Araştırma Makalesi / Research Article

Gauss Karma Modellerin Özellikleri ve Değişken Parçalanmalarına Dayalı Kümeleme

Maruf GÖGEBAKAN*

Bandırma OnYedi Eylül Üniversitesi, Denizcilik Fakültesi, Denizcilik İşletmeleri ve Yönetimi Bölümü, Bandırma, Balıkesir
(ORCID: 0000-0003-0447-8311)

Öz

Bu çalışmada çok değişkenli verideki homojenlik ve heterojenlik durumları incelenmiş ve heterojen değişkenler belirlenmiştir. Değişkenlerdeki parçalanmaların (heterojenlik) normal karma dağılımlardaki bileşenlere denk geldiği gösterilmiş ve alt grup sayıları belirlenmiştir. k-ortalamalar (k-means) algoritması ile değişkenlerdeki parçalanmalara atanan gözlemler belirlenmiş ve veri gruplandırma yapılmıştır. Değişkenlerdeki her bir parçalanmanın Gauss Karma Modeldeki (GMM) bir kümelenmeye karşılık geldiği varsayımı altında muhtemel küme sayıları ve küme sayıları için aralık elde edilmiş ve küme sayılarına bağlı olarak model sayıları belirlenmiştir. Parçalanma (bileşen) sayısına bağlı model sayıları Genetik Algoritmalarla (GA) belirlenmiş ve En Çok Olabilirlik Kestirimi (MLE) algoritması ile parametreler tahmin edilmiştir. Modele dayalı kümeleme yöntemi ile Gauss Karma Modeller arasından veri yapısına uyan en iyi modelin seçimi log-olabilirlik, AIC ve BIC gibi bilgi kriterleri ile belirlenmiştir.

Anahtar kelimeler: Gauss karma model, değişken veri parçalama, genetik algoritmalar, modele dayalı kümeleme, bilgi kriteri.

Properties of Gaussian Mixture Models and Variable Segmentations Based Clustering

Abstract

In this study, homogeneity and heterogeneity in multivariate data were examined and heterogeneous variables were determined. Partitions in the variables (heterogeneity) were shown to coincide with the components of normal mixed distributions and the number of subgroups was determined. K-means algorithms were used to determine the observations of the partition of the variables and the data was grouped. Under the assumption that each partition in the variables corresponds to a cluster in the Gaussian Mixture Model (GMM), the range for the possible number of clusters and clusters was obtained and the model numbers were determined based on the clusters. The number of models based on the number of components (partition) was determined by Genetic Algorithms (GA) and the parameters were estimated with Maximum Likelihood Estimation (MLE) algorithm. With the model based clustering method, the selection of the best model matching the data structure from the Gaussian Mixture Models was determined by information criteria such as log- olabilirlik, AIC and BIC.

Keywords: Gaussian mixture model, variable data segmentation, genetics algorithm, model based clustering, information criteria.

1. Giriş

Sonlu karma modeller verilerin kümelenmesinde en etkili yöntemler arasındadır [1]. Sonlu karma dağılımlar verileri modele dayalı kümelenmesinde kullanılır. Modele dayalı kümeleme p -boyutlu çok değişkenli heterojen veriyi anlamlı altgruplara bölmek için kullanılan yöntemlerden biridir. Çok

*Sorumlu yazar: mgogebakan@bandirma.edu.tr
Geliş Tarihi: 29.11.2019, Kabul Tarihi: 20.03.2020

değişkenli karma modellerin kümelenmesinde modele dayalı kümelemenin yanında diskriminant analizi karma modellerin kümelenmesinde kullanılır [2]. Çok değişkenli normal dağılımların karmasının her bileşeni, sonlu karma dağılımlarda heterojen verideki bir kümeye denk gelir [3]. Çok değişkenli heterojen verideki kümelenmenin n tane $p-$ boyutlu $x_1, x_2, \dots, x_n$ gözleminde her biri bilinmeyen $\pi_1, \dots, \pi_k$ olasılıkları ile sonlu sayıdaki k grup koşullu olasılık yoğunluklarının karmasından geldiği varsayılır [4]. $j = 1,2, \dots, n$ için j. gözlem değeri ve x_j değişkeni için normal dağılımların karma modeli,

$$f(x_j;\theta) = \sum_{i=1}^{k} \pi_i \, f_i(x_j;\psi_i) \quad (1)$$

şeklinde yazılır. Burada $i = 1,2, \dots, k$ için π_i, $0 < \pi_i < 1$ ve $\sum_{i=1}^{k} \pi_i = 1$ olacak biçimde i. kümenin karma olasılık oranını göstermektedir. $j = 1,2, \dots, n$ için grup koşullu yoğunluk $f_i(x_j;\psi_i)$, bilinmeyen parametreler vektörü ψ_i'ye bağlıdır. Burada çok değişkenli normal karma dağılımların bilinmeyen parametrelerin $\theta = (\pi_1, \dots, \pi_k, \psi_1, \dots, \psi_k)$ vektörü Ω parametre uzayında tümünü temsil eden vektördür. Bu çalışmada $f_i(x_j;\psi_i)$'nin μ_i ortalamalı Σ_i varyans-kovaryans matrisli çok değişkenli normal dağılım olduğu varsayılır. Burada $\psi_i = (\mu_i, \Sigma_i)$ olacak şekilde bileşenlerin parametrelerinin vektörüdür. $f_i = (x_j;\mu_i,\Sigma_i)$ çok değişkenli olasılık yoğunluğu,

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

şeklinde gösterilir. Burada p veri setindeki değişken sayısıdır.

Verideki değişkenlerin bağımsız altkümelerinin çoklu kümelenme yapıları modele dayalı olarak normal karma modellerle belirlenir [5]. Veri setindeki çoklu kümelenme yapılarını tanımlamada değişkenlerin yapıları ve parçalanmaları kullanılır [6]. Normal dağılımların karma modelinde verideki alt grup yapılarını tanımlamak için kök seçim metodu kullanılır [7]. Modele dayalı kümelemede değişkenlere dayalı küme sayısının belirlenmesi ve bu küme sayılarının sınırının bulunması için normal karma modellerde farklı stratejiler elde edilmiştir[8]. Ancak bu stratejide kümelerin sayısının üst sınırının belirlenmesi için güçlü bir yöntem geliştirilememiştir. Verilerdeki heterojenliğe dayalı küme sayısının belirlenmesi için EM algoritmaları ile k-ortalamalar algoritmaları kullanılarak tek değişkenli normal karma dağılımlar kullanılabilir [9]. Veri için en iyi kümelenme modeli Akaike Bilgi Kriteri (AIC) [10] ve Bayesçi Bilgi Kriteri (BIC) [11] gibi iyi bilinen bilgi kriterleri kullanılarak optimizasyon ile belirlenir. Normal karma dağılımlar çoklu doğrusal regresyonların hata terimlerinin belirlenmesi ve gruplanmasında kullanılır [12]. Değişkenlerdeki veri parçalanmaya dayalı normal karma model (GMM) veri madenciliğinde yeni bir kümeleme yaklaşımıdır [13]. Normal karma dağılımların kullanıldığı kümeleme yöntemlerinde verinin değişkenleri üzerinde heterojenlik testi uygulanarak değişken ayıklama veya boyut indirgeme yapılır [14]. Veri madenciliğinde normal karma dağılımların modele dayalı kümelenmesinde en iyi modelin seçilebilmesi için bilgi kriterleri yaygın olarak kullanılır [15].

2. Materyal ve Metot

Bu çalışmada geliştirilen veri madenciliği yönteminin çalışma prensibini göstermek amacıyla prototip çalışma yapılmıştır. İki değişkenli ve heterojen yapıya sahip sentetik veri seti üretilip, üzerinde geliştirilen metot uygulanarak yöntem gösterilmiştir.

Normal dağılımların karmasına dayalı kümeleme verinin yapısına uygun en iyi küme sayısı ve yapısını ortaya çıkarmaya yarayan etkili bir kümeleme yöntemlerinden birisidir [3]. Bu çalışmada verilerin modele dayalı olarak kümelenmesi için normal karma dağılımlar kullanılmıştır. Literatürdeki çalışmalardan farklı olarak değişkenlerdeki heterojen yapılar tek değişkenli normal karma dağılımlar yardımıyla EM algoritmaları kullanılarak belirlenmiştir. Değişkenlerdeki heterojenliğe göre boyut indirgeme yapılmış ve genetik algoritmalar yardımıyla karma normal modeller oluşturulmuştur. Karma modeller arasından veriye en uygun kümelenme yapısı ve yeri optimizasyon yaparak belirlenmiştir.

2.1. Karma Dağılım Testi ile Heterojen Değişkenlerin Seçimi ve Karma Modellerin Oluşturulması

Sonlu karma dağılımlarda verilerin homojen veya heterojen yapıda olması oluşabilecek küme sayısı ve yerini belirlemektedir. Verideki herbir değişkenin sahip olduğu parçalanma sayısı değişkenin küme oluşumundaki rolünü belirlemektedir. Çok değişkenli veride heterojen veya parçalanmaya (normal olmayan) sahip değişkenler belirlenip parçalanmanın olmadığı homojen (normal) değişkenler kümelenmeye etkisi olmadığından elenir. Normal yapıda olmayan karma değişkenlerin parçalanma sayıları verideki kümelenme merkez sayılarını verir. $s=1,\ldots,p$ olmak üzere verideki her bir X_s değişkeninin heterojenliği incelendiğinde p-boyutlu veride X_s değişkeninin yapısına göre parçalanmalarının sayısı $k_S \geq 1$ olur. Homojen yapıdaki normal dağılan değişkende $k_S = 1$, heterojen yapıdaki karma dağılıma sahip değişken için $k_S > 1$ olur. Verideki her bir değişken için k_S değeri tek değişkenli normal karma dağılıma göre EM algoritmaları kullanılarak belirlenir. Tek değişkenli normal karma dağılım,

$$f(x;\theta) = \sum_{i=1}^{k} \pi_i f_i(x;\mu_i,\sigma_i) \quad (3)$$

şeklinde ifade edilir. Burada $f(x_j;\theta)$ normal karma dağılımın olasılık yoğunluk fonksiyonunu, k karma dağılımdaki bileşen sayısı ve π_i karma olasılık ağırlıklarını göstermektedir. EM algoritması iteratif bir algoritma olduğundan değişkenlerdeki parametreler tahmin edilir. EM algoritmasında z tamamlanmış verinin etiketleme vektörünü temsil edecek şekilde $\{X_1, X_2, \ldots, X_n, Z_1, Z_2, \ldots, Z_n\}$ verideki en çok olabilirlik (likelihood) fonksiyonu,

$$L = f(x_1, x_2, \ldots, x_n, z_1, z_2, \ldots, z_n; \pi, \psi) = \prod_{j=1}^{n} \prod_{i=1}^{k} \left[\pi_i f_i(x_j;\psi_i)\right]^{z_{ij}} \quad (5)$$

denklemiyle bulunur. Fonksiyonun logaritması alındığında,

$$lnL(\pi, \psi; x_1, x_2, \ldots, x_n, z_1, z_2, \ldots z_n) = \sum_{j=1}^{n} \sum_{i=1}^{k} z_{ij} ln\left[\pi_i f_i(x_j;\psi_i)\right] \quad (6)$$

Log-olabilirlik fonksiyonu elde edilir. EM algoritmasında iterasyon ile olabilirlik fonksiyonunu maksimum yapan etiketleme vektörü elde edilir. EM algoritmasının ilk adımı (E) beklenti ve ikinci (M) en büyük yapma adımıdır.

E (Beklenti) Adımı: Algoritmada z_{ij} etiket değerlerini tahmin etmek için grup koşullu beklenen değer,

$$\hat{z}_{ij} = E(z_{ij}|x_j;\pi_i,\psi_i) = \frac{\pi_i f_i(x_j;\psi_i)}{\sum_{i=1}^{k} f_i(x_j;\psi_i)} \quad (7)$$

eşitliği ile bulunur.

M (En büyük yapma) Adımı: Koşullu olasılıklar toplamı $\sum_{i=1}^{k} \pi_i = 1$ olduğundan ve olabilirlik fonksiyonunun en büyük değerini elde etmek için parametreler,

$$lnL(\pi, \psi; x_1, x_2, \ldots, x_n, z_1, z_2, \ldots z_n) = \sum_{j=1}^{n} \sum_{i=1}^{k} \hat{z}_{ij} ln\left[\pi_i f_i(x_j;\psi_i)\right] \quad (8)$$

denklemi ile elde edilir. EM algoritmasında parametrelerde değişim duruncaya kadar iterasyon devam eder ve hata minimum olduğunda etiket vektörü elde edilmiş olur.

Her bir heterojen değişkendeki parçalanmayı model tabanlı belirlemek amacıyla parametreleri tahmin edilen log-olabilirlik fonksiyonlarından AIC ve BIC değerleri hesaplanır. Log-olabilirlik fonksiyon değerinin maksimum, AIC ve BIC değerlerinin minimum olduğu modelde parçalanma sayısı optimum olarak bulunur.

Simülasyon ile üretilen normal dağılımlardan gelen sentetik veri seti iki değişkenli ($p = 2$) ve her bir değişkende 300 gözlem bulunmaktadır. Veride X_1 değişkeninde X_{11} , X_{12} ve X_2 değişkeninde X_{21}, X_{22} alt grupları olacak şekilde üretilmiştir. İki değişkenli veri setinde üç küme oluşacak şekilde

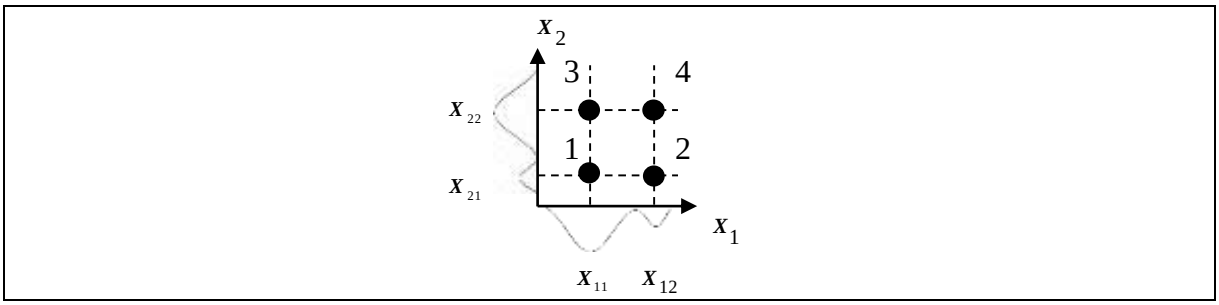
ortalamaları $\mu_{11} = 25$, $\mu_{12} = 60$ ve $\mu_{21} = 20$, $\mu_{22} = 70$, standart sapmaları sabit olmak üzere $\sigma = 5$ parametre değerlerinden elde edilmiştir.

Tablo 1. Değişken veri parçalamaya dayalı karma dağılımdan elde edilen Log-olabilirlik, AIC ve BIC değerleri parçalanma sayısı

Değişken	Küme sayısı	Log-Likelihood	AIC	BIC	Parçalanma sayısı
X_1	k=1	-1327.4	2658.9	2666.3	2
	k=2	**-1318.0**	**2646.0**	**2664.5**	
	k=3	-1317.1	2650.2	2679.8	
X_2	k=1	-1365.8	2735.6	2743.0	2
	k=2	**-1297.1**	**2604.2**	**2622.8**	
	k=3	-1297.1	2610.2	2639.9	

k-ortalamalar algoritması gibi bir ayrıştırma algoritması kullanılarak gözlem değerleri heterojen verideki alt gruplara atanır. X_1 ve X_2 değişkenlerindeki parçalanmalar karma dağılımlar kullanılarak sırasıyla $k_1 = 2$ ve $k_2 = 2$ olarak elde edilir. X_1 değişkeninde parçalanmalar X_{11} ve X_{12}, X_2 değişkeninde parçalanmalar ise X_{21} ve X_{22} olarak belirlenir. Değişkenlerdeki parçalanmalara bağlı olarak oluşabilecek en az küme sayısı $C_{\min} = \max\{k_s\} = \max\{2,2\} = 2$ ve en çok küme sayısı $C_{\max} = \prod_{s=1}^{p} k_S = \prod_{s=1}^{2} k_S = k_1 k_2 = 2.2 = 4$ olarak elde edilir.

Çok değişkenli verideki parçalanmalar ve bu parçalanmalara karşılık gelen küme merkezleri modellenmiş ve Şekil 1. de gösterilmiştir.



Şekil 1. Değişkenlerde meydana gelen parçalanmalar ve bunlara karşılık gelen küme merkezleri

Değişkenlerdeki parçalanma sayıları ve bu parçalanmalardan oluşan maksimum ve minimum küme sayılarına göre karma normal modeller elde edilir [14]. Değişkenlerdeki her bir parçalanmaya en az bir kümelenme merkezi karşılık gelir. Veride her bir değişkenin ikiye parçalanması durumunda kümelenme merkezleri için M_{Toplam} ile gösterilen normal dağılımların karma modelleriyle oluşturulabilecek toplam model sayısı,

$$M_{Toplam} = 2^{C_{Max}} - 1 \qquad (9)$$

eşitliğinden elde edilir. Değişkenlerdeki parçalanma sayılarından elde edilen en az ve en çok küme sayılarına göre 1 küme merkezli, 2 küme merkezli, 3 küme merkezli ve 4 küme merkezli (full model) normal karma modellerin sayısı $M_{Toplam} = 2^{C_{Max}} - 1 = 2^4 - 1 = 15$ olarak elde edilir. Böylece küme sayılarını veren bağıntı ile karma normal modellerin fonksiyonları arasında birebir ve örten bir ilişki meydana gelir.

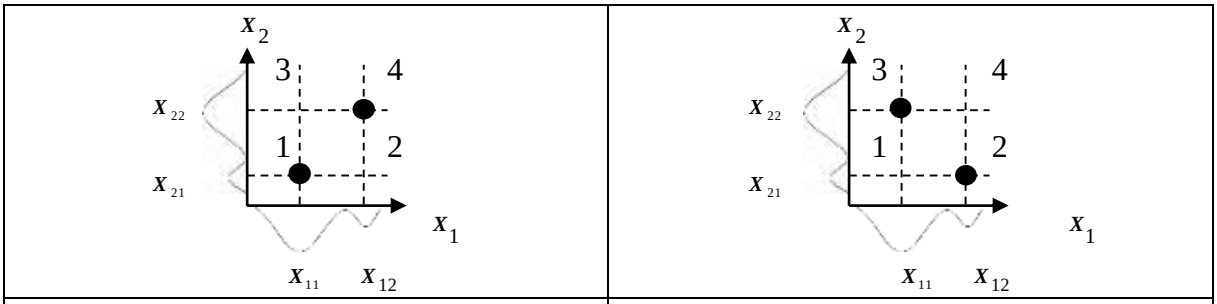
Geçerli modellerin sayısının belirlenmesi karma normal modellerin zor problemlerindendir. Karma modellerin yapısındaki satır ve sütun yapısına bakılarak bire-bir ve örten bir fonksiyonda her satır ve sütunda en az bir kümenin bulunması gerekmektedir. Kombinatorik bir problem olan geçerli model sayısı için $\binom{n}{r}$ ifadesi n elemanlı kümedeki r elemanlı kombinasyonları göstermektedir. Bir,

iki, üç ve dört kümelenme merkezli oluşturulabilecek normal dağılımların karma modellerinin sayısı sırasıyla $\binom{4}{1}=4$, $\binom{4}{2}=6$, $\binom{4}{3}=4$ ve $\binom{4}{4}=1$ olmak üzere toplam modellerin sayısı $\binom{4}{1}+\binom{4}{2}+\binom{4}{3}+\binom{4}{4}=4+6+4+1=15$ olarak hesaplamalı şekilde elde edilir. Karma modellerin küme merkezlerinin seçimine göre genetik kodları, küme sayısına göre karma model sayıları ve bu modeller arasından uygun karma modellerin sayıları Tablo 2'de gösterilmiştir.

Tablo 2. Modellerin temsili gösterimleri, kümelenmelere göre model sayıları, modellerin uygun olup olmadıkları ve uygun model sayıları

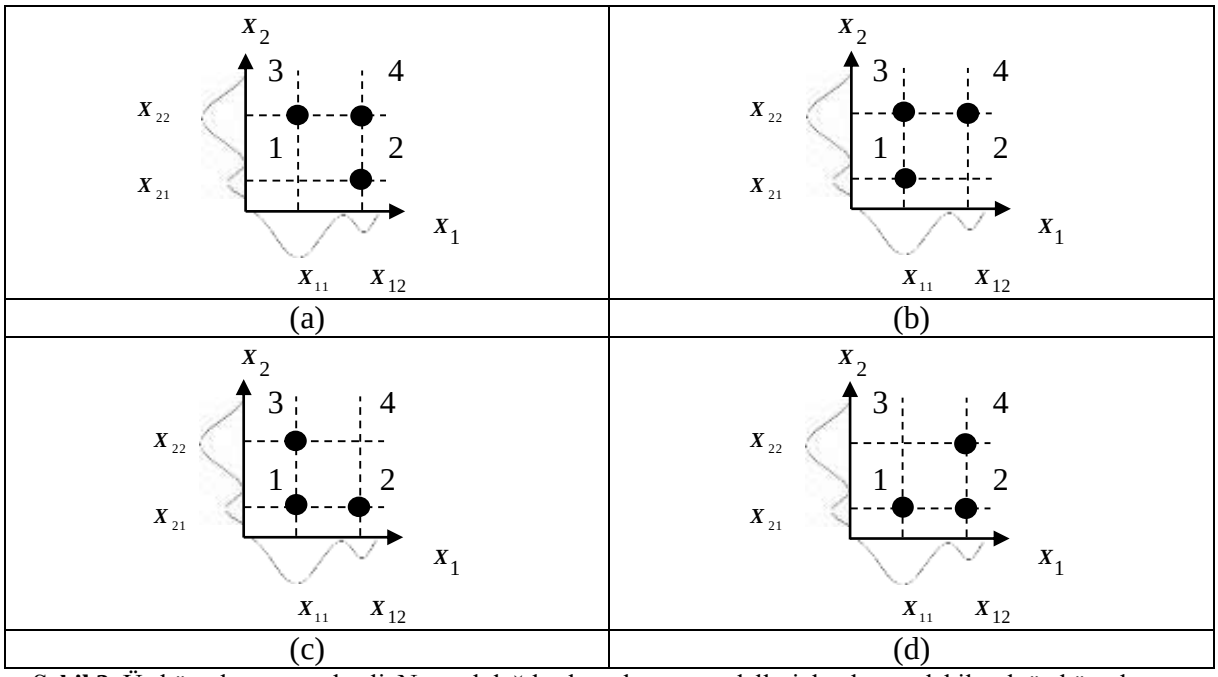
Küme Sayısı	**Modellerin Genetik Kodları**	**Model Sayısı**	**Uygun Model**	**Uygun Model Sayısı**
1 Merkezli Modeller	1 0 0 0	4	0	0
	0 1 0 0		0	
	0 0 1 0		0	
	0 0 0 1		0	
2 Merkezli Modeller	1 1 0 0	6	0	2
	1 0 1 0		0	
	1 0 0 1		1	
	0 1 1 0		1	
	0 1 0 1		0	
	0 0 1 1		0	
3 Merkezli Modeller	1 1 1 0	4	1	4
	1 1 0 1		1	
	1 0 1 1		1	
	0 1 1 1		1	
4 Merkezli Model	1 1 1 1	1	1	1
Toplam		15		7

İki değişkenli veride her bir değişkenin ikiye parçalanması durumunda veride normal dağılımların karma modelleriyle oluşturulabilecek ve $M_{Muhtemel}$ ile gösterilen muhtemel modeller Şekil 2 - Şekil 4'te gösterilmiştir.

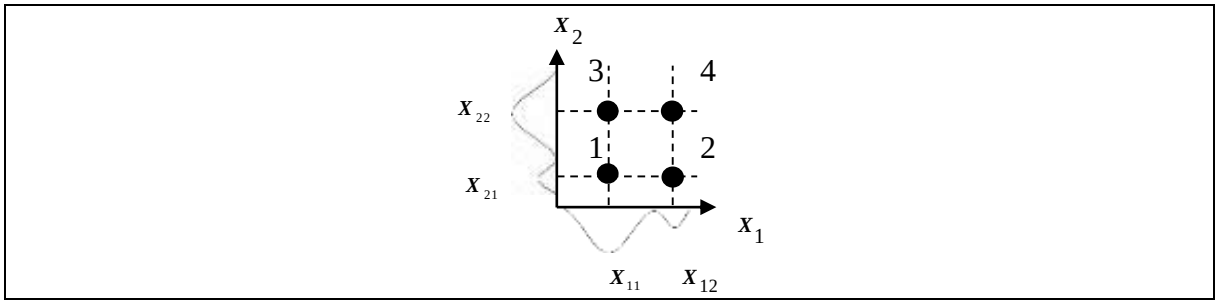


(a) (b)

Şekil 2. İki kümelenme merkezli. Normal dağılımların karma modelleriyle oluşturulabilecek iki kümelenme merkezli iki muhtemel modelin kümelenme merkezleri



Şekil 3. Üç kümelenme merkezli. Normal dağılımların karma modelleriyle oluşturulabilecek üç kümelenme merkezli dört muhtemel modelin kümelenme merkezleri



Şekil 4. Dört kümelenme merkezli. Normal dağılımların karma modelleriyle oluşturulabilecek dört kümelenme merkezli bir muhtemel modelin kümelenme merkezleri

İki değişkenli heterojen veride her bir değişkenin ikiye parçalanması durumunda veri seti için normal dağılımların karma modelleriyle oluşturulabilecek 15 karma model arasından uygun aday model sayısı 7 dir.

2.2. Normal Karma Modellerin Yapısı ve Özellikleri

Çok değişkenli normal karma modellerde değişken parçalama yöntemine dayalı elde edilen modelde küme merkezleri ve bu küme merkezlerine bağlı oluşabilecek normal karma modeller elde edilmiştir. Normal karma modeller değişkenlerdeki gözlemlerin atandığı küme merkezlerindeki verilerin ortalama vektörleri μ_i ve varyans-kovaryans matrisleri Σ_i den elde edilen merkezi eğilim ölçülerine göre kümeleme yapmaktadır.

Şekil 1'deki bir numaralı kümelenme merkezi X_1 değişkeninin X_{11} ve x_2 değişkeninin X_{21} parçalarının oluşturduğu kümelenme merkezidir. Bu merkezin ortalama vektörü μ_1 ve varyans-kovaryans matrisi $\boldsymbol{\Sigma}_1$ sırasıyla $\boldsymbol{\mu}_1 = \begin{bmatrix} \mu_{11} \\ \mu_{21} \end{bmatrix}$ ve $\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_{11}^2 & \rho_1\sigma_{11}\sigma_{21} \\ \rho_1\sigma_{21}\sigma_{11} & \sigma_{21}^2 \end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_1 = Corr(X_{11}, X_{21})$ olmak üzere x_{11} ve x_{21} parçalanmaları arasındaki korelasyon katsayısını

göstermekte ve $\rho_1 = \frac{\sigma_{1121}}{\sigma_{11}\sigma_{21}}$ olarak tanımlanmaktadır. Bir numaralı küme merkezi $N(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$ normal dağılımına sahiptir. İki numaralı kümelenme merkezi x_1 değişkeninin x_{12} ve x_2 değişkeninin x_{21} parçalarının oluşturduğu kümelenme merkezdir. Bu merkezin ortalama vektörü $\boldsymbol{\mu}_2$ ve varyans-kovaryans matrisi $\boldsymbol{\Sigma}_2$ sırasıyla $\boldsymbol{\mu}_2 = \begin{bmatrix}\mu_{12}\\ \mu_{21}\end{bmatrix}$ ve $\boldsymbol{\Sigma}_2 = \begin{bmatrix}\sigma_{12}^2 & \rho_2\sigma_{12}\sigma_{21}\\ \rho_2\sigma_{21}\sigma_{12} & \sigma_{21}^2\end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_2 = Corr(X_{12}, X_{21})$ olmak üzere x_{12} ve x_{21} parçalanmaları arasındaki korelasyon katsayısını göstermekte ve $\rho_2 = \frac{\sigma_{1221}}{\sigma_{12}\sigma_{21}}$ olarak tanımlanmaktadır. İki numaralı küme merkezi $N(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$ normal dağılımına sahiptir. Üç numaralı kümelenme merkezi x_1 değişkeninin x_{11} ve x_2 değişkeninin x_{22} parçalarının oluşturduğu kümelenme merkezdir. Bu merkezin ortalama vektörü $\boldsymbol{\mu}_3$ ve varyans-kovaryans matrisi $\boldsymbol{\Sigma}_3$ sırasıyla $\boldsymbol{\mu}_3 = \begin{bmatrix}\mu_{11}\\ \mu_{22}\end{bmatrix}$ ve $\boldsymbol{\Sigma}_3 = \begin{bmatrix}\sigma_{11}^2 & \rho_3\sigma_{11}\sigma_{22}\\ \rho_3\sigma_{22}\sigma_{11} & \sigma_{22}^2\end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_3 = Corr(X_{11}, X_{22})$ olmak üzere x_{11} ve x_{22} parçalanmaları arasındaki korelasyon katsayısını göstermekte ve $\rho_3 = \frac{\sigma_{1122}}{\sigma_{11}\sigma_{22}}$ olarak tanımlanmaktadır. Üç numaralı küme merkezi $N(\mathbf{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$ normal dağılımına sahiptir. Dört numaralı kümelenme merkezi x_1 değişkeninin x_{11} ve x_2 değişkeninin x_{22} parçalarının oluşturduğu kümelenme merkezdir. Bu merkezin ortalama vektörü $\boldsymbol{\mu}_4$ ve varyans-kovaryans matrisi $\boldsymbol{\Sigma}_4$ sırasıyla $\boldsymbol{\mu}_4 = \begin{bmatrix}\mu_{12}\\ \mu_{22}\end{bmatrix}$ ve $\boldsymbol{\Sigma}_4 = \begin{bmatrix}\sigma_{12}^2 & \rho_4\sigma_{12}\sigma_{22}\\ \rho_4\sigma_{22}\sigma_{12} & \sigma_{22}^2\end{bmatrix}$ şeklinde tanımlansın. Burada $\rho_4 = Corr(X_{12}, X_{22})$ olmak üzere x_{12} ve x_{22} parçalanmaları arasındaki korelasyon katsayısını göstermekte ve $\rho_4 = \frac{\sigma_{1222}}{\sigma_{12}\sigma_{22}}$ olarak tanımlanmaktadır. Dört numaralı küme merkezi $N(\mathbf{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$ normal dağılımına sahip tir.

Çok değişkenli sonlu karma modellerde modele dayalı kümeleme yapabilmek için modellerin normal dağılımların karmalarından geldiği varsayılır ve bu modeller normal dağılımlara gere belirlenirler.

Tek bileşenli normal karma modeller : $f(\boldsymbol{x};\boldsymbol{\theta}_1) = N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ i = 1, 2, 3, 4 olmak üzere dört tanedir. Bunların hiçbiri varsayımlara uymadığından tamamı uygun model değildir. İki kümelenme merkezli normal dağılımların karma modelleri: $i = 1,..,4$ için $0 < \pi_i < 1$ ve $\sum_{i=1}^{k}\pi_i = 1$ olmak üzere $f(\boldsymbol{x};\boldsymbol{\theta}_2) = \pi_i N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) + \pi_j N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ i, j = 1, 2, 3, 4 olmak üzere altı tanedir. İki bileşenli karma modellerden varsayıma uyan uygun modeller $f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_1 N(\boldsymbol{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_3 N(\boldsymbol{x};\boldsymbol{\mu}_3,\boldsymbol{\Sigma}_3)$ ve $f(\boldsymbol{x};\boldsymbol{\theta}) = \pi_2 N(\boldsymbol{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2) + \pi_4 N(\boldsymbol{x};\boldsymbol{\mu}_4,\boldsymbol{\Sigma}_4)$ dür. Üç bileşenli normal karma modeller: $i = 1,..,4$ için $0 < \pi_i < 1$ ve $\sum_{i=1}^{k}\pi_i = 1$ olmak üzere $f(\boldsymbol{x};\boldsymbol{\theta}_3) = \pi_i N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) + \pi_j N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) + \pi_k N(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ i, j, k = 1, 2, 3, 4 olmak üzere 4 adettir. Üç bileşenli normal karma modellerin tamamı varsayıma uyan modellerdir. Dört bileşenli normal karma model: $i = 1,..,4$ için $0 < \pi_i < 1$ ve $\sum_{i=1}^{k}\pi_i = 1$ olmak üzere $f(\boldsymbol{x};\boldsymbol{\theta}_4) = \pi_i N(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) + \pi_j N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) + \pi_k N(\mathbf{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) + \pi_t N(\mathbf{x};\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)$ i, j, k, t = 1, 2, 3, 4 olmak üzere 1 adettir ve

bu model varsayıma uyan uygun modeldir. Dolayısıyla oluşturulabilecek toplam model sayısı 15 ve uygun model sayısı 7 dir.

Karma modellerdeki parametreler veri setindeki gözlem değerlerinin bulunduğu parçalanmalardan elde edilen gözlemlere dayalı olarak belirlenmektedir.

3. Bulgular ve Tartışma

Çok değişkenli veride heterojenlik testi ile elde edilen değişkenlerin parçalanmalara ve bu parçalanmalara ait özellikler elde edilmiştir. Verideki parçalanmalara düşen gözlemler ve bu gözlemlerin ait olduğu kümeler ayrıştırma algoritmaları ile elde edilir. Elde edilen gözlem değerleri modelleri oluşturan normal dağılımların hesaplanması için kullanılır. Bu normal karma modeldeki parametreler verideki parçalanmaya düşen gözlemlerden elde edilir.

3.1. k-ortalamalar Algoritması ile Ayrıştırma ve Parametre Değerlerinin Elde Edilmesi

Çok değişkenli heterojen veri setindeki parçalanmalar ve bu parçalanmalara düşen gözlemler çeşitli kümeleme ve ayrıştırma algoritmaları ile belirlenebilir. Bunların en önemlilerinde birisi k-ortalamalar algoritmasıdır. k-ortalamalar algoritması Gauss karma modellerinden farklı olarak oluşan kümelere farklı sayıda gözlem atamaktadır. Bu çalışmada değişkenlerde oluşacak küme sayısında farklı sayıda gözlem bulunması her bir kümenin olasılık ağırlıklarını etkilediğinden k-ortalamalar algoritması tercih edilmiştir. Heterojen $X_1 \, ve \, X_2$ değişkenlerindeki alt gruplar sırasıyla $X_{11}, X_{12} \; ve \; X_{21}, X_{22}$ olarak belirlenmiştir. k-ortalamalar algoritmasının çalışma prensibine göre daha önceden belirlenen parçalanma sayış kadar küme merkezi seçilir ve algoritma kullandığı uzaklık metriğine göre en uygun gözlemleri yakın bulunduğu kümeye atar ve iterasyon ile grup üyelikleri değişmeyinceye kadar devam eder. Seçilen giriş küme merkezi değeri ile gözlemler arasındaki uzaklık

$$argmin_s \sum_{i=1}^{k} \sum_{j \in S_i} \left\| x_j - \mu_i \right\|^2 \tag{10}$$

denklemi ile elde edilir. Belirlenen kümeler arası uzaklığın maksimum (heterojenlik) aynı zamanda parçalanmalara atanan gözlemler arası uzaklığın minimum (homojenlik) olduğu kümeleme en iyi kümeleme olarak kabul edilir. Çok değişkenli veride değişkenlerin parçalanmalarına atanan gözlemlerin etiketleri ve sayıları Tablo 3'de verilmiştir.

Tablo 3. Çok değişkenli veri setindeki değişkenlerin parçalanmalarına düşen verilerin etiketleri ve parçalanmalara atanan gözlem sayıları

Değişkenler	X_1 **değişkenindeki parçalanma ve atanan gözlem sayıları**		X_2 **değişkenindeki parçalanma ve atanan gözlem sayıları**	
Parçalanma ve Etiketleri	n_{11} $(\hat{n}_{11})$	n_{12} $(\hat{n}_{12})$	n_{21} $(\hat{n}_{21})$	n_{22} $(\hat{n}_{22})$
Gözlem sayıları	89	211	74	226
Toplam gözlem sayısı	300		300	

Çok değişkenli heterojen veri setindeki parçalanmalardan elde edilen küme merkezleri ve bu merkezlere atanan gözlemlerden elde edilen olasılık ağırlıkları, ortalama vektörleri ve varyans-kovaryans matrisleri popülasyondan hesaplanmıştır. Elde edilen parametre değerleri Tablo 4 – Tablo 6'da gösterilmiştir.

Tablo 4. Oluşturulan iki kümelenme merkezli karma modellerindeki karma oranları, ortalama vektörleri ve varyans-kovaryans matrislerinin parametre tahminleri

Modeller	$\hat{\boldsymbol{\pi}}$	$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Sigma}}$	$\hat{\rho}$
İki bileşenli birinci uygun model (Şekil 2(a))	$\pi_1 = 0{,}5416$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 28{,}2664 \\ 59{,}2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 109{,}277 & 23{,}113 \\ 23{,}113 & 67{,}624 \end{bmatrix}$	$\rho_1 = 0{,}269$
	$\pi_4 = 0{,}4583$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 62{,}2378 \\ 17{,}7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 134{,}152 & -13{,}605 \\ -13{,}605 & 146{,}631 \end{bmatrix}$	$\rho_4 = -0{,}097$
İki kümelenme merkezli ikinci uygun model (Şekil 2(b))	$\pi_2 = 0{,}5183$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 62{,}2378 \\ 59{,}2654 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 134{,}152 & 18{,}965 \\ 18{,}965 & 67{,}624 \end{bmatrix}$	$\rho_2 = 0{,}199$
	$\pi_3 = 0{,}4816$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 28{,}2664 \\ 17{,}7130 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 109{,}277 & 1{,}616 \\ 1{,}616 & 146{,}631 \end{bmatrix}$	$\rho_3 = 0{,}013$

Tablo 5. Oluşturulan üç kümelenme merkezli karma modellerindeki karma oranları, ortalama vektörleri ve varyans-kovaryans matrislerinin parametre tahminleri

Modeller	$\hat{\boldsymbol{\pi}}$	$\hat{\boldsymbol{\mu}}$	$\hat{\boldsymbol{\Sigma}}$	$\hat{\rho}$
Üç bileşenli üçüncü uygun model (Şekil 3(a))	$\pi_2 = 0{,}2748$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23{,}5331 \\ 20{,}1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92{,}963 & 18{,}051 \\ 18{,}051 & 136{,}873 \end{bmatrix}$	$\rho_2 = 0{,}160$
	$\pi_3 = 0{,}3037$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67{,}3698 \\ 68{,}5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108{,}335 & 15{,}393 \\ 15{,}393 & 80{,}099 \end{bmatrix}$	$\rho_3 = 0{,}0165$
	$\pi_4 = 0{,}4214$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23{,}5331 \\ 68{,}5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92{,}963 & 6{,}761 \\ 6{,}761 & 80{,}099 \end{bmatrix}$	$\rho_4 = 0{,}078$
Üç bileşenli dördüncü uygun model (Şekil 3(b))	$\pi_1 = 0{,}1781$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67{,}3698 \\ 20{,}1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108{,}335 & -2{,}386 \\ -2{,}386 & 136{,}873 \end{bmatrix}$	$\rho_1 = -0{,}20$
	$\pi_3 = 0{,}3442$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67{,}3698 \\ 68{,}5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108{,}335 & 15{,}393 \\ 15{,}393 & 80{,}099 \end{bmatrix}$	$\rho_3 = 0{,}0165$
	$\pi_4 = 0{,}4775$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23{,}5331 \\ 68{,}5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92{,}963 & 6{,}761 \\ 6{,}761 & 80{,}099 \end{bmatrix}$	$\rho_4 = 0{,}078$
Üç bileşenli beşinci uygun model (Şekil 3(c))	$\pi_1 = 0{,}2136$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67{,}3698 \\ 20{,}1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108{,}335 & -2{,}386 \\ -2{,}386 & 136{,}873 \end{bmatrix}$	$\rho_1 = -0{,}20$
	$\pi_2 = 0{,}3735$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23{,}5331 \\ 20{,}1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92{,}963 & 18{,}051 \\ 18{,}051 & 136{,}873 \end{bmatrix}$	$\rho_2 = 0{,}160$
	$\pi_3 = 0{,}4128$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 67{,}3698 \\ 68{,}5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 108{,}335 & 15{,}393 \\ 15{,}393 & 80{,}099 \end{bmatrix}$	$\rho_3 = 0{,}0165$
Üç bileşenli altıncı uygun model (Şekil 3(d))	$\pi_1 = 0{,}1841$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 67{,}3698 \\ 20{,}1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 108{,}335 & -2{,}386 \\ -2{,}386 & 136{,}873 \end{bmatrix}$	$\rho_1 = -0{,}20$
	$\pi_2 = 0{,}3220$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 23{,}5331 \\ 20{,}1245 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 92{,}963 & 18{,}051 \\ 18{,}051 & 136{,}873 \end{bmatrix}$	$\rho_2 = 0{,}160$
	$\pi_4 = 0{,}4937$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 23{,}5331 \\ 68{,}5173 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 92{,}963 & 6{,}761 \\ 6{,}761 & 80{,}099 \end{bmatrix}$	$\rho_4 = 0{,}078$

Tablo 6. Oluşturulan dört kümelenme merkezli karma modellerindeki karma oranları, ortalama vektörleri ve varyans-kovaryans matrislerinin parametre tahminleri

Model	$\hat{\pi}$	$\hat{\mu}$	$\hat{\Sigma}$	$\hat{\rho}$
Dört bileşenli yedinci uygun model (Şekil 4)	$\pi_1 = 0{,}2158$	$\boldsymbol{\mu}_1 = \begin{bmatrix} 26{,}6768 \\ 73{,}0997 \end{bmatrix}$	$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 114{,}677 & -14{,}283 \\ -14{,}283 & 100{,}252 \end{bmatrix}$	$\rho_1 = -0{,}133$
	$\pi_2 = 0{,}2441$	$\boldsymbol{\mu}_2 = \begin{bmatrix} 71{,}2125 \\ 73{,}0997 \end{bmatrix}$	$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 132{,}427 & 1{,}616 \\ 1{,}616 & 100{,}252 \end{bmatrix}$	$\rho_2 = 0{,}014$
	$\pi_3 = 0{,}2558$	$\boldsymbol{\mu}_3 = \begin{bmatrix} 26{,}6768 \\ 24{,}8489 \end{bmatrix}$	$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 114{,}677 & 22{,}735 \\ 22{,}735 & 104{,}498 \end{bmatrix}$	$\rho_3 = 0{,}208$
	$\pi_4 = 0{,}22841$	$\boldsymbol{\mu}_4 = \begin{bmatrix} 71{,}2125 \\ 24{,}8489 \end{bmatrix}$	$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 132{,}427 & 25{,}461 \\ 25{,}461 & 104{,}498 \end{bmatrix}$	$\rho_4 = 0{,}216$

3.2. Heterojen veride en iyi kümelenme yapısını veren normal dağılımların karma modelleri için bilgi kriterlerinin hesaplanması

Heterojen veride en iyi kümelenme yapısını normal dağılımların karma modellerini kullanarak modele dayalı belirlemek amacıyla her uygun model için log-olabilirlik fonksiyonu hesaplanır. Normal dağılımların karma modeli için l olabilirlik fonksiyonu,

$$L(\pi, \mu, \Sigma) = \prod_{j=1}^{n} f_i(x_j; \theta_i) = \prod_{j=1}^{n} \sum_{i=1}^{k} \pi_i \, f_i(x_j; \mu_i, \Sigma_i) \tag{11}$$

olarak tanımlanır. İşlem kolaylığı için (11)'deki eşitlikte her iki tarafın logaritması alındığında log-olabilirlik fonksiyonu,

$$log\ L(\pi, \mu, \Sigma) = \sum_{j=1}^{n} log\left(\sum_{i=1}^{k} \pi_i \, f_i(x_j; \mu_i, \Sigma_i)\right) \tag{12}$$

olarak elde edilir. Normal dağılımların karma modellerinde her uygun model için log-likelihood fonksiyon değeri sırasıyla $\hat{\pi}_i$, $\hat{\mu}_i$ ve $\hat{\Sigma}_i$ tahmin edilmiş değerleri kullanılarak hesaplanır. Heterojen veride en iyi kümelenme yapısını normal dağılımların karma modellerini kullanarak modele dayalı belirlemek amacıyla her uygun model için bilgi kriteri AIC ve BIC değerleri,

$$AIC = -2logL(\hat{\pi}, \hat{\mu}, \hat{\Sigma}) + 2d \tag{13}$$

$$BIC = -2logL(\hat{\pi}, \hat{\mu}, \hat{\Sigma}) + dlog(n) \tag{14}$$

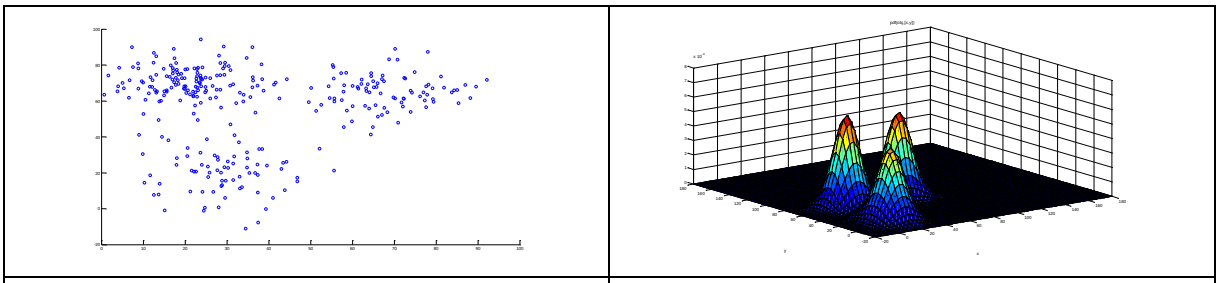
Eşitlikleri ile hesaplanır. Normal dağılımların karma modellerinde her uygun model için AIC ve BIC değeleri hesaplanırken $\hat{\pi}_i$, $\hat{\mu}_i$ ve $\hat{\Sigma}_i$tahmin edilmiş değerleri ve normal dağılımların karma modelindeki parametre sayısı d ve gözlem sayısı n kullanılır.

Veri seti için normal dağılımların karma modellerinde log-olabilirlik fonksiyon, AIC ve BIC değerleri heterojen veride kümelenme yapısını ortaya çıkarmak ve en iyi modeli belirlemek amacıyla hesaplanır. Çok değişkenli karma modeller için elde edilen log-olabilirlik, AIC ve BIC değerleri ve parametre sayıları Tablo 7'deverilmiştir.

Çok değişkenli heterojen veri seti için normal dağılımların karma modelleri kullanılarak bilgi kriterlerine göre belirlenen en iyi modelin üç kümelenme merkezli beşinci uygun model (Şekil 3(b)) olduğu belirlenmiştir. Simülasyon ile elde edilen sentetik veri seti için belirlenen en iyi modelin verdiği kümelenme yapısı Şekil 5'de gösterilmiştir.

Tablo 7. Çok değişkenli veri setinde uygun normal karma modeller için log-olabilirlik fonsiyonu (Log-L), Akaike bilgi kriteri (AIC), Bayesçi bilgi kriteri (BIC) değerleri ve parametre sayıları

Uygun Modeller	**Log-L**	**AIC**	**BIC**	d
İki bileşenli birinci uygun model	-3675.3	7372.7	7350.7	11
İki bileşenli ikinci uygun model	-3460.4	6942.8	6920.8	11
Üç bileşenli üçüncü uygun model	-3146.3	6326.6	6292.6	17
Üç bileşenli dördüncü uygun model	-3767	7568	7534	17
Üç bileşenli beşinci uygun model	**-2641.8**	**5317.7**	**5283.7**	**17**
Üç bileşenli altıncı uygun model	-3054.3	6142.5	6108.5	17
Dört bileşenli yedinci uygun model	-2705.2	5456.4	5410.4	23



(a) (b)

Şekil 5. Simülasyon ile elde edilen sentetik veri setinde uygun modeller için oluşturulan normal dağılımların karma modelleri kullanılarak bilgi kriterlerine göre belirlenen en iyi modelin verdiği kümelenme yapısının (a) Saçılım grafiği, (b) Yüzey grafiği

4. Sonuç ve Öneriler

Bu çalışmada çok değişkenli veride heterojenlik testi ile veri setindeki değişkenlerin yapısının belirlenmesine dayalı yeni bir kümeleme yöntemi önerilmiş ve özellikleri anlatılmıştır. Heterojenlik testi ile ortaya çıkan değişkenlerin verideki kümelemeyi belirlediği ortaya konulmuş ve oluşacak küme merkezlerinin yeri, sayısı ve yapısı hakkında bilgiler elde edilmiştir. Oluşan küme merkezlerine bağlı olarak k-ortalamalar algoritması ile değişkenlerdeki parçalanmalara atanan gözlemler belirlenmiştir. Gauss karma modellerin sayısı ve bunların arasından modele dayalı kümelenmeye uyan karma modeller Genetik Algoritmalar ile belirlenip genetik kodlara dönüştürülmüştür. Gauss karma modellerin yapısı ve her bir uygun modelin özellikleri ortaya çıkarılmıştır. Karma modellerdeki parametreler tahmin edilip her bir model için bilgi kriterleri elde edilmiştir.

Çok değişkenli veride her bir değişkendeki parçalanmaya dayalı karma model kümelemede iki değişkendeki parçalanmalar belirlenmiş ve bu parçalanmalara düşen gözlemler ve oluşturduğu kümeler elde edilmiştir. Genetik Algoritmalar ile belirlenen 7 uygun karma model arasından üç bileşenli Gauss karma modelin veri yapısına uyan en iyi model olduğu bilgi kriterleri yardımıyla elde edilmiştir. Genetik algoritmalarla belirlenen karma modelin hesaplamalar ile elde edilen üç bileşenli modelin grafikleri elde edilmiş ve modeldekine uyduğu görülmüştür.

Sezgisel olarak verideki değişken sayısı ne olursa olsun verideki heterojen değişkenlerdeki parçalanmaların verideki kümelenme sayısını ve yapısını etkilediği ya da belirlediği söylenebilir. Bu çalışmadaki önerilen kümelenme yöntemi yukarıdaki varsayımdan hareketle büyük veride modele dayalı kümelenme için yeni bir genetik algoritmayla birlikte veri madenciliğinde kümelenme yöntemi olarak geliştirilebilir. Ayrıca kullanılan bu yeni modele dayalı karma kümeleme yöntemi farklı dağılımlardan gelen karma modellerin kümelenmesi için kullanılabilir.

Yazarların Katkısı

Çalışmada tüm katkı yazara aittir.

Çıkar Çatışması Beyanı

Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

Araştırma ve Yayın Etiği Beyanı

Yapılan çalışmada araştırma ve yayın etiğine uyulmuştur.

Kaynaklar

[1] McLachlan G.J., Peel D. 2000. Finite Mixture Models. Wiley, New York.
[2] Fraley C., Raftery A.E. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97 (458): 611-631.
[3] McLachlan G.J., Chang S.U. 2004. Mixture Modelling for Cluster Analysis. Statistical Methods in Medical Research, 13: 347-361.
[4] Fraley C., Raftery A.E. 1998. How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. The Computer Journal, 41 (8): 578-588.
[5] Soffritti G. 2003. Identifying multiple cluster structures in a data matrix. Communications in Statistics, Simulation & Computation, 32 (4): 1151-1181.
[6] Galimberti G., Soffritti G. 2007. Model-based methods to identify multiple cluster structures in a data set. Computational Statistics and Data Analysis, 52: 520-536.
[7] Seo B., Kim D. 2012. Root selection in normal mixture models. Computational Statistics and Data Analysis. 56: 2454-2470.
[8] Servi T., Erol H. 2007. On Total Number of Candidate Component Cluster Centers and Total Number of Candidate Mixture Models in Model Based Clustering. Selçuk Journal of Applied Mathematics 8 (2): 57-69.
[9] Gogebakan M., Erol H. 2018. A New Semi-supervised Classification Method Based on Mixture Model Clustering for Classification of Multispectral Data. Journal of the Indian Society of Remote Sensing, 46 (8): 1323-1331.
[10] Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19 (6): 716-723.
[11] Schwarz G. 1978. Estimating the dimension of a model, Ann. Statist., 6 (2): 461-464.
[12] Ranciati S., Galimberti G., Soffritti G. 2019. Bayesian variable selection in linear regression models with non-normal errors. Statistical Methods & Applications, 28 (2): 323-358.
[13] Erol H., Gogebakan M., Erol R. 2017. Grid Structures and Orientations Of Clusters Using Discretization Of Variables In Big Data. Proceedings of International Conference on Engineering, Technology, and Applied Science ICETA 2017, ISSN 2411-9318: 16-31.
[14] Gogebakan M., Erol H. 2019. Mixture Model Clustering Using Variable Data Segmentation and Model Selection: A Case Study of Genetic Algorithm. Mathematics Letters, 5 (2): 23-32.
[15] Akogul S., Erisoglu M. 2017. An Approach for Determining the Number of Clusters in a Model Based Cluster Analysis. Entropy, 19 (9): 452.