düşük boyutlara indirgenmesi gerekir ya da bu değişkenlerin izdüşümleri kullanılır. Burada dikkat edilmesi

gereken, verinin daha düşük boyutlara indirgenmesi durumunda küme yapısının nasıl ve ne kadar etkileneceğidir.

Bir veya iki boyutlu verilerde kümelerin grafiksel olarak belirlenmesi için genel düşünce: tek tepeli dağılımın, kümelenme göstermeyen homojen bir kitleyi temsil ettiğidir. Buna karşın birden fazla ve farklı tepelerin varlığı kümelenebilir, heterojen kitlenin varlığını göstermektedir (Everitt ve ark. 2001). Burada her bir tepenin bir kümeyi temsil ettiği düşünülür. Verinin grafiksel gösterimindeki çok tepeli yapı, veride küme yapısının bulunduğu yönünde kuvvetli kanıt oluşturmaktadır. Bir veya iki boyutlu verideki tepe (mod) yapısı, grafiksel olarak belirlenebilir. Tek boyutlu verilerde tepelerin ya da kümelerin belirlenmesi için histogramdan faydalanılır. Çok boyutlu verilerde ise yüzey grafikleri, izdüşüm grafikleri ve kontur grafikleri oluşturulabilir.

Sonlu karma dağılım modelleri, modele dayalı kümeleme analizinde yoğun olarak çalışılmaktadır (Wolfe 1970, Edwards ve Cavalli-Sforza 1965, Day 1969, Scott ve Symons 1971, Duda ve Hart 1973, Binder 1978). Sonlu karma dağılım modelleri, kümeleme analizinde kullanılan yöntemlerin uygulamalarında ortaya çıkan sorunların çözümüne olanak sağlar (McLachlan ve Basford 1988, Banfield ve Raftery 1993, Cheeseman ve Stutz 1995, Everitt ve Dunn 2001).

Kümeleme analizi verinin anlamlı gruplara ayrılmasıyla ilgilidir (Fraley ve Raftery 1998). Kümeleme analizinde çalışılan en önemli problem, verideki kümelenmelerin sayısının ve yapısının belirlenmesidir. Verideki kümelenmelerin sayısını ve yapısını belirlemenin bir yöntemi de modele dayalı kümeleme analizidir. Kümelerin çok değişkenli karma dağılım modelleriyle analizi; karma model kümeleme analizi ya da modele dayalı kümeleme analizi olarak adlandırılır (Bozdogan 1994). Modele dayalı kümeleme analizinde veri, çok değişkenli karma dağılım modelleriyle temsil edilir. Bu temsilde her bir bileşen bir kümelenmeye karşılık gelir (Fraley ve Raftery 1998). Kümeleme analizinde amaç her bir grup içinde homojenliğin maksimum yapılması ve aynı zamanda gruplar arasındaki farkın maksimum yapılmasıdır (Arabie ve ark. 1996).

Çok değişkenli karma dağılım modellerini kullanarak modele dayalı kümeleme analizi, istatistikteki en zor problemlerden biridir (Bozdogan 1994). Bu

zorluk, çok değişkenli karma dağılım modelinin bileşen yığın sayısının belirlenmesindeki güçlükten kaynaklanır. Modele dayalı kümeleme analizinde veriyi temsil eden çok değişkenli karma dağılım modelinde bileşen yığın sayısının belirlenmesi amacıyla bilgi kriteriyle birlikte model seçimi Bozdogan(1994) tarafından çalışılmıştır.

Bu tez çalışmasında modele dayalı kümeleme analizi için çok değişkenli karma dağılım modelindeki bileşenlerin çok değişkenli normal dağılım olduğu varsayımı altında bileşen yığın sayısının belirlenmesi için yöntemler (Duda ve ark. 2001) incelenmiştir. Çok değişkenli karma dağılım modelindeki bileşen yığın sayısının belirlenmesinden sonra modeldeki parametrelerin tahminleri EM algoritması (Mclachlan ve Peel 2000) kullanılarak elde edilmiştir. Karma dağılım modeline dayalı kümeleme analizinin uygulama alanları araştırılmıştır.

Bu tez çalışması dokuz bölümden oluşmaktadır. İkinci bölümde çok değişkenli kümeleme analizi ve karma kümeleme analizi ile ilgili yapılan önceki çalışmalar incelenmiştir. Üçüncü bölümde çok değişkenli kümeleme analizinde kavramlar, tanımlar, gösterimler ve yöntemler araştırılmıştır. Dördüncü bölümde çok değişkenli karma normal dağılım modeline dayalı kümeleme analizi ele alınmıştır. Beşinci bölümde karma kümeleme analizinde kullanılan model seçimi kriterleri incelenmiştir. Altıncı bölümde karma kümeleme analizinde aday bileşen küme sayısının belirlenmesi için yeni bir yöntem oluşturulmuştur. Yedinci bölümde çok değişkenli veri setindeki grupların inceltilerek karma normal kümeleme analizi için yeni bir yöntem geliştirilmiştir. Sekizinci bölümde çok değişkenli veri setindeki grupların inceltilerek karma normal kümeleme analizi, uzaktan algılanmış çok bantlı uydu görüntü verisinin kümelenmesi ve alan bazında eğitimli sınıflandırılması uygulaması verilmiştir. Bu bölümde yeni geliştirilen bir yöntemin uygulaması gerçekleştirilmiştir. Dokuzuncu ve son bölümde sonuçlar ve öneriler belirtilmiştir.

2. ÖNCEKİ ÇALIŞMALAR

Karma dağılım modeliyle ilgili ilk çalışma olarak Pearson (1894) tarafından yapılan çalışma kabul edilmektedir. Pearson (1894) çalışmasında aynı coğrafik alanda yaşayan farklı yengeç türlerini modellemek için iki tek değişkenli normal dağılımın karmasını kullanmıştır. İki tek değişkenli normal dağılımın karmasındaki parametreleri, momentler yöntemini kullanarak tahmin etmiştir. Pearson (1894) tarafından yapılan bu çalışma ayrıca karma dağılımların genetik çalışmalarda kullanıldığı ilk çalışma olarak da kabul edilir (Schork ve ark., 1996).

Wolfe (1965, 1967), Day (1969) ve Binder (1978), çok değişkenli normal dağılımların karmasını kümeleme analizinde istatistiksel model olarak kullanmışlardır.

Symons (1981), kümeleme kriterleri ve çok değişkenli normal karmalar (Clustering Criteria and Multivariate Normal Mixtures) başlıklı çalışmasında çok değişkenli normal dağılımların bir karmasının kullanılmasıyla oluşabilecek dört yeni kümeleme kriteri üzerinde durmuştur. Bu çalışmada önerilen dört kümeleme kriteri, karma modelin bileşen kovaryans matrisleri üzerinde farklı kabullere bağlı olarak, en çok olabilirlik ve Bayes yaklaşımlarından elde edilmiştir. Bu dört kriterden ikisi, Friedman ve Rubin (1967) tarafından geliştirilen grup içi kareler toplamı matrisinin determinantına dayalı yöntemin yeniden düzenlenmiş halidir. Diğer kriterlerde farklı şekildeki kümeleme yapıları için önerilmiştir. Bu kriterlerin performansı daha önce Reaven ve Miller (1979) tarafından incelenen 145 yetişkin bireye ait 3 değişkenli diyabet verisi üzerinde gösterilmiştir. Çalışmanın sonucunda kümeleme analizi için en uygun kriterin ya da yaklaşımın veri yapısına bağlı olduğu bu nedenle de bir araştırmacıya kümeleme analizinde hangi kriterin kullanılması gerektiğine dair bir önermede bulunmanın zor olduğu belirtilmiştir.

Render ve Walker (1984); karma yoğunluklar, en çok olabilirlik ve EM algoritması (mixture densities, maximum likelihood and the EM algorithm) başlıklı çalışmada üstel dağılım ailesindeki dağılımların meydana getirdiği karma modellerin parametrelerini EM algoritması (Demster ve ark. 1977) kullanılarak en çok olabilirlik tahminlerinin nasıl elde edileceğini açıklamıştır. EM (Expectation – Maximization)

algoritmasının, üstel dağılımlardan meydana gelen karma modellerde teorik ve pratik özellikleri tartışılmıştır. EM algoritması kullanılarak karma modellerin parametre tahminlerinde ortaya çıkan bazı sorunlar özetlenmiştir. Sonuç olarak EM algoritmasının dezavantajları olarak; EM algoritma için alınan parametre başlangıç değerlerinin seçiminin oldukça önemli olduğu, EM algoritmasının ardışık bir algoritma olması nedeniyle yakınsamanın bazı durumlarda oldukça yavaş olabileceği ve bazı durumlarda da EM algoritmasının parametrenin global bir maksimum değeri yerine yerel bir maksimum değerine yakınsayabileceğinin mümkün olduğu şeklinde belirtilmiştir. EM algoritmasının bu tür dezavantajlarının algoritmada yapılacak bazı yeniden düzenlemeler ile parametre başlangıç değerlerinin seçimi içinde ayrı bir algoritma veya yöntem kullanılması durumunda yerel maksimuma yakınsama tehlikesinin ortadan kalkabileceği ve hibrit algoritmalar kullanılırsa algoritmanın yakınsama hızının artabileceği belirtilmiştir.

Basford ve McLachlan (1985), normal karma modellerle olabilirlik tahmini (likelihood estimation with normal mixture models) başlıklı çalışmalarında çok değişkenli normal dağılımların bir karmasındaki parametrelerin olabilirlik tahmini yöntemiyle çalışıldığında ortaya çıkan problemleri ortaya koymuşlardır. Karma modeldeki bileşenlerin kovaryans matrislerinin eşit alınması durumda, olabilirlik eşitliklerinden elde edilen köklerin seçiminin doğrudan yapılabileceğini, parametrelerin en çok olabilirlik tahmin edicilerinin bulunabileceğini ve bu tahmin edicilerin tutarlı olduğunu göstermişlerdir. Karma modeldeki bileşenlerin kovaryans matrislerinin birbirinden farklı alınması durumunda ise en çok olabilirlik eşitliklerinden elde edilen birden fazla köklerden hangisinin parametre tahmininde kullanılacağının belli olmadığını belirtmişlerdir.

Ganesalingam (1989), sınıflandırma ve karma yaklaşımlarda en çok olabilirliğe dayalı kümeleme (classification and mixture approaches to clustering via maximum likelihood) başlıklı çalışmasında en çok olabilirliğe dayalı iki kümeleme yöntemini, iki farklı kitleden elde edilen gözlemler olması durumunda ve gözlemlerin kitlelerden hangisine ait olduklarına ilişkin bir bilgi bulunmadığında ortaya çıkan sınıflandırma problemi çerçevesinde karşılaştırmıştır. Bu yöntemler en çok olabilirliğe dayalı ve MA (Mixture Approach) ile gösterilen karma yaklaşım, ve

en çok olabilirliğe dayalı ve CA (Classification Approach) ile gösterilen sınıflandırma yaklaşımıdır. Bu iki yöntem, karma örnekleme ve ayrık örnekleme gibi iki farklı örnekleme yöntemi ile elde edilen ortalamaları farklı ve bileşen kovaryans matrisleri aynı olan iki çok değişkenli normal dağılımın bir karması için oluşturulan simülasyon verisi üzerinde karşılaştırılarak temel farklar ortaya konmuştur. Simülasyon verisinin elde edildiği karma dağılım için iki varsayım ayrı ayrı incelenmiştir. Bu varsayımlardan birincisi karmanın bileşen oranlarının biliniyor olması diğer varsayım ise karmanın bileşen oranlarının bilinmiyor olmasıdır. Sınıflandırma yaklaşımında, karma yaklaşımdan farklı olarak gözlemlere ait bileşen etiket vektörleri bilinmeyen parametreler olarak alınır. Karma oranları arasındaki farkın 0,5'ten daha büyük olduğu karma örneklemede veya bileşen kümelerinin hacmi farklı olduğu durumlarda ayrık örneklemede karma yaklaşımın sınıflandırma yaklaşımına göre tercih edilebileceğini belirtmiştir. Özellikle karmanın bileşen kümelerindeki gözlem sayısı eşit ya da birbirine çok yakın olduğu her iki örneklemede, sınıflandırma yaklaşımının daha iyi sonuçlar verdiğini gözlemlemiştir.

Bu iki yöntem ayrıca daha önce Habbema ve ark. (1974) tarafından incelenen hemofili gerçek veri kümesi üzerinde de karşılaştırılmıştır. Hemofili verisi hemofili taşıyıcısı olan ve olmayan 75 bayan hasta üzerinden elde edilen antihemofilik düzeyiyle ilişkili iki değişkenli bir veridir. Çalışmanın sonucunda MA ve CA yaklaşımlarıyla elde edilen hemofili verisinin 75 biriminin tahmin edilmiş sonraki olasılıkları ile verinin gerçek etiket vektörü arasındaki fark oldukça fazladır. Sonuç olarak özellikle MA yaklaşımı ile iyi alt kitle parametreleri tahmin edilebilmesi için n gözlem sayısının oldukça fazla olması gerektiğini belirtmiştir.

Celeux ve Govaert (1992), kümeleme için bir sınıflandırma EM algoritması ve iki stokastik versiyonu (a classification EM algorithm for clustering and two stochastic versions) başlıklı çalışmalarında sınıflandırma maksimum olabilirlik yaklaşımı altında optimizasyona dayalı kümeleme yöntemleri için genel bir sınıflandırma EM algoritması tanımlamışlardır. Bu genel sınıflandırma EM algoritmasından klasik optimizasyon kümeleme algoritmalarının parametre başlangıç değerlerine olan bağımlılığını azaltacak şekilde rasgele karmaşıklığı da içeren iki stokastik algoritma elde etmişlerdir. Elde ettikleri bu iki stokastik algoritmanın

performansını sayısal denemeler üzerinden sınıflandırma EM algoritmasının özel bir versiyonu olan standart k-ortalamalar algoritması ve varyans kriterine göre karşılaştırmışlardır. Tanımladıkları genel sınıflandırma EM algoritmasını kısaca CEM (Classification EM) ile gösterdiler. Bu algoritmaya bağlı olarak elde ettikleri stokastik algoritmayı SEM (Stokastik EM) ve Sınıflandırma yumuşatmayı CAEM (Classification Annealing EM) olarak adlandırmışlardır. Bu algoritmaların performanslarını, küçük ve büyük örneklem hacimli simülasyon verisini ve 3641 hastadan elde edilen gerçek serum verisini (Sandor ve Lechevallier, 1977) kullanarak test etmişlerdir. SEM ve CAEM algoritmalarında iterasyon sayısı oldukça fazladır. Bunun nedeni algoritmaların parametre başlangıç değerlerine olan bağımlılığını azaltacak şekilde adımların artmasıdır. Çalışmalarının sonucunda küçük hacimli örneklemler için CAEM, büyük hacimli örneklemler için SEM önerilmiştir.

Celeux ve Govaert (1993), karma ve sınıflandırma en çok olabilirliklerinin kümeleme analizlerinde karşılaştırılması (comparison of the mixture and the classification Maximum Likelihood in cluster analysis) başlıklı çalışmalarında en çok olabilirliğe dayalı karma ve sınıflandırma yaklaşımları ile yapılan kümeleme yöntemlerini farklı varsayımlar altında incelemişlerdir. Bu iki yaklaşım, bir karma modelin karma oranları eşit veya karma oranları bilinmeyen şeklinde iki farklı varsayım altında Monte Carlo sayısal denemeleri üzerinde karşılaştırmışlardır. Kullandıkları simülasyon denemelerinde özellikle küçük hacimli örneklemler için karma oranları üzerindeki kısıtın kümeleme yönteminin seçiminde oldukça etkili bir faktör olduğunu göstermişlerdir. Sayısal denemeler sonucunda küçük hacimli örneklemlerde sınıflandırma yaklaşımının orta ya da büyük hacimli örneklemlerde karma yaklaşımın daha iyi sonuç verdiğini belirtmişlerdir. Çalışmanın sonucunda karma modellerle kümeleme işlemine başlamadan önce, veri kümesi hakkında herhangi bir önbilginin bulunmamasından dolayı, karma oranlarının başlangıçta nasıl alınacağı ve kümeleme için hangi yaklaşımın seçileceğinin belirsiz olduğunu belirtmişlerdir. Karma oranlarının eşit olması varsayımlı modelin daha cimri bir model olduğunu ve başlangıç değerlerinden daha az etkilendiğini belirtmişler ve bu nedenlerle de kümelemenin daha dayanıklı (robust) sonuçlar verdiğini belirtmişlerdir.

Banfield ve Raftery (1993), normal dağılım modeline ve normal dağılım olmayan modellere dayalı kümeleme (model based Gaussian and non-Gaussian clustering) başlıklı çalışmalarında, çok değişkenli normal dağılımların karmasındaki bileşen kovaryans matrislerinin özdeğer ayrışımı kullanılarak parametrik ifade edilmesiyle, kümeler üzerinde bazı özelliklerin de dikkate alınmasını sağlamışlardır. Bu özellikler kümelerin geometrik olarak ifade edilmesine de olanak sağlayan kümelerin yön, hacim ve şekil bilgileridir. Bu sayede Friedman ve Rubin (1967) tarafından önerilen kriterden daha genel ve, Scott ve Symons (1971) tarafından önerilen modellere göre daha cimri modelleri içeren yeni kriterleri önermişlerdir. Veri kümesi yapısı önceden tanımlanmış kümelerden meydana gelebilir. Bununla birlikte bu tanımlanan veri kümelerinin yapısına uygun olmayan gözlemler bulunabilir. Bu gözlemler gürültü gözlemleri olarak adlandırılır. Banfield ve Raftery (1993) tarafından yapılan çalışmada bu tür gözlemlerin bir Poisson sürecinden ortaya çıktığı kabul edilmiş ve etiketleri sıfır almıştır. Gürültü gözlemleri dikkate alınarak küme sayısının belirlenmesi için bir Bayesci tahmin yaklaşımı (an approximate Bayessian Approach) önermişlerdir. Bu yaklaşımla elde ettikleri bilgi kriteri ağırlıklı ortalama kanıt ya da AWE (Average Weight of Evidence) olarak adlandırılmıştır. Önerdikleri kümeleme kriterlerinin performansını simülasyon verisi ve gerçek veri kümeleri üzerinde Ward'ın kareler toplamı kriteri (Ward 1963) ve tek-bağlantı (single-linkage) yöntemi ile karşılaştırmışlardır. Bu çalışmada simülasyon verisinin yanında iki gerçek veri kümesi de kullanmışlardır. İncelenen gerçek veri kümelerinden birincisi Reaven ve Miller (1979) tarafından daha önce incelenen 145 bireyden ölçülmüş üç değişken değerini içeren diyabet verisidir. Diyabet verisinde klinik çalışmalar sonucu kimyasal diyabetikler, belirgin diyabetikler ve diyabetik olmayanlar şeklinde 3 sınıfın olduğunu belirlemişlerdir. Klinik çalışmalarla elde edilen gözlemlerin sınıflandırma etiket vektörü doğru kabul edilerek çalışmalarında önerilen kümeleme kriterleri karşılaştırılmıştır. İkinci gerçek veri kümesi ise Manyetik rezonans görüntü (MRI) verisidir. Manyetik rezonans görüntüleme doku içerisindeki hidrojen çekirdeklerinin manyetik rezonansa dayalı olarak vücut dokusunun kimyasal karakteristiklerini ölçen bir yöntemdir (Oldendorf ve Oldendorf 1988). MRI voksel olarak adlandırılan küçük hacimli elementler içerisindeki farklı

doku tiplerini çok ayrıntılı olmayacak şekilde ayırabilir. Harekete geçirilen hidrojen çekirdeği tipine (situmule) bağlı olarak her bir voksel içerisindeki farklı karakteristikler ölçülebilir. Bunun sonucunda her bir voksel üzerinden çok değişkenli ölçümler elde edilir. MRI verisinde 26100 voksel bulunmaktadır. Banfield ve Raftery (1993) bu veride fazla sayıda voksel olmasından dolayı voksellerin bir rastgele örneklemini almışlar ve bu rastgele örneklem içerindeki vokselleri kümelemişlerdir. Daha sonra görüntüdeki diğer vokselleri elde ettikleri bu kümelere sınıflandırmışlardır. Aldıkları rastgele örnekleme dayalı olarak görüntü kümesindeki küme sayısının 7 ile 21 arasında olabileceğini belirtmişlerdir.

Bozdogan (1994), model seçim kriteri ve karmaşıklığın yeni bir bilgisel ölçütü kriteri kullanılarak karma model kümeleme analizi (mixture model cluster analysis using model selection criteria and a new informational measure of complexity) başlıklı çalışmasında karma dağılım modellerini kullanarak uygulanan kümeleme analizini karma model kümeleme analizi olarak adlandırmış ve istatistikte çalışılan en zor problemlerden biri olduğunu belirtmiştir. Bu çalışması temel olarak bileşen küme sayılarının seçimi ile ilgilidir. Bileşen kovaryans matris yapılarını dikkate alarak dört genel durum için bileşen küme sayısının bulunmasını amaçlamıştır. Bu genel durumlardan birincisi, karma dağılım modelindeki bileşen kümelerinin birbirinden farklı kovaryans matris yapısına sahip olması durumudur. İkincisi, karma dağılım modelindeki bileşen kümelerinin aynı kovaryans matris yapısına sahip olması durumudur. Üçüncüsü, karma dağılım modelindeki bileşen kümelerinin aynı fakat köşegen kovaryans matris yapısına sahip olması durumudur. Dördüncüsü, karma dağılım modelindeki bileşen kümelerinin aynı fakat değişkenlerin ikişerli bağımsız olacak şekilde köşegen kovaryans yapısına sahip olması durumudur.

Bozdogan (1994) tarafından yapılan çalışmada, Akaike (1973) tarafından oluşturulan ve AIC (Akaike's Information Criterion) olarak adlandırılan bilgi kriterini; Bozdogan (1987) tarafından tanımlanan ve CAIC (Consistent Akaike's Information Criterion -) olarak adlandırılan tutarlı Akaike bilgi kriterini ve, Bozdogan (1988, 1990a, 1990b, 1993) tarafından geliştirilen ve ICOMP (Informational Measure of Complexity Criterion) olarak adlandırılan karmaşıklığın

bilgisel ölçütü kriterini incelemiştir. AIC, CAIC ve ICOMP kriterlerinin performanslarını farklı bileşen kovaryans matris yapısında karma dağılım kümeleme sonucu doğru küme sayısını bulma bakımından karşılaştırmıştır. Model seçim kriterlerinin kullanıldığı bu çalışmada ayrıca karma model kümeleme analizinin uygulamasına yönelik diyabet gerçek veri kümesini incelemiştir. Bu veride, gizli olmayan diyabetik, kimyasal diyabetik ve normal bireylerin gerçek sınıfları bilinmeden nasıl kümelendiklerini göstermiştir. Obez olmayan yetişkin 145 birey üzerinden beş değişkene bağlı olarak klinik yöntemlerle elde edilmiş sınıflandırmalara ilişkin Reaven ve Miller (1979) tarafından yapılan orijinal çalışmada elde edilen sonuçlarla, karma model kümeleme analiziyle elde edilen sonuçları karşılaştırmıştır. Küme sayıları bilinen ve simülasyon yoluyla üretilen çok değişkenli normal veri kümelerine dayalı sayısal örnekler incelenmiş, en uygun modelin ve küme sayılarının seçimi için model seçim kriterlerinin etkinliğini araştırmıştır. Bu yöntemlerde eş anlı olarak kümelemede uyumun eksikliğini, parametre sayısını, örneklem büyüklüğünü ve artan küme sayısıyla ortaya çıkan karmaşıklığı dikkate almıştır.

Celeux ve Govaert (1995), tutumlu ya da cimri normal dağılım modellerine dayalı kümeleme (Gaussian parsimonious clustering models) başlıklı çalışmasında Banfield ve Raftery (1993) tarafından yapılan çalışmaya benzer şekilde karma modelin bileşen kovaryans matrislerinin özdeğer ayrışımı ile parametrik olarak ifade etmiş ve bu sayede ortaya çıkan yeni küme yapılarını incelemiştir. Kovaryans matrisinin parametrik ifade edilmesi, bileşen kümelerinin yönü, şekli ve hacmi gibi özelliklerinin de dikkate alınmasını sağlamıştır.

Celeux ve Govaert (1995) tarafından yapılan çalışmada, karma dağılım modelinin bileşen oranları eşit ve farklı olması durumlarının da dikkate alındığı toplam 28 model önermişlerdir. Her bir model için parametrelerin en çok olabilirlik tahminlerinin nasıl elde edildiğini göstermişlerdir. Verinin çok değişkenli normal dağılımların karması ile modellenebileceği kabulüne dayalı olarak elde edilen farklı küme yapılarında parametre tahminlerini simülasyon ve gerçek veri kümelerini kullanarak karşılaştırmışlardır. Kullanılan gerçek veri kümesi daha önce Soubiran (1993) tarafından incelenen hareketli yıldız verisidir (stellar kinematics data). Bu veri

kümesi 2370 yıldızın samanyolu merkezine ve eksenine olan hızlarına ait iki değişken değerlerini içerir. Hareketli yıldız veri kümesini farklı hacimli küresel üç bileşen kümeli karma dağılım modeliyle modellemişlerdir. Celeux ve Govaert (1995), çalışmalarının sonunda önerdikleri küme yapıları ile çok değişkenli normal dağılımların karmasının birçok veriyi kümelemede karmaşık algoritmaların kullanılmasına gerek kalmaksızın başarıyla kümelediklerini belirtmişler ve pratik uygulamalarda özellikle küçük hacimli veri kümeleri için eşit hacimli modellerin seçimini önermişlerdir.

McLachlan ve Peel (1996), eğitimsiz öğrenme için karma normal dağılım modellerine dayalı bir algoritma (an algorithm for unsupervised learning via normal mixture models) başlıklı çalışmalarında etiket vektörü bulunmayan veriye bir normal karma modeli uyduran ve uydurulan bu modelin bileşenlerine gözlemleri sınıflandıran bir algoritma geliştirmişlerdir. Gözlemler birbirinden ayrık g bileşenli bir kümenin her bir bileşenine tahmin edilmiş bileşene ait olma sonraki olasılıklarına bakılarak yapılmıştır. Verideki her bir gözlem bileşene ait olma sonraki olasılığı yüksek olan karmanın bileşenine atanmıştır. McLachlan ve Peel (1996) bu çalışmalarında EM algoritmasını kullanarak parametrelerin en çok olabilirlik tahminlerinin bulunması için en çok olabilirlik fonksiyonun çözümünden elde edilen çok katlı köklerden hangisinin seçileceği problemi üzerinde durmuşlardır. Birden fazla grup içeren veriye karma modeli, EM algoritmasının parametre başlangıç değerlerinin seçimini de yapacak şekilde uyduran NMM (Normal Mixture Model) adında bir algoritma geliştirmişlerdir. Bu algoritmada parametre başlangıç değerleri: (a) rastgele olarak, (b) aşamalı hiyerarşik kümelemeye dayalı olarak ve (c) genelleştirilmiş öğrenme vektör niceleme adında ve GLVQ (Generalized Learning Vector Quantization) ile gösterilen kümeleme yöntemi (Pal ve ark. 1993) ile seçilmektedir. Geliştirdikleri NMM algoritması için oluşturdukları programın özelliklerini tanıtmışlardır. Başlangıç değerleri rastgele alınan yöntemle NMM algoritmasının performansının diğer parametre başlangıç değerlerinin seçim yöntemleri ile uygulanan NMM algoritmasına karşı daha iyi performans gösterdiğini simülasyon verileri üzerinde göstermişlerdir.

Bensmail ve ark. (1997), modele dayalı kümeleme analizinde sonuçlar (inference in model based cluster analysis) başlıklı çalışmalarında; Banfield ve Raftery (1993) ve, Celeux ve Govaert (1995) tarafından önerilen çok değişkenli normal dağılımların karmasına dayalı kümeleme analizinde karmanın bileşen kovaryans matrislerinin geometrik olarak modellenmesi yöntemi ile iyi sonuçlar elde edilmesine karşın bazı sınırlamalara sahip olduğunu belirtmişlerdir. Bu sınırlamalar parametre tahminlerinin yanlı olabileceğini, bileşen kovaryans matris yapısının geometrik şeklinin araştırmacı tarafından analize başlanmadan önce seçilmesi gerektiğini, değişik mümkün modeller arasından hangisinin hangi durumlarda seçilebileceğine dair kesin bir yöntemin olmaması olarak belirtmişlerdir. En iyi modeli ve küme sayısını bir adımda belirleyen ve daha önce Banfield ve Raftery(1993) tarafından önerilen AWE kriterinden daha doğru bir Bayes faktörü yaklaşımını önermişlerdir. Yaklaşık Bayes faktörlerini, Laplace-Metropolis tahmin ediciyi (Lewis ve Raftery 1997) kullanarak Gibbs örnekleyici çıktısı (Gibbs sampler output) ile hesaplamışlardır.

Bensmail ve ark. (1997), çalışmalarında simülasyon veri kümelerinin yanında; Kelebek verisi (Celeux ve Robert 1993) ve Hareketli yıldız verisi (Soubiran 1993) gibi gerçek veri kümelerini de incelemişlerdir. Kelebek verisi 23 kelebeğin 4 kanat uzunluğuna ait bir veri kümesidir. Bu veride amaç farklı kelebek türlerini tespit ederek sınıflandırmaktır. Bensmail ve ark. (1997), bu veri seti için karma dağılım modelini; bileşen kovaryans matrislerinin hacimleri farklı, yön ve şekilleri aynı olan 4 bileşenli karma dağılım modeli olarak oluşturmuşlardır. Her bir kelebeği daha sonra bu kümelere sınıflandırmışlardır. Sınıflandırma sonucu 4 kümeden birinde yalnız bir kelebek olduğu görüldüğünden bu kümeyi ihmal etmişlerdir.Kelebeklerin gerçek sınıfları ile buldukları sınıf sonuçları oldukça yakındır. Hareketli yıldız verisi için de benzer karma dağılım model yapısını kullanmışlar ve veriyi 3 küme olarak sınıflandırmışlardır.

Fraley ve Raftery (1998), modele dayalı kümeleme analizi ile küme sayısı kaçtır? Hangi kümeleme yöntemi? sorularına cevaplar (how many clusters? which clustering method? answers via model based cluster analysis) başlıklı çalışmalarında bir veri kümesinin içerdiği küme sayısı ve bileşen küme yapısı hakkında önbilgiye

sahip olmadan küme yapısının nasıl belirlenebileceği üzerine durmuşlardır. Homojen olmayan bir veri kümesinin karma dağılımlar ile temsil edilebileceğini ve bu karma dağılımın da her bir bileşenin farklı bir kümeye karşılık geldiğini belirtmişlerdir. Karma dağılımın bileşen yoğunluk fonksiyonunun çok değişkenli normal dağılım olması durumunda ise bileşen kovaryans matrislerinin parametrik özelliklerine göre farklı geometrik yapıya sahip modellerin elde edilebileceğini söylemişlerdir. Verideki parçalanmalar durumunda her bir parçalanmaya karşılık gelen parametreleri, en çok olabilirlik için EM algoritmasını kullanarak tahmin etmişlerdir. EM algoritması için parametre başlangıç değerleri olarak eklemeli hiyerarşik kümeleme yönteminden elde edilen parametre tahminlerini kullanmışlardır. Farklı bileşen kovaryans yapısına sahip modelleri, Bayesci bilgi kriterine BIC (Bayesian Information Criterion) (Schwarz, 1978) dayalı olarak hesaplanan bir Bayes faktör yaklaşımını kullanarak karşılaştırmışlardır. Bu yöntem, önem testlerinden farklı olarak aynı anda ikiden fazla modelin karşılaştırılmasında kullanılmaktadır. Bu sayede küme sayısını ve küme yapısını belirleme problemleri eş anlı olarak çözülür.

Fraley ve Raftery (1998) tarafından önerilen modele dayalı kümeleme yönteminin adımları aşağıda verilmiştir.

1. Maksimum küme sayısı M ile gösterilsin. M değeri seçilir. Bu sayı mümkün olduğunca küçük alınmalıdır.
2. Küme sayısı 2'den başlayarak maksimum küme sayısı M'ye kadar eklemeli hiyerarşik kümeleme analizi uygulanılarak kısıtsız karma normal model için gözlemler sınıflandırılır.
3. Hiyerarşik kümeleme analizinden elde edilen sınıflandırmalara dayalı olarak her bir kümenin parametreleri hesaplanır. Hesaplanan bu parametreler EM algoritması için parametre başlangıç değerleri olarak kabul edilir ve EM algoritması küme sayısı 2,…, M olmak üzere her bir küme sayısında farklı geometrik yapılı tüm normal kümeleme modelleri için uygulanılır.
4. Küme sayısı 1'de dahil olmak üzere küme sayısı M'ye kadar olan her bir model için BIC değerleri bulunur.
5. Her bir küme sayısında, modellerin her biri için elde edilen BIC değerlerinin bir matrisi oluşturulur.

6. Modellerin her biri için, BIC değerlerine ait grafikleri çizilir. Bu grafiklerde ilk belirleyici yerel maksimum değere karşılık gelen küme sayısı ve yapısı en uygun küme sayısı ve küme modeli olarak alınır.

şeklindedir. Burada belirleyici BIC değeri için Kass ve Raftery (1995)'nin çalışmasında belirtilen:

i. BIC değerleri arasındaki fark ≤ 2 ise karşılaştırılan modeller arasında fark vardır hipotezi için yeterli bir kanıt yoktur.

ii. $2 <$ BIC değerleri arasındaki fark ≤ 6 ise karşılaştırılan modeller arasında fark vardır hipotezi için pozitif yönde bir kanıt vardır.

iii. $6 <$ BIC değerleri arasındaki fark ≤ 10 ise karşılaştırılan modeller arasında fark vardır hipotezi için kuvvetli bir kanıt vardır.

iv. BIC değerleri arasındaki fark > 10 ise karşılaştırılan modeller arasında fark vardır hipotezi için oldukça kuvvetli bir kanıt vardır.

önermeleri dikkate alınır.

Fraley ve Raftery (1998), çalışmalarında önerdikleri bu yöntemi 6 farklı bileşen kovaryans matris yapısına sahip çok değişkenli normal dağılımların karmasını kullanarak 3 değişken ve 145 gözlemden oluşan diyabet verisi üzerinde incelemişlerdir. Bu inceleme sonucunda diyabet verisi için en iyi modelin, ilk yerel maksimum BIC değerine karşılık gelen 3 bileşenli kısıtlanmamış bileşen kovaryans matris yapısına sahip karma model olduğunu belirtmişlerdir. Bileşen küme kovaryans matris yapısına göre ortaya çıkan diğer karma yapılarını incelediklerinde küme sayısının 8 veya 9 olabileceğini belirtmişlerdir. Küme sayısı 9 olması durumunda, karma modelin bir bileşenine 3 gözlem sınıflanmaktadır. Bu 3 gözlem kullanılarak karmanın bu bileşeni için kovaryans matrisinin tahmin edilemeyeceğini ve ayrıca karmanın bileşen kovaryans matrisinin singüler olma probleminin diğer bir ifadeyle kötü koşulluluk durumunun ortaya çıkabileceğini belirtmişlerdir.

Biernacki ve Govaert (1999), modele dayalı kümelemede ve ayrıştırma analizinde model seçimi (choosing models in model based clustering and discriminant analysis) başlıklı çalışmalarında daha önce Celeux ve Govaert (1993) tarafından geliştirilen bileşen küme kovaryans matrislerinin özdeğer ayrışımına dayalı olarak parametrik ifade edilmesiyle ortaya çıkan küme modellerinin seçiminde

kullanılan AIC (Akaike 1973), AIC3 (Bozdogan 1992), BIC (Schwarz 1978), ICOMP (Bozdogan 1988), NEC (Celeux ve Soromenho 1996), CLM (Biernacki ve Govaert 1997) ve çapraz geçerlilik kriterinin (Smyth, 2000) performansını Monte Carlo simülasyon çalışmaları ile karşılaştırmışlardır. Çalışmanın sonucunda verideki küme sayısı G bilindiğinde AIC3 ve BIC kriterleri en iyi karma modelin belirlenmesinde daha başarılı olmalarına karşın, BIC kriterinin kullanılmasını önermişlerdir. Verinin bileşen küme sayısı bilinmediği durumlarda yapılan simülasyon çalışmasında her iki kriterinde küme sayısını olduğundan daha fazla bulmaya eğilimli olduğunu belirtmişlerdir. CLM ve NEC bilgi kriterleri, küme sayısını bulmada genellikle daha başarılı kriterler olmasına karşın istenilen en iyi modelin seçiminde başarısız olduğunu belirlemişlerdir. Bu sonuçlara göre eş anlı olarak küme sayısının seçimi ve en iyi kümeleme yönteminin belirlenmesi için belirli bir kriterin önerilmesinin zor olduğunu belirtmişlerdir. Bu bilgi kriterlerinin performansları ayrıca ayrıştırma analizi için de karşılaştırılmış ve özellikle küçük hacimli örneklem durumunda çapraz geçerlilik kriterinin başarılı sonuçlar verdiği sonucuna varmışlardır. Örneklem hacminin büyüdükçe AIC3 ve BIC kriterlerinin daha basit hesaplanabilir olmasından dolayı, çapraz geçerlilik kriterine karşı tercih edilebileceğini önermişlerdir.

Yeung ve ark.(2001), gen ifade verisi için veri dönüşümleri ve modele dayalı kümeleme (model based clustering and data transformations for gene expression data) başlıklı çalışmalarında temel amaç olarak modele dayalı kümeleme yönteminin gen ifade verilerinde uygulanabilirliğini ve BIC kriterine göre elde edilen küme modellerinin yorumlanabilirliğini göstermişlerdir. Modele dayalı kümelemenin başarısını deneye dayalı yöntemler içerisinde oldukça başarılı sonuçlar veren CAST (Ben-Dor ve Yakhini 1999) yöntemi ile karşılaştırmışlar ve kümeleme sonuçlarını da Uyarlanmış Rand İndeksi (Hubert ve Arabie 1985) kriterini kullanarak değerlendirmişlerdir. Bu çalışmada gen ifade verilerine literatürde yaygın olarak kullanılan 3 dönüşüm yöntemi uygulanmıştır. Uygulanan yöntemler logaritmik, karekök ve her bir gen için ham ifade seviyelerinin merkezi standartlaştırılması dönüşümleridir. Her bir sınıf içerisindeki veri için çok değişkenli normallik kabulünü ya da varsayımını Aitchison (1986) tarafından önerilen testlerle incelemişlerdir. Bu

yöntemlerden en basiti sınıf içerisindeki gözlemlerin marjinal dağılımlarının tek değişkenli normallikten sapmalarının test edilmesidir. Bu çalışmada ikisi gerçek, biri yapay olmak üzere 3 veri kümesini kullanmışlardır. İnceledikleri gerçek veri kümelerinden birincisi Schummer ve ark. (1999) tarafından elde edilen ovary verisinin bir alt kümesidir. Ovary verisinin incelenen alt kümesinde 235 klon ve 24 doku örneklemi (deneme) bulunmaktadır. Diğer gerçek veri kümesi yeast cell cycle verisidir (Cho ve ark. 1998). Yeast cell cycle verisi 6000 genin 17 farklı zaman noktasındaki ifade seviyelerini içermektedir. Bu veri kümesinden alınan iki farklı alt küme incelenmiştir. Yeung ve ark.(2001) modele dayalı kümeleme yönteminin gen ifade verilerine uygulanmasında karşılaştıkları zorlukları da belirtmişlerdir. Bunlar karmanın bileşenine sadece birkaç gözlem düşmesi durumunda parametre tahminin zorlaşması, ayrıca BIC değerlerine göre belirlenen fazla bileşenli en iyi karma model yapılarında parametre sayısının verideki gözlem sayısından fazla olması durumlarıdır. Çalışmanın sonucunda Modele dayalı kümelemenin gen ifade verileri için oldukça kullanışlı bir teknik olduğunu belirtmişlerdir. Gen ifade verilerine uygulanılan dönüşümler sonucu normalliğin arttığını ve farklı dönüşümlerle elde edilen kümeleme performanslarının değiştiğini gözlemişlerdir. Gerçek veri kümeleri üzerinde karma normallik kabulünü bazı dönüşümler altında bile mükemmel derecede sağlamasa da modele dayalı kümeleme yönteminin doğru küme sayısını belirlemede ve daha yüksek kalitede küme yapıları elde etmede CAST yöntemine göre az da olsa daha iyi sonuçlar verdiğini belirtmişlerdir. Yapay veri kümeleri üzerinde BIC değerine göre elde edilen küme sayısı ve köşegen bileşen küme kovaryans matris yapısına sahip karma model ile oldukça yüksek kalitede kümeler elde etmişlerdir. Kümelenmenin amacına uygun olacak şekilde gen ifade verilerine bir dönüşümün uygulanmasının yararlı olacağını önermişlerdir.

Figueiredo ve Jain (2002), sonlu karma modellerin eğitimsiz öğrenmesi (unsupervised learning of finite mixture models) başlıklı çalışmalarında çok değişkenli verinin sonlu bir karma dağılım modelinin eğitilmesi için eğitimsiz bir algoritma geliştirmişlerdir. Geliştirdikleri algoritmanın iki önemli amacı bulunmaktadır. Birincisi bileşen küme sayısının seçimi, ikincisi ise bu algoritmanın EM algoritmasından farklı olarak başlangıç değerlerinden etkilenmemesidir.

Önerdikleri bu algoritmanın EM algoritması gibi parametre uzayında tekil bir noktaya yakınsama ihtimalinin olmadığını iddia etmişlerdir. Bu algoritmada, aday modeller arasından birinin seçimine yönelik bir model seçim kriteri kullanılmamıştır. Bu algoritmanın, EM algoritması yazılabilen her türlü parametrik karma model yapısında kullanılabileceğini belirtmişlerdir. Çalışmanın sonunda karma modellerde sapan değerlerin belirlenmesinin önemli bir konu olduğunu, sapan değerlerin ekstra bir bileşenle (büyük varyanslı düzgün veya normal dağılım) temsil edilebileceğini iddia etmişler ve bu konu üzerinde halen çalışmaya devam ettiklerini belirtmişlerdir.

Fraley ve Raftery (2002), modele dayalı kümeleme, ayrıştırma analizi ve yoğunluk tahmini (model based clustering, discriminant analysis, and density estimation) başlıklı çalışmalarında kümeleme analizini bir veri içerisinde gözlemlere ait grupların bulunmasında kullanılan bir yöntem olarak tanımlamışlardır. Pratik uygulamalarda kullanılan birçok kümeleme yönteminin denemeye dayalı olmasına karşın temelinde mantıklı yöntemler olduğunu belirtmişlerdir. Kümeleme analizinde ortaya çıkan örneğin veride kaç küme vardır?, hangi kümeleme yöntemi kullanılmalıdır? ve sapan değerler nasıl incelenmelidir? gibi problemlerin çözümlerini içeren çok az sistematik bir kılavuz bilgisinin olduğuna dikkat çekmişlerdir. İstatistiksel yaklaşım temelli, modele dayalı kümeleme ile bütün bu sorunların çözümlerine yönelik genel bir metodoloji araştırması yapmışlardır. Elde ettikleri sonuçların ayrıştırma analizi ve yoğunluk tahmini gibi çok değişkenli istatistiksel problemler içinde önemli olduğunu belirtmişlerdir. Bu çalışmalarında kullanılan gerçek veri kümeleri; göğüs kanseri teşhis verisi (diagnostic breast cancer data) verisi (http://archive.ics.uci.edu/ml/datasets.html) ve mayın belirleme verisidir (http://archive.ics.uci.edu/ml/datasets.html).

Fraley ve Raftery (2002) tarafından yapılan çalışmada, göğüs kanseri teşhisi verisi için gerçekte 30 değişkenli olmasına karşın daha önce yapılmış birçok çalışma sonrasında 3 önemli özellik değişkeni (Mangasarian ve ark. 1995) bu çalışmada da alınmıştır. Bu veri kümesinin sınıfları olan teşhisler yani veri etiket vektörü bilinmektedir. Çalışmaları sonunda modele dayalı kümeleme analizini kullanarak maksimum BIC değerine göre 3 gruplu kısıtsız karma modeli elde etmişlerdir. İki ve 3 gruplu kısıtsız karma modellerin BIC değerleri arasındaki farkın oldukça küçük

olduğunu ve bu durumda veri kümesi için 2 veya 3 kümeli karma modelin alınabileceğini belirtmişlerdir. Klinik çalışmalar sonucu elde edilen teşhis sınıfları ile iki gruplu kısıtsız karma model sonuçlarının 519 gözlemde 29 gözlem hariç birbirine benzediği sonucuna varmışlardır. Bu veri kümesi için elde edilen 3 gruplu karma model yapısının da klinik çalışmalar için önemli olabileceğini, bu yapıda ortaya çıkan ara sınıfın klinik çalışmalarla ayrıca incelenmesi gerektiğini önermişlerdir.

Fraley ve Raftery (2002) tarafından yapılan çalışmada incelenen bir diğer veri mayın belirleme verisidir. Mayın belirleme verisi Amerikan deniz kuvvetlerince mayınların teşhisi için yapılan bir projede Witherspoon ve ark. (1995) tarafından geliştirilen COBRA (The Coastal Battlefield Reconnaissance and Analysis) programında kullanılan bir görüntü verisidir. COBRA programı ile incelenen görüntü içerisinde mayın olabilecek 173 bölge tespit edilmiştir. Tespit edilen bu bölgelerden yalnız 35 tanesi gerçekte mayın içeren bölgedir. Bu veriye modele dayalı kümeleme yöntemini uyguladılar. Analiz sonucunda BIC değerlerine göre küresel sabit kovaryans matris yapılı 4 bileşen kümeli karma model elde etmişlerdir. Bu karma modelin bir bileşen kümesi 89 gözlemden oluşmaktadır. Bu gözlemlerden 35 tanesi mayınlı bölgedir. Bu sayede içerisinde mayın olan bölgelerin tespitinde COBRA programı ile olan başarı oranı $35/173$ iken modele dayalı kümeleme analizi sonucunda başarı oranını $35/89$ olarak elde etmişlerdir.

Soffritti (2003), bir veri matrisinden birden fazla küme yapısının belirlenmesi (identifying multiple cluster structures in a data matrix) başlıklı çalışmasında her türlü değişken tipinde uygulanılabilen hiyerarşik aşamalı kümelemeye benzer bir yöntem önermiştir. Önerilen yöntemde verinin her bir değişkeninin küme yapısını içerip içermediğini ayrı ayrı incelemiş ve küme yapısı gösteren değişkenlerin oluşturduğu değişken alt kümesi üzerinde genel küme yapısını araştırmıştır.

Goldberger ve Roweis (2004), bir karma modelin hiyerarşik kümelenmesi (hierarchical clustering of a mixture model) başlıklı çalışmalarında çok fazla bileşeni bulunan bir karma modeli, bileşen yapılarını koruyacak şekilde daha az bileşen sayılı bir karma modele indirgeyecek bir algoritma önermişlerdir. Bunun için asıl karma modelin bileşenlerini gruplayacak iki karma model arasındaki uzaklığı hesaplamada kullanılan yeni bir uzaklık ölçütü kullanmışlardır. Bu uzaklık ölçütüyle iki çok

değişkenli normal karma model arasındaki Kullback Leibler yakınsama kriterinden farklı biçimde analitik olarak hesaplamışlardır.

Fraley ve Raftery (2007), kemometrikte Mclust yazılım programını kullanarak modele dayalı sınıflandırma yöntemleri (model based methods of classification using the Mclust software in chemometrics) başlıklı çalışmalarında modele dayalı kümeleme analizi için son zamanlarda geliştirilen yöntem ve yazılımlar sayesinde kümeleme ve sınıflandırma çalışmalarında deneye dayalı yöntemlere göre olasılık modellerine dayalı kümeleme yöntemlerinin tercih edilmesine sebep olduğunu belirtmişlerdir. Modele dayalı kümeleme ve sınıflandırma analizlerinde yer alan kemometrik çalışmaların çok değişkenli görüntü analizlerinin (Wehrens ve ark. 2004, Tran ve ark. 2006), Manyetik salınım görüntülemeyi (Fraley ve ark. 2005), mikrodizilim görüntü bölütlemeyi (Li ve ark. 2005), istatistiksel işlem kontrolünü (Thissen ve ark. 2005) ve gıda güvenilirliğini (Toher ve ark. 2005, Dean ve ark. 2006) içerdiğini belirtmişlerdir. Fraley ve Raftery (2007) çalışmalarında, kemometrik ölçümleri içeren veri kümeleri üzerinde Mclust yazılımının modele dayalı kümelemede, yoğunluk tahmininde ve ayrıştırma analizinde nasıl kullanılacağını göstermişlerdir.

Modele dayalı kümeleme analizi çerçevesinde önemli konulardan biri de özellik ya da değişken seçimi olmuştur. En iyi özellik ya da değişken alt kümesinin seçimi, özellik seçimi olarak adlandırılır. Son zamanlarda fen ve mühendislik alanlarında ortaya çıkan gelişmelere dayalı olarak daha büyük boyutlu veri kümeleri ile çalışma zorunluluğu doğmuştur. En iyi özellik alt kümesi seçimi bu alanlarda büyük sayıda özellikler içeren veri kümelerinin analizleri için önemlidir (Han ve Kamber 2001, Jin ve ark. 2002, Bradley ve ark. 2000). Buna bir örnek moleküler biyolojideki binlerce özelliği içeren verilerin sınıflandırılması (Baldi ve Hatfield 2002, Xing ve ark. 2001) verilebilir. Başka bir örnekte, bir web sitesindeki binlerce farklı anahtar kelime ile temsil edilen web sayfalarının sınıflandırılmasıdır (Vaithyanathan ve Dom 1999). Başka bir örnekte, görsel tabanlı görüntü sınıflandırma yöntemlerinde her pikselin bir özellik olarak kullanılabilir olmasından dolayı binlerce özelliğin görüntü sınıflandırılmasında (Bins ve Draper 2001) kullanılması sayılabilir.

Özellik seçimi, eğitimli sınıflandırmada verinin en yüksek doğrulukla kümelenmesini sağlayacak şekilde, yoğun olarak çalışılmaktadır (Blum ve langley 1997, Jain ve Zongker 1997, Kohavi ve John 1997, Koller ve Sahami 1997). En uygun küme sayısı ve en uygun özellik alt kümesi birbirleriyle ilişkilidir (Dy ve Brodley 2000). Örneğin temel bileşenler analizi gibi varyans analizlerine dayalı yöntemlerin kullanılmasıyla kümelemenin amacına uygun olacak şekilde iyi özellikler her zaman seçilemez. Büyük varyansa sahip özellikler, verideki gerçek gruplamadan bağımsız olabilir. Özellik seçimi algoritmalarının (Caruana ve Freitag 1994, Kohavi ve John 1997, Pudil ve ark. 1994) çoğu, tüm özellik alt kümeleri üzerinde inceleme yapan tümleşik bir araştırma içerisinde yer alır. Bu algoritmalarda genellikle deneye dayalı ya da sezgisel yöntemler benimsenmiştir.

Çok boyutlu veri kümeleri ile çalışılan görüntü analizinde önemli bir problem görüntü bölütlemedir. Görüntü bölütleme, bir kümeleme problemi olarak ifade edilebilir (Frugui ve Krishapuram 1999, Jain ve Flynn 1996, Shi ve Malik 2000). Temelde her desen hakkında daha fazla bilgiye sahip olunması durumunda bir kümeleme algoritmasının daha iyi performans göstermesi beklenir. Bu durumda desenleri temsil edecek mümkün olduğunca çok özelliğin kullanılması önerilir. Fakat bu pratikte her zaman kolay olmaz. Eğitimli sınıflandırmada özellik seçiminin sınırlı miktarda veriden oluşan eğitici sınıflandırıcıların performanslarını artırabileceği bilinmektedir (Raudys ve Jain 1991). Bu da daha ekonomik sınıflandırıcıların ve birçok durumda yorumlanabilir modellerin ortaya çıkmasını sağlar.

Law ve ark. (2003), karma dağılımlara dayalı kümeleme için özellik seçimi (feature selection in mixture based clustering) başlıklı çalışmalarında normal dağılımların karmasına dayalı kümeleme analizinde kullanılabilecek iki özellik seçim yöntemini önermişlerdir. Birinci yöntemde özeliklerin önem derecesi tahmin edilmektedir. Bunun için bir EM algoritması geliştirmişlerdir. İkinci yöntemde ise Koller ve Sahami (1996) tarafından önerilen karşılıklı bilgiye dayalı özellik ilişkisi kriterinin (mutual information based feature relavance criterion) eğitimsiz öğrenme durumuna uyarlanmasıdır. Özellik önem derecesinin tahminine dayalı özellik seçim yöntemi, şarap gerçek veri kümesi (http://archive.ics.uci.edu/ml/datasets/Wine) kullanılarak elde edilen sonuçlar eğitimli sınıflandırma için geliştirilmiş LNKnet

yazılım programının (http://www.ll.mit.edu/mission/communications/ist/lnknet.html) sonuçları ile karşılaştırılmıştır. Şarap verisi 3 sınıftan oluşan 13 değişkenli bir veri kümesidir. LNKnet yazılım programı ile 9 değişken seçilmiştir. Law ve ark. (2003) önerdikleri özellik önem derecesine dayalı yöntemi kullanarak değişkenleri önem sırasına sıraladıklarında belirledikleri ilk dokuz değişken içerisindeki 8 değişken LNKnet ile belirlenen değişkenlerle aynıdır. Law ve ark. (2003), bu sonucu eğitimsiz bir yöntemde elde ettiklerinden dolayı önerdikleri algoritmanın başarılı olduğunu iddia etmişlerdir. Law ve ark. (2003) tarafından yapılan çalışmada önerilen ikinci özellik seçim kriteri, küme sayısının bilindiği ve bileşen küme kovaryans matrisi üzerinde herhangi bir kısıtın olmadığını kabul ederek bir simülasyon verisi üzerinde test etmeyi içerir. Law ve ark. (2003) çalışmalarının sonucunda önerdikleri iki yeni özellik seçim yönteminin birbirine karşı avantaj ve dezavantajlarının olduğunu belirtmişlerdir.

Raftery ve Dean (2006), modele dayalı kümelemede değişken seçimi (variable selection for model based clustering) başlıklı çalışmalarında modele dayalı kümeleme için değişken ya da özellik seçimi problemi üzerinde durmuşlardır. Değişken seçiminde içeren küme sayısının tahmin edilmesinde ve en iyi kümeleme modelinin belirlenmesinde karmanın bileşen kovaryans yapısı ile ilgili kullanılabilecek bir arama (greedy search) algoritmasının kullanılmasını önermişlerdir. Değişken seçimi ve kümelenme yöntemlerinin bir kombinasyonu olan arama (greedy search) algoritmasının adımları aşağıda verilmiştir.

1. Tek değişkenli kümeleme yapısını en yüksek oranda veren değişken, birinci kümeleme değişkeni olarak alınır.
2. Birinci adımda seçilen birinci kümeleme değişkeni ile birlikte iki değişkenli kümeleme yapısını en yüksek oranda elde edilen değişken, ikinci kümeleme değişkeni olarak alınır.
3. Bir önceki adımlarda elde edilen değişkenlerle birlikte bir önceki adımda elde edilen kümeleme oranı da dikkate alınarak en yüksek çok değişkenli kümeleme oranını veren değişken, kümeleme değişkeni olarak seçilir.
4. Değişken bir önceki değişken kümesine eklendiğinde kümelemeye katkısı olmuyorsa en iyi özellik vektörüne alınmaz ve modelden atılır.

5. Algoritmaya 3. ve 4. adımların her ikisi de aynı anda ret edildiği duruma kadar devam edilir.

Bu algoritmada kümeleme ölçüt kriteri olarak BIC kullanılır. Algoritmanın her bir adımında en yüksek BIC değerine göre küme sayısı ve kümeleme modeli belirlenir. Raftery ve Dean (2006) çalışmalarında bu algoritmanın etkinliğini simülasyon ve gerçek veri kümeleri üzerinde test etmişlerdir. Kullandıkları gerçek veri kümelerinden ilki Leptograspsus yengeç verisidir (Campbell ve Mahon 1974). Yengeç verisinde 100 portakal renkli ve 100 mavi renkli yengeç türlerinde eşit sayıda dişi ve erkek yengeç bulunmaktadır. Toplam 200 yengece ait 5 özellik değeri ölçülmüştür. Veri 4 sınıf içermektedir. Bunlar: 1) Mavi erkek, 2) mavi dişi, 3) portakal renkli erkek, 4) portakal renkli dişi şeklindedir. Değişken seçim yöntemi kullanılmadan uygulanılan modele dayalı kümeleme yöntemi ile 7 bileşenli karma model elde etmişlerdir. Küme sayısı sınıf sayısından fazladır ve elde edilen kümelerde homojen sınıf bireyleri bulunmamaktadır. Sınıflandırma hata oranını %42.5 olarak bulmuşlardır. Raftery ve Dean (2006) önerdikleri arama (greedy search) algoritmasını bu veriye uygulamışlar ve uygulama sonucunda 4 değişkenin kümeleme değişkeni olarak alınabileceğini ve bu değişkenler kullanılarak bu veri için 4 bileşenli karma modelin en iyi model olduğunu göstermişlerdir. Bu arama algoritması sayesinde yengeç verisinin sınıflandırma hata oranını %7.5 olarak hesaplamışlardır.

Raftery ve Dean (2006) tarafından kullanılan ikinci gerçek veri kümesi Iris veri kümesidir (Anderson 1935). Iris veri kümesi 3 farklı türdeki: 1) Iris setosa, 2) Iris versicolor, 3) Iris virginica Iris çiçeğinin dört özelliğine ait her bir tür için eşit sayıda olmak üzere toplam 150 gözlem değerleri içermektedir. En yüksek BIC değerine göre elde edilen küme sayısı iki veya üç olmaktadır. Üç bileşenli karma model için elde edilen BIC değeri, iki bileşenli karma model için elde edilen BIC değerine göre bir birim daha büyüktür. İki bileşenli karma model yapısı incelendiğinde, karmanın bir bileşen kümesinde Iris setosa türü oldukça iyi ayırt edilmiş, diğer bileşen kümesi ise Iris versicolor ve Iris virginica türlerini içermiştir. Bu durumda minimum sınıflandırma hata oranını %33 olarak elde etmişlerdir. Üç bileşenli karma model incelendiğinde ise hatalı sınıflandırma oranını %3.3 olarak

bulmuşlardır. Hatalı sınıflandırma oranı üç bileşenli karma model yapısı için oldukça düşük olmasına karşın araştırmacının bu modeli seçmesine sebep olacak açık bir kanıt yoktur. Raftery ve Dean (2006), bu veri kümesine uyguladıkları arama (greedy search) algoritması sonucu üç küme değişkeni belirlemişler ve bu küme değişkenlerini kullanarak en yüksek BIC değerine göre 3 bileşenli karma model elde etmişlerdir. Bu karma model yapısına göre hatalı sınıflandırma oranını %4 olarak hesaplamışlardır. En yüksek ikinci BIC değerinde 4 bileşenli karma model yapısını elde etmişlerdir. Iris verisinin 3 değişken kullanıldığı durumda elde edilen dört bileşenli karma model yapısı ile 3 bileşenli karma model yapısının BIC değerleri arasındaki fark 14 birim olmaktadır. Bu farkın 3 bileşenli karma model yapısının seçilmesi için kuvvetli bir kanıt olduğunu belirtmişlerdir.

Galimberti ve Soffritti (2007), bir verideki birden fazla küme yapısını belirlemek için modele dayalı yöntemler (model based methods to identify multiple cluster structures in a data set) çalışmalarında farklı değişken alt kümelerine göre gözlemlerdeki farklı parçalanışları belirleme problemi üzerinde durmuşlar ve genelleştirilmiş modele dayalı yöntem ile değişken kümeleme yöntemlerinin bir kombinasyonundan oluşan yeni bir yöntem önermişlerdir. Önerdikleri yöntemin etkinliğini simülasyon ve gerçek veri kümeleri üzerinde göstermişlerdir. Çalışmada kullandıkları gerçek veri kümesi bir İtalyan finans gazetesinde (www.ilsole24ore.com) yayınlanan İtalya'nın 103 bölgesindeki hayat kalitesiyle ilgili 7 değişken değerini içeren veridir. Bu verideki küme sayısını AIC, AIC3, ICL, BIC, CAIC gibi farklı bilgi kriterlerini kullanarak ve 7 değişkenini farklı 2 veya 3 değişken alt grubuna kümeleyerek modele dayalı kümeleme sonuçlarını incelemişledir.

3. KÜMELEME ANALİZİNDE KULLANILAN YÖNTEMLER

Kümeleme analizi yöntemleri nesneleri, desenleri, bireyleri, birimleri, olayları veya gözlemleri belirli sayıdaki kümelere, gruplara, alt kümelere veya kategorilere böler. Everitt ve ark. (2001), küme (cluster) terimi için genel olarak ifade edilebilecek tek bir tanımın olmadığını, böyle bir tanımın yapılmasının zor olmasının yanında yanlış anlamlar çıkarılabileceğini belirtmiştir. Küme terimi için yapılan bazı tanımlar:

1) Küme (cluster) birbirine benzer bireylerin bir kümesi olup, bu kümedeki bireylerin diğer kümedeki bireylerden farklı olması durumudur.
2) Bir küme test uzayında noktalar toplamı olup içerisindeki herhangi iki nokta çifti arasındaki uzaklığın, içerisindeki bir nokta ile dışındaki bir nokta arasındaki uzaklıktan küçük olması durumudur.
3) Kümeler p boyutlu özellikler uzayında oldukça yüksek yoğunluklu noktaları içeren ve nispeten daha düşük yoğunluktaki noktaları içeren diğer bölgelerden ayrılan sürekli bölgelerdir.

şeklinde özetlenebilir (Everitt 1980).

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 (Gordon 1999). Kümeleme yöntemleri başlıca iki sınıfa ayrılır. Bunlar bölünmeli (partitional) kümeleme ve hiyerarşik kümeleme yöntemleridir (Everitt ve ark. 2001, Jain ve Dubes 1988).

Bu bölümde genel anlamda kümeleme analizinde kullanılan uzaklık ve benzerlik ölçüleri incelenecektir. Hiyerarşik kümeleme yöntemlerinde kullanılan algoritmalar tanıtılacaktır. Kümeleme geçerlik kriterleri incelenecek ve özellikleri araştırılacaktır. Kümeleme yöntemleri performansları bakımından karşılaştırılacaktır.

3.1. Tanım ve Gösterimler

Bu kesimde çok değişkenli verinin bir matris ile nasıl ifade edildiği gösterilecektir. Birçok kümeleme analiz yönteminin ortaya çıkmasına neden olan değişken tipleri ve değişken ölçekleri açıklanacaktır.

3.1.1. Çok Değişkenli Verinin Matris Gösterimi

Bir araştırmada veya deneyde sosyal ya da fiziksel bir olayın anlaşılması için n sayıda birey, nesne ya da deneysel gözlemin her biri üzerinden gözlenen $p > 1$ sayıda değişken ya da karakteristiğin kayıt edilmiş ölçüm değerlerine çok değişkenli veri denir.

	1. değişken	*2. değişken*		*j. değişken*		*(p-1). değişken*	*p. değişken*
1. gözlem	x_{11}	x_{12}	…	x_{1j}	…	$x_{1(p-1)}$	x_{1p}
2. gözlem	x_{21}	x_{22}	…	x_{2j}	…	$x_{2(p-1)}$	x_{2p}
	⋮	⋮		⋮		⋮	⋮
i. gözlem	x_{i1}	x_{i2}	…	x_{ij}	…	$x_{i(p-1)}$	x_{ip}
	⋮	⋮		⋮		⋮	⋮
n. gözlem	x_{n1}	x_{n2}	…	x_{nj}	…	$x_{n(p-1)}$	x_{np}

Şekil 3.1. Çok değişkenli veride gözlem ve değişken vektörleri.

Böyle bir veri kümesi genellikle,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1(p-1)} & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2(p-1)} & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{i(p-1)} & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{n(p-1)} & x_{np} \end{bmatrix} \tag{3.1}$$

biçimindeki $n \times p$ tipinde bir $\mathbf{X}$ matrisi ile ifade edilir. Burada x_{ij}, $i.$gözlem ya da nesne üzerinden gözlenen $j.$ değişken yada karakteristik değerini ifade eder. Bu durumda (3.1)'deki eşitlikte gösterilen $\mathbf{X}$ veri matrisi $i = 1,2,\ldots,n$ ve $j = 1,2,\ldots,p$ olmak üzere her biri $1 \times p$ boyutlu $\mathbf{x}_i$ gözlem vektörleri,

$$\mathbf{X} = \left(\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_{n.}^T\right)^T \tag{3.2}$$

şeklinde de ifade edilebilir.

3.1.2. Değişken Tipleri

Bir rastgele değişken kesikli ve sürekli olmak üzere 2 kategoriye sınıflandırılabilir.

Tanım 3.1: Bir rastgele değişkenin alabileceği değerlerin sayısı sonlu veya sayılabilir sonsuzlukta ise bu rastgele değişkene kesikli rastgele değişken denir (Akdeniz 2004).

Tanım 3.2: Bir rastgele değişken bir aralıkta ya da birden çok aralıkta her değeri alabiliyorsa bu rastgele değişkene sürekli rastgele değişken denir (Akdeniz 2004).

3.1.3. Değişken Ölçekleri

Birçok istatistiksel analizde olduğu gibi kümeleme analizinde de değişkenlerin ölçek türleri büyük önem taşır. Bir değişkenin ölçüm düzeyi bu ölçümlerin matematiksel olarak nasıl ele alınacağına karşılık gelir. Stevens (1946), ölçüm düzeylerini isimsel ya da kategorik (nominal), sıralı (ordinal), aralık (interval) ve oran (ratio) olmak üzere dört sınıfta toplamıştır.

3.1.3.1. İsimsel Ölçek

Bu ölçek ile ölçülmüş değişkenler durum, etiket, kategori veya isimlerle temsil edilir. Örneğin saç rengi, futbol kulübü, doğum yeri, cinsiyet, kan grubu gibi. Her ne kadar isimsel değerler alan değişkenler sayısal değerlerle ilişkilendirilseler de bu sayısal değerlerle herhangi bir matematiksel işlem yapılamaz. Örneğin saç rengi siyah olanların 1, kahverengi olanların 2, sarı olanların 3 ile kodlanması gibi. Bu değerler arasında bir sıralanma söz konusu değildir. Örneğin sarı saçın 3 yerine 1 ile kodlanması, siyah saçın 1 yerine 2 ile kodlanması veya kahverengi saçın 2 yerine 3 ile kodlanması arasında bir fark yoktur.

3.1.3.2. Sıralı Ölçek

İsimsel ölçekte olduğu gibi değişkenler durum, etiket, kategori veya isimlerle temsil edilir. Bu ölçekle ölçülen değişken değerleri arasında sıralanma sırası önemlidir. Örneğin bir kişinin çalıştığı iş yerinden memnuniyetinin; çok memnun olmasının 1, memnun olmasının 2 ve memnun olmamasının 3 ile kodlanması gibi. Durumlarda değişken değerleri arasında bir önem sırası vardır. Diğer bir ifadeyle sıralanma en az önemli durumdan en çok önemli duruma doğru ya da tam tersi biçimde yapılır. Burada işinden çok memnun olan birinin sadece memnun olan birine göre iki kat fazla memnun olduğu söylenemez. Sadece bu kişinin diğerine göre daha fazla memnun olduğu iddia edilebilir.

3.1.3.3. Aralık Ölçeği

Aralık ölçekte ölçülmüş değişken değerleri pozitif veya negatif gerçel sayı değerler alabilir. Ölçümler arası farklar eşit öneme sahiptir. Örneğin $10^{\circ}C$ ile $20^{\circ}C$ arasındaki fark ile $-40^{\circ}C$ ile $-30^{\circ}C$ arasındaki fark eşit öneme sahiptir. Aralık ölçekte mutlak bir sıfır değeri yoktur. Bu nedenden dolayı değerler arasındaki oran anlamsızdır. Örneğin $20^{\circ}C$ ile $10^{\circ}C$ arasındaki oran ile $40^{\circ}C$ ile $20^{\circ}C$ arasındaki

oran sayısal olarak aynı olmasına karşın sıcaklığın $20^{\circ}C$ olması, $10^{\circ}C$'den iki kat daha sıcak olduğu anlamına gelmez.

3.1.3.4. Oran Ölçeği

Oran ölçeği, aralık ölçüm düzeyini kapsamakla birlikte değerler arasında oransal önem bulunmaktadır. Bu ölçekte bir değerin diğerinden ne kadar oranda az veya çok olduğu anlamlıdır. Bunun nedeni bu ölçekte mutlak bir sıfır değeri vardır. Bu düzeyde sıfır değerini alan bir değişken için bu değişkenin ölçüldüğü gözlemde bulunmadığı anlamına gelir. Örneğin aylık geliri 2000TL olan bir meslek sahibinin aylık geliri 1000 TL olan diğer bir meslek sahibinden ayda iki kat fazla kazandığı söylenebilir. Aylık geliri 0 TL olan bir kişinin herhangi bir geliri olmadığı anlaşılır.

3.2. Kümeleme Analizinde Kullanılan Benzerlik ve Farklılık Ölçüleri

Benzerliğin ölçülmesi oldukça zordur. Temel olarak benzerlik nicel olup, iki nesne veya iki özellik arasındaki ilişkinin kuvveti olarak açıklanabilir. Bu nicel değer ya -1 ile 1 arasında ya da normalleştirilerek 0 ile 1 arasında ölçeklendirilir. *i.* ve *j.* nesne veya gözlem arasındaki benzerlik s_{ij} ile ifade edilir. Bu nicel değer alınan ölçeğe veya veri tipine göre değişik yollardan elde edilir.

Uzaklık, farklılıkları ölçer. Farklılıklarda çeşitli özelliklere dayalı olarak iki nesne arasındaki zıtlık ya da uyumsuzlukların bir ölçümüdür. Farklılık aynı zaman da iki nesne arasındaki düzensizliğin veya bozukluğun bir ölçüsü olarak düşünülebilir. Benzerlik ölçümleri bir gözlemin diğerlerinden ayırt edilmesini sağlar. Bu sayede gözlemler benzerlik veya farklılıklarına dayalı olarak gruplara ayrılabilir. Gözlemler alt gruplara veya kümelere bir kere atandıktan sonra her grubun karakteristikleri anlaşılabilir ve kümelerin özellikleri açıklanabilir. Gruplama ile bilginin düzenlenmesi ve sonuçların elde edilmesi daha etkin sağlanır. Yeni gözlemler bu alt grup veya kümelere sınıflandırılabilir ve özellikleri tahmin edilebilir. Veri bu sayede basitleştirilerek daha etkin değişkenler üzerinde çalışılabilir. Veri kümesi içerisindeki

belirgin olmayan gizli yapı bulunabilir. Verinin içerisindeki kümelenme yapısına ve bu kümelerin parametre tahminine dayalı olarak plan ve kararlar alınabilir. Bu bilgiler ışığında aşağıda bazı tanımlar verilmiştir.

Tanım 3.3: (Uzaklık) $\mathbf{X}$ bir küme olmak üzere $d : \mathbf{X} \times \mathbf{X} \to \Re$ şeklinde tanımlanan bir fonksiyon eğer tüm $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ için

$$1.\ d(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \tag{3.3}$$

$$2.\ d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i) \tag{3.4}$$

$$3.\ d(\mathbf{x}_i, \mathbf{x}_i) = 0 \tag{3.5}$$

koşullarını sağlıyorsa d, $\mathbf{X}$ üzerinde bir **uzaklık** (veya farklılık) olarak adlandırılır.

Tanım 3.4: (Uzaklık Uzayı) Bir d uzaklığı ile ilişkilendirilmiş $\mathbf{X}$'in $(\mathbf{X}, d)$ kümesine bir uzaklık uzayı denir.

Tanım 3.5: (Benzerlik) $\mathbf{X}$ bir küme olmak üzere negatif olmayan, simetrik $s : \mathbf{X} \times \mathbf{X} \to \Re$ fonksiyonu tüm $\mathbf{x}, \mathbf{y} \in \mathbf{X}$ için $s(\mathbf{x}, \mathbf{y}) \leq s(\mathbf{x}, \mathbf{x})$ koşulunu sağlıyorsa ve eşitlik ancak ve ancak $\mathbf{x} = \mathbf{y}$ olduğunda gerçekleşiyorsa s fonksiyonu **benzerlik** (yakınlık) olarak adlandırılır.

Tanım 3.6: (Yarı Metrik) $\mathbf{X}$ bir küme olmak üzere $d : \mathbf{X} \times \mathbf{X} \to \Re$ şeklinde tanımlanan negatif olmayan, simetrik bir fonksiyon tüm $\mathbf{x}_i \in \mathbf{X}$ ve $d(\mathbf{x}_i, \mathbf{x}_i) = 0$ eşitliğini sağlasın. Eğer tüm $, \mathbf{x}_j, \mathbf{x}_k \in \mathbf{X}$, için d fonksiyonu

$$d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_j, \mathbf{x}_k) \tag{3.6}$$

koşulunu (Üçgen eşitsizliği) sağlıyorsa d fonksiyonuna $\mathbf{X}$ üzerinde bir yarı metrik'tir denir.

Tanım 3.7: (Metrik) $\mathbf{X}$ bir küme olmak üzere $d : \mathbf{X} \times \mathbf{X} \to \Re$ şeklinde tanımlanan d fonksiyonu, (3.3)-(3.6)'deki koşullarını tüm $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{X}$ için sağlıyorsa $\mathbf{X}$ üzerinde metrik olarak adlandırılır.

Tanım 3.8: (Metrik Uzayı) Bir $\mathbf{X}$ kümesi üzerinde tanımlanan bir d metrik ile oluşturduğu $(\mathbf{X}, d)$ kümesi metrik uzayı olarak adlandırılır.

Bir $n \times p$ boyutlu $\mathbf{X}$ veri matrisinin her biri $1 \times p$ boyutlu $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ gözlem vektörleri için $i = 1,2,\ldots,n$ ve $j = 1,2,\ldots,n$ olmak üzere $n \times n$ boyutlu $\mathbf{D} = d_{ij}$ uzaklıklar matrisi hesaplanabilir. Burada $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$'dır. $\mathbf{D}$ uzaklıklar matrisi köşegen elemanları sıfır olan simetrik bir matristir.

Sonlu uzaklık ölçümleri (metrikleri) birçok alanda temel bir araç olmuştur. Bu alanlar arasında geometri, olasılık, istatistik, grafik teorisi, kümeleme, veri analizi, desen tanımlama, görüntü analizleri, genetik, mühendislik, yapay sinir ağları, astronomi, moleküler biyoloji sayılabilir.

3.2.1. Kesikli Değişkenler için Benzerlik Ölçüleri

Değişkenlerinin tümü isimsel ya da kategorik olan verilerde gözlem çiftleri arasındaki yakınlık eşleştirme tipli benzerlik ölçüleri ile belirlenir. Ölçümler genellikle [0,1] aralığında ölçeklenmiş olsa da bazı durumlarda yüzdelik olarak da ifade edilebilir. *i.* ve *j.* gözlem arasındaki benzerlik katsayısı ya da değeri $s_{i,j}$ ile ifade edilebilir. Bu değerin sıfır olması iki gözlemin tüm değişkenleri cinsinden maksimum farklı olduğu anlamına gelir. İki gözlem arasındaki benzerlik yerine farklılık da ölçülebilir. İki gözlem arasındaki $d_{i,j}$ uzaklığı $s_{i,j}$ benzerlik katsayısından,

$$d = 1 - s,\ d = \frac{1-s}{s},\ d = \sqrt{1-s},\ d = \sqrt{2\left(1-s^2\right)},\ d = -\ln s \text{ ve } d = \arccos s \tag{3.7}$$

dönüşümler kullanılarak elde edilebilir.

3.2.1.1. İki Sonuçlu (Binary) Değişkenler İçin Kullanılan Başlıca Benzerlik Ölçüleri

Verideki tüm değişkenler sadece iki sonuçlu değer alıyorsa bu verideki gözlemler arasındaki yakınlık için birçok benzerlik ölçüsü tanımlanmıştır. İki sonuçlu değişkenlerin aldığı iki değerin öneminin eşit olup olmamasına göre bir gözlem çifti arasındaki benzerlik ölçüleri simetrik ve simetrik olmayan ölçüler şeklinde iki kategoriye ayrılır (Kaufman ve Rousseeuw 1990). İki sonuçlu değişkenler içeren gözlem çiftleri arasındaki benzerlik ölçüsü hesaplanırken eşleştirme tipli benzerlik ölçütleri kullanılır. Her biri p değişkenli iki gözlem çifti arasındaki eşleştirme tipli benzerlik ölçüsü bulunurken çaprazlık tablosunda faydalanılır. Çaprazlık tablosu, iki sonuçlu değişkenler içeren gözlem çiftinin karşılıklı eşleşen değişken değerlerinin sayısından oluşur. Tablo 3.1 ile verilen çaprazlık tablosunda her bir gözlem çiftinin karşılıklı değişkenlerinin aynı anda 1 değerini alanlarının sayısı a, 0 değerlerini alanların sayısı d, değişkenleri biri 1 diğeri 0 alanların sayısı b, tam tersi durumda c ile ifade edilmiştir.

İki sonuçlu p değişkenden oluşan veriler için birçok eşleştirme tipli benzerlik ölçüsü kullanılmakla birlikte en yaygın olarak kullanılanların bir listesi Tablo 3.2 ile verilmiştir. İki sonuçlu p değişkenden oluşan gözlemler için bu kadar çok eşleştirme tipli benzerlik ölçüsünün kullanılmasının nedeni $0-0$ eşleştirmesinin nasıl yorumlandığı ile ilgilidir. Bazı durumlarda $0-0$ eşleştirmesi, $1-1$ eşleştirmesi ile eşdeğer olabilir. Örneğin cinsiyetin erkek için 1 değeri ve bayan için 0 ile değeri kodlanması veya tam tersi şeklinde kodlanması arasında bir fark yoktur. Bazı durumlarda ise $0-0$ eşleştirmesi bir gözlem üzerinde herhangi bir özelliğin bulunup bulunmadığı ile ilgili olabilir. Örneğin bir böceğin bir özelliği kanat olup olmaması ise kanatlı olmasının 1, olmamasının 0 ile kodlanması gibi.

Tablo 3.1. İki sonuçlu p sayıda değişken içeren bir gözlem çiftinin çapraz tablosu.

		i **. gözlem**		
	Değişken değeri	1	0	Toplam
j **. gözlem**	1	a	b	$a+b$
	0	c	d	$c+d$
	Toplam	$a+c$	$b+d$	$p=a+b+c+d$

Tablo 3.2. İki sonuçlu p değişken içeren verinin gözlem çiftleri arasındaki yakınlık için kullanılan bazı eşleştirme tipli benzerlik ölçüleri.

Benzerlik Kodu	**Ölçüm**	**Formül**
S1	Eşleştirme Katsayısı (Zubin 1938)	$s_{ij}=\frac{a+d}{p}$
S2	Jaccard Katsayısı (Jaccard 1908)	$s_{ij}=\frac{a}{a+b+c}$
S3	Rogers ve Tanimoto (1960)	$s_{ij}=\frac{a+d}{a+2(b+c)+d}$
S4	Sokal ve Sneath (1963)	$s_{ij}=\frac{a}{a+2(b+c)}$
S5	Sokal ve Sneath (1963)	$s_{ij}=\frac{2(a+d)}{2(a+d)+(b+c)}$
S6	Gower ve Legendre (1986)	$s_{ij}=\frac{a+d}{a+\frac{1}{2}(b+c)+d}$
S7	Gower ve Legendre (1986)	$s_{ij}=\frac{a}{a+\frac{1}{2}(b+c)}$
S8	Dice (1945)	$s_{ij}=\frac{2a}{2a+b+c}$

3.2.1.2. İki Sonuçtan Fazla Değer Alan Kesikli Değişkenler İçin Benzerlik Ölçüleri

Tümü iki sonuçtan daha fazla değerler alan kesikli değişkenlerden oluşan veride bir gözlem çifti arasındaki yakınlık benzerlik katsayıları (skor değerleri) ile belirlenir (Everitt ve ark. 2001). Her biri p sayıda iki sonuçtan fazla değerler alan kesikli $\mathbf{x}_i.$ ve $\mathbf{x}_j.$ gözlem çifti arasındaki benzerlik için

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{p}\sum_{k=1}^{p} s_{ijk} \tag{3.8}$$

şeklinde tanımlanan bir skor değeri kullanılır. Burada s_{ijk},

$$s_{ijk} = \begin{cases} 0 & \text{Eğer } \mathbf{x}_i \text{ ve } \mathbf{x}_j \ k. \text{ değişkende eşleşmiyorsa} \\ w & \text{Eğer } \mathbf{x}_i. \text{ ve } \mathbf{x}_j. \ k. \text{ değişkende eşleşiyorsa} \end{cases} \tag{3.9}$$

şeklinde tanımlanır.

3.2.2. Sürekli Değişkenler İçin Uzaklık Ölçüleri

Sürekli değişkenlerden oluşan verilerde gözlemlerin yakınlığı uzaklık ölçüleri ve korelasyon tipli benzerlik ölçüleri ile bulunur. Uzaklık ölçüleri arasında en yaygın olarak kullanılan ve L_2 normu olarak ta bilinen Öklid uzaklığıdır. Her biri p tane sürekli değişken içeren $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlem çifti arasındaki Öklid uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{\frac{1}{2}} \right)^2 \tag{3.10}$$

ile verilir. Öklid uzaklığı kullanılarak oluşturulan kümelerin yapısı değişken değerlerine dönüşüm uygulansa da özellik uzayı içerisinde değişmezdir (invariant) (Duda ve ark. 2001).

Öklid uzaklığı L_r normu ya da Minkowski uzaklığı olarak adlandırılan bir metrik ailesinin özel bir durumudur. Her biri p tane sürekli değişken içeren $i.$ ve $j.$ gözlem çifti arasındaki Minkowski uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{\frac{1}{r}} \right)^r \tag{3.11}$$

ile verilir. Minkowski uzaklığında r'nin aldığı bazı değerlere göre farklı uzaklıklar tanımlanmıştır. Minkowski uzaklığında $r = 2$ alınması durumunda (3.10) eşitliğinde tanımlanan ile Öklid uzaklığı elde edilir. Minkowski uzaklığında $r = 1$ alınmasıyla city-block uzaklığı ya da Manhattan uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \tag{3.12}$$

ve $r \to \infty$ için L_∞ normu

$$d(\mathbf{x}_i, \mathbf{x}_j) = \max_{1 \le k \le p} \left| x_{ik} - x_{jk} \right| \tag{3.13}$$

şeklinde elde edilir.

Aynı zamanda bir metrik olan karesel Mahalanobis uzaklığı da sürekli değişkenler arasındaki yakınlığın bulunmasında kullanılabilir. Her biri p boyutlu $i.$ ve $j.$ gözlem vektörleri arasındaki karesel Mahalanobis uzaklığı,

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{3.14}$$

ile ifade edilir. Burada $\mathbf{S}$ örneklem yada küme içi kovaryans matrisidir. Mahalanobis uzaklığı hiper-elips yapıda kümeler oluşturur. Bu yapı tekil olmayan lineer dönüşümler altında değişmezdir (Duda ve ark. 2001). Mahalanobis uzaklığının hesaplanmasında $\mathbf{S}$ örneklem kovaryans matrisinin tersinin bulunmasında problemlerle karşılaşılabilir. Verideki değişkenler ilişkili değilse $\mathbf{S}$ bir birim matris yapısına sahip olur ve Mahalanobis uzaklığı karesel Öklid uzaklığına eşdeğer olur (Mao ve Jain 1996).

Küme yapısında simetri kabulüne dayalı olarak noktasal simetri uzaklığı

$$d(\mathbf{x}_i, \mathbf{x}_r) = \min_{j=1,\ldots,n \text{ ve } i \neq j} \frac{\left\|(\mathbf{x}_i - \mathbf{x}_r) + (\mathbf{x}_j - \mathbf{x}_r)\right\|}{\left\|(\mathbf{x}_i - \mathbf{x}_r)\right\| + \left\|(\mathbf{x}_j - \mathbf{x}_r)\right\|} \tag{3.15}$$

şeklinde tanımlanır. Burada $\mathbf{x}_r$ bir referans noktasıdır ve $\|\cdot\|$ öklidyen normunu ifade eder. Noktasal simetri uzaklığı, $\mathbf{x}_i$ gözlemi ile $\mathbf{x}_r$ referans noktası arasındaki uzaklığı diğer $n-1$ gözlemi de dikkate alarak hesaplar.

Gözlemler arasındaki uzaklık Pearson Korelasyon katsayısı gibi bir korelasyon katsayısından da hesaplanabilir (Cliff ve ark. 1995). Pearson korelasyon katsayısı,

$$r_{ij} = \frac{\sum_{l=1}^{p}(x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\left(\sum_{l=1}^{p}(x_{il} - \bar{x}_i)^2 \sum_{l=1}^{p}(x_{jl} - \bar{x}_j)^2\right)^{\frac{1}{2}}} \tag{3.16}$$

şeklinde ifade edilir. Burada $\bar{x}_i$, i. gözlem üzerinden ölçülen tüm p değişken değerinin ortalaması olup,

$$\bar{x}_i = \frac{1}{p}\sum_{l=1}^{p}\mathbf{x}_{il} \tag{3.17}$$

şeklinde ifade edilir. Korelasyon katsayısı $[-1,+1]$ değişim aralığında bir değer alır. Korelasyon katsayısı -1 değerine yakınlaştıkça iki gözlemin değişkenleri arasında negatif yönde kuvvetli bir ilişki, $+1$ değerine yakınlaştıkça pozitif yönde kuvvetli bir ilişki olduğunu anlamına gelir. Korelasyon katsayısı kullanılarak iki gözlem vektörü arasındaki [0,l] değişim aralığında değerler alan uzaklık ölçüsü,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{(1 - r_{ij})}{2} \tag{3.18}$$

şeklinde tanımlanır.

Sürekli değişkenler içeren bir gözlem çifti doğrudan benzerlik ölçüleri ile karşılaştırılabilir. Böyle bir benzerlik ölçüsü

$$s(\mathbf{x}_i, \mathbf{x}_j) = \cos\alpha = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\left\|\mathbf{x}_i\right\| + \left\|\mathbf{x}_j\right\|} \tag{3.19}$$

olarak ifade edilen kosinüs benzerliğidir. Kosinüs benzerliği normalleştirilmiş iç çarpım olarak da düşünülebilir. Birbirine daha çok benzeyen iki gözlem, değişken uzayında birbirine daha paralel olacağından kosinüs değeri daha yüksek olacaktır. Kosinüs benzerliğine,

$$d(\mathbf{x}_i, \mathbf{x}_j) = 1 - s(\mathbf{x}_i, \mathbf{x}_j) \tag{3.20}$$

dönüşümü uygulandığında bir uzaklık ölçüsü elde edilir. Kosinüs benzerliği iki gözlemin farkının büyüklüğü üzerinden bir bilgi vermez.

3.2.3. Sürekli ve Kategorik Değişkenler İçeren Gözlem Çiftleri İçin Benzerlik Ölçüleri

Bazı değişkenleri sürekli, bazı değişkenleri kategorik olan verideki gözlem çiftleri arasındaki benzerliğin bulunması için çeşitli yaklaşımlar önerilmiştir. Bunlardan biri tüm değişkenlerin ikili değerlere dönüştürülmesi ve iki sonuçlu değişkenler içeren gözlem çiftleri için önerilen bir benzerlik ölçüsünü kullanmaktır. Başka bir yaklaşım ise her değişken tipine uygun uzaklıklar hesaplanır ve bunlar birleştirilir. Daha karmaşık bir yaklaşım, Gower (1971) tarafından önerilmiştir. Gower kategorik değişkenler için genel benzerlik ölçüsünü,

$$s_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}} \tag{3.21}$$

biçiminde ifade etmiştir. Burada s_{ijk}, k. değişken değerine göre i. ve j. gözlemler arasındaki benzerlik ölçüsüdür. w_{ijk} ise i. ve j. gözlem k. değişkene göre karşılaştırmasında değişken değeri bulunmuyorsa 0 diğer durumlarda 1 değerini almaktadır. Gower (1971), verideki sürekli değişkenler için benzerlik ölçüsünü,

$$s_{ij} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k} \tag{3.22}$$

biçiminde tanımlamıştır. Burada R_k, i. ve j. gözlemin k. değişken değerlerinin değişim aralığı (range) olarak tanımlanır. Gower (1971) katsayısının kullanımına ilişkin bir örnek aşağıda verilmiştir.

Örnek 3.1: Bahçe çiçeği verisi (Kaufman ve Rousseeuw 1990) 18 bahçe çiçeğinin yedi özelliğine ilişkin değerler içerir. Çiçekler için ilgilenilen bu özellikler;

1: Bitki donduğunda bahçede bırakıldı mı? (1=evet, 0=hayır)

2. Bitki gölge altında mı? (1=evet, 0=hayır)

3. Bitki yumrulumu? (1=evet, 0=hayır)

4. Bitkinin çiçek rengi nedir? (beyaz=1, sarı=2, pembe=3, kırmızı=4, mavi=5)

5. Bitkinin ekildiği toprak türü nedir? (kuru=1, normal=2, nemli=3)

6. Bitkinin boyu kaç cm dir?

7. Bitkiler arası ideal uzaklık kaç cm dir?

Bu özelliklerde sadece bitki türü yumrulu mu sorusunda incelenen iki bahçe çiçeği de yumrulu değil ise 0 diğer tüm durumlarda 1 ile ağırlıklandırılmıştır.

Bahçe çiçeği verisi (Kaufman ve Rousseeuw 1990) için Gower(1971) tarafından önerilen benzerlik matrisi elde etmek amacıyla matlab kodu Şekil 3.2'de verilmiştir.

Tablo 3.3. Bahçe çiçeği verisi (Kaufman ve Rousseeuw 1990).

	Değişkenler						
	Kategorik					**Sürekli**	
Bitki Türleri	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Begonya	0	1	1	4	3	25	15
2. Katırtırnağı	1	0	0	2	1	150	50
3. Kamelya	0	1	0	3	3	150	50
4. Yıldız çiçeği	0	0	1	4	2	125	50
5. Myosotis sylvatica	0	1	0	5	2	20	15
6. Küpe çiçeği	0	1	0	4	3	50	40
7. Sardunya	0	0	0	4	3	40	20
8. Kılıç çiçeği	0	0	1	2	2	100	15
9. Süpürge otu	1	1	0	3	1	25	15
10. Hortensis	1	1	0	5	2	100	60
11. Süsen çiçeği	1	1	1	5	3	45	10
12. Zambak	1	1	1	1	2	90	25
13. Vadi zambağı	1	1	0	1	2	20	10
14. Şakayık	1	1	1	4	2	80	30
15. Karanfil	1	0	0	3	2	40	20
16. Gül	1	0	0	4	2	200	60
17. İskoç gülü	1	0	0	2	2	150	60
18. Lale	0	0	1	2	1	25	10

Tablo 3.4. Bahçe çiçeği verisinin Gower (1971) tarafından önerilen benzerlik ölçüsüne göre ilk 10 bitki türü için elde edilen benzerlik matrisi.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1.00									
2	0.09	1.00								
3	0.52	0.33	1.00							
4	0.53	0.41	0.41	1.00						
5	0.57	0.10	0.43	0.39	1.00					
6	0.77	0.21	0.71	0.48	0.56	1.00				
7	0.69	0.30	0.46	0.56	0.46	0.76	1.00			
8	0.51	0.43	0.29	0.74	0.51	0.32	0.51	1.00		
9	0.43	0.43	0.43	0.11	0.50	0.39	0.30	0.23	1.00	
10	0.24	0.42	0.42	0.38	0.61	0.39	0.14	0.30	0.45	1.00

```
load bahçebitkisiverisi;
X=bahçebitkisiverisi;
for i=1:18;
   for j=1:18;
      for k=1:7;
         if X(i,k)==X(j,k)
            w(i,j,1:2)=[1,1];
            s(i,j,k)=1;
         else
            w(i,j,4:7)=[1,1,1,1];
            s(i,j,k)=0;
            if   (X(i,3)==0) && ( X(j,3)==0)
               w(i,j,3)=0;
            else
               w(i,j,3)=1;
           end
         end
      end
   end
end
for i=1:18;
   for j=1:18;
      for k=6:7;
         s(i,j,k)=1-(abs(X(i,k)-X(j,k)))/range(X(:,k));
      end
   end
end
```

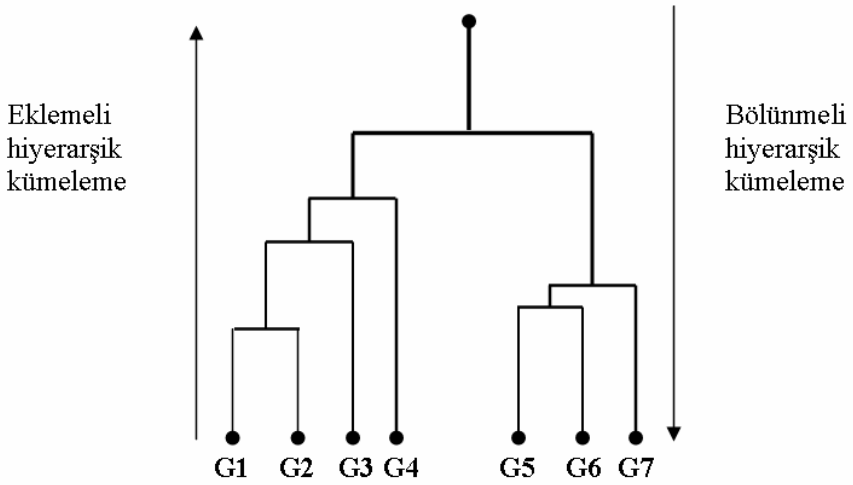
Şekil 3.2. Bahçe çiçeği verisi için Gower (1971) tarafından önerilen benzerlik matrisinin elde edilmesi amacıyla matlab kodu.

3.3. Hiyerarşik Kümeleme Yöntemleri

Hiyerarşik kümelemede, verideki gözlemler iç içe parçalanmaların veya kaynaştırmaların bir dizisi şeklinde gruplanır. Hiyerarşik kümeleme yöntemleri etkin bir hesaplama tekniği kullanarak verideki iyi tanımlanmış kümeleri bulmaya çalışır. n sayıda gözlemden oluşan bir veri g farklı küme yapısı içeriyorsa, n gözlemin g farklı kümeye parçalanmalarının sayısı,

$$N(n,g) = \frac{1}{g!}\sum_{k=1}^{g}\binom{g}{k}(-1)^{g-k}k^{n} \tag{3.23}$$

ile verilir (Duran ve Odel 1974, Jensen 1969, Seber 1984). Bu değer $\frac{g^{n}}{g!}$ ile yaklaşık olarak da bulunabilir. Örneğin, 10 gözlem ve 3 farklı kümeye 9330 farklı şekilde parçalanır. Bu değer oldukça büyüktür. Hiyerarşik kümeleme yöntemleri, olanaklı tüm sonuçların incelenmesine gerek kalmaksızın mantıklı sonuçların bulunmasını sağlar. Hiyerarşik kümeleme algoritmaları, ardışık bir işlem süreci içerir. Hiyerarşik kümeleme yöntemlerinin sonuçları, dendrogram olarak bilinen bir ağaç diyagramı ile grafiksel olarak gösterilebilir. Bu grafikte birleştirilen küme çiftleri ve aralarındaki uzaklık görülebilir.



Şekil 3.3. Hiyerarşik kümelemenin ağaç diyagramı ile grafiksel gösterimi.

Gözlemler kümelere iki farklı şekilde hiyerarşik olarak gruplanır. Bu hiyerarşik kümeleme yöntemleri yığılmalı (agglomerative) ya da kaynaştırmalı hiyerarşik kümeleme ve bölünmeli ya da parçalanmalı veya paylaştırmalı hiyerarşik kümelemedir.

3.3.1. Yığılmalı Hiyerarşik Kümeleme Yöntemi

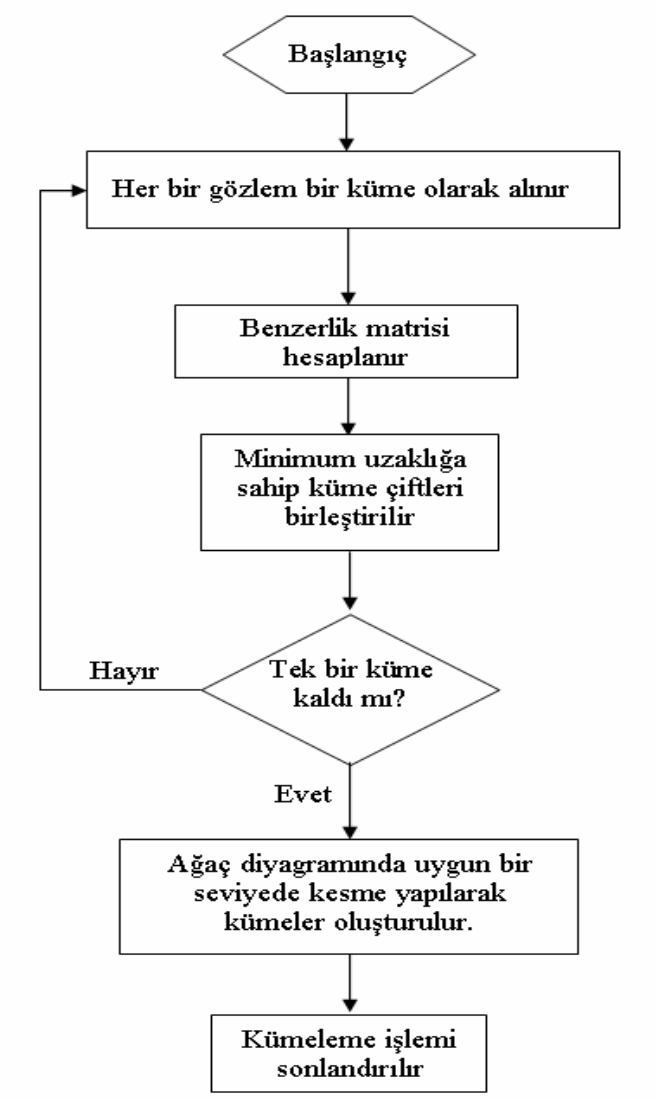
Yığılmalı hiyerarşik kümeleme yönteminde verideki n sayıda gözlemin her biri bir küme olarak düşünülür. Bu kümeler bir seri birleştirme işlemleri ile bu kümeler tek bir küme oluşturuncaya kadar uygulanır. Bu yönteme yığılmalı (agglomerative) ya da kaynaştırmalı hiyerarşik kümeleme denir (Hansen ve ark. 1997). Genel bir yığılmalı hiyerarşik kümeleme yönteminin adımları aşağıda verilmiştir.

1. Küme sayısı n olarak alınır. $G_1, G_2, \ldots, G_n$ ile ifade edilen bu n sayıda küme için benzerlik matrisi hesaplanır.
2. Benzerlik matrisinde n küme sayısı, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,n$ ve $i \neq j$ olmak üzere tüm $\min D(G_i, G_j)$ uzaklığına sahip iki küme belirlenir. Bu iki küme birleştirilerek yeni bir küme oluşturulur.
3. Benzerlik matrisi güncellenir. Diğer bir ifade ile yeni oluşan küme de dikkate alınarak benzerlik matrisi hesaplanır.
4. Tek bir küme yapısı elde edilinceye kadar 2. ve 3. adımlar tekrarlanır.

Yığılmalı hiyerarşik kümeleme de küme çiftlerinin birleştirilerek yeni bir kümenin elde edilmesi ve daha sonra elde edilen bu küme ile diğer bir kümenin birleştirilerek yine bir küme elde edilmesi kümeler arasında benzerliğin hesaplanmasında tanımlanan uzaklık fonksiyonuna bağlıdır. $i = 1,2,\ldots,n$, $j = 1,2,\ldots,n$ ve $i \neq j$ olmak üzere G_i ve G_j gibi iki küme arasındaki benzerliğin bulunmasında kullanılan oldukça fazla uzaklık tanımı bulunmaktadır (Jain ve Dubes 1988). Lance ve Williams (1967), G_i, G_j ve G_l üç küme olmak üzere kümeler arası benzerliğin bulunmasında genel yinelenmeli bir formülü,

$$D\left(G_l,(G_i,G_j)\right)=\beta_i D(G_l,G_i)+\beta_j D(G_l,G_j)+\eta D(G_i,G_j)$$
$$+\gamma\left|D(G_l,G_i)-D(G_l,G_j)\right| \quad (3.24)$$

şeklinde önermiştir. Burada $D(.,.)$ gözlemler arası benzerliği bulmada kullanılan bir uzaklık fonksiyonu β_i, β_j, η ve γ katsayıları ise seçilen yığılmalı hiyerarşik kümeleme tekniğine göre değerler alan parametreleri ifade eder.



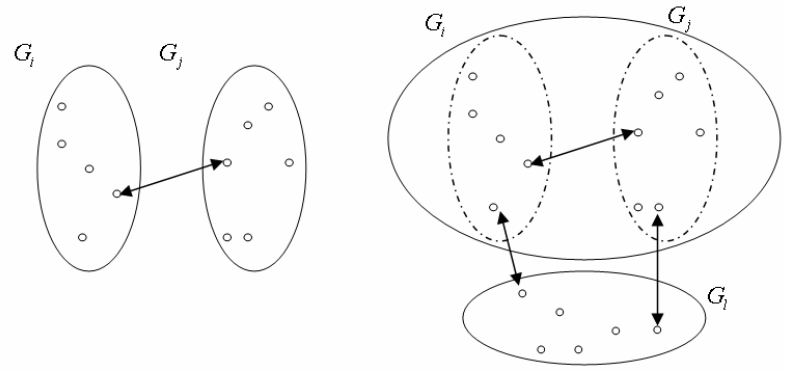
Şekil 3.4. Yığılmalı hiyerarşik kümeleme yönteminin akış çizelgesi.

Tablo 3.5. Lance ve Williams (1967) tarafından önerilen (3.24)'deki eşitlik ile verilen kümeler arası benzerlik için genel yinelenmeli formülündeki β_i, β_j, η ve γ parametrelerinin kullanılan tekniğe göre aldığı değerler. n_i, n_j, ve n_l sırasıyla G_i, G_j ve G_l kümelerindeki gözlem sayısıdır.

Küme Bağlantı Tekniği	β_i	β_j	η	γ
Tek Bağlantı	0.5	0.5	0	-0.5
Tam Bağlantı	0.5	0.5	0	0.5
Ortalama Bağlantı	$\frac{n_i}{n_i+n_j}$	$\frac{n_i}{n_i+n_j}$	0	0
Ağırlıklandırılmış Ortalama Bağlantı	0.5	0.5	0	0
Medyan Bağlantı	0.5	0.5	-0.25	0
Küme Merkezleri Bağlantı	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
Ward Yöntemi	$\frac{n_i+n_l}{n_i+n_j+n_l}$	$\frac{n_j+n_l}{n_i+n_j+n_l}$	$\frac{-n_l}{n_i+n_j+n_l}$	0

3.3.1.1. Tek Bağlantı Algoritması

Tek bağlantı iki farklı kümedeki en yakın gözlemlere göre belirlenen iki küme arasındaki uzaklıktır. Tek bağlantı kümeleme aynı zamanda en yakın komşuluk yöntemi olarak da adlandırılır (Jain ve Dubes 1988).



Şekil 3.5. Tek bağlantı kümeleme yöntemine göre iki küme arasındaki uzaklığın belirlenmesi.

Tablo 3.5 ile verilen parametrelerin tek bağlantı tekniğinde aldığı değerleri, (3.24)'deki eşitlikte yerine konulursa iki küme arasındaki uzaklık,

$$D\left(G_l,(G_i,G_j)\right)=\min\left(D(G_l,G_i),D(G_l,G_j)\right) \tag{3.25}$$

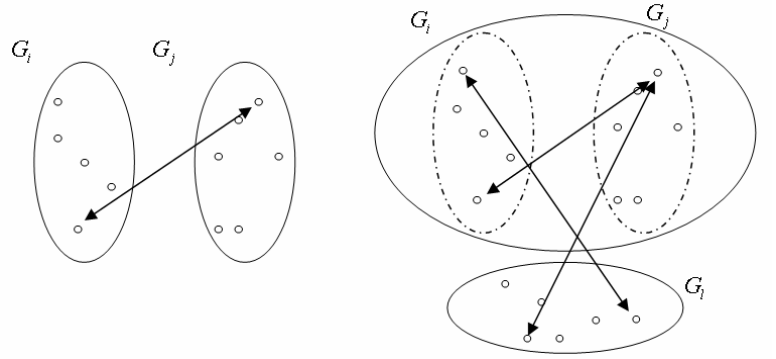
şeklinde elde edilir. Tek bağlantı kümeleme yöntemi gittikçe uzayan kümeler oluşturma eğilimindedir (Everitt ve ark. 2001). Bu yöntem sapan aykırı değerlerden oldukça fazla etkilenir. Bu sebepten birbirinden oldukça farklı özellikleri bulunan iki küme birleştirilebilir. Kümeler birbirinden oldukça ayrık iseler tek bağlantı yöntemi iyi sonuçlar verir (Everitt ve ark. 2001).

3.3.1.2. Tam Bağlantı Algoritması

Tam bağlantı yönteminde, iki farklı kümedeki birbirine en uzak gözlem çifti arasındaki uzaklık değeri iki küme arasındaki uzaklık olarak alınır. Tablo 3.5 ile verilen parametrelerin tek bağlantı tekniğinde aldığı değerleri, (3.24)'deki eşitlikte yerine konulursa iki küme arasındaki uzaklık,

$$D\left(G_l,(G_i,G_j)\right)=maksimum\left(D(G_l,G_i),D(G_l,G_j)\right) \tag{3.26}$$

şeklinde elde edilir. Tam bağlantı kümeleme özellikle belirgin olmayan, küçük ve yoğun küme yapısına sahip verilerde etkin bir yöntemdir.



Şekil 3.6. Tam bağlantı kümeleme yönteminde iki küme arasındaki uzaklığın belirlenmesi.

3.3.1.3. Ortalama Bağlantı Algoritması

Ortalama bağlantı yönteminde ayrı gruplarda yer alan gözlem çiftleri arasındaki ortalama uzaklık iki küme arasındaki uzaklık olarak alınır (Everitt ve ark. 2001). Bu yöntem aynı zamanda ağırlıksız grup çiftleri yöntemi olarak da adlandırılır (Jain ve Dubes 1988). Ortalama bağlantı tekniğindeki parametrelerin Tablo 3.5 ile verilen değerleri, (3.24)'deki eşitlikte yerine konulursa ortalama bağlantı tekniğine göre iki küme arasındaki uzaklık,

$$D\left(G_l,(G_i,G_j)\right)=\frac{1}{2}\left(D(G_l,G_i)+\beta_j D(G_l,G_j)\right) \tag{3.27}$$

şeklinde ifade edilir.

3.3.1.4. Ağırlıklı Ortalama Bağlantı Yöntemi

İki küme arasındaki uzaklığın bulunması için ortalama bağlantı tekniğinde 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 ağırlıklandırılır. Ağırlıklı ortalama bağlantı yöntemindeki parametrelerin Tablo 3.5 ile verilen değerleri, (3.24)'deki eşitlikte yerine konulursa iki küme arasındaki uzaklık,

$$D\left(G_l,(G_i,G_j)\right)=\frac{n_i}{n_i+n_j}D(G_l,G_i)+\frac{n_j}{n_i+n_j}D(G_l,G_j) \tag{3.28}$$

şeklinde ifade edilir.

3.3.1.5. Merkezi Bağlantı Yöntemi

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,

$$m_i = \frac{1}{n_i} \sum_{\mathbf{x} \in G_i} \mathbf{x} \tag{3.29}$$

ile bulunur. Burada n_i, G_i kümesindeki gözlem sayısını ifade eder. Merkezi bağlantı yöntemindeki parametrelerin Tablo 3.5 ile verilen değerleri, (3.24)'deki eşitlikte yerine konulursa iki küme arasındaki uzaklık,

$$D\big(G_l,(G_i,G_j)\big) = \frac{n_i}{n_i+n_j} D(G_l,G_i) + \frac{n_j}{n_i+n_j} D(G_l,G_j) - \frac{n_i n_j}{(n_i+n_j)^2} D(G_i,G_j) \tag{3.30}$$

şeklinde elde ifade edilir. Bu ifade,

$$D\big(G_l,(G_i,G_j)\big) = \left\| m_l - m_{(ij)} \right\|^2 \tag{3.31}$$

şeklinde tanımlanan iki kümenin merkezleri arasındaki karesel Öklid uzaklığı ile eşdeğerdir.

3.3.1.6. Medyan Bağlantı Yöntemi

Merkezi bağlantı yönteminden farklı olarak, medyan bağlantı yönteminde iki küme arasındaki uzaklık, iki kümenin merkezleri arasındaki uzaklığın eşit ağırlıklı olarak hesaplanmasıyla elde edilir (Gower 1967). Medyan bağlantı yöntemindeki parametrelerin Tablo 3.5 ile verilen değerleri, (3.24)'deki eşitlikte yerine konulursa iki küme arasındaki uzaklık,

$$D\left(G_l,(G_i,G_j)\right)=\frac{1}{2}D(G_l,G_i)+\frac{1}{2}D(G_l,G_j)-\frac{1}{4}D(G_i,G_j) \tag{3.32}$$

şeklinde ifade edilir.

3.3.1.7. Ward Yöntemi

Artırımlı (incremental) hata kareler toplamı yöntemi olarak da adlandırılan Ward'ın hiyerarşik kümeleme yönteminde amaç artan sınıf içi hata kareler toplamının minimum yapılmasıdır (Everitt ve ark. 2001, Ward 1963). Sınıf içi hata kareleri toplamı,

$$E=\sum_{k=1}^{K}\sum_{\mathbf{x}_i\in G_k}\left\|\mathbf{x}_i-\mathbf{m}_k\right\|^2 \tag{3.33}$$

ile ifade edilir. Burada K, küme sayısını ve $k=1,2,\ldots,K$ olmak üzere m_k, eşitlik 3.29 ile tanımlanan $k.$ küme ortalamasını gösterir. İki kümenin birleşimi ile oluşan küme ile diğer kümeler arasındaki hata kareler farkı,

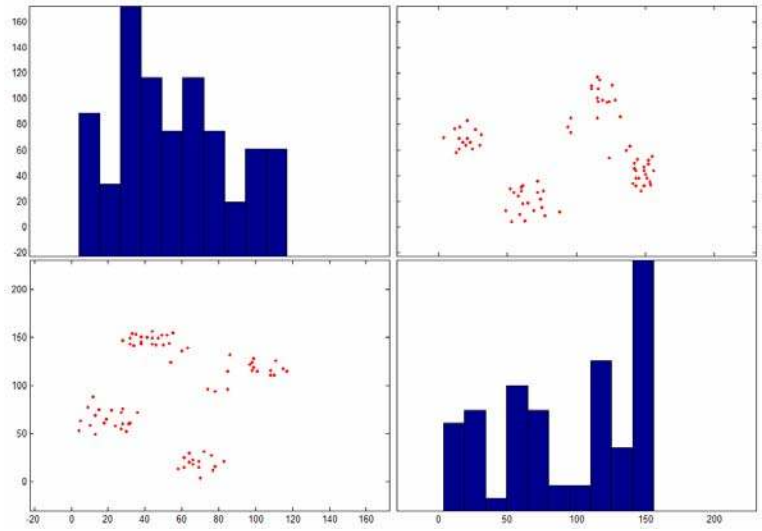
$$\Delta E_{ij}=\frac{n_i n_j}{n_i+n_j}\left\|\mathbf{m}_i-\mathbf{m}_j\right\|^2 \tag{3.34}$$

şeklinde hesaplanır. Ward (1963) tarafından önerilen hiyerarşik kümeleme yöntemdeki parametrelerin Tablo 3.5 ile verilen değerleri, (3.24)'deki eşitlikte yerine konulursa iki küme arasındaki uzaklık,

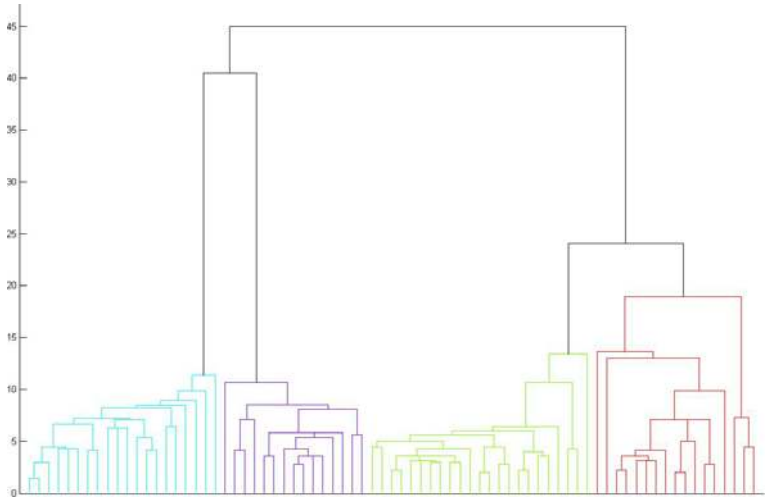
$$D\left(G_l,(G_i,G_j)\right)=\frac{n_i+n_j}{n_i+n_j+n_l}D(G_l,G_i)+\frac{n_j+n_l}{n_i+n_j+n_l}D(G_l,G_j)$$
$$-\frac{n_l}{(n_i+n_j)^2}D(G_i,G_j) \tag{3.35}$$

şeklinde ifade edilir.

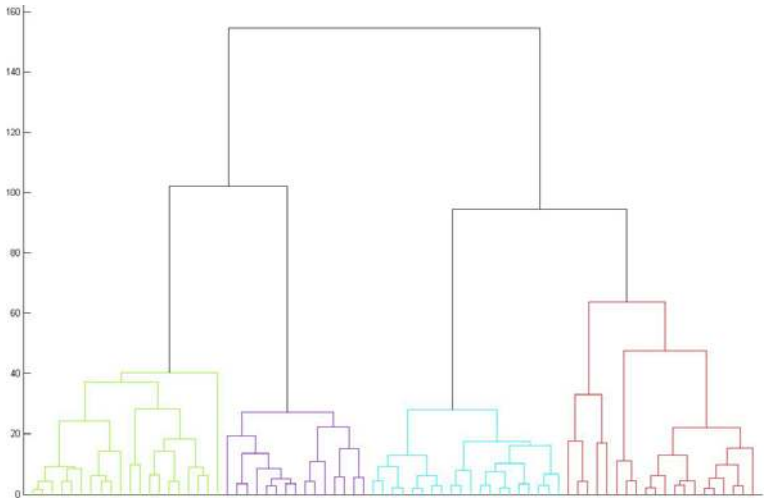
Örnek 3.2: Kümeleme analiz yöntemlerinin açıklanmasında oldukça yaygın olarak kullanılan Ruspini verisi (Ruspini 1970) incelenecektir. Ruspini verisi 75 gözlem üzerinden gözlenmiş iki değişkenin değerlerini içerir. Bu veride doğal olarak 4 grup bulunmaktadır. Ruspini verisine ait saçılım grafiği Şekil 3.7 ile verilmiştir. Ruspini verisine uygulanan yığılmalı hiyerarşik kümeleme yöntemlerine ait ağaç grafiği Şekil 3.8 ile verilmiştir. Yığılmalı hiyerarşik kümeleme yöntemleri için Öklid uzaklığı kullanılmıştır.



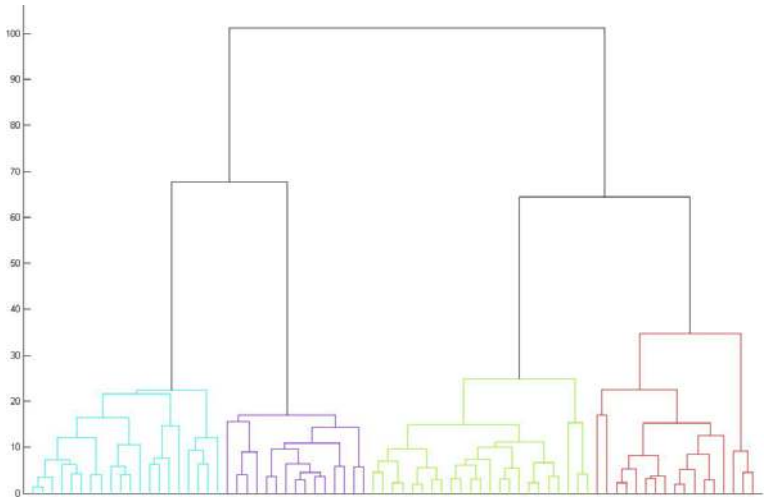
Şekil 3.7. Ruspini (Ruspini 1970) verisinin saçılım grafiği.



Şekil 3.8. Ruspini (Ruspini 1970) verisinin tek bağlantı kümeleme sonucu için ağaç grafiği ya da dendrogramı.

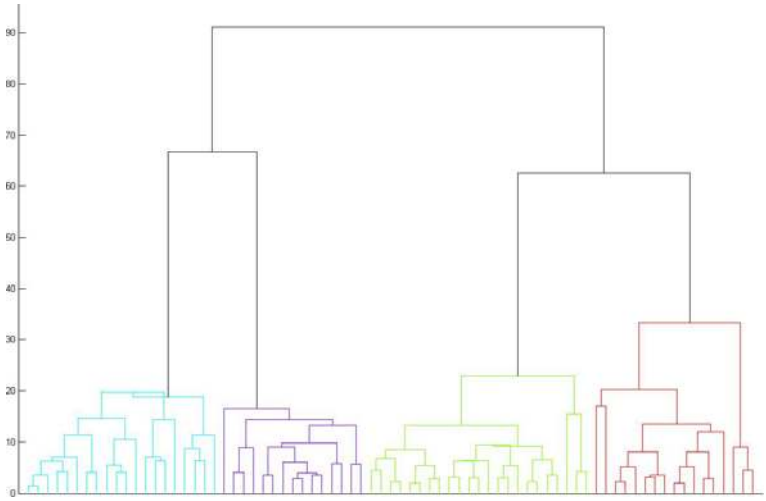


Şekil 3.9. Ruspini (Ruspini 1970) verisinin tam bağlantı kümeleme sonucu için ağaç grafiği ya da dendrogramı.

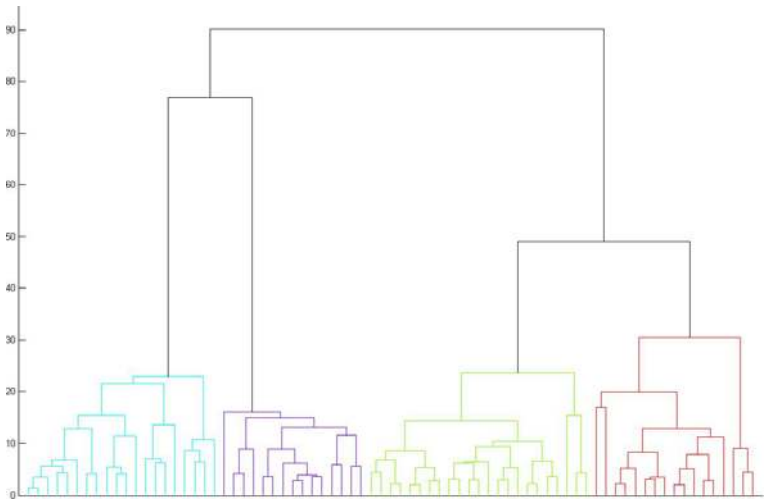


Şekil 3.10. Ruspini (Ruspini 1970) verisinin ortalama bağlantı kümeleme sonucu için ağaç grafiği ya da dendrogramı.

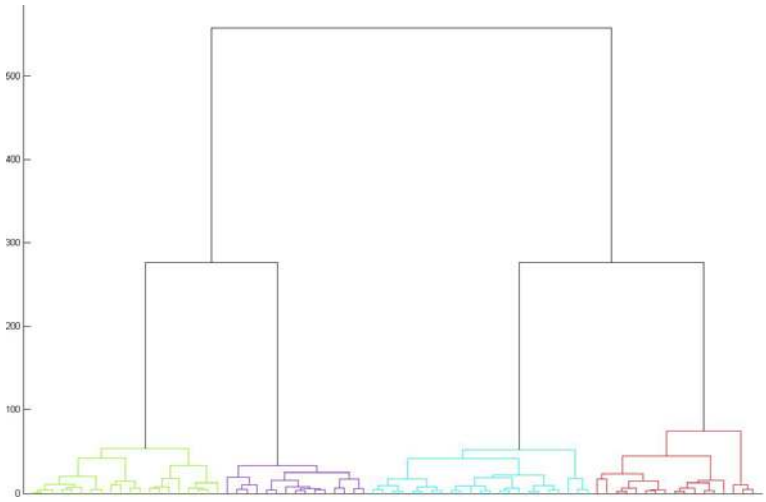
Şekil 3.11. Ruspini (Ruspini 1970) verisinin ağırlıklı ortalama kümeleme sonucu için ağaç grafiği ya da dendgramı.



Şekil 3.12. Ruspini (Ruspini 1970) verisinin merkezi bağlantı kümeleme sonucu için ağaç grafiği ya da dendgramı.



Şekil 3.13. Ruspini (Ruspini 1970) verisinin medyan bağlantı kümeleme sonucu için ağaç grafiği ya da dendgramı.

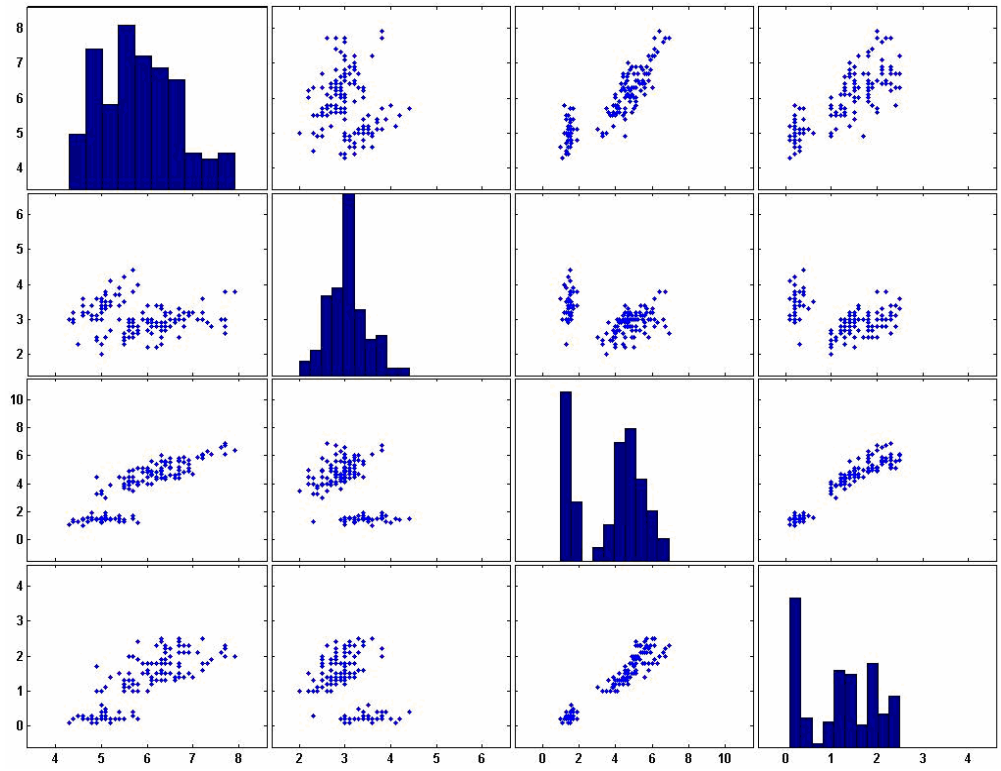


Şekil 3.14. Ruspini (Ruspini 1970) verisinin Ward kümeleme sonucu için ağaç grafiği ya da dendgramı.

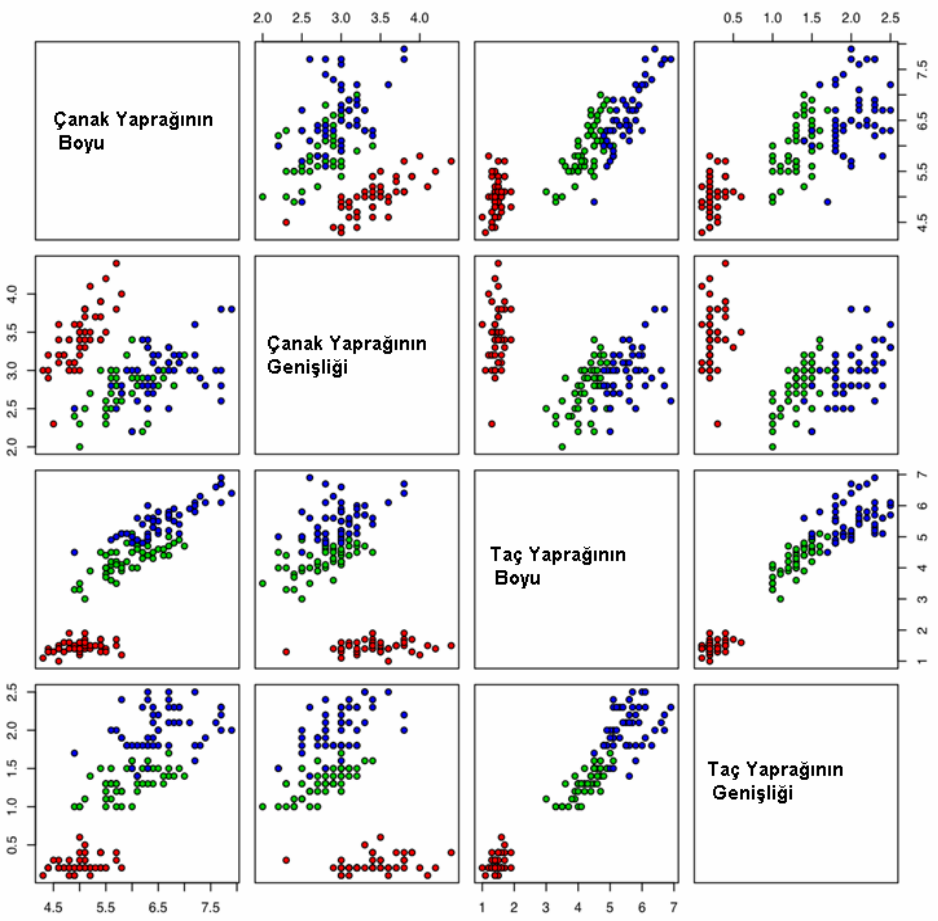
Örnek 3.3: Bu örnek üzerinde hiyerarşik kümelemede seçilen kümeleme yönteminin ve uzaklık ölçüsünün kümelemeye olan etkisi gösterilecektir. Bu amaçla Iris (süsen) çiçeği verisi (Fisher 1936) kullanılacaktır. Süsen çiçeği verisi 150 gözlem üzerinden gözlenen dört değişken değerlerini içerir. Bu değişkenler çanak ve taç yapraklarının uzunluk ve genişliğidir. Süsen çiçeği verisinde Setosa, Versicolor ve Virginica olarak adlandırılan 3 süsen çiçeği sınıfı vardır. Her bir sınıfta eşit sayıda gözlem bulunur.

Iris verisine yığılmalı hiyerarşik kümeleme de en çok kullanılan yöntemler uygulandığında elde edilen genel sınıflandırma hata oranları Tablo 3.6 ile verilmiştir.

Tablo 3.6 incelendiğinde kümelemede kullanılan yöntemlerde seçilen uzaklık ölçüsü kümeleme sonuçlarına etki etmektedir. Tablo 3.6 ile verilen kümeleme yöntemlerinin genel sınıflandırma hata oran değerleri Iris çiçeğinde üç küme olduğu bilindiği kabulü altında elde edilmiştir.



Şekil 3.15. Iris (süsen) çiçeği verisinin saçılım grafiği.



Şekil 3.16. Iris (süsen) çiçeği verisinin türlerine göre saçılım grafiği. (Kırmızı: Setosa, Yeşil: Versicolor, Mavi: Virginica).

Tablo 3.6. Iris (süsen) çiçeği verisine uygulanan yığılmalı hiyerarşik kümeleme sonuçları genel hata sınıflandırma yüzdeleri.

	Öklid Uzaklığı	Cityblok Uzaklığı	Mahalanobis Uzaklığı	Pearson Korelasyon	Kosinüs Uzaklığı
Tek Bağlantı	%33,3	%33,3	%66	%33,3	%33,3
Tam Bağlantı	%15,3	%10	%48	%14,6	%16
Ortalama Bağlantı	%9,3	%10	%65,3	%4,6	%33,3
Ağırlıklı Ortalama	%10	%4,6	%32,6	%31	%4
Merkezi Bağlantı	%6	%64	%66	%5,3	%34
Medyan Bağlantı	%24,6	%24,6	%42	%32	%19,3
Ward Yöntemi	%10,6	%11,3	%17,3	%11,3	%2

3.3.2. Hiyerarşik Bölünmeli (Divisive) Kümeleme Yöntemleri

Hiyerarşik bölünmeli kümeleme yöntemleri, yığılmalı hiyerarşik kümelemenin tam tersi yolla yapılmasıdır. Bölünmeli hiyerarşik kümeleme yönteminde başlangıçta verinin n sayıda gözlemi tek bir küme içerisinde yer alır. Bir seri küme bölme işlemleri uygulanmasıyla n birimlik bu küme her bir gözlem farklı kümede olacak şekilde n sayıda kümeye bölünür (Gordon 1999). Bölünmeli bir hiyerarşik kümeleme algoritması n sayıda gözlemden oluşan bir veriye uygulandığında daha ilk adımda, $2^{n-1}-1$ sayıda mümkün yolla boş olmayan iki alt kümeye bölünebilir. Bölünmeli hiyerarşik kümeleme algoritmaları hesaplama maliyetleri oldukça yüksek olduğu için pek tercih edilmez. Bununla birlikte verinin ana yapısı bu algoritmanın ilk adımlarında ortaya çıkmaktadır.

Bölünmeli hiyerarşik kümeleme genel olarak iki kategoriye ayrılır. Bunlar tek etkili (monothetic) ve eş etkili (polythetic) bölünmeli hiyerarşik kümeleme yöntemleridir (Everitt 1993, Willett 1988). Tek etkili hiyerarşik kümelemede bir küme, iki alt kümeye yalnız tek bir değişken dikkate alınarak bölünürken, eş etkili hiyerarşik kümelemede ise bölünme, tüm p değişkeninde dikkate alınmasıyla yapılır.

İki sonuçlu değişken değerleri içeren verilerde tek etkili yaklaşım daha kolay uygulanabilir. Bu tür verilerde iki gruba bölünme bir özelliğin bulunması veya bulunmamasına dayalı olarak yapılır. Değişken (özellik) bir ki-kare istatistiğini veya bir enformasyon istatistiğini maksimize edecek şekilde seçilir (Everitt 1993, Gordon 1999).

Eş etkili hiyerarşik kümeleme için bir yöntem, MacNaughton-Smith ve ark. (1964) tarafından önerilmiştir. Bu yöntemde bir parçalama (splinter) grubu oluşturulur. Bu parçalama grubun içerisine belirli bir kritere uygun olarak seçilen bir gözlem ya da gözlemler aktarılır. Parçalama grubundaki bu gözlem ya da gözlemler kök olarak ifade edilir. Bu gözlemlerin dışındaki diğer gözlemler kalan grupta yer alır. Kalan gruptaki her bir gözlemin kalan gruptaki diğer gözlemlere olan ortalama uzaklıkları ve kalan gruptaki her gözlemlerin parçalama grubundaki gözleme veya gözlemlere olan ortalama uzaklık değerleri hesaplanır. Kalan gruptaki en yüksek

ortalama uzaklık değerine sahip gözlemler belirlenir. En yüksek ortalama uzaklık değerine sahip gözlemin kalan gruptaki ortalama uzaklığı ile parçalama grubundaki gözlemlere olan ortalama uzaklığı arasındaki fark pozitif ise bu gözlem parçalama grubuna dahil edilir ve tüm uzaklıklar yeniden hesaplanır. Fark negatif ise kalan grupta bulunmaya devam eder ve kalan gruptaki en yüksek ikinci ortalama uzaklık değerine sahip gözlem incelenir.

Örnek 3.4: Atletizmde sekiz ülkenin sporcularına ait farklı mesafelerdeki ülke rekor süreleri Tablo 3.7. ile verilmiştir (Dawkins 1989). Burada dikkate alınan koşu mesafeleri sırasıyla 1-100 m(s), 2-200 m(s), 3-400 m(s), 4-800 m(dk), 5-1500 m(dk), 6-5000 m(dk), 7-10000 m(dk), 8-maraton'dur. Bu sekiz ülke atletlerinin rekor sürelerine göre ülkeler gruplandırılmak isteniyor. Sekiz ülkenin koşu sürelerine göre birbirlerine olan uzaklıkla Tablo 3.7'de verilmektedir. Tablo 3.7'deki değerler incelendiğinde ABD diğer yedi ülkeye göre daha fazla ortalama uzaklığına sahiptir. Bu sonuçtan dolayı parçalama grubun ilk elemanı olarak ABD alınır. Bundan sonraki adımda kalan gruptaki yedi ülkenin birbirlerine olan ortalama uzaklıkları ve parçalama grubundaki gözlem arasındaki uzaklıkları hesaplanır. Bu iki ortalama uzaklıklar arasındaki farklara bakıldığında Avustralya'nın pozitif farka sahip olduğu görülür. Avustralya, ABD'nin bulunduğu parçalama grubuna aktarılır.

Tablo 3.7. Sekiz ülkeye ait çeşitli mesafelerde elde edilen atletizm rekorları (Dawkins 1989).

Ülke	1	2	3	4	5	6	7	8
Avustralya	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
Belçika	10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95
Kanada	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
Almanya	10.12	20.33	44.87	1.73	3.56	13.17	27.42	129.92
İngiltere	10.11	20.21	44.93	1.70	3.51	13.01	27.51	129.13
Kenya	10.46	20.66	44.92	1.73	3.55	13.10	27.80	129.75
ABD	9.93	19.75	43.86	1.73	3.53	13.20	27.53	128.22
Rusya	10.07	20.00	44.60	1.75	3.59	13.20	27.53	130.55

Tablo 3.10'dan görüleceği gibi kalan gruptaki ülkelerin ortalama uzaklıkları ile parçalama grubundaki ülkelere olan ortalama uzaklıklarının farkı negatiftir. Bu

sebepten parçalama grubunun güncellenmesi burada sonlandırılır. Sonuç olarak; Grup A={ABD, Avustralya} ve Grup B={Belçika, Kanada, Almanya, İngiltere, Kenya, Rusya} şeklinde 8 ülke iki gruba bölünür.

Tablo 3.8. Sekiz ülkenin her birinin diğer ülkelere olan ortalama uzaklıkları.

Ülke	Diğer ülkelere olan ortalama Uzaklık
Avustralya	1.6403
Belçika	1.1690
Kanada	1.5911
Almanya	1.0834
İngiltere	1.1639
Kenya	1.1538
ABD	2.0602
Rusya	1.5130

Tablo 3.9. Kalan grupta yer alan yedi ülkenin birbirlerine ve parçalama grubundaki ülkeye olan ortalama uzaklıkları.

Ülke	Kalanlar arası ortalama uzaklık (1)	Splinter gruba olan ortalama uzaklık (2)	Fark (1)-(2)
Avustralya	1.7287	1.1101	0.6186
Belçika	0.9755	2.3302	-1.3547
Kanada	1.3919	2.7862	-1.3943
Almanya	0.9185	2.0728	-1.1543
İngiltere	1.1076	1.5016	-0.3940
Kenya	0.9865	2.1580	-1.1715
Rusya	1.3547	2.4622	-1.1075

Tablo 3.10. Kalan gruptaki altı ülkenin, birbirlerine ve parçalama grubundaki ülkelere olan ortalama uzaklıkları.

Ülke	Kalan gruptakilerin birbirlerine olan ortalama uzaklıkları (1)	Parçalama gruba olan ortalama uzaklıklar (2)	Fark (1)-(2)
Belçika	0.814	2.036	-1.2220
Kanada	1.239	2.453	-1.2140
Almanya	0.761	1.872	-1.1110
İngiltere	1.138	1.225	-0.0870
Kenya	0.860	1.856	-0.9960
Rusya	1.163	2.386	-1.2230

3.4. Paylaştırmalı (Partitional) Kümeleme Yöntemleri

Paylaştırmalı, taksimli, ayırmalı veya bölmeli yöntemlerde gözlemler benzerlik ya da uzaklık matrislerine dayalı olan bir hiyerarşik yaklaşım kullanılmadan K kümeye ayrıştırılır. Bu yöntemler aynı zamanda "optimizasyon yöntemleri" olarak da adlandırılır. Küme sayısı K daha önceden belirlenmiş olabileceği gibi yöntem sürecinde de hesaplanabilir.

Kümeleme yöntemlerinde amaç n sayıda nesnenin K kümeye parçalanması için tüm mümkün yolların incelenmesi ve belli bir kritere göre en uygun kümelenmenin bulunmasıdır. Yalnız burada sorun parçalanma sayılarının oldukça fazla olmasıdır. (3.23)'deki eşitlik kullanılarak elde edilebilecek parçalanmalarının sayısı hesaplanabilir. Bu nedenden dolayı veride gözlem sayısı, küme sayısı veya her ikisinin de fazla olması durumunda, kümeleme için basit yöntemlere ihtiyaç duyulur.

3.4.1. Kümeleme Kriterleri

Kümeleme analizinde amaç veriyi, aynı küme içinde bulunan gözlemleri homojen farklı kümelerde yer alan gözlemleri de iyi ayrışımlı olacak şekilde kümelere gruplayan bir bölünmeyi araştırmaktır. Gözlemlerin homojenliği ve ayrışımı için birçok kriter önerilmiştir. Paylaştırmalı kümeleme içinde en yaygın olarak kullanılan kriter hata kareleri toplamıdır (Jain ve Dubes 1988).

Hata kareleri toplam kriteri $j = 1,2,\ldots,n$ olmak üzere $\mathbf{x}_j \in \Re^p$ gözlemi $G_1, G_2, \ldots, G_K$ ile ifade edilen K kümeye paylaştırılırsa hata kareleri toplamı,

$$E_s(\boldsymbol{\Gamma}, \mathbf{M}) = \sum_{i=1}^{K}\sum_{j=1}^{n} \gamma_{ij} \left\| \mathbf{x}_j - \mathbf{m}_i \right\|^2 = \sum_{i=1}^{K}\sum_{j=1}^{n} \gamma_{ij} (\mathbf{x}_j - \mathbf{m}_i)^T (\mathbf{x}_j - \mathbf{m}_i) \qquad (3.36)$$

formunda olup burada, $j = 1,2,\ldots,n$ ve $i = 1,2,\ldots,K$ olmak üzere $\boldsymbol{\Gamma} = \{\gamma_{ij}\}$ şeklinde bir parçalanma matrisi, γ_{ij}

$$\gamma_{ij} = \begin{cases} 1 & x_j \in G_i \\ 0 & \text{diğer yerlerde} \end{cases} \tag{3.37}$$

şeklinde ifade edilen ve

$$\sum_{i=1}^{K} \gamma_{ij} = 1, \; j = 1,2,\ldots,n \tag{3.38}$$

koşulunu sağlayan etiket değeri, $\mathbf{m}_i$,

$$\sum_{j=1}^{n} \gamma_{ij} \mathbf{m}_i \tag{3.39}$$

şeklinde ifade edilen $i.$ küme ortalaması ve $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K]$ küme merkezleri (cenroids) matrisini göstermektedir.

Hata kareleri toplamını minimum yapacak şekilde elde edilen parçalanma en uygun parçalanma olarak düşünülür ve paylaştırmaya minimum varyans paylaştırması denir. Hata kareleri toplamı kriteri özellikle birbirine daha yakın gözlemler ve birbirinden iyi ayrışımlı küme yapısı içeren veriler için uygun bir kriterdir. Veride sapan aykırı değerler bulunması durumunda hata kareleri toplamı kriteri geniş kümeleri daha küçük parçalara ayırma eğilimindedir (Duda ve ark. 2001).

Hata kareleri toplam kriteri saçılım matrislerinde de elde edilebilir (Duda ve ark. 2001).

$$\mathbf{S}_T = \mathbf{S}_W + \mathbf{S}_B \tag{3.40}$$

Burada, $\mathbf{S}_T$, $\mathbf{S}_W$ ve $\mathbf{S}_B$ sırasıyla toplam saçılım matrisi, küme içi saçılım matrisi, kümeler arası saçılım matrisidir ve

$$\mathbf{S}_T = \sum_{j=1}^{n} (\mathbf{x}_j - \mathbf{m})(\mathbf{x}_j - \mathbf{m})^T \tag{3.41}$$

$$\mathbf{S}_W = \sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_{ij} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \tag{3.42}$$

$$\mathbf{S}_B = \sum_{i=1}^{K} N_i (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \tag{3.43}$$

olarak ifade edilir. Burada $\mathbf{m}$ verinin tamamı için toplam ortalama vektörü olup

$$\mathbf{m} = \frac{1}{n} \sum_{i=1}^{K} n_i \mathbf{m}_i \tag{3.44}$$

biçiminde ifade edilir. $\mathbf{S}_W$ matrisinin izi

$$iz(\mathbf{S}_W) = \sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_{ij} (\mathbf{x}_j - \mathbf{m}_i)^T (\mathbf{x}_j - \mathbf{m}_i) \tag{3.45}$$

formumdadır. $\mathbf{S}_W$ matrisinin izinin minimum yapılması hata kareler toplamı kriterine eşdeğerdir. Benzer düşünceyle $iz(\mathbf{S}_T) = iz(\mathbf{S}_W) + iz(\mathbf{S}_B)$ olduğundan $iz(\mathbf{S}_B)$ ifadesinin maksimum yapılması hata kareleri toplamı kriterine eşdeğerdir.

Saçılım matrislerinin determinantına dayalı olarak da hata kareleri toplamı kriterine eşdeğer ifadeler elde edilir. Bunun için $\mathbf{S}_W$ matrisinin determinantı,

$$\left|\mathbf{S}_W\right| = \left| \sum_{i=1}^{K} \sum_{j=1}^{n} \gamma_{ij} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T \right| = \left| E_s(\mathbf{\Gamma}, \mathbf{M}) \right| \tag{3.46}$$

olduğundan bu ifade minimum yapılır. Küme içindeki gözlem sayısının verideki değişken sayısından az olması durumunda $\mathbf{S}_B$ matrisinin determinantı sıfır olacağından $\mathbf{S}_B$ matrisini kullanmak tercih edilmez. Determinanta dayalı olarak hata karelerin minimum yapılması lineer dönüşümlerden etkilenmez.

$iz(\mathbf{S}_W)$ ifadesinin minimum yapılması kriterinde, veride lineer dönüşümler öncesi ve sonrası farklı kümelenme sonuçları elde edilmektedir. Bu durumu ortadan kaldırmak için kümeler arası ve küme içi saçılım matrislerinin çarpımından elde edilen matrisin izinin maksimum yapılmasına dayalı bir kriter kullanılabilir. Kümeler arası ve küme içi saçılım matrislerinin çarpımından elde edilen matrisin izi,

$$tr\left(\mathbf{S}_W^{-1}\mathbf{S}_B\right)=\sum_{i=1}^{p}\lambda_i \tag{3.47}$$

şeklinde ifade edilir. Burada $\lambda_1,\lambda_2,\ldots,\lambda_p$, $\mathbf{S}_W^{-1}\mathbf{S}_B$ çarpım matrisinin özdeğerleridir. Benzer şekilde,

$$tr\left(\mathbf{S}_T^{-1}\mathbf{S}_W\right)=\sum_{i=1}^{p}\frac{1}{1+\lambda_i} \tag{3.48}$$

ve

$$\frac{\left|\mathbf{S}_T\right|}{\left|\mathbf{S}_W\right|}=\prod_{i=1}^{p}(1+\lambda_i) \tag{3.49}$$

kriterleri de kullanılabilir (Everitt ve ark. 2001).

(3.36)'daki eşitlik ile verilen hata kareler toplamı kriterinden, ortalama vektörleri ihmal edilirse

$$E_s(\mathbf{\Gamma})=\sum_{i=1}^{K}\frac{1}{n_i}\sum_{\mathbf{x}_m\in G_i}^{n}\sum_{\mathbf{x}_j\in G_i}\left\|\mathbf{x}_m-\mathbf{x}_j\right\|^2 \tag{3.50}$$

eşitliği elde edilir. Bu ifade G_i kümesinin homojenliği için bir ölçüdür. Bu eşitlik genelleştirilerek,

$$E_s(\mathbf{\Gamma}) = \sum_{i=1}^{K} \psi_1(i) \tag{3.51}$$

biçiminde yazılır. Burada $\psi_1(i)$,

$$\psi_1(i) = \frac{1}{n_i} \sum_{\mathbf{x}_m \in G_i} \sum_{\mathbf{x}_j \in G_i} D_{mj} \tag{3.52}$$

şeklinde ifade edilir. D_{mj}, $\mathbf{x}_m$ ve $\mathbf{x}_j$ gözlem çifti arasında bir uzaklık tanımlar. 3.41 eşitliği ile verilen kriter bir küme içerisinde yer alan gözlemlerin farklılıklarına ilişkin bir ölçüdür.

Homojenliğe dayalı diğer indisler (Everitt ve ark. 2001, Hansen ve Jaumard 1997) aşağıda verilmiştir.

1. G_i kümesindeki gözlemlerin aralarındaki maksimum uzaklığı ölçen G_i kümesinin çapı

$$\psi_2(i) = \max_{\substack{\mathbf{x}_m \in G_i, \mathbf{x}_j \in G_i \\ m \neq j}} (D_{mj}) \tag{3.53}$$

2. G_i kümesindeki bir $\mathbf{x}_j$ gözlemi ile bu kümedeki diğer gözlemler arasındaki uzaklıklar toplamının minimumuna dayalı star indeksi

$$\psi_3(i) = \min_{\mathbf{x}_m \in G_i} \left(\sum_{\mathbf{x}_j \in G_i} D_{mj} \right) \tag{3.54}$$

3. G_i kümesindeki tüm $\mathbf{x}_j$ gözlemleri için, $\mathbf{x}_j$ gözlemi ile diğer gözlemler arası maksimum uzaklığın minimumunu ölçen G_i kümesinin yarıçapı indeksi

$$\psi_4(i) = \min_{\mathbf{x}_m \in G_i} \left(\max_{\mathbf{x}_j \in G_i} D_{mj} \right) \tag{3.55}$$

4. G_i kümesinin bir gözlemi ile G_i kümesinin dışındaki tüm gözlemler arasındaki uzaklıklar toplamının bir ölçüsü olan G_i kümesinin kesme indeksi

$$\psi_5(i) = \sum_{\mathbf{x}_m \in G_i} \sum_{\mathbf{x}_j \notin G_i} D_{mj} \tag{3.56}$$

5. G_i kümesinin bir gözlemi ile G_i kümesinin dışındaki tüm gözlemler arasındaki minimum uzaklığı ölçen, G_i kümesinin ayrılma indeksi

$$\psi_6(i) = \min_{\mathbf{x}_m \in G_i, \mathbf{x}_j \notin G_i} \min(D_{mj}) \tag{3.57}$$

3.4.2. *k* -ortalamalar Algoritması

k -ortalamalar algoritması (MacQueen 1967) kümeleme analizinde en yaygın olarak kullanılan bir kümeleme algoritmasıdır (Duda ve ark. 2001). *k* -ortalamalar algoritması, (3.36)'daki eşitlik ile verilen hata kareler toplamının minimum yapılmasına dayalı olarak verideki en uygun parçalanmayı bulmayı amaçlayan adımsal optimizasyon yöntemidir. *k* -ortalamalar algoritmasının adımları aşağıda verilmiştir.

1. Veri hakkında sahip olunan bazı önbilgilere dayalı olarak verinin bir K parçalanması rasgele yapılır. Oluşturulan K kümenin küme ortalama matrisi $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K]$ hesaplanır.
2. Verideki her bir gözlem,

$$\left\|\mathbf{x}_j - \mathbf{m}_l\right\| < \left\|\mathbf{x}_j - \mathbf{m}_i\right\|,\ j = 1,2,\ldots,n,\ i = 1,2,\ldots,K \text{ ve } i \neq l \tag{3.58}$$

koşulunu sağlıyorsa kendine en yakın G_l kümesine atanır.

3. Küme ortalama matrisi 2. adımdaki atanmalar dikkate alınarak yeniden,

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x}_j \in G_i} \mathbf{x}_j \tag{3.59}$$

şeklinde hesaplanır.

4. Algoritmanın 2. ve 3. adımlarına küme yapısı sabitleninceye (gözlemler bulundukları kümeden başka bir kümeye atanamaz hale gelinceye) kadar ardışık olarak uygulanılır.

k-ortalamalar algoritmasının 1. adımında verinin parçalanması ve bu parçalanmaya göre küme merkezlerinin oluşturulması kümeleme sonucuna oldukça etki eder. Bu etkinin azaltılması için başlangıçta verinin nasıl parçalanacağına ilişkin genel bir yöntem yoktur. Bu amaçla k-ortalamalar algoritması rastgele alınan başlangıç değerlerle birkaç kez uygulanır (Jain ve Dubes 1988). k-ortalamalar algoritmasının 1. adımında K farklı parçalanma alma yerine K gözlem değerleride alınabilir.

Örnek 3.5: Avrupa'da 25 ülkenin tüketilen dokuz gıda ürününe göre elde edilen protein verisi (Hand ve ark. 1994) Tablo 3.11 ile verilmiştir. İncelenen gıda türleri yani değişkenler; Y1: Kırmızı et, Y2: Beyaz et, Y3: Yumurta, Y4: Süt, Y5: Balık, Y6: Tahıl, Y7: Nişastalı ürünler, Y8: Fındık ve Y9: Meyve ve Sebzeler. Protein verisi Tablo ile verilmiştir.

Tablo 3.11 Protein verisi; Avrupanın 25 ülkesinin dokuz farklı gıda cinsine göre protein tüketimi (Hand ve ark. 1994).

	G1	G2	G3	G4	G5	G6	G7	G8	G9
Arnavutluk	10.10	1.40	0.50	8.90	0.20	42.30	0.60	5.50	1.70
Avusturya	8.90	14.00	4.30	19.90	2.10	28.00	3.60	1.30	4.30
Belçika	13.50	9.30	4.10	17.50	4.50	26.60	5.70	2.10	4.00
Bulgaristan	7.80	6.00	1.60	8.30	1.20	56.70	1.10	3.70	4.20
Çek. Cum.	9.70	11.40	2.80	12.50	2.00	34.30	5.00	1.10	4.00
Danimarka	10.60	10.80	3.70	25.00	9.90	21.90	4.80	0.70	2.40
D.Almanya	8.40	11.60	3.70	11.10	5.40	24.60	6.50	0.80	3.60
Finlandiya	9.50	4.90	2.70	33.70	5.80	26.30	5.10	1.00	1.40
Fransa	18.00	9.90	3.30	19.50	5.70	28.10	4.80	2.40	6.50
Yunanistan	10.20	3.00	2.80	17.60	5.90	41.70	2.20	7.80	6.50
Macaristan	5.30	12.40	2.90	9.70	0.30	40.10	4.00	5.40	4.20
İrlanda	13.90	10.00	4.70	25.80	2.20	24.00	6.20	1.60	2.90
İtalya	9.00	5.10	2.90	13.70	3.40	36.80	2.10	4.30	6.70
Hollanda	9.50	13.60	3.60	23.40	2.50	22.40	4.20	1.80	3.70
Norveç	9.40	4.70	2.70	23.30	9.70	23.00	4.60	1.60	2.70
Polonya	6.90	10.20	2.70	19.30	3.00	36.10	5.90	2.00	6.60
Portekiz	6.20	3.70	1.10	4.90	14.20	27.00	5.90	4.70	7.90
Romanya	6.20	6.30	1.50	11.10	1.00	49.60	3.10	5.30	2.80
İspanya	7.10	3.40	3.10	8.60	7.00	29.20	5.70	5.90	7.20
İsveç	9.90	7.80	3.50	24.70	7.50	19.50	3.70	1.40	2.00
İsviçre	13.10	10.10	3.10	23.80	2.30	25.60	2.80	2.40	4.90
İngiltere	17.40	5.70	4.70	20.60	4.30	24.30	4.70	3.40	3.30
Rusya	9.30	4.60	2.10	16.60	3.00	43.60	6.40	3.40	2.90
B.Almanya	11.40	12.50	4.10	18.80	3.40	18.60	5.20	1.50	3.80
Yugoslavya	4.40	5.00	1.20	9.50	0.60	55.90	3.00	5.70	3.20

Bu ülkelerin tüketilen gıda türleri cinsinden beş kümede toplandığı düşünülürse bu beş kümenin k-ortalamalar algoritması kullanılarak elde etmek için beş gözlem değeri rasgele olarak seçilir. Rastgele seçilen beş gözlem Tablo 3.12 ile verilmiştir.

Tablo 3.12 Protein tüketimi verisinden rastgele seçilen beş ülke ve protein tüketimleri.

Seeds	G1	G2	G3	G4	G5	G6	G7	G8	G9
İrlanda	13.90	10.00	4.70	25.80	2.20	24.00	6.20	1.60	2.90
İngiltere	17.40	5.70	4.70	20.60	4.30	24.30	4.70	3.40	3.30
Polonya	6.90	10.20	2.70	19.30	3.00	36.10	5.90	2.00	6.60
Yunanistan	10.20	3.00	2.80	17.60	5.90	41.70	2.20	7.80	6.50
D.Almanya	8.40	11.60	3.70	11.10	5.40	24.60	6.50	0.80	3.60

Seçilen bu beş rasgele ülke ile diğer ülkelerin protein tüketimleri cinsinden aralarındaki Öklid uzaklığı hesaplanır. Bu uzaklık değerlerine göre diğer ülkeler Tablo 3.13 ile belirtildiği gibi beş farklı kümeye gruplanır.

Tablo 3.13 Protein tüketimi verisinden Avrupa'dan 25 ülkenin, seçilen 5 rastgele ülke kümesine göre gruplanması.

Ülkeler	C1	C2	C3	C4	C5	Kümeler
Danimarka	8.94	11.74	17.81	24.47	15.18	C1
Finlandiya	11.57	15.95	19.36	24.11	23.83	C1
İrlanda	0.00	8.24	16.01	22.98	16.18	C1
Hollanda	6.83	11.99	15.34	24.06	13.27	C1
İsveç	8.88	10.90	19.15	25.16	15.58	C1
İsviçre	5.10	7.97	13.50	20.04	14.65	C1
B.Almanya	9.71	11.10	18.51	26.28	10.56	C1
Belçika	9.15	6.83	12.20	18.25	8.93	C2
Fransa	10.15	6.99	14.01	18.25	13.86	C2
Norveç	10.88	10.52	16.91	21.42	14.99	C2
İngiltere	8.24	0.00	17.07	20.32	14.78	C2
Avusturya	10.04	12.92	9.94	19.53	10.74	C3
Çek. Cum.	17.58	16.50	8.26	15.03	10.61	C3
Macaristan	25.02	24.17	11.99	14.93	17.52	C3
Polonya	16.01	17.07	0.00	12.36	14.87	C3
Bulgaristan	38.36	36.42	24.49	19.32	33.61	C4
Yunanistan	22.98	20.32	12.36	0.00	22.11	C4
İtalya	20.04	17.26	9.07	7.98	15.62	C4
Romanya	31.30	29.73	17.43	13.34	26.73	C4
Rusya	23.02	21.61	10.81	8.18	21.37	C4
Yugoslavya	37.95	36.41	23.73	18.75	33.28	C4
Arnavutluk	27.90	24.31	17.64	12.14	22.87	C4
D.Almanya	16.18	14.78	14.87	22.11	0.00	C5
Portekiz	27.14	22.77	21.68	22.16	15.16	C5
İspanya	21.81	17.66	15.50	16.27	11.78	C5

Tablo 3.13 ile belirtildiği gibi gözlemlerin atandığı 5 küme yapısı için küme merkezleri Tablo 3.14 verildiği gibi hesaplanır.

Tablo 3.14 Protein tüketimi verisi için yeni küme merkezleri.

	G1	G2	G3	G4	G5	G6	G7	G8	G9
C1	11.13	9.96	3.63	25.03	4.80	22.61	4.57	1.49	3.01
C2	14.58	7.40	3.70	20.23	6.05	25.50	4.95	2.38	4.13
C3	7.70	12.00	3.18	15.35	1.85	34.63	4.63	2.45	4.78
C4	8.14	4.49	1.80	12.24	2.19	46.66	2.64	5.10	4.00
C5	7.23	6.23	2.63	8.20	8.87	26.93	6.03	3.80	6.23

Protein verisindeki gözlemler, Tablo 3.14'de verilen yeni küme merkezlerine olan Öklid uzaklıklarına dayalı olarak yeniden beş farklı kümeye Tablo 3.15'de belirtildiği gibi gruplanır.

Tablo 3.15. Protein tüketimi verisindeki güncellenmiş beş küme yapısı.

Ülke	C1	C2	C3	C4	C5	Kümeler
Danimarka	5.34	9.15	18.40	30.19	19.18	G1
Finlandiya	11.05	14.99	22.12	30.34	26.44	G1
İrlanda	4.56	7.75	16.54	28.20	21.02	G1
Hollanda	4.96	9.88	14.91	28.54	19.14	G1
İsveç	5.01	9.34	19.50	30.93	19.21	G1
İsviçre	5.33	6.47	13.75	25.36	18.76	G1
B.Almanya	8.04	9.69	16.97	30.52	16.80	G1
Belçika	9.03	4.08	10.94	22.60	12.83	G2
Fransa	11.04	5.58	13.72	23.59	16.51	G2
Norveç	7.66	8.18	17.93	27.54	16.45	G2-> G1
İngiltere	9.22	4.28	16.71	26.03	17.32	G2
Avusturya	9.33	10.06	8.60	22.79	16.30	G3
Çek. Cum.	17.54	13.99	3.91	15.02	12.89	G3
Macaristan	24.87	21.88	8.93	11.27	17.37	G3
Polonya	15.86	14.05	5.33	14.96	16.15	G3
Bulgaristan	38.78	34.76	24.29	11.14	31.22	G4
Yunanistan	22.98	18.74	14.01	9.39	19.11	G4
İtalya	19.69	15.26	8.53	10.55	13.34	G4-> G3
Romanya	31.55	27.92	17.17	4.50	24.68	G4
Rusya	23.58	19.82	12.23	6.98	20.03	G4
Yugoslavya	38.25	34.64	23.83	10.55	30.70	G4
Arnavutluk	28.04	23.36	16.23	7.67	20.02	G4
D.Almanya	14.59	12.06	11.82	24.21	8.56	G5
Portekiz	24.87	20.37	20.29	24.81	7.26	G5
İspanya	20.39	15.64	13.96	19.11	4.74	G5

Tablo 3.15 ile özetlenen küme yapılarında gözlemlerin bazıları bulundukları kümelerin dışında diğer kümelere atandıkları için k-ortalamalar algoritmasına devam edilir. Tablo 3.15 ile verilen kümeleme yapısına göre oluşturulan yeni küme merkezleri Tablo 3.16 ile verilmiştir.

Tablo 3.16. Protein tüketimi verisi için yeni küme ortalama vektörleri.

	G1	G2	G3	G4	G5	G6	G7	G8	G9
C1	10.91	9.30	3.51	24.81	5.41	22.66	4.58	1.50	2.98
C2	16.30	8.30	4.03	19.20	4.83	26.33	5.07	2.63	4.60
C3	7.96	10.62	3.12	15.02	2.16	35.06	4.12	2.82	5.16
C4	8.00	4.38	1.62	12.00	1.98	48.30	2.73	5.23	3.55
C5	7.23	6.23	2.63	8.20	8.87	26.93	6.03	3.80	6.23

Tablo 3.17 Protein tüketimi verisinden Avrupa'dan 25 ülkesinin yeni küme ortalamalar vektörlerine olan uzaklıkları ve gruplandıkları kümeler.

Ülke	C1	C2	C3	C4	C5	Kümeler
Danimarka	4.91	11.25	18.78	31.73	19.18	C1
Finlandiya	10.83	16.84	22.19	31.64	26.44	C1
İrlanda	5.16	8.38	16.96	29.71	21.02	C1
Hollanda	5.64	10.74	15.66	30.10	19.14	C1
İsveç	4.40	11.65	19.71	32.53	19.21	C1
İsviçre	5.71	6.91	13.97	26.93	18.76	C1
B.Almanya	8.28	10.28	17.55	32.16	16.80	C1
Norveç	6.71	10.90	17.77	29.12	16.45	C1
Belçika	8.92	3.60	11.00	24.23	12.83	C2
Fransa	11.04	3.70	13.60	25.15	16.51	C2
İngiltere	9.05	4.15	16.48	27.61	17.32	C2
Avusturya	9.65	10.10	9.52	24.33	16.30	C3
Çek. Cum.	17.51	13.18	3.97	16.50	12.89	C3
Macaristan	24.84	21.17	8.66	12.27	17.37	C3
Polonya	15.79	14.06	5.24	16.41	16.15	C3
İtalya	19.41	14.92	6.82	12.31	13.34	C3
Bulgaristan	38.62	33.98	23.43	9.64	31.22	C4
Yunanistan	22.65	18.58	12.58	10.68	19.11	C4
Romanya	31.38	27.35	16.29	3.33	24.68	C4
Rusya	23.35	19.49	11.22	7.97	20.03	C4
Yugoslavya	38.06	34.09	22.99	8.94	30.70	C4
Arnavutluk	27.78	22.75	14.89	8.45	20.02	C4
D.Almanya	14.42	12.20	12.20	25.78	8.56	C5
Portekiz	24.25	20.98	19.53	26.20	7.26	C5
İspanya	19.90	15.89	12.96	20.69	4.74	C5

Tablo 3.16 ile verilen yeni küme merkezlerine göre gözlemlerin hangi kümeye atandıkları Tablo 3.17 ile verilmiştir. Tablo 3.17'den görüldüğü gibi hiçbir gözlem bulunduğu kümeden başka bir kümeye atanmamıştır. Bu sebepten k-

ortalamalar algoritması sonlandırılır. Protein verisindeki 5 farklı küme yapısı Tablo 3.17'de gösterildiği gibi elde edilmiş olur.

3.5. Kümelerin Değerlendirilmesi

Kümeleme analizinde kullanılan farklı algoritmalarla veya aynı algoritmada alınan farklı parametre başlangıç değeri ile farklı birçok kümeleme yapısı elde edilebilir. Aynı veri kümesinden elde edilen farklı kümeleme yapılarının anlamlı olup olmadığı kümeleme geçerliliği olarak adlandırılır (Gordon 1998, Halkidi ve ark. 2002a, Halkidi ve ark. 2002b). Herhangi bir kümeleme yapısı içermeyen bir veri kümesine uygulanılan bir kümeleme algoritması sonucu elde edilen kümeleme yapısı anlamsız olacaktır. Kümeleme sonucu ileri aşama analizlere başlanmadan önce veride bir kümeleme yapısının varlığından emin olmak için bazı testler uygulanabilir.

Kümeleme geçerlilik kriterleri kümeleme algoritmalarının tipine bağlı olarak üç kategoride toplanmaktadır (Theodoridis ve Koutroumbas 2006). Bunlar dışsal (external) kriterler, içsel (Internal) kriterler ve oransal (Relative) kriterlerdir. Bir veri kümesi $\mathbf{X}$, bu veri kümesinden elde edilen kümeleme yapısı $\mathbf{C}$ ile gösterilsin. Dışsal (external) kriterlerde veri hakkında sahip olunan ön bilgi ile kümeleme algoritması sonunda elde edilen $\mathbf{C}$ kümeleme yapısı karşılaştırılır. Örneğin kümeleme sonunda elde edilen gözlemlerin küme etiketleri ile daha önceden bilgi sahibi olunan gözlemlerin kategori etiketleri karşılaştırılır. İçsel (Internal) kriterlerde ise veri hakkında bir önbilgi kullanılmaz. Kümeleme yapısının geçerliliği elde edilen yapı içerisinden değerlendirilir. Örneğin $\mathbf{X}$ veri kümesinin $\mathbf{C}$ kümeleme yapısı benzerlik matrisi kullanılarak değerlendirilir. Oransal (Relative) kriterlerde ise farklı kümeleme algoritmaları ile elde edilen kümeleme yapıları birbiri arasında veya aynı kümeleme algoritması fakat farklı parametre değerleri ile elde edilen kümeleme yapıları birbirleriyle karşılaştırılır.

Dışsal (external) ve içsel (Internal) kümeleme geçerlilik kriterleri istatistiksel hipotez testleri ile ilişkilidir (Jain ve Dubes 1988). Jain ve Dubes(1988), kümeleme geçerlilik kriterlerinde ortak olarak kullanılan sıfır hipotezlerini üç grupta toplamıştır.

a) Sıfır hipotezi (Rastgele pozisyon hipotezleri): p boyutlu uzayın belirli bir bölgesinde n gözlem vektörünün tüm yerleri eşit olasılıklıdır.
b) Sıfır hipotezi (Rastgele grafik hipotezleri): Tüm $n \times n$ boyutlu rank sıralı benzerlik matrisleri eşit olasılığa sahiptir.
c) Sıfır Hipotezi (Rastgele etiket hipotezleri): n gözlemin etiketlerinin tüm permütasyonları eşit olasılıklıdır.

3.5.1. Dışsal (External) Kriterler

Bir **X** veri kümesinde n gözlem vektörü bulunsun. Bu **X** veri kümesi hakkında bilgi sahibi olunan parçalanma **P** ve bu veri kümesine uygulanılan bir kümeleme algoritması sonu elde edilen kümeleme yapısı **C** ile gösterilsin. **X** veri kümesinde $\mathbf{x}_i$ ve $\mathbf{x}_j$ gözlem çiftinin **C** ve **P** parçalanmalarındaki yerlerine göre dört durum vardır.

1. Durum: $\mathbf{x}_i$ ve $\mathbf{x}_j$, **C** parçalanmasında aynı küme içerisinde ve **P** parçalanmasında aynı kategori içerisinde yer almasıdır
2. Durum: $\mathbf{x}_i$ ve $\mathbf{x}_j$, **C** parçalanmasında aynı küme içerisinde ve **P** parçalanmasında farklı kategoriler içerisinde yer almasıdır
3. Durum: $\mathbf{x}_i$ ve $\mathbf{x}_j$, **C** parçalanmasında farklı küme içerisinde ve **P** parçalanmasında aynı kategori içerisinde yer almasıdır
4. Durum: $\mathbf{x}_i$ ve $\mathbf{x}_j$, **C** parçalanmasında farklı küme içerisinde ve **P** parçalanmasında farklı kategoriler içerisinde yer almasıdır

Gözlem vektör çiftlerinin yukarıda belirtilen dört durumdaki eşleşme sayıları sırasıyla a, b, c ve d ile gösterilsin. Toplam gözlem çiftlerinin sayısı

$$M = \frac{n(n-1)}{2} \tag{3.60}$$

ifade edilsin. Bu durumda $M = a + b + c + d$ olarak ifade edilir. **C** ve **P** parçalanmaları arasında eşleşmenin ölçüsü için kullanılan yaygın dışsal (external) indisleri:

1. Rand indeksi (Rand 1971)

$$R = \frac{(a+d)}{M} \tag{3.61}$$

olarak tanımlanır ve $[0,1]$ aralığında değerler alır. İndeks değerleri bire yakınlaştıkça **C** ve **P** parçalarının birbirine olan benzerliği artmaktadır.

2. Jaccard katsayısı (Jaccard 1908)

$$J = \frac{a}{a+b+c} \tag{3.62}$$

olarak tanımlanır ve $[0,1]$ aralığında değerler alır. Katsayı bire yakınlaştıkça **C** ve **P** parçalarının birbirine olan benzerliği artmaktadır

3. Fowlkes ve Mallows indeksi (Fowlkes ve Mallows 1983)

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}} \tag{3.63}$$

olarak tanımlanır. İndeks değerleri bire yakınlaştıkça **C** ve **P** parçalarının birbirine olan benzerliği artmaktadır

4. Γ istatistiği (Xu ve Wunsch 2009).

$$\Gamma = \frac{Ma - m_1 m_2}{\sqrt{m_1 m_2 (M - m_1)(M - m_2)}} \tag{3.64}$$

olarak tanımlanır ve $[-1, 1]$ aralığında değerler alır. Burada $m_1 = a + b$ ve $m_2 = a + c$ olarak tanımlanır. İndeks değerleri 1'e yakınlaştıkça **C** ve **P** parçalarının birbirine olan benzerliği, -1'e yaklaştıkça farklılığı artmaktadır.

3.5.2. İçsel (Internal) Kriterler

Hiyerarşik kümeleme yapılarının karşılaştırılmasın da en yaygın olarak kullanılan kümeleme geçerlilik ölçüsü Cophenetic korelasyon katsayısıdır (Jain ve Dubes 1988). Cophenetic korelasyon katsayısı, $\mathbf{X}$ veri kümesi için hesaplanan benzerlik matrisi $\mathbf{P} = \{p_{ij}\}$ ile hiyerarşik kümeleme yöntemine göre elde edilen ağaç diyagramında veri gözlem çiftlerinin ilk defa aynı kümede gruplandığı yakınlık seviyeleri q_{ij} değerlerinden oluşan Cophenetic matrisi $\mathbf{Q} = \{qi_j\}$ arasındaki benzerliğin bir ölçüsüdür. μ_p ve μ_q sırasıyla $\mathbf{P} = \{p_{ij}\}$ ve $\mathbf{Q} = \{qi_j\}$ 'nun ortalamaları,

$$\mu_P = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_{ij} \tag{3.65}$$

$$\mu_Q = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} q_{ij} \tag{3.66}$$

şeklinde ifade edilir. Burada M, (3.60)'daki eşitlikte tanımlandığı gibidir. Cophenetic korelasyon katsayısı, CCC kısaltmasıyla gösterilmektedir,

$$\mathrm{CCC} = \frac{\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=1}^{n} p_{ij}q_{ij} - \mu_P\mu_Q}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=1}^{n} p_{ij}^2 - \mu_P^2\right)}\sqrt{\left(\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=1}^{n} q_{ij}^2 - \mu_Q^2\right)}} \tag{3.67}$$

formu ile hesaplanır. Burada CCC, $[-1, 1]$ arasında değer alır. CCC değerinin 1 olması verideki hiyerarşi yapıdan elde edilen $\mathbf{P} = \{p_{ij}\}$ ve $\mathbf{Q} = \{qi_j\}$ arasında yüksek derecede benzerlik olduğunu gösterir. Küme kesikli ölçek değerini içeren gözlem vektörleri içeriyorsa hiyerarşik kümeleme yönteminden elde edilen $\mathbf{P} = \{p_{ij}\}$ ve $\mathbf{Q} = \{qi_j\}$ arasındaki benzerlik için Goodman-Kruskal istatistiği (Hubert 1974) ve Kendall istatistiği (Cunningham ve Ogilvie, 1972) kullanılabilir.

Tibshirani ve ark. (2001), bir veri kümesinde küme sayısı K'yı tahmin etmek için gap istatistik yöntemini önermişlerdir. Veriyi gruplamada kullanılan tüm yöntemlerde uygulanabilir olan gap istatistik yönteminde verinin yapısına uygun bir referans sıfır dağılımı oluşturulur. Bu referans dağılımından orjinal veri ile aynı hacimli rastgele örneklem üretilir. Bu rastgele örnekleme orijinal veriye uygulanan kümeleme yöntemi uygulanarak orijinal verinin küme içi saçılımı ile rastgele örneklemden elde edilen kümeleme yapısındaki küme içi saçılımlar karşılaştırılır. Gap istatistik değerini maksimum yapan K sayısı optimal küme sayısı olarak alınır. Gap istatistiğinde gözlemlerin bulundukları küme merkezlerine olan karesel uzaklıklarının toplamı olan küme içi kareler toplamı W_k,

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k \tag{3.68}$$

şeklinde ifade edilir. Burada D_k, $k.$ kümedeki gözlem çiftleri arasındaki uzaklıkların toplamı olmak üzere,

$$D_k = \sum_{\mathbf{x}_i, \mathbf{x}_j \in G_k} d_{ij} \tag{3.69}$$

şeklinde tanımlanır.

Gap istatistik yönteminin adımları aşağıdaki gibi özetlenebilir.

1. n gözlem içeren veriye bir kümeleme yöntemi, küme sayısı k $(k = 1,2,...,K)$ alınarak kümeler elde edilir.

2. Birinci adımda alınan k küme sayısına göre $\log(W_k)$ değeri hesaplanır.

3. n hacimlik bir rasgele örneklem uygun bir yöntemle üretilir. Bu rasgele örneklem $\mathbf{X}^s$ ile gösterilir.

4. Birinci adımda hangi kümeleme yöntemi uygulanmışsa $\mathbf{X}^s$ rasgele örneklemi için uygulanılır ve kümeler elde edilir.

5. Bu örneklem içinde küme içi saçılım ölçüsü hesaplanır bu değer $W_{k,b}^s$ ile ifade edilmek üzere $\log(W_{k,b}^s)$ değeri bulunur.

6. $b = 1,2,...,B$ olmak üzere 3-5 adımları B defa tekrar edilir. Bu şekilde B sayıda $\log(W_{k,b}^s)$ değeri hesaplanır.

7. Altıncı adımda hesaplanan B sayıda $\log(W_{k,b}^s)$ değerinin ortalaması,

$$\overline{W_k} = \frac{1}{B}\sum_{b=1}^{B}\log(W_{k,b}^s) \tag{3.70}$$

eşitliğiyle ve standart sapması,

$$sd_k = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left[\log(W_{k,b}^s) - \overline{W_k}\right]^2} \tag{3.71}$$

eşitliğiyle hesaplanır.

8. Tahmin edilmiş gap istatistiği,

$$gap(k) = \overline{W_k} - \log(W_k) \tag{3.72}$$

9. Bu adımda

$$s_k = sd_k\sqrt{1+\frac{1}{B}} \tag{3.73}$$

tanımlanır ve

$$gap(k) \geq gap(k+1) - s_{k+1} \tag{3.74}$$

koşulunu sağlayan en küçük k değeri küme sayısı olarak alınır.

3.5.3. Oransal (Relative) Kriterler

Dışsal (external) ve içsel (Internal) kriterler istatistiksel testlerin kullanımını gerektirdiğinden hesaplamalara oldukça duyarlıdır. Oransal (Relative) kriterler de bu gereklilik yoktur ve farklı kümeleme algoritmalarının ya da aynı kümeleme algoritmasının farklı başlangıç parametreleri ile elde edilen karşılaştırılması için önerilmişlerdir.

Bir kümeleme algoritmasında verideki küme sayısı K bilinmesi gerekir. Küme sayısı K bilinmiyorsa bu değer için mümkün minimum değer olan $K_{\min}$ ve maksimum değer olan $K_{\max}$ arasında algoritma uygulanılarak kümeleme yapıları elde edilir. Elde edilen bu farklı kümeleme yapıları belirlenen bir indise göre değerlendirilir ve en iyi indis değerine sahip kümeleme yapısı seçilir. Hiyerarşik kümeleme yapısında bu indisler bir durdurma kuralı olarak düşünülebilir. Kümeleme analizi içerisinde, kümelerin geometrik ve istatistiksel yapılarını da dikkate alarak belirli bir standartta kümelerin elde edilmesinde bu şekilde birçok indis önerilmiştir.

1. Calinski ve Harabasz indeksi (Calinski ve Harabasz 1974) CH kısaltmasıyla gösterilmektedir,

$$CH(K) = \frac{iz(\mathbf{S_B})(n-K)}{iz(\mathbf{S_W})(K-1)} \tag{3.75}$$

şeklinde ifade edilir. Burada $iz(\mathbf{S_B})$ ve $iz(\mathbf{S_W})$ sırasıyla kümeler arası ve küme içi saçılım matrislerinin iz değerleridir. Küme sayısı K'nın bir seçimi için bu indeksi maksimum yapan değerin seçimi önerilir.

2. Davies-Bouldin indeksi (Davies-Bouldin 1979), küme içindeki gözlemlerin küme merkezine olan uzaklıklarını minimum yapmayı ve kümeler arası uzaklıkları maksimum yapmayı amaçlar. $i = 1,2,\ldots,k$ ve $j = 1,2,\ldots,k$ olmak üzere *i.* ve diğer kümeler arasındaki maksimum karşılaştırma oranı her bir küme için R_i ile gösterilen küme indeksi,

$$R_i = \max_{i \neq j} \left(\frac{e_i + e_j}{d_{ij}} \right) \tag{3.76}$$

şeklinde hesaplanır. Burada d_{ij} *i.* ve *j.* küme merkezleri arasındaki uzaklığı, e_i ve e_j sırasıyla *i.* ve *j.* kümenin ortalama hatalarını gösterir. Davies-Bouldin indeksi, DB kısaltmasıyla gösterilmektedir,

$$DB(K) = \frac{1}{K} \sum_{i=1}^{K} R_i \tag{3.77}$$

şeklinde tanımlanır. Minimum Davies-Bouldin indeks değerinde küme sayısı K belirlenebilir.

3. Dunn indeksi (Dunn 1974) yüksek yoğunlu iyi ayrımlı kümelerin belirlenmesinde oldukça etkin bir kriterdir. G_i ve G_j ile gösterilen iki küme arasındaki $D(G_i, G_j)$ uzaklığı sırasıyla G_i ve G_j kümelerinde yer alan **x** ve **y** gözlem çiftleri arasındaki minimum uzaklık olarak,

$$D(G_i, G_j) = \min_{\mathbf{x} \in G_i, \mathbf{y} \in G_j} d(\mathbf{x}, \mathbf{y}) \tag{3.78}$$

biçiminde tanımlanır. G_i kümesinin çapı ise bu küme içerisinde yer alan bir gözlem çifti arasındaki maksimum uzaklık olarak,

$$çap(G_i) = \max_{\mathbf{x},\mathbf{y}\in\mathbf{G}_i} d(\mathbf{x},\mathbf{y}) \tag{3.79}$$

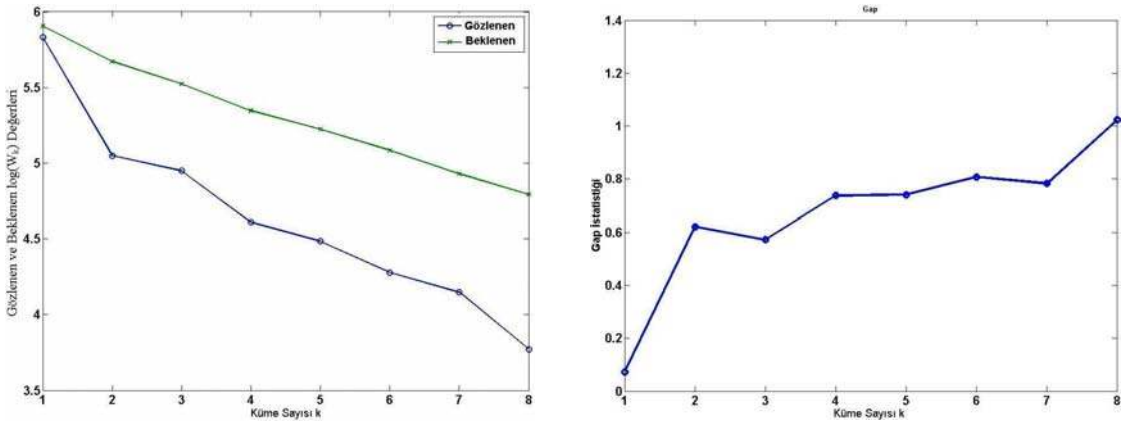
olarak tanımlanır. Dunn indeksi ($DN(K)$),

$$DN(K) = \min_{i=1,2,...,K}\left(\min_{j=i+1,...,K}\left(\frac{D(G_i,G_j)}{\max_{l=1,2,...,K} çap(G_l)}\right)\right) \tag{3.80}$$

şeklinde ifade edilir.

Örnek 3.6: Hiyerarşik kümeleme yöntemlerinden Ortalama bağlantı ve Ward kümeleme yöntemleri Iris (süsen) çiçeği verisindeki (Fisher 1936) kümeleneme yapısını bulmadaki performansları bakımından gap istatistiğine dayalı olarak karşılaştırılacaktır.

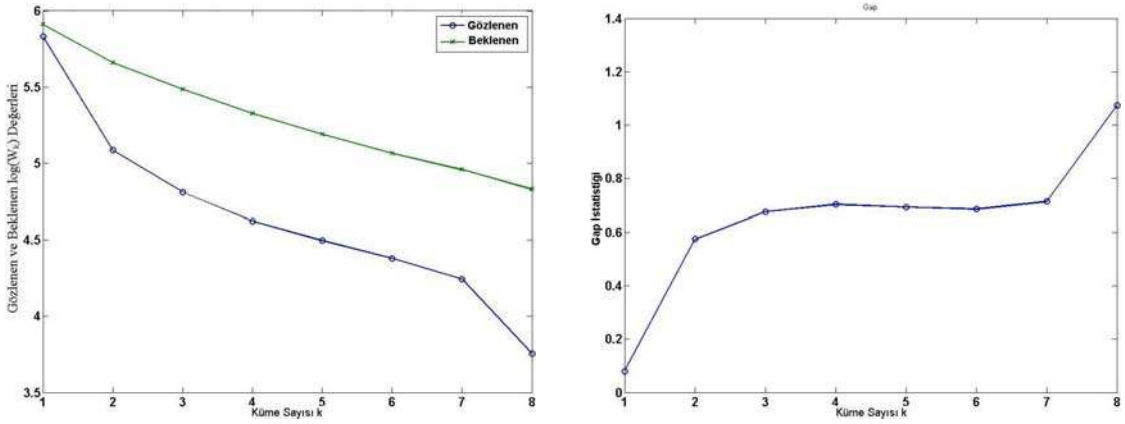
Iris (süsen) çiçeği verisi için city block uzaklığına dayalı ortalama bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.17. Iris (süsen) çiçeği verisine city block uzaklığına dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

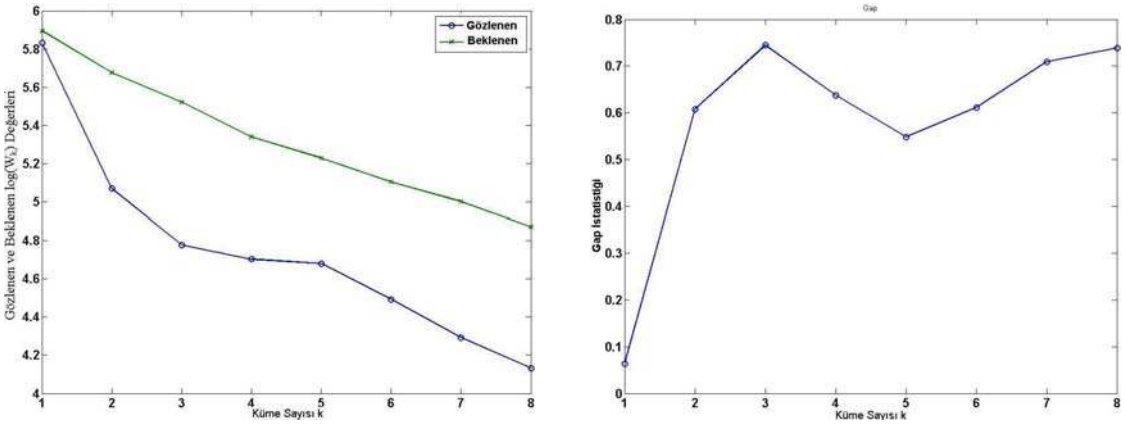
Iris (süsen) çiçeği verisi için city block uzaklığına dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.18. Iris (süsen) çiçeği verisine city Block uzaklığına dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

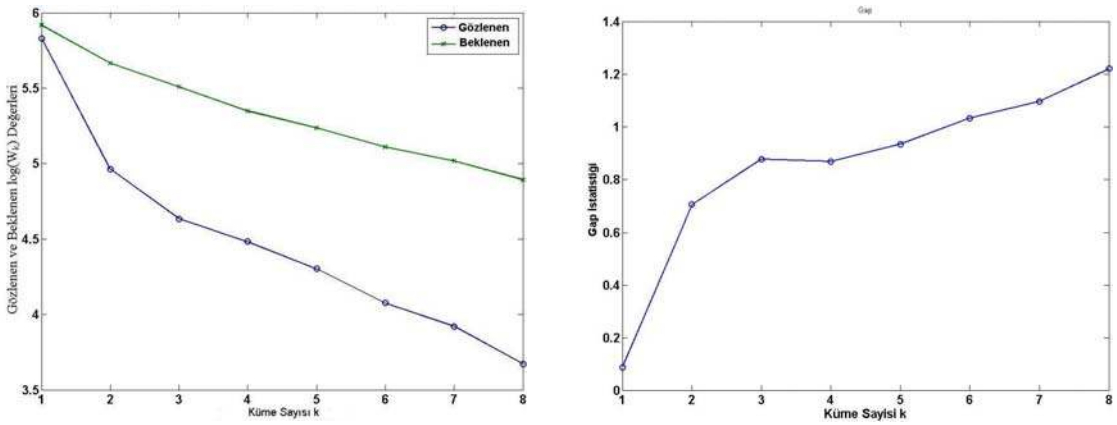
Iris (süsen) çiçeği verisi için Öklid uzaklığına dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.19. Iris (süsen) çiçeği verisine Öklid uzaklığına dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

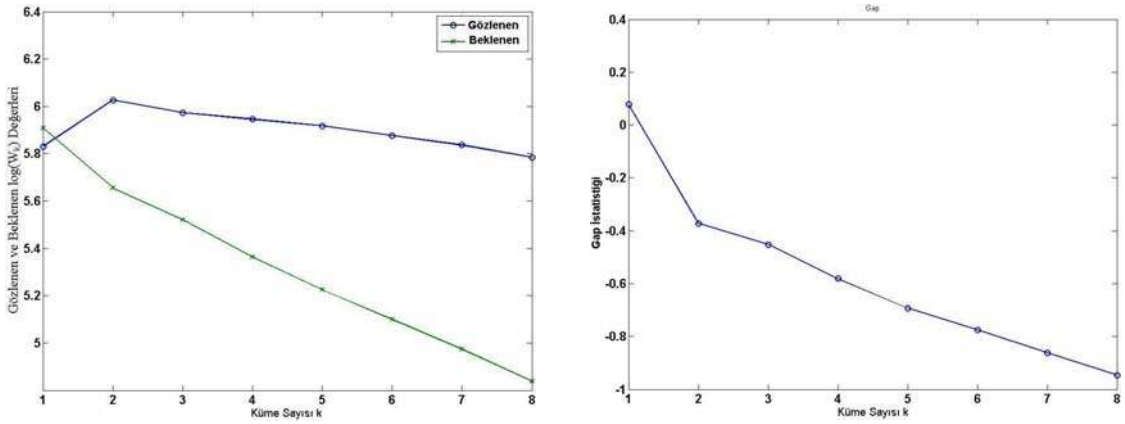
Iris (süsen) çiçeği verisi için Öklid uzaklığına dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.20. Iris (süsen) çiçeği verisine Öklid uzaklığına dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

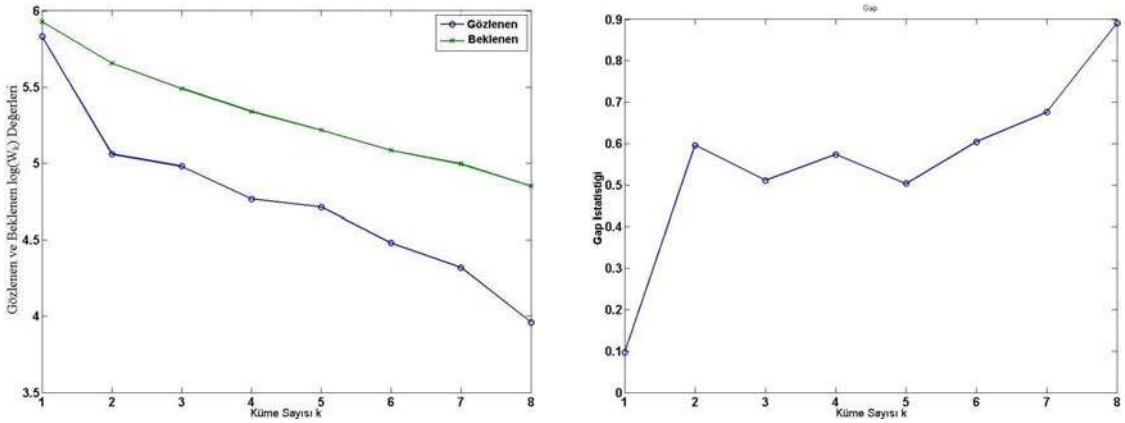
Iris (süsen) çiçeği verisi için Mahalanobis uzaklığına dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.21. Iris (süsen) çiçeği verisine Mahalanobis uzaklığına dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

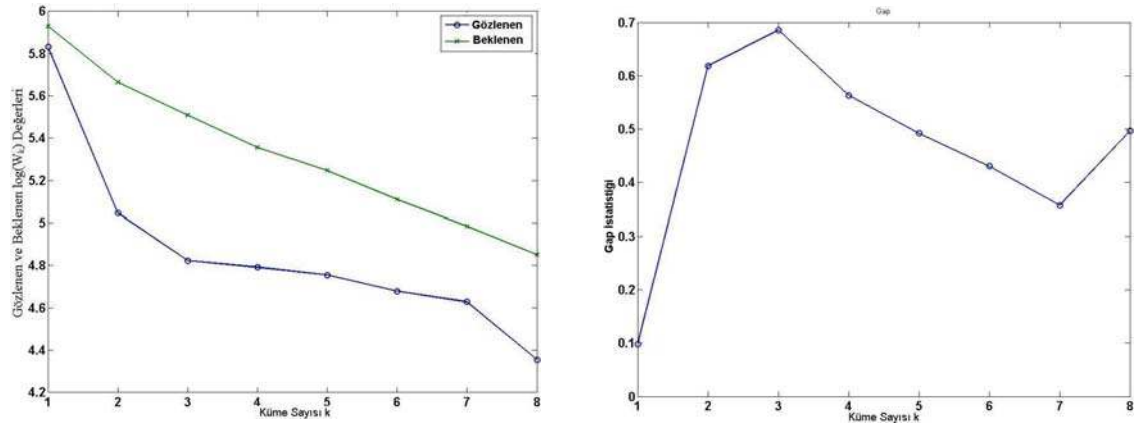
Iris (süsen) çiçeği verisi için Mahalanobis uzaklığına dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.22. Iris (süsen) çiçeği verisine Mahalanobis uzaklığına dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

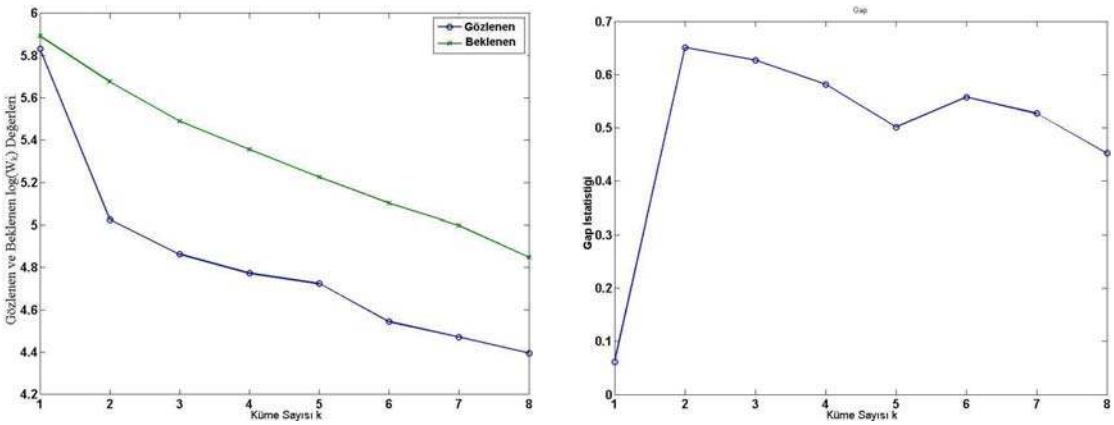
Iris (süsen) çiçeği verisi için Pearson korelasyon ölçüsüne dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.23. Iris (süsen) çiçeği verisine Pearson korelasyon ölçüsüne dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

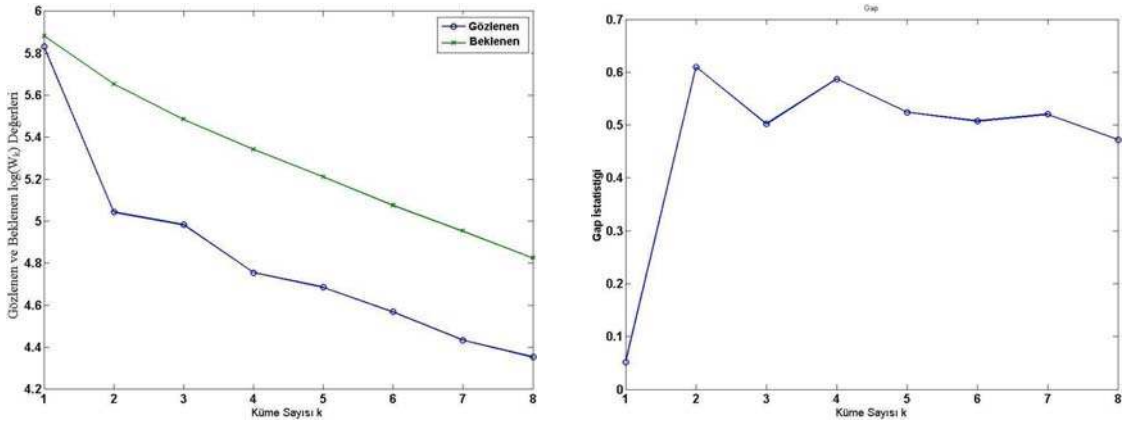
Iris (süsen) çiçeği verisi için Pearson korelasyon ölçüsüne dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.24. Iris (süsen) çiçeği verisine Pearson korelasyon ölçüsüne dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

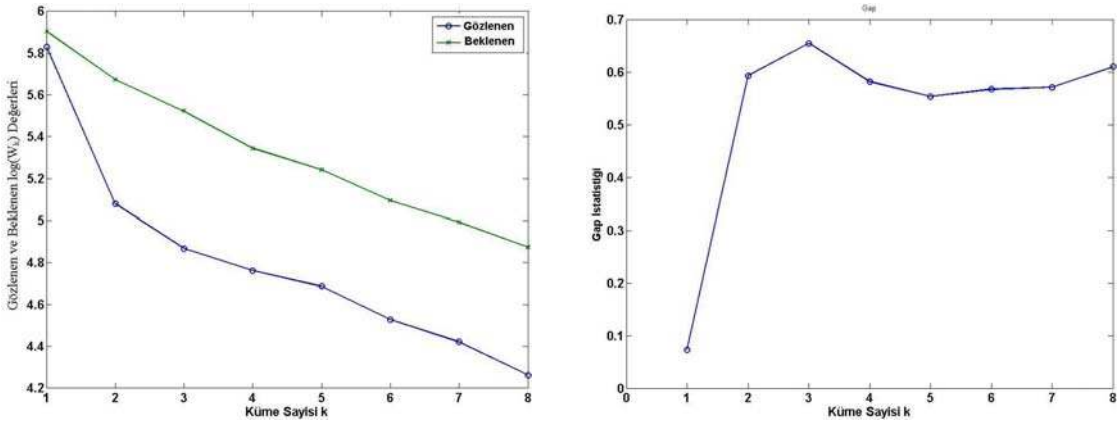
Iris (süsen) çiçeği verisi için Kosinüs benzerliğine dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.25. Iris (süsen) çiçeği verisine Kosinüs benzerliğine dayalı Ortalama Bağlantı kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

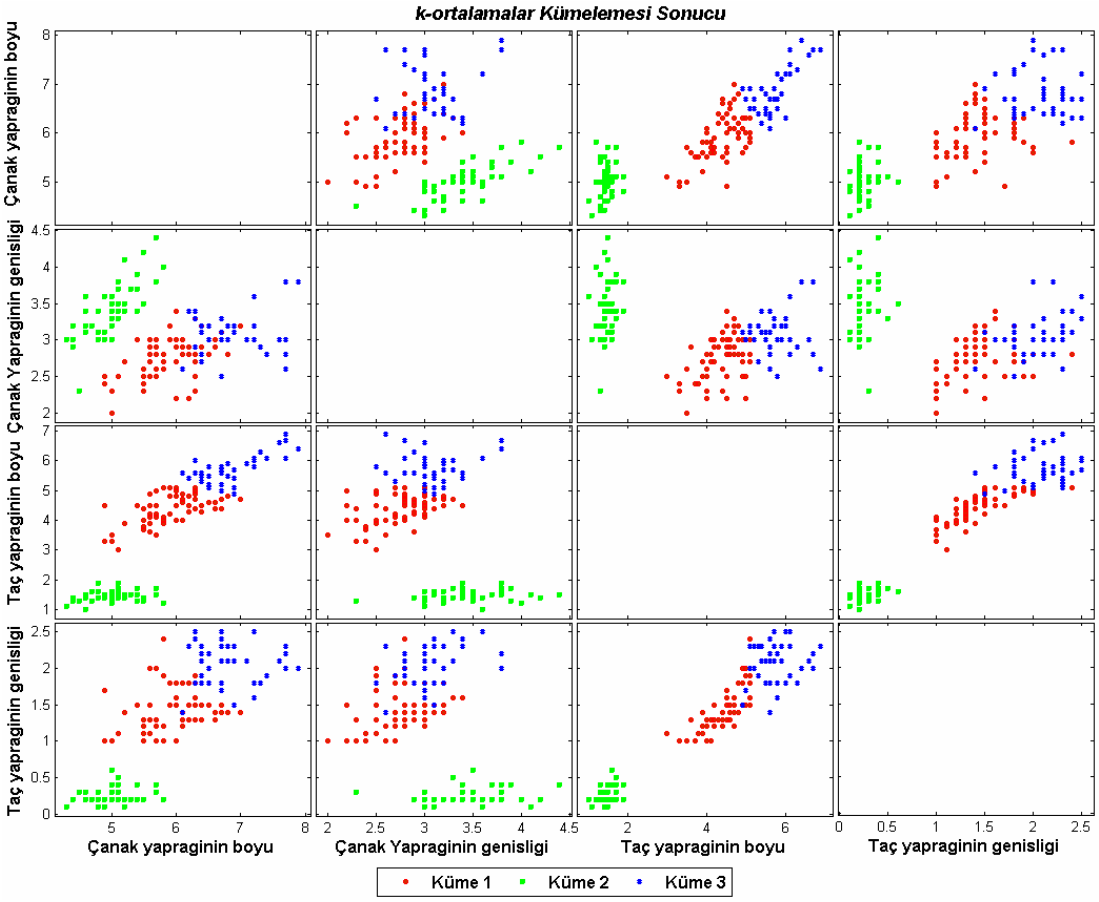
Iris (süsen) çiçeği verisi için Kosinüs benzerliğine dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı aşağıda verilmiştir.



(a) Gap istatistiği (b) Küme Sayısı

Şekil 3.26. Iris (süsen) çiçeği verisine Kosinüs benzerliğine dayalı Ward kümeleme yöntemi uygulandığında farklı küme sayılarında elde edilen gap istatistiği ve bu istatistik değerlerine göre en uygun küme sayısı.

Örnek 3.7: Iris (süsen) çiçeği verisinin k-ortalamalar algoritması uygulandığında elde edilen kümeleme sonucu ile verinin gerçek sınıflandırması karşılaştırıldığında Rand indeks değeri 0.8797 olarak bulunur. Rand indeks değerine göre k-ortalamalar algoritması kullanılarak elde edilen sınıflandırma ile verinin gerçek sınıflandırması arasında iyi bir uyumun olduğu söylenir.



Şekil 3.27. Iris (süsen) çiçeği verisinin k-ortalamalar kümeleme algoritması ile elde edilen üç küme yapısına göre saçılım grafiği.

Örnek 3.8: Iris (süsen) çiçeği verisine uygulanan hiyerarşik kümeleme yöntemlerin farklı metrik ölçümlerine göre elde edilen Cophenetic korelasyon katsayı değerleri aşağıdaki tabloda verilmiştir.

Tablo 3.18. Iris (süsen) çiçeği verisi için hiyerarşik kümeleme yöntemlerinin farklı uzaklık ölçülerine göre elde edilen Cophenetic katsayı değerleri.

	Hiyerarşik Kümeleme Yöntemleri					
Ölçütler	Tek Bağlantı	Tam Bağlantı	Ortalama Bağlantı	Merkezi Bağlantı	Medyan	Ward
Öklid	0.8639	0.7276	**0.8770**	0.8768	0.7537	0.8728
Mahalonabis	0.6451	0.4721	**0.6768**	0.6516	0.4980	0.5250
Minkowski	0.8639	0.7276	**0.8770**	0.8768	0.7537	0.8728
Jaccard	0.3744	0.6368	**0.6864**	0.3936	0.3548	0.4210

4. ÇOK DEĞİŞKENLİ SONLU KARMA DAĞILIM MODELLERİNE DAYALI KÜMELEME

Sonlu karma dağılımlar tek değişkenli ve çok değişkenli veriler için oldukça uygun ve güçlü bir modelleme aracıdır (Fraley ve Raftery 2002, McLachlan ve Peel 2000). Çok değişkenli veriye istatistiksel modellemenin yapıldığı her alanda sonlu karma dağılımlar kullanılabilir (Fraley 1998). Modele dayalı kümelemede gözlenen çok değişkenli verinin, olasılık dağılımlarının bir karmasından elde edildiği kabul edilir ve çok değişkenli dağılımların bir karmasıyla modellenir (McLachlan ve Peel 2000). Modele dayalı kümeleme de karma dağılım modelindeki her bir bileşen dağılımı bir grup ya da kümeye karşılık gelir (Fraley ve Raftery 2002).

4.1. Sonlu Karma Dağılım Modelleri İçin Temel Gösterim ve Tanımlar

4.1.1. Temel gösterimler

Bir n hacimlik bir rastgele örneklem $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ ile gösterilsin ve burada her bir $\mathbf{Y}_j$, $\Re^p$ uzayında tanımlı, p-boyutlu bir rastgele vektör olsun. Rastgele vektör $\mathbf{Y}_j$'nin olasılık yoğunluk fonksiyonu $f(\mathbf{y}_j)$ ile gösterilsin. Bu durumda rastgele vektör $\mathbf{Y}_j$, ilgilenilen j. nesne ya da birey üzerinden gözlenen p tane özelliğe ait ölçüm değerlerini içerir. Tüm örneklem $\mathbf{Y}$ ile gösterilirse, $\mathbf{Y}=\left(\mathbf{Y}_1^T,\ldots\mathbf{Y}_n^T\right)^T$ şeklinde yazılır. Burada üst indis T, vektör devriğini veya matris devriğini göstermektedir. Gözlenmiş bir rastgele örneklem $\mathbf{y}=\left(\mathbf{y}_1^T,\ldots\mathbf{y}_n^T\right)^T$ ile ifade edilsin. Burada $\mathbf{y}_j$, $\mathbf{Y}_j$ rasgele vektörünün gözlenen değeridir. Kabul edelim ki rastgele vektör $\mathbf{Y}_j$'nin olasılık yoğunluk fonksiyonun formu,

$$f(\mathbf{y}_j)=\sum_{i=1}^{g}\pi_i f_i(\mathbf{y}_j) \quad j=1,2,\ldots,n \tag{4.1}$$

biçiminde verilsin. Burada $f_i(\mathbf{y}_j)$ bir olasılık yoğunluk fonksiyonunu ve π_i katsayı değerleri negatif olmayan ve toplamları 1 olan oranları ifade etsin. π_i katsayı değerleri

$$0 \leq \pi_i \leq 1 \quad (i = 1,\ldots,g) \tag{4.2}$$

ve

$$\sum_{i=1}^{g} \pi_i = 1 \tag{4.3}$$

kısıtlarını sağlasın. Bu durum da $f_1(\mathbf{y}_j),\ldots,f_g(\mathbf{y}_j)$ fonksiyonları birer olasılık yoğunluk fonksiyonlarını ifade ettiği gibi (4.1)'deki eşitlik ile verilen $f(\mathbf{y}_j)$ fonksiyonu da bir olasılık yoğunluk fonksiyonunu ifade eder. Bu fonksiyon, karma olasılık yoğunluk fonksiyonu olarak adlandırılır. Karma olasılık yoğunluk fonksiyonunun (4.1)'deki eşitlik ile verilen ifadesinde g, karma olasılık yoğunluk fonksiyonun bileşen sayısını, π_i. bileşenin karma oranını ya da ağırlığını ve $f_i(\mathbf{y}_j)$ karma dağılımın i. bileşen olasılık yoğunluk fonksiyonunu ifade eder.

Karma dağılım modelinde bileşen sayısı g, sabit olarak yani önceden belirlenmiş bir değer olarak düşünülür. Fakat birçok uygulamada g değeri bilinmez ve veriden tahmin edilmek zorundadır.

4.1.2. Karma Dağılım Modelinin Parametrik Gösterimi

Birçok uygulamada karma dağılımın $f_i(\mathbf{y}_j)$ bileşen olasılık yoğunluk fonksiyonları belirli bir parametrik olasılık dağılım ailesinden alınır. Bu durumda $f_i(\mathbf{y}_j)$ bileşen olasılık yoğunluk fonksiyonu, $f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)$ şeklinde parametrik olarak ifade edilir. Burada $\boldsymbol{\theta}_i$ karma dağılım modelinin i. Bileşen olasılık yoğunluk

fonksiyonun bilinmeyen parametrelerini içeren bir vektördür. Karma olasılık dağılımı $f(\mathbf{y}_j)$ parametrik olarak,

$$f(\mathbf{y}_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j;\boldsymbol{\theta}_i) \tag{4.4}$$

şeklinde ifade edilir. Burada $\boldsymbol{\Psi}$, karma olasılık dağılım modelinde yer alan tüm bilinmeyen parametrelerin bir vektörüdür ve

$$\boldsymbol{\Psi} = \left(\pi_1,\ldots,\pi_{g-1},\boldsymbol{\xi}^T\right)^T \tag{4.5}$$

şeklinde açık olarak yazılır. Burada $\boldsymbol{\xi}$, $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_g$ bileşen dağılımların bilinmeyen parametre vektörlerini içeren bir vektördür.

4.1.3. Karma Modellerin Elde Edilmesi

Bir rastgele $\mathbf{Y}_j$ vektörü için g bileşenli ve $f(\mathbf{y}_j)$ ile ifade edilen bir karma olasılık yoğunluk fonksiyonu elde etmek amacıyla gerekli tanımlamalar ve kabuller aşağıdaki gibi yapılır.

Z_j, 1,2,…,g değerlerini sırasıyla $\pi_1,\pi_2,\ldots,\pi_g$ olasılıkları ile alan kategorik rastgele değişken olsun. $i = 1,2,\ldots,g$ olmak üzere $Z_j = i$ verildiğinde $\mathbf{Y}_j$ rastgele vektörünün koşullu olasılık yoğunluk fonksiyonu $f_i(\mathbf{y}_j)$ ile gösterilsin. Bu durumda $\mathbf{Y}_j$ rastgele vektörünün marjinal yoğunluk fonksiyonu olan koşulsuz olasılık yoğunluk fonksiyonu $f(\mathbf{y}_j)$ ile verilir. Burada kategorik Z_j rastgele değişkeni, $\mathbf{Y}_j$ değişken vektörünün bileşen etiketi gibi düşünülür. $\mathbf{Z}_j$, g boyutlu bileşen etiket vektörü olmak üzere i. elemanı $Z_{ij} = (Z_j)_i$ olup, j. rasgele değişken vektörü $\mathbf{Y}_j$ karma modelin i. bileşeninden elde edilmişse 1 diğer durumlarda 0 değerini alır. Bu durumda bileşen etiket vektörünü $\mathbf{Z}_j$'nin dağılımı $\pi_1,\ldots,\pi_g$ olasılıklı g farklı

kategori üzerinden bir çekim ile ilgili olacağından çok terimli bir dağılım olacaktır. Bu durumda $\mathbf{Z}_j$'nin olasılığı,

$$\Pr\{\mathbf{Z}_j = z_j\} = \pi_1^{z_{1j}}, \pi_2^{z_{2j}}, \ldots, \pi_g^{z_{gj}} \tag{4.6}$$

ve

$$\mathbf{Z}_j = Mult_g(1, \pi) \tag{4.7}$$

biçiminde yazılır. Burada $\pi = (\pi_1, \ldots, \pi_g)^T$ şeklinde ifade edilen vektördür. Buna göre (4.1)'deki eşitlikle verilen bir karma dağılım, $G_1, G_2, \ldots, G_g$ gibi g grubu sırasıyla $\pi_1, \ldots, \pi_g$ oranlarıyla içeren bir G kitlesinden elde edilmiş olan $\mathbf{Y}_j$ rastgele vektörünü modellemede kullanılabilir.

4.2. Modele Dayalı Kümeleme

Sonlu karma modeller gibi olasılık modellerine dayalı olarak kümeleme yöntemleri geliştirilebilir. Modele dayalı kümeleme de verinin olasılık dağılımların bir karmasından elde edildiği düşünülür. Diğer bir ifadeyle verinin, her bir bileşen dağılımı verideki bir kümeye karşılık gelen olasılık dağılımlarının bir karmasından üretildiği kabul edilir.

Modele dayalı kümelemede verideki karmaşık kümelemenin ve küme yapılarının modellenmesinde iki yaklaşım kullanılır. Bunlar sınıflandırma olabilirlik yaklaşımı ve karma olabilirlik yaklaşımıdır (Fraley 1998). Sınıflandırma olabilirlik fonksiyonu,

$$L_C(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_g; \gamma_1, \gamma_2, \ldots, \gamma_n \mid \mathbf{y}) = \prod_{j=1}^{n} f_{\gamma_j}(\mathbf{y}_j \mid \boldsymbol{\theta}_{\gamma_j}) \tag{4.8}$$

şeklinde ifade edilir. Burada γ_j , $\mathbf{y}_j$ rastgele vektörü karmanın i. bileşeninden elde edilmişse $\gamma_j = i$ değerini alan sınıf etiketleridir.

Karma olabilirlik yaklaşımında bir olasılık fonksiyonu ağırlıklandırılmış bileşen fonksiyonlarının toplamı şeklinde ifade edildiği kabul edilir. Kümeleme için karma olabilirlik yaklaşımı kullanılırsa kümeleme problemi karma dağılım modelinin parametrelerinin kestirimi problemine dönüşür. En çok olabilirlik fonksiyonu,

$$L_M(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_g; \pi_1, \ldots, \pi_g \mid \mathbf{y}) = \prod_{j=1}^{n} \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j \mid \boldsymbol{\theta}_i) \qquad (4.9)$$

şeklinde ifade edilir. Burada π_i, bir gözlemin *i.* bileşene ait olma olasılığıdır. $L_M(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_g; \pi_1, \ldots, \pi_g \mid \mathrm{y})$, (4.2) ve (4.3)'deki eşitliklerde verilen kısıtlara sahiptir. Karma olabilirlik yaklaşımında sonlu karma olasılık yoğunluklarının karmasının parametre tahminleri için en yaygın olarak kullanılan yöntem, EM (Demster ve ark. 1977) algoritmasıdır.

4.2.1. Karma Olabilirlik Yaklaşımında Çok Değişkenli Karma Dağılım Modelinde Parametre Tahmini

Karma olabilirlik yaklaşımı ile yapılan kümelemede karma dağılım modelindeki parametreleri tahmin etmek için genellikle EM algoritması kullanılır. EM algoritması verinin tamamlanmamış veri olması durumunda en çok olabilirlik kestirim için genel bir istatistiksel yöntemdir. EM algoritması döngüsel (iteratif) bir algoritmadır. EM algoritması E beklenti (Expectation) ve M maksimum yapma (maximization) adımlarından oluşur. Bu adımlar belirli bir yakınsama kriteri sağlanana kadar ardışık olarak gerçekleştirilir.

4.2.1.1. En Çok Olabilirlik Kestirimi

Bir $\mathbf{y} = \left(\mathbf{y}_1^T, \ldots \mathbf{y}_n^T\right)^T$ rastgele örnekleminin $f(\mathbf{y}_j; \boldsymbol{\Psi})$ olasılık yoğunluk fonksiyonun formu (4.4)'deki eşitlik ile verilen karma dağılım modeline sahip olsun. Gözlenmiş veriden elde edilen karma dağılım fonksiyonunun $\boldsymbol{\Psi}$ parametre vektörü için olabilirlik fonksiyonu,

$$L(\boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\mathbf{y}_j; \boldsymbol{\Psi}) \tag{4.10}$$

ile verilir. Logaritması alınmış olabilirlik fonksiyonu,

$$log\, L(\Psi) = \sum_{j=1}^{n} log\left(f(\mathrm{y}_j; \Psi)\right) = \sum_{j=1}^{n} log\left(\sum_{i=1}^{g} \pi_i f_i(\mathrm{y}_j; \boldsymbol{\theta}_i)\right) \tag{4.11}$$

şeklinde ifade edilir. (4.9)'deki eşitliğinin $\boldsymbol{\Psi}$ vektörüne göre türevi alınıp sıfıra eşitlenirse,

$$\frac{\partial \log L(\boldsymbol{\Psi})}{\partial \boldsymbol{\Psi}} = 0 \tag{4.12}$$

şeklinde elde edilen olabilirlik eşitliği bulunur. Olabilirlik eşitliğinin çözümü ile $\boldsymbol{\Psi}$ parametre vektörünün $\hat{\boldsymbol{\Psi}}$ ile gösterilen en çok olabilirlik kestirici elde edilir. Karma dağılım modeli içinde (4.12)'deki eşitliği $\boldsymbol{\Psi}$ vektörünün $\hat{\boldsymbol{\Psi}}$ ile gösterilen en çok olabilirlik kestiricisi,

$$\hat{\pi}_i = \frac{\sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \hat{\boldsymbol{\Psi}})}{n} \quad i = 1,2,\ldots,g \tag{4.13}$$

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\frac{\tau_i(\mathbf{y}_j;\hat{\boldsymbol{\Psi}})\partial\log f_i(\mathbf{y}_j,\boldsymbol{\theta}_i)}{\partial\boldsymbol{\xi}}=0 \tag{4.14}$$

eşitlikleri ile ifade edilen koşulları sağlayacak şekilde düzenlenir. Burada $\tau_i(\mathbf{y}_j;\hat{\boldsymbol{\Psi}})$, y_j gözleminin karmanın i. bileşenine sonraki (posterior) ait olma olasılığıdır. Sonraki (posterior) ait olma olasılığı,

$$\tau_i(\mathbf{y}_j;\hat{\boldsymbol{\Psi}})=\frac{\pi_i f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)}{\sum_{m=1}^{g}\pi_h f_h(\mathbf{y}_j;\boldsymbol{\theta}_h)} \tag{4.15}$$

biçiminde ifade edilir.

Bileşenleri belirli bir dağılım ailesinin üyesi olan karma dağılım modelinde olabilirlik eşitliğinin çözümü için çok sayıda çalışma yapılmıştır (Hasselblad 1966, 1969; Wolfe 1965, 1967, 1970; Day 1969). Yapılan bu çalışmalardan esinlenerek, (4.13) ve (4.14) eşitliklerini sağlayacak şekilde (4.12)'deki eşitlik ile verilen olabilirliğin döngüsel bir çözüm yöntemine bağlı olarak Dempster ve ark. (1977), EM algoritmasını önermişlerdir.

4.2.1.2. Karma Modellerde Tamamlanmamış Veri Yapısı

Bağımsız ve özdeş dağılımlı, n tane rastgele vektör $\mathbf{Y}_1,\ldots\mathbf{Y}_n$ ile gösterilsin ve bu rastgele vektörlerin aldıkları değerler sırasıyla $\mathbf{y}_1,\ldots\mathbf{y}_n$ ile ifade edilsin. $\mathbf{Y}_1,\ldots\mathbf{Y}_n$ rastgele vektörlerinin $f(\mathbf{y}_j)$ dağılım fonksiyonu, (4.1)'deki eşitlik ile verilen karma dağılım olasılık yoğunluk fonksiyonu olsun. Bu durumda

$$\mathbf{Y}_1,\ldots\mathbf{Y}_n \overset{böd}{\sim} \mathbf{F} \tag{4.16}$$

burada $\mathbf{F}(\mathbf{y}_j)$, $f(\mathbf{y}_j)$ olasılık yoğunluk fonksiyonuna karşılık gelen dağılım fonksiyonudur.

EM yapısında gözlenen $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)^T$ verisinde gözlemlerle ilişkilendirilen $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ bileşen etiket vektörleri bulunmadığından dolayı $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)^T$ verisi tamamlanmamış veri olarak düşünülür. Karma olabilirlik yaklaşımında verinin her bir $\mathbf{y}_j$ gözleminin karmanın hangi bileşenine ait olduğu önemlidir. $i = 1,2,\ldots,g$ ve $j = 1,2,\ldots,n$ olmak üzere $\mathbf{y}_j$ gözlemi karma dağılım modelinin *i.* bileşeninden gözlemlenmişse veya bu bileşenine ait ise g boyutlu $\mathbf{z}_j$ bileşen etiket vektörünün *i.* elamanı olan $z_{ij} = (\mathbf{z}_j)_i = 1$ değerini alırken diğer elamanları sıfır olacaktır. Bu durumda $\mathbf{z}_j$ bileşen etiket vektörleri ile ilişkilendirilen $\mathbf{y}_1, \ldots \mathbf{y}_n$ özellik verisi ile tamamlanmış veri olarak adlandırılır. Tamamlanmış veri,

$$\mathbf{y}_c = (\mathbf{y}^T, \mathbf{z}^T)^T \tag{4.17}$$

şeklinde gösterilir. Burada

$$\mathbf{y} = \left(\mathbf{y}_1^T, \ldots \mathbf{y}_n^T\right)^T \tag{4.18}$$

gözlenmiş veri vektörü ya da tamamlanmamış veri vektörünü ve

$$\mathbf{z} = \left(\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_n^T\right)^T \tag{4.19}$$

bileşen etiket değişkenlerinin vektörünü göstermektedir. $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ bileşen etiket vektörleri sırasıyla $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ rasgele vektörlerinin gözlenen değerleridir. Bağımsız özellik verisi için $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ rasgele vektörlerinin dağılımının,

$$\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n \overset{böd}{\sim} Mult_g(1, \pi) \tag{4.20}$$

şeklinde bağımsız ve özdeş çok terimli dağılım olduğu varsayılır. Bu varsayım $\mathbf{Y}_c$ tamamlanmış değişken veri vektörünün dağılımının $\mathbf{Y}$ tamamlanmamış değişken veri vektörünün uygun dağılımı olacağını belirtir. Bu durumda $\boldsymbol{\Psi}$ için logaritması alınmış tamamlanmış veri olabilirlik fonksiyonu,

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij}\left[\log \pi_i + \log f_i(\mathbf{y}_j; \boldsymbol{\theta}_i)\right] \tag{4.21}$$

şeklinde olur. Bu probleme EM algoritması uygulanırken z_{ij}'ler kayıp (missing) gözlemler olarak düşünülür.

4.2.1.3. EM algoritmasının E ve M adımları

Bu kesimde EM algoritmasının karma dağılım modeli için oluşturulan E ve M adımları incelenecektir.

E adımı: Tamamlanmış verinin logaritmik olabilirlik fonksiyonu, z_{ij} etiket değerleri cinsinden lineer olduğundan E adımda; $\mathbf{y}$ gözlenmiş değeri verilmişken $\boldsymbol{Z}_{ij}$ kategorik değişkeninin anlık koşullu beklenen değeri hesaplanır. Burada $\mathbf{Z}_{ij}$, z_{ij}'lere karşılık gelen rastgele değişkenlerdir. $\boldsymbol{\Psi}$ parametre vektörü için $\boldsymbol{\Psi}^{(0)}$ ile ifade edilen başlangıç değeri atanır. EM algoritmasının birinci döngüsünde E adımı için $\mathbf{y}$ verilmişken $\log L_c(\boldsymbol{\Psi})$ ifadesinin koşullu beklenen değeri, $\boldsymbol{\Psi}^{(0)}$ başlangıç değerleri ile

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(0)}) = E_{\boldsymbol{\Psi}^{(0)}}\left\{\log L_c(\boldsymbol{\Psi}) \mid y\right\} \tag{4.22}$$

biçiminde hesaplanır. EM algoritmasının $(k+1)$. döngüsündeki E adımında $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ ifadesinin bulunmasını gerektirir. Burada $\boldsymbol{\Psi}^{(k)}$, EM algoritmasının k. Adımında elde edilen $\boldsymbol{\Psi}$ vektörünün değeridir. EM algoritmasının $(k+1)$. döngüsündeki E adımında $i = 1,2,\ldots,g$ ve $j = 1,2,\ldots,n$ olmak üzere

$$E_{\Psi^{(k)}}(Z_{ij} \mid \mathrm{y}) = pr_{\Psi^{(k)}}\left\{Z_{ij} = 1 \mid \mathrm{y}\right\} = \tau_i(\mathrm{y}_j;\Psi^{(k)}) = \frac{\pi_i^{(k)} f_i(\mathrm{y}_j;\boldsymbol{\theta}_i^{(k)})}{\sum_{m=1}^{g} \pi_m^{(k)} f_m(\mathrm{y}_j;\boldsymbol{\theta}_i^{(k)})} \quad (4.23)$$

değeri hesaplanır. Burada $\tau_i(\mathbf{y}_j;\boldsymbol{\Psi}^{(k)})$ ifadesi j. gözlem vektörü $\mathbf{y}_j$'nin karmanın i. bileşenine sonraki ait olma olasılığıdır. (4.23)'deki eşitlik kullanılarak (4.21)'deki ifadenin $\mathbf{y}$ verilmişken koşullu beklenen değeri,

$$Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)}) = \sum_{i=1}^{g}\sum_{j=1}^{n} \tau_i(\mathbf{y}_j;\boldsymbol{\Psi}^{(k)})\left\{\log \pi_i + \log f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)\right\} \quad (4.24)$$

ile hesaplanır.

M adım: EM algoritmasının $(k+1)$. döngüsündeki M adımında $\boldsymbol{\Omega}$ parametre uzayında tanımlı $\boldsymbol{\Psi}$ vektörünün $Q(\boldsymbol{\Psi};\boldsymbol{\Psi}^{(k)})$ fonksiyonunu maksimum yapan $\boldsymbol{\Psi}^{(k+1)}$ güncellenmiş tahmin değeri bulunur. Sonlu karma olasılık dağılım modelinde π_i'nin $\pi_i^{(k+1)}$ güncel tahmini, bileşen yoğunluklarındaki bilinmeyen parametrelerin vektörü $\boldsymbol{\xi}$'nın güncellenmesinden bağımsız olarak yapılır.

z_{ij}'ler gözlenmiş olsaydı tamamlanmış veri için π_i'nin en çok olabilirlik tahmini,

$$\hat{\pi}_i = \sum_{j=1}^{n} \frac{z_{ij}}{n} \quad (i = 1,2,\ldots,g) \quad (4.25)$$

şeklinde bulunurdu. Tamamlanmış verinin logaritması alınmış olabilirliğinde EM algoritmasının E-adımında z_{ij} ifadesi yerine bu ifadenin anlık koşullu beklenen değerleri $\tau_i(\mathbf{y}_j;\boldsymbol{\Psi}^{(k)})$ kullanıldığı gibi benzer şekilde π_i'nin güncel tahmin değeri $\pi_i^{(k+1)}$ hesaplanırken (4.25)'deki eşitlikteki z_{ij} ifadesinin yerine $\tau_i(\mathbf{y}_j;\boldsymbol{\Psi}^{(k)})$ ifadesi kullanılır ve

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_i(\mathbf{y}_j;\boldsymbol{\Psi}^{(k)})}{n} \quad (i = 1,2,\ldots,g) \tag{4.26}$$

şeklinde elde edilir.

EM algoritmasının $(k+1)$. iterasyonundaki M-adımda $\boldsymbol{\xi}$'nın, $\boldsymbol{\xi}^{(k+1)}$ güncel değeri,

$$\sum_{i=1}^{g}\sum_{j=1}^{n} \frac{\tau_i(\mathbf{y}_j;\boldsymbol{\Psi}^{(k)})\partial \log f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)}{\partial \boldsymbol{\xi}} = \mathbf{0} \tag{4.27}$$

şeklinde elde edilen eşitliğin uygun bir köküne eşittir.

EM algoritmasında E ve M adımları bir yakınsama kriteri sağlanıncaya kadar tekrarlanır. Yakınsama için uygun bir durdurma kuralı olarak $\|.\|$ uygun bir norm olmak üzere $\left\|\boldsymbol{\Psi}^{(k+1)} - \boldsymbol{\Psi}^{(k)}\right)\| < \varepsilon$ ve $\varepsilon > 0$ olma koşulunu sağlayan bir ε değeri alınabilir veya $L(\boldsymbol{\Psi}^{(k+1)}) - L(\boldsymbol{\Psi}^{(k)})$ farkı oldukça küçükse ya da durağanlaşıyorsa algoritma sonlandırılır.

4.2.1.4. EM algoritmasının Özellikleri

EM algoritmasının bazı özellikleri aşağıdaki gibi sıralanabilir.

1. Her EM algoritmasının tekrarında olabilirlik fonksiyon değeri artar.
2. Genel düzgünlük koşulları altında güvenilir genel yakınsamaya sahiptir.

3. Hesaplamalar anlaşılır ve analitik olarak algoritmanın adımları kolay gerçekleştirilir.
4. EM algoritmasının tekrarlarındaki maliyet diğer hesaplama yöntemleri ile karşılaştırıldığında oldukça düşüktür.
5. Kayıp verinin tahminlerini de bulur.

EM algoritmasının bazı dezavantajları da vardır. Bunlar:

1. Parametre tahminlerinin kovaryans matrisinin bir tahminini otomatik olarak bulamaz.
2. Bazı durumlarda yakınsama oldukça yavaş olur.
3. Bazı problemlerde E ve M adımlar analitik olarak ifade edilemez.

4.2.2. Sınıflandırma En Çok Olabilirlik Yaklaşımı

Sınıflandırma en çok olabilirlik yaklaşımında $j = 1,2,\ldots,n$ ve $i = 1,2,\ldots,g$ olmak üzere z_{ij}, $\mathbf{y}_j$ gözlemi, karmanın *i*. bileşeninden elde ediliyorsa 1 diğer bileşenlerinden elde ediliyorsa sıfır değerini alan elemanlardan oluşan g boyutlu $\mathbf{z}_j$ 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, $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n$ 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'lerin eşit olduğu kabul edilir. Bu durumda kısıtlı sınıflandırma en çok olabilirlik kriteri,

$$L_{CR} = \sum_{i=1}^{g} \sum_{y_j \in P_i} \log\left(f(\mathbf{y}_j;\boldsymbol{\theta}_i)\right) \tag{4.28}$$

şeklinde yazılır. Burada $P=\left(P_1,P_2,\ldots,P_g\right)$, $\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n$ bileşen etiket vektörleriyle ilişkilendirilmiş $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n$ gözlemlerinin bir parçalanışıdır. Buna göre *i*. parçalanış $P_i=\left\{\mathbf{y}_j \mid z_{ij}=1\right\}$ olacak şekilde gözlem değerlerinden oluşur.

Karma örnekleme yapısı içerisinde, $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n$ örneklemi, (4.4)'deki eşitlikte verilen karma dağılımdan, bileşenlerinden alınan gözlemlerin sayıları örneklem hacmi n ve olasılık parametreleri $\pi_1,\pi_2,\ldots,\pi_g$ olan çok terimli bir dağılıma sahip olacak şekilde rastgele elde edilirler. Bu durumda sınıflandırma en çok olabilirlik kriteri (Symons 1981),

$$L_{CR}=\sum_{i=1}^{g}\sum_{x_j\in P_i}\log\left(\pi_i f(\mathbf{x}_j;\boldsymbol{\theta}_i)\right) \tag{4.29}$$

şeklinde tanımlanır. Burada (4.25) ve (4.26)'daki eşitliklerle verilen her iki sınıflandırma en çok olabilirlik kriteri iteratif parçalama algoritmaları ile optimize edilebilirler. Bu kriterlerin optimizasyonu için CEM, sınıflandırma EM (Classification EM) algoritması ve SEM, stokastik EM (Stochastic EM) algoritması kullanılabilir (Celeux ve Govaert 1992).

4.2.2.1. CEM Algoritması ve Özellikleri

CEM algoritması iteratif bir algoritma olup her bir iterasyon E, C ve M ile adlandırılan 3 adımdan oluşur. Algoritmanın adımları aşağıdaki gibi özetlenir.

Algoritmaya P^0 ile gösterilen verinin bir parçalanışının alınmasıyla başlar. Algoritmanın $m>0$ olmak üzere m. İterasyonda E, C ve M adımları aşağıdaki gibi uygulanır.

E-Adımı: $j=1,2,\ldots,n$ ve $i=1,2,\ldots,g$ olmak üzere $\mathbf{y}_j$ gözlemlerinin P_i. Parçalanışına sonraki ait olması olasılıkları anlık parametre tahminleri için

$$t_i^m(\mathbf{y}_j) = \frac{\pi_i^m f(\mathbf{y}_j, \boldsymbol{\theta}_i^m)}{\sum_{h=1}^{g} \pi_h^m f(\mathbf{y}_j, \boldsymbol{\theta}_h^m)} \tag{4.30}$$

şeklinde hesaplanır.

C-Adımı: Her bir $\mathbf{y}_j$ gözlemi $1 \le i \le g$ olmak üzere maksimum sonraki ait olma olasılığına sahip olacak şekilde kümeye atanır. Maksimum sonraki ait olması olasılığı tek değil ise en küçük indeks değerli kümeye gözlem atanır. Gözlemlerin kümelere atanma işleminden sonra elde edilen parçalanış P^m ile gösterir.

M-Adımı: $i = 1,2,\ldots,g$ olmak üzere $\left(\pi_i^{m+1}, \boldsymbol{\theta}_i^{m+1}\right)$ en çok olabilirlik kestirimleri P_i^m alt örneklemleri kullanılarak hesaplanır.

CEM sonlu sayıda iterasyon sonucu yakınsayan K-means algoritmasına benzer bir algoritmadır. CEM algoritması gözlenmiş olabilirlik fonksiyonunu maksimize etmez. Bunun yerine her örneklem noktasının kayıp bileşen etiketleri $\mathbf{z}_j$'lerinde veri kümesine dahil edildiği durumda

$$CL(\boldsymbol{\Psi} \mid \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n) = \sum_{i=1}^{g} \sum_{j \mid z_j = i} \log\left[\pi_i f(\mathbf{y}_j \mid \boldsymbol{\theta}_j)\right] \tag{4.31}$$

şeklinde ifade edilen logaritması alınan tamamlanmış olabilirlik fonksiyonunu $\boldsymbol{\theta}$ ve $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ parametreleri ile maksimize etmeyi amaçlar. Sonuç olarak CEM algoritması özellikle karma dağılımın bileşenleri birbiri üstüne katlandığı durumlarda, parametreler için yanlı kestirimlerde bulunur (Celeux ve Govaert 1993).

4.2.2.2. SEM Algoritması ve Özellikleri

SEM algoritması, EM algoritmasının bir stokastik biçimidir. EM algoritmasının E ve M adımları arasında $j = 1,2,\ldots,n$ olmak üzere $\mathbf{z}_j$ bilinmeyen bileşen etiket vektörlerini, onların anlık koşullu dağılımlarından rastgele olarak elde

ederek yeniden düzenleyen stokastik bir algoritmadır. Algoritmaya parametreler için $\boldsymbol{\theta}^0$ ile gösterilen değerler alınarak başlanır.

E-Adımı: $j = 1,2,\ldots,n$ ve $i = 1,2,\ldots,g$ olmak üzere $\boldsymbol{\theta}$ parametre vektörünün anlık değeri için $t_i^m(\mathbf{y}_j)$ koşullu olasılıkları EM algoritmasının E adımında olduğu gibi hesaplanır.

S-Adım: $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ gözlemlerinin $P^m = \left(P_1^m, P_{2,}^m \ldots, P_g^m\right)$ ile gösterilen bir P parçalanması her bir $\mathbf{y}_j$ gözleminin $t_i^m(\mathbf{y}_j)$ parametresi ile çok terimli dağılıma uygun olacak şekilde karmanın bir bileşenine rasgele olarak atanmasıyla elde edilir.

M-Adımı: $\boldsymbol{\theta}$ parametresinin en çok olabilirlik kestiricisi karmanın i. Bileşeninin bir alt örneklemi gibi düşünülen P_i^m parçalanma kümesi kullanılarak hesaplanır.

SEM veri ekleme tipli bir algoritmadır (Wei ve Tanner 1990). SEM noktasal olarak yakınsamaz (Celeux ve Govaert 1992).

4.2.3. Çok Değişkenli Normal Dağılımların Karması

Çok değişkenli Normal dağılımların karma yoğunluk fonksiyonu,

$$f(\mathbf{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \Phi_i(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{4.32}$$

ile verilir. Burada $i = 1,2,\ldots,g$ ve $j = 1,2,\ldots,n$ olmak üzere $\Phi_i(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ fonksiyonu ortalama vektörü $\boldsymbol{\mu}_i$ kovaryans matrisi $\boldsymbol{\Sigma}_i$ olan,

$$\Phi_i(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}} \left|\boldsymbol{\Sigma}_i\right|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_i)\right\}} \tag{4.33}$$

şeklinde ifade edilen çok değişkenli normal dağılım fonksiyonudur. Bu durumda karma modelin tüm bilinmeyen parametreler vektörü $\boldsymbol{\Psi} = \left(\pi_1, \ldots, \pi_{g-1}, \boldsymbol{\xi}^T\right)^T$ şeklinde yazılır. Burada $\boldsymbol{\xi}$, karma dağılım modelindeki bileşen olasılık yoğunluk

fonksiyonlarının parametreleri olan bileşen ortalama vektörleri $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_g)$ ve bileşen kovaryans matrisleri $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_g)$'dan oluşur. $i = 1,2,\ldots,g$ ve $j = 1,2,\ldots,n$ olmak üzere $\mathbf{y}_j$ gözlem vektörünün karmanın i. bileşenine sonraki ait olması olasılığı $\tau_i(\mathbf{y}_j; \boldsymbol{\Psi})$,

$$\tau_i(\mathbf{y}_j; \boldsymbol{\Psi}) = \frac{\pi_i \Phi(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{m=1}^{g} \pi_m \Phi(\mathbf{y}_j; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \tag{4.34}$$

şeklinde ifade edilir. EM algoritmasının $(k+1)$. iterasyonundaki M-adımında güncellenmiş karma oranları π_i, ve ortalama vektörü $\boldsymbol{\mu}_i$'nin en çok olabilirlik kestiricileri sırasıyla,

$$\pi_i^{(k+1)} = \sum_{j=1}^{n} \frac{\tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)})}{n} \tag{4.35}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)}) \mathbf{y}_j}{\sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)})} \tag{4.36}$$

ifadeleri ile bulunur. Bileşen olasılık yoğunluklarının kovaryans matrisi $\boldsymbol{\Sigma}_i$'nin güncel tahmini de

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)})(\mathbf{y}_j - \boldsymbol{\mu}_i^{(k+1)})(\mathbf{y}_j - \boldsymbol{\mu}_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)})} \quad (i = 1,2,\ldots,g) \tag{4.37}$$

ifadesi ile hesaplanır. Karma modeldeki parametrelerin en çok olabilirlik kestiricileri $W_{i1}^{(s)}$, $W_{i2}^{(s)}$ ve $W_{i3}^{(s)}$ yeter istatistikleri cinsinden de ifade edilebilir. Yeterli istatistikler,

$$W_{i1}^{(k)} = \sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)}) \tag{4.38}$$

$$W_{i2}^{(k)} = \sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)}) \mathbf{y}_j \tag{4.39}$$

$$W_{i3}^{(k)} = \sum_{j=1}^{n} \tau_i(\mathbf{y}_j; \boldsymbol{\Psi}^{(k)}) \mathbf{y}_j \mathbf{y}_j^T \tag{4.40}$$

şeklinde tanımlanmak üzere parametre tahminleri,

$$\pi_i^{(k+1)} = \frac{W_{i1}^{(s)}}{n} \tag{4.41}$$

$$\boldsymbol{\mu}_i^{(k+1)} = \frac{W_{i2}^{(s)}}{W_{i1}^{(s)}} \tag{4.42}$$

$$\boldsymbol{\Sigma}_i^{(k+1)} = \frac{W_{i3}^{(k)} - W_{i1}^{(k)^{-1}} W_{i2}^{(k)} W_{i2}^{(k)^T}}{W_{i1}^{(k)}} \tag{4.43}$$

eşitlikleri kullanılarak hesaplanır. Bazı uygulamalarda bileşen olasılık yoğunluklarının kovaryans matrislerinin,

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma} \quad (i = 1,2,\ldots,g) \tag{4.44}$$

şeklinde eşit olduğu kabul edilir. Burada $\boldsymbol{\Sigma}$ bilinmez. Bu durum, karma normal modelin bileşenlerinin homojen yapıda olduğunu gösterir. Ortak bileşen kovaryans matrisinin güncellenmiş tahmini,

$$\boldsymbol{\Sigma}^{(k+1)} = \sum_{i=1}^{g} \frac{W_{i1}^{(k)} \boldsymbol{\Sigma}_i^{(k+1)}}{n} \tag{4.45}$$

ile hesaplanır. Burada $\Sigma_i^{(k+1)}$, (4.43)'deki eşitlik ile bulunur.

Örnek 4.1: Fisher (1936), Iris verisi olarak da adlandırılan Iris (süsen) çiçek verisini ilk defa ayrıştırma analizi ile ilgili çalışmalarda kullanmıştır. Iris çiçek verisi, süsen çiçeğinin Iris setosa, Iris virginica ve Iris versicolor olarak adlandırılan 3 farklı türünü içerir. Her bir türe ait 4 özellik üzerinden ölçülmüş 50 gözlem değeri bulunur. Ölçüm yapılan özellikler çanak yaprağı uzunluğu, çanak yaprağı genişliği çiçek yaprağı uzunluğu ve çiçek yaprağı genişlikleridir. Bu veri için üç bileşenli karma normal dağılım modeli EM algoritması kullanılarak elde edilen parametre değerleri ile oluşturulacaktır. EM algoritması için parametre başlangıç değerleri Tablo 4.1 ile verilmiştir. Bu başlangıç değerleri alınarak uygulanan EM algoritmasındaki her bir döngü sonucu elde edilen logaritması alınmış olabilirlik fonksiyonun değerleri Tablo 4.2 ile verilmiştir.

Tablo 4.1. Üç bileşenli karma normal dağılımın parametrelerinin EM algoritması kullanılarak en çok olabilirlik kestirimi ile bulunmasında alınan parametre başlangıç değerleri.

Parametre	Birinci Bileşen	İkinci Bileşen	Üçüncü Bileşen
Ağırlık	1/8	4/8	3/8
Ortalama	$[5\ 3\ 1\ 0]^T$	$[5\ 2\ 4\ 1]^T$	$[7\ 3\ 6\ 2]^T$
Kovaryans	$\begin{bmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{bmatrix}$	$\begin{bmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{bmatrix}$	$\begin{bmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{bmatrix}$

Tablo 4.2 EM Algoritmasında döngü sayısına göre olabilirlik değerleri.

Döngü Sayısı	Log-likelihood Değerleri	Döngü Sayısı	Log-likelihood Değerleri
1	-289,089	9	-180,538
2	-208,409	10	-180,306
3	-203,025	11	-180,222
4	-198,333	12	-180,196
5	-192,471	13	-180,188
6	-186,291	14	-180,186
7	-182,424	15	-180,185
8	-181,072	16	-180,185

Tablo 4.2'den görüldüğü gibi logaritmik olabilirlik fonksiyonu 9. iterasyondan sonra birbirine oldukça yakın değerler alarak yakınsamaya başlamış ve 15. ve 16. iterasyonlarda eşit değerler almıştır. Bundan sonraki iterasyonları uygulamaya gerek kalmamıştır. EM algoritması kullanılarak elde edilen üç bileşenli karma normal dağılımın en çok olabilirlik kestirimleri Tablo 4.3'de verilmiştir. Bu algoritma için oluşturulan Matlab program kodu şekil 4.1 ile verilmiştir.

Tablo 4.3 Iris (süsen) çiçeği verisi (Fisher 1936) için oluşturulan karma dağılım modelindeki parametrelerin tahmin edilen değerleri.

Bileşen	Parametre	Tahmin Değerleri
1	Ağırlık	0.333
	Ortalama	$[5.006 \quad 3.428 \quad 1.462 \quad 0.246]^T$
	Kovaryans	$\begin{bmatrix} 0.121 & 0.097 & 0.016 & 0.010 \\ 0.097 & 0.140 & 0.011 & 0.009 \\ 0.016 & 0.011 & 0.029 & 0.006 \\ 0.010 & 0.009 & 0.006 & 0.010 \end{bmatrix}$
2	Ağırlık	0.298
	Ortalama	$[5.914 \quad 2.777 \quad 4.199 \quad 1.296]^T$
	Kovaryans	$\begin{bmatrix} 0.275 & 0.097 & 0.184 & 0.054 \\ 0.097 & 0.092 & 0.091 & 0.043 \\ 0.184 & 0.091 & 0.200 & 0.060 \\ 0.054 & 0.043 & 0.060 & 0.031 \end{bmatrix}$
3	Ağırlık	0.368
	Ortalama	$[6.542 \quad 2.948 \quad 5.476 \quad 1.982]^T$
	Kovaryans	$\begin{bmatrix} 0.387 & 0.092 & 0.303 & 0.062 \\ 0.092 & 0.110 & 0.084 & 0.056 \\ 0.303 & 0.084 & 0.329 & 0.075 \\ 0.062 & 0.056 & 0.075 & 0.086 \end{bmatrix}$

```
load ('data.mat', 'x');         % Iris verisi matlab editörüne yüklenir.
[p n]=size(x); nu1=[5 3 1 0]'; nu2=[5 2 4 1]'; nu3=[7 3 6 2 ]'; a1=1/8; a2=4/8; a3=3/8;
cov1=[1 0 0 0 ; 0 1 0 0 ; 0 0 1 0 ;  0 0 0 1]; cov2=[1 0 0 0 ; 0 1 0 0 ; 0 0 1 0 ; 0 0 0 1];
cov3=[1 0 0 0 ; 0 1 0 0 ; 0 0 1 0 ; 0 0 0 1]; j=1;

for i=1:n
   f1(i)=((2*pi)^(-p/2))*((det(cov1))^(-1/2))*exp((-1/2)*(x(:,i)-nu1)'*inv(cov1)*(x(:,i)-nu1));
   f2(i)=((2*pi)^(-p/2))*((det(cov2))^(-1/2))*exp((-1/2)*(x(:,i)-nu2)'*inv(cov2)*(x(:,i)-nu2));
   f3(i)=((2*pi)^(-p/2))*((det(cov3))^(-1/2))*exp((-1/2)*(x(:,i)-nu3)'*inv(cov3)*(x(:,i)-nu3));
   L(i)=a1*f1(i)+a2*f2(i)+a3*f3(i);
   i=i+1;
end

likelihoode=sum(L); likelihoody=likelihoode+0.0006;
while abs(likelihoody-likelihoode)>=0.0005
    likelihoode=likelihoody;

for i=1:n
   f1(i)=((2*pi)^(-p/2))*((det(cov1))^(-1/2))*exp((-1/2)*(x(:,i)-nu1)'*inv(cov1)*(x(:,i)-nu1));
   f2(i)=((2*pi)^(-p/2))*((det(cov2))^(-1/2))*exp((-1/2)*(x(:,i)-nu2)'*inv(cov2)*(x(:,i)-nu2));
   f3(i)=((2*pi)^(-p/2))*((det(cov3))^(-1/2))*exp((-1/2)*(x(:,i)-nu3)'*inv(cov3)*(x(:,i)-nu3));
   tau1(i)=a1*f1(i)/(a1*f1(i)+a2*f2(i)+a3*f3(i));
   tau2(i)=a2*f2(i)/(a1*f1(i)+a2*f2(i)+a3*f3(i));
   tau3(i)=a3*f3(i)/(a1*f1(i)+a2*f2(i)+a3*f3(i));
   pay1(i,:)=tau1(i)*x(:,i); pay2(i,:)=tau2(i)*x(:,i); pay3(i,:)=tau3(i)*x(:,i);
   varpay1(:,:,i)=tau1(i)*(x(:,i)-nu1)*(x(:,i)-nu1)'; varpay2(:,:,i)=tau2(i)*(x(:,i)-nu2)*(x(:,i)-nu2)';
   varpay3(:,:,i)=tau3(i)*(x(:,i)-nu3)*(x(:,i)-nu3)';     i=i+1;
end
a1y=sum(tau1)/n; a2y=sum(tau2)/n; a3y=sum(tau3)/n;
nu1y=(sum(pay1)/sum(tau1))'; nu2y=(sum(pay2)/sum(tau2))'; nu3y=(sum(pay3)/sum(tau3))';
b=[0 0 0 0;0 0 0 0;0 0 0 0;0 0 0 0]; c=[0 0 0 0;0 0 0 0;0 0 0 0;0 0 0 0];
d=[0 0 0 0;0 0 0 0;0 0 0 0;0 0 0 0];
for i=1:n
   b=varpay1(:,:,i)+b; c=varpay2(:,:,i)+c; d=varpay3(:,:,i)+d; i=i+1;
end
cov1y=b/sum(tau1); cov2y=c/sum(tau2); cov3y=d/sum(tau3);
for i=1:n
   f1y(i)=((2*pi)^(-p/2))*((det(cov1y))^(-1/2))*exp((-1/2)*(x(:,i)-nu1y)'*inv(cov1y)*(x(:,i)-nu1y));
   f2y(i)=((2*pi)^(-p/2))*((det(cov2y))^(-1/2))*exp((-1/2)*(x(:,i)-nu2y)'*inv(cov2y)*(x(:,i)-nu2y));
   f3y(i)=((2*pi)^(-p/2))*((det(cov3y))^(-1/2))*exp((-1/2)*(x(:,i)-nu3y)'*inv(cov3y)*(x(:,i)-nu3y));
   Ly(i)=a1y*f1y(i)+a2y*f2y(i)+a3y*f3y(i);     i=i+1;
end
likelihoody=sum(Ly);
   a1=a1y;  a2=a2y; a3=a3y;     nu1=nu1y; nu2=nu2y; nu3=nu3y;
   cov1=cov1y; cov2=cov2y; cov3=cov3y;     j=j+1;
end
```

Şekil 4.1. Üç bileşenli karma normal dağılımın EM algoritması kullanılarak parametrelerinin en çok olabilirlik kestiricilerinin bulunması için oluşturulan Matlab kodu.

4.2.4. Modele Dayalı Yığılmalı Hiyerarşik Kümeleme

Modele dayalı yığılmalı (agglomerative) hiyerarşik kümeleme, modele dayalı kümelemede küme sayısına göre parametre başlangıç değerlerinin bulunması amacıyla kullanılır. Modele dayalı yığılmalı (agglomerative) hiyerarşik kümeleme, Bölüm 3.3.1 ile verilen yığılmalı (agglomerative) hiyerarşik kümeleme yapısına benzer. Bununla birlikte yığılmalı (agglomerative) hiyerarşik kümeleme yönteminde olduğu gibi bir uzaklık tanımlanmaz. Modele dayalı yığılmalı (agglomerative) hiyerarşik kümelemede amaç fonksiyonu olarak sınıflandırma olabilirlik fonksiyonu kullanılır.

Modele dayalı yığılmalı (agglomerative) hiyerarşik kümeleme için sınıflandırma olabilirlik fonksiyonu,

$$L_{CL}(\boldsymbol{\theta}_i, \gamma_j \mid \mathbf{y}_j) = \prod_{j=1}^{n_{\gamma_j}} f_{\gamma_j}(\mathbf{y}_j \mid \boldsymbol{\theta}_{\gamma_j}) \tag{4.46}$$

biçiminde tanımlanır (Fraley 1998). Burada γ_j, $\mathbf{y}_j$ gözlem vektörünün ait olduğu, elde edildiği, çekildiği ya da gözlemlendiği kümenin indeks değeri ve n_{γ_j} ise j. bileşendeki gözlemlerin toplam sayısıdır. Modele dayalı yığılmalı (agglomerative) hiyerarşik kümeleme yönteminde sınıflandırma olabilirlik fonksiyonunu maksimum yapacak şekilde parametre değerleri bulunmak istenir. Yığılmalı (agglomerative) hiyerarşik kümelemede olduğu gibi başlangıçta her bir gözlemin ayrı bir kümede olduğu yani her bir gözlemin bir küme olduğu düşünülür. Sınıflandırma olabilirlik fonksiyonunda maksimum artışı sağlayan gözlem (küme) çiftleri birleştirilir. Bu şeklinde tüm gözlemler tek bir küme içerisinde yer alıncaya kadar kümelemeye devam edilir.

Fraley (1998), modele dayalı hiyerarşik kümeleme yapısında başlangıçta her bir gözlemin ayrı bir kümede yani her bir gözlemin bir küme olması düşüncesi yerine gözlemlerin belirli bir parçalanışı ile kümeleme yöntemine başlanabileceğini önermiştir. Çok değişkenli normal dağılımların karmasına dayalı olarak yapılan

yığılmalı (agglomerative) hiyerarşik kümeleme için Fraley (1998), Banfield ve Raftery (1993) tarafından önerilen bileşen kümelerinin kovaryans matris yapısına göre dört farklı kriter önermiştir. Çok değişkenli karma model yapısında bileşen olasılık yoğunluğu $f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)$, çok değişkenli normal olasılık yoğunluk fonksiyonu olarak alınırsa (4.46)'daki eşitlik ile verilen sınıflandırma olabilirlik fonksiyonu,

$$L_{CL}(\mu_1,\mu_2,\ldots,\mu_g;\Sigma_1,\Sigma_2,\ldots,\Sigma_g,\gamma \mid \mathrm{y}) = \prod_{i=1}^{g}\prod_{j\in I_i}(2\pi)^{-\frac{p}{2}}\left|\Sigma_i\right|^{-\frac{1}{2}}e^{\left\{-\frac{1}{2}(\mathrm{y}_j-\mu_i)^T\Sigma_i^{-1}(\mathrm{y}_j-\mu_i)\right\}} \quad (4.47)$$

biçiminde ifade edilir. Burada I_i, i. kümede yer alan gözlemlere ait indis kümesidir. $I_i = \{j : \gamma_j = i\}$ şeklinde açık olarak ifade edilir. (4.47)'deki eşitlikte bileşen ortalama vektörü $\boldsymbol{\mu}_i$'nin en çok olabilirlik kestiricisi, $\hat{\boldsymbol{\mu}}_i$,

$$\hat{\mu}_i = \overline{\mathrm{y}}_i = \frac{s_i}{n_i} \quad (4.48)$$

şeklinde küme ortalamasına eşittir. Burada $\overline{\mathbf{y}}_i$, i. küme ortalamasıdır. s_i, i. kümedeki gözlem değerlerinin toplamıdır. n_i, i. kümedeki gözlemlerin toplam sayısıdır. (4.47)'deki eşitlikte $\boldsymbol{\mu}_i$ yerine en çok olabilirlik kestiricisi $\hat{\boldsymbol{\mu}}_i = \overline{\mathbf{y}}_i$ ifadesi kullanılırsa yoğunlaştırılmış (concentrated) log olabilirlik fonksiyonu,

$$L_{CL}(\Sigma_1,\Sigma_2,\ldots,\Sigma_g,\gamma \mid \mathrm{y};\hat{\mu}_1,\hat{\mu}_2,\ldots,\hat{\mu}_g) = -\frac{np\log(2\pi)}{2} - \frac{1}{2}\sum_{i=1}^{g}\left\{Iz(\mathbf{W}_i\boldsymbol{\Sigma}_i^{-1}) + n_i\log\left|\boldsymbol{\Sigma}_i\right|\right\} \quad (4.49)$$

biçiminde ifade edilir. Burada $\mathbf{W}_i$ küme içi saçılım matrisi olup, G_i, i. kümeyi temsil etmek üzere

$$\mathbf{W}_i = \sum_{\mathbf{y} \in G_i} (\mathbf{y} - \overline{\mathbf{y}}_i)(\mathbf{y} - \overline{\mathbf{y}}_i)^T \tag{4.50}$$

ifadesi ile hesaplanır. Yoğunlaştırılmış (concentrated) log olabilirlik fonksiyonunun (4.49)'daki eşitlik ile verilen formunda bileşen varyans kovaryans matrisi $\mathbf{\Sigma}_i$ ve indeks vektörü $\boldsymbol{\gamma}$ bilinmeyen parametreler olarak kalır. (4.49)'daki eşitlik ile verilen olabilirlik fonksiyonunu maksimum yapabilmenin bir yolu Fraley (1998) tarafından önerilen normal karma dağılım modelindeki bileşen dağılımlarının dört farklı kovaryans yapısına göre iki kümeyi birleştirmek için tanımladığı kriterleri minimum yapmaktır. Bu modeller ve kriterler Tablo 4.4'de verilmiştir.

Tablo 4.4. Normal karma modellerde bileşen kovaryans matrisinin parametrik yapısı ve minimum yapılacak kriterler.

Bileşen Kovaryans Modeli ($\mathbf{\Sigma}_i$)	Kriter
$\sigma^2\mathbf{I}$	$\dot{I}z\left(\sum_{i=1}^{g}\mathbf{W}_i\right)$
$\sigma_i^2\mathbf{I}$	$\sum_{i=1}^{g} n_i \log\left[\dot{I}z\left(\frac{\mathbf{W}_i}{n_i}\right)\right]$
$\mathbf{\Sigma}$	$\left\lvert\sum_{i=1}^{g}\mathbf{W}_i\right\rvert$
$\mathbf{\Sigma}_i$	$\sum_{i=1}^{g} n_i \log\left\lvert\frac{\mathbf{W}_i}{n_i}\right\rvert$

$\mathbf{\Sigma}_i = \sigma^2\mathbf{I}$ **modeli:** Bu modelde kovaryans matrisi, karma normal dağılım modelindeki tüm bileşenler için eşit ve köşegen olacak şekilde kısıtlanmıştır. Bu model için minimum yapılacak olan kriter,

$$\dot{I}z\left(\sum_{i=1}^{g}\mathbf{W}_i\right) = \sum_{i=1}^{g}\dot{I}z(\mathbf{W}_i) \tag{4.51}$$

biçiminde ifade edilir. Modele dayalı eklemeli hiyerarşik kümeleme de G_i ve G_k ile temsil edilen iki kümeyi birleştirmede kullanılacak olan maliyet fonksiyonu,

$$\begin{aligned} C(i,k) &= \dot{I}z(\mathbf{W}_{\langle i,k\rangle}) - \dot{I}z(\mathbf{W}_i) - \dot{I}z(\mathbf{W}_k) \\ &= \dot{I}z(\mathbf{W}_{\langle i,k\rangle} - \mathbf{W}_i - \dot{I}z\mathbf{W}_k) \\ &= \mathbf{w}_{ik}^T \mathbf{w}_{ki} \end{aligned} \tag{4.52}$$

şeklinde tanımlanır. Burada $\mathbf{w}_{ik} = r_{ki}s_i - r_{ik}s_k$ ve $r_{ik} = \sqrt{\dfrac{n_i}{n_k(n_i+n_k)}}$ ile hesaplanır. Amaç fonksiyonunu minimum yapan kümeler algoritmanın her adımında birleştirilir.

$\boldsymbol{\Sigma}_i = \sigma_i^2\mathbf{I}$ modeli: Bu modelde kovaryans matrisi karma normal dağılım modelindeki her bileşen dağılım için köşegen yapıda olmakla birlikte tüm kümeler için eşit olma koşulu yoktur. Bu model için minimum yapılacak olan kriter,

$$\sum_{i=1}^{g} n_i \log\left[\dot{I}z\left(\frac{\mathbf{W}_i}{n_i}\right)\right] \tag{4.53}$$

biçiminde ifade edilir. Bu kriter için iki kümeyi birleştirmede kullanılacak olan maliyet fonksiyonu,

$$\begin{aligned} C(i,k) = (n_i+n_k)\log\left(\frac{\dot{I}z(\mathrm{W}_i) - \dot{I}z(\mathrm{W}_k) + \mathrm{w}_{ik}^T\mathrm{w}_{ik}}{n_i+n_k}\right) \\ - n_i\log\left(\frac{\dot{I}z(\mathbf{W}_i)}{n_i}\right) - n_k\log\left(\frac{\dot{I}z(\mathbf{W}_k)}{n_k}\right) \end{aligned} \tag{4.54}$$

şeklinde verilir.

$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ Modeli: Bu modelde karma normal dağılım modelindeki her bileşen dağılım için kovaryans matrisi için eşit olma koşulu vardır. Bu model için minimum yapılacak olan kriter,

$$\left|\sum_{i=1}^{g}\mathbf{W}_i\right| \tag{4.55}$$

biçiminde ifade edilir. Herhangi bir maliyet fonksiyonuna ihtiyaç yoktur.

Kısıtsız model: Bu model de normal karma dağılım modelindeki her bileşen dağılım için kovaryans matrisi üzerinde herhangi bir kısıt bulunmaz ve kümeler için eşit olma koşulu yoktur. Bu model için minimum yapılacak olan kriter,

$$\sum_{i=1}^{g} n_i \log\left|\frac{\mathbf{W}_i}{n_i}\right| \tag{4.56}$$

biçimimde ifade edilir.

4.2.5. Çok Değişkenli Karma Normal Modeller

Celeux ve Govaert (1995), modele dayalı kümeleme analizi için çok değişkenli karma normal dağılımları önermiştir. Burada model kelimesi karma dağılımın bileşen olasılık yoğunluk fonksiyonlarındaki kovaryans matrisi için kabul edilen bazı kısıtlamaları ve kovaryans matrisinin geometrik özellikleri temsil edilir (Martinez ve Martinez 2005). Celeux ve Govaert (1995), çok değişkenli normal dağılımların bileşen dağılımlarındaki kovaryans matrisi $\mathbf{\Sigma}_i$'yi,

$$\mathbf{\Sigma}_i = \lambda_i \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T \tag{4.57}$$

şeklinde özdeğer ayrışımı ile ifade etmiştir. Burada $\lambda_i = \left|\mathbf{\Sigma}_i\right|^{1/p}$, p değişken sayısını göstermek üzere $\mathbf{D}_i$, $\mathbf{\Sigma}_i$'nin özvektörlerinin matrisi, $\mathbf{A}_i$, köşegen bir matris olup, köşegen üzerindeki elamanları $\mathbf{\Sigma}_i$ bileşen kovaryans matrisinin normalleştirilmiş

özdeğerlerini azalan sırada içeriri ve determinantı $|\mathbf{A}_i| = 1$'dir. λ_i, i-inci kümenin hacmini, $\mathbf{D}_i$, yönünü ve $\mathbf{A}_i$ şeklini belirler.

Modellerin yorumlanması daha kolay hale getiren bu karakteristiklerin tümü aynı anda olmamak üzere λ_i, $\mathbf{D}_i$ ve $\mathbf{A}_i$ parametreleri üzerindeki kabul varyasyonları sonucu Celeux ve Govaert (1995), farklı 14 varyans kovaryans matris yapısı önermişlerdir. Bu 14 farklı kovaryans matris yapısına göre de karma dağılımdaki bileşen olasılıklarının eşit veya farklı alınması durumu da dikkate alınarak 28 çok değişkenli karma normal modeli elde edilir. Bu modeller üç kategori altında toplanır. Bu kategoriler genel, köşegen ve küresel varyans kovaryans model aileleri olarak adlandırılır. 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 4.5 - Tablo 4.7 ile verilmiştir.

Tablo 4.5. Karma normal modellerin genel bileşen varyans-kovaryans yapıları.

Model	Hacmi	Şekli	Yönü
$\mathbf{\Sigma}_i = \lambda \mathbf{D}\mathbf{A}\mathbf{D}^T$	Sabit	Sabit	Sabit
$\mathbf{\Sigma}_i = \lambda_i \mathbf{D}\mathbf{A}\mathbf{D}^T$	Farklı	Sabit	Sabit
$\mathbf{\Sigma}_i = \lambda \mathbf{D}\mathbf{A}_i\mathbf{D}^T$	Sabit	Farklı	Sabit
$\mathbf{\Sigma}_i = \lambda \mathbf{D}_i\mathbf{A}\mathbf{D}_i^T$	Sabit	Sabit	Farklı
$\mathbf{\Sigma}_i = \lambda_i \mathbf{D}\mathbf{A}_i\mathbf{D}^T$	Farklı	Farklı	Sabit
$\mathbf{\Sigma}_i = \lambda_i \mathbf{D}_i\mathbf{A}\mathbf{D}_i^T$	Farklı	Sabit	Farklı
$\mathbf{\Sigma}_i = \lambda \mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T$	Sabit	Farklı	Farklı
$\mathbf{\Sigma}_i = \lambda_i \mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T$	Farklı	Farklı	Farklı

Tablo 4.6. $\mathbf{B}$ köşegen bir matris olmak üzere karma normal modellerin köşegen bileşen varyans-kovaryans yapıları.

Model	Hacmi	Şekli	Yönü
$\mathbf{\Sigma}_i = \lambda \mathbf{B}$	Sabit	Sabit	Eksen
$\mathbf{\Sigma}_i = \lambda_i \mathbf{B}$	Farklı	Sabit	Eksen
$\mathbf{\Sigma}_i = \lambda \mathbf{B}_i$	Sabit	Farklı	Eksen
$\mathbf{\Sigma}_i = \lambda_i \mathbf{B}_i$	Farklı	Farklı	Eksen

Tablo 4.7. I birim matris olmak üzere karma normal modellerin küresel bileşen varyans-kovaryans yapıları.

Model	Hacmi	Şekli	Yönü
$\boldsymbol{\Sigma}_i = \lambda \mathbf{I}$	Sabit	Sabit	----------
$\boldsymbol{\Sigma}_i = \lambda_i \mathbf{I}$	Farklı	Sabit	----------

4.2.5.1. Çok Değişkenli Karma Normal Modellerin Grafiksel Gösterimi İçin Uygulama

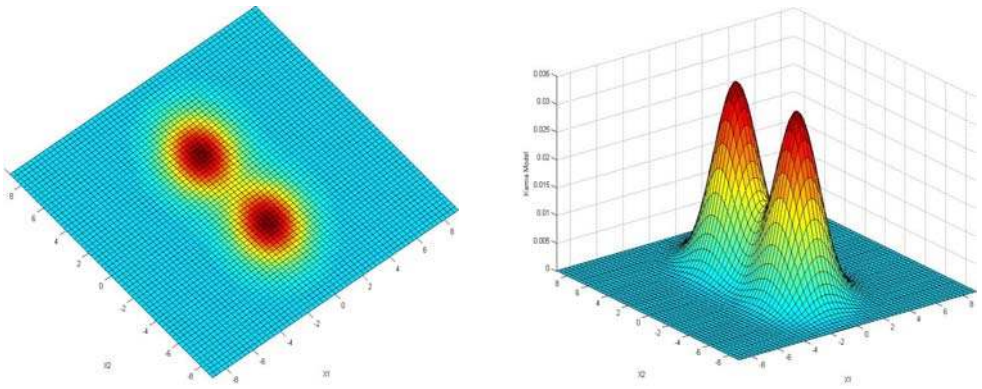
Çok değişkenli normal dağılımların karmasında bileşen olasılık yoğunluk fonksiyonlarının Celeux ve Govaert (1995) tarafından önerilen 14 farklı kovaryans matris yapısına göre oluşabilecek karma normal modeller iki değişken ve iki bileşen durumlarında gösterilecektir. Burada bileşen oranları eşit alınmıştır.

1. $[\pi\lambda\mathbf{C}]$ **modeli için grafiksel gösterim:** $[\pi\lambda\mathbf{C}]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.8'de verilmiştir. $[\pi\lambda\mathbf{C}]$ modelinde $\mathbf{C} = \mathbf{DAD}^T$ ve $\boldsymbol{\Sigma} = \lambda\mathbf{C}$ dir.

Tablo 4.8. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{C}]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	Birinci Bileşen ($i=1$)	İkinci Bileşen ($i=2$)
π_i	0.5	0.5
μ_i	$\begin{bmatrix}0 & -3\end{bmatrix}^T$	$\begin{bmatrix}0 & 3\end{bmatrix}^T$
Σ_i	$\begin{bmatrix}2 & 0.5\\ 0.5 & 3\end{bmatrix}$	$\begin{bmatrix}2 & 0.5\\ 0.5 & 3\end{bmatrix}$
λ_i	2.3979	2.3979
D_i	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$
A_i	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$

Tablo 4.8'deki parametre değerleri ile $[\pi\lambda\mathbf{C}]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.2'de gösterilmiştir.



(a) Kontur grafiği **(b)** Olasılık yoğunluk fonksiyon grafiği

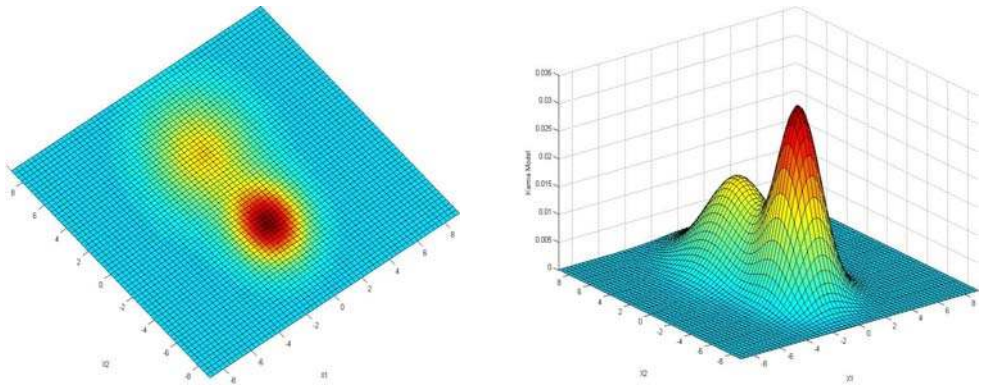
Şekil 4.2. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{C}]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

2. $[\pi\lambda_i\mathbf{C}]$ **modeli için grafiksel gösterim:** $[\pi\lambda_i\mathbf{C}]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.9'da verilmiştir. $[\pi\lambda_i\mathbf{C}]$ modelinde $\mathbf{C}=\mathbf{DAD}^T$ ve $\mathbf{\Sigma}_i=\lambda_i\mathbf{C}$, $i=1,2$ dir.

Tablo 4.9. Bileşen Kovaryans matrisleri $[\pi\lambda_i\mathbf{C}]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
μ_i	$[0 \quad -3]^T$	$[0 \quad 3]^T$
Σ_i	$\begin{bmatrix} 2 & 0.5 \\ 0.5 & 3 \end{bmatrix}$	$\begin{bmatrix} 4.1703 & 1.0426 \\ 1.0426 & 6.25542 \end{bmatrix}$
λ_i	2.3979	5
D_i	$\begin{bmatrix} 0.3827 & -0.9239 \\ 0.9229 & 0.3827 \end{bmatrix}$	$\begin{bmatrix} 0.3827 & -0.9239 \\ 0.9229 & 0.3827 \end{bmatrix}$
A_i	$\begin{bmatrix} 1.3375 & 0 \\ 0 & 0.7477 \end{bmatrix}$	$\begin{bmatrix} 1.3375 & 0 \\ 0 & 0.7477 \end{bmatrix}$

Tablo 4.9'daki parametre değerleri ile $[\pi\lambda_i \mathbf{C}]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.3'te gösterilmiştir.



(a) Kontur grafiği **(b)** Olasılık yoğunluk fonksiyon grafiği

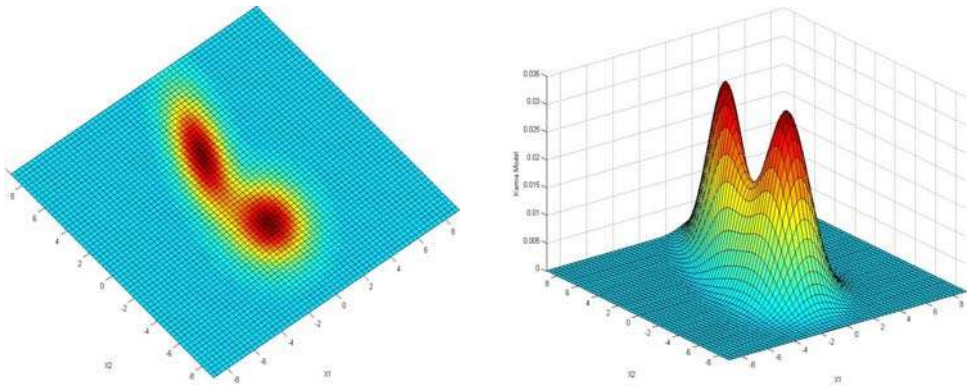
Şekil 4.3. Bileşen Kovaryans matrisleri $[\pi\lambda_i \mathbf{C}]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

3. $[\pi\lambda \mathrm{DA}_i \mathrm{D}^T]$ **modeli için grafiksel gösterim:** $[\pi\lambda \mathrm{DA}_i \mathrm{D}^T]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.10'da verilmiştir. $[\pi\lambda \mathrm{DA}_i \mathrm{D}^T]$ modelinde $\mathbf{\Sigma}_i = \lambda \mathbf{D}\mathbf{A}_i\mathbf{D}^T$, $i = 1,2$ dir.

Tablo 4.10 Bileşen Kovaryans matrisleri $[\pi\lambda \mathbf{D}\mathbf{A}_i\mathbf{D}^T]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	Birinci Bileşen ($i=1$)	İkinci Bileşen ($i=2$)
π_i	0.5	0.5
μ_i	$[0 \quad -3]^T$	$[0 \quad 3]^T$
Σ_i	$\begin{bmatrix} 2 & 0.5 \\ 0.5 & 3 \end{bmatrix}$	$\begin{bmatrix} 1.7357 & 2.2608 \\ 2.2608 & 6.2573 \end{bmatrix}$
λ_i	2.3979	2.3979
D_i	$\begin{bmatrix} 0.3827 & -0.9239 \\ 0.9229 & 0.3827 \end{bmatrix}$	$\begin{bmatrix} 0.3827 & -0.9239 \\ 0.9229 & 0.3827 \end{bmatrix}$
A_i	$\begin{bmatrix} 1.3375 & 0 \\ 0 & 0.7477 \end{bmatrix}$	$\begin{bmatrix} 3 & 0 \\ 0 & 0.3333 \end{bmatrix}$

Tablo 4.10'daki parametre değerleri ile $\left[\pi\lambda\mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.4'de gösterilmiştir.



(a) Kontur grafiği **(b)** Olasılık yoğunluk fonksiyon grafiği

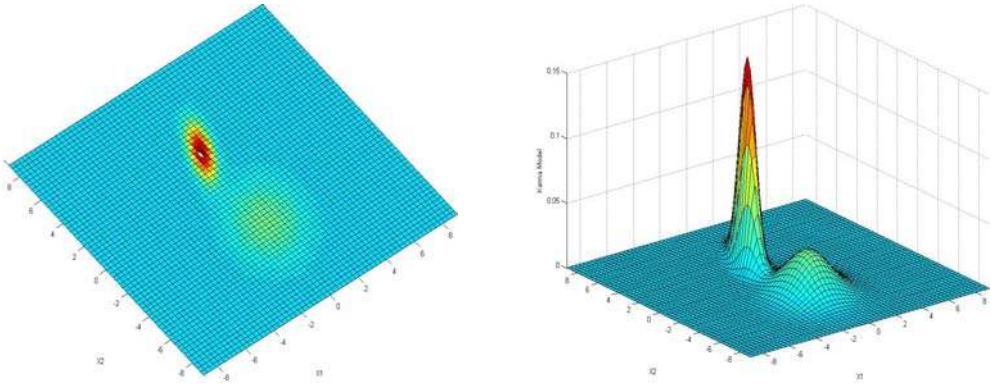
Şekil 4.4. Bileşen Kovaryans matrisleri $\left[\pi\lambda\mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

4. $\left[\pi\lambda_i\mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ **modeli için grafiksel gösterim:** $\left[\pi\lambda_i\mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.11'de verilmiştir. $\left[\pi\lambda_i\mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ modelinde $\mathbf{\Sigma}_i = \lambda_i\mathbf{D}\mathbf{A}_i\mathbf{D}^T$, $i = 1,2$.

Tablo 4.11 Bileşen Kovaryans matrisleri $\left[\pi\lambda_i\mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
μ_i	$\begin{bmatrix}0 & -3\end{bmatrix}^T$	$\begin{bmatrix}0 & 3\end{bmatrix}^T$
Σ_i	$\begin{bmatrix}2 & 0.5\\ 0.5 & 3\end{bmatrix}$	$\begin{bmatrix}0.3619 & 0.4714\\ 0.4714 & 1.3047\end{bmatrix}$
λ_i	2.3979	5
D_i	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$
A_i	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$	$\begin{bmatrix}3 & 0\\ 0 & 0.3333\end{bmatrix}$

Tablo 4.11'daki parametre değerleri ile $\left[\pi\lambda_i \mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.5'de gösterilmiştir.



(a) Kontur grafiği **(b)** Olasılık yoğunluk fonksiyon grafiği

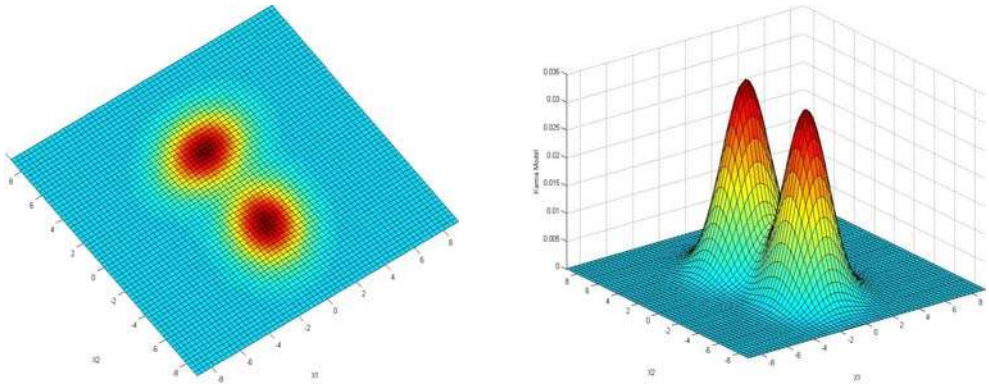
Şekil 4.5. Bileşen Kovaryans matrisleri $\left[\pi\lambda_i \mathbf{D}\mathbf{A}_i\mathbf{D}^T\right]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

5. $\left[\pi\lambda\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeli için grafiksel gösterim: $\left[\pi\lambda\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.12'de verilmiştir. $\left[\pi\lambda\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modelinde $\boldsymbol{\Sigma}_i = \lambda\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T$, $i = 1,2$ dir.

Tablo 4.12 Bileşen Kovaryans matrisleri $\left[\pi\lambda\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
μ_i	$\begin{bmatrix}0 & -3\end{bmatrix}^T$	$\begin{bmatrix}0 & 3\end{bmatrix}^T$
Σ_i	$\begin{bmatrix}2 & 0.5\\ 0.5 & 3\end{bmatrix}$	$\begin{bmatrix}2.9384 & 0.5548\\ 0.5548 & 2.0616\end{bmatrix}$
λ_i	2.3979	2.3979
D_i	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$	$\begin{bmatrix}-0.9 & 0.4359\\ -0.4359 & -0.9\end{bmatrix}$
A_i	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$

Tablo 4.12'daki parametre değerleri ile $\left[\pi\lambda\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.6'da gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

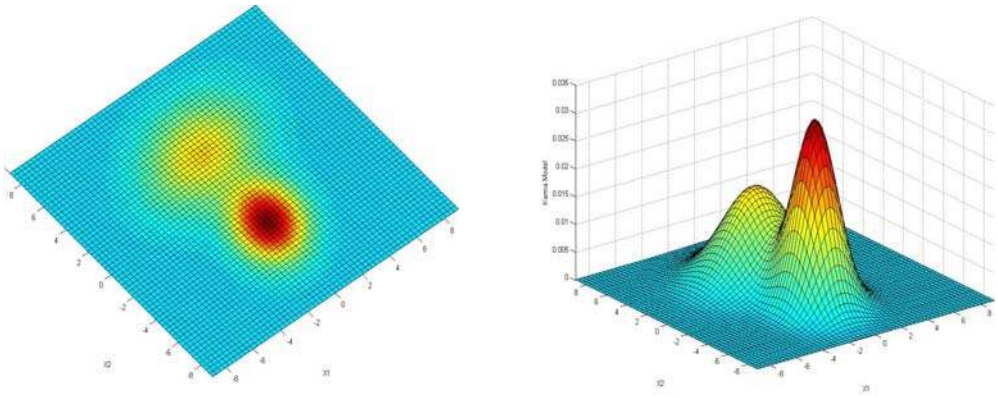
Şekil 4.6. Bileşen Kovaryans matrisleri $\left[\pi\lambda\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

6. $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeli için grafiksel gösterim: $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.13'te verilmiştir. $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modelinde $\mathbf{\Sigma}_i = \lambda_i\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T$, $i = 1,2$.

Tablo 4.13 Bileşen Kovaryans matrisleri $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}\mathbf{D}_i^T\right]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
μ_i	$\begin{bmatrix}0 & -3\end{bmatrix}^T$	$\begin{bmatrix}0 & 3\end{bmatrix}^T$
Σ_i	$\begin{bmatrix}2 & 0.5\\ 0.5 & 3\end{bmatrix}$	$\begin{bmatrix}6.127 & 1.1569\\ 1.1569 & 4.2988\end{bmatrix}$
λ_i	2.3979	5
D_i	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$	$\begin{bmatrix}-0.9 & 0.4359\\ -0.4359 & -0.9\end{bmatrix}$
A_i	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$

Tablo 4.13'teki parametre değerleri ile $\left[\pi\lambda_i \mathbf{D}_i \mathbf{A} \mathbf{D}_i^T\right]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.7'de gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

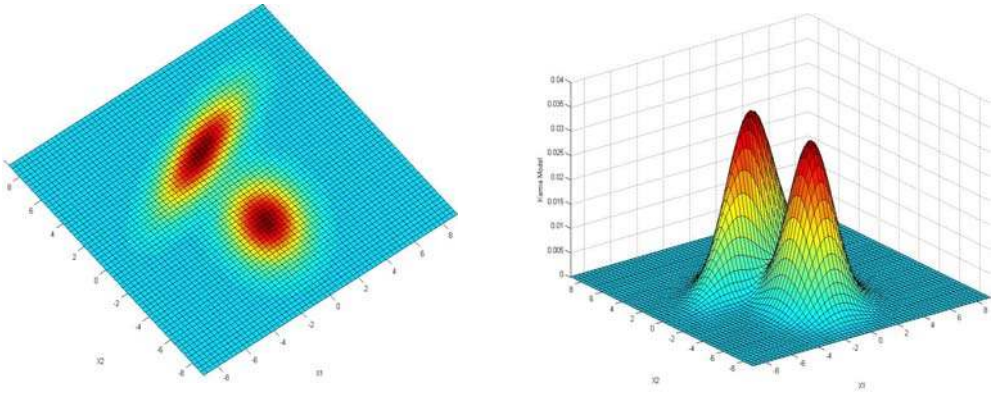
Şekil 4.7. Bileşen Kovaryans matrisleri $\left[\pi\lambda_i \mathbf{D}_i \mathbf{A} \mathbf{D}_i^T\right]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

7. $\left[\pi\lambda \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T\right]=\left[\pi\lambda \mathbf{C}_i\right]$ modeli için grafiksel gösterim: $\left[\pi\lambda \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T\right]=\left[\pi\lambda \mathbf{C}_i\right]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.14'te verilmiştir. $\left[\pi\lambda \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T\right]=\left[\pi\lambda \mathbf{C}_i\right]$ modelinde $\mathbf{C}_i = \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T$ ve $\mathbf{\Sigma}_i = \lambda \mathbf{C}_i$, $i = 1,2$ dir.

Tablo 4.14. Bileşen Kovaryans matrisleri $\left[\pi\lambda \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T\right]=\left[\pi\lambda \mathbf{C}_i\right]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
μ_i	$\begin{bmatrix}0 & -3\end{bmatrix}^T$	$\begin{bmatrix}0 & 3\end{bmatrix}^T$
Σ_i	$\begin{bmatrix}2 & 0.5\\ 0.5 & 3\end{bmatrix}$	$\begin{bmatrix}5.9788 & 2.5086\\ 2.5086 & 2.0143\end{bmatrix}$
λ_i	2.3979	2.3979
D_i	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$	$\begin{bmatrix}-0.9 & 0.4359\\ -0.4359 & -0.9\end{bmatrix}$
A_i	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$	$\begin{bmatrix}3 & 0\\ 0 & 0.3333\end{bmatrix}$

Tablo 4.14'deki parametre değerleri ile $\left[\pi\lambda\mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T\right]=\left[\pi\lambda\mathbf{C}_i\right]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.8'de gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

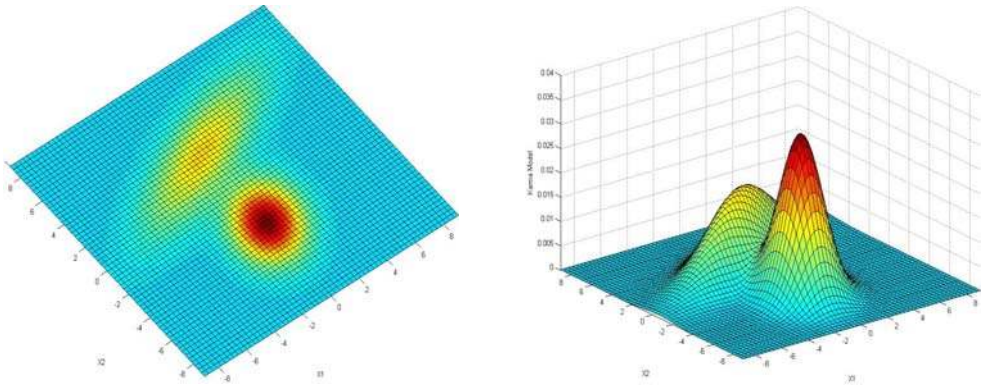
Şekil 4.8. Bileşen Kovaryans matrisleri $\left[\pi\lambda\mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T\right]=\left[\pi\lambda\mathbf{C}_i\right]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

8. $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T\right]=\left[\pi\lambda_i\mathbf{C}_i\right]$ modeli için grafiksel gösterim: $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T\right]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.15'te verilmiştir. $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T\right]$ modelinde $\mathbf{C}_i=\mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T$ ve $\mathbf{\Sigma}_i=\lambda_i\mathbf{C}_i$, $i=1,2$ dir.

Tablo 4.15. Bileşen Kovaryans matrisleri $\left[\pi\lambda_i\mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T\right]=\left[\pi\lambda_i\mathbf{C}_i\right]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
μ_i	$\begin{bmatrix}0 & -3\end{bmatrix}^T$	$\begin{bmatrix}0 & 3\end{bmatrix}^T$
Σ_i	$\begin{bmatrix}2 & 0.5\\ 0.5 & 3\end{bmatrix}$	$\begin{bmatrix}12.4667 & 5.2308\\ 5.2308 & 4.2001\end{bmatrix}$
λ_i	2.3979	5
D_i	$\begin{bmatrix}0.3827 & -0.9239\\ 0.9229 & 0.3827\end{bmatrix}$	$\begin{bmatrix}-0.9 & 0.4359\\ -0.4359 & -0.9\end{bmatrix}$
A_i	$\begin{bmatrix}1.3375 & 0\\ 0 & 0.7477\end{bmatrix}$	$\begin{bmatrix}3 & 0\\ 0 & 0.3333\end{bmatrix}$

Tablo 4.15'teki parametre değerleri ile $[\pi\lambda_i \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T]=[\pi\lambda_i \mathbf{C}_i]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.9'da gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

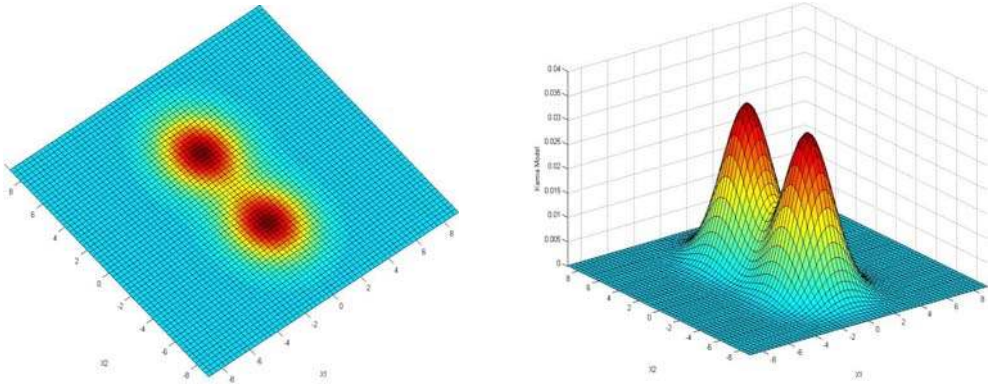
Şekil 4.9. Bileşen Kovaryans matrisleri $[\pi\lambda_i \mathbf{D}_i \mathbf{A}_i \mathbf{D}_i^T]=[\pi\lambda_i \mathbf{C}_i]$ modeline sahip karma normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

9. $[\pi\lambda\mathbf{B}]$ modeli için grafiksel gösterim: $[\pi\lambda\mathbf{B}]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.16'da verilmiştir. $[\pi\lambda\mathbf{B}]$ modelinde $\mathbf{B}=\mathbf{DAD}^T$ ve $\boldsymbol{\Sigma}_i=\lambda\mathbf{B}$, $i=1,2$ dir.

Tablo 4.16. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{B}]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
$\boldsymbol{\mu}_i$	$[0 \quad -3]^T$	$[0 \quad 3]^T$
$\boldsymbol{\Sigma}_i$	$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$
λ_i	2.4495	2.4495
$\mathbf{D}_i$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
$\mathbf{A}_i$	$\begin{bmatrix} 1.2247 & 0 \\ 0 & 0.8165 \end{bmatrix}$	$\begin{bmatrix} 1.2247 & 0 \\ 0 & 0.8165 \end{bmatrix}$
$\mathbf{B}_i$	$\begin{bmatrix} 0.8165 & 0 \\ 0 & 1.2247 \end{bmatrix}$	$\begin{bmatrix} 0.8165 & 0 \\ 0 & 1.2247 \end{bmatrix}$

Tablo 4.16'daki parametre değerleri ile $[\pi\lambda\mathbf{B}]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.10'da gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

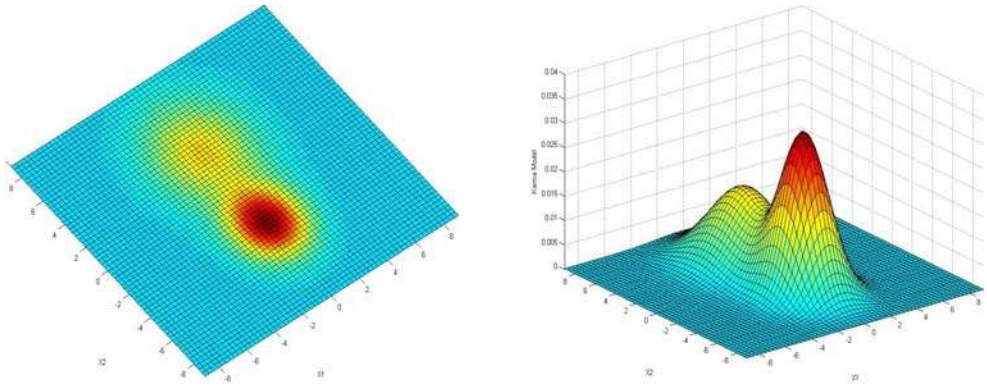
Şekil 4.10. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{B}]$ modeline sahip karma normal normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

10. $[\pi\lambda_i\mathbf{B}]$ **modeli için grafiksel gösterim:** $[\pi\lambda_i\mathbf{B}]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.17'de verilmiştir. $[\pi\lambda_i\mathbf{B}]$ modelinde $\mathbf{B}=\mathbf{DAD}^T$ ve $\mathbf{\Sigma}_i=\lambda_i\mathbf{B}$, $i=1,2$ dir.

Tablo 4.17'deki parametre değerleri ile $[\pi\lambda_i\mathbf{B}]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.11'de gösterilmiştir.

Tablo 4.17. Bileşen Kovaryans matrisleri $[\pi\lambda_i\mathbf{B}]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	Birinci Bileşen ($i=1$)	İkinci Bileşen ($i=2$)
π_i	0.5	0.5
$\boldsymbol{\mu}_i$	$[0 \quad -3]^T$	$[0 \quad 3]^T$
$\boldsymbol{\Sigma}_i$	$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	$\begin{bmatrix} 4.0825 & 0 \\ 0 & 6.1237 \end{bmatrix}$
λ_i	2.4495	5
$\mathbf{D}_i$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
$\mathbf{A}_i$	$\begin{bmatrix} 1.2247 & 0 \\ 0 & 0.8165 \end{bmatrix}$	$\begin{bmatrix} 1.2247 & 0 \\ 0 & 0.8165 \end{bmatrix}$
$\mathbf{B}_i$	$\begin{bmatrix} 0.8165 & 0 \\ 0 & 1.2247 \end{bmatrix}$	$\begin{bmatrix} 0.8165 & 0 \\ 0 & 1.2247 \end{bmatrix}$



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

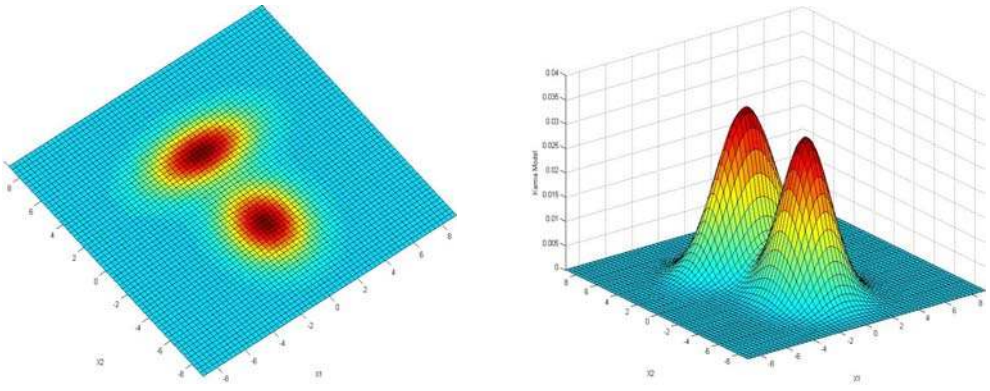
Şekil 4.11. Bileşen Kovaryans matrisleri $[\pi\lambda_i\mathbf{B}]$ modeline sahip karma normal normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

11. $[\pi\lambda\mathbf{B}_i]$ modeli için grafiksel gösterim: $[\pi\lambda_i\mathbf{B}]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.18'de verilmiştir. $[\pi\lambda\mathbf{B}_i]$ modelinde $\mathbf{B}_i = \mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T$ ve $\boldsymbol{\Sigma}_i = \lambda\mathbf{B}_i$, $i = 1,2$ dir.

Tablo 4.18. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{B}_i]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	Birinci Bileşen ($i=1$)	İkinci Bileşen ($i=2$)
π_i	0.5	0.5
$\boldsymbol{\mu}_i$	$[0 \quad -3]^T$	$[0 \quad 3]^T$
$\boldsymbol{\Sigma}_i$	$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	$\begin{bmatrix} 4.2966 & 0 \\ 0 & 1.23965 \end{bmatrix}$
λ_i	2.4495	2.4495
$\mathbf{D}_i$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$\mathbf{A}_i$	$\begin{bmatrix} 1.2247 & 0 \\ 0 & 0.8165 \end{bmatrix}$	$\begin{bmatrix} 1.7541 & 0 \\ 0 & 0.5701 \end{bmatrix}$
$\mathbf{B}_i$	$\begin{bmatrix} 0.8165 & 0 \\ 0 & 1.2247 \end{bmatrix}$	$\begin{bmatrix} 1.7541 & 0 \\ 0 & 0.5701 \end{bmatrix}$

Tablo 4.18'deki parametre değerleri ile $[\pi\lambda\mathbf{B}_i]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.12'de gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

Şekil 4.12. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{B}_i]$ modeline sahip karma normal normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

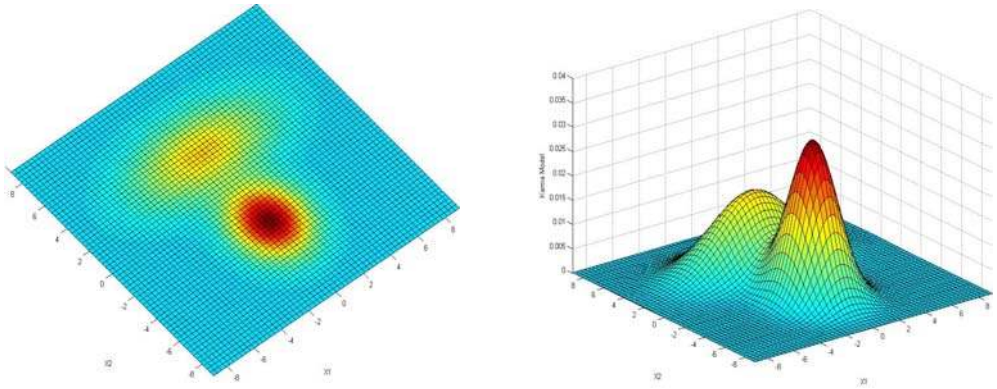
12. $[\pi\lambda_i\mathbf{B}_i]$ modeli için grafiksel gösterim: $[\pi\lambda_i\mathbf{B}_i]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre

değerleri Tablo 4.19'da verilmiştir. $[\pi\lambda_i\mathbf{B}_i]$ modelinde $\mathbf{B}_i = \mathbf{D}_i\mathbf{A}_i\mathbf{D}_i^T$ ve $\mathbf{\Sigma}_i = \lambda_i\mathbf{B}_i$, $i = 1,2$ dir.

Tablo 4.19. Bileşen Kovaryans matrisleri $[\pi\lambda_i\mathbf{B}_i]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
$\boldsymbol{\mu}_i$	$[0 \quad -3]^T$	$[0 \quad 3]^T$
$\mathbf{\Sigma}_i$	$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$	$\begin{bmatrix} 8.7705 & 0 \\ 0 & 2.8505 \end{bmatrix}$
λ_i	2.4495	5
$\mathbf{D}_i$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$\mathbf{A}_i$	$\begin{bmatrix} 1.2247 & 0 \\ 0 & 0.8165 \end{bmatrix}$	$\begin{bmatrix} 1.7541 & 0 \\ 0 & 0.5701 \end{bmatrix}$
$\mathbf{B}_i$	$\begin{bmatrix} 0.8165 & 0 \\ 0 & 1.2247 \end{bmatrix}$	$\begin{bmatrix} 1.7541 & 0 \\ 0 & 0.5701 \end{bmatrix}$

Tablo 4.19'daki parametre değerleri ile $[\pi\lambda_i\mathbf{B}_i]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.13'te gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

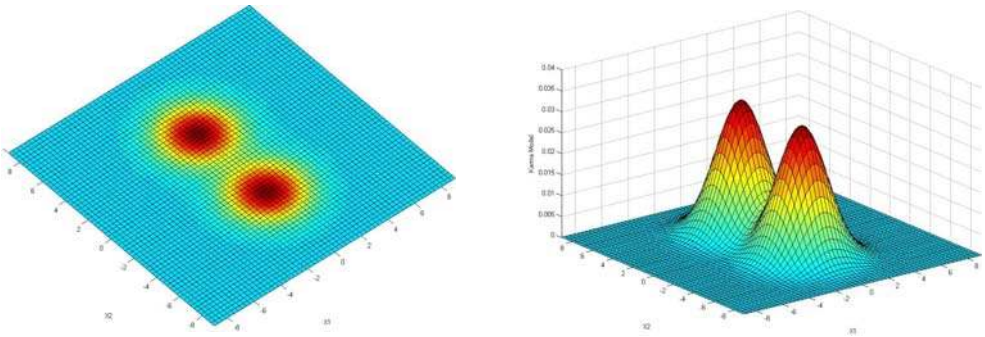
Şekil 4.13. Bileşen Kovaryans matrisleri $[\pi\lambda_i\mathbf{B}_i]$ modeline sahip karma normal normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

13. $[\pi\lambda\mathbf{I}]$ modeli için grafiksel gösterim: $[\pi\lambda\mathbf{I}]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.20'de verilmiştir. $[\pi\lambda\mathbf{I}]$ modelinde $\mathbf{I}=\mathbf{DAD}^T$ ve $\boldsymbol{\Sigma}_i=\lambda\mathbf{I}$, $i=1,2$ dir.

Tablo 4.20. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{I}]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen ($i=1$)**	**İkinci Bileşen ($i=2$)**
π_i	0.5	0.5
$\boldsymbol{\mu}_i$	$[0 \quad -3]^T$	$[0 \quad 3]^T$
$\boldsymbol{\Sigma}_i$	$\begin{bmatrix} 2.5 & 0 \\ 0 & 2.5 \end{bmatrix}$	$\begin{bmatrix} 2.5 & 0 \\ 0 & 2.5 \end{bmatrix}$
λ_i	2.5	2.5
$\mathbf{D}_i$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$
$\mathbf{A}_i$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

Tablo 4.20'deki parametre değerleri ile $[\pi\lambda\mathbf{I}]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.14'de gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

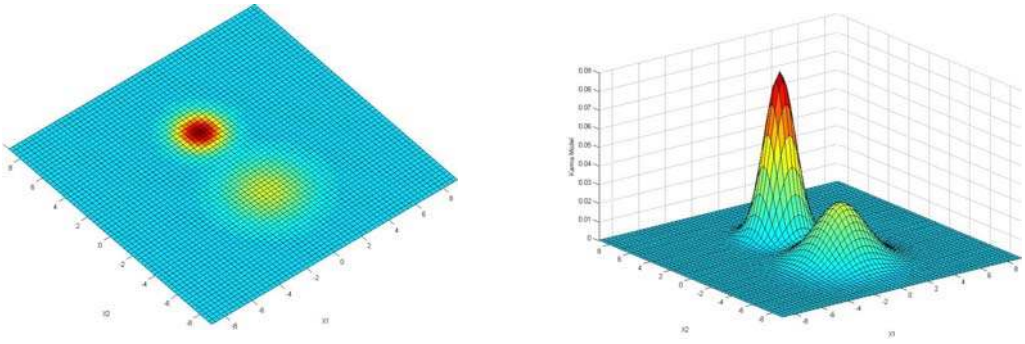
Şekil 4.14. Bileşen Kovaryans matrisleri $[\pi\lambda\mathbf{I}]$ modeline sahip karma normal normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

14. $[\pi\lambda_i\mathbf{I}]$ **modeli için grafiksel gösterim:** $[\pi\lambda_i\mathbf{I}]$ modeline sahip olacak şekilde alınan karma normal dağılım modelinde bileşenler ve bileşenlerin parametre değerleri Tablo 4.21'de verilmiştir. $[\pi\lambda_i\mathbf{I}]$ modelinde $\mathbf{I}=\mathbf{DAD}^T$ ve $\mathbf{\Sigma}_i=\lambda_i\mathbf{I}$, $i=1,2$ dir.

Tablo 4.21. Bileşen Kovaryans matrisleri $[\pi\lambda_i\mathbf{I}]$ modeline sahip olacak şekilde alınan karma normal dağılımın parametre değerleri.

Parametreler	**Birinci Bileşen** ($i=1$)	**İkinci Bileşen** ($i=2$)
π_i	0.5	0.5
$\boldsymbol{\mu}_i$	$\begin{bmatrix}0 & -3\end{bmatrix}^T$	$\begin{bmatrix}0 & 3\end{bmatrix}^T$
$\mathbf{\Sigma}_i$	$\begin{bmatrix}2.5 & 0\\ 0 & 2.5\end{bmatrix}$	$\begin{bmatrix}0.9 & 0\\ 0 & 0.9\end{bmatrix}$
λ_i	2.5	0.9
$\mathbf{D}_i$	$\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$	$\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$
$\mathbf{A}_i$	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$	$\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$

Tablo 4.21'deki parametre değerleri ile $[\pi\lambda_i\mathbf{I}]$ modeline sahip karma normal dağılımın kontur grafiği ve olasılık yoğunluk fonksiyon grafiği Şekil 4.15'te gösterilmiştir.



(**a**) Kontur grafiği (**b**) Olasılık yoğunluk fonksiyon grafiği

Şekil 4.15. Bileşen kovaryans matrisleri $[\pi\lambda_i\mathbf{I}]$ modeline sahip karma normal normal dağılımın (a) Kontur grafiği ve (b) Olasılık yoğunluk fonksiyon grafiği.

5. KARMA KÜMELEME ANALİZİNDE MODEL SEÇİM KRİTERLERİ

Çok değişkenli karma dağılım modeline dayalı kümeleme analizinde model seçimi genellikle olabilirlik fonksiyonuna dayalı olarak yapılmaktadır. Olabilirlik fonksiyonuna ceza terimi eklenmesiyle oluşan form kullanılır. Bu yapıda bir karmaya eklenen bir bileşenle büyüyen olabilirlik değeri, modele eklenen bu bileşendeki bilinmeyen parametrelerle artan ceza teriminin farkının alınmasıyla hesaplanır. Olabilirlik fonksiyonu yerine genellikle logaritmik olabilirlik fonksiyonu kullanılır. Ceza terimli logaritmik olabilirlik fonksiyonuna dayalı olarak yapılan model seçimi Bilgi kriterleri olarak adlandırılır.

5.1. Model Seçimi için Bayes Yaklaşımı

Bayes yaklaşımının model seçiminde kullanılması Jeffreys (1935, 1939) çalışmaları ile başlamıştır. Kass ve Raftery (1995), Bayes çarpanlarının model seçimi için farklı alanlarda uygulanabilirliğini göstermiştir.

Bayes yaklaşımında M_1 ve M_2 ile gösterilen iki model durumunda gözlenen $\mathbf{y}$ verisinin M_1 modelinden veya M_2 modelinden çekildiği kabul edilir. $p(M_1)$ ve $p(M_2)$ önsel (prior) olasılıklar olmak üzere, Bayes teoremine göre gözlenen $\mathbf{y}$ verisi verilmişken M_i modelinin doğru olması hipotezinin sonraki olasılığı,

$$p(M_i \mid \mathbf{y}) = \frac{p(M_i)p(y \mid M_i)}{p(M_1)p(y \mid M_1) + p(M_2)p(y \mid M_2)} \quad i = 1,2\,. \tag{5.1}$$

ile verilir. Burada iki modelin sonraki olasılıklarında paydadaki terimleri aynı olacağından bu iki modelin sonraki olasılıklarının oranı,

$$\frac{p(M_1 \mid \mathbf{y})}{p(M_2 \mid \mathbf{y})} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \cdot \frac{p(M_1)}{p(M_2)} \tag{5.2}$$

ile verilir. (5.2)'deki eşitliğinin sağ tarafındaki ilk çarpan Bayes çarpanı olarak adlandırılır. Bu iki model için Bayes çarpanı B_{12} olmak üzere,

$$B_{12} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \tag{5.3}$$

şeklinde ifade edilir. M_i modeli bilinmeyen parametreler içeriyorsa sonraki olasılık $p(\mathbf{y} \mid M_i)$,

$$p(\mathbf{y} \mid M_i) = \int p(\mathbf{y} \mid \boldsymbol{\theta_i}, M_i) p(\boldsymbol{\theta_i} \mid M_i) d\boldsymbol{\theta_i} \tag{5.4}$$

ile bulunur. Bu eşitliğe M_i modelinin integrallenmiş olabilirliği denir. Gözlenen $\mathbf{y}$ verisi verilmişken en büyük olabilirlik değerine sahip model doğru model olarak seçilir. $p(M_1)$ ve $p(M_2)$ olasılıkları eşit olması durumunda (5.2)'deki eşitlikten,

$$\frac{p(M_1 \mid \mathbf{y})}{p(M_2 \mid \mathbf{y}} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \tag{5.5}$$

olarak elde edilir. Bu durumda en yüksek integrallenmiş olabilirlik değerine sahip model en iyi model olarak seçilir. Model sayısı ikiden fazla durumlar içinde bir Bayes çözüm mümkündür. (Dasgupta ve Raftery 1998, Fraley ve Raftery, 2002).

5.1.1. Bayesci Bilgi Kriteri

İntegrallenebilir olabilirlik fonksiyonu modellerin önsel (prior) olasılıklarına bağlı olduğundan, modeller için logaritmik integrallenebilir olabilirlik fonksiyonun değeri yaklaşık olarak bulunabilir. Bu yaklaşık değer, kısaca Bayesci bilgi kriteri ya da kısaca BIC (Bayesian Information Criterion) ile ifade edilir (Schwarz 1978). Sonlu karma dağılım modeli, bileşen sayısı g ve bu g bileşenin bilinmeyen

parametrelerini içeren $\boldsymbol{\Psi} = \{\pi_1,\ldots,\pi_g,\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_g\}$ vektörü ile belirlenir. Sonlu karma dağılım modellerinde model seçimi ile gözlenen $\mathbf{y}$ verisinin kümelenme yapısını en iyi belirleyen modelin bulunması amaçlanır.

Sonlu karma dağılım modellerinde model seçimi için klasik bir yöntem, integrallenebilir olabilirlik fonksiyonu $f(\mathbf{y} \mid M, g)$'nin maksimum yapılmasıdır (Mclachlan ve Peel 2000).

$$(\hat{M}, \hat{g}) = \arg\max_{M,g} f(\mathbf{y} \mid M, g) \tag{5.6}$$

integrallenebilir olabilirlik fonksiyonu,

$$f(\mathbf{y} \mid M, g) = \int_{\Theta_{M,g}} f(\mathbf{y} \mid M, g, \boldsymbol{\Psi}) g(\boldsymbol{\Psi} \mid M, g) d\boldsymbol{\Psi} \tag{5.7}$$

şeklinde tanımlanır. Burada $f(\mathbf{y} \mid M, g, \boldsymbol{\Psi})$,

$$f(\mathbf{y} \mid M, g, \boldsymbol{\Psi}) = \prod_{j=1}^{n} f(\mathbf{y}_j \mid M, g, \boldsymbol{\Psi}) \tag{5.8}$$

dir. Burada $\Theta_{M,g}$, bileşen sayısı g olan M modelinin parametre uzayını, $g(\boldsymbol{\Psi} \mid M, g)$ ise bu modelde $\boldsymbol{\Psi}$ için eğitici olmayan ya da çok az eğitici olan önsel dağılımı göstermektedir. Düzgünlük koşulları altında integrallenebilir olabilirlik fonksiyonunun asimptotik değeri için bir yaklaşım Schwarz (1978) tarafından

$$\log f(\mathbf{y} \mid M, g) \approx \log f(\mathbf{y} \mid M, v, \hat{\boldsymbol{\Psi}}) - \frac{\nu_{M,g}}{2} \log(n) \tag{5.9}$$

olarak önerilmiştir. Burada $\hat{\boldsymbol{\Psi}}$, $\boldsymbol{\Psi}$'nın en çok olabilirlik tahminidir. $\hat{\boldsymbol{\Psi}}$,

$$\hat{\boldsymbol{\Psi}} = \arg\max_{\boldsymbol{\Psi}} f(\mathbf{y} \mid M, g, \boldsymbol{\Psi}) \tag{5.10}$$

şeklinde tanımlanır. (5.9)'daki eşitlikte $\nu_{M,g}$, bileşen sayısı g olan M modelindeki serbest parametre sayısını, n ise gözlenen verinin gözlem sayısını ifade eder. (5.9)'daki eşitlik ile verilen ifade BIC kriteri olarak adlandırılır. Karma dağılım modelindeki serbest parametre sayısı $\nu_{M,g} = d$ olarak alınırsa BIC kriteri,

$$BIC_{M,g} = -2L_{M,g} + d\ln n \tag{5.11}$$

ile verilir. Burada $L_{M,g}$, M ve g için logaritmik en çok olabilirlik fonksiyonudur. BIC kriterini minimum yapan model, en iyi model olarak seçilir. BIC kriteri, karma dağılım modelleri için düzgünlük koşullarını sağlamasa da tutarlı ve pratik uygulamalarda etkin olduğu ispatlanmıştır (Keribin 2000).

5.2. Logaritmik Olabilirlik Fonksiyonun Yanlılık Düzeltmesi

Model seçiminde, oluşturulan modeller ile ilişkili olarak doğru modelin seçimi, Kullback-Leibler bilgisi (Kullback ve Leibler 1951) cinsinden bir yaklaşımla bulunabilir. $f(m)$, doğru yoğunluk fonksiyonunu gösterirse oluşturulan yoğunluk fonksiyonu, $f(m;\hat{\boldsymbol{\theta}})$ ile ilişkili olan yoğunluk fonksiyonu $f(m)$'nin Kullback-Leibler bilgisi,

$$I\left\{f(m); f(m;\hat{\boldsymbol{\Psi}})\right\} = \int f(m)\log f(m)dm - \int f(m)\log f(m;\hat{\boldsymbol{\Psi}})dm \tag{5.12}$$

şeklindedir. (5.12)'deki eşitlikten elde edilen bilgi değeri, doğru yoğunluk fonksiyonu $f(m)$'nin oluşturulan yoğunluk fonksiyonu $f(m;\hat{\boldsymbol{\Psi}})$'dan sapmasının bir ölçüsüdür. Burada amaç Kullback-Leibler bilgi değerini mümkün olduğunca

küçültmektir. (5.12)'deki eşitliğinin sağ tarafındaki ikinci terim, model seçimi ile ilişkilidir ve

$$\upsilon(y;F)=\int f(m)\log f(m;\hat{\Psi})dm=\int \log f(m;\hat{\Psi})dF(m) \tag{5.13}$$

şeklinde ifade edilir. Burada F, gözlenen $\mathbf{y}=(\mathbf{y}_1^T,\ldots\mathbf{y}_n^T)^T$ verisinin doğru dağılımını göstermektedir. (5.13)'deki eşitlikte doğru dağılımın fonksiyonu F, deneysel dağılım fonksiyonu $\hat{F}_n$ ile değiştirilirse $\upsilon(y;F)$ 'nin bir tahmin edicisi,

$$\upsilon(\mathrm{y};\hat{F}_n)=\frac{1}{n}\sum_{j=1}^{n}\log f(\mathrm{y}_j;\hat{\Psi})=\frac{1}{n}\log L(\hat{\Psi}) \tag{5.14}$$

olarak elde edilir. $\upsilon(\mathrm{y};\hat{F}_n)$ genellikle,

$$\int \log f(m)dF(m) \tag{5.15}$$

olarak tanımlanan beklenen logaritmik yoğunluk fonksiyonu olduğundan daha büyük tahmin eder. Bunun nedeni, deneysel dağılım fonksiyonu $\hat{F}_n$'nin doğru dağılım F'ye göre oluşturulan dağılım $F_{\hat{\Psi}}$'ya daha yakın olmasıdır. (5.9)'daki eşitliğinin bir tahmin edicisi olan $\upsilon(\mathbf{y};\hat{F}_n)$ 'nin yanlılığı fonksiyonel olarak,

$$\begin{aligned} b(F) &= E_F\left\{\upsilon(Y;\hat{F}_n)-\upsilon(Y;F)\right\} \\ &= E_F\left\{\frac{1}{n}\sum_{j=1}^{n}\log(Y_j;\hat{\mathbf{\Psi}})-\int\log f(m;\hat{\mathbf{\Psi}})dF(m)\right\} \end{aligned} \tag{5.16}$$

ile ifade edilir. Burada $b(F)$, doğru dağılım F'nin yanlılık miktarını göstermektedir. E_F, $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ bağımsız rastgele değişkenlerinin ortak dağılım fonksiyonu F'nin beklenen değeridir.

Olabilirlik fonksiyonundan yanlılık terimi $b(F)$'nin çıkartılmasıyla logaritmik olabilirlik fonksiyonunun yanlılık düzeltmesi,

$$\log L(\hat{\boldsymbol{\Psi}}) - b(F) \tag{5.17}$$

şeklinde yapılır. Model seçimi için bir bilgi kriteri yanlılığı giderilmiş logaritmik olabilirlik fonksiyonuna dayalı olarak oluşturulur. Model seçiminde doğru model olarak (5.11)'deki ifadesini maksimum yapan, yani (5.12)'deki eşitlik ile verilen Kullback-Leibler bilgisini minimum yapan model en iyi model olarak seçilir.

Literatürde bilgi kriterleri genellikle bu yanlılık düzeltmesinin iki katının negatif değerlisi alınarak,

$$-2\log L(\hat{\boldsymbol{\Psi}}) + 2C \tag{5.18}$$

şeklinde düzenlenir. Burada birinci terim uyumun yetersizliğinin, ikinci terim C ise modelin karmaşıklığının bir ölçüsüdür. C genellikle ceza terimi olarak adlandırılır. En iyi model olarak, $-2\log L(\hat{\boldsymbol{\Psi}}) + 2C$ ifadesini minimum yapan model seçilir.

5.2.1. Akaike Bilgi Kriteri

Akaike (1974), sonlu karma dağılım modellerinde (5.16)'da verilen eşitlikteki yanlılık terimi $b(F)$'nin asimptotik olarak modeldeki parametre sayısına eşit olduğunu göstermiştir. Buna göre modeldeki parametre sayısı d ile gösterilirse, kısaca AIC (Akaike's Information Criterion) olarak adlandırılan Akaike bilgi kriteri,

$$-2\log L(\hat{\boldsymbol{\Psi}}) + 2d \tag{5.19}$$

olarak tanımlanır. AIC değeri minimum olan model en iyi model olarak seçilir.

Soromenho (1993) ve, Celeux ve Soromenho (1996), çalışmalarında AIC değerinin karma dağılım modellerindeki bileşen sayısını olduğundan daha fazla tahmin etmeye ya da seçmeye meyilli olduğunu göstermiştir. Buna karşın AIC, pratikte halen yaygın olarak kullanılmaktadır (Fraley ve Raftery 1998).

5.2.2. Bootsrap Yöntemine Dayalı Bilgi Kriteri

Ishiguro ve ark. (1997), (5.16)'daki eşitlikte tanımlanan yanlılık terimi $b(F)$ 'yi, Efron'un yeniden örnekleme (bootstrap) yöntemini (Efron 1979) kullanarak tahmin etmeyi önermişlerdir. Efron'un (bootstrap) bilgi kriteri EIC (Efron's Information Criterion) olarak da adlandırılan bu yönteme göre karma dağılım modelinin bileşen sayısı g,

$$-2\log L(\hat{\boldsymbol{\Psi}}) + 2b(\hat{F}_n) \tag{5.20}$$

ifadesine dayalı olarak seçilir. Burada $b(\hat{F}_n)$ şeklinde gösterilen parametrik olmayan yeniden örnekleme (bootstrap) yanlılık terimidir. $b(\hat{F}_n)$, bağımsız B tane yeniden örneklemeye dayalı Monte Carlo yöntemleri ile yaklaşık olarak hesaplanır. Bu durumda (5.16)'daki eşitlikte $b(\hat{F}_n)$ değeri,

$$b(\hat{F}_n) = E_{\hat{F}_n}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(Y_j^*;\hat{\boldsymbol{\Psi}}^*) - \frac{1}{n}\sum_{j=1}^{n}\log f(Y_j;\hat{\boldsymbol{\Psi}}^*)\right\} \tag{5.21}$$

ile bulunur. Burada $\hat{\boldsymbol{\Psi}}^*$ terimi, $\hat{F}_n$ dağılımından yeniden örnekleme ile oluşturulan bağımsız ve özdeş $\boldsymbol{Y}_1^*,\ldots,\boldsymbol{Y}_n^*$ rastgele değişkenlerinin örnekleminden edilen en çok olabilirlik tahminidir.

Yeniden örneklemenin yanlılığı, bağımsız B yeniden örnekleme ile elde edilen örneklemelerden yaklaşık olarak hesaplanabilir. Bağımsız B yeniden örnekleme ile elde edilen örneklemler,

$$\boldsymbol{Y}_{1b}^{*},\ldots,\boldsymbol{Y}_{nb}^{*} \underset{\sim}{böd} \hat{\boldsymbol{F}}_{n} \quad (b=1,2,\ldots,B) \tag{5.22}$$

şeklinde ifade edilir. $b=1,2,\ldots,B$ için $\hat{\boldsymbol{\Psi}}_{b}^{*}$ $(b=1,2,\ldots,B)$, yeniden örnekleme yöntemi ile oluşturulan *b.* örneklemden elde edilen en çok olabilirlik tahmini olmak üzere yanlılık terimi yaklaşık olarak

$$b(\hat{F}_{n}) \approx \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(y_{jb}^{*};\hat{\boldsymbol{\Psi}}_{b}^{*})-\frac{1}{n}\sum_{j=1}^{n}\log f(y_{j};\hat{\boldsymbol{\Psi}}_{b}^{*})\right\} \tag{5.23}$$

ifadesine eşit olur. Konishi ve Kitagawa (1996), yeniden örnekleme simülasyonlarında bir varyans indirgeme tekniği kullanıldığında yeniden örnekleme ile oluşturulan örneklem sayısının oldukça azaltılabileceğini göstermiştir. Konishi ve Kitagawa (1996), (5.16)'daki eşitliğin sağ tarafındaki ifadenin beklentisini,

$$b(F)=b_{1}(F)+b_{2}(F)+b_{3}(F) \tag{5.24}$$

olarak üç parçanın toplamı şeklinde yazmışlardır. Burada

$$b_{1}(F)=E_{F}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(Y_{j};\hat{\boldsymbol{\Psi}})-\frac{1}{n}\sum_{j=1}^{n}\log f(Y_{j};\boldsymbol{\Psi})\right\} \tag{5.25}$$

$$b_{2}(F)=E_{F}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(Y_{j};\boldsymbol{\Psi})-\int\log f(m;\boldsymbol{\Psi})dF(m)\right\} \tag{5.26}$$

$$b_{3}(F)=E_{F}\left\{\int\log f(m;\boldsymbol{\Psi})dF(m)-\int\log f(m;\hat{\boldsymbol{\Psi}})dF(m)\right\} \tag{5.27}$$

dir. $b_{2}(F)$ ifadesi sıfır olacağından (5.24)'deki eşitliğin sağ tarafındaki yalnız birinci ve üçüncü ifadeler,

$$b(\hat{F}_n) = b_1(\hat{F}_n) + b_3(\hat{F}_n) \tag{5.28}$$

eşitliği sağlayacak şekilde tahmin edilmesi gerekir. Burada $b_1(\hat{F}_n)$ ve $b_3(\hat{F}_n)$ sırasıyla,

$$b_1(\hat{F}_n) \approx \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{y}^*_{jb};\hat{\boldsymbol{\Psi}}^*) - \frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{y}^*_{jb};\hat{\boldsymbol{\Psi}}_b)\right\} \tag{5.29}$$

$$b_3(\hat{F}_n) \approx \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{y}_j;\hat{\boldsymbol{\Psi}}) - \frac{1}{n}\sum_{j=1}^{n}\log f(\mathbf{y}_j;\hat{\boldsymbol{\Psi}}^*_b)\right\} \tag{5.30}$$

ile yaklaşık olarak hesaplanır. Konishi ve Kitagawa (1996), yaptıkları uygulamalarda (5.29)'daki eşitlikte $b_2(\hat{F}_n)$ 'nin yeniden örnekleme teriminin tahmin değerinin her ne kadar küçük olsa da $b_1(F)$ ve $b_3(F)$'in yeniden örnekleme tahminlerinden daha büyük varyanslı olduğunu göstermişlerdir. Yeniden örnekleme tekrarlarını azaltmak için sadece $b_1(F)$ ve $b_3(F)$ tahminlerinin kullanıldığı düzenlenmiş yeniden örnekleme yaklaşımı kullanılır (McLachlan 2000, McLachlan ve Peel 2000).

5.2.3. Çapraz Geçerliliğe Dayalı Bilgi Kriteri

Smyth (2000), logaritmik olabilirlik fonksiyonunun yanlılık düzeltmesi için CVIC (Cross Validation based Information Criterion) çapraz geçerlilik kriterini önermiştir. Bu kritere göre karma dağılım modelindeki bileşen sayısı g, çapraz geçerlilik logaritmik olabilirliğe dayalı olarak belirlenir. Çapraz geçerlilik logaritmik olabilirliği,

$$\sum_{j=1}^{n}\log f(\mathbf{y}_j;\hat{\boldsymbol{\Psi}}_{(j)}) \tag{5.31}$$

şeklinde ifade edilir. Burada $\hat{\boldsymbol{\Psi}}_{(j)}$, $\boldsymbol{\Psi}$'nın, gözlenen $\mathbf{y}_1,\ldots,\mathbf{y}_n$ verisinden j. gözlemi $\mathbf{y}_j$'nin $(j=1,2,\ldots,n)$ atılması ile oluşan örneklemden elde edilen en çok olabilirlik tahminidir. Çapraz geçerlilik kriteri $\boldsymbol{\Psi}$'ye dayalı olarak elde edilen ve hacmi gerçek örneklem ya da eğitici örneklem hacmi ile aynı olan bir test örneklemi kullanılarak model oluşturmaya alternatif bir yöntem olarak görülebilir.

Burada her bir defada yalnız bir gözlemin silinmesine dayalı "bir çıkar at" çapraz geçerlilik yapısı oldukça zaman alıcı bir yöntemdir. Bu yüzden $v>1$ olmak üzere her defasında v gözlemin atıldığı v-misli (fold) çapraz geçerliliği geliştirilmiştir. Bu yöntemde gözlenen veri, her biri n/v hacimli olan v ayrık alt kümeye ayrılır. Diğer bir yöntemde Monte Carlo çapraz geçerlilik yöntemidir. Bu yöntemde veriden B sayıda parçalanmalar üretilerek $\boldsymbol{\Psi}$'yi tahmin etmek için örneklem hacmi γm olan bir test alt örneklemi ve hacmi $(1-\gamma)n$ olan eğitici örneklemi oluşturulur. Bu yöntemde klasik v-misli çapraz geçerlilik yönteminden farklı olarak test kümesinde her bir gözlemin birden fazla kullanılabilir olmasıdır. Smyth (2000), γ değerinin 0.5 alınmasıyla daha güçlü ya da dayanıklı (robust) sonuçlar elde edileceğini ve B değerinin de 20 ve 50 arasında olmasının birçok uygulama için yeterli olacağını belirtmiştir.

BIC ve CVIC karşılaştırıldığı çalışmalarda değişik veri kümeleri üzerinde karma normal modelleri kullanıldığında bileşen sayısını aynı bulmuşlardır (Smyth 2000, Mclachlan 1992, Fraley ve Raftery 1998).

5.3. Sınıflandırmaya Dayalı Bilgi Kriterleri

5.3.1. Sınıflandırma Olabilirlik Kriteri

Biernacki ve Govaert (1997), bir karma normal modelin veriye uydurulmasında belirleyici olan küme sayısının seçimi için 4. Bölümde (4.10)'daki eşitlik ile verilen olabilirlik fonksiyonu $L(\boldsymbol{\Psi})$ ile (4.21)'deki eşitlik ile verilen $\log L_c(\boldsymbol{\Psi})$ tamamlanmış logaritmik olabilirlik fonksiyonu arasındaki ilişkiyi kullanarak bir kriter önermiştir. Hathaway (1986) karma logaritmik olabilirliği,

$$\log L(\boldsymbol{\Psi}) = \log L_c(\boldsymbol{\Psi}) - \log k(\boldsymbol{\Psi}) \tag{5.32}$$

şeklinde ifade etmiştir. Burada $\log k(\boldsymbol{\Psi})$,

$$\log k(\boldsymbol{\Psi}) = \sum_{i=1}^{g}\sum_{j=1}^{n} z_{ij} \log \tau_{ij} \tag{5.33}$$

ile bulunur. Burada τ_{ij}, gözlenen $\mathbf{y}$ veri vektöründeki j. gözlemin karmanın i. bileşenine ait olması olasılığıdır. $k(\boldsymbol{\Psi})$ fonksiyonu ise gözlenen veri $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T, \ldots, \mathbf{y}_n^T)^T$ verilmişken bileşen etiket değişkenleri $\mathbf{z} = (\mathbf{z}_1^T, \mathbf{z}_2^T, \ldots, \mathbf{z}_1^T)^T$'nin koşullu yoğunluk fonksiyonudur. Gözlenen veri $\mathbf{y}$ verilmişken $\log k(\boldsymbol{\Psi})$'nin koşullu beklentisi $-EN(\tau)$ olmak üzere

$$EN(\tau) = -\sum_{i=1}^{g}\sum_{j=1}^{n} \tau_{ij} \log \tau_{ij} \tag{5.34}$$

şeklinde ifade edilir. ve $-EN(\tau)$, bulanık sınıflandırma matrisi $C = \{(\tau_{ij})\}$'nin entropisidir. Burada

$$\tau = \left(\tau_1^T, \tau_2^T, \ldots, \tau_n^T\right)^T \tag{5.35}$$

ve

$$\tau_j = (\tau_1(\mathbf{y}_j; \boldsymbol{\Psi}), \tau_2(\mathbf{y}_j; \boldsymbol{\Psi}), \ldots, \tau_g(\mathbf{y}_j; \boldsymbol{\Psi}))^T \tag{5.36}$$

formunda olup $j = 1,2,\ldots,n$ olmak üzere $\mathbf{y}_j$ gözlem vektörünün bileşen üyeliğinin sonraki olasılıklarının bir vektörüdür.

Tamamlanmış veri olabilirliğinin gözlem vektörü $\mathbf{y}$'ye ek olarak bileşen etiket vektörü $\mathbf{z}$'ye de bağlı olduğunu göstermek için $L_c(\mathbf{\Psi};\mathbf{z})$ şeklinde yazılsın. $L_c(\mathbf{\Psi};\mathbf{z})$'de $\mathbf{z}=\hat{\tau}$ alınırsa (5.32)'deki eşitlikten

$$L_c(\hat{\mathbf{\Psi}};\hat{\tau}) = L(\hat{\mathbf{\Psi}}) - 2EN(\hat{\tau}) \tag{5.37}$$

elde edilir. Burada Burada $\hat{\tau}$, τ'nun en çok olabilirlik kestiricisidir ve (5.36)'daki eşitlikte τ_{ij} yerine

$$\hat{\tau}_{ij} = \tau_i(\mathbf{y}_j;\hat{\mathbf{\Psi}}) \quad i=1,2,\ldots,g \quad j=1,2,\ldots,n \tag{5.38}$$

alınarak hesaplanır.

(5.37)'deki eşitlikte CLC (Classification Likelihood Criterion) sınıflandırma olabilirlik bilgi kriteri,

$$-2\log L(\hat{\mathbf{\Psi}}) + 2EN(\hat{\tau}) \tag{5.39}$$

olarak tanımlanır. Bu kriterin amacı, CLC'yi minimum yapan küme sayısı g değerini seçmektir. Model için tahmin edilen entropi $EN(\hat{\tau})$, modelin karmaşıklığına karşılık gelen ceza terimi olarak kullanılır.

Karma dağılım modelinin bileşenleri arasındaki ayrım fazla ise $EN(\hat{\tau})$ minimum değeri olan sıfıra yakınsayacaktır. Karma dağılım modelinde bileşenler arası ayırım zayıf ise $EN(\hat{\tau})$ ifadesi daha büyük değerler alır. Biernacki ve ark. (1999), bu kriterin karma dağılım modelinde bileşen olasılıklarının eşit olma kısıtı altında iyi sonuçlar verdiğini belirtmişlerdir.

Banfield ve Raftery (1993), sınıflandırma olabilirlik yaklaşımını kullanarak sınıf sayısının belirlenmesinde Bayesci çözüme yakın bir yöntem önermişlerdir.

Kanıtların yaklaşık olarak ağırlığı kriteri AWE (Approximate Weight of Evidence) olarak isimlendirilen bu kriter,

$$AWE_{(g)} = -2\log Lc + 2d(3/2 + \log n) \tag{5.40}$$

şeklinde ifade edilir.

5.3.2. Normalleştirilmiş Entropi Kriteri

Celeux ve Soromenho (1996), tahmin edilmiş entropi $EN(\hat{\tau})$'yi kullanarak küme sayısının seçimi için NEC (Normalized Entropy Criterion) normalleştirilmiş entropy kriterini önermişlerdir. Karma kümeleme analizinde uygulanan en çok olabilirlik yaklaşımı ile sınıflandırma en çok olabilirlik yaklaşımları arasındaki farkların bir ilişkisine dayalıdır. Karma dağılım modellerinde $\log L(\hat{\boldsymbol{\Psi}})$ değeri, bileşen sayısı g arttıkça büyüdüğü için tahmin edilmiş entropi $EN(\hat{\tau})$ doğrudan kullanılamaz. Normalleştirilmiş formu,

$$NEC_g = \frac{EN(\hat{\tau})}{\log L(\hat{\boldsymbol{\Psi}}) - \log L(\hat{\boldsymbol{\Psi}}^*)} \tag{5.41}$$

şeklinde ifade edilir. Burada $\hat{\boldsymbol{\Psi}}^*$, tek bileşen durumda $(g=1)$ $\boldsymbol{\Psi}$'nin en çok olabilirlik tahminidir. NEC kriterinin minimum olduğu bileşen sayısı g küme sayısı olarak seçilir. Entropi $g=1$ için sıfır değerini almaktadır. Bu durumu da dikkate alarak Biernacki ve ark. (1999), $NEC_g < 1$ olmak üzere NEC_g değerinin minimum olduğu bileşen sayısı $g > 1$ seçimini önermişlerdir.

6. MODELE DAYALI KÜMELEMEDE ADAY BİLEŞEN KÜME MERKEZLERİNİN SAYISI İÇİN BİR ARALIK

Çok değişkenli normal dağılımların karmasına dayalı kümeleme analizinde en önemli problemlerden biri gözlenen verinin yapısına uygun olarak aday karma modellerin sayısının belirlenmesi ve bu aday karma modelin bileşen sayısının seçimi olmuştur (Servi ve Erol 2007). Bu bölümde araştırılan konu iki alt başlık olarak verilecektir. Birincisi, aday bileşen küme merkezlerinin toplam sayısı kavramı açıklanacak ve çok değişkenli verinin her bir değişkenindeki parçalanmaların sayısı kullanılarak aday bileşen küme merkezlerinin sayısı için bir aralık önerilmiştir. İkincisi, çok değişkenli karma normal modeline dayalı kümeleme yapısında aday karma modellerin sayısı hakkında bilgi verilmiştir. Aday karma modellerin sayısı, her biri farklı bileşen sayılı olabilecek mümkün karma modellerin toplamı olarak tanımlanmıştır.

6.1. Giriş

Her biri p-boyutlu $\mathbf{y}_j$ gözlem vektörlerinden ($j = 1,2,\ldots,n$) meydana gelen çok değişkenli $\mathbf{y} = \left(\mathbf{y}_1^T,\ldots\mathbf{y}_n^T\right)^T$ gözlenen verisinin kümelenmesi için birçok yöntem geliştirilmiştir. Bu yöntemlerden biri de olasılık dağılımlarına dayalı olan ve sonlu karma dağılım modelleri kullanılarak oluşturulan kümeleme yöntemidir. Sonlu karma dağılım modeline dayalı kümeleme analizinde her bir bileşen çok değişkenli $\mathbf{y}$ gözlenen verisindeki bir kümeye karşılık gelir. Sonlu karma dağılım modelinde bileşen yoğunluklarının çok değişkenli normal dağılım olarak alınması ile yapılan kümeleme, çok değişkenli karma normal modellere dayalı kümeleme olarak adlandırılır. Çok değişkenli karma normal modeller, kümeleme analizi dışında ayrıca yoğunluk tahminleri ve ayrıştırma analizlerinde de yaygın olarak kullanılır.

Karma normal modellere dayalı kümeleme analizinde her biri p-boyutlu $\mathbf{y}_j$ gözlem vektörlerinden ($j = 1,2,\ldots,n$) meydana gelen çok değişkenli $\mathbf{y} = \left(\mathbf{y}_1^T,\ldots\mathbf{y}_n^T\right)^T$

gözlenen verisinin sonlu g sayıda çok değişkenli normal dağılımların bir karmasından elde edildiği kabul edilir. Çok değişkenli karma normal dağılım modeli için olasılık yoğunluk fonksiyonu $f(\mathbf{y}_j;\boldsymbol{\Psi})$,

$$f(\mathbf{y}_j;\boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j;\boldsymbol{\theta}_i) \tag{6.1}$$

ile gösterilir. Burada $i = 1,2,\ldots,g$ ve $j = 1,2,\ldots,n$ için π_i karma oranı, $0 \le \pi_i \le 1$ ve $\sum_{i=1}^{g} \pi_i = 1$ kısıtlarına sahip olmak üzere $\mathbf{y}_j$ gözlenen vektörünün karma modelin *i.* bileşeninden elde edilmesi olasılığını göstermektedir. $f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)$ bileşen olasılık yoğunluk fonksiyonudur. $\boldsymbol{\Psi}$ parametre vektörü, π_i karma oranlarını ve karma dağılım modelindeki her bir bileşen yoğunluğundaki $\boldsymbol{\theta}_i$ parametrelerini içermektedir. Diğer bir ifadeyle $\boldsymbol{\Psi} = \{\pi_1,\pi_2,\ldots,\pi_g,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_g\}$ şeklinde açık olarak yazılır.

(6.1)'deki eşitlik ile verilen karma dağılım modelinde $f_i(\mathbf{y}_j;\boldsymbol{\theta}_i)$ bileşen olasılık yoğunluk fonksiyonları,

$$f_i(\mathbf{y}_j;\boldsymbol{\theta}_i) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} e^{\left\{-\frac{1}{2}(\mathbf{x}_j-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_j-\boldsymbol{\mu}_i)\right\}} \tag{6.2}$$

şeklinde çok değişkenli normal dağılım modeli alındığında karma dağılım modeli, karma normal dağılım modeli olur. Burada $\boldsymbol{\theta}_i$ vektörü, $\boldsymbol{\mu}_i$ ortalama vektörünü ve $\boldsymbol{\Sigma}_i$ kovaryans matrisini içermektedir.

Karma normal dağılım modelindeki bilinmeyen parametreler, olabilirlik fonksiyonuna dayalı olarak EM algoritması ile tahmin edilir. Modele dayalı kümeleme de en iyi model ve küme sayısı bir bilgi kriterine dayalı olarak bulunur. Literatürde en yaygın olarak kullanılan bilgi kriterleri Akaike'nin bilgi kriteri AIC ve Bayes bilgi kriteri BIC' dir. AIC kriterinin $\boldsymbol{y}$ gözlenen verisinin karma normal

dağılım olarak modellenmesinde bileşen sayısını olduğundan daha fazla seçmeye eğilimli olduğu (over estimation) ispatlanmış olsa da pratik uygulamalarda halen yaygın olarak kullanılmaktadır (Celeux ve Soromenho 1996, McLachlan ve Peel 2000).

Dasgupta ve Raftery (1998), yığılmalı hiyerarşik modele dayalı kümeleme yöntemi ile oluşturulan parçalanmalar baz alınarak elde edilen parametre tahmin değerini, EM algoritmasında parametre başlangıç değerleri olarak alınabileceğini belirtmişlerdir. Bu şekilde parametre başlangıç değerleri alınan EM algoritmasının karma olabilirlik fonksiyonunu maksimum yapacak parametre tahminleri bulmada daha etkin olduğunu da göstermişlerdir. Fraley ve Raftery (1998), Dasgupta ve Raftery (1998) tarafından yapılan çalışmadaki bu sonucu kullanarak modele dayalı kümeleme için bir yöntem önermişlerdir. Geliştirdikleri bu yönteme $\mathbf{y}$ gözlenen verisi için maksimum küme sayısı M 'nin keyfi bir seçimi ile başlanır. Karma normal dağılım modeline dayalı yığılmalı hiyerarşik kümeleme yöntemi, küme sayısı birden başlayarak keyfi seçilen maksimum küme sayısı M 'ye kadar sınıfların elde edilmesinde kullanılır. Bu M sınıftan elde edilen parametre değerleri EM algoritması için başlangıç değerleri olarak alınır. EM algoritması, küme sayısı g 1'den başlayarak maksimum küme sayısı M 'ye kadar olan modeller için uygulanır.

EM algoritması sonucu elde edilen modellerin parametre tahminleri ve her bir modele karşılık gelen BIC değerleri bulunur. BIC değerlerinin küme sayısına göre grafiği oluşturulur. Veriye en uygun küme sayısı, oluşturulan BIC değerlerinin grafiğinde ilk yerel maksimum BIC değerine karşılık gelen küme sayısı olarak alınır. Bu şekilde belirlenen en uygun küme sayısı, veriyi modellemek için kullanılan karma normal dağılım modelinde bileşen sayısı olarak kabul edilir. Veri en iyi modelleyen karma normal model de aday karma dağılım modeller arasından bu bileşen sayılı model olarak belirlenir. Fraley ve Raftery (1998) tarafından önerilen bu yöntemin bazı dezavantajları bulunmaktadır (Servi ve Erol 2007). Dezavantajlardan birincisi, mümkün olduğunca küçük alınması önerilen maksimum küme sayısı M 'nin nasıl bulunacağının belirsiz olmasıdır. Dezavantajlardan ikincisi en iyi modelin belirlenmesi amacıyla BIC değerlerinin: küme sayısı g için minimum 1'den

başlanarak, maksimum M 'ye kadar hesaplanmasıdır. Bu yöntem, veride gereksiz ya da olduğundan daha fazla hesaplamaların yapılmasına neden olmaktadır.

Bu bölümde, Fraley ve Raftery (1998) tarafından önerilen modele dayalı kümelemede ortaya çıkan dezavantajları da ortadan kaldıracak şekilde iki ana konu üzerinde durulacaktır. Birincisi, toplam aday bileşen küme merkez sayısı $TNCCCC$ (Total Number of Candidate Component Cluster Centers) tanımlanması ve çok değişkenli veride her bir değişkendeki parçalanmaların sayısı kullanılarak $TNCCCC$ için bir aralık oluşturulmasıdır. Karma normal dağılım modelinin bileşenleri bu aday bileşen küme merkezleri etrafında şekillenecektir. İkincisi, çok değişkenli karma normal modele dayalı kümelemede model seçimi için toplam aday karma modellerin sayısı $TNCMM$ (Total Number of Candidate Mixture Models) hakkında bir eşitliğin verilmesidir. Aday karma modellerin sayısının $NCMM$ (Number of Candidate Mixture Models), farklı bileşen küme sayıları ile mümkün karma modellerin sayılarının $NPMM$ (Number of Possible Mixture Models) toplamı olarak tanımlanmasıdır. Çok değişkenli veriyi temsil edecek en iyi ya da en uygun modelin mümkün karma modeller arasından seçilmesidir.

6.2. Aday Bileşen Küme Merkezlerinin Toplam Sayısı İçin Bir Aralık

Her biri p-boyutlu $\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_n$ gözlenen veri için bileşen yoğunlukları (6.2)'deki eşitlik ile verilen çok değişkenli normal dağılım olmak üzere (6.1)'deki eşitlik ile ifade edilen karma dağılım modeline sahip olsun. p-boyutlu çok değişkenli gözlenen verinin her bir $\mathbf{Y}_s$ bağımsız rastgele değişkeninin $s = 1,2,\ldots,p$ olmak üzere,

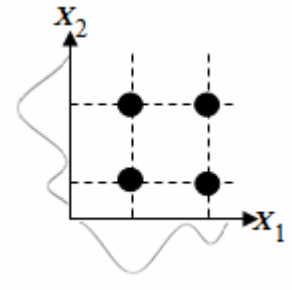
$$\mathbf{Y}_s = \begin{cases} k_s > 1,\ \mathbf{Y}_s \text{ rastgele değişkeni karma yapıya sahipse} \\ \qquad \text{(Tek Değişkenli Karma Normal Dağılım Durumu)} \\ k_s = 1,\ \mathbf{Y}_s \text{ rastgele değişkeni karma yapıya sahip değilse} \\ \qquad \text{(Tek Değişkenli Normal Dağılım Durumu)} \end{cases} \tag{6.3}$$

şeklinde iki kategoriden birine sınıflandığı varsayılsın. Burada k_s, p-boyutlu çok değişkenli veride $\mathbf{Y}_s$ rastgele değişkenindeki bileşen sayısıdır. p-boyutlu çok değişkenli veride s. rastgele değişken $\mathbf{Y}_s$ üzerinde gözlenmiş j. Gözlem, y_{sj} ile ifade edilsin. Bu durumda p-boyutlu çok değişkenli verinin g bileşenli karma normal modelinde $TNCCCC$ için aralık,

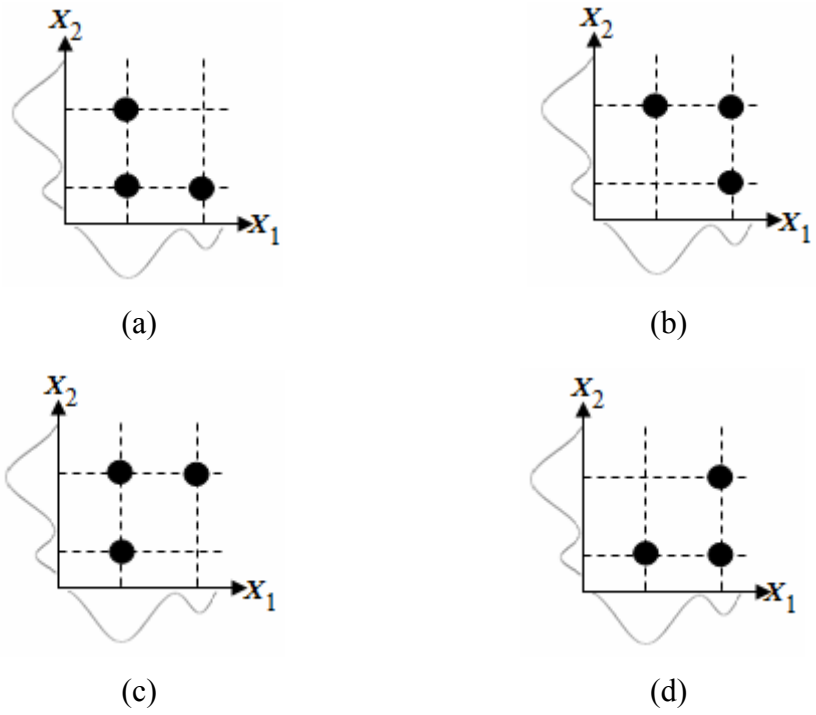
$$\max\{k_1,\ldots,k_p\} \le TNCCCC \le \prod_{s=1}^{p} k_s \qquad (6.4)$$

şeklinde oluşturulur. Bu eşitlikte $\max\{k_1,\ldots,k_p\}$ terimi, toplam aday bileşen küme merkez sayısının minimumunu ve $\prod_{s=1}^{p} k_s$ terimi, toplam aday bileşen küme merkez sayısının maksimumunu ifade eder. Buna göre $TNCCCC_{\min} = \max\{k_1,\ldots,k_p\}$ ve $TNCCCC_{\max} = \prod_{s=1}^{p} k_s$ şeklinde ifade edilir.

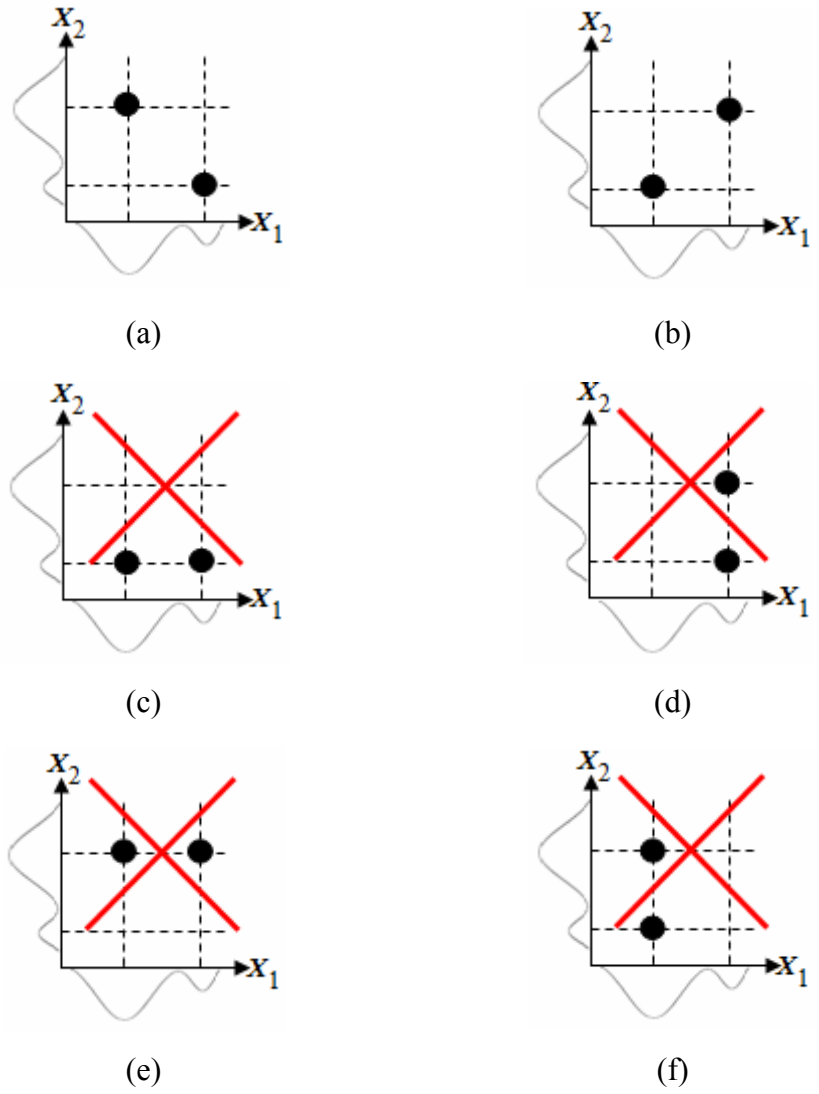
$TNCCCC_{\min}$ ve $TNCCCC_{\max}$ ifadelerinin mantığını iki değişkenli bir karma normal model örneği üzerinde gösterelim. İki değişkenli karma modelin birinci rastgele değişkeni Y_1 için $k_1 = 2$ ve ikinci rastgele değişkeni Y_2 için $k_2 = 2$ olsun. Bu durumda $TNCCCC_{\min} = \max\{2,2\} = 2$ ve $TNCCCC_{\max} = k_1.k_2 = 4$ olarak elde edilir. Bu örnek için oluşabilecek aday karma model durumları Şekil 6.1 - Şekil 6.4'de verilmiştir. Aday bileşen küme merkezlerinin sayısı toplam 4 olduğunda: Şekil 6.1'de gösterildiği gibi 4 bileşen kümeli 1 tane, Şekil 6.2'de gösterildiği gibi 3 bileşen kümeli 4 tane, Şekil 6.3'te gösterildiği gibi 2 bileşen kümeli 6 tane iki değişkenli karma normal model; ve Şekil 6.4'de gösterildiği gibi 1 bileşenli 4 tane normal model oluşturulabilir.



Şekil 6.1. İki değişkenli bir normal karma modelin X_1 değişkeni için $k_1 = 2$ ve X_2 değişkeni için $k_2 = 2$ alınmasıyla oluşabilecek farklı aday bileşen küme merkezlerinin sayısı toplam 4 olduğunda: 4 bileşen kümeli 1 karma normal model durumu. Mümkün karma normal model sayısı 1 dir.

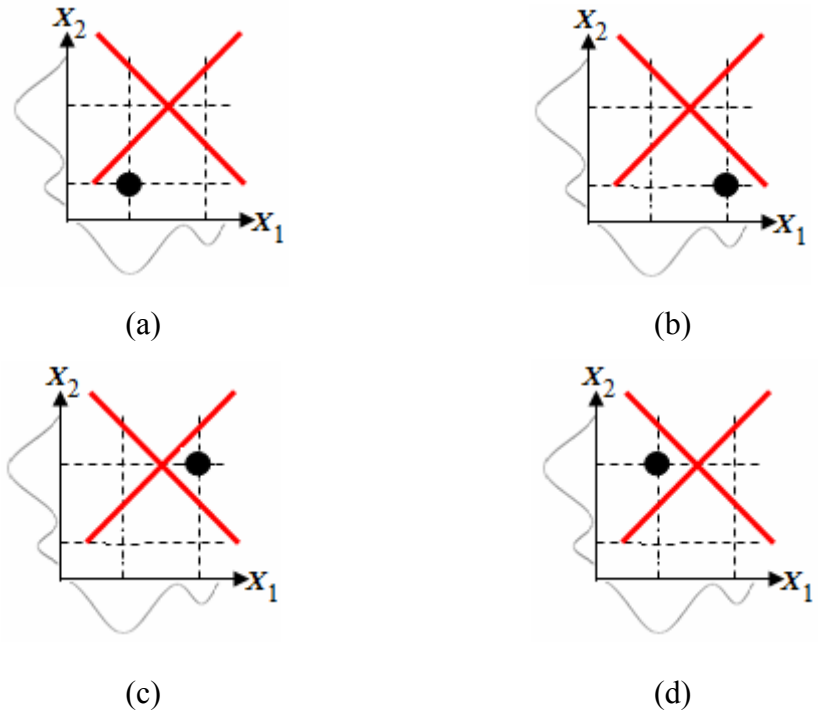


Şekil 6.2. İki değişkenli bir normal karma modelin X_1 değişkeni için $k_1 = 2$ ve X_2 değişkeni için $k_2 = 2$ alınmasıyla oluşabilecek farklı aday bileşen küme merkezlerinin sayısı toplam 4 olduğunda: 3 bileşen kümeli 4 karma normal model durumu. Mümkün karma normal model sayısı 4 dür.



Şekil 6.3. İki değişkenli bir normal karma modelin X_1 değişkeni için $k_1 = 2$ ve X_2 değişkeni için $k_2 = 2$ alınmasıyla oluşabilecek farklı aday bileşen küme merkezlerinin sayısı toplam 4 olduğunda: 2 bileşen kümeli 6 karma normal model durumu. Mümkün karma normal model sayısı 2 dir.

Şekil 6.3'te (c) ve (e) durumlarında ikinci rastgele değişken Y_2 için $k_2 = 2$ olması koşulu ya da varsayımı ile çelişmektedir. Benzer biçimde Şekil 6.3'te (d) ve (f) durumlarında birinci değişken Y_1 için $k_1 = 2$ olması koşulu ya da varsayımı ile çelişki gösterdiğinden bu durumlar mümkün olamayan iki değişkenli karma normal yapılardır.

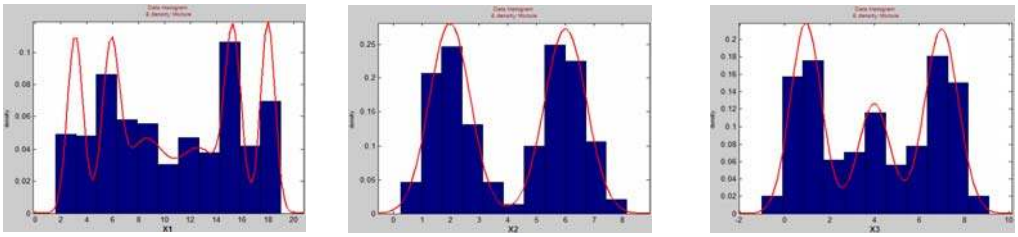


Şekil 6.4. İki değişkenli bir normal karma modelin X_1 değişkeni için $k_1 = 2$ ve X_2 değişkeni için $k_2 = 2$ alınmasıyla oluşabilecek farklı aday bileşen küme merkezlerinin sayısı toplam 4 olduğunda: 1 bileşen kümeli 4 normal model durumu. Mümkün karma normal model sayısı 0 dır.

Şekil 6.4'de (a) - (d) durumlarında aynı anda rastgele değişkenler Y_1 ve Y_2 için sırasıyla $k_1 = 2$ ve $k_2 = 2$ olması koşulu ya da varsayımı ile çelişki gösterdiğinden bu durumlar mümkün olamayan iki değişkenli karma normal yapılardır.

TNCCCC için (6.4)'deki ifadeyle önerilen aralığın etkinliği Tablo 6.1'de verilen simülasyon verisi kullanılarak açıklanmıştır. Tablo 6.1'de verilen simülasyon verisi, 3 değişkenlidir. GENMIX yazılımı (Martinez ve Martinez 2005) kullanılarak oluşturulan simülasyon verisinin birinci rastgele değişkeni Y_1 için $k_1 = 6$, ikinci rastgele değişkeni Y_2 için $k_2 = 2$ ve üçüncü rastgele değişkeni Y_3 için $k_3 = 3$ olarak alınmıştır. Bu simülasyon verisi için *TNCCCC* 'nin alabileceği değerler $6 \le TNCCCC \le 36$ şeklinde bulunmuştur. MIXMOD yazılımı (Biernacki ve ark.

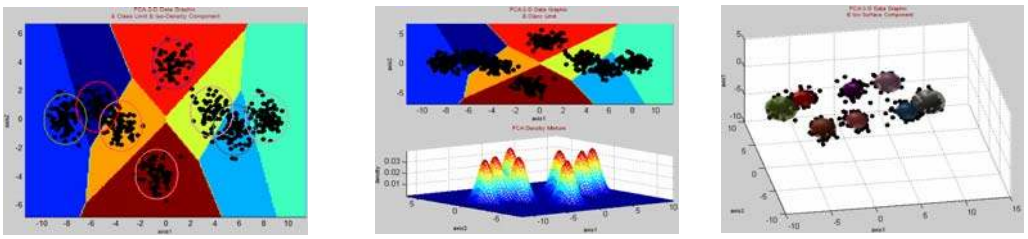
2006) kullanılarak simülasyon verisindeki küme sayısı $g = 8$ olarak elde edilmiştir. Simülasyon verisinin her bir değişkeni için MIXMOD yazılımı (Biernacki ve ark. 2006) kullanılarak oluşturulan tek değişkenli normal dağılımların karma yoğunluk fonksiyonlarının grafikleri Şekil 6.5 (a) – (c)'de gösterilmiştir.



(a) (b) (c)

Şekil 6.5. Tablo 6.1'de verilen simülasyon verisinin (a) X_1 değişkeni, (b) X_2 değişkeni ve (c) X_3 değişkeni için MİXMOD yazılımı kullanılarak elde edilen tek değişkenli karma model grafikleri.

Tablo 6.1'deki simülasyon verisinde küme sayısı $g = 8$ 'dir. Diğer bir ifadeyle elde edilen en iyi çok değişkenli karma normal model sekiz bileşenlidir. Simülasyon verisi için oluşturulan 8 bileşenli karma normal model için saçılım grafiği, 8 bileşenli karma normal olasılık yoğunluk fonksiyonunun grafiği ve üç boyutlu kümelenme görünümü Şekil 6.6'da verilmiştir.



(a) (b) (c)

Şekil 6.6. Tablo 6.1'deki simülasyon verisi için MIXMOD yazılımı kullanılarak oluşturulan çok değişkenli karma normal modelin (a) Saçılım grafiği $g = 8$ bileşen kümesine ait 2 boyutlu gösterim, (b) $g = 8$ bileşenli karma normal modelin olasılık yoğunluk fonksiyonunun grafiği, (c) Simülasyon verisindeki bileşen kümelerinin 3 boyutlu gösterimi.

Modele dayalı kümeleme analizinde bilinen klasik yöntemler için geliştirilen algoritmalarda karma normal model için en uygun bileşen sayısı araştırılırken bileşen sayısı 1'den başlanır ve maksimum değer kabul edilen M 'ye kadar test edilir. Tablo 6.1'deki simülasyon verisi için karma normal model oluşturulurken (6.4)'deki eşitlik kullanılarak elde edilen aralık dikkate alındığında bu algoritmalar için Servi ve Erol (2007) tarafından önerilen bileşen küme sayısı için başlangıç değeri 6 olmaktadır. Karma normal model oluşturulurken işlemlerin en uygun bileşen küme sayısı başlangıç değerinin 6 alınması durumunda, bileşen küme sayısının 1, 2, 3, 4 ve 5 değerleri için hesaplama yapılmasına gerek yoktur. Bu nedenle de en uygun bileşen küme sayının belirlenmesi daha hızlı hesaplanmıştır.

Tablo 6.1'de incelenen diğer veri kümeleri için oluşturulan $TNCCCC$ aralıklarının bazıları çok geniştir. Bu aralıkların çok geniş olması modele dayalı karma kümeleme analizi için bir sorun oluşturmaz. Bunun nedeni modele dayalı kümeleme analizinde kullanılan algoritmalarda en uygun bileşen küme sayısı için $TNCCCC_{\min}$ değerinin alınmasıdır. Örneğin, en uygun bileşen küme sayısının belirlenmesi amacıyla simülasyon verisinde başlangıç bileşen küme sayısı için 6 değeri alınmış ve en uygun bileşen küme sayısı 8 değerinde elde edilmiştir. Bu nedenle en uygun bileşen küme sayısını bulmak amacıyla $TNCCCC_{\max}$ değeri olan 36'ya kadar deneme yapmaya gerek yoktur.

6.3. Aday Karma Modellerin Toplam Sayısı

$TNCCCC$ için geliştirilen ve (6.4)'deki eşitlikte önerilen aralık aynı zamanda p-boyutlu çok değişkenli veri için model seçimi kriterlerinin kullanıldığı karma kümeleme analizinde $TNCMM$ hakkında bir önbilgi de sağlar. $TNCMM$ için eşitlik,

$$TNCMM = 2^{TNCCCC_{\max}} \tag{6.5}$$

ile verilir. Burada $TNCCCC_{\max}$, toplam aday bileşen küme merkez sayısıdır. Bölüm 6.2'de verilen iki değişkenli karma normal dağılım modeli örneği için toplam aday

karma model sayısı (6.5)'deki eşitlik kullanılarak $TNCMM = 2^{TNCCCC_{\max}} = 2^4 = 16$ bulunur. $NCMM$ aday karma model sayısı ise 7'dir. $NCMM$ aday karma model sayısı bu örnek için 4, 3 ve 2 bileşen kümeli $NPMM$ mümkün karma modellerin sayılarının toplamından elde edilir. Tablo 6.1'deki simülasyon verisi için $NPMM\_4CC$, 4 bileşen kümeli (CC - Component Cluster) $NPMM$ mümkün karma modellerin sayısını, $NPMM\_3CC$, 3 bileşen kümeli (CC - Component Cluster) $NPMM$ mümkün karma modellerin sayısını, $NPMM\_2CC$, 2 bileşen kümeli (CC - Component Cluster) $NPMM$ mümkün karma modellerin sayısını ve $NPMM\_1CC$, 1 bileşen kümeli (CC - Component Cluster) $NPMM$ mümkün karma modellerin sayısını göstermek üzere $NCMM$ aday karma modellerin sayısı,

$$NCMM = NPMM\_4CC + NPMM\_3CC + NPMM\_2CC + NPMM\_1CC \quad (6.6)$$

ile hesaplanır. Bu durumda incelenen örnek için Şekil 6.1'de gösterildiği gibi $NPMM\_4CC$=1, Şekil 6.2'de gösterildiği gibi $NPMM\_3CC$=4, Şekil 6.3'te gösterildiği gibi $NPMM\_2CC$=2 ve Şekil 6.4'de gösterildiği gibi $NPMM\_1CC$=0 olduğundan $NCMM$ aday karma modellerin sayısı, $NCMM$=1+4+2+0=7 olarak belirlenir.

6.4. Toplam Aday Küme Bileşen Merkez Sayıları İçin Geliştirilen Yöntemin Farklı Veri Setlerindeki Performansı

Bu bölümde $TNCCCC$ için (6.4)'deki eşitlik ile önerilen aralık, çok değişkenli karma normal kümeleme analizinde literatürde oldukça sık kullanılan veri kümeleri üzerinde test edilecektir. İncelecek veri kümeleri ve bazı karakteristikleri Tablo 6.1'de özetlenmiştir. Uygulamada kullanılan veri kümelerinin her bir değişkeninin bileşen sayıları ve bilgi kriterler değerlerinin bulunması için gerekli hesaplamalar, MIXMOD yazılımı (Biernacki ve ark. 2005) kullanılarak yapılmıştır.

Tablo 6.1. İncelenen veri kümelerinin bazı özellikleri ve *TNCCCC* için aralık değerleri. (n : gözlem sayısı, p :verinin boyutu, Y_s : s . değişken, k_s : s . değişkenin karma dağılım modelinin bileşen sayısı, g : verideki gerçek küme sayısı).

Veri Seti	n	p	$\mathbf{X}_s$	k_s	*TNCCCC* **için aralık**	g
Geyser (Venableys ve Ribley 1994)	272	2	$\mathbf{X}_1$	3	$3 \le TNCCCC \le 6$	3
			$\mathbf{X}_2$	2		
Ruspini (Kaufman ve Rousseeuw 1990)	75	2	$\mathbf{X}_1$	2	$4 \le TNCCCC \le 8$	4
			$\mathbf{X}_2$	4		
Diabetes (Banfield ve Raftery 1993)	145	3	$\mathbf{X}_1$	2	$3 \le TNCCCC \le 12$	3
			$\mathbf{X}_2$	3		
			$\mathbf{X}_3$	2		
Simulated Data Set (Servi ve Erol 2007)	600	3	$\mathbf{X}_1$	6	$6 \le TNCCCC \le 36$	8
			$\mathbf{X}_2$	2		
			$\mathbf{X}_3$	3		
Iris (Biernacki ve ark. 2005)	150	4	$\mathbf{X}_1$	3	$3 \le TNCCCC \le 18$	3
			$\mathbf{X}_2$	3		
			$\mathbf{X}_3$	2		
			$\mathbf{X}_4$	2		
Artificial Data (Galimberti ve Soffritti 2007)	300	4	$\mathbf{X}_1$	3	$3 \le TNCCCC \le 36$	7
			$\mathbf{X}_2$	3		
			$\mathbf{X}_3$	2		
			$\mathbf{X}_4$	2		
Liver Disorders (Halbe ve Aladjem 2005)	345	6	$\mathbf{X}_1$	1	$2 \le TNCCCC \le 64$	2
			$\mathbf{X}_2$	2		
			$\mathbf{X}_3$	2		
			$\mathbf{X}_4$	2		
			$\mathbf{X}_5$	2		
			$\mathbf{X}_6$	2		
			$\mathbf{X}_6$	1		
			$\mathbf{X}_7$	2		
			$\mathbf{X}_8$	2		

Tablo 6.1 (Devamı). İncelenen veri kümelerinin bazı özellikleri ve *TNCCCC* için aralık değerleri. (n: gözlem sayısı, p:verinin boyutu, Y_s: s. değişken, k_s: s. değişkenin karma dağılım modelinin bileşen sayısı, g: verideki gerçek küme sayısı).

Veri Seti	n	p	$\mathbf{X}_s$	k_s	*TNCCCC* **için aralık**	g
Yeast (Newman ve ark. 1998)	1484	8	$\mathbf{X}_1$	2	$2 \le TNCCCC \le 32$	10
			$\mathbf{X}_2$	1		
			$\mathbf{X}_3$	2		
			$\mathbf{X}_4$	2		
			$\mathbf{X}_5$	1		
			$\mathbf{X}_6$	1		
			$\mathbf{X}_7$	2		
			$\mathbf{X}_8$	2		
Glass (Halbe ve Aladjem 2005)	214	9	$\mathbf{X}_1$	2	$2 \le TNCCCC \le 32$	6
			$\mathbf{X}_2$	2		
			$\mathbf{X}_3$	1		
			$\mathbf{X}_4$	2		
			$\mathbf{X}_5$	2		
			$\mathbf{X}_6$	1		
			$\mathbf{X}_7$	2		
			$\mathbf{X}_8$	1		
			$\mathbf{X}_9$	1		
Wine (Newman ve ark. 1998)	178	13	$\mathbf{X}_1$	2	$3 \le TNCCCC \le 256$	3
			$\mathbf{X}_2$	2		
			$\mathbf{X}_3$	1		
			$\mathbf{X}_4$	1		
			$\mathbf{X}_5$	3		
			$\mathbf{X}_6$	2		
			$\mathbf{X}_7$	2		
			$\mathbf{X}_8$	2		
			$\mathbf{X}_9$	1		
			$\mathbf{X}_{10}$	3		
			$\mathbf{X}_{11}$	2		
			$\mathbf{X}_{12}$	2		
			$\mathbf{X}_{13}$	2		

Tablo 6.1'deki sonuçlara göre Geyser (Venableys ve Ribley 1994), Ruspini (Kaufman ve Rousseeuw 1990), Diabetes (Banfield ve Raftery 1993), Simulated Data Set (Servi ve Erol 2007), Iris (Biernacki ve ark. 2005), Artificial Data (Galimberti ve Soffritti 2007), Liver Disorders (Halbe ve Aladjem 2005), Yeast (Newman ve ark. 1998) ve Glass (Halbe ve Aladjem 2005) veri setleri için oluşturulan *TNCCCC* aralıkları kabul edilebilir genişliktedir. Bununla birlikte Tablo 6.1'deki sonuçlara göre Wine (Newman ve ark. 1998) veri seti için oluşturulan *TNCCCC* aralığı kabul edilebilir genişliğin çok üstündedir. Bunun nedenleri olarak değişken sayısının fazla olması ve değişkenler arasında ilişki bulunması verilebilir.

7. ÇOK DEĞİŞKENLİ VERİDEKİ GRUPLARIN İNCELTİLEREK KARMA KÜMELEME ANALİZİ

Bu bölümde çok değişkenli veri setinin heterojen yapıda olması, alt gruplar içermesi, kısmen ya da tamamıyla iç içe geçmiş grupların bulunması durumlarında çok değişkenli normal karma dağılım modeline dayalı olarak kümelenme sayısının ve yapısının belirlenmesi amacıyla yeni bir algoritma geliştirilmiştir.

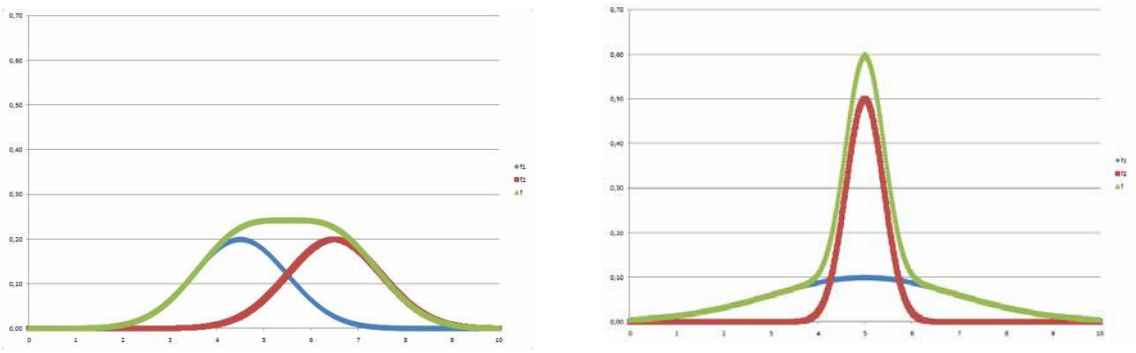
Çok değişkenli veride küme sayısının veya değişken sayısının fazla olması durumunda kümelerin belirlenmesi işlemi zorlaşır (Fraley 1998). Çok değişkenli veriyi kümelemek amacıyla kullanılan birçok kümeleme yöntemi bulunmaktadır (Jain ve Dubes 1988, Kaufman ve Rousseeuw 1990). Karma kümeleme analizi olarak da adlandırılan çok değişkenli normal karma modele dayalı kümeleme, çok değişkenli veriyi birbirinden farklı ve anlamlı grup ya da kümelere parçalamanın bir yoludur (Fraley ve Raftery 2002).

Çok değişkenli normal karma dağılım modeline dayalı olarak kümelenme sayısının ve yapısının belirlenmesi amacıyla yeni geliştirilen algoritmanın işleyiş prensibi Cam belirleme (glass identification) verisi (Halbe ve Aladjem 2005) üzerinde açıklanmıştır. Çok değişkenli veri için oluşturulan karma normal dağılım modelindeki parametre tahminlerinin hesaplanması amacıyla MIXMOD yazılımı (Biernacki ve ark. 2006) kullanılmıştır. Karma kümeleme analizinde çok değişkenli karma normal dağılım modelindeki parametreler, ençok olabilirlik yöntemiyle birlikte EM algoritması (McLachlan ve Krishnan 1997) kullanılarak tahmin edilmiştir. En uygun kümeleme yapısı, bir bilgi kriterinin kullanılmasıyla elde edilir. Farklı bilgi kriterleri bulunmakla birlikte bu çalışmada Bayesci bilgi kriteri BIC (Schwarz 1978) kullanılmıştır. BIC, literatürde en yaygın olarak kullanılan bilgi kriteridir (McLachlan ve Peel 2000).

7.1. Çok Değişkenli Veride İç İçe Geçmiş Gruplar Yapıları

Çok değişkenli veri setinin heterojen yapıda olması, diğer bir ifadeyle alt gruplar içermesi ve bu alt grupların kısmen ya da tamamıyla iç içe geçmiş yapıda

olması durumlarında çok değişkenli normal karma dağılım modeline dayalı olarak kümelenme sayısının ve yapısının belirlenmesi oldukça zordur. Çok değişkenli veride iç içe alt grup yapısının bulunduğu bir örnek Şekil 7.1'de gösterilmiştir.



(a) (b)

Şekil 7.1. Çok değişkenli veride iç içe alt grup yapısının bulunması durumları (a) farklı ortalamalı, eşit varyanslı alt grup durumu. (b) eşit ortalamalı, farklı varyanslı alt grup durumu.

7.2. Çok Değişkenli Veride Karma Yapının Belirlenmesi ve Grupların İnceltilmesi Algoritması

Çok değişkenli veride karma yapının belirlenmesi için bilgi kriterlerine dayalı olarak model seçimi yöntemi uygulanır. Karma dağılım modellerinde model seçimi konusu 5. Bölümde incelenmiştir.

Çok değişkenli verideki grupların inceltilerek karma kümeleme analizi için önerilen algoritmanın adımları aşağıda verilmiştir.

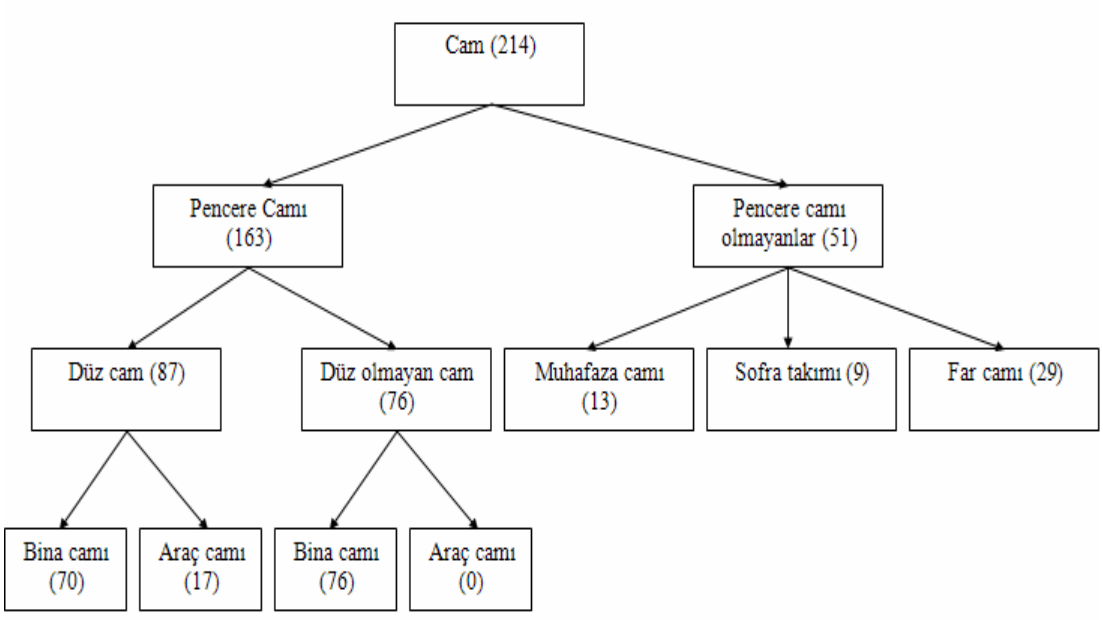
1. Çok değişkenli verinin çok değişkenli normal dağılımların bir karmasından geldiği varsayımı yapılır ve veriye karma kümeleme analizi uygulanır. Bu durumda veri ya homojendir, karma yapı yoktur ya da veri alt gruplara ayrılır, karma yapı vardır.
2. Veri karma yapıya sahipse ya da veride alt grup bulunuyorsa elde edilen her bir alt gruptaki veri homojenlik için test edilir. Diğer bir ifadeyle her bir alt gruptaki veri için tekrar karma yapı araştırılır.

3. Eğer alt gruptaki veri homojen ise bu alt grup bir küme olarak belirlenir ve bu alt gruptaki ya da kümedeki veri için çok değişkenli normal dağılım modeli oluşturulur. Eğer alt gruptaki veri homojen değil ise 1. adıma geri dönülür. İşlemler homojen olmayan bu alt gruptaki veri için tekrarlanır.
4. Bu işlemlere ortaya çıkan tüm alt gruplardaki verinin homojenliği sağlanıncaya kadar devam edilir.

7.3. Çok Değişkenli Verideki Grupların İnceltilmesi Ve Karma Kümeleme Analizi

Bu bölümde çok değişkenli verideki grupların inceltilerek karma kümeleme analizi için önerilen yöntemin çalışma prensibi Cam belirleme (Glass identification) verisi (Halbe ve Aladjem 2005) üzerinde gösterilmiştir. Cam belirleme verisi, bir suç mahallinde bulunan cam parçacıklarının doğru sınıflandırılması durumunda bir kanıt olarak kullanılabileceği düşüncesiyle oluşturulmuştur. Cam belirleme verisi, 214 gözlem değerinden ve 9 değişkenden oluşmaktadır. Bu değişkenler sırasıyla: 1) Kırılma katsayısı RI (Refractive Index), 2) Sodyum (Na) miktarı, 3) Magnezyum (Mg) miktarı, 4) Alüminyum (Al) miktarı, 5) Silisyum (Si) miktarı, 6) Potasyum (K) miktarı, 7) Kalsiyum (Ca) miktarı, 8) Baryum (Ba) miktarı ve 9) Demir (Fe) miktarı şeklindedir. Cam belirleme verisinin laboratuar çalışması sonucunda 7 cam kategorisinden; bunlar: 1. Bina camı, 2. Araç camı, 3. Düz olmayan bina camı, 4. Düz olmayan araç camı, 5. Koruma camı, 6. Mutfak eşyası camı ve Otomobil far camı olmak üzere düz olmayan araç camı dışındaki kategorilerden 6'sına sınıflandırılmıştır. Cam belirleme verisinin cam sınıflarına olan dağılımı Şekil 7.2'de gösterilmiştir.

Cam belirleme verisinin her bir değişkenine uygulanılan karma kümeleme analizi sonucu elde edilen bileşen sayıları ve (6.4)'deki ifade kullanılarak $TNCCCC$ toplam aday bileşen küme merkezi sayısı (Servi ve Erol 2007) için oluşturulan aralık Tablo 7.1'de verilmiştir.

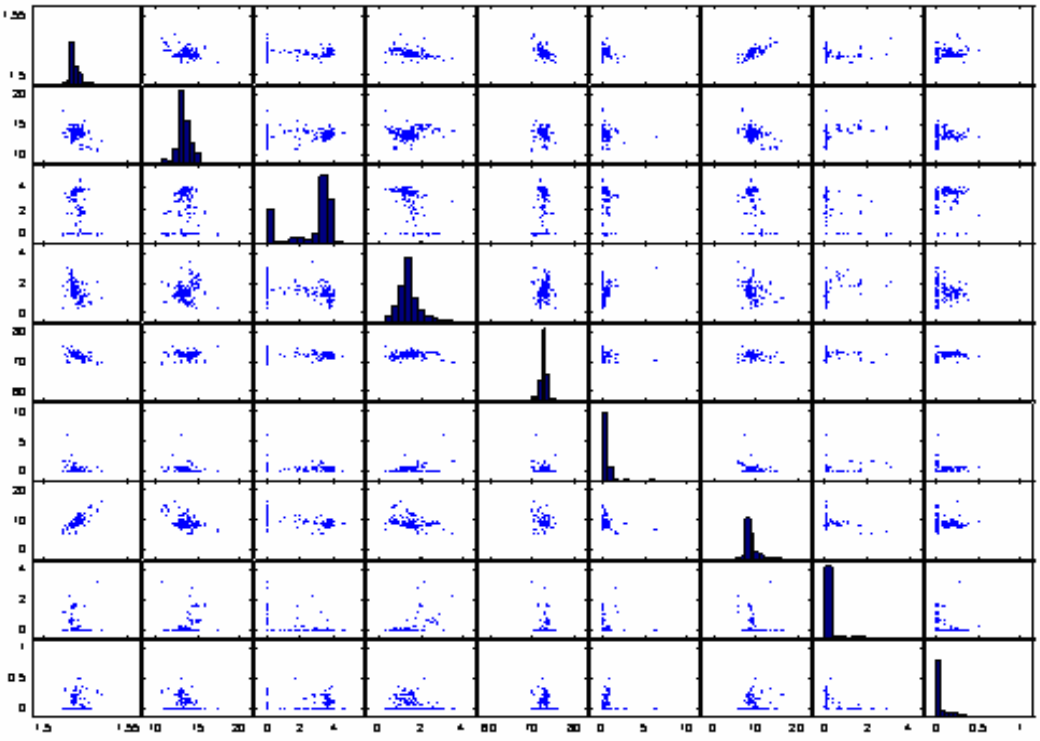


Şekil 7.2. Cam belirleme verisinin cam kategorilerine dağılımı. Parantez içindeki değerler gözlem sayılarını göstermektedir (Halbe ve Aladjem 2005).

Tablo 7.1. Çok değişkenli Cam belirleme verisinin bazı özellikleri ve $TNCCCC$ için oluşturulan aralıklar. (n: Gözlem ya da nesne sayısı, p: Değişken ya da özellik sayısı, $\mathbf{X}_s$: Değişkenler, k_s: Her bir değişkendeki bileşen sayısı, g: Karma modeldeki gerçek bileşen küme sayısı).

Veri Seti	n	p	$\mathbf{X}_s$	k_s	$TNCCCC$ **için aralık**	g
Glass (Halbe ve Aladjem 2005)	214	9	$\mathbf{X}_1$	2	$2 \le TNCCCC \le 32$	6
			$\mathbf{X}_2$	2		
			$\mathbf{X}_3$	1		
			$\mathbf{X}_4$	2		
			$\mathbf{X}_5$	2		
			$\mathbf{X}_6$	1		
			$\mathbf{X}_7$	2		
			$\mathbf{X}_8$	1		
			$\mathbf{X}_9$	1		

Cam belirleme verisindeki 9 değişken için MATLAB yazılımı kullanılarak oluşturulan saçılım grafiği Şekil 7.3'te verilmiştir.



Şekil 7.3. Cam belirleme verisindeki 9 değişken için MATLAB yazılımı kullanılarak oluşturulan saçılım grafiği.

Başlangıç aşamasında yapılan işlemler:

Cam belirleme verisinin tamamının yer aldığı grup G_0 olarak adlandırılsın. Cam belirleme verisinin $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_4$, $\mathbf{X}_5$ ve $\mathbf{X}_7$ değişkenleri alınarak ve MIXMOD yazılımı kullanılarak başlangıç aşamasında oluşturulan karma normal modeller için elde edilen BIC değerleri Tablo 7.2'de verilmiştir.

Tablo 7.2'deki sonuçlara göre G_0 olarak adlandırılan cam belirleme verisi için başlangıç aşamasında karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. G_{11} ve G_{12} olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.3'de verilmiştir.

Tablo 7.2. Cam belirleme verisi için MIXMOD yazılımı kullanılarak başlangıç aşamasında oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-144.545	261.770	2716.119
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**-776.932**	-340.200	1963.862
Çok değişkenli üç normal dağılımın karması için BIC değerleri	-721.299	-448.825	1814.975

Tablo 7.3. Cam belirleme verisinde G_0 grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	G_{11}	G_{12}	**Toplam**
G_0	140	74	214

Tablo 7.3'teki sonuçlara göre G_0 grubundaki 214 gözlem değerinden 140 tanesi G_{11} alt grubuna ve 74 tanesi G_{12} alt grubuna sınıflandırılmıştır.

Birinci aşamada yapılan işlemler:

Başlangıç aşamasının sonunda elde edilen G_{11} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{11} grup olarak alınsın. G_{11} grubundaki veriler için MIXMOD yazılımı kullanılarak birinci aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.4'de verilmiştir.

Tablo 7.4'deki sonuçlara göre G_{11} olarak adlandırılan gruptaki veriler için birinci aşamada karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. G_{2111} ve G_{2112} olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.5'te verilmiştir.

Tablo 7.4. Cam belirleme verisinin G_{11} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-1316.647	-709.082	738.0631
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**-1362.698**	-1054.407	425.495
Çok değişkenli üç normal dağılımın karması için BIC değerleri	-1297.876	-1109.620	357.311

Tablo 7.5. Cam belirleme verisinde G_{11} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	G_{2111}	G_{2112}	**Toplam**
G_{11}	83	57	140

Tablo 7.5'teki sonuçlara göre G_{11} grubundaki 140 gözlem değerinden 83 tanesi G_{2111} alt grubuna ve 57 tanesi G_{2112} alt grubuna sınıflandırılmıştır.

Başlangıç aşamasının sonunda elde edilen G_{12} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{12} grup olarak alınsın. G_{12} grubundaki veriler için MIXMOD yazılımı kullanılarak birinci aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.6'da verilmiştir.

Tablo 7.6'daki sonuçlara göre G_{12} olarak adlandırılan gruptaki veriler için birinci aşamada karma normal dağılım modeli oluşturulduğunda üç alt grup yapısı elde edilmiştir. G_{2121}, G_{2122} ve G_{2123} olarak adlandırılan bu üç alt gruptaki gözlemlerin dağılımı Tablo 7.7'de verilmiştir.

Tablo 7.6. Cam belirleme verisinin G_{12} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	233.261	383.076	1246.604
Çok değişkenli iki normal dağılımın karması için BIC değerleri	210.288	337.218	1130.583
Çok değişkenli üç normal dağılımın karması için BIC değerleri	**164.125**	323.187	1038.236
Çok değişkenli dört normal dağılımın karması için BIC değerleri	267.472	235.517	995.975

Tablo 7.7. Cam belirleme verisinde G_{12} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı			
Veri	G_{2121}	G_{2122}	G_{2123}	**Toplam**
G_{12}	28	28	18	74

Tablo 7.7'deki sonuçlara göre G_{12} grubundaki 74 gözlem değerinden 28 tanesi G_{2121} alt grubuna, 28 tanesi G_{2122} alt grubuna ve 18 tanesi G_{2123} alt grubuna sınıflandırılmıştır.

İkinci aşamada yapılan işlemler:

Birinci aşama sonunda elde edilen G_{2111} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{2111} grup olarak alınsın. G_{2111} grubundaki veriler için MIXMOD yazılımı kullanılarak ikinci aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.8'de verilmiştir.

Tablo 7.8. Cam belirleme verisinin G_{2111} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-1084.937	-877.465	23.845
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**-1093.863**	-975.916	-34.640
Çok değişkenli üç normal dağılımın karması için BIC değerleri	-1045.093	-1000.732	-105.0924

Tablo 7.8'deki sonuçlara göre G_{2111} olarak adlandırılan gruptaki veriler için ikinci aşamada karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. G_{321111} ve G_{321112} olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.9'da verilmiştir.

Tablo 7.9. Cam belirleme verisinde G_{2111} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	G_{321111}	G_{321112}	**Toplam**
G_{2111}	37	46	83

Tablo 7.9'daki sonuçlara göre G_{2111} grubundaki 83 gözlem değerinden 37 tanesi G_{321111} alt grubuna ve 46 tanesi G_{321112} alt grubuna sınıflandırılmıştır.

Birinci aşama sonunda elde edilen G_{2112} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{2112} grup olarak alınsın. G_{2112} grubundaki veriler için MIXMOD yazılımı kullanılarak ikinci aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.10'da verilmiştir.

Tablo 7.10. Cam belirleme verisinin G_{2112} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-428.137	-166.838	407.312
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-459.350	-302.922	320.089
Çok değişkenli üç normal dağılımın karması için BIC değerleri	**-491.792**	-320.488	298.880
Çok değişkenli dört normal dağılımın karması için BIC değerleri	-376.454	-311.400	259.235

Tablo 7.10'daki sonuçlara göre G_{2112} olarak adlandırılan gruptaki veriler için ikinci aşamada karma normal dağılım modeli oluşturulduğunda üç alt grup yapısı elde edilmiştir. G_{321121}, G_{321122} ve G_{321123} olarak adlandırılan bu üç alt gruptaki gözlemlerin dağılımı Tablo 7.11'de verilmiştir.

Tablo 7.11. Cam belirleme verisinde G_{2112} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı			
Veri	G_{321121}	G_{321122}	G_{321123}	**Toplam**
G_{2112}	26	18	13	57

Tablo 7.11'deki sonuçlara göre G_{2112} grubundaki 57 gözlem değerinden 26 tanesi G_{321121} alt grubuna, 18 tanesi G_{321122} alt grubuna ve 13 tanesi G_{321123} alt grubuna sınıflandırılmıştır.

Birinci aşama sonunda elde edilen G_{2121} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{2121} grup olarak alınsın. G_{2121} grubundaki veriler için MIXMOD yazılımı kullanılarak ikinci aşamada

oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.12'de verilmiştir.

Tablo 7.12. Cam belirleme verisinin G_{2121} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	115.7468	147.803	513.259
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**70.1632**	133.0939	472.680
Çok değişkenli üç normal dağılımın karması için BIC değerleri	87.356	153.150	452.322

Tablo 7.12'deki sonuçlara göre G_{2121} olarak adlandırılan gruptaki veriler için ikinci aşamada karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. G_{321211} ve G_{321212} olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.13'te verilmiştir.

Tablo 7.13. Cam belirleme verisinde G_{2121} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	G_{321211}	G_{321212}	**Toplam**
G_{2121}	20	8	28

Tablo 7.13'teki sonuçlara göre G_{2121} grubundaki 28 gözlem değerinden 20 tanesi G_{321211} alt grubuna ve 8 tanesi G_{321212} alt grubuna sınıflandırılmıştır.

Birinci aşama sonunda elde edilen G_{2122} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{2122} grup olarak alınsın. G_{2122} grubundaki veriler için MIXMOD yazılımı kullanılarak ikinci aşamada

oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.14'de verilmiştir.

Tablo 7.14. Cam belirleme verisinin G_{2122} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-106.485	75.671	366.292
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**-131.540**	-43.899	246.304
Çok değişkenli üç normal dağılımın karması için BIC değerleri	-110.935	-69.302	211.827

Tablo 7.14'deki sonuçlara göre G_{2122} olarak adlandırılan gruptaki veriler için ikinci aşamada karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. G_{321221} ve G_{321222} olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.15'te verilmiştir.

Tablo 7.15. Cam belirleme verisinde G_{2122} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	G_{321221}	G_{321222}	**Toplam**
G_{2122}	21	7	28

Tablo 7.15'teki sonuçlara göre G_{2122} grubundaki 28 gözlem değerinden 21 tanesi G_{321221} alt grubuna ve 7 tanesi G_{321222} alt grubuna sınıflandırılmıştır.

Birinci aşama sonunda elde edilen G_{2123} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{2123} grup olarak alınsın. G_{2123} grubundaki veriler için MIXMOD yazılımı kullanılarak ikinci aşamada

oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.16'da verilmiştir.

Tablo 7.16. Cam belirleme verisinin G_{2123} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	56.358	111.001107	324.958409
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**41.080**	90.028213	286.031117
Çok değişkenli üç normal dağılımın karması için BIC değerleri	66.281	98.218471	281.443275

Tablo 7.16'daki sonuçlara göre G_{2123} olarak adlandırılan gruptaki veriler için ikinci aşamada karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. G_{321231} ve G_{321232} olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.17'de verilmiştir.

Tablo 7.17. Cam belirleme verisinde G_{2123} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	G_{321231}	G_{321232}	**Toplam**
G_{2123}	7	11	18

Tablo 7.17'teki sonuçlara göre G_{2123} grubundaki 18 gözlem değerinden 7 tanesi G_{321231} alt grubuna ve 11 tanesi G_{321232} alt grubuna sınıflandırılmıştır.

Üçüncü aşamada yapılan işlemler:

İkinci aşama sonunda elde edilen G_{321111} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321111} grup olarak alınsın. G_{321111} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.18'de verilmiştir.

Tablo 7.18. Cam belirleme verisinin G_{321111} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-513.835	-483.728	-31.380
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**-545.419**	-500.599	-50.240
Çok değişkenli üç normal dağılımın karması için BIC değerleri	-471.731	-510.721	-75.872

Tablo 7.18'deki sonuçlara göre G_{321111} olarak adlandırılan gruptaki veriler için üçüncü aşamada karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. $G_{43211111}$ ve $G_{43211112}$ olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.19'da verilmiştir.

Tablo 7.19. Cam belirleme verisinde G_{321111} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	$G_{43211111}$	$G_{43211112}$	**Toplam**
G_{321111}	10	27	37

Tablo 7.19'daki sonuçlara göre G_{321111} grubundaki 37 gözlem değerinden 10 tanesi $G_{43211111}$ alt grubuna ve 27 tanesi $G_{43211112}$ alt grubuna sınıflandırılmıştır.

İkinci aşama sonunda elde edilen G_{321112} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321112} grup olarak alınsın. G_{321112} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.20'de verilmiştir.

Tablo 7.20. Cam belirleme verisinin G_{321112} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-653.812	-582.331	18.440
Çok değişkenli iki normal dağılımın karması için BIC değerleri	**-694.525**	-597.743	-40.716
Çok değişkenli üç normal dağılımın karması için BIC değerleri	-651.952	-622.306	-69.525

Tablo 7.20'deki sonuçlara göre G_{321112} olarak adlandırılan gruptaki veriler için üçüncü aşamada karma normal dağılım modeli oluşturulduğunda iki alt grup yapısı elde edilmiştir. $G_{43211121}$ ve $G_{43211122}$ olarak adlandırılan bu iki alt gruptaki gözlemlerin dağılımı Tablo 7.21'de verilmiştir.

Tablo 7.21. Cam belirleme verisinde G_{321112} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı		
Veri	$G_{43211121}$	$G_{43211122}$	**Toplam**
G_{321112}	15	31	46

Tablo 7.21'daki sonuçlara göre G_{321112} grubundaki 46 gözlem değerinden 15 tanesi $G_{43211121}$ alt grubuna ve 31 tanesi $G_{43211122}$ alt grubuna sınıflandırılmıştır.

İkinci aşama sonunda elde edilen G_{321121} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321121} grup olarak alınsın. G_{321121} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.22'de verilmiştir.

Tablo 7.22. Cam belirleme verisinin G_{321121} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-243.290**	-158.627	121.205
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-206.021	-161.093	118.482

Tablo 7.22'deki sonuçlara göre G_{321121} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321121} grubu bir kümelenme olarak elde edilmiştir.

İkinci aşama sonunda elde edilen G_{321122} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321122} grup olarak alınsın. G_{321122} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.23'te verilmiştir.

Tablo 7.23. Cam belirleme verisinin G_{321122} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-204.082**	-157.577	63.146
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-195.737	-167.418	19.687

Tablo 7.23'deki sonuçlara göre G_{321122} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321122} grubu bir kümelenme olarak elde edilmiştir.

İkinci aşama sonunda elde edilen G_{321123} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321123} grup olarak alınsın. G_{321123} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.24'de verilmiştir.

Tablo 7.24. Cam belirleme verisinin G_{321123} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-156.963	-108.814	37.356
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-146.788	-114.352	19.871

Tablo 7.24'deki sonuçlara göre G_{321123} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321123} grubu bir kümelenme olarak elde edilmiştir.

İkinci aşama sonunda elde edilen G_{321211} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321211} grup olarak alınsın. G_{321211} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.25'te verilmiştir.

Tablo 7.25. Cam belirleme verisinin G_{321211} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	46.245	104.005	348.923
Çok değişkenli iki normal dağılımın karması için BIC değerleri	29.864	89.041	315.320
Çok değişkenli üç normal dağılımın karması için BIC değerleri	**21.649**	84.661	306.726
Çok değişkenli dört normal dağılımın karması için BIC değerleri	33.862	97.908	306.368

Tablo 7.25'teki sonuçlara göre G_{321211} olarak adlandırılan gruptaki veriler için üçüncü aşamada karma normal dağılım modeli oluşturulduğunda üç alt grup yapısı elde edilmiştir. $G_{43212111}$, $G_{43212112}$ ve $G_{43212113}$ olarak adlandırılan bu üç alt gruptaki gözlemlerin dağılımı Tablo 7.26'da verilmiştir.

Tablo 7.26. Cam belirleme verisinde G_{321211} grubundaki gözlemlerin alt gruplara dağılımı.

	Verideki Alt Grup Yapısı			
Veri	$G_{43212111}$	$G_{43212112}$	$G_{43212113}$	**Toplam**
G_{321211}	6	8	6	20

Tablo 7.26'daki sonuçlara göre G_{321211} grubundaki 20 gözlem değerinden 6 tanesi $G_{43212111}$ alt grubuna, 8 tanesi $G_{43212112}$ alt grubuna ve 6 tanesi $G_{43212113}$ alt grubuna sınıflandırılmıştır.

İkinci aşama sonunda elde edilen G_{321212} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321212} grup olarak alınsın. G_{321212} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.27'de verilmiştir.

Tablo 7.27. Cam belirleme verisinin G_{321212} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**18.814**	39.595	132.421
Çok değişkenli iki normal dağılımın karması için BIC değerleri	26.723	21.280	133.423

Tablo 7.27'deki sonuçlara göre G_{321212} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321212} grubu bir kümelenme olarak elde edilmiştir.

İkinci aşama sonunda elde edilen G_{321221} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321221} grup olarak alınsın. G_{321221} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.28'de verilmiştir.

Tablo 7.28. Cam belirleme verisinin G_{321221} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-147.724**	-45.088	190.240
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-142.016	-36.160	168.364

Tablo 7.28'deki sonuçlara göre G_{321221} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321221} grubu bir kümelenme olarak elde edilmiştir.

İkinci aşama sonunda elde edilen G_{321222} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321222} grup olarak alınsın. G_{321222} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.29'da verilmiştir.

Tablo 7.29'daki sonuçlara göre G_{321222} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321222} grubu bir kümelenme olarak elde edilmiştir.

Tablo 7.29. Cam belirleme verisinin G_{321222} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-52.117**	38.895	115.068
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-48.234	40.626	119.609

İkinci aşama sonunda elde edilen G_{321231} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321231} grup olarak alınsın. G_{321231} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.30'da verilmiştir.

Tablo 7.30. Cam belirleme verisinin G_{321231} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-49.315**	27.518	112.302
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-35.611	17.819	102.416

Tablo 7.30'daki sonuçlara göre G_{321231} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321231} grubu bir kümelenme olarak elde edilmiştir.

İkinci aşama sonunda elde edilen G_{321232} alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. G_{321232} grup olarak alınsın. G_{321232} grubundaki veriler için MIXMOD yazılımı kullanılarak üçüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.31'de verilmiştir.

Tablo 7.31. Cam Belirleme verisinin G_{321232} grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**18.248**	56.918	186.041
Çok değişkenli iki normal dağılımın karması için BIC değerleri	21.380	47.863	162.153

Tablo 7.31'deki sonuçlara göre G_{321232} grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda G_{321232} grubu bir kümelenme olarak elde edilmiştir.

Dördüncü aşamada yapılan işlemler:

Üçüncü aşama sonunda elde edilen $G_{43211111}$ alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. $G_{43211111}$ grup olarak tanımlansın. $G_{43211111}$ grubundaki veriler için MIXMOD yazılımı kullanılarak dördüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.32'de verilmiştir.

Tablo 7.32. Cam belirleme verisinin $G_{43211111}$ grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-153.015	**-157.471**	-28.202
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-144.622	-124.131	-15.371

Tablo 7.32'deki sonuçlara göre $G_{43211111}$ grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda $G_{43211111}$ grubu bir kümelenme olarak elde edilmiştir.

Üçüncü aşama sonunda elde edilen $G_{43211112}$ alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. $G_{43211112}$ grup olarak tanımlansın. $G_{43211112}$ grubundaki veriler için MIXMOD yazılımı kullanılarak dördüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.33'te verilmiştir.

Tablo 7.33. Cam belirleme verisinin $G_{43211112}$ grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-441.829**	-422.343	-59.472
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-412.389	-411.693	-116.848

Tablo 7.33'teki sonuçlara göre $G_{43211112}$ grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda $G_{43211112}$ grubu bir kümelenme olarak elde edilmiştir.

Üçüncü aşama sonunda elde edilen $G_{43211121}$ alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. $G_{43211121}$ grup olarak tanımlansın. $G_{43211121}$ grubundaki veriler için MIXMOD yazılımı kullanılarak dördüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.34'de verilmiştir.

Tablo 7.34. Cam belirleme verisinin $G_{43211121}$ grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-228.259**	-206.997	-11.084
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-205.655	-202.456	-18.564

Tablo 7.34'deki sonuçlara göre $G_{43211121}$ grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda $G_{43211121}$ grubu bir kümelenme olarak elde edilmiştir.

Üçüncü aşama sonunda elde edilen $G_{43211122}$ alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. $G_{43211122}$ grup olarak tanımlansın. $G_{43211122}$ grubundaki veriler için MIXMOD yazılımı kullanılarak dördüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.35'te verilmiştir.

Tablo 7.35. Cam belirleme verisinin $G_{43211122}$ grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-527.689**	-443.599	-29.466
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-512.459	-443.116	-60.460

Tablo 7.35'teki sonuçlara göre $G_{43211122}$ grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda $G_{43211122}$ grubu bir kümelenme olarak elde edilmiştir.

Üçüncü aşama sonunda elde edilen $G_{43212111}$ alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. $G_{43212111}$ grup olarak tanımlansın. $G_{43212111}$ grubundaki veriler için MIXMOD yazılımı kullanılarak dördüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.36'da verilmiştir.

Tablo 7.36'daki sonuçlara göre $G_{43212111}$ grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda $G_{43212111}$ grubu bir kümelenme olarak elde edilmiştir.

Tablo 7.36. Cam belirleme verisinin $G_{43212111}$ grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	**-33.378**	8.648	89.521
Çok değişkenli iki normal dağılımın karması için BIC değerleri	7.831	78.419	95.677

Üçüncü aşama sonunda elde edilen $G_{43212112}$ alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. $G_{43212112}$ grup olarak tanımlansın. $G_{43212112}$ grubundaki veriler için MIXMOD yazılımı kullanılarak dördüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.37'de verilmiştir.

Tablo 7.37'deki sonuçlara göre $G_{43212112}$ grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda $G_{43212112}$ grubu bir kümelenme olarak elde edilmiştir.

Tablo 7.37. Cam belirleme verisinin $G_{43212112}$ alt grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	-8.428	92.881	**-49.907**
Çok değişkenli iki normal dağılımın karması için BIC değerleri	-6.187	-17.020	75.402

Üçüncü aşama sonunda elde edilen $G_{43212113}$ alt grubundaki verilerin çok değişkenli karma normal dağılımdan geldiği varsayılsın. $G_{43212113}$ grup olarak tanımlansın. $G_{43212113}$ grubundaki veriler için MIXMOD yazılımı kullanılarak dördüncü aşamada oluşturulan karma normal modeller ve elde edilen BIC değerleri Tablo 7.38'de verilmiştir.

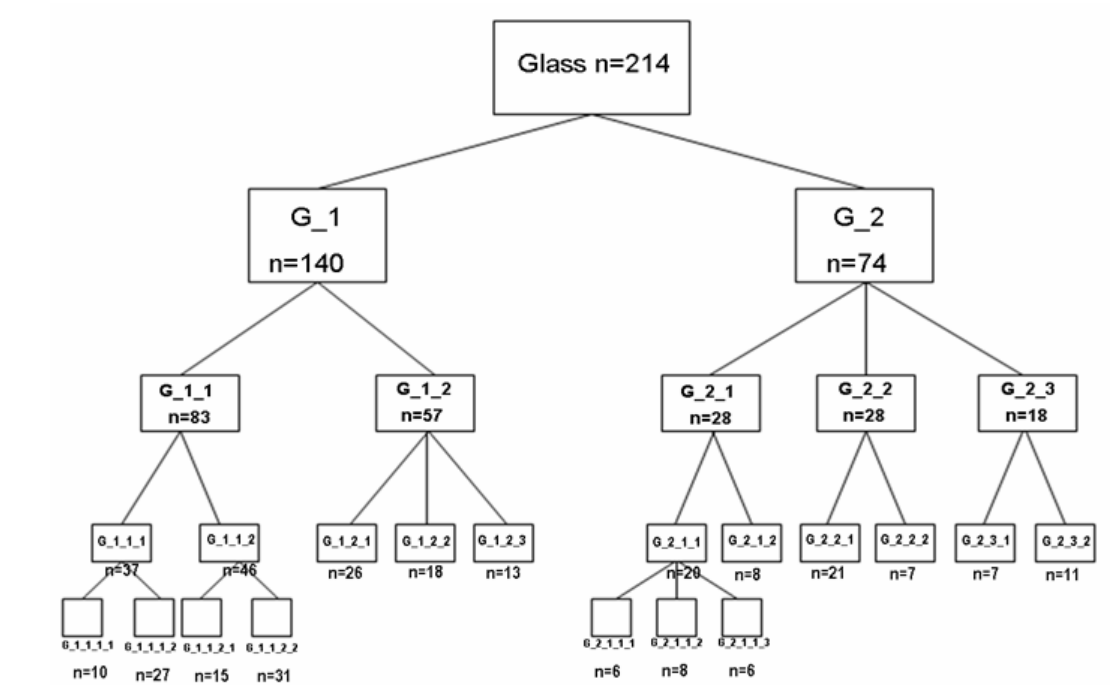
Tablo 7.38. Cam belirleme verisinin $G_{43212113}$ grubundaki veriler için MIXMOD yazılımı kullanılarak oluşturulan karma normal modeller ve elde edilen BIC değerleri.

	Normal Model Tipi		
Karma Normal Model	Elipsoidal pk_Lk_Ck	Köşegen pk_Lk_Bk	Küresel pk_Lk_I
Çok değişkenli normal model için BIC değerleri	23.953	109.889	**-31.282**
Çok değişkenli iki normal dağılımın karması için BIC değerleri	29.720	50.264	79.947

Tablo 7.38'deki sonuçlara göre $G_{43212113}$ grubu alt grup içermemektedir. Bu nedenle çok değişkenli verideki grupların inceltilerek karma kümeleme analizi sonucunda $G_{43212113}$ grubu bir kümelenme olarak elde edilmiştir.

Dördüncü aşama sonunda cam belirleme verisindeki tüm kümelenmeler belirlenmiştir. Elde edilen tüm kümelenmeler homojen yapıdadır. Bu çalışma sonucunda cam belirleme verisi için çok değişkenli verideki grupların inceltilerek

kümelenmesi yöntemini kullanarak oluşturulan 15 kümelenme yapısı ve her kümedeki gözlem sayısı Şekil 7.4'de gösterilmiştir. Cam belirleme verisi için çok değişkenli verideki grupların inceltilerek kümelenmesi yöntemini kullanarak oluşturulan 15 kümelenme yapısı için ortalama vektörleri ve kovaryans matrisleri Tablo 7.39'da verilmiştir.



Şekil 7.4. Cam belirleme verisi için çok değişkenli verideki grupların inceltilerek kümelenmesi yöntemini kullanarak oluşturulan 15 kümelenme yapısı ve her kümedeki gözlem sayısı.

Tablo 7.39. Cam belirleme verisi için çok değişkenli verideki grupların inceltilerek kümelenmesi yöntemini kullanarak oluşturulan 15 kümelenme yapısı için ortalama vektörleri ve kovaryans matrisleri.

Küme	Küme Adı	Gözlem Sayısı	Ortalama Vektörü	Kovaryans Matrisi				
1	$G_{43211111}$	10	1.5164	0.0000	0.0000	0.0000	0.0000	0.0000
			13.1930	0.0000	0.0416	0.0000	0.0000	0.0000
			1.2740	0.0000	0.0000	0.0027	0.0000	0.0000
			72.7590	0.0000	0.0000	0.0000	0.0163	0.0000
			8.5190	0.0000	0.0000	0.0000	0.0000	0.0445
2	$G_{43211112}$	27	1.5163	0.0000	0.0000	0.0000	0.0000	0.0000
			13.0941	0.0000	0.0620	-0.0015	-0.0450	-0.0106
			1.5230	0.0000	-0.0015	0.0035	-0.0027	-0.0014
			72.9459	0.0000	-0.0450	-0.0027	0.0671	-0.0036
			8.0356	0.0000	-0.0106	-0.0014	-0.0036	0.0150
3	$G_{43211121}$	15	1.5183	0.0000	0.0000	0.0000	0.0000	0.0000
			13.2073	0.0000	0.0489	0.0061	-0.0300	-0.0310
			1.2393	0.0000	0.0061	0.0087	-0.0105	-0.0064
			72.5313	0.0000	-0.0300	-0.0105	0.0764	-0.0089
			8.3840	0.0000	-0.0310	-0.0064	-0.0089	0.0421
4	$G_{43211122}$	31	1.5176	0.0000	0.0000	0.0000	0.0000	0.0000
			12.8913	0.0000	0.0786	-0.0085	-0.0315	-0.0431
			1.2516	0.0000	-0.0085	0.0113	-0.0013	-0.0010
			73.0268	0.0000	-0.0315	-0.0013	0.0520	-0.0089
			8.5274	0.0000	-0.0431	-0.0010	-0.0089	0.0521
5	G_{321121}	26	1.5175	0.0000	0.0002	-0.0002	-0.0001	0.0002
			13.2358	0.0002	0.1831	-0.0702	-0.0939	-0.0313
			1.4323	-0.0002	-0.0702	0.1130	-0.0088	-0.0895
			72.8327	-0.0001	-0.0939	-0.0088	0.1185	0.0496
			8.4862	0.0002	-0.0313	-0.0895	0.0496	0.1642
6	G_{321122}	18	1.5219	0.0000	0.0000	-0.0001	0.0000	0.0001
			13.7222	0.0000	0.1980	-0.0281	-0.1111	-0.0763
			0.8006	-0.0001	-0.0281	0.0337	0.0069	-0.0273
			71.8150	0.0000	-0.1111	0.0069	0.0795	0.0478
			9.6150	0.0001	-0.0763	-0.0273	0.0478	0.1170
7	G_{321123}	13	1.5185	0.0000	0.0002	-0.0001	-0.0001	0.0001
			13.4454	0.0002	0.1547	-0.0370	-0.1077	-0.0445
			1.4408	-0.0001	-0.0370	0.0289	0.0198	-0.0147
			72.1215	-0.0001	-0.1077	0.0198	0.0916	0.0331
			8.8000	0.0001	-0.0445	-0.0147	0.0331	0.0753
7	G_{321123}	13	1.5185	0.0000	0.0002	-0.0001	-0.0001	0.0001
			13.4454	0.0002	0.1547	-0.0370	-0.1077	-0.0445
			1.4408	-0.0001	-0.0370	0.0289	0.0198	-0.0147
			72.1215	-0.0001	-0.1077	0.0198	0.0916	0.0331
			8.8000	0.0001	-0.0445	-0.0147	0.0331	0.0753
8	$G_{43212111}$	6	1.5183	0.0000	0.0004	0.0004	-0.0038	0.0027
			14.3767	0.0004	0.1776	0.0149	-0.1823	0.0060
			1.8083	0.0004	0.0149	0.0419	-0.1542	0.1270
			72.4850	-0.0038	-0.1823	-0.1542	1.6973	-0.9072
			9.5683	0.0027	0.0060	0.1270	-0.9072	1.2678

Tablo 7.39 (Devamı)**.** Cam belirleme verisi için çok değişkenli verideki grupların inceltilerek kümelenmesi yöntemini kullanarak oluşturulan 15 kümelenme yapısı için ortalama vektörleri ve kovaryans matrisleri.

Küme	Küme Adı	Gözlem Sayısı	Ortalama Vektörü	Kovaryans Matrisi				
9	$G_{43212112}$	8	1.5212	0.0000	0.0007	-0.0003	-0.0020	0.0021
			13.2513	0.0007	0.1386	-0.0436	-0.2937	0.1903
			1.5563	-0.0003	-0.0436	0.0287	0.1128	-0.1045
			72.7438	-0.0020	-0.2937	0.1128	0.7361	-0.6601
			11.7013	0.0021	0.1903	-0.1045	-0.6601	0.8715
10	$G_{43212113}$	6	1.5279	0.0000	-0.0008	0.0000	-0.0015	0.0042
			12.4700	-0.0008	1.2194	0.0618	-0.9680	-1.6740
			0.8367	0.0000	0.0618	0.0348	-0.0848	-0.0818
			71.6983	-0.0015	-0.9680	-0.0848	1.3730	0.5895
			13.7767	0.0042	-1.6740	-0.0818	0.5895	3.6524
11	G_{321212}	8	1.5211	0.0000	0.0013	-0.0009	-0.0026	0.0024
			13.9300	0.0013	1.0882	0.1566	-1.1578	-0.3451
			1.4563	-0.0009	0.1566	0.2284	-0.0089	-0.5048
			72.4800	-0.0026	-1.1578	-0.0089	1.8681	0.2765
			8.1763	0.0024	-0.3451	-0.5048	0.2765	1.5848
12	G_{321221}	21	1.5161	0.0000	-0.0006	0.0003	-0.0006	0.0004
			14.7390	-0.0006	0.4694	-0.3065	0.2811	-0.3447
			2.2119	0.0003	-0.3065	0.2785	-0.2073	0.2707
			73.3205	-0.0006	0.2811	-0.2073	0.2802	-0.2427
			8.6652	0.0004	-0.3447	0.2707	-0.2427	0.3129
13	G_{321222}	7	1.5212	0.0000	-0.0051	0.0010	-0.0037	0.0057
			13.6257	-0.0051	1.5394	-0.4213	1.0166	-1.6835
			1.4757	0.0010	-0.4213	0.2711	-0.1637	0.4144
			71.9171	-0.0037	1.0166	-0.1637	0.7873	-1.1757
			9.9814	0.0057	-1.6835	0.4144	-1.1757	1.8986
14	G_{321231}	7	1.5201	0.0000	-0.0015	0.0008	-0.0011	0.0034
			12.4429	-0.0015	1.4266	-0.5403	0.4784	-0.3700
			1.1857	0.0008	-0.5403	0.2427	-0.2806	0.2808
			73.3029	-0.0011	0.4784	-0.2806	0.4612	-0.3348
			10.9186	0.0034	-0.3700	0.2808	-0.3348	2.0242
15	G_{321232}	11	1.5202	0.0000	-0.0007	0.0005	-0.0007	0.0024
			12.8091	-0.0007	1.3203	-0.3720	0.1630	-0.2480
			1.3700	0.0005	-0.3720	0.2689	-0.2706	0.0399
			73.1382	-0.0007	0.1630	-0.2706	0.4354	-0.0828
			10.6827	0.0024	-0.2480	0.0399	-0.0828	1.5829

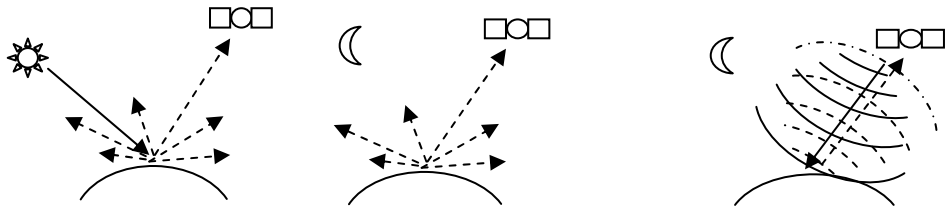
8. ÇOK DEĞİŞKENLİ VERİDEKİ GRUPLARIN İNCELTİLEREK KARMA KÜMELEME ANALİZİ İLE ÇOK BANTLI UYDU GÖRÜNTÜSÜNÜN ALAN BAZINDA EĞİTİMLİ SINIFLANDIRILMASI

Bu bölümde gelişmiş uydu teknolojileri kullanılarak uzaktan algılanmış ve sayısallaştırılmış, birbirinden ayırt edilebilen bölümler ya da alanlar içeren, çok bantlı uydu görüntü verisi grupların inceltilerek karma kümeleme analizi ile alan-bazında eğitimli sınıflandırılmıştır.

Uzaktan algılama dünya üzerindeki nesne ya da alanlar hakkında bilgiyi bu nesne yada alanlarla doğrudan yani fiziksel bir temas olmaksızın bilgi toplamadır. Bu anlamda insanlar görerek, işiterek, kokuları algılayarak uzaktan algılama gerçekleştirirler. Geliştirilen birçok uzaktan algılama aracı ile nesnelerin yüzeylerinden yansıttıkları veya yaydıkları elektromanyetik enerji ölçülerek kayıt edilir.

Uzaktan algılama teknikleri elektromanyetik spektrumun değişik dalga boylarında dünya yüzeyinden görüntüler almamızı sağlar. Bir uzaktan algılanmış görüntünün en önemli karakteristiklerinden biri dalga boyu bölgesidir. Dünya gözlem uydu sensörleri dünya yüzeyi tarafından elektromanyetik yayılım ya da yansımayı ölçer. Ölçülen radyomanyetik yansımalar yeryüzü yüzeyinin karakteristiklerine bağlıdır. Görüntü verisinden arazi bilgisinin elde edilmesini yeryüzü örtüsü ile ölçüm değerleri arasındaki ilişki sağlar.

Aktif algılayıcılar yeryüzüne elektromanyetik spektrumun mikrodalga değişim aralığında ışın sinyalleri gönderirler ve geri dönen miktarı ardışık zaman aralıklarında ölçerler. Pasif algılayıcılar ise dünya yüzeyinden yayılan termal infrared ışınları ya da güneşten çıkıp yüzeyinden yansıyan görülebilir ve infrarede yakın ışımaları ölçer. Thematic Mapper sensor (TM), Landsat Earth gözlem uydu programında kullanılmaktadır. TM bir pasif algılayıcıdır ve yeryüzüne çarpıp yansıyan güneş ışığı yayılımlarını ölçer.



(a)	(b)	(c)
Güneş ışığı yansıması (Görülebilir infra-red)	Radyasyon yayılımı (Thermal infra-red)	Aktif sensörden radyosyon yayılmı (Mikrodalga)

Şekil 8.1. Algılayıcı tipleri.

8.1. Landsat Uydusu ve Algılayıcı Özellikleri

Yeryüzü kaynakları sivil uydu projesi ilk defa 1960'lı yılların ortasında gelişmeye başlamış ve Landsat programının oluşması sağlanmıştır. İlk Landsat uydusu NASA tarafından 1972 yılında uzaya gönderilmiştir. Bu uyduyu daha sonra Landsat 2, 3, 4 ve 5 uyduları sırasıyla 1975, 1978, 1982 ve 1984 yıllarında takip etmiştir. Bunlardan Landsat 5 uydusu halen aktif durumdadır. 1993 yılında gönderilen Landsat 6 uydusu aktif olamamış bunun üzerine 1999 yılında Landsat 7 uydusu gönderilmiştir. Zaman içerisinde ticarileşen Landsat uyduları 1992 yılından itibaren NASA ve ABD savunma birimi taraflarından idare edilmektedir.

Çok bantlı algılayıcılar elektromanyetik spektrumun farklı bölümlerindeki yansımaları eş zamanlı olarak ölçerler. Bir algılayıcının spectral çözünürlüğü dalga boyları ya da dalga boyu aralıkları ile ifade edilen band sayısı ve onların spektrum içerisindeki konumları ile belirlenir. Landsat 1, Landsat 2 ve Landsat 3 uydularının ana algılayıcısı olan Multi Spectral Scanner (MSS) spektrumun görülebilir ve infrarede yakın değişim aralığında 4 band üzerinden işlem yapmaktadır. Landsat 4 ve 5 uydularındaki Thematic Mapper (TM) sensörü ise yedi spektral band üzerinden ölçümler yapmaktadır. MSS ve TM bandlarının karakteristikleri farklı tipteki yeryüzü kaynaklarının tespit edilmesi ve izlenmesini maksimum kapasitede gerçekleştirecek şekilde seçilir.

Tablo 8.1. Landsat 4 ve Landsat 5 uydusunun Multi Spectral Scanner (MSS) algılayıcısının bantları, veri topladıkları dalga boyları ve çözünürlükleri.

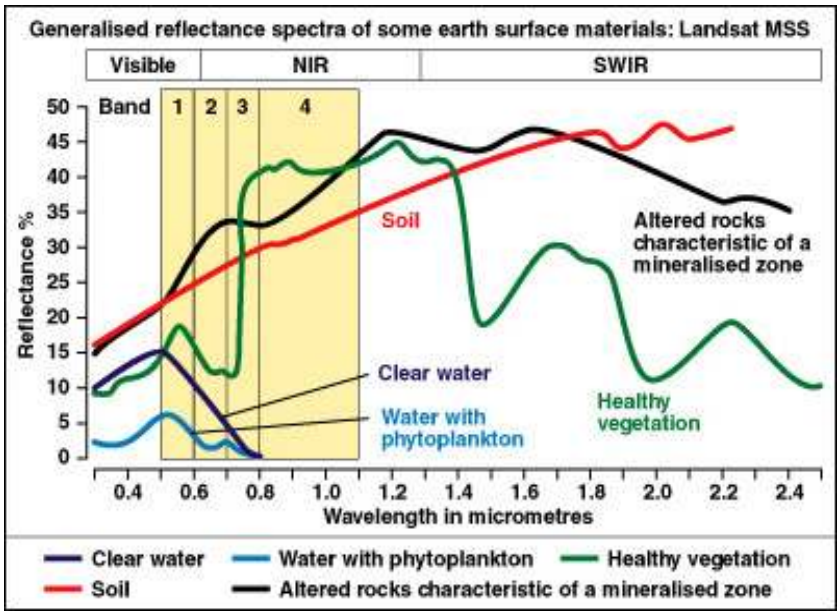
Uydu	Kanal (Bant)	Veri Topladığı Dalga Boyu (Mikrometre)	Çözünürlük (Metre)
LANDSAT MSS	Bant1	0.5 - 0.6	80
	Bant2	0.6 - 0.7	80
	Bant3	0.7 - 0.8	80
	Bant4	0.8 - 1.1	80

Tablo 8.2. Landsat 4 ve 5 uydusunun Thematic Mapper (TM) algılayıcısının bantları, veri topladıkları dalga boyları ve çözünürlükleri.

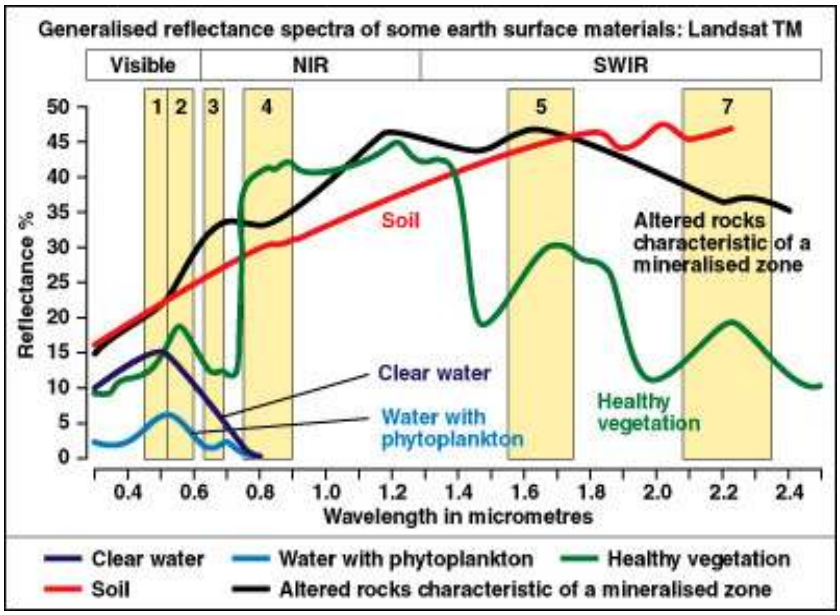
Uydu	Kanal (Bant)	Veri Topladığı Dalga Boyu (Mikrometre)	Çözünürlük (Metre)
LANDSAT TM	Bant1	0.45 - 0.52	30
	Bant2	0.52 - 0.60	30
	Bant3	0.63 - 0.69	30
	Bant4	0.75 - 0.90	30
	Bant5	1.55 - 1.75	30
	Bant6	10.40 - 12.5	120
	Bant7	2.08 - 2.35	30

TM bandları ve bunların belirli kompozisyonları farklı türde yeryüzü kaynaklarını belirlemede kullanılır. Örneğin 4. band 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 kombinasyonları 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.



Şekil 8.2. Landsat MSS Uydusunun bantları ve veri topladığı dalga boyları.



Şekil 8.3. Landsat TM Uydusunun bantları ve veri topladığı dalga boyları.

8.2. YÖNTEM

Bu bölümde çok bantlı uydu görüntü verisinin çok değişkenli verideki grupların inceltilerek karma kümeleme analizi incelenmiştir. Çok değişkenli karma normal dağılım modellerinin oluşturulması, parametre tahminlerinin hesaplanması ve grafiksel gösterimler için MIXMOD yazılımı (Biernacki ve ark. 2006) kullanılmıştır. Ayrıca oluşturulan çok değişkenli karma normal dağılım modellerinin 3 boyutlu grafiklerinin çiziminde EMMIX yazılımı (McLachlan ve ark. 1999) kullanılmıştır. Karma kümeleme analizinde çok değişkenli karma normal dağılım modelindeki parametreler, karma en çok olabilirlik yöntemiyle birlikte EM algoritması (McLachlan ve Krishnan 1997) kullanılarak tahmin edilmiştir. En iyi ya da uygun kümeleme yapısı, bir bilgi kriterinin kullanılmasıyla elde edilir. Farklı bilgi kriterleri bulunmakla birlikte bu çalışmada Bayesci bilgi kriteri (BIC) (Schwarz 1978) kullanılmıştır.

Alan bazında eğitimli sınıflandırma yönteminde ayrıştırma fonksiyonu olarak Bhattacharyya uzaklığı (Mac ve Barnard 1996, Pichler ve ark. 1996) kullanılmıştır. Bhattacharyya uzaklığı farklı kovaryansa sahip iki çok değişkenli normal dağılım arasındaki uzaklığın ölçümünde kullanılabilir (Mak ve Barnard 1996, Pichler ve ark. 1996). Sınıflandırma için karar kuralı elde edilen Bhattacharyya uzaklık değerlerine göre oluşturulmuştur. Çok değişkenli verideki grupların inceltilerek karma kümeleme analizi ile alan bazında eğitimli sınıflandırmada elde edilen sınıflandırma hata matrisi, genel doğruluk, kullanıcı doğruluğu ve üretici doğruluğu (Richards 1986) belirtilmiştir.

8.2.1. Çok Bantlı Uydu Görüntü Verisinin Grupların İnceltilmesiyle Karma Kümeleme Analizi İle Alan Bazında Eğitimli Sınıflandırılması İçin Kontrol Ve Test Gruplarının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntü verisinin iç içe karma kümeleme analizi kullanılarak alan bazında eğitimli sınıflandırılması için öncelikle görüntü verisindeki test grupları (TG) ve kontrol grupları (CG) belirlenmelidir. Eğitimli

sınıflandırma çerçevesinde kontrol grupları, bölgede ilgilenilen türdeki alanlardır. Bölgedeki belirlenen kontrol sınıfları dışındaki tüm sınıflar test sınıfları olarak adlandırılır. Test sınıfları için sadece bölgenin uzaktan algılanmış ve sayısallaştırılmış çok bantlı uydu görüntü verisi kullanılır ve yansıma özellikleri belirlenir. Test sınıfları için bölgede arazi çalışmaları yapılmaz.

8.2.2. Çok Bantlı Uydu Görüntü Verisinin Grupların İnceltilmesiyle Karma Kümeleme Analizi İle Alan Bazında Eğitimli Sınıflandırılması İçin Ayrıştırma Fonksiyonu

Çok değişkenli verinin alan bazında eğitimli sınıflandırılmasında ayrıştırma fonksiyonu olarak Bhattacharyya uzaklığı kullanılabilir. Diğer bir ifadeyle alan bazında eğitimli sınıflandırmada test grupları için oluşturulan çok değişkenli normal dağılım modelleriyle kontrol grupları için oluşturulan çok değişkenli normal dağılım modellerini karşılaştırmak amacıyla Bhattacharyya uzaklığı kullanılır. Bhattacharyya uzaklığı farklı kovaryans matrislerine sahip iki çok değişkenli normal dağılım modeli arasındaki uzaklığı hesaplamada kullanılır. i-inci TG ve j-inci CG arasındaki $D_{i,j}^{TG,CG}$ ile gösterilen Bhattacharyya uzaklığı (Mak and Barnard 1996, Pichler ve ark. 1996),

$$D_{i,j}^{TG,CG} = \frac{1}{8}\left(\boldsymbol{\mu}_j^{CG} - \boldsymbol{\mu}_i^{TG}\right)^T \left[\frac{\boldsymbol{\Sigma}_i^{TG} + \boldsymbol{\Sigma}_j^{CG}}{2}\right]^{-1} \left(\boldsymbol{\mu}_j^{CG} - \boldsymbol{\mu}_i^{TG}\right) + \frac{1}{2}\ln \frac{\left|\frac{\boldsymbol{\Sigma}_i^{TG} + \boldsymbol{\Sigma}_j^{CG}}{2}\right|}{\sqrt{\left|\boldsymbol{\Sigma}_i^{TG}\right|}\sqrt{\left|\boldsymbol{\Sigma}_j^{CG}\right|}} \quad (8.1)$$

formundadır. Burada $\boldsymbol{\mu}_i$ ve $\boldsymbol{\Sigma}_i$ çok değişkenli normal dağılımın sırasıyla ortalama vektörü ve kovaryans matrisidir. Test grupları ile kontrol grupları arasındaki Bhattacharyya uzaklık değerleri kullanılarak bir uzaklıklar matrisi oluşturulur.

8.2.3. Çok Bantlı Uydu Görüntü Verisinin Grupların İnceltilmesiyle Karma Kümeleme Analizi İle Alan Bazında Eğitimli Sınıflandırılması İçin Karar Kuralı

Alan bazında eğitimli sınıflandırmada karar kuralı, Bhattacharyya uzaklıklarına dayalı olarak tanımlanır. Karar kuralına göre bir test grubu aralarındaki Bhattacharyya uzaklığı minimum olan bir kontrol grubuna atanır. Alan bazında eğitimli sınıflandırmanın son aşamasında sınıflandırma hata matrisi oluşturulur. Bu sayede üretici doğruluğu, kullanıcı doğruluğu ve genel doğruluk cinsinden sınıflandırmanın sonucu yorumlanır.

8.2.4. Çok Bantlı Uydu Görüntü Verisinin Grupların İnceltilmesiyle Karma Kümeleme Analizi İle Alan Bazında Eğitimli Sınıflandırılması İçin Oluşturulan Algoritma

Çok bantlı uydu görüntü verisinin karma kümeleme analizi kullanılarak alan bazında eğitimli sınıflandırılması için önerilen algoritmanın adımları aşağıdaki gibidir.

1. Çok bantlı uydu görüntü verisindeki kontrol grupları belirlenir ve kontrol grupları için çok değişkenli normal dağılım modelleri oluşturulur.
2. Çok bantlı uydu görüntü verisinin çok değişkenli normal dağılımların bir karmasından geldiği varsayımı yapılır ve veriye karma kümeleme analizi uygulanır. Bu sayede veri alt gruplara ayrılır.
3. Elde edilen her bir alt grubun homojenliği test edilir.
4. Eğer alt grup homojen ise bu alt grup bir küme olarak belirlenir ve bu alt grup ya da küme için çok değişkenli normal dağılım modeli oluşturulur ve 5. adıma geçilir. Eğer alt grup homojen değil ise bu alt gruptaki veri için çok değişkenli normal dağılımların bir karmasından elde edildiği varsayımı yapılır ve alt gruptaki veriye tekrar karma kümeleme analizi uygulanır. Bu sayede alt gruptaki veri daha küçük hacimli alt gruplara ayrılır ve 3. adıma

geri dönülür. Bu işleme elde edilen tüm alt grupların homojenliği sağlanıncaya kadar devam edilir.

5. Test gurubu olarak adlandırılan her bir homojen alt grup ya da küme için oluşturulan çok değişkenli normal dağılım modeliyle her bir kontrol grubu ya da kümesi için oluşturulan çok değişkenli normal dağılım modeli arasındaki Bhattacharyya uzaklıkları hesaplanır. Bu uzaklık değerlerine göre bir uzaklık matrisi oluşturulur.
6. Uzaklık matrisinin girdileri, homojen test kümeleri ve kontrol kümeleri arasındaki Bhattaracyya uzaklıklarından oluşmaktadır. Bir test kümesi Bhattaracyya uzaklığı cinsinden kendine en yakın bir kontrol kümesine sınıflandırılır. Diğer bir ifadeyle bir test kümesi aralarındaki Bhattaracyya uzaklığı minimum olan kontrol kümesine sınıflandırılır.

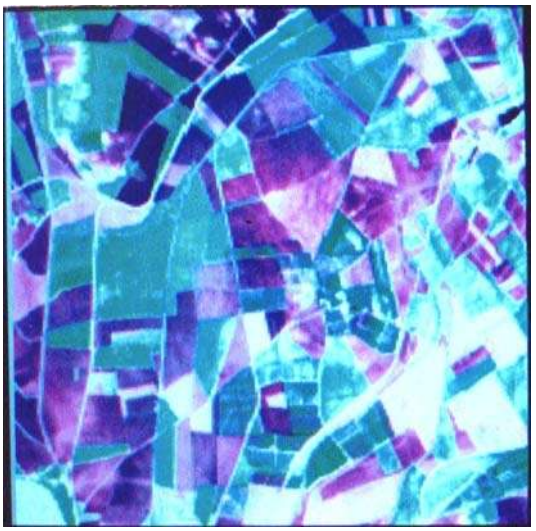
8.3. Çok Bantlı Uydu Görüntü Verisinin Grupların İnceltilmesiyle Karma Kümeleme Analizi İle Alan Bazında Eğitimli Sınıflandırılması İçin Bir Uygulama

Bu tez çalışmasında önerilen ya da oluşturulan alan bazındaki eğitimli sınıflandırma yönteminin çalışma prensibi, bir tarımsal bölgedeki parselleri sınıflandırma uygulaması üzerinde açıklanacaktır.

8.3.1. Verinin Yapısı ve Özelliği

Çok değişkenli normal dağılımların karmasına dayalı iç içe kümeleme analizi kullanılarak alan bazında eğitimli sınıflandırma yöntemi bir tarımsal bölgenin uzaktan algılanmış Landsat Thematic Mapper (Landsat TM) görüntü verisine uygulanacaktır. Çok bantlı uydu görüntü verisi olarak Adana il sınırları içinde bulunan bir bölgeye ait 27.03.1992 tarihli, yörünge ve satır numarası 175-34 olan Landsat TM uydusunun 3., 4. ve 5. bantlarındaki piksel değerleri kullanılmıştır (Erol ve Akdeniz 2005). Landsat TM uydusunun 7 bandı bulunmasına karşın 3., 4. ve 5. bantların kullanılmasının nedeni, bu bantların bitki örtüsünü, suyu ve toprağı iyi

belirlemesidir yada ayırt etmesidir. Çok bantlı uydu görüntü verisinin her bir bandı 198×200 pikselden oluşmaktadır ($n = 39600$). İncelenen alanın uydu görüntüsü Şekil 8.4'de gösterilmiştir.

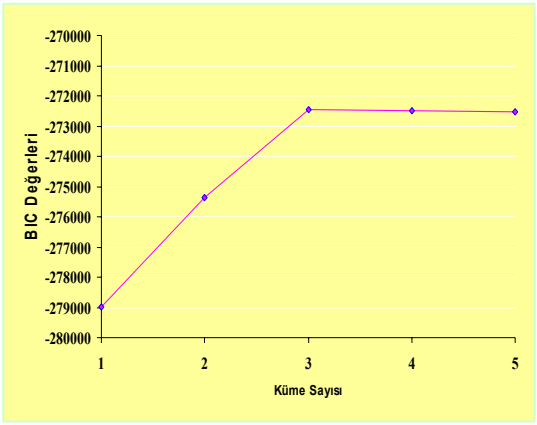


Şekil 8.4. Tarımsal bölgenin Landsat TM görüntüsü (Erol ve Akdeniz 2005).

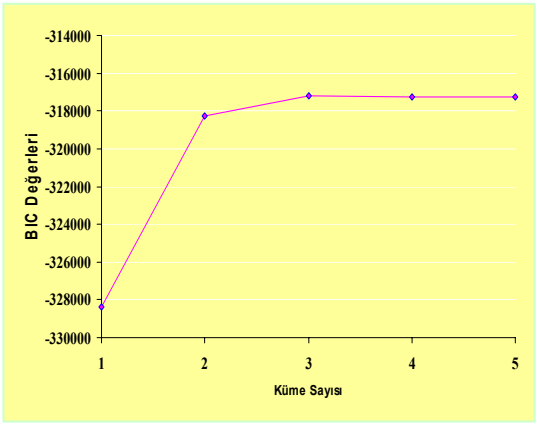
8.3.2. Görüntü Verisindeki Aday Küme Merkezlerinin Sayısı Hakkında Bir Aralık Oluşturulması

Alan bazında eğitimli sınıflandırma için Şekil 8.4'deki görüntü hakkında herhangi bir önbilgi bulunmayan görüntü verisinde aday küme merkezlerinin sayısına ilişkin bir aralık Servi ve Erol (2007) tarafından önerilen yöntemle elde edilmiştir. Bu yöntem de çok değişkenli verinin her bir X1, **X**2 ve **X**3 değişkenlerindeki optimum bileşen sayıları MIXMOD yazılımı kullanılarak farklı bileşen sayılarında elde edilen BIC kriterine göre sırasıyla $k_1 = 3$, $k_2 = 3$ ve $k_3 = 6$ bulunmuştur. Bu değişkenlere ait farklı bileşen sayılarında elde edilen BIC değerleri Şekil 8.5'te gösterilmiştir. Her bir değişken için elde edilen karma model yapıları Tablo 8.3 - Tablo 8.5 ile verilmiştir. Oluşturulan bu karma model yapıları değişkenlerdeki gözlem değerlerine göre çizilen histogramlarla birlikte Şekil 8.6'da verilmiştir. Çok değişkenli verideki küme sayısı g için aralık, $mak\{k_1, k_2, k_3\} \le g \le k_1 k_2 k_3$ bağıntısından $mak\{3,3,6\} \le g \le 3.3.6$ ya da $6 \le g \le 54$

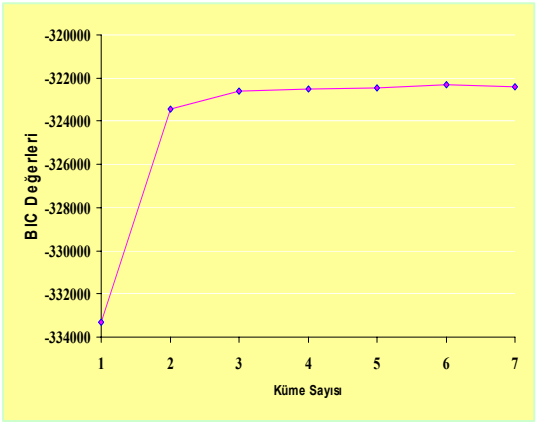
olarak bulunmuştur. Bulunan bu aralık sayesinde çok değişkenli verinin normal dağılımların karmasına dayalı iç içe kümeleme analizinde model seçimi için bilgi kriterleri hesaplanırken bileşen sayısı Fraley ve Raftery (1998)'de olduğu gibi birden başlayarak değil, Servi ve Erol (2007)'de olduğu gibi altıdan başlayarak yapılmıştır ve ilk yerel maksimum BIC değeri elde edilinceye kadar bir arttırılmıştır. Burada amaç $6 \le g \le 54$ aralığında gerçek küme sayısının belirlenmesidir.



(a) X1 için bileşen sayısı



(b) X2 için bileşen sayısı



(c) X3 için bileşen sayısı

Şekil 8.5. Uydu görüntü verisinin X1, X2 ve X3 değişkenlerinin bileşen sayıları ve bileşen sayılarına karşılık gelen BIC değerleri.

Tablo 8.3. X1 değişkeni için MIXMOD programı kullanılarak elde edilen karma normal modelin parametreleri.

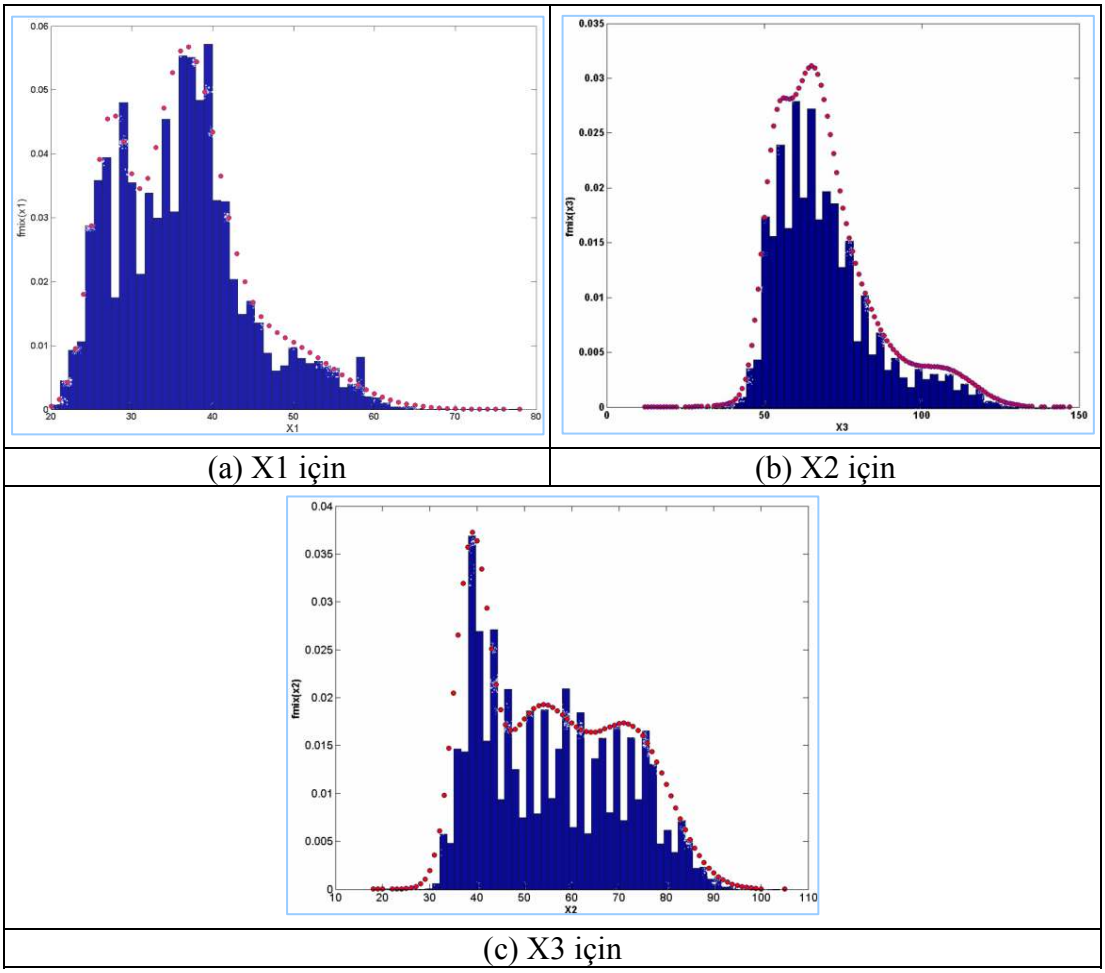
Number of samples : 39600 Criterion Type : BIC Initial start parameters method : RANDOM Type : EM Stopping rule : NBITERATION_EPSILON Number of iterations : 100 Set tolerance (xml criterion) : 0.0001 Number of Clusters : 3 Model Type : Gaussian Ellipsoidal Model : pk_Lk_C Criterion Value : 272463.735530			
Bileşen	Karma Oranı	Ortalaması	Varyansı
1	0.537643	36.457967	16.704443
2	0.211430	46.901973	56.878098
3	0.250927	27.181136	5.792948

Tablo 8.4. X2 değişkeni için MIXMOD programı kullanılarak elde edilen karma normal modelin parametreleri.

Number of samples : 39600.000000 Criterion Type : BIC Initial start parameters method : RANDOM Type : EM Stopping rule : NBITERATION_EPSILON Number of iterations : 100 Set tolerance (xml criterion) : 0.0001 Number of Clusters : 3 Model Type : Gaussian Ellipsoidal Model : pk_Lk_C Criterion Value : 317227.586002			
Bileşen	Karma Oranı	Ortalaması	Varyansı
1	0.290019	38.756577	12.533709
2	0.323932	72.886278	65.402818
3	0.386049	52.906968	70.275182

Tablo 8.5. X3 değişkeni için MIXMOD programı kullanılarak elde edilen karma normal modelin parametreleri.

Number of samples : 39600.000000 Criterion Type : BIC Initial start parameters method : RANDOM Type : EM Stopping rule : NBITERATION_EPSILON Number of iterations : 100 Set tolerance (xml criterion) : 0.000100 Number of Clusters : 6 Model Type : Gaussian Ellipsoidal Model : pk_Lk_C Criterion Value : 322304.340183			
Bileşen	Karma Oranı	Ortalaması	Varyansı
1	0.060427	108.297037	84.839228
2	0.198078	53.272003	18.083439
3	0.199287	63.929819	35.209413
4	0.182998	65.693424	58.652770
5	0.205348	73.848768	108.118363
6	0.153862	78.768566	306.092891



(a) X1 için (b) X2 için

(c) X3 için

Şekil 8.6. Uydu görüntü verisinin X1, X2 ve X3 değişkenlerinin optimum bileşen sayılarına göre karma dağılım modelleri.

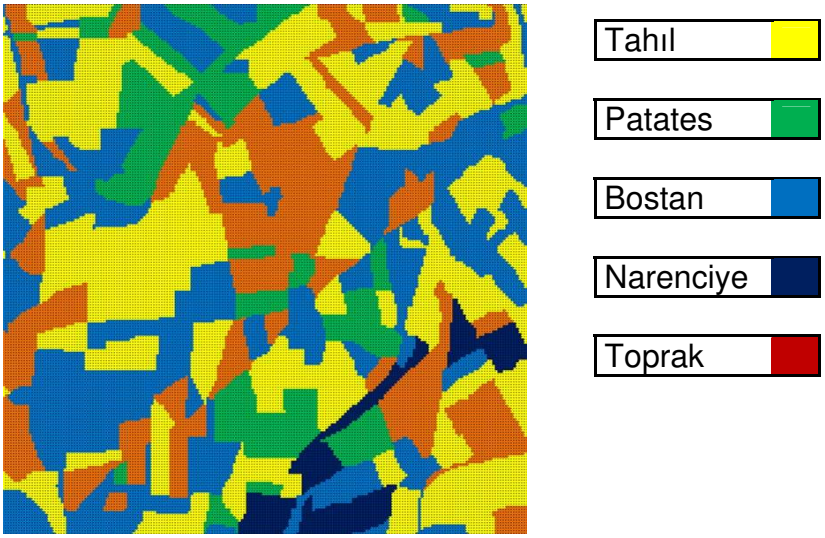
8.3.3. Çok Bantlı Uydu Görüntü Verisinin Grupların İnceltilmesiyle Karma Kümeleme Analizi İle Alan Bazında Eğitimli Sınıflandırılması İçin Kontrol Ve Test Gruplarının Belirlenmesi

8.3.3.1 Kontrol Gruplarının Belirlenmesi

Eğitimli sınıflandırma çerçevesinde: Her bir kontrol sınıfının çok değişkenli normal dağılımların bir karmasından geldiği varsayımı altında, 6 Tahıl kontrol sınıfının her biri çok değişkenli normal dağılım modeline sahip homojen yapıda 11 alt kümesi, 4 Patates kontrol sınıfının her biri çok değişkenli normal dağılım

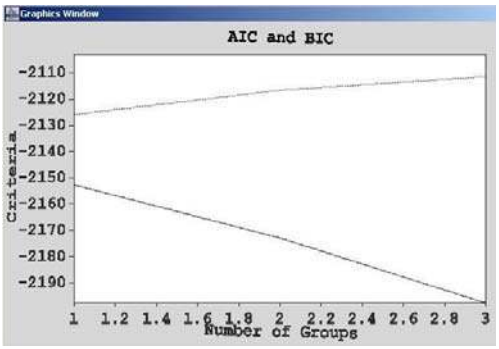
modeline sahip homojen yapıda 7 alt kümesi, 6 Bostan kontrol sınıfının her biri çok değişkenli normal dağılım modeline sahip homojen yapıda 12 bostan alt kümesi, 3 Narenciye kontrol sınıfının her biri çok değişkenli normal dağılım modeline sahip homojen yapıda 6 Narenciye alt kümesi, 5 Toprak kontrol sınıfının her biri çok değişkenli normal dağılım modeline sahip homojen yapıda 9 toprak alt kümesi belirlenmiştir.

Diğer bir ifadeyle 24 kontrol sınıfı için 45 homojen yapıda alt küme elde edilmiştir. Kontrol sınıflarına ait toplam piksel (gözlem) sayısı $n_c = 3682$'dir. Kontrol sınıfları için oluşturulan karma modelin bileşen sayısı 45, $6 \le g \le 54$ aralık içerisine düşmektedir. Şekil 8.4. ile verilen uydu görüntü verisinin yeryüzü gerçeğine göre elde edilen sınıflandırma konusal haritası Şekil 8.7'de gösterilmiştir.

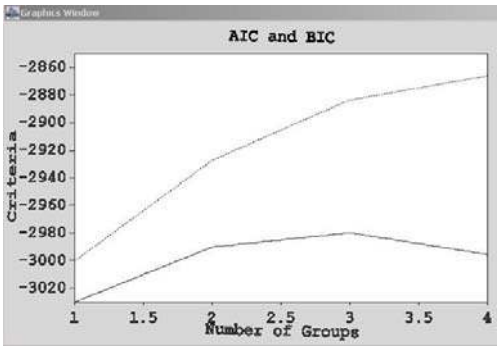


Şekil 8.7. Uydu görüntü verisinin yeryüzü gerçeğine uygun olarak elde edilmiş sınıflandırma konusal haritası.

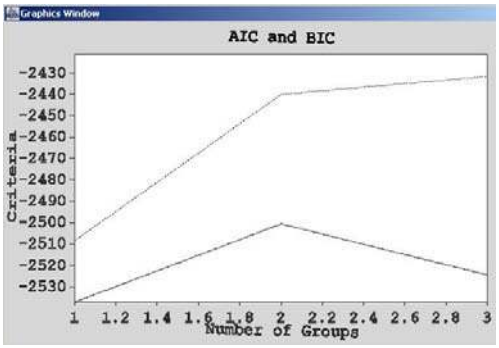
Tahıl kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri Şekil 8.8'de gösterilmiştir.



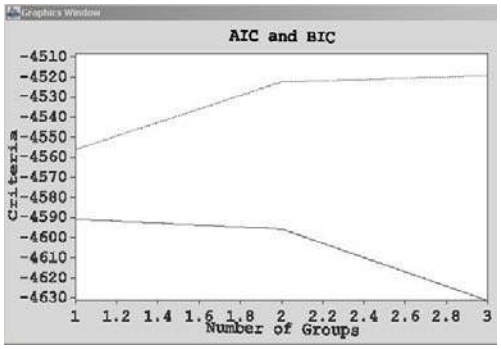
(a) Tahıl Kontrol Alanı 1



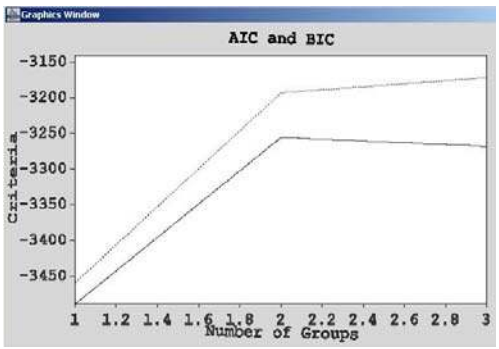
(b) Tahıl Kontrol Alanı 2



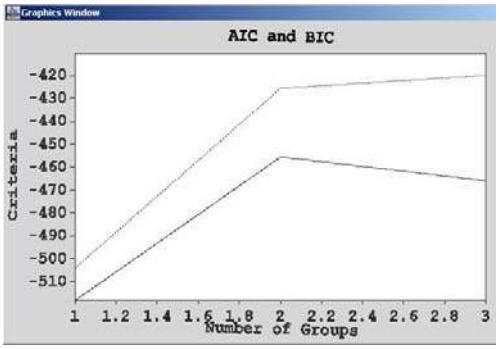
(c) Tahıl Kontrol Alanı 3



(d) Tahıl Kontrol Alanı 4



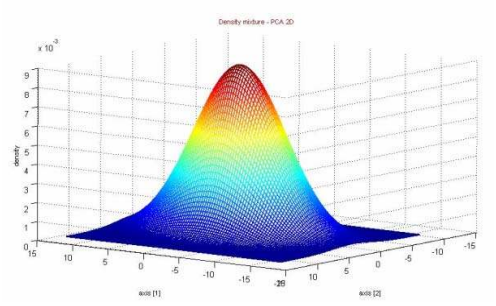
(e) Tahıl Kontrol Alanı 5



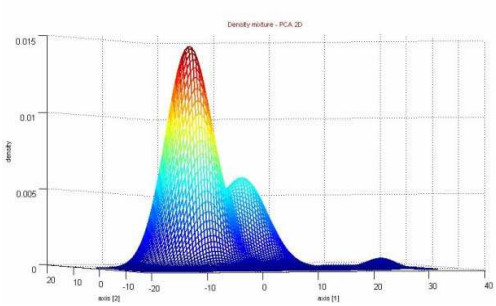
(f) Tahıl Kontrol Alanı 6

Şekil 8.8. Tahıl kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri.

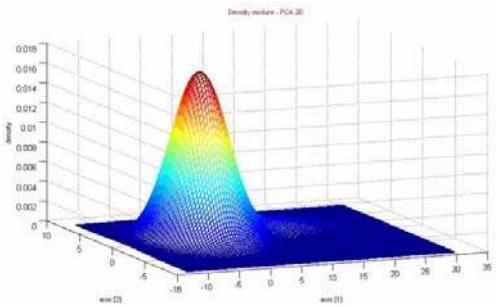
Tahıl kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri Şekil 8.9'da gösterilmiştir.



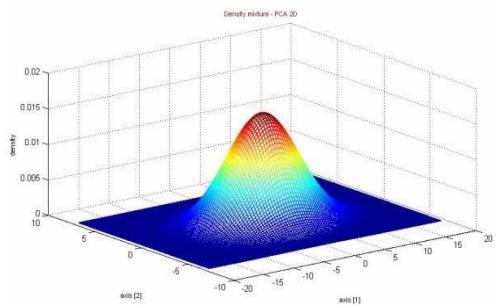
(a) Tahıl Kontrol Alanı 1



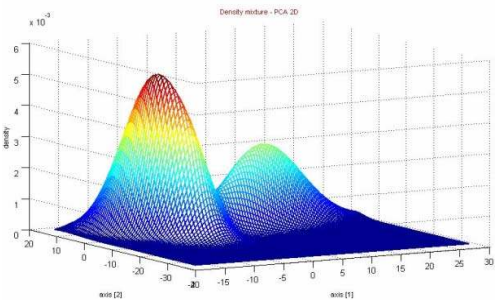
(b) Tahıl Kontrol Alanı 2



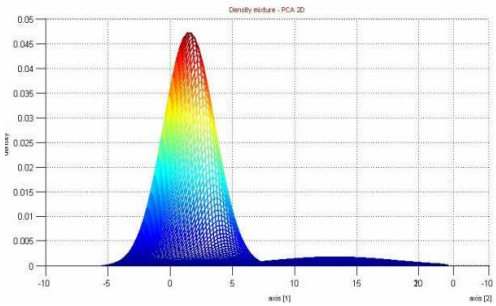
(c) Tahıl Kontrol Alanı 3



(d) Tahıl Kontrol Alanı 4



(e) Tahıl Kontrol Alanı 5



(f) Tahıl Kontrol Alanı 6

Şekil 8.9. Tahıl kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri.

Tahıl kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi Tablo 8.6'da verilmiştir.

Tablo 8.6. Tahıl kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi.

		T1	T2			T3		T4	T5		T6	
		1	2	3	4	5	6	7	8	9	10	11
T1	1	0.00	13.66	2.54	0.40	2.72	0.77	0.29	1.28	0.71	1.92	2.87
T2	2	13.66	0.00	7.25	13.98	2.12	9.27	26.94	47.76	10.08	23.21	102.69
	3	2.54	7.25	0.00	1.48	4.37	2.19	4.87	6.98	0.67	6.53	13.17
	4	0.40	13.98	1.48	0.00	3.62	1.16	0.86	2.01	0.53	4.78	6.98
T3	5	2.72	2.12	4.37	3.62	0.00	1.45	4.60	7.30	3.53	4.13	10.18
	6	0.77	9.27	2.19	1.16	1.45	0.00	1.90	4.78	1.78	2.74	6.03
T4	7	0.29	26.94	4.87	0.86	4.60	1.90	0.00	0.60	1.03	3.22	3.08
T5	8	1.28	47.76	6.98	2.01	7.30	4.78	0.60	0.00	1.74	6.33	2.19
	9	0.71	10.08	0.67	0.53	3.53	1.78	1.03	1.74	0.00	4.81	3.72
T6	10	1.92	23.21	6.53	4.78	4.13	2.74	3.22	6.33	4.81	0.00	6.04
	11	2.87	102.69	13.17	6.98	10.18	6.03	3.08	2.19	3.72	6.04	0.00

Tahıl kontrol sınıflarının alt gruplarına ait bileşen ortalamaları Tablo 8.7'de verilmiştir.

Tablo 8.7. Tahıl kontrol sınıflarının alt gruplarına ait bileşen ortalamaları.

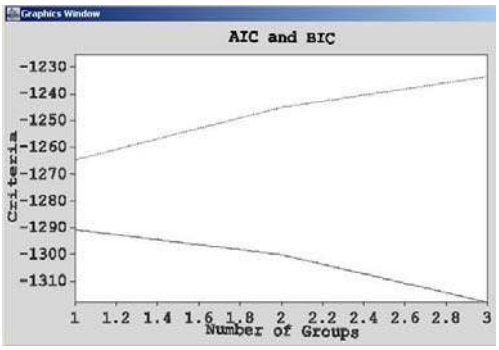
TAHIL KONTROL SINIFI	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN ORTALAMASI		
	T1	1	27.750	72.347	56.813
	T2	1	37.250	61.000	85.000
		2	31.396	74.988	69.041
		3	27.560	76.104	58.457
	T3	1	34.992	64.224	73.603
		2	30.075	68.565	64.816
	T4	1	25.846	75.094	52.917
	T5	1	23.501	80.839	47.691
		2	29.426	77.894	63.877
	T6	1	30.500	56.250	60.500
		2	25.813	63.781	51.875

Tahıl kontrol sınıflarının alt gruplarına ait bileşen kovaryans matrisleri Tablo 8.8'de verilmiştir.

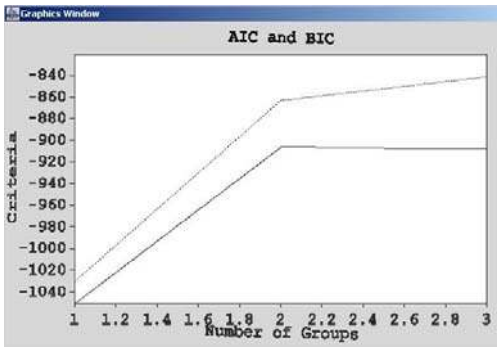
Tablo 8.8. Tahıl kontrol sınıflarının alt gruplarına ait bileşen kovaryans matrisleri

	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN KOVARYANSI		
TAHIL KONTROL SINIFI	T1	1	3.701	-7.212	4.189
			-7.212	27.129	-9.192
			4.189	-9.192	14.263
	T2	1	0.188	-1.250	-0.250
			-1.250	9.500	0.500
			-0.250	0.500	5.000
		2	3.832	-2.503	3.646
			-2.503	4.608	-3.006
			3.646	-3.006	12.271
		3	2.221	-2.258	3.616
			-2.258	6.320	-4.878
			3.616	-4.878	12.672
	T3	1	3.411	-3.820	8.625
			-3.820	9.623	-21.527
			8.625	-21.527	54.133
		2	2.419	-2.598	2.814
			-2.598	7.115	-3.065
			2.814	-3.065	10.319
	T4	1	2.278	-5.031	3.332
			-5.031	18.581	-9.696
			3.332	-9.696	10.463
	T5	1	2.678	-9.628	2.144
			-9.628	48.530	-5.784
			2.144	-5.784	6.638
		2	8.926	3.526	18.702
			3.526	12.893	17.239
			18.702	17.239	63.854
	T6	1	14.750	0.000	0.000
			0.000	18.188	0.000
			0.000	0.000	2.250
		2	0.527	0.000	0.000
			0.000	4.858	0.000
			0.000	0.000	1.922

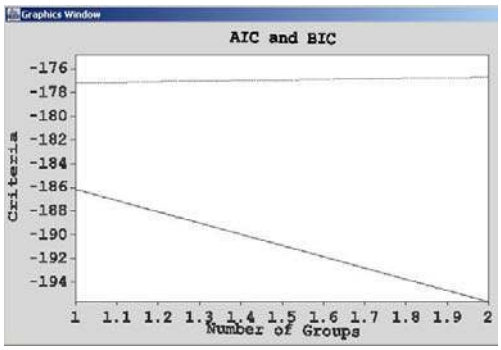
Patates kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri Şekil 8.10'da gösterilmiştir.



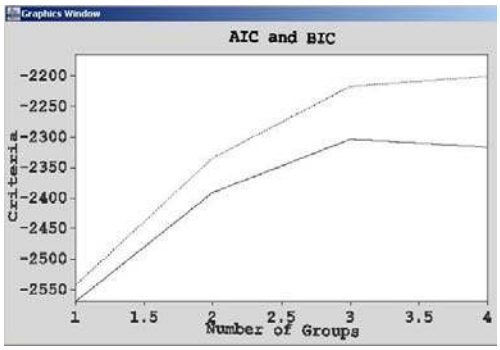
(a) Patates Kontrol Alanı 1



(b) Patates Kontrol Alanı 2



(c) Patates Kontrol Alanı 3



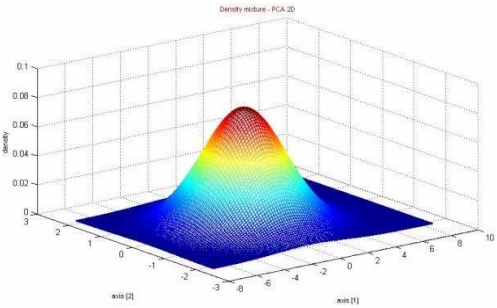
(d) Patates Kontrol Alanı 4

Şekil 8.10. Patates kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri.

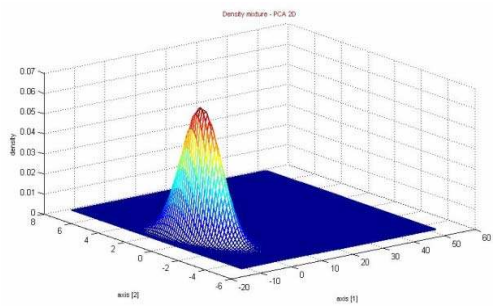
Patates kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri Şekil 8.11'de gösterilmiştir.

Patates kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi Tablo 8.9'da verilmiştir.

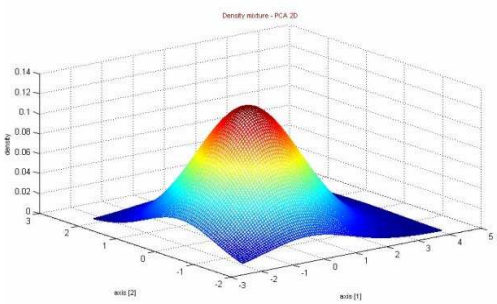
Patates kontrol sınıflarının alt gruplarına ait bileşen ortalamaları Tablo 8.10'da verilmiştir.



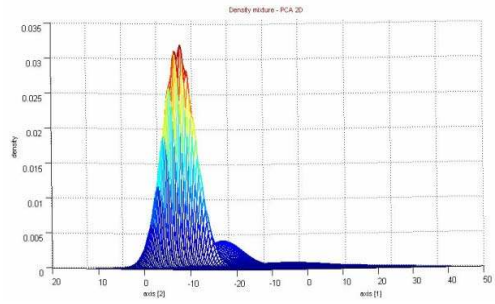
(a) Patates Kontrol Alanı 1



(b) Patates Kontrol Alanı 2



(c) Patates Kontrol Alanı 3



(d) Patates Kontrol Alanı 4

Şekil 8.11. Patates kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri.

Tablo 8.9. Patates kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi.

		P1	P2		P3	P4		
		1	2	3	4	5	6	7
P1	1	0.00	0.57	2.39	0.51	8.37	2.14	0.14
P2	2	0.57	0.00	2.10	0.17	7.18	1.72	0.24
	3	2.39	2.10	0.00	2.56	2.27	1.53	1.93
P3	4	0.51	0.17	2.56	0.00	8.05	2.04	0.24
P4	5	8.37	7.18	2.27	8.05	0.00	2.26	7.34
	6	2.14	1.72	1.53	2.04	2.26	0.00	1.59
	7	0.14	0.24	1.93	0.24	7.34	1.59	0.00

Tablo 8.10. Patates kontrol sınıflarının alt gruplarına ait bileşen ortalamaları.

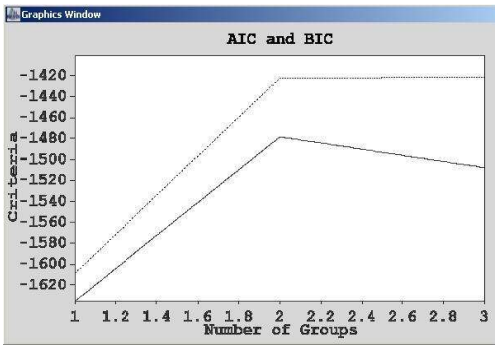
	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN ORTALAMASI		
PATATES KONTROL SINIFI	P1	1	30.182	31.963	50.3401
	P2	1	31.817	33.060	50.137
		2	35.339	43.649	66.420
	P3	1	31.400	33.000	50.500
	P4	1	31.084	63.197	62.165
		2	31.206	42.484	55.899
		3	30.812	32.847	51.153

Patates kontrol sınıflarının alt gruplarına ait bileşen kovaryans matrisleri Tablo 8.11'de verilmiştir.

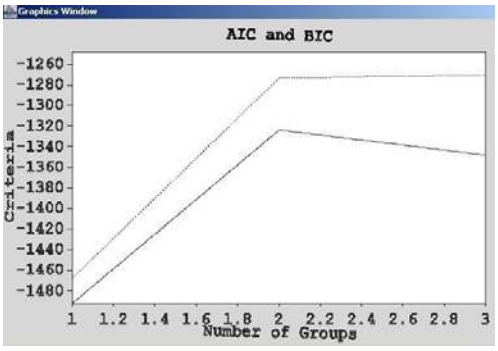
Tablo 8.11. Patates kontrol sınıflarının alt gruplarına ait bileşen kovaryans matrisleri.

	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN KOVARYANSI		
PATATES KONTROL SINIFI	P1	1	1.005	0.318	1.199
			0.318	0.791	1.094
			1.199	1.094	5.128
	P2	1	1.339	0.552	1.087
			0.552	1.095	1.596
			1.087	1.596	5.640
		2	7.781	25.198	23.255
			25.198	111.972	107.687
			23.255	107.687	111.886
	P3	1	0.840	0.000	0.000
			0.000	0.500	0.000
			0.000	0.000	2.050
	P4	1	15.486	-14.398	1.850
			-14.398	50.345	-0.101
			1.850	-0.101	46.345
		2	2.717	-4.004	0.861
			-4.004	28.512	4.438
			0.861	4.438	7.604
		3	1.542	0.542	1.941
			0.542	1.115	1.511
			1.941	1.511	6.662

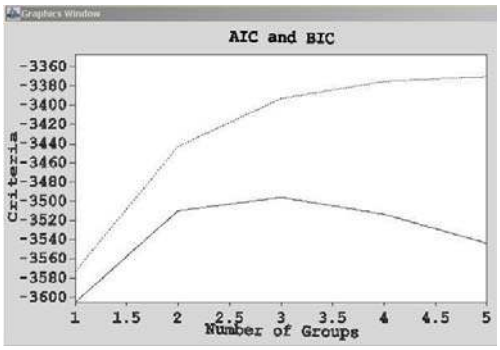
Bostan kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri Şekil 8.12'de gösterilmiştir.



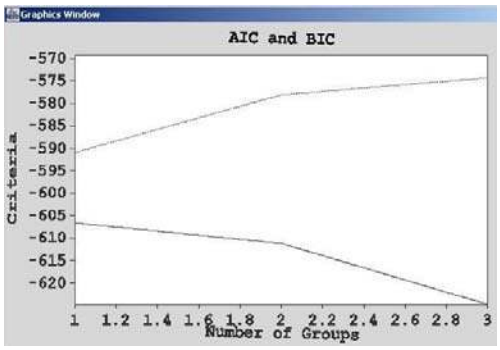
(a) Bostan Kontrol Alanı 1



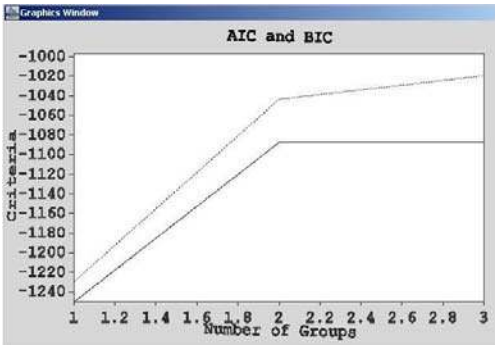
(b) Bostan Kontrol Alanı 2



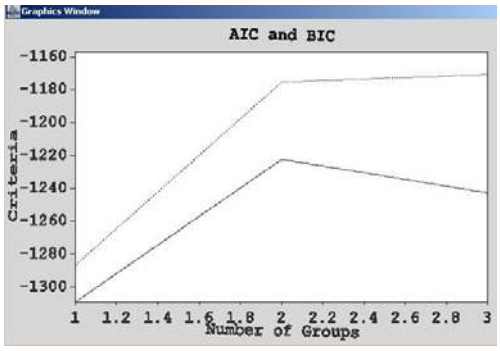
(c) Bostan Kontrol Alanı 3



(d) Bostan Kontrol Alanı 4



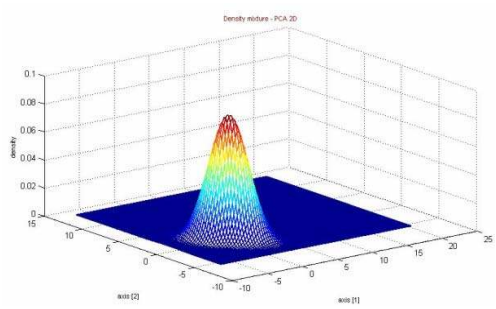
(e) Bostan Kontrol Alanı 5



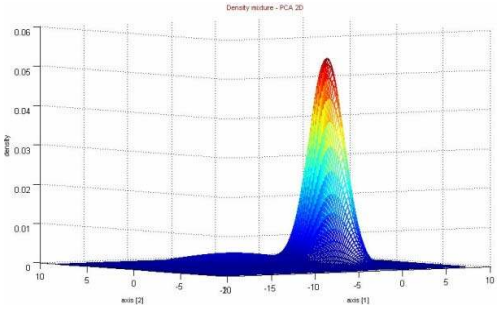
(f) Bostan Kontrol Alanı 6

Şekil 8.12. Bostan kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri.

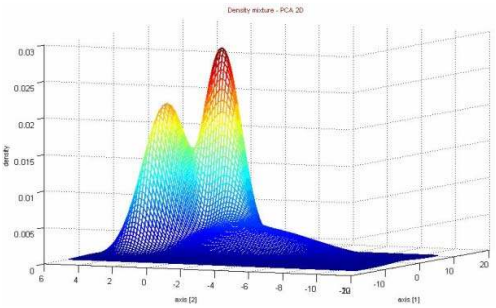
Bostan kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri Şekil 8.13'te gösterilmiştir.



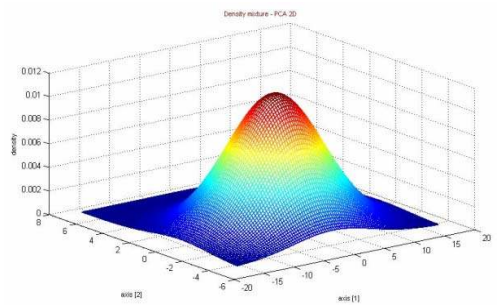
(a) Bostan Kontrol Alanı 1



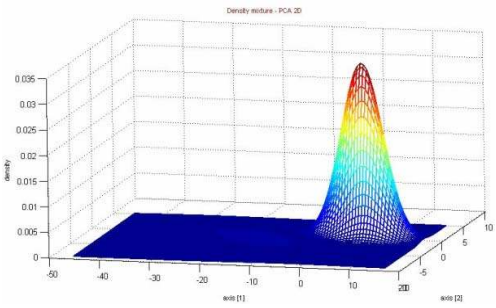
(b) Bostan Kontrol Alanı 2



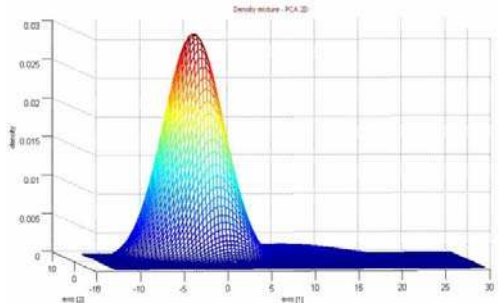
(c) Bostan Kontrol Alanı 3



(d) Bostan Kontrol Alanı 4



(e) Bostan Kontrol Alanı 5



(f) Bostan Kontrol Alanı 6

Şekil 8.13. Bostan kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri.

Bostan kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi Tablo 8.12'de verilmiştir.

Tablo 8.12. Bostan kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi.

		B1		B2		B3			B4	B5		B6	
		1	2	3	4	5	6	7	8	9	10	11	12
B1	1	0	4.36	4.43	1.68	3.94	0.4	4.26	4.43	17.76	1.43	0.9	2.08
	2	4.36	0	5.58	7.36	1.78	1.62	0.44	5.15	18.93	1.93	4.13	3.54
B2	3	4.43	5.58	0	4.76	5.94	2.22	2.15	2.54	24.73	3.18	7.09	0.86
	4	1.68	7.36	4.76	0	7.69	1.51	5.96	5.86	25.55	2.23	1.26	3.4
B3	5	3.94	1.78	5.94	7.69	0	1.83	1	2.68	11.47	1.44	3.9	3.66
	6	0.4	1.62	2.22	1.51	1.83	0	1.45	2.64	13.8	1.04	0.87	0.89
	7	4.26	0.44	2.15	5.96	1	1.45	0	3.57	12.82	1.78	4.93	1.64
B4	8	4.43	5.15	2.54	5.86	2.68	2.64	3.57	0	5.39	1.59	5.17	2.16
B5	9	17.76	18.93	24.73	25.55	11.47	13.8	12.82	5.39	0	2.38	8.59	21.02
	10	1.43	1.93	3.18	2.23	1.44	1.04	1.78	1.59	2.38	0	0.96	2.05
B6	11	0.9	4.13	7.09	1.26	3.9	0.87	4.93	5.17	8.59	0.96	0	3.15
	12	2.08	3.54	0.86	3.4	3.66	0.89	1.64	2.16	21.02	2.05	3.15	0

Bostan kontrol sınıflarının alt gruplarına ait bileşen ortalamaları Tablo 8.13'te verilmiştir.

Tablo 8.13. Bostan kontrol sınıflarının alt gruplarına ait bileşen ortalamaları.

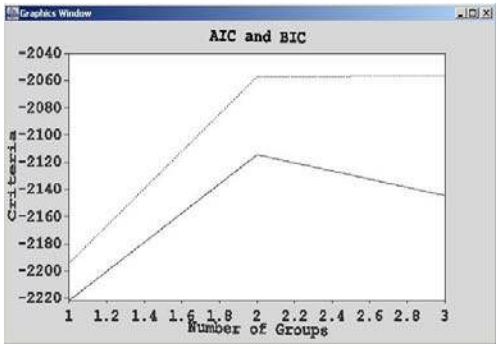
	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN ORTALAMASI		
BOSTAN KONTROL SINIFI	B1	1	39.353	46.297	69.114
		2	41.507	41.304	68.222
	B2	1	36.717	37.582	70.598
		2	35.443	45.347	64.475
	B3	1	43.421	44.108	75.195
		2	39.402	45.723	70.006
		3	41.216	40.885	70.496
	B4	1	43.571	45.857	87.952
	B5	1	54.319	54.664	96.629
		2	45.514	50.100	79.140
	B6	1	39.003	52.097	76.011
		2	37.397	41.083	71.322

Bostan kontrol sınıflarının alt gruplarına ait bileşen kovaryans matrisleri Tablo 8.14'te verilmiştir.

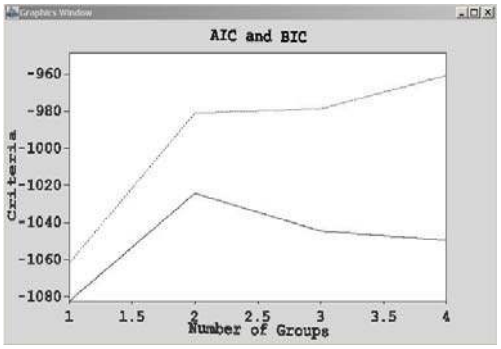
Tablo 8.14. Bostan kontrol sınıflarının alt gruplarına ait bileşen kovaryans matrisleri.

	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN KOVARYANSI		
BOSTAN KONTROL SINIFI	B1	1	6.509	11.673	10.051
			11.673	23.731	20.619
			10.051	20.619	20.728
		2	0.782	0.243	0.720
			0.243	0.900	0.625
			0.720	0.625	3.663
	B2	1	1.862	1.144	2.272
			1.144	1.525	1.775
			2.272	1.775	5.745
		2	2.813	5.169	2.604
			5.169	19.468	10.754
			2.604	10.754	11.282
	B3	1	1.582	0.092	1.497
			0.092	1.298	0.698
			1.497	0.698	5.653
		2	2.245	1.618	1.116
			1.618	11.408	3.907
			1.116	3.907	10.829
		3	2.785	1.897	2.529
			1.897	2.113	2.285
			2.529	2.285	5.760
	B4	1	6.197	5.986	12.027
			5.986	8.170	11.350
			12.027	11.350	34.760
	B5	1	2.342	1.567	4.288
			1.567	1.765	3.707
			4.288	3.707	15.946
		2	17.251	14.012	24.436
			14.012	43.625	67.554
			24.436	67.554	151.110
	B6	1	19.669	11.240	12.131
			11.240	20.146	23.404
			12.131	23.404	33.532
		2	1.183	0.723	1.044
			0.723	4.988	4.225
			1.044	4.225	8.002

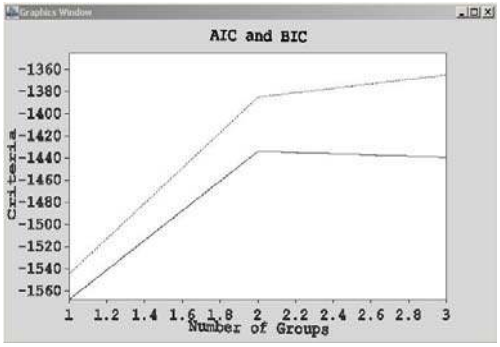
Narenciye kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri Şekil 8.14'de gösterilmiştir.



(a) Narenciye Kontrol Alanı 1



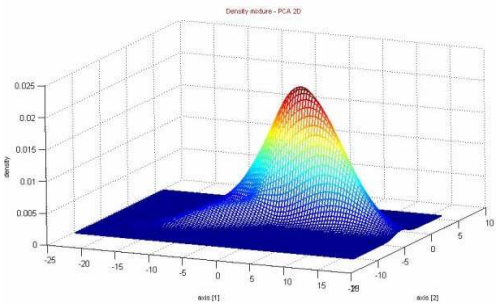
(b) Narenciye Kontrol Alanı 2



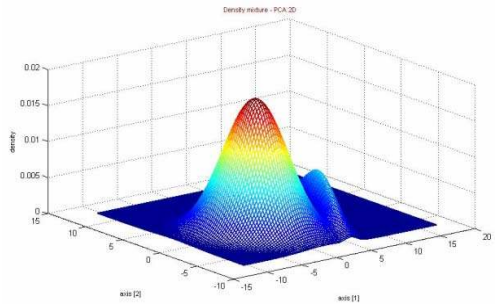
(c) Narenciye Kontrol Alanı 3

Şekil 8.14. Narenciye kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri.

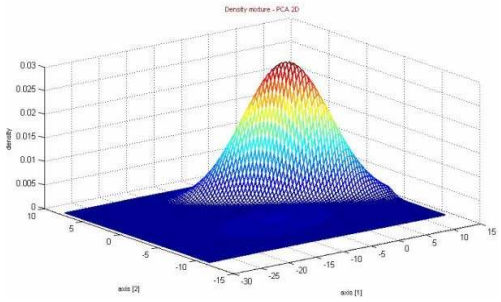
Narenciye kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri Şekil 8.15'te gösterilmiştir.



(a) Narenciye Kontrol Alanı 1



(b) Narenciye Kontrol Alanı 2



(c) Narenciye Kontrol Alanı 3

Şekil 8.15. Narenciye kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri.

Narenciye kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi Tablo 8.15'te verilmiştir.

Tablo 8.15. Narenciye kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi.

		N1		N2		N3	
		1	2	3	4	5	6
N1	1	0.00	1.12	28.50	29.23	1.12	0.74
	2	1.12	0.00	34.17	66.76	3.65	0.61
N2	3	28.50	34.17	0.00	3.34	19.18	29.05
	4	29.23	66.76	3.34	0.00	41.64	71.80
N3	5	1.12	3.65	19.18	41.64	0.00	1.95
	6	0.74	0.61	29.05	71.80	1.95	0.00

Narenciye kontrol sınıflarının alt gruplarına ait bileşen ortalamaları Tablo 8.16'da verilmiştir.

Tablo 8.16. Narenciye kontrol sınıflarının alt gruplarına ait bileşen ortalamaları.

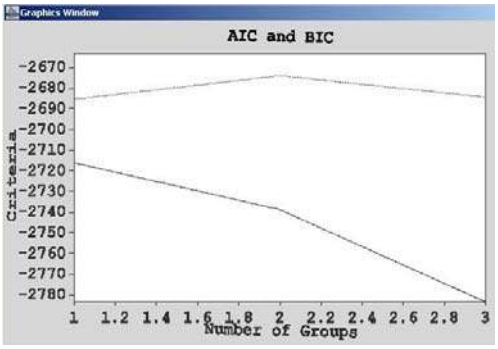
	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN ORTALAMASI		
NARENCİYE KONTROL SINIFI	N1	1	52.955	58.737	104.782
		2	55.896	57.636	112.909
	N2	1	27.576	65.019	52.977
		2	25.419	63.601	42.665
	N3	1	46.521	61.983	100.817
		2	55.046	58.462	108.294

Narenciye kontrol sınıflarının alt gruplarına ait bileşen kovaryansları Tablo 8.17'de verilmiştir.

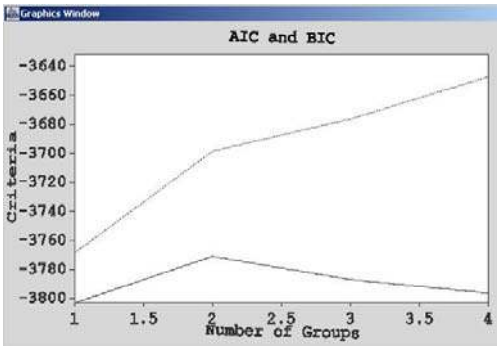
Tablo 8.17. Narenciye kontrol sınıflarının alt gruplarına ait bileşen kovaryansları.

	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN KOVARYANSI		
NARENCİYE KONTROL SINIFI	N1	1	11.358	-3.644	13.363
			-3.644	2.454	-5.294
			13.363	-5.294	25.844
		2	5.912	4.823	11.206
			4.823	5.152	10.849
			11.206	10.849	30.754
	N2	1	22.588	-7.167	5.234
			-7.167	6.762	-5.079
			5.234	-5.079	4.606
		2	1.216	0.003	1.593
			0.003	8.071	2.816
			1.593	2.816	7.779
	N3	1	31.722	2.693	35.239
			2.693	5.485	8.751
			35.239	8.751	47.894
		2	4.680	4.452	6.998
			4.452	5.716	8.595
			6.998	8.595	26.712

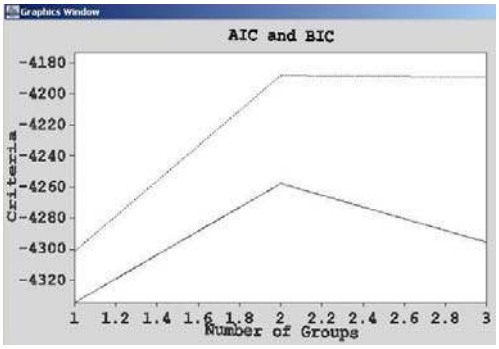
Toprak kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri Şekil 8.16'da gösterilmiştir.



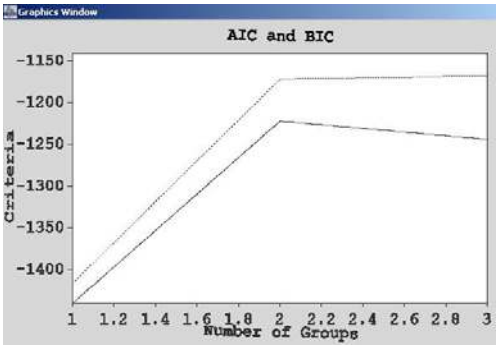
(a) Toprak Kontrol Alanı 1



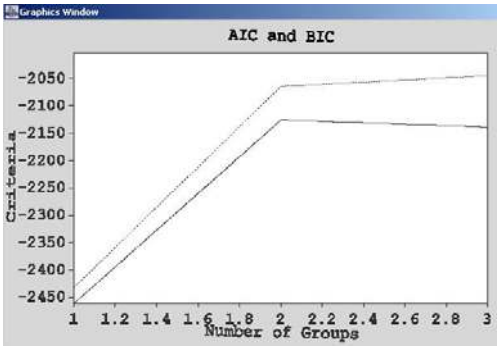
(b) Toprak Kontrol Alanı 2



(c) Toprak Kontrol Alanı 3



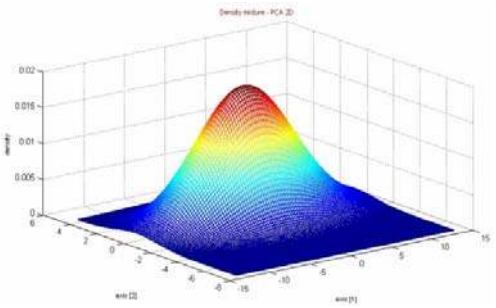
(d) Toprak Kontrol Alanı 4



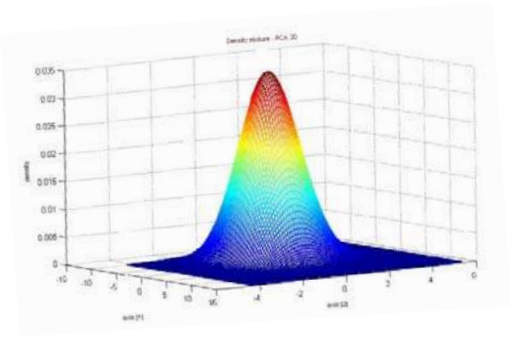
(e) Toprak Kontrol Alanı 5

Şekil 8.16. Toprak kontrol alanları için bileşen küme sayıları ve karşılık gelen AIC ve BIC değerleri.

