T.C.
ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

GAUSS KARMA REGRESYON ANALİZİ'NİN, TÜRETİLMİŞ VERİLERDE ETKİNLİĞİNİN ARAŞTIRILMASI

DOKTORA TEZİ

EYLEM ITIR AYDEMİR

TEZ YÖNETİCİSİ
Doç. Dr. SETENAY ÖNER

EKİM 2009

T.C.
ESKİŞEHİR OSMANGAZİ ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ
BİYOİSTATİSTİK ANABİLİM DALI

GAUSS KARMA REGRESYON ANALİZİ'NİN, TÜRETİLMİŞ VERİLERDE ETKİNLİĞİNİN ARAŞTIRILMASI

DOKTORA TEZİ

EYLEM ITIR AYDEMİR

TEZ YÖNETİCİSİ
Doç. Dr. SETENAY ÖNER

KABUL VE ONAY SAYFASI

Eylem Itır AYDEMİR'in Doktora tezi olarak hazırladığı **"Gauss Karma Regresyon Analizinin Türetilmiş Verilerde Etkinliğinin Araştırılması"** başlıklı bu çalışma Eskişehir Osmangazi Üniversitesi Lisans Üstü Eğitim ve Öğretim yönetmeliği'nin ilgili maddesi uyarınca değerlendirilerek **"KABUL"** edilmiştir.

Tarih: 09.10.2009

Üye: Prof. Dr. Kazım ÖZDAMAR

Üye: Prof. Dr. İsmet DOĞAN

Üye: Doç. Dr. Setenay ÖNER (Danışman)

Üye: Doç. Dr. Fezan MUTLU

Üye: Yrd. Doç. Dr. Canan BAYDEMİR

Eskişehir Osmangazi Üniversitesi Sağlık Bilimleri Enstitüsü Yönetim Kurulu'nun 22./10./2009 tarih ...801../..3737.. Sayılı kararı ile onaylanmıştır.

Prof. Dr. Ferruh YÜCEL
Enstitü Müdürü

ÖZET

Araştırmacılar doğrusal olmayan örneklerde esnekliği elde etmek için parametrik modellerin geliştirilmesine yönelmişlerdir. Veri modellemede parametrik modeller çok boyutlu problemlere başarı ile uygulanırken esneklik varsayımlarını sağlayamayıp, yanlı tahminler vermektedir. Parametrik olmayan yöntemler ise esnekliği sağlar fakat yüksek boyutlarda güçlük çekmektedirler.

Esneklik varsayımlarını sağlayan GMR yöntemi, verideki heterojeniteyi belirlemek ya da aşırı yayılmayı açıklamak için regresyon modellerinin sonlu karmaları kullanılarak oluşturulur. Regresyon modeli oluşturulurken yoğunluk fonksiyonunu modelleme amacıyla kullanılır, verinin bileşik yoğunluğu modellenir ve GMM'den regresyon fonksiyonu türetilir.

Bu çalışmanın amacı, GMR yönteminin teorik temellerini açıklamak, Doğrusal ve Karesel Ayırma Analizi ile Esnek Ayırma Analizi yöntemlerinden MARS ve BRUTO Analizlerinin sonuçlarını karşılaştırmak, türetilmiş veriler kullanarak ayırma problemlerinde kullanımını göstermek ve Model Tabanlı Kümeleme yöntemlerini açıklamaktır.

Veri türetimi ve analizlerde; ortalama vektörleri, kovaryans matrisleri ve grup gözlem sayıları değiştikçe ayırma yöntemlerinin doğruluk oranları arasındaki değişiklikler, grup sayısının artışı ile birlikte ayırma yöntemlerinin doğruluk oranları arasındaki değişimler ve GMM'e Model Tabanlı Kümeleme yöntemi uygulayarak en iyi modelin belirlenmesi ve Poisson gürültü (Poisson noise) uygulandığındaki ayırma yönteminin nasıl uygulandığı gösterilmiştir.

Sonuç olarak Kovaryans matrisinin parametrizasyonuna göre, grup ortalama vektörleri arasındaki farka göre doğruluk oranlarının değiştiği, grup gözlem sayılarına göre doğruluk oranlarının değişmediği, büyük gözlem sayılarında GMR'nin yüksek

doğruluk oranları verdiği gözlenmiştir. GMR parametrik olmayan regresyon modellemede diğer yöntemlerin yerine kullanılabilir.

Anahtar Kelimeler: Gauss Karma Regresyon, Sonlu Karma Modeller, BIC, EM, Model Tabanlı Kümeleme Analizi, Çok Boyutluluk

SUMMARY

The analysts have gone towards to developing the parametric methods to get flexibility in non linear samples. Parametric models could be applied in high dimensions but could not supply the flexibility assumptions, this results in biassed forecasting. Non parametric methods have problems in high dimensions.

GMR is a flexible method used to determine the heterogenity in data and explain the overdispersion with Finite Mixtures. The GMM is used to model the density function and the joint density of the data and derived the regression function.

The main goal of this research is to explain the theoretic basis of GMR and to compare the analysis results with Lineer, Quadratic, MARS, BRUTO discriminant analysis and to show how the analysis work in discrimination when the simulation data is used and to explain the model-based DA.

For simulations and analysis; it is shown that, how the accuracy proportions of the discriminant methods changed when the mean vector, covariance matrix and group observation sizes are changed, how the accuracy proportions of discrimant methods are changed when the groups are increased, what is the best model when the model-based cluster method is applied to GMM, how the discriminant method works after poisson noise added to the model.

Finally the accuracy ratios are changed due to the covariance matrix parametrization and the difference between mean vectors, but the group observation sizes. The GMR should be used in non parametric regression modeling instead of other methods.

Key Words: Gaussian Mixture Regression, Finite Mixture Models, BIC, EM, Model-Based Clustering, High Dimension

İÇİNDEKİLER

ÇİZELGE DİZİNİ

ŞEKİL DİZİNİ

KISALTMALAR

1. BIC:Bayesian Information Criterion
2. BRUTO:Adaptive Additive Modeling
3. CA:Cluster Analysis (Kümeleme Analizi)
4. CART:Classification and Regression Trees (Sınıflandırma ve Regresyon Ağaçları)
5. DA:Discriminant Analysis (Ayırma Analizi)
6. EDA:Exploratory Data Analysis (Açıklayıcı Veri Analizi)
7. EM:Expectation Maximization
8. FA:Factor Analysis (Faktör Analizi)
9. FDA:Flexible Discriminant Analysis (Esnek Ayırma Analizi)
10. GAM:Generalized Additive Models (Genelleştirilmiş Eklemeli Modeller)
11. GCV:Generalized Cross Validation
12. GMC:Gaussian Mixture Classification (Gauss Karma Sınıflandırması)
13. GMM:Gaussian Mixture Models (Gauss Karma Modeller)
14. GMR: Gaussian Mixture Regression (Gauss Karma Regresyon)
15. HMM:Hidden Markov Models (Hidden Markov Modeller)
16. LDