İKİ PARAMETRELİ RAYLEIGH DAĞILIMLARININ SONLU KARMALARINDA PARAMETRE TAHMİNİ

Hayrinisa DEMİRCİ BİÇER [1] *Cenker BİÇER* [2]

Öz

Heterojen yapıda bir popülasyondan elde edilmiş verilerin istatistiksel analizinde oldukça kullanışlı modeller olan sonlu karma dağılımlar için parametre tahmin problemi istatistikte oldukça önemli bir problemdir. Bu çalışma, iki parametreli Rayleigh dağılımlarının sonlu karmaları için parametre tahmin problemini ele almaktadır. Bu kapsamda, iki parametreli Rayleigh dağılımlarının sonlu karmalarında mevcut bilinmeyen parametreler için en çok olabilirlik tahmin edicileri E-M algoritmasına göre elde edilmektedir. Bununla birlikte çalışmada, elde edilen en çok olabilirlik tahmin edicilerinin karma dağılımın bilinmeyen parametrelerini tahmin etmedeki performansını ortaya koymak için, karma oran parametresinin ve karma bileşen dağılımlarındaki parametrelerin farklı değerlerini göz önünde bulunduran ve tahmin edicilere ait hata kareler ortalamalarını, yanlılık miktarlarını ve standart sapmalarını ortaya koyan simülasyon çalışması sonuçlarına yer verilmektedir. Buna ek olarak, açıklayıcı amaçlar için gerçek bir veri seti kullanılarak yapılan bir de örneğe yer verilmektedir.

Anahtar Kelimeler: *İki Parametreli Rayleigh Dağılımı, Karma Dağılımlar, En Çok Olabilirlik Yöntemi, E-M Algoritması*

JEL Sınıflandırması: *C690, C020, C610*

PARAMETER ESTIMATION IN FINITE MIXTURES OF TWO-PARAMETER RAYLEIGH DISTRIBUTIONS

Abstract

The problem of parameter estimation for finite mixture distributions, which are very useful models for the statistical analysis of data obtained from a heterogeneous population, is quite important problem in statistics. This paper focuses on the parameter estimation problem for finite mixtures of the two-parameter Rayleigh distributions. In this context, the maximum likelihood estimators for the unknown parameters in the finite mixtures of the two-parameter Rayleigh distributions are obtained according to the E-M algorithm. Furthermore, in order to demonstrate the estimation performance of the obtained maximum likelihood estimators, a simulation study which gives the mean square error, bias and standard deviations of the estimators, is carried out by considering for different values of the mixing ratios and parameters of the component distributions. Also, an actual data set is analysed for illustrative purposes.

Keywords: *Two-Parameter Rayleigh Distribution, Mixture Distributions, Maximum Likelihood Method, E-M Algorithm*

JEL Classification: *C690, C020, C610*

[1] *Yrd. Doç. Dr., Kırıkkale Üniversitesi, Fen Ed. Fak., İstatistik Bölümü, hdbicer@hotmail.com*

[2] *Yrd. Doç. Dr., Kırıkkale Üniversitesi, Fen Ed. Fak., İstatistik Bölümü, cbicer@kku.edu.tr*

DOI: 10.18092/ulikidince.352176

Makalenin Geliş Tarihi (Recieved Date): 14-11-2017

Yayına Kabul Tarihi (Acceptance Date): 15-01-2018

1. Giriş

Sonlu karma dağılım modelleri heterojen yapıdaki bir popülasyondan elde edilmiş verilerin istatistiksel analizi için standart parametrik dağılım ailelerine nazaran oldukça yararlı sonuçlar sunan dağılım modelleri olup yaygın bir kullanım alanına sahiptir. X bir rasgele değişken olmak üzere; X rasgele değişkeninin olasılık yoğunluk fonksiyonu;

$$f(x) = \sum_{j=1}^{k} p_j f_j(x, \theta_j) \quad (1)$$

biçiminde ise X rasgele değişkenine k bileşenli f karma dağılımına sahiptir denir. Burada k bileşen sayısını, $f_j(x, \theta_j)$ j. bileşenin olasılık (yoğunluk) fonksiyonunu ve p_j, $\sum_{j=1}^{k} p_j = 1$ ve $p_j \geq 0$ olmak üzere, karma dağılımın j. bileşeni için karma oranını göstermektedir (Açıkgöz, 2007). Sonlu karma dağılım modellerinde temelde üç tip parametre söz konusudur. Bunlardan ilki bileşen sayısı (k), ikincisi bileşenlere ait p_j karma oranları ve üçüncü tip parametreler ise bileşen dağılımlarına ait parametrelerdir.

Bu çalışmanın amacı, k bileşen sayısının bilindiği varsayımı altında, iki parametreli Rayleigh dağılımlarının k bileşenli sonlu karma dağılımlarının bilinmeyen parametrelerini tahmin etmektir. İki parametreli Rayleigh dağılımı fen, müdendislik ve sağlık alanlarından birçok problemin modellenmesinde başarı ile kullanılan çarpık bir dağılımdır. İki parametreli Rayleigh dağılımının olasılık yoğunluk fonksiyonu

$$f(x, \lambda, \mu) = 2\lambda(x - \mu)e^{-\lambda(x-\mu)^2}, x > \mu \quad (2)$$

ve dağılım fonksiyonu

$$F(x, \lambda, \mu) = 1 - e^{-\lambda(x-\mu)^2} \quad (3)$$

biçimindedir. Burada,$\lambda > 0$ dağılımın ölçek parametresi, $\mu \in \mathbb{R}$ ise dağılımın konum parametresidir (Dey vd., 2014). Eşitlik (1) ile verilen karma dağılımın olasılık yoğunluk fonksiyonunun göz önüne alınmasıyla k bileşenli iki parametreli Rayleigh karma dağılımın olasılık yoğunluk fonksiyonu

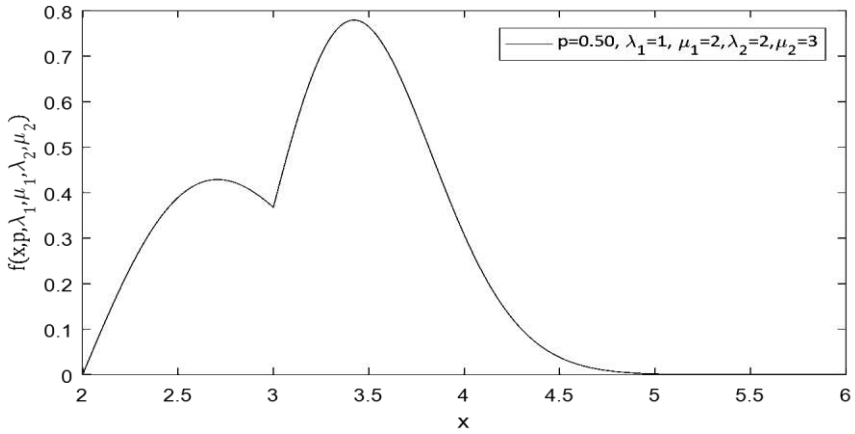
$$f(x_i, \Phi) = \sum_{j=1}^{k} p_j f_j(x_i, \theta_j) = \sum_{j=1}^{k} p_j 2\lambda_j (x_i - \mu_j) e^{-\lambda_j (x_i - \mu_j)^2} I_{x_i > \mu_j}, \quad x_i > \exists \mu_j, \mu_j \in \mathbb{R}, \lambda_j > 0 \quad (4)$$

biçiminde kolayca yazılabilir. Burada λ_j j. bileşenin ölçek parametresi, μ_j j. bileşenin konum parametresi, $I_{x_i > \mu_j}$

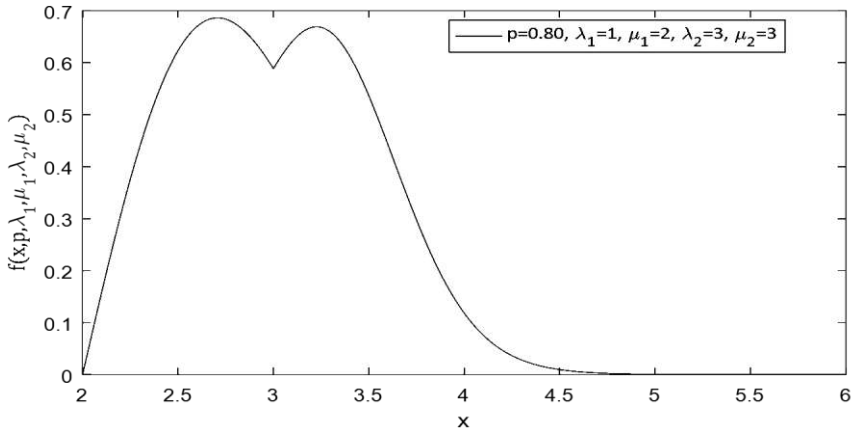
$$I_{x_i > \mu_j} = \begin{cases} 1 & x_i > \mu_j \\ 0 & x_i \leq \mu_j \end{cases} \quad (5)$$

biçiminde tanımlı indikatör fonksiyonudur. Burada p_j j. bileşen için karma oran parametresi olup, bu dağılımda toplam $(3k)$ tane bilinmeyen parametre söz konusu olmaktadır. (4) eşitliği ile verilen olasılık yoğunluk fonksiyonunun biçiminin anlaşılabilmesi için Şekil 1. ve Şekil 2. de farklı parametre değerleri için iki bileşenli iki parametreli Rayleigh karma dağılımın olasılık yoğunluk fonksiyonunun grafikleri verilmiştir.

Şekil 1: **İki Bileşenli İki Parametreli Rayleigh Karma Dağılımın Olasılık Yoğunluk Fonksiyonunun Grafiği** $p = 0.50,\ \lambda_1 = 1,\ \lambda_2 = 2,\ \mu_1 = 2,\ \mu_2 = 3$



Şekil 2: **İki Bileşenli İki Parametreli Rayleigh Karma Dağılımın Olasılık Yoğunluk Fonksiyonunun Grafiği** $p = 0.80,\ \lambda_1 = 1,\ \mu_1 = 2,\ \lambda_2 = 3,\ \mu_2 = 3$



Bu çalışmanın diğer bölümleri ise şu şekilde düzenlenmiştir: Karma dağılımlarda parametre tahmini üzerine yapılan çalışmalara ait literatür özeti ikinci bölümde verilmektedir. Üçüncü bölümde, k bileşenli iki parametreli Rayleigh karma dağılımının bilinmeyen parametrelerinin en çok olabilirlik tahmin edicileri E-M algoritması kullanılarak elde edilmektedir. Dördüncü bölümde, üçüncü bölümde elde edilen en çok olabilirlik tahmin edicilerinin parametreleri tahmin etmedeki etkinliğini ortaya koymak için yapılan simülasyon çalışmalarına ve gerçek bir veri seti kullanılarak yapılan bir uygulamaya yer verilmektedir. Çalışmanın beşinci bölümünde ise elde edilen sonuçlar tartışılmaktadır.

2. Literatür Özeti

Literatürde karma dağılımlar üzerine yapılmış geniş bir çalışma yelpazesi vardır. Karma dağılımlarda parametre tahmin problemi üzerine bilinen ilk çalışma; Pearson (1894) tarafından yapılan, iki bileşenli karma normal dağılımın parametreleri için momentler tahmin edicilerinin elde edildiği çalışmadır. Sonlu Karma dağılımlar için olabilirlik fonksiyonun karmaşıklığından ve en çok olabilirlik tahmin edicilerinin analitik olarak elde edilememesinden dolayı literatürde karşılaşılan ilk çalışmalar çoğunlukla momentler yöntemine dayalı tahmin edicilerin elde edilmesi üzerinedir. Üstel, Gamma, Weibull, Binom, Negatif Binom, Geometrik ve Poisson, dağılımlarının karmalarına ait parametrelerin momentler yöntemine dayalı tahmin edicileri Summ ve Oommen (1995) tarafından yapılan çalışma ile verilmiştir.

Hesaplama araçlarındaki gelişmelerin, sayısal yöntemleri daha kullanılabilir kılmasının bir sonucu olarak karma dağılımların parametrelerinin tahmininde daha etkin sonuçlar sunan en çok olabilirlik yöntemi daha çok kullanım olanağı bulmaktadır. Günümüzde en çok olabilirlik yönteminin tercih edilmesinin diğer bir önemli sebebide, Dempster vd. (1977), tarafından E-M (Expectation–Maximization) algoritmasının tanıtılmasıdır. Literatürde, E–M algoritmasının tanıtılmasından sonra, sonlu karma dağılımlar için parametre tahmin problemini göz önüne alan ve parametrelerin en çok olabilirlik tahmin edicilerinin elde edildiği çok sayıda çalışma mevcuttur. Bu çalışmaların bazıları şunlardır: Dick ve Bowden (1973) tarafından iki bileşenli karma normal dağılımın parametreleri için en çok olabilirlik tahminlerini Newton-Raphson yöntemine göre elde etmiştir. Leytham (1984) aynı problemi yaptığı çalışma ile ele almış, iki bileşenli karma normal dağılımın parametreleri için E–M algoritmasına dayalı en çok olabilirlik tahmin edicileri elde edilerek tahmin edicilerin küçük örneklem özelliklerini Monte–Carlo simülasyonuna göre araştırmıştır. Liu vd. (2006) tarafından internet trafiği Gamma dağılımlarının sonlu karmaları kullanılarak modellenmiş ve Gamma dağılımlarının sonlu karmalarına ait parametreler araştırmacılar tarafından E-M algoritması kullanılarak en çok olabilirlik yöntemine göre tahmin edilmiştir. Wang ve Wang (2014) üstel dağılımın belirli parametrik fonksiyonlara göre karmalarını göz önüne almışlar ve E–M algoritmasına dayalı en çok olabilirlik tahmin edicileri elde etmişlerdir. Afify (2011) bir parametreli Rayleigh dağılımının karmaları için 1. tip sansürleme olması durumunda en çok olabilirlik tahmin edicilerini elde etmiştir. Elmahdy ve Aboutahoun (2013) üç parametreli Weibul dağılımlarının sonlu karmaları için E-M algoritmasına dayalı en çok olabilirlik tahmin edicilerini yaptıkları çalışma ile vermişlerdir.

İleri düzey okuyucular sonlu karma dagılım modelleri ve E-M algoritması hakkında daha kapsamlı bilgi için (McLachlan ve Krishnan, 1997) ve (Everitt ve Hand, 1981), kaynaklarına başvurulabilir.

3. Yöntem

Bu bölümde, k bileşenli iki parametreli Rayleigh karma dağılımının bilinmeyen parametrelerinin tahminleri en çok olabilirlik yöntemine göre E-M algoritması kullanılarak elde edilmektedir.

E-M algoritması belirli bir örnekleme dayalı olarak bir karma dağılımdaki parametrelerin en çok olabilirlik tahminlerini parametrelerin bir Φ^0 başlangıç tahmin değerinden başlayarak yakınsama sağlanıncaya kadar Φ'yi tekrar tekrar güncelleyerek bulan yinelemeli bir tahmin yöntemdir. Karma dağılımlar için E-M algoritmasındaki her yineleme,

E- aşaması: $Q\left(\Phi, \Phi^{(t)}\right) = E_{\Phi^{(t)}}[\ln L(\Phi)|Y]$
M- aşaması: $\Phi^{(t+1)} = \underset{\Phi}{\arg\max} Q\left(\Phi, \Phi^{(t)}\right)$

biçiminde tanımlanan bir E-aşaması ve bir M-aşamasından oluşur (McLachlan ve Krishnan. 1997:22). Burada t yineleme sayısını göstermektedir. Ayrıca z_i, x_i gözleminin hangi bileşenden olduğunu belirtmek için kullanılan ancak gözlenemeyen etiket vektörlerini temsil etsin. x_i, $(i = 1,2,\dots,n)$gözlemine karşılık gelen etiket vektörü $z_i = [z_{i1}, z_{i2}, \dots, z_{ik}]$ ve $j = 1,2,\dots,k$ için

$$z_{ij} = \begin{cases} 1, & x_i, j.\text{bileşen dağılımından bir gözlem} \\ 0, & \text{aksi durumlarda} \end{cases} \quad (6)$$

olmak üzere $Y = (X, Z)$biçimindedir.

3.1. İki Parametreli Rayleigh Dağılımlarının Karmaları için En Çok Olabilirlik Tahmin Edicileri

k bileşenli iki parametreli Rayleigh karma dağılımının olasılık yoğunluk fonksiyonu (4) eşitliğinde verildiği biçimde ve $X_1, X_2, \dots, X_{\mathrm{n}}$ rasgele değişkenleri bu dağılımdan rasgele bir örneklem olmak üzere, $X_1, X_2, \dots, X_n$ rasgele değişkenlerine ait logaritmik olabilirlik fonksiyonu

$$\ln L(\Phi) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} p_j f_j\left(x_i, \theta_j\right) \quad (7)$$

dir. Burada θ_j, $\theta_j = [\lambda_j, \mu_j]'$ olup *j.* bileşen dağılımı için parametre vektörünü, Φ ise $\Phi = [p_1, p_2, \ldots p_k, \theta_1, \theta_2, \ldots \theta_k]'$ biçiminde olup karma dağılım için parametre vektörünü göstermektedir. (7) eşitliği ile verilen logaritmik olabilirlik fonksiyonunu maksimize etmek için

$$g(\Phi, \tau) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} p_j f_j(x_i, \theta_j) + \tau\left(\sum_{j=1}^{k} p_j - 1\right) \quad (8)$$

Lagrange fonksiyonu tanımlansın. Burada $\tau\left(\sum_{j=1}^{k} p_j - 1\right)$ Lagrange çarpanı olarak alınmıştır. Dağılımın karma oran parametrelerini tahmin etmek için (8) eşitliğinin p_j' ye türevi alınırsa

$$\frac{\partial}{\partial p_j}\sum_{i=1}^{n} \ln \sum_{j=1}^{k} p_j f_j(x_i, \theta_j) + \tau\left(\sum_{j=1}^{k} p_j - 1\right) = \sum_{i=1}^{n} \frac{f_j(x_i, \theta_j)}{\sum_{j=1}^{k} p_j f_j(x_i, \theta_j)} + \tau$$

$$= \sum_{i=1}^{n} \left(\frac{p_j f_j(x_i, \theta_j)}{\sum_{j=1}^{k} p_j f_j(x_i, \theta_j)}\right) \frac{1}{p_j} + \tau \quad (9)$$

elde edilir. (9) eşitliğinin sıfıra eşitlenmesiyle

$$\sum_{i=1}^{n} W(j|x_i) \frac{1}{p_j} + \tau = 0 \quad (10)$$

$$\sum_{i=1}^{n} W(j|x_i) = -p_j \tau \quad (11)$$

elde edilir. $-\tau = c$ olarak alınıp (11) eşitliğinde yerine yazılırsa

$$\sum_{i=1}^{n} W(j|x_i) = c p_j \quad (12)$$

olur. Böylece dağılımın j. bileşenine karşılık gelen karma oran parametresine ait tahmin edici

$$p_j = \frac{1}{c}\sum_{i=1}^{n} W(j|x_i) \quad (13)$$

olarak elde edilir. Burada, c bir sabit ve,

$$W(j|x_i) = \frac{p_j f_j(x_i, \theta_j)}{\sum_{j=1}^{k} p_j f_j(x_i, \theta_j)} \quad (14)$$

dir. Diğer taraftan $\theta_j = [\lambda_j \mu_j]'$ parametrelerini tahmin etmek için (8) eşitliğinin θ_j'ye göre türevi alınıp sıfıra eşitlenirse

$$\frac{\partial}{\partial \theta_j}\sum_{i=1}^{n} \ln \sum_{j=1}^{k} p_j f_j(x_i, \theta_j) = \sum_{i=1}^{n} \frac{p_j \frac{\partial}{\partial \theta_j} f_j(x_i, \theta_j)}{\sum_{j=1}^{k} p_j f_j(x_i, \theta_j)} = 0 \quad (15)$$

elde edilir. (14) eşitliğinin hem payı hemde paydasının $f_j(x_i, \theta_j)$ ile çarpılmasıyla

$$\sum_{i=1}^{n} W(j|x_i) \frac{\partial}{\partial \theta_j} \ln\left(f_j(x_i, \theta_j)\right) = 0 \quad (16)$$

elde edilir. $\theta_j = (\lambda_j, \mu_j)$ olduğu göz önünde bulundurularak (4) eşitliği ile verilen olasılık yoğunluk fonksiyonun doğal logaritmasının λ_j ve μ_j' ye göre türevleri alınıp (16) eşitliğinde yerine konsun. λ_j için

$$\frac{\partial}{\partial \lambda_j} \ln\left(f_j(x_i, \lambda_j, \mu_j)\right) = \left(\frac{1}{\lambda_j} - (x_i - \mu_j)^2\right) I_{x_i > \mu_j} \quad (17)$$

olmak üzere,

$$\sum_{i=1}^{n} W(j|x_i) \left(\frac{1}{\lambda_j} - \left(x_i - \mu_j\right)^2\right) I_{x_i > \mu_j} = 0 \quad (18)$$

$$\frac{\sum_{i=1}^{n} W(j|x_i)}{\lambda_j} = \sum_{i=1}^{n} W(j|x_i)(x_i - \mu_j)^2 I_{x_i > \mu_j} \quad (19)$$

eşitliğine ulaşılır. (19) eşitliğinden λ_j çekilirse

$$\lambda_j = \frac{\sum_{i=1}^{n} W(j|x_i)}{\sum_{i=1}^{n} W(j|x_i)(x_i-\mu_j)^2 I_{x_i>\mu_j}} \quad (20)$$

elde edilir. Benzer biçimde (4) eşitliğinden μ_j için

$$\frac{\partial}{\partial \mu_j} \ln\left(f_j(x_i, \lambda_j, \mu_j)\right) = \left(2\lambda_j(x_i-\mu_j) - (x_i-\mu_j)^{-1}\right) I_{x_i>\mu_j} \quad (21)$$

dir. (21) eşitliğinde hesaplanan türevin (16) eşitliğinde yerinde kullanılmasıyla

$$\sum_{i=1}^{n} W(j|x_i)\left(2\lambda_j\left(x_i-\mu_j\right) - \left(x_i-\mu_j\right)^{-1}\right) I_{x_i>\mu_j} = 0 \quad (22)$$

$$\sum_{i=1}^{n} W(j|x_i)\left(x_i-\mu_j\right) I_{x_i>\mu_j} = \frac{1}{2\lambda_j}\sum_{i=1}^{n} \frac{W(j|x_i)}{\left(x_i-\mu_j\right) I_{x_i>\mu_j}} \quad (23)$$

$$\sum_{i=1}^{n} W(j|x_i) x_i\, I_{x_i>\mu_j} - \sum_{i=1}^{n} W(j|x_i)\mu_j = \frac{1}{2\lambda_j}\sum_{i=1}^{n} \frac{W(j|x_i)}{(x_i-\mu_j) I_{x_i>\mu_j}} \quad (24)$$

elde edilir. Böylece (24) eşitliğinden μ_j,

$$\mu_j = \frac{1}{\sum_{i=1}^{n} W(j|x_i)}\left(\sum_{i=1}^{n} W(j|x_i) x_i I_{x_i>\mu_j} - \frac{1}{2\lambda_j}\sum_{i=1}^{n} \frac{W(j|x_i)}{(x_i-\mu_j) I_{x_i>\mu_j}}\right) \quad (25)$$

olarak bulunur. Böylece bir Φ^0 başlangıç tahmin değeri ile birlikte (13), (20), (25) eşitliklerinin göz önüne alınması ve $c = n$ seçilmesiyle, k bileşenli iki parametreli Rayleigh karma dağılımının parametreleri için E-M algoritmasına dayalı en çok olabilirlik tahmin edicileri

$$\hat{p}_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} W^{(t)}(j|x_i) \quad (26)$$

$$\hat{\lambda}_j^{(t+1)} = \frac{\sum_{i=1}^{n} W^{(t)}(j|x_i)}{\sum_{i=1}^{n} W^{(t)}(j|x_i)\left(x_i-\hat{\mu}_j^{(t)}\right)^2 I_{x_i>\mu_j}} \quad (27)$$

$$\hat{\mu}_j^{(t+1)} = \frac{1}{\sum_{i=1}^{n} W^{(t)}(j|x_i)}\left(\sum_{i=1}^{n} W^{(t)}(j|x_i) x_i I_{x_i>\mu_j} - \frac{1}{2\hat{\lambda}_j^{(t)}}\sum_{i=1}^{n} \frac{W^{(t)}(j|x_i)}{\left(x_i-\hat{\mu}_j^{(t)}\right) I_{x_i>\mu_j}}\right) \quad (28)$$

olarak elde edilir. Burada $W^{(t)}(j|x_i) = \frac{\hat{p}_j^{(t)} f_j(x_i,\widehat{\theta_j})}{\sum_{j=1}^{k} \hat{p}_j^{(t)} f_j(x_i,\widehat{\theta_j})}$ dır.

4. Bulgular

4.1. Simülasyon Çalışması

Bu bölümde, bir önceki bölümde elde edilen tahmin edicilerin karma dağılımın parametrelerini tahmin etmedeki performansını ortaya koymak için iki bileşenli iki parametreli Rayleigh karma dağılım göz önünde bulundurularak yapılan simülasyon çalışmasına yer verilmektedir. Simülasyon çalışmasında karma oran parametresinin $p = 0.50$ ve $p = 0.70$ olduğu iki farklı durum düşünülmüştür.

Durum 1: $p = 0.50$ durumu için $\lambda_1, \mu_1, \lambda_2$ ve $\mu_2$parametrelerine ait değerler sırasıyla $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 0$ ve $\mu_2 = 1{,}3{,}5$ olarak seçilmiş ve ilgili dağılımlardan *n*= 50,100 ve 200 birimlik rasgele örneklemler üretilmiştir. 1000 defa tekrarla gerçekleştirilen simülasyon çalışmasıyla, parametre tahmin değerleri (Ort.), tahminlere ait simülasyon hata kareler ortalaması (HKO) ve standart sapma (S.Sapma) değerleri elde edilmiştir. Simülasyon çalışmasından elde edilen sonuçlar Tablo 1-3 de verildiği gibidir.

Tablo 1: $p = 0.50, \lambda_1 = 1, \mu_1 = 0, \lambda_2 = 2$ ve $\mu_2 = 1$ **Değerleri için Simülasyon Sonuçları**

n	50			100			200		
Parametre	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma
p	0.628	0.024	0.090	0.609	0.017	0.070	0.567	0.007	0.053
λ_1	3.302	108.300	10.200	1.785	1.588	0.990	1.257	0.281	0.466
μ_1	0.079	0.014	0.086	0.064	0.009	0.073	0.025	0.002	0.037
λ_2	2.407	1.233	1.039	2.102	0.236	0.478	2.197	0.285	0.499
μ_2	0.989	0.019	0.139	0.982	0.009	0.094	1.014	0.005	0.070

Tablo 2: p**= 0.50,** λ_1 **=1,** μ_1 **=0,** λ_2 **=2 ve** μ_2 **=3 Değerleri için Simülasyon Sonuçları**

n	50			100			200		
Parametre	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma
p	0.501	0.000	0.003	0.501	0.000	0.006	0.501	0.000	0.003
λ_1	1.221	0.147	0.315	1.107	0.051	0.199	1.073	0.021	0.126
μ_1	0.066	0.010	0.074	0.040	0.004	0.046	0.030	0.002	0.038
λ_2	2.361	0.584	0.677	2.135	0.163	0.382	2.060	0.080	0.279
μ_2	3.029	0.007	0.079	3.016	0.002	0.043	3.007	0.001	0.034

Tablo 3: p**= 0.50,** λ_1 **=1,** μ_1 **=0,** λ_2 **=2 ve** μ_2 **=5 Değerleri için Simülasyon Sonuçları**

n	50			100			200		
Parametre	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma
p	0.501	0.000	0.003	0.501	0.000	0.003	0.501	0.000	0.003
λ_1	1.202	0.241	0.449	1.110	0.048	0.191	1.070	0.028	0.154
μ_1	0.063	0.012	0.090	0.041	0.004	0.051	0.030	0.002	0.034
λ_2	2.304	0.619	0.729	2.210	0.227	0.429	2.075	0.072	0.258
μ_2	5.033	0.005	0.063	5.022	0.002	0.042	5.009	0.001	0.029

Tablo 1 – 3 ile verilen sonuçlara bakıldığında gözlem sayısı arttıkça tahmin değerleri gerçek değerlere yaklaşmakta ve tahmin edicilerin hepsi her durumda daha küçük HKO ve S.sapma değerlerine sahip olmaktadır. Bununla birlikte yine Tablo 1 – 3 de verilen sonuçlara göre gözlem sayısı 50 olduğunda, karma dağılımın konum parametreleri μ_1 ve μ_2'nin değerleri birbirine yaklaştıkça λ_1 ve λ_2 parametrelerinin tahminleri için hesaplanan HKO ve S.sapma değerleri artmaktadır. Ancak bu durum gözlem sayısının artmasıyla ortadan kalkmaktadır.

Durum 2: p =0.70 durumu için Durum 1'de olduğu gibi $\lambda_1, \mu_1, \lambda_2$ ve $\mu_2$parametrelerine ait değerler sırasıyla $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 0$ ve $\mu_2 = 1,3,5$ olarak seçilmiş ve ilgili dağılımlardan n =50,100 ve 200 birimlik rasgele örneklemler üretilmiştir.1000 tekrarlı simülasyon çalışmasıyla elde edilen sonuçlar Tablo 4 – 6'da sunulmuştur.

Tablo 4: **p= 0.70, λ_1 =1, μ_1 =0, λ_2 =2 ve μ_2 =1 Değerleri için Simülasyon Sonuçları**

n	50			100			200		
Parametre	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma
p	0.646	0.015	0.108	0.642	0.013	0.098	0.642	0.012	0.093
λ_1	2.214	7.103	2.385	1.587	0.773	0.658	1.475	0.761	0.735
μ_1	0.066	0.010	0.072	0.051	0.006	0.054	0.026	0.002	0.029
λ_2	2.799	5.507	2.218	2.223	0.934	0.945	2.368	3.600	1.871
μ_2	0.957	0.058	0.239	0.980	0.035	0.186	0.951	0.031	0.169

Tablo 5: **p= 0.70, λ_1 =1, μ_1 =0, λ_2 =2 ve μ_2 =3 Değerleri için Simülasyon Sonuçları**

n	50			100			200		
Parametre	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma
p	0.698	0.000	0.009	0.699	0.000	0.004	0.698	0.000	0.007
λ_1	1.162	0.093	0.258	1.066	0.027	0.151	1.037	0.013	0.108
μ_1	0.049	0.006	0.061	0.035	0.003	0.042	0.018	0.001	0.027
λ_2	2.461	1.533	1.155	2.237	0.401	0.590	2.167	0.175	0.385
μ_2	3.027	0.015	0.121	3.024	0.005	0.063	3.015	0.003	0.056

Tablo 6: **p= 0.70, λ_1 =1, μ_1 =0, λ_2 =2 ve μ_2 =5 Değerleri için Simülasyon Sonuçları**

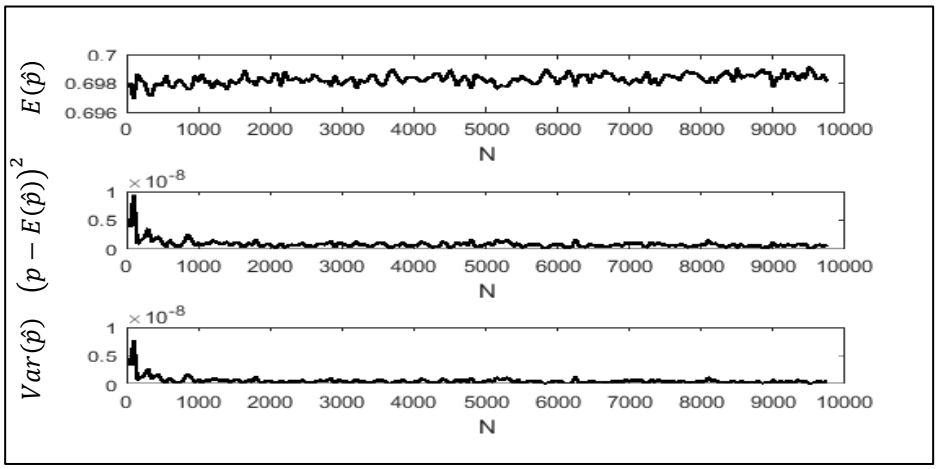
n	50			100			200		
Parametre	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma	Ort.	HKO	S.Sapma
p	0.695	0.001	0.025	0.696	0.001	0.024	0.698	0.000	0.021
λ_1	1.083	0.081	0.273	1.045	0.029	0.166	1.026	0.014	0.116
μ_1	0.045	0.007	0.069	0.031	0.003	0.041	0.024	0.002	0.032
λ_2	2.258	0.996	0.969	2.113	0.425	0.645	2.023	0.252	0.504
μ_2	5.003	0.022	0.150	4.987	0.016	0.126	4.989	0.012	0.109

Durum 2 için yapılan ve Tablo 4 – 6 ile verilen simülasyon çalışması sonuçları Durum 1 için yapılan simülasyon çalışması sonuçları ile tamamen uyum içerisindedir. Dolayısıyla simülasyon çalışmasına ait sonuçlara dayanarak; yeterli gözlem olması halinde elde edilen tahmin edicilerin HKO ölçütüne göre oldukça iyi bir tahmin performansına sahip oldukları söylenebilir.

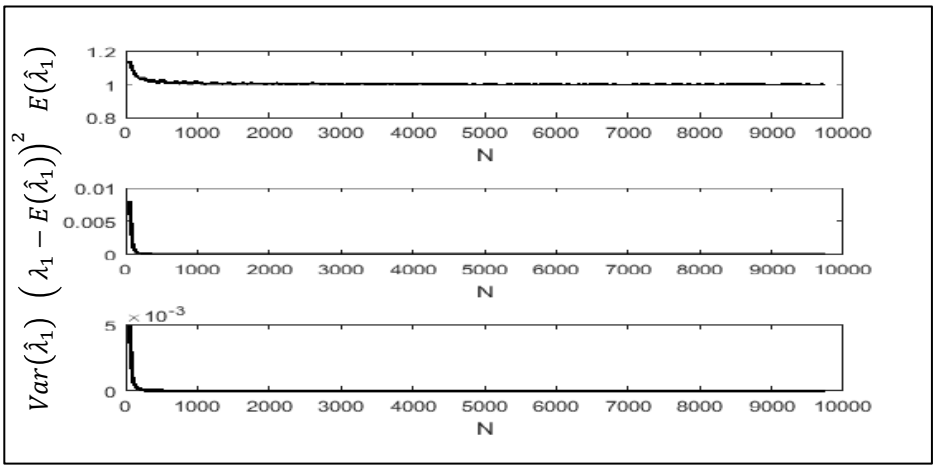
Üçüncü bölümünde elde edilen tahmin edicilerin tahmin performanslarının gösterilmesinin

yanında bu kısımda ayrıca, tahmin edicilerin asimptotik özelliklerini araştıran ikinci bir simülasyon çalışmasına yer verilmiştir. Simülasyon çalışmasında, $p = 0.70, \lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0$ ve $\mu_2 = 3$ parametre değerlerine sahip iki parametreli Rayleigh dağılımının iki bileşenli karması göz önüne alınmış ve ilgili dağılımdan $n = 50{,}100{,}150, \ldots, 10.000$ birimlik rasgele örneklemler üretilmiştir. Her n değeri için 1000 tekrarlı olarak gerçekleştirilen simülasyon çalışmasında sırasıyla $\hat{p}, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\mu}_1, \hat{\mu}_2$ tahmin edicilerine ait beklenen değerler ($E(\hat{p})$, $E(\hat{\lambda}_1)$, $E(\hat{\lambda}_2)$, $E(\hat{\mu}_1)$, $E(\hat{\mu}_2)$), yanlılık miktarlarının karesi ($[p - E(\hat{p})]^2$, $\left[\lambda_1 - E(\hat{\lambda}_1)\right]^2$, $\left[\lambda_2 - E(\hat{\lambda}_2)\right]^2$, $[\mu_1 - E(\hat{\mu}_1)]^2$, $[\mu_2 - E(\hat{\mu}_2)]^2$) ve varyanslar ($Var(\hat{p})$, $Var(\hat{\lambda}_1)$, $Var(\hat{\lambda}_2)$, $Var(\hat{\mu}_1)$, $Var(\hat{\mu}_2)$) elde edilmiştir. Elde edilen sonuçlar Şekil 3 – Şekil 7 de çizilerek verilmiştir.

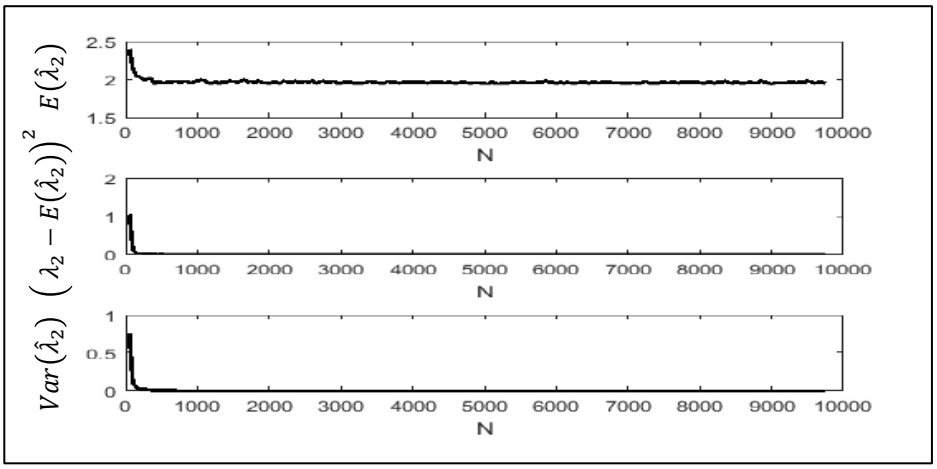
Şekil 3: p **Parametresi için Asimptotik Sonuçlar**



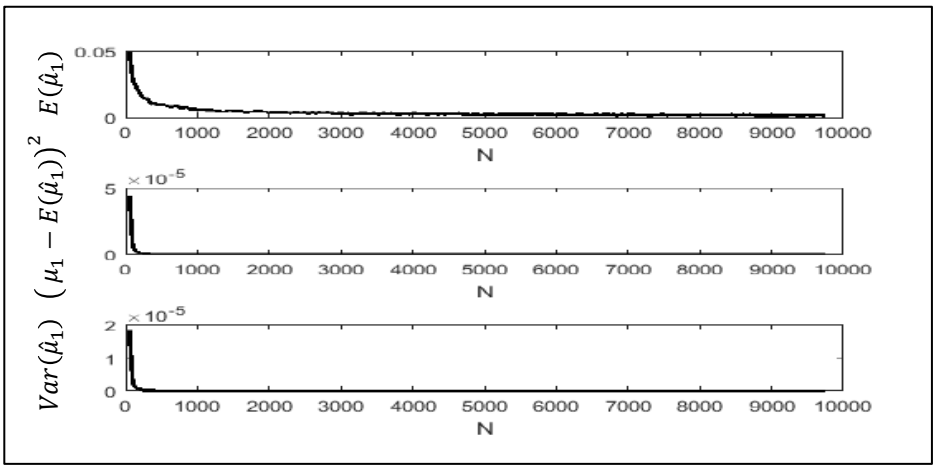
Şekil 4: λ_1 **Parametresi için Asimptotik Sonuçlar**



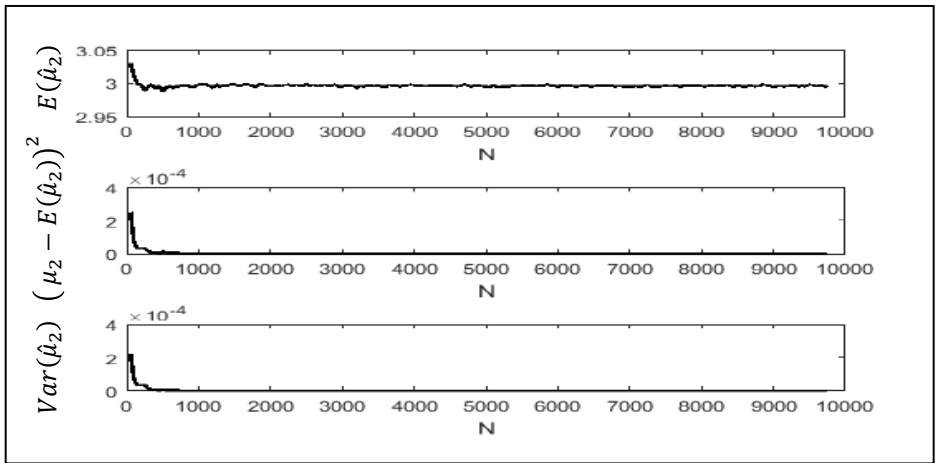
Şekil 5: λ_2 **Parametresi için Asimptotik Sonuçlar**



Şekil 6: μ_1 **Parametresi için Asimptotik Sonuçlar**



Şekil 7: μ_2 **Parametresi için Asimptotik Sonuçlar**



Şekil 3 – 7 incelendiğinde tüm tahmin edicilerin asimptotik olarak yansız ve tutarlı oldukları söylenebilir.

4.2. Uygulama

Bu kısımda, çalışmada elde edilen tahmin ediciler kullanılarak, klima arıza süreleri veri seti iki parametreli Rayleigh dağılımının karmaları ile analiz edilmektedir.

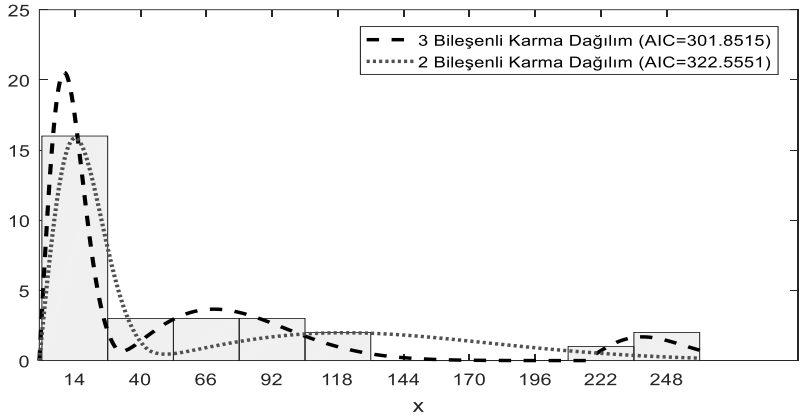
Klima arıza süreleri veri seti, bir uçağın klima sisteminin ardışık arızaları arasındaki süreleri (uçuş saati) içermektedir. Veriler Proschan (1963) tarafından yapılan çalışma ile elde edilmiş ve üstel dağılım kullanılarak modellenmiştir. Gupta ve Kundu (2003) bu veri setinin modellenmesinde Weibull ve genelleştirilmiş üstel dağılımı kullanmış ve Weibull dağılımının bu veri seti için daha uygun olduğunu belirtmiştir. Klima arıza süreleri veri seti için model olarak iki parametreli Rayleigh dağılımının 2 ve 3 bileşenli karmaları düşünüldüğünde, her iki karma dağılama ait bilinmeyen parametre değerleri Tablo 7 de verildiği gibi bulunmuştur.