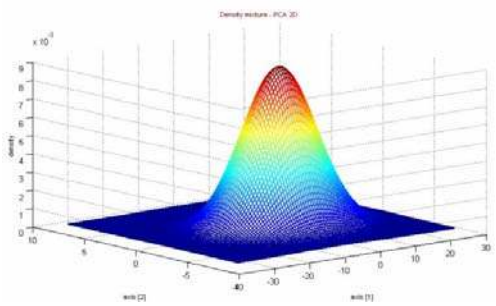
Toprak kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri Şekil 8.17'de gösterilmiştir.



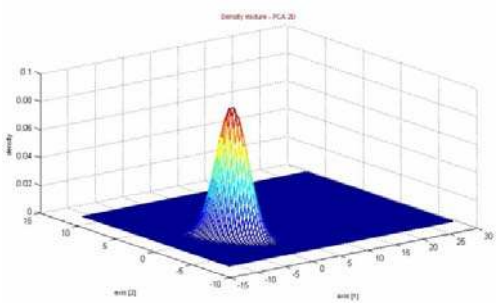
(a) Toprak Kontrol Alanı 1



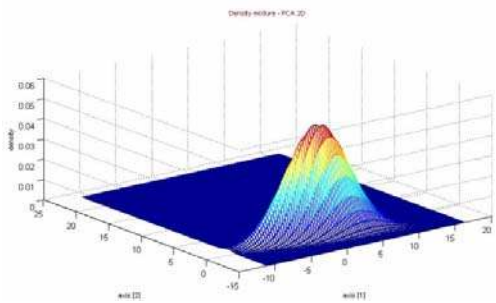
(b) Toprak Kontrol Alanı 2



(c) Toprak Kontrol Alanı 3



(d) Toprak Kontrol Alanı 4



(e) Toprak Kontrol Alanı 5

Şekil. 8.17. Toprak kontrol alanlarının BIC kriterine göre elde edilen bileşen küme sayıları için 3 boyutlu karma yoğunluk fonksiyon grafikleri.

Toprak kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi Tablo 8.18'de verilmiştir.

Tablo 8.18. Toprak kontrol gruplarının birbirlerine olan Bhattacharyya uzaklık matrisi.

		TO1	TO2		TO3		TO4		TO5	
		1	2	3	4	5	6	7	8	9
TO1	1	0.00	2.74	2.41	3.62	4.42	2.14	2.04	2.94	2.29
TO2	2	2.74	0.00	0.40	9.74	7.43	3.17	7.35	0.63	0.80
	3	2.41	0.40	0.00	9.90	9.91	3.96	10.15	1.23	0.28
TO3	4	3.62	9.74	9.90	0.00	0.58	5.25	5.75	10.04	9.91
	5	4.42	7.43	9.91	0.58	0.00	2.99	5.85	6.16	9.78
TO4	6	2.14	3.17	3.96	5.25	2.99	0.00	1.49	2.62	5.08
	7	2.04	7.35	10.15	5.75	5.85	1.49	0.00	8.48	10.57
TO5	8	2.94	0.63	1.23	10.04	6.16	2.62	8.48	0.00	1.25
	9	2.29	0.80	0.28	9.91	9.78	5.08	10.57	1.25	0.00

Toprak kontrol sınıflarının alt gruplarına ait bileşen ortalamaları Tablo 8.19'da verilmiştir.

Tablo 8.19. Toprak kontrol sınıflarının alt gruplarına ait bileşen ortalamaları.

	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN ORTALAMASI		
TOPRAK KONTROL SINIFI	TO1	1	41.871	42.330	81.674
	TO2	1	36.705	38.012	62.341
		2	36.782	37.016	64.280
	TO3	1	55.987	57.309	113.025
		2	54.756	58.428	110.742
	TO4	1	42.625	51.622	80.536
		2	44.074	44.301	77.088
	TO5	1	36.198	43.972	64.472
		2	35.746	36.446	64.933

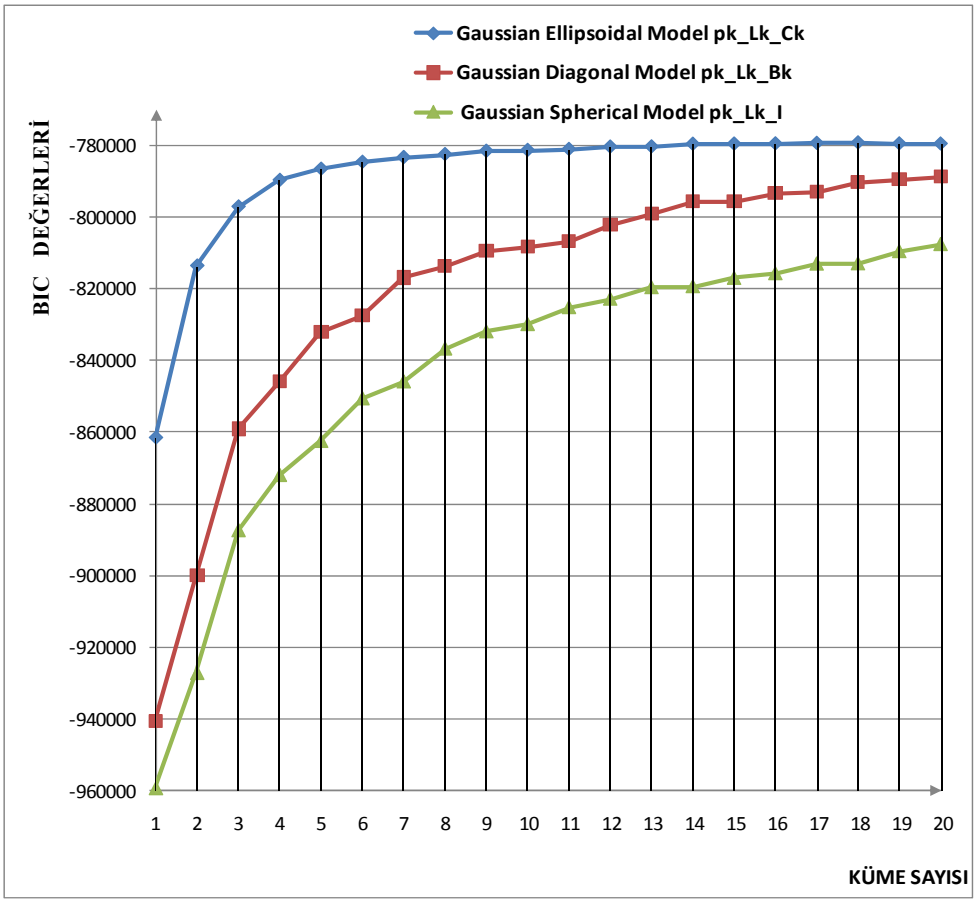
Toprak kontrol sınıflarının alt gruplarına ait bileşen kovaryansları Tablo 8.20'de verilmiştir.

Tablo 8.20. Toprak kontrol sınıflarının alt gruplarına ait bileşen kovaryansları.

	ALT SINIFI	ALT SINIFLARIN BİLEŞEN SAYISI	ALT SINIFLARIN KOVARYANSI		
TOPRAK KONTROL SINIFI	TO1	1	4.470	4.369	7.944
			4.369	5.132	9.139
			7.944	9.139	24.041
	TO2	1	1.162	0.803	2.576
			0.803	4.680	3.628
			2.576	3.628	17.299
		2	0.990	0.536	-0.082
			0.536	0.870	0.870
			-0.082	0.870	13.388
	TO3	1	9.933	10.200	20.874
			10.200	11.763	24.379
			20.874	24.379	61.494
		2	26.188	16.306	33.005
			16.306	14.091	26.984
			33.005	26.984	83.235
	TO4	1	3.599	-1.009	3.314
			-1.009	47.817	38.163
			3.314	38.163	50.064
		2	0.833	0.582	0.660
			0.582	1.038	0.602
			0.660	0.602	2.759
	TO5	1	1.372	0.316	4.269
			0.316	55.180	9.323
			4.269	9.323	22.855
		2	1.128	0.686	2.516
			0.686	1.127	2.917
			2.516	2.917	13.768

8.3.3.2 Test Gruplarının Belirlenmesi

Uzaktan algılanmış çok bantlı uydu görüntü verisine çok değişkenli normal karma kümeleme analizi uygulandığında BIC değerlerine göre 18 grup belirlenmiştir. Karma kümeleme analizinin birinci aşamasında her bir grup sayısına göre elde edilen BIC değerlerine ait grafik Şekil 8.18'de verilmiştir. Grafikten de görüleceği gibi ilk maksimum BIC değerine küme sayısı 18'de karşılaşılmaktadır.



Şekil 8.18. Uzaktan algılanmış çok bantlı uydu görüntü verisindeki farklı küme sayılarına göre elde edilen BIC değerleri.

Başlangıç aşamasında elde edilen 18 kümeye ait karma dağılım modeli Tablo 8.21'de verilmiştir. Bu 18 küme test grubu olarak ele alınır ve Bölüm 8.2.2'de verilen ayrıştırma fonksiyonu kullanılarak Bölüm 8.3.3.1'de elde edilen kontrol sınıflarının her bir grubuna olan uzaklıkları hesaplanır. Hesaplanan uzaklıklar matrisi Tablo 8.22'de verilmiştir. Bölüm 8.2.3'te belirtilen karar kuralına uygun olarak her bir test grubunun kontrol sınıfına olan ataması yapılır. Başlangıç aşamasında elde edilen ve test grupları olarak adlandırılan 18 kümenin kontrol sınıflarına atama sonucu Tablo 8.23'te gösterilmiştir. Başlangıç aşamasının sonunda elde edilen sınıflandırma doğruluk tablosu Tablo 8.24'de verilmiştir. Başlangıç aşamasında eğitimli sınıflandırma sonucu elde edilen renkli sınıflandırma konusal haritası Şekil 8.19'da verilmiştir.

Tablo 8.21. Veri için MIXMOD yazılımı kullanılarak başlangıç aşamasında elde edilen kümeler, karma oranları, ortalama vektörleri ve varyans kovaryans matrisleri.

Model Type : Gaussian Ellipsoidal Model: pk_Lk_Ck			Number of Clusters: 18		
Number of Free Parameters: 179			Log-Likelihood: -388542.366173		
Complete Log-Likelihood: -406516.523			Entropy: 35802.890523		
Küme	**Karma Oranı**	**Ortalama Vektörü**	**Varyans Kovaryans Matrisi**		
1	0.0461	37.9261	9.7180	6.2326	18.4511
		42.8184	6.2326	10.3622	16.0148
		67.6189	18.4511	16.0148	58.7453
2	0.0532	54.1866	15.8630	15.2940	31.5593
		56.0187	15.2940	16.5834	34.3135
		107.5190	31.5593	34.3135	96.7725
3	0.0649	41.3218	10.9182	10.1553	18.4868
		42.0827	10.1554	10.4425	18.3817
		78.3168	18.4868	18.3817	46.5743
4	0.0579	32.0740	9.1678	-6.8654	12.6453
		71.6038	-6.8654	20.9163	-2.1622
		69.2251	12.6453	-2.1622	33.1441
5	0.0801	25.3510	3.2934	-9.5084	4.3391
		75.7207	-9.5084	61.7535	-12.4810
		52.2222	4.3391	-12.4810	11.0515
6	0.0496	49.0458	50.3423	36.8058	83.5374
		57.2575	36.8058	38.6786	66.5806
		94.7103	83.5374	66.5806	227.6409
7	0.0404	42.1992	10.4249	9.8982	18.0966
		54.5544	9.8982	40.3401	43.5324
		79.0155	18.0966	43.5324	122.6577
8	0.0718	30.3840	7.7407	-14.5813	8.2780
		65.4261	-14.5813	82.9573	-12.6480
		60.7459	8.2780	-12.6480	36.3077
9	0.0653	37.0121	3.1349	2.6424	6.8627
		38.9230	2.6424	4.7657	10.8541
		65.1657	6.8627	10.8541	42.8888

Tablo 8.21 (Devamı)**.** Veri için MIXMOD yazılımı kullanılarak başlangıç aşamasında elde edilen kümeler, karma oranları, ortalama vektörleri ve varyans kovaryans matrisleri.

Model Type : Gaussian Ellipsoidal Model: pk_Lk_Ck			Number of Clusters: 18		
Number of Free Parameters: 179			Log-Likelihood: -388542.366173		
Complete Log-Likelihood: -406516.523			Entropy: 35802.890523		
Küme	**Karma Oranı**	**Ortalama Vektörü**	**Varyans Kovaryans Matrisi**		
10	0.0353	40.0721	40.0020	50.7902	77.7874
		64.5107	50.7902	182.2201	215.2684
		78.5193	77.7874	215.2684	348.1850
11	0.0516	40.3957	10.0264	9.9013	21.5270
		40.1654	9.9013	10.7470	22.4757
		67.0840	21.5270	22.4757	53.0860
12	0.0439	28.4999	4.9049	-9.3436	9.5504
		53.5608	-9.3436	41.9080	-18.1916
		50.9132	9.5504	-18.1916	30.6329
13	0.0476	35.0883	5.8309	-1.1022	13.7967
		50.7432	-1.1022	32.6875	0.5670
		64.5833	13.7967	0.5670	56.8705
14	0.0567	35.6741	18.2913	-17.6142	10.2657
		66.3361	-17.6141	75.2327	-8.4777
		70.0672	10.2657	-8.4777	83.5611
15	0.0436	43.6111	24.8654	24.7187	36.8821
		46.3808	24.7187	28.0274	41.9623
		81.6569	36.8821	41.9623	119.1329
16	0.0852	35.9409	3.7024	2.8085	6.9711
		36.5426	2.8085	3.1413	7.2141
		63.7452	6.9711	7.2141	28.4940
17	0.0841	27.2671	3.7479	-10.9649	5.7594
		75.4851	-10.9649	55.8484	-12.9064
		57.8695	5.7594	-12.9064	21.2215
18	0.0228	30.9058	1.6559	0.6574	1.3396
		32.4795	0.6574	1.2445	1.3616
		50.7165	1.3396	1.3616	5.3862

Tablo 8.22. Başlangıç aşaması sonunda elde edilen 18 test grubunun 45 kontrol grubuna olan Bhattacharyya uzaklık değerlerinden oluşan uzaklık matrisi.

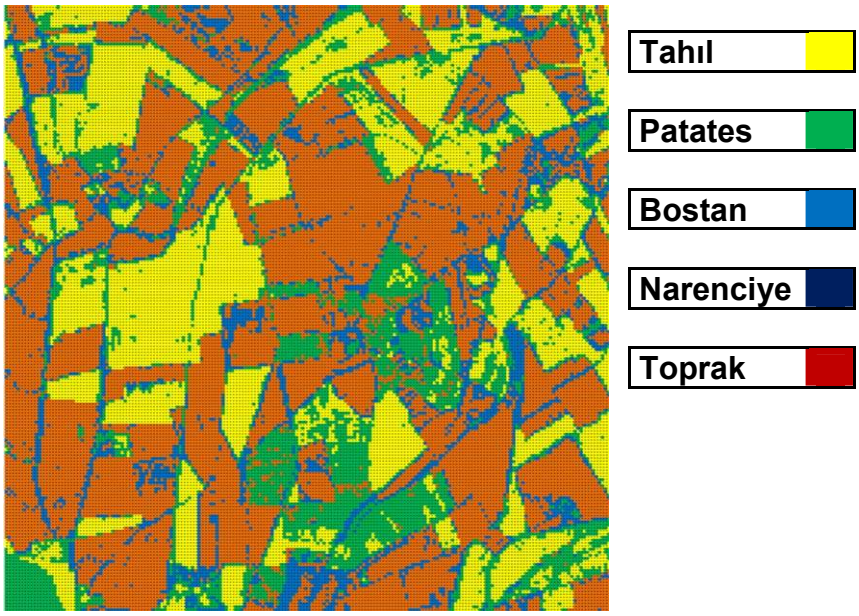
			Başlangıç Aşama Sonunda Elde Edilen 18 Test Grubu																	
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Kontrol Sınıfları	Tahıl	1	**0.1**	1.6	0.3	11.6	14.7	9.9	1.1	7.4	12.3	5.4	0.7	2.8	10.5	9.6	3.2	3.2	11.4	1.2
		2	10.5	2.4	20.2	12.5	13.4	7.1	3.4	5.7	17.8	3.3	4.5	14.4	9.9	9.1	3.9	2.8	14.6	4.3
		3	1.5	1.2	3.7	35.7	32.4	22.6	1.4	14.4	43.4	8.8	1.4	9.2	28.4	35.6	5.8	3.0	51.7	0.3
		4	0.4	2.0	0.8	41.9	40.1	27.3	1.5	20.0	47.9	11.3	1.5	5.8	35.4	33.4	7.0	4.9	46.6	1.1
		5	2.7	1.1	4.7	15.3	12.4	7.0	1.1	6.1	30.1	3.0	1.4	7.2	11.3	13.2	2.9	1.7	29.6	1.3
		6	0.6	1.3	2.0	23.9	24.1	15.9	1.2	10.9	27.1	6.2	0.8	4.9	19.7	19.3	3.7	2.8	27.1	0.7
		7	0.6	2.5	0.2	20.0	22.7	15.7	1.9	12.8	21.6	8.3	1.6	3.5	18.0	16.5	5.6	4.9	19.3	2.2
		8	1.4	3.4	0.4	14.1	19.4	13.0	2.5	10.4	12.4	8.0	2.5	2.5	13.8	13.2	5.8	5.8	13.6	3.9
		9	0.6	1.2	1.0	40.5	38.9	27.6	1.1	17.9	44.1	10.7	1.1	3.6	32.1	35.3	6.3	3.4	45.7	0.6
		10	1.7	1.7	2.4	7.1	11.1	5.9	1.7	3.5	6.7	2.4	0.7	1.7	5.6	5.5	1.2	1.6	7.1	3.2
		11	2.7	3.5	1.4	57.9	58.7	30.1	3.7	23.4	51.0	8.8	2.2	1.9	47.2	45.1	6.3	4.9	55.4	4.9
	Patates	12	11.7	8.2	10.3	3.0	9.4	4.1	12.8	2.0	0.9	6.2	5.5	4.5	4.2	5.7	4.6	3.9	3.4	29.6
		13	10.2	7.2	10.5	2.6	8.6	3.9	11.5	1.8	0.7	5.7	4.8	4.2	3.6	3.3	3.8	3.4	2.8	25.9
		14	4.8	1.5	7.3	2.2	7.0	2.3	1.9	0.5	1.3	1.4	1.8	2.5	2.1	1.3	0.7	1.3	1.6	4.1
		15	11.6	7.9	11.0	3.1	9.3	4.2	13.1	2.1	1.0	6.1	5.3	4.5	4.2	4.6	4.3	3.9	3.6	29.5
		16	0.6	0.5	1.3	4.0	6.8	3.8	0.4	2.3	4.4	1.8	0.1	1.5	3.3	3.4	0.9	1.0	4.4	0.9
		17	3.4	3.3	3.4	3.0	8.4	3.2	4.2	1.3	1.6	2.3	1.6	0.7	3.3	2.6	1.1	1.8	2.5	6.6
		18	10.2	7.4	9.7	2.6	8.7	3.7	11.3	1.7	0.6	5.6	4.8	3.8	3.7	4.0	3.9	3.5	2.6	26.3
	Bostan	19	6.9	2.2	9.5	2.7	6.6	2.1	2.3	0.5	2.0	1.6	2.5	3.8	1.7	1.1	1.6	2.1	2.0	7.3
		20	11.9	3.7	18.3	1.1	4.5	2.8	4.2	2.9	1.4	4.1	4.1	9.6	1.8	2.6	4.2	3.8	3.4	14.2
		21	7.2	4.7	9.8	1.0	5.1	3.0	5.0	2.3	2.8	4.7	3.8	4.0	1.5	1.2	2.4	4.4	0.4	17.0
		22	4.5	2.1	6.8	3.3	8.0	2.6	2.4	0.6	2.3	1.6	1.5	2.2	2.7	1.8	0.5	1.6	2.5	5.6
		23	12.1	2.9	18.0	0.9	3.1	1.9	2.9	2.4	2.4	3.1	4.3	10.5	1.1	3.5	4.0	3.6	4.9	11.3
		24	6.4	1.8	10.2	1.5	4.5	1.6	2.0	0.4	2.0	1.4	2.1	4.6	0.9	0.9	1.1	1.8	1.8	6.7
		25	9.1	3.4	12.8	0.6	3.9	2.5	3.7	2.4	1.1	3.9	3.7	6.1	1.4	1.5	3.1	3.7	1.4	13.6
		26	8.9	2.6	12.3	0.9	1.4	1.0	2.7	2.2	4.4	2.2	4.4	6.3	0.5	2.0	2.8	3.9	2.7	10.2
		27	32.1	4.5	40.7	3.8	1.2	1.4	5.5	7.5	9.7	2.6	12.4	30.0	4.3	19.6	14.5	5.5	21.3	13.9
		28	6.4	1.2	8.7	1.2	2.0	0.5	1.5	0.7	2.0	0.4	2.6	4.2	0.7	1.7	1.8	1.3	2.5	5.9
		29	4.8	1.2	7.3	4.1	6.5	2.3	1.4	0.6	4.1	0.6	1.9	3.4	2.3	2.8	1.0	1.2	4.8	4.8
		30	6.4	3.4	10.1	1.2	4.7	1.8	3.5	1.1	2.1	2.9	2.9	4.0	1.0	0.4	1.4	3.2	0.8	11.7
	Narenciye	31	13.8	3.4	19.4	6.4	1.2	1.1	6.0	3.6	17.0	1.4	8.6	20.3	5.6	22.3	7.8	3.5	37.7	7.2
		32	21.7	5.2	27.9	4.3	0.2	1.3	7.4	7.6	12.1	2.6	12.4	20.4	4.5	12.3	10.7	6.7	14.7	14.1
		33	2.1	2.6	1.7	17.7	19.6	13.1	2.4	9.6	21.8	5.3	1.6	2.4	11.7	19.0	3.5	3.3	25.9	4.5
		34	3.4	3.9	2.8	40.6	45.0	26.6	4.4	20.6	38.5	8.6	2.7	1.0	33.9	34.3	5.5	5.2	42.5	6.2
		35	8.8	2.2	15.3	7.9	2.9	1.8	4.0	3.3	15.2	1.2	5.6	12.4	6.3	12.9	4.9	2.4	20.6	7.2
		36	23.5	4.6	30.9	4.4	0.5	0.7	6.8	6.6	11.5	1.9	12.3	22.2	4.1	13.2	11.0	5.7	15.4	13.3
	Toprak	37	9.0	3.2	12.5	0.4	2.3	2.1	3.1	3.0	3.1	3.4	4.4	6.0	0.9	2.0	2.9	4.4	1.7	12.4
		38	6.7	3.8	10.0	0.9	5.2	2.2	4.8	1.0	0.6	3.4	2.6	3.5	1.2	0.2	1.6	2.7	0.5	13.7
		39	7.4	4.7	10.4	0.8	5.5	3.1	5.9	1.9	0.7	4.7	3.4	3.9	1.7	0.7	2.4	3.6	0.2	18.5
		40	16.5	4.6	21.0	3.1	0.2	1.3	5.8	7.2	7.7	2.5	9.7	13.9	3.8	8.5	8.4	6.4	8.6	13.6
		41	10.4	3.2	13.4	3.6	0.4	0.3	4.1	3.4	8.1	1.2	6.5	8.9	2.6	5.3	4.8	4.2	8.3	10.0
		42	7.8	1.0	11.9	1.3	2.9	1.2	1.4	0.7	2.4	0.7	2.7	7.0	0.9	2.5	1.9	1.3	3.6	4.0
		43	15.6	3.3	23.4	1.0	3.1	2.4	3.3	3.6	2.9	3.7	5.4	13.4	1.7	5.3	5.3	4.4	5.5	12.0
		44	4.5	1.7	7.8	1.7	5.7	2.3	1.9	0.6	1.5	1.7	1.2	2.7	1.6	0.7	0.4	1.5	1.1	3.6
		45	7.2	5.1	9.5	1.0	5.7	3.1	6.0	1.8	1.0	4.9	3.6	3.4	1.8	1.1	2.4	3.9	0.1	19.4

Tablo 8.23. Başlangıç aşamasında 18 test grubunun atandıkları kontrol sınıfları.

Test Grubu	Kontrol Sınıfı
1	Tahıl
2	Patates
3	Tahıl
4	Toprak
5	Toprak
6	Toprak
7	Patates
8	Bostan
9	Toprak
10	Bostan
11	Patates
12	Patates
13	Bostan
14	Toprak
15	Toprak
16	Patates
17	Toprak
18	Tahıl

Tablo 8.24. Başlangıç aşama sonunda elde edilen sınıflandırma hata matrisi.

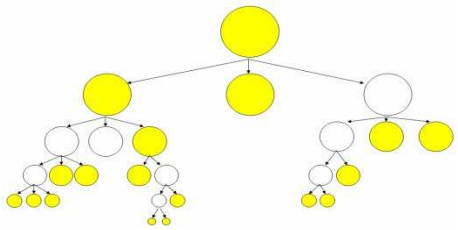
	Tahıl	Patates	Bostan	Narenciye	Toprak	Toplam	Kullanıcı Doğruluğu
Tahıl	9691	285	178	21	78	10253	0.95
Patates	3085	1986	1068	195	473	6807	0.29
Bostan	353	418	2633	381	1218	5003	0.53
Narenciye	0	0	0	0	0	0	----
Toprak	401	1952	7428	970	6786	17537	0.39
Toplam	13530	4641	11307	1567	8555	39600	
Üretici Doğruluğu	0.72	0.43	0.23	----	0.79		0.53



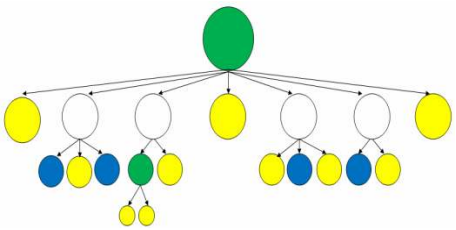
Şekil 8.19. Başlangıç aşaması sonunda elde edilen sınıflandırma sonucu oluşturulan renkli sınıflandırma konusal haritası.

8.3.4. Çok Bantlı Uydu Görüntü Verisinin Grupların İnceltilmesiyle Karma Kümeleme Analizi İle Alan Bazında Eğitimli Sınıflandırma Sonucu

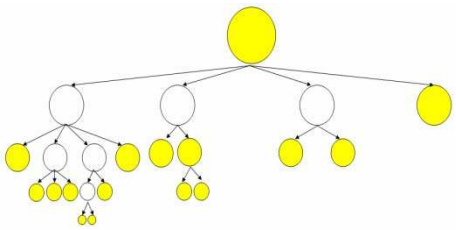
Başlangıç aşamasında elde edilen 18 test grubunun her birine Bölüm 8.2.4'de verilen verideki grupların inceltilmesiyle karma kümeleme analizi için algoritmanın adımları uygulandığında elde edilen grup yapıları Şekil 8.20'de verilmiştir. Karma kümeleme sonucu elde edilen homojen grupların, Bölüm 8.3.3.1'de verilen kontrol gruplarına olan uzaklıkları hesaplanmıştır. Bölüm 8.2.3'te verilen karar kuralına göre her bir test grubu bir kontrol sınıfına atanır. Tüm bu işlemlerin sonucunda elde edilen renkli sınıflandırma konusal haritası Şekil 8.21'de verilmiştir. Karma kümeleme sonucu elde edilen sınıflandırma hata matrisi Tablo 8.25'te verilmiştir.



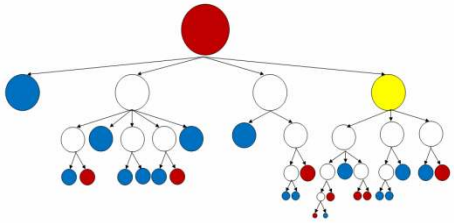
(a) Test 1 Grubu



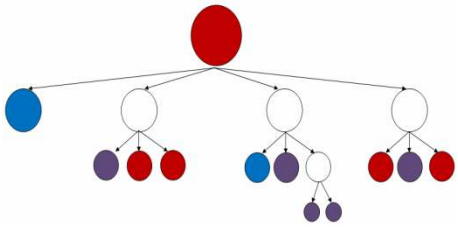
(b) Test 2 Grubu



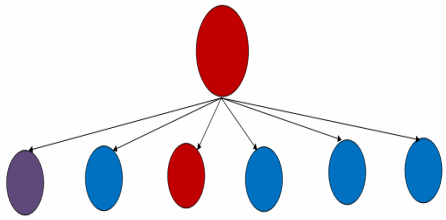
(c) Test 3 Grubu



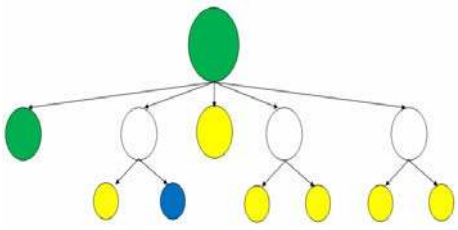
(d) Test 4 Grubu



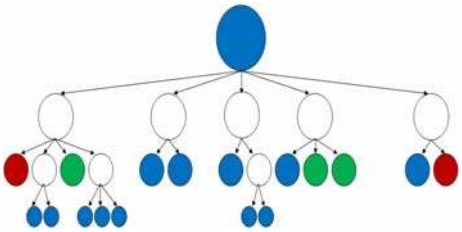
(e) Test 5 Grubu



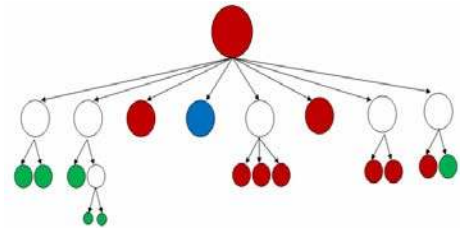
(f) Test 6 Grubu



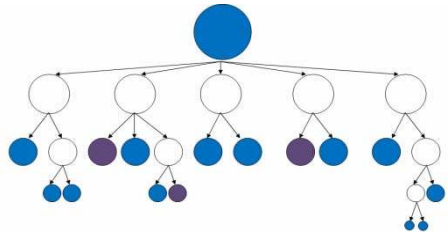
(g) Test 7 Grubu



(h) Test 8 Grubu

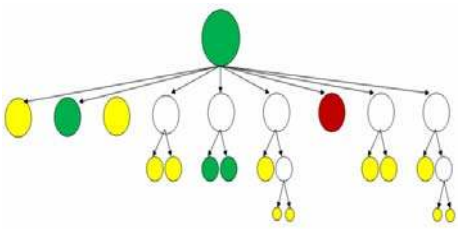


(i) Test 9 Grubu

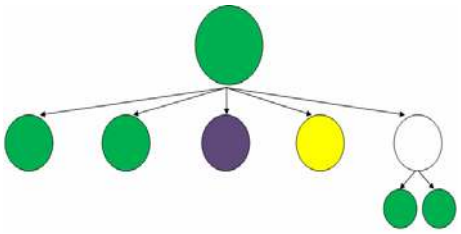


(j) Test 10 Grubu

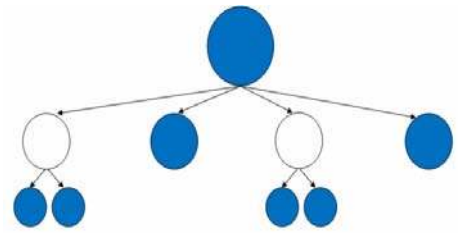
Şekil 8.20. Verideki grupların inceltilerek karma kümeleme analizinde başlangıç aşamasında elde edilen test gruplarının alt grup yapısı.



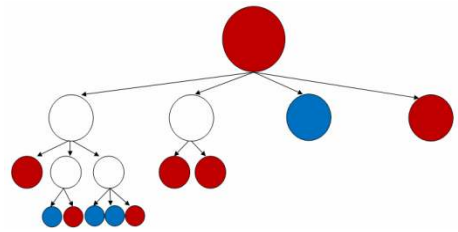
(k) Test 11 Grubu



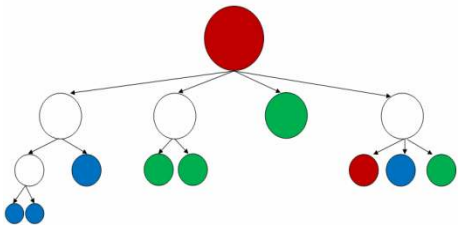
(l) Test 12 Grubu



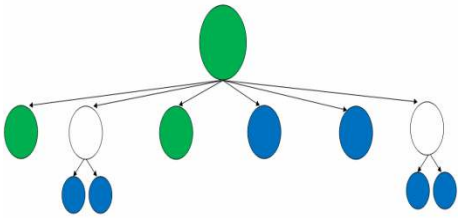
(m) Test 13 Grubu



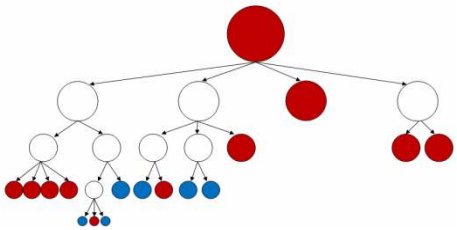
(n) Test 14 Grubu



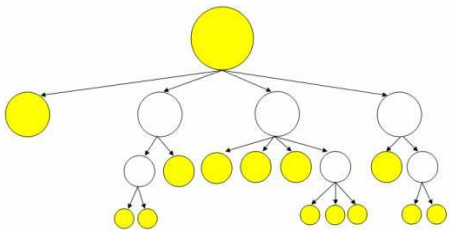
(o) Test 15 Grubu



(p) Test 16 Grubu



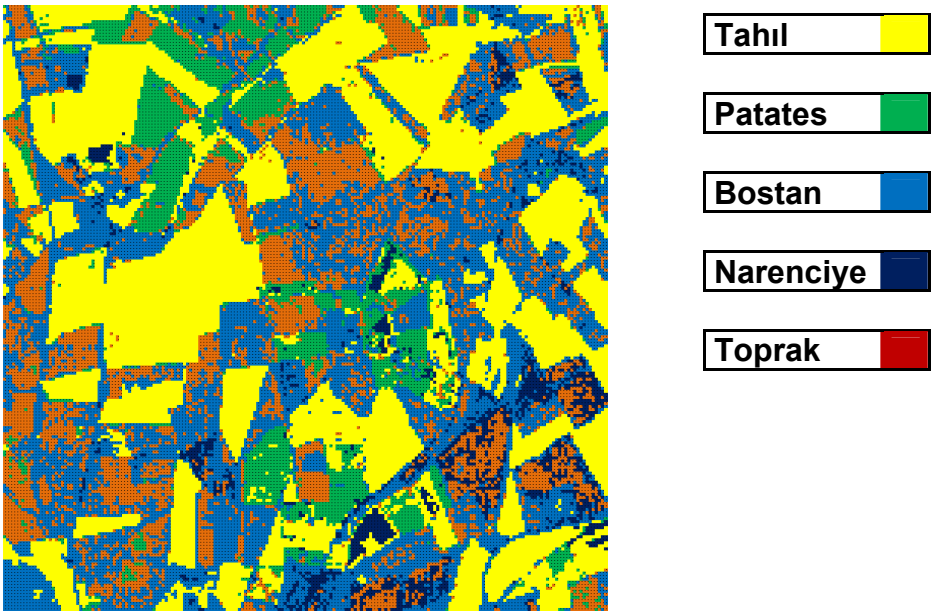
(r) Test 17 Grubu



(s) Test 18 Grubu

Şekil 8.20 (Devamı). Verideki grupların inceltilerek karma kümeleme analizinde başlangıç aşamasında elde edilen test gruplarının alt grup yapısı.

Çok bantlı uydu görüntü verisine, verideki grupların inceltilerek karma kümeleme analizi sonunda elde edilen renkli sınıflandırma konusal haritası Şekil 8.21'de gösterilmiştir. Sınıflandırma hata matrisi Tablo 8.25'de verilmiştir.



Şekil 8.21. Çok bantlı uydu görüntü verisine, verideki grupların inceltilerek karma kümeleme analizi sonunda elde edilen renkli sınıflandırma konusal haritası.

Tablo 8.25. Çok bantlı uydu görüntü verisine, verideki grupların inceltilerek karma kümeleme analizi sonunda elde edilen sınıflandırma hata matrisi.

	TAHIL	PATATES	BOSTAN	NARENCİYE	TOPRAK	TOPLAM	Kullanıcı Doğruluğu
TAHIL	11887	875	593	52	285	13692	0.87
PATATES	492	2764	328	10	538	4132	0.67
BOSTAN	746	482	7997	557	3132	12914	0.62
NARENCİYE	196	88	323	699	578	1884	0.37
TOPRAK	209	432	2066	249	4022	6978	0.58
TOPLAM	13530	4641	11307	1567	8555	39600	
Üretici Doğruluğu	0.88	0.60	0.71	0.45	0.47		0.69

Tablo 8.25'teki sonuçlara göre herhangi bir ön bilgi olmadan çok bantlı uydu görüntü verisine, verideki grupların inceltilerek karma kümeleme analizi uygulandığında elde edilen doğru sınıflandırma yüzdesi %69'dur. Bu oran bu tür sınıflandırma için iyi bir sonuçtur.

9. SONUÇ VE ÖNERİLER

Bu tez çalışması dokuz bölümden oluşmaktadır. Birinci bölümde karma dağılım modelinin uygulama alanları ana hatlarıyla verilmiştir. İkinci bölümde çok değişkenli kümeleme analizi ve karma kümeleme analizi ile ilgili yapılan önceki çalışmalar incelenmiştir. Üçüncü bölümde çok değişkenli kümeleme analizinde kavramlar, tanımlar, gösterimler ve yöntemler araştırılmıştır. Dördüncü bölümde çok değişkenli karma normal dağılım modeline dayalı kümeleme analizi ele alınmıştır. Beşinci bölümde karma kümeleme analizinde kullanılan model seçimi kriterleri incelenmiştir.

Altıncı bölümde karma kümeleme analizinde aday bileşen küme sayısının belirlenmesi için yeni bir yöntem oluşturulmuştur. Oluşturulan yöntem literatürde yer alan veri setlerinin çoğu için iyi sonuçlar vermiştir. Bazı veri setleri için aralık oldukça geniştir. Bu sorun verinin boyutunun yüksek olması ve verideki değişkenler arasındaki ilişkiden kaynaklanmaktadır. Değişken seçimi ya da veri indirgemeyle bu sorunun çözülebileceği öngörülmektedir.

Ayrıca bu bölümde önerilen mantıksal yapı daha da geliştirilerek bileşen küme sayısı ve aday modeller için bir genetik algoritmanın oluşturulması sağlanabilir.

Yedinci bölümde çok değişkenli veri setindeki grupların inceltilerek karma normal kümeleme analizi için yeni bir yöntem geliştirilmiştir. Geliştirilen yöntemin farklı veri setlerindeki sonuçları incelenmelidir. Geliştirilen yöntemin performansının diğer yöntemlerin performanslarıyla karşılaştırılmalıdır.

Çok değişkenli veride karmaşık ya da iç içe kümelenme bulunduğunda, önerilen yöntemin kullanılması durumunda grupların inceltilerek karma dağılım modeline dayalı kümelemenin etkinliği daha da artmaktadır.

Sekizinci bölümde çok değişkenli veri setindeki grupların inceltilerek karma normal kümeleme analizi, uzaktan algılanmış çok bantlı uydu görüntü verisinin kümelenmesi ve alan bazında eğitimli sınıflandırılması uygulaması verilmiştir. Bu bölümde yeni geliştirilen bir yöntemin uygulaması gerçekleştirilmiştir. Çalışmada kenar çizgi belirleme algoritması gibi bir ön bilgi ya da ek bilgi olmaksızın uzaktan

algılanmış çok bantlı uydu görüntü verisinin sınıflandırılmasında verideki grupların inceltilerek karma kümeleme analizi yapılmıştır. Karma kümeleme analizine dayalı olarak elde edilen genel sınıflandırma doğruluk yüzdesi %69 dür. Bu değer, bu tür sınıflandırma için kabul edilebilir bir doğruluk değeridir.

Dokuzuncu ve son bölümde sonuçlar ve öneriler belirtilmiştir.

KAYNAKLAR

AITCHISON, J. (1986). The statistical analysis of compositional data. Methuen, New York. 416p.

AKAIKE, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In: B. N. PETROV and F. CSAKI, eds. Second International Symposium on Information Theory. Budapest: Academia Kiado, pp. 267–281.

AKDENİZ, F. (2004). Olasılık istatistik. Nobel Kitapevi. Adana.

ANDERSON, E. (1935). The Irises of the Gaspé peninsula. Bulletin of the American Iris Society 59, 2-5.

ARABIE, P., HUBERT, L.J. AND DE SOETE, G. (1996). Clustering and Classification. World Scientific, Singapore.

BALDI, P. AND HATFIELD, G.W. (2002). DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling, Cambridge University Press, Cambridge, UK.

BANFIELD, J.D. AND RAFTERY, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803-822.

BASFORD, K.E. AND MC LACHLAN,G J.(1985). Likelihood Estimation with Normal Mixture Models Journal of the Royal Statistical Society, Vol. 34 Issue 3, p:282-289.

BEN-DOR, A. AND YAKHINI, Z. (1999). Clustering gene expression patterns. RECOMB, 33-42.

BENSMAIL, H., CELEUX, G., RAFTERY, A.E., ROBERT, C.P. (1997). Inference in model-based cluster analysis. Statistics and Computing, Volume 7, Number 1, pp. 1-10(10).

BIERNACKI, C. AND GOVAERT, G. (1997). Using the Classification Likelihood to Choose the Number of Clusters. Computing Science and Statistics, 29 (2), 451-457.

BIERNACKI, C., AND GOVAERT, G. (1999). Choosing Models in Model-Based Clustering and Discriminant Analysis, Journal of Statistical Computation and Simulation, 64, 49–71.

BIERNACKI, C., CELEUX, G., AND GOVAERT, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. Pattern Recognition Letters 20, 267-272.

BIERNACKI, C., CELEUX, G., GOVAERT, G. AND LANGROGNET, F. (2005). Model-based cluster analysis and discriminant analysis with the MIXMOD software. INRIA Institute National de Recherche en Informatique et en Automatique. No:0302, January 2005, Rapport Technique.

BIERNACKI, C., CELEUX, G., GOVAERT, G. AND LANGROGNET, F. (2006). Model-based cluster analysis and discriminant analysis with the MIXMOD software. Computational Statistics and Data Analysis, 51, 587-600.

BINDER, D.A. (1978). Bayesian Cluster Analysis, Biometrika, 65, 31-38.

BINS, J. AND DRAPER, B. A. (2001). Feature Selection from Huge Feature Sets. International Conference on Computer Vision, Vancouver, IEEE CS Press.

BLUM, A.L. AND LANGLEY, P. (1997). Selection of Relevant Features and Examples in Machine Learning. Artificial Intelligence, 97(1–2), 245–271.

BOZDOGAN, H. (1987). Model selection and Akaike's Information criterion (AIC): The general theory and its analytical extentions, Psychometrika, 52, No. 3, Special Section(invited paper), 345-370.

BOZDOGAN, H. (1988). ICOMP: A new model selection criterion, in H. H. Bock(ed.), Classification and Related Methods of Data Analysis, North-Holland, Amsterdam, April, 599-608.

BOZDOGAN, H. (1990A). On the information based measure of covariance complexity and its application to evaluation of multivariate linear models, Communications in Statistics, Theory and Methods, 19(1), 221-278.

BOZDOGAN, H. (1990B). Multi-sample cluster analysis of the common principal component model in K groups using an entropic statistical complexity criterion, invited paper presented at the international Symposium on Theory and Practice of Classification, 16-19 December, Puschino, Soviet Union.

BOZDOGAN, H. (1993). Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix. In Information and Classification, O. Opitz, B. Lausen, and R. Klar (eds.), Heidelberg: Springer-Verlag, pp. 40-54.

BOZDOGAN, H. (1994). Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Information Measure of Complexity. In Proceedings of the Fist US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, H. Bozdogan (Eds.), Vol. 2, Boston, Kluwer Academic Publishers, pp. 69-113.

BRADLEY, P., FAYYAD, U. AND REINA, C. (2000). Clustering very large databases using EM mixture models, in: Proceedings of 15th International Conference on Pattern Recognition, Vol. 2, pp. 76–80.

CALINSKI, R.B. AND HARABASZ, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3, 1-27.

CAMPBELL, N. A. AND MAHON, R. J. (1974). A multivariate study of variation in two species of rock crab of the genus Leptograpsus. Australian Journal of Zoology, 22, 417–425.

CARUANA, R. AND FREITAG, D. (1994). Greedy attribute selection. In Proceedings of the Eleventh International Conference on Machine Learning, pages 180-189.

CELEUX, G. AND GOVAERT, G. (1992). A Classification EM Algorithm for Clustering and Two Stochastic versions. Computational Statistics and Data Analysis, 14, 315-332.

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.

CELEUX, G. AND GOVAERT, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition, 28, 781-793.

CELEUX, G. AND ROBERT, C.L. (1993). Une histoire de discrétisation (avec commentaires). La Revue de Modulad 11, 7-42.

CELEUX, G., AND SOROMENHO, G. (1996). An Entropy Criterion for Assessing the Number of Clusters in a Mixture, Journal of Classification, 13,195–212.

CHEESEMAN, P. AND STUTZ, J. (1995). Bayesian classification (Autoclass C); Theory and results, in Advances in Knowledge Discovery and Data Mining (U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds), pp. 61-83. American Association for Artificial Intelligence Press, Menlo Park, CA.

CHO, R.J., CAMPBELL, M.J., WINZELER, E.A., STEINMETZ, L., CONWAY, A., WODICKA, L., WOLFSBERG, T.G., GABRIELIAN, A.E., LANDSMAN, D., LOCKHART, D.J. and DAVIS, R.W. (1998). A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Molecular Cell, Vol. 2, 65–73.

CLIFF, A.D., HAGGETT, P., SMALLMAN-RAYNOR, M.R., STROUP, D.F. AND WILLIAMSON, G. D. (1995). The application of multidimensional scaling methods to epidemiological data. Statistical methods in medical research. 4(2):102-23.

CONWAY, J.A., BROWN, L.M.J., VECK, N.J., WIELOGORSK, A. AND BORGEAUD, M. (1991). A model based system for crop classification from radar imagery. Antennas and Propagation (ICAP 91) Seventh International Conference on (IEE), Vol. 2, pp.616-619.

CUNNINGHAM, K. and OGILVIE, J. (1972). Evaluation of hierarchical grouping techniques: A preliminary study . Computer Journal , 15 : 209 – 213.

DASGUPTA, A. AND RAFTERY, A.E. (1998). Detecting Features in Spatial Point Processes with Clutter via Model-based Clustering. Journal of the American Statistical Association, 93, 294-302.

DAVIES, D. AND BOULDIN, D. (1979). A cluster separation measure . IEEE Transactions on Pattern Analysis and Machine Intelligence , 1 : 224 – 227.

DAWKINS, R. (1989). The Selfish Gene, new edn, Oxford University Press, Oxford.

DAY, N.E. (1969). Estimating the components of a mixture of normal distributions. Biometrika, 56, 463-474.

DEAN, N., MURPHY, T.B. and DOWNEY, G. (2006). Updating Classification Rules with Unlabeled Data with Applications in Food Authenticity Studies. Journal of the Royal Statistical Society, Series C (Applied Statistics), 55, 1–14.

DELLAPORTAS, P. (1998). Bayesian classification of Neolithic tools. Applied Statistics, 47, 279–297.

DEMPSTER, A. P., LAIRD, N. AND RUBIN, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39, 1–38.

DICE, L.R. (1945). Measures of the amount of ecologic association between species. Ecology 26, 297–302.

DUDA, R.O. AND HART, P.E. (1973). Pattern Classification and Scene Analysis, Wiley, New York.

DUDA, R.O., HART, P.E. AND STORK, D.G. (2001). Pattern Classification. John Wiley & Sons, INC. New York.

DUNN, J. (1974). Well separated clusters and optimal fuzzy partitions . Journal of Cybernetics, 4 : 95 – 104.

DURAN, B.S. AND ODELL, P.L. (1974). Cluster Analysis, New York: Springer-Verlag.

DY, J. G., AND BRODLEY, C. E. (2000). Feature subset selection and order identification for unsupervised learning. Proceedings of the Seventeenth International Conference on Machine Learning (pp. 247-- 254). Morgan Kaufmann.

EDWARDS, A.W.F. AND CAVALLI-SFORZA, L.L. (1965). A method for cluster analysis. Biometrics, 21, 362-375.

EFRON, B. (1979). Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics, 7, 1-26.

ELLIS, T.E., RUDD, M.D., RAJAB M.H. AND WEHRLY T.E. (1996). Cluster analysis of MCMI scores of suicidal psychiatric patients: Four personality profiles. Journal of Clinical Psychology, 52, 411-22.

EROL, H. and AKDENIZ, F. (2005). A per-field classification method based on mixture distribution models and an application to Landsat Thematic Mapper data. International Journal of Remote Sensing Vol. 26, No. 6, 20, 1229–1244.

EVERITT, B. S. (1980). Cluster analysis. Halsted Press, New York, p. 25-27

EVERITT, B. S. (1993). Cluster analysis. 3rd ed. London: Edward Arnold.

EVERITT, B.S. AND DUNN, G. (2001). Applied Multivariate Data Analysis. Arnold, London.

EVERITT, B.S., LANDAU, S. AND LEESE, M. (2001). Cluster Analysis. Oxford University Press Inc., New York, NY.

FAUNDEZ-ABANS, M., ORMENO M.I. AND DE OLIVEIRA-ABANS, M. (1996). Classification of planetary nebulae by cluster analysis and artificial neural Networks. Astronomy and Astrophysics Supplement Series, 116, 395-402.

FARMER, A.E., MCGUFFIN, P. AND SPITZNAGEL E.L. (1983). Heterogeneity in schizophrenia: A cluster analytic approach. Psychiatric Research, 8, 1-12.

FIGUEIREDO, M. AND JAIN, A.K. (2002). Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence - PAMI, vol. 24, no. 3, pp. 381-396, March 2002.

FISHER, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics(7), pp. 179-188.

FRALEY, C. (1998). Algorithms for model-based Gaussian hierarchical clustering. SIAM Journal on Scientific Computing. (20). 270-281.

FRALEY, C. AND RAFTERY, A.E. (1998). How many clusters? Which clustering method?-Answers via model-based cluster analysis. Computer Journal, 41, 578-588.

FRALEY, C. AND RAFTERY, A.E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97, 611-631.

FRALEY, C. AND RAFTERY, A.E. (2007). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. Journal of Classification, 24, 155-181.

FRALEY C., RAFTERY A.E. AND WEHRENS R. (2005). Incremental Model-based Clustering for Large Datasets with Small Clusters." Journal of Computational and Graphical Statistics, 14, 1–18.

FRIEDMAN, H.P. AND RUBIN, J. (1967). On some invariant criteria for grouping data. Journal of American Statistical Association, 62, 1159-1178.

FRIGUI H. AND KRISHNAPURAM R., (1999). A robust competitive clustering algorithm with applications in computer vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 5, pp 450-465.

GALIMBERTI, G. AND SOFFRITTI, G. (2007). Model-based methods to identify multiple cluster structures in a data set, Computational Statistics & Data Analysis, 52, 520-536.

GANESALINGAM, S. (1989). Classification and Mixture Approach to Clustering via Maximum Likelihood, Applied Statistics, 38, 455-466.

GOLDBERGER, J. AND ROWEIS, S. (2005). Hierarchical Clustering of a Mixture Model. Neural Information Processing Systems 17 (NIPS'04). pp 505-512.

GORDON, A.D. (1998). Cluster validation. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.H. Bock and Y. Baba, Editors, Studies in Classification, Data Analysis, and Knowledge Organization: Data Science, Classification, and Related Methods, Springer, Tokyo (1998), pp. 22–39.

GORDON, A.D. (1999). Classification, (2nd edn). Chapman & Hall, New York, NY.

GOWER, J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27, 857-872

GOWER, J. C. AND LEGENDRE, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. Journal of classification. New York: Springer. 5-48.

GREEN, P.E., FRANK, R.E. AND ROBINSON, P.J. (1967). Cluster analysis in test market selection. Management science, 13, 387-400.

HABBEMA, J.D.F., HERMANS, J. AND VAN DEN BROEK, K. (1974), A stepwise discriminant analysis program using density estimation. In Compstat 1974, Proc. Computational Statistics, pp. 101-110. Vienna: Physica.

HALBE, Z. AND ALADJEM, M. (2005). Model-based mixture discriminant analysis: an experimental study. Pattern Recognition. 38, 437-440.

HALKIDI, M., BATISTAKIS, Y., AND VAZIRGIANNIS, M. (2002A). Cluster Validity Methods: Part I. SIGMOD Record 31(2): 40-45.

HALKIDI, M., BATISTAKIS, Y., AND VAZIRGIANNIS, M. (2002B). Clustering Validity Checking Methods: Part II. SIGMOD Record 31(3): 19-27.

HAN, J. AND KAMBER, M. (2000). Data mining: concepts and techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA.

HAND, D.J., DALY, F., LUNN, A.D., McCONWAY, K.J. and OSTROWSKI, E. (eds.) (1994). A Handbook of Small Data Sets, London: Chapman & Hall.

HANSEN, P., HERTZ, A. AND QUINODOZ, N. (1997). Splitting trees. Discrete Mathematics 165-166: 403-419.

HANSEN, P. AND JAUMARD, B. (1997). Cluster analysis and mathematical programming. Mathematical Programming 79, pp. 191–215.

HASSELBLAD, V. (1966). Estimation of parameters for a mixture of normal distributions. Technometrics 8, 431-444.

HASSELBLAD, V. (1969). Estimation of finite mixture distributions from the exponential family. Journal of the American Statistical Association 64, 1459-1471.

HATHAWAY, R.J. (1986). Another interpretation of the EM algorithm for mixture distributions. Statistics & Probability Letters 4, 53-56.

HOUDSON, F.R. (1971). Numerical typology and prehistoric archaeology, in Mathematics in the Archaeological and Historical Sciences, Edinburgh University Press, Edinburgh.

HUBERT, L. (1974). Approximate evaluation techniques for the single - link and complete -link hierarchical clustering procedures . Journal of the American Statistical Association, 69 : 698 – 704.

HUBERT, L. AND ARABIE, P. (1985). Comparing partitions. J. Of Classification, pages 193–218, 1985.

ISHIGURO, M., SAKAMOTO, Y. AND KITAGAWA, G. (1997). Bootstrapping log-likelihood and EIC, an extension of AIC. Annals of the Institute of Statistical Mathematics 49,411-434.

JACCARD, P. (1908). Nouvelles recherches sur la distribution florale. Bull. Soc. Vaud. Sci. Nat. 44 (1908), pp. 223–270.

JAIN, A.K. AND DUBES, R.C. (1988). Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

JAIN, A. K. AND FLYNN, P.J. (1996). Image segmentation using clustering. In Advances in Image Understanding: A Festschrift for Azriel Rosenfeld, N. Ahuja and K. Bowyer, Eds, IEEE Press, Piscataway, NJ, 65-83.

JAIN, A.K. AND ZONGKER, D. (1997). Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2), 153–158.

JEFFREYS, H. (1935). Some tests of significance, treated by the theory of probability. Proceedings of the Cambridge Philosophy Society, 31:203–222.

JEFFREYS, H. (1939). Theory of Probability, Oxford: Claredon Press.

JENSEN, R.E. (1969). A Dynamic Programming Algorithm for Cluster Analysis, Operations Research Vol. 17, No. 6, November-December 1969, pp. 1034-1057 DOI: 10.1287/opre.17.6.1034.

JIN, H.-D., LEUNG, K.-S. AND WONG, M.-L. (2002). Scaling-up model-based clustering algorithm by working on clustering features. In: Proceeding of Third International Conference on Intelligent Data Engineering and Automated Learning. pp. 569-575.

KASS, R. E. AND RAFTERY, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90:773–795.

KAUFMAN, L., AND ROUSSEEUW, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, Inc.

KERIBIN, C. (2000). Consistent estimation of the order of mixture models, Sankhya, Series A 62 (1), pp. 49–66.

KOHAVI, RON. AND JOHN, G. H. (1997). Wrappers for Feature Subset Selection. Artificial Intelligence, 97(1–2), 273–324.

KOLLER, D., AND SAHAMI, M. (1996). Toward Optimal Feature Selection. In: Proceedings of the Thirteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, pp. 284–292.

KONISHI, S. AND KITAGAWA, G. (1996). Generalized information criteria in model selection. Biometrika 83, 875-890.

KULLBACK, S. AND LEIBLER, R.A. (1951). On information and sufficiency. Annals of Mathematical Statistics 22, 79-86.

KURTZ, A., MOLLER, H.J., BAINDL, G., BURK, F., TORHORST, A., WACHTLER, C. AND LAUTER, H. (1987). Classification of parasuicide by cluster analysis. British Journal of Psychiatry, 150, 520-525.

LANCE, G.N. AND WILLIAMS, W.T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems. Comput. J. 9, 60–64.

LAW, M.H., JAIN, A.K. AND FIGUEIREDO, M.A.T. (2003). Feature Selection in Mixture-Based Clustering," Advances in Neural Information Processing Systems 15, pp. 625-632, Cambridge, Mass.: MIT Press.

LEWIS, S.M. AND RAFTERY, A.E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. Journal of the American Statistical Assocation, 92, 648-655.

LI Q., FRALEY C., BUMGARNER R.E., YEUNG K.Y. and RAFTERY A.E. (2005). Donuts, Scratches and Blanks: Robust Model-based Segmentation of Microarray Images. Bioinformatics, 21, 2875–2882.

MACNAUGHTON-SMITH, P., WILLIAMS, W.T., DALE, M.B. and MOCKETT, L.G. (1964). Dissimilarity analysis: a new technique of hierarchical sub-division. Nature 202, 1034-1035.

MACQUEEN. J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, eds. L. M. L. Cam and J. Neyman, Berkeley, CA: University of California Pres, pp, 281-297.

MAK, B. AND BARNARD, E. (1996) "Phone clustering using the Bhattacharyya distance," in Proc. Int. Conf. Spoken Language Processing, vol. 4, pp. 2005–2008.

MALLORY-GREENOUGH, J.M. AND GREENOUGH, J.D. (1998). New data for old pots: Trace element characterization of Ancient Egyptian pottery using ICP-MS. Journal of Archaeological Science, 25, 85-97.

MAO, J. AND JAIN, A. K. (1996). A self-organizing network for hyperellipsoidal clustering (hec). IEEE Transactions on Neural Networks, vol. 7, pp. 16-29.

MARTINEZ, W.L. AND MARTINEZ, A.R. (2005). Explanatory Data Analysis with MatLab, Computer Science and Data Analysis Series, Chapman & Hall/CRC. NewYork.

MCLACHLAN, G.J. (1992). Discriminant Analysis and Statistical Pattern Recognition. New York: Wiley.

MCLACHLAN, G.J. (2000). On the grouping of treatments following an ANOVA. Journal of Agricultural, Biological, and Environmental Statistics

MCLACHLAN, G.J. AND BASFORD, K.E. (1988). Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

MCLACHLAN, G.J. AND KRISHNAN, T. (1997). The EM Algorihm and Extensions. John Wiley & Sons, Inc. New York.

MCLACHLAN, G.J. AND PEEL, D. (1996). An algorithm for unsupervised learning via normal mixture models. In ISIS: Information, Statistics and Induction in Science, D.L. Dowe, K.B. Korb, and J.J. Oliver (Eds.). Singapore: World Scientific Publishing, pp. 354-363.

MCLACHLAN, G.J., PEEL, D., BASFORD, K.E., and ADAMS, P. (1999). The EMMIX software for the fitting of mixtures of normal and t-components. Journal of Statistical Software 4, No. 2.

MCLACHLAN, G. AND PEEL, D. (2000). Finite Mixture Models. John Wiley & Sons, Inc. New York.

NEWMAN, D. J., HETTICH, S., BLAKE, C. L. AND MERZ, C. J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/ ~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.Golub, G. H., Heath, C. G., and Wahba, W. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21, 215-223.

OLDENDORF, W. AND OLDENDORF, W. JR. (1988). Basics of Magnetic Resonance Imaging. Boston: Martinus Nijhoff.

PAL, N.R., BEZDEK, J.C., and TSAO, E.C.K. (1993). Generalized clustering networks and Kohonen's self organizing scheme. IEEE Trans. Neural Networks 4, 549-557.

PAYKEL, E.S. (1971). Classification of Depressed Patients: A Cluster Analysis Derived Grouping. British Journal of Psychiatry, 118, 275-288.

PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. London* A 185 , 71-110.

PICHLER, O., TEUNER, A., and HOSTICKA, B. J. (1996). A comparison of texture future extraction using adaptivegabor filtering, Pyramidal and Tree Structured Wavelet Transforms, Elsevier Science on Pattern Recognition, Vol.29 No. 5, pp. 733-742.

PILOWSKY, I., LEVINE, S. AND BOULTON, D.M. (1969). The classification of depression by numerical taxonomy. British Journal of Psychiatry, 115, 937–945.

PUDIL, P., NOVOVICOVA, J. AND KITLER, J. (1994). Floating search methods in feature selection. Pattern Recognition Letters, 15(11):1119-1125.

RAFTERY, A.E., DEAN, N. (2006). Variable Selection for Model-based Clustering. Journal of the American Statistical Association, 101, 168–178.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66: 846–850. doi:10.2307/2284239.

RAUDYS, S. AND JAIN, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, no. 3, pp. 252—264.

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.

RENDER, R.A AND WALKER, H. F. (1984) Mixture Densities, Maximum Likelihood and The EM Algorithm. Society for Industrial and Applied Mathematics vol. 26, no. 2, pp.195-239.

RICHARDS, J.A., 1986, Remote Sensing Digital Image Analysis, pp. 173–187 (New York: Springer-Verlag).

ROGERS, D.J., TANIMOTO, T.T. (1960). A computer program for classifying plants. Science 132, 1115-1118.

RUSPINI, E. H. (1970). Numerical methods for fuzzy clustering. Inform. Sci., 2, 319–350.

SANDOR, G. AND LECHEVALLIER, Y. (1977). Decoupage optimal de variables quatitatives et application a la definition d'une grille de diagnostic tiree de l'etudes des proteines seriques, First International Symposium on Data Analysis and Informatics, 665-674.

SCHORK, N. J., S. P. NATH, K. LINDPAINTNER, AND H. J. JACOB. (1996). Extensions of quantitative trait locus mapping in experimental organisms. Hypertension 28: 1104-1111.

SCHUMMER, M., NG, W.V., BUMGARNER, R.E., NELSON, P.S., SCHUMMER, B., BEDNARSKI, D.W., HASSELL, L., BALDWIN, R.L., KARLAN, B.Y. and HOOD, L. (1999). Comparative hybridization of an array of 21, 500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas. Gene 238:, 375-385.

SCHWARZ, G., (1978). Estimating the dimension of a model. Annals of Statistics 6(2):461-464.

SCOTT, A.J. AND SYMONS, M.J. (1971). Clustering methods based on likelihood ratio criteria. Biometrics, 27, 387-398.

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 J. Appl. Math. 8, no. 2, 57-69.

SEBER, C. A. F. (1984). Multivariate Observations. John Wiley & Sons, New York.

SHI, J. AND MALIK, J. (2000). Normalized Cuts and Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888-905.

SOFFRITTI, G. (2003). Identifying multiple cluster structures in a data matrix, Communications in Statistics: Simulation and Computation, 32, 1151–1177.

SOKAL, R.R. AND SNEATH, P.N.A. (1963). Principles of Numerical Taxonomy. Freeman, San Francisco.

SOROMENHO, G. (1993). Comparing approaches for testing the number of components in a finite mixture model. Computational Statistics 9, 65-78.

SOUBIRAN, C. (1993). Kinematics of the Galaxy's stellar populations from a proper motion survey. Astronomy and Astrophysics , 274, 181-188.

STEVENS, S.S. (1946).On the theory of scales of measurement. Science 103, pp. 677–680.

SYMONS, M.J. (1981). Clustering Criteria and Multivariate Normal Mixtures. Biometrics, 37, 35-43.

SMYTH, P. (2000). Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood. Statistics and Computing, 10, 63–72.

TIBSHIRANI, R., WALTHER, G. AND HASTIE, T. (2001). Estimating the number of clusters via the gap statistic, J Royal Statist Soc B, (63), 411-423.

THEODORIDIS, S. And KOUTROUMBAS, K. (2006). Pattern Recognition, third ed. Elsevier Academic Pres.

THISSEN U., SWIERENGA H., DE WEIJER A.P., WEHRENS R., MELSSEN W.J., BUYDENS L.M.C. (2005). Multivariate Statistical Process Control Using Mixture Modeling. Journal of Chemometrics, 19, 23–31.

TOHER D., DOWNEY G. and MURPHY T. (2005). A Comparison of Model-based and Regression Classification Techniques Applied to Near-Infrared Spectroscopic Data in Food Authentication Studies." Technical Report 5/10, Department of Statistics, Trinity College Dublin.

TRAN, T.N., WEHRENS, R. AND BUYDENS, L.M.C. (2006). SMIXTURE: a strategy for mixture models clustering of multivariable images J. Chemom., 19, 607-614.

VAITHYANATHAN, S. AND DOM, B. (1999). Model selection in unsupervised learning with applications to document clustering, in: Proc. 16th Int'l Conf. on Machine Learning, 1999, pp. 433–443, Morgan Kaufmann, San Francisco, CA.

VENABLES, W. N. AND RIPLEY, B. D. (1994). Modern Applied Statistics with S-Plus. Springer, 462 pages, ISBN 0 387 94350 1.

WALLACE, C.S. AND BOULTON, D.M. (1968). An information measure for classification. Computer Journal, 11, 185–194.

WARD, J. H. (1963). “Hierarchical Grouping to Optimize an Objective Function”, Journal of the American Statistical Association. 58, 234-244.

WEI, G.C.G. AND TANNER, M.A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms Journal of the American Statistical Association, Vol. 85, No. 411. pp. 699-704.

WEHRENS, R., BUYDENS, L., FRALEY, C., and RAFTERY, A.E. (2004). Model-based Clustering for Image Segmentation and Large Datasets via Sampling. Journal of Classification 21, 231–253.

WILLETT, P. (1988). Recent trends in hierarchical document clustering: a critical review. Information Processing & Management, 24(5):577-597.

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.

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.

WOLFE, J.H. (1967). NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixture Distributions, Technical Bulletin USNPRA SRM 68-2, U.S. Naval Personnel Research Activity, San Diego.

WOLFE, J.H. (1970). Pattern Clustering by Multivariate Mixture Analysis, Multivariate Behavioral Research, 5, 329-350.

XING, E.P., JORDAN. M.I. AND KARP, R.M. (2001). Feature selection for high-dimensional genomic microarray data, Proceedings of the Eighteenth International Conference on Machine Learning, (ICML2001).

XU, R. And WUNSCH, D.C. II. (2009). Clustering, John Wiley & Sons, Inc., Hoboken, New Jersey.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. AND RUZZO W. L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics Vol. 17 no. 10, Pages 977-987.

ZUBIN, J. (1938). The Mentally Ill and Mentally Handicapped in Institutions: By. Public Health Reports; supplement, No. 146. Washington: United States Government Printing Office.

ÖZGEÇMİŞ

1972 yılında Mersin'de doğdum. İlköğretimimi Abdulkadir Perşembe İlkokulunda, Orta öğretimimi Trabzon Kanuni Orta Okulunda ve Lise öğretimimi Zonguldak Fener Lisesinde aldım. 1991 yılında Çukurova Üniversitesi Fen Edebiyat Fakültesi Matematik Bölümünde başladığım lisans eğitimimi 1997 yılında tamamlayarak mezun oldum ve aynı yıl Karaman ilinin Ermenek ilçesine bağlı Çamlıca Köyü İlköğretim okulunda Matematik Öğretmeni olarak göreve başladım. 1998 yılında Milli Eğitim Bakanlığının açmış olduğu yurt dışı yüksek lisans bursluluk sınavını kazandığım için öğretmenlik görevimden ayrıldım. 1998-1999 yılları arasında Orta Doğu Teknik Üniversitesi Yabancı Diller bölümünde İngilizce hazırlık kursuna devam ettim. 1999 yılında Amerika Birleşik Devletleri'nin Florida Üniversitesinin Elektrik ve Bilgisayar Mühendisliği Bölümünde yüksek lisans eğitimime başladım. 2002 yılında yüksek lisans eğitimimi tamamlayarak Çukurova Üniversitesi Rektörlük Bilgisayar Bilimleri ve Enformatik Bölümünde Araştırma Görevlisi olarak çalışmaya başladım. 2003 yılında Ç.Ü. Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda doktora eğitimime başladım. 2003 yılında Enformatik bölümündeki görevimden ayrılarak Fen Edebiyat Fakültesi İstatistik Bölümünde Araştırma Görevlisi olarak göreve başladım. Evli ve bir çocuk babasıyım.