Tablo 7: **Klima Arıza Süreleri Veri Seti için Parametre Tahminleri**

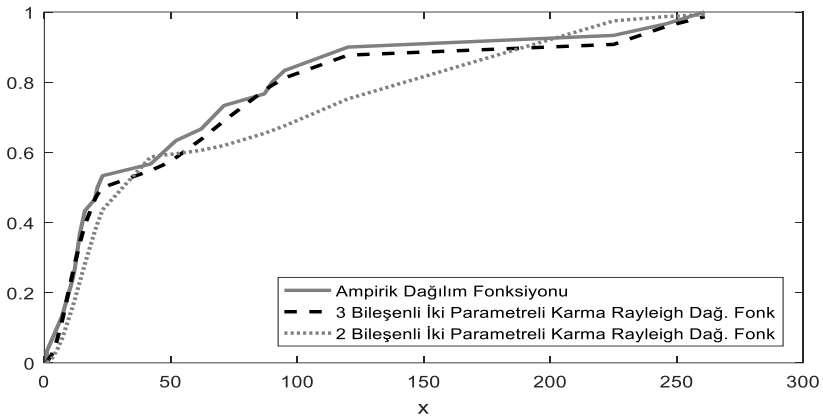
Model			
2 Bileşenli Karma Dağılım		3 Bileşenli Karma Dağılım	
Parametre	Tahmin	Parametre	Tahmin
p_1	0.5941	p_1	0.5337
p_2	0.4059	p_2	0.3663
λ_1	0.0025	p_3	0.1000
μ_1	0.0000	λ_1	0.0052
λ_2	0.0001	μ_1	0.0182
μ_2	43.2637	λ_2	0.0004
		μ_2	31.3123
		λ_3	0.0010
		μ_3	216.1029

Tablo 7'de verilen tahmin değerleri göz önünde bulundurularak klima arıza süreleri veri setinin iki parametreli Rayleigh dağılımlarının 2 veya 3 bileşenli karmaları ile modellenip modellenemeyeceğini ortaya koymak için veri setine Kolmogorov-Smirnov testi uygulanmıştır, 2 bileşenli karma dağılım için Kolmogorov-Smirnov istatistiğinin değeri 0.1576 ve ilgili *p*-değeri 0.40390 olarak, 3 bileşenli karma dağılım için de Kolmogorov-Smirnov istatistiğinin değeri 0.1157 ve ilgili *p*-değeri 0.7742 olarak bulunmuştur. Dolayısıyla her iki dağılımında bu veri seti için bir model olarak kullanılabileceği söylenebilir. Veri setini modellemede kullanılan 2 ve 3 bileşenli dağılımlar arasından hangisinin veri setine daha uygun olduğu hakkında bir fikir sahibi olabilmek için verilerin histogramları üzerine model olarak kullanılacak dağılımların olasılık yoğunluk fonksiyonları çizdirilebilir, bakınız Şekil 8. Benzer şekilde, verilerin ampirik dağılım fonksiyonu ve model dağılımların dağılım fonksiyonları aynı grafikte birlikte çizdirilebilir, bakınız Şekil 9.

Şekil 8: **Verilerin Histogramı ve Veri Setini Modellemede Kullanılan Dağılımlara Ait Olasılık Yoğunluk Fonksiyonları**



Şekil 9: **Klima Arıza Veri Setine Ait Ampirik Dağılım Fonksiyonu ve Model Dağılım Fonksiyonları**



Şekil 8 ve Şekil 9'dan açıkça görülmektedir ki, iki parametreli Rayleigh dağılımının 3 bileşenli karması bu veri setini 2 bileşenli karma dağılıma göre daha iyi açıklamaktadır. Şekil 8 de verilen Akaike bilgi kriteri (AIC) değerleri de bu sonucu desteklemektedir. Dahası, klima arıza süreleri veri setini modellemede kullanılabilecek olası dağılımlar için hesaplanan Akaike bilgi kriteri değerleri Tablo 8 de verilmiştir.

Tablo 8: **Klima Arıza Süreleri Veri Setini Modellemede Kullanılabilecek Olası Dağılımlar için Akaike Bilgi Kriteri Değerleri**

Model Dağılım	AIC
2 Bileşenli Karma Rayleigh	322.5551
3 Bileşenli Karma Rayleigh	301.8515
Üstel	307.2593
Weibull	307.8738
Gamma	308.3347
Log-Normal	307.2587

Tablo 8 de verilen AIC değerleri içerisinde en küçük olanın 3 bileşenli iki parametreli Rayleigh dağılımlarının karması olan dağılım olduğu görülmektedir. Dolayısıyla AIC kriterine göre, bu veri setini modellemede 3 bileşenli iki parametreli Rayleigh dağılımlarının karması olan dağılımın veri setini modellemede kullanılabilecek olası dağılımlar arasından en uygun dağılım olduğu sonucuna varılabilir.

5.Sonuç

Bu çalışmada, E-M algoritması kullanılarak, iki parametreli Rayleigh dağılımlarının sonlu karmalarında mevcut bilinmeyen parametrelerin en çok olabilirlik tahmin edicileri elde edildi. Elde edilen tahmin edicilerin bilinmeyen parametre değerlerini tahmin etmedeki performansı, simülasyon çalışmaları ile simülasyon hata kareler ortalaması ölçütüne göre değerlendirilmiştir. Simülasyon çalışması sonuçları elde edilen tahmin edicilerin, bileşen dağılımlarının konum parametrelerinin birbirinden yeterince uzak olduğu durumlarda, küçük örneklem çaplarında bile oldukça etkin olduğunu göstermiştir. Bileşen dağılımlarının konum parametrelerinin birbirine yakın olduğu durumlarda ise gözlem sayısının artmasıyla orantılı olarak tahmin edicilerin tahmin performansının arttığı gözlenmiştir. Ayrıca yapılan ikinci simülasyon çalışması ile elde edilen tahmin edicilerin asimptotik olarak tutarlı tahmin ediciler olduğu gösterilmiştir. Dolayısıyla E – M algoritmasına dayalı olarak elde edilen en çok olabilirlik tahmin edicilerinin Rayleigh dağılımlarının sonlu karmalarında mevcut bilinmeyen parametre değerlerinin tahmininde kullanılmasının uygun olduğu söylenebilir.

Kaynakça

Açıkgöz, İ. (2007). Sonlu Karma Dağılımlarda Parametre Tahmini. (Yayımlanmamış Doktora Tezi). Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Afify, W. M. (2011). Classıcal Estimation of Mixed Rayleigh Distribution in Type I Progressive Censored. *Journal of Statistical Theory and Applications,* 10(4), 619-632.

Dempster, A. P., Laird, N. M., ve Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1), 1-38.

Dey, S., Dey, T. ve Kundu, D. (2014). Two-parameter Rayleigh distribution: different methods of estimation. *American Journal of Mathematical and Management Sciences*, 33(1), 55-74.

Dick, N. P. ve Bowden, D. C. (1973). Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics*, 29(4), 781-790.

Elmahdy, E. E. ve Aboutahoun, A. W. (2013). A New Approach for Parameter Estimation of Finite Weibull Mixture Distributions for Reliability Modeling. *Applied Mathematical Modelling*, 37(4), 1800-1810.

Everitt, B. S. ve Hand, D. J. (1981). *Finite Mixture Distributions.* London: Monographs on Applied Probability and Statistics. Chapman and Hall.

Gupta, R. D. ve Kundu, D. (2003). Discriminating Between Weibull and Generalized Exponential Distributions. *Computational statistics & data analysis*, 43(2), 179-196.

Leytham, K. M. (1984). Maximum likelihood estimates for the parameters of mixture distributions. *Water resources research*, 20(7), 896-902.

Liu, Z., Almhana, J., Choulakian, V. ve McGorman, R. (2006). Traffic modeling with gamma mixtures and dynamical bandwidth provisioning. *Communication Networks and Services Research Conference,* 123-130, Canada.

McLachlan, G. J. ve Krishnan, T. (1997). *The EM Algorithm and Extensions, (2nd ed.)*. Canada: Wiley series in probability and statistics.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London.* 185(A), 71-110.

Proschan, F. (1963). Theoretical Explanation of Observed Decreasing Failure Rate. *Technometrics*, 5(3), 375-383.

Sum, S. T. ve Oommen, B. J. (1995). Mixture decomposition for distributions from the exponential family using a generalized method of moments. *IEEE transactions on systems, man, and cybernetics*, 25(7), 1139-1149.

Wang, Y. ve Wang, J. (2014). The EM Algorithm for The Finite Mixture of Exponential Distribution Models. *Int. J. Contemp. Math. Sciences*, 9(2), 57-64.

PARAMETER ESTIMATION IN FINITE MIXTURES OF TWO-PARAMETER RAYLEIGH DISTRIBUTIONS

Extended Abstract

Aim: The problem of parameter estimation for finite mixture distributions, which are very useful models in the statistical analysis of data obtained from heterogeneous populations, is a very important problem in statistics. In this work, maximum likelihood estimators for the finite mixtures of two parameter Rayleigh distributions, which can be used in the statistical analysis of a data set observed from a possible heterogeneous population, are obtained. The estimation performance and some asymptotic properties such as asymptoticaly unbiasedness and consistency of the estimators are shown by simulation study. Finally, by using the estimators obtained in study, a real data set is analysed with two and three component mixtures of two-parameter Rayleigh distribution.

Method(s): The two-parameter Rayleigh distribution is a skewed distribution that is used successfully in modeling many problems from science, medicine and health. The probability density function of the k-component mixtures of two-parameter Rayleigh distributions is

$$f(x,\Phi)=\sum_{j=1}^{k}p_j f_j(x,\theta_j)=\sum_{j=1}^{k}p_j 2\lambda_j(x-\mu_j)e^{-\lambda_j(x-\mu_j)^2}I_{\mathrm{x}>\mu_j},\quad x>\exists\mu_j,\mu_j\in\mathbb{R},\lambda_j>0 \qquad (1)$$

where λ_j is a scale parameter for the component *j,* $\mu_j$is location parameter for the component *j*, $I_{x_i>\mu_j}$ is a indicator function defined as follow

$$I_{x_i>\mu_j}=\begin{cases}1 & x_i>\mu_j\\ 0 & x_i\le\mu_j\end{cases} \qquad (2)$$

Also, where $p_j,\ p_j>0$ and $\sum_{j=1}^{k}p_j=1$, is the mixing ratio parameter for the component *j*. There are totally $(3\mathrm{k})$ unknown parameters in this distribution. Let we assume that $X_1,X_2,\ldots,X_{\mathrm{n}}$ is a random sample from mixture density given by equiation (1), then logarithmic likelihood function for the sample $X_1,X_2,\ldots,X_n$ is

$$\ln L(\Phi)=\sum_{i=1}^{n}\ln\sum_{j=1}^{k}p_j f_j(x_i,\theta_j) \qquad (3)$$

where $\theta_j,\ \theta_j=[\lambda_j,\mu_j]'$ is the parameter vector of the *j*th component density,$\Phi=[p_1,p_2,\ldots p_k,\theta_1,\theta_2,\ldots\theta_k]'$ is the parameter vector for the mixture distribution with k components. To obtain the maximum likelihood estimators for the k-component mixture of two-parameter Rayleigh distributions, firstly, let us define a Lagrange function,

$$g(\Phi,\tau)=\sum_{i=1}^{n}\ln\sum_{j=1}^{k}p_j f_j(x_i,\theta_j)+\tau\left(\sum_{j=1}^{k}p_j-1\right), \qquad (4)$$

for maximizing the logarithmic likelihood function given by equation (3). Where $\tau\left(\sum_{j=1}^{k}p_j-1\right)$ is a Lagrange multiplier. From solution of the Lagrange function given by equation (4), maximum likelihood estimators based on Expectation-Maximization (E-M) algorithm for the parameters of k-component mixtures of two-parameter Rayleigh distributions are obtained as follow, with initial estimation Φ^0,

$$\hat{p}_j^{(t+1)}=\frac{1}{n}\sum_{i=1}^{n}W^{(t)}(j|x_i) \qquad (5)$$

$$\hat{\lambda}_j^{(t+1)} = \frac{\sum_{i=1}^{n} W^{(t)}(j|x_i)}{\sum_{i=1}^{n} W^{(t)}(j|x_i)\left(x_i - \hat{\mu}_j^{(t)}\right)^2 I_{x_i>\mu_j}} \tag{6}$$

$$\hat{\mu}_j^{(t+1)} = \frac{1}{\sum_{i=1}^{n} W^{(t)}(j|x_i)}\left(\sum_{i=1}^{n} W^{(t)}(j|x_i)x_i I_{x_i>\mu_j} - \frac{1}{2\hat{\lambda}_j^{(t)}}\sum_{i=1}^{n}\frac{W^{(t)}(j|x_i)}{\left(x_i - \hat{\mu}_j^{(t)}\right)I_{x_i>\mu_j}}\right) \tag{7}$$

where $W^{(t)}(j|x_i) = \frac{\hat{p}_j^{(t)} f_j(x_i,\widehat{\theta_j})}{\sum_{j=1}^{k}\hat{p}_j^{(t)} f_j(x_i,\widehat{\theta_j})}$.

Findings: In order to demonstrate the estimation performance of the estimators obtained in this paper, a simulation study have carried out by considering the two-component mixtures of two-parameter Rayleigh distributions. We have considered two cases in simulation study in which ratio parameter $p = 0.50$ and $p = 0.70$, respectively. In both cases, values of the other parameters have been set as $\lambda_1 = 1$, $\lambda_2 = 2$, $\mu_1 = 0$ ve $\mu_2 = 1, 3, 5$. Sample sizes have been chosen as 50, 100,200 in both cases and it has been generated the random samples from the both configurations. By 1000 replicated Monte-Carlo simulations, for both cases, it has been computed the maximum likelihood estimations of parameters. Mean square errors and biases of estimations based on Monte-Carlo simulation have also been computed. The results of simulation study for both cases are show that the estimation performances of the estimators obtained in this paper are quite satisfactory in meaning to the mean square errors.

Conclusion: In this study, using the E-M algorithm, the maximum likelihood estimators of the parameters of the finite mixtures of the two-parameter Rayleigh distributions were obtained. The performance of the obtained estimators in estimating the unknown parameter values was evaluated according to the mean square error based on Monte-Carlo simulation. The simulation study results show that the estimation performances of the estimators obtained in this paper are quite satisfactory. In addition, this result is also valid for the small sample sizes, if the location parameters of the component distributions are far enough away from to each other. From simulation, we observe that when the location parameters of the component distributions are close to each other, the estimation performances slowly decrease. However, this decreasing disappears with increasing the sample size. Therefore, in order to estimating the unknown parameter values in the finite mixtures of the two-parameter Rayleigh distributions, it can be said that it is appropriate to using the maximum likelihood estimators based on the E-M algorithm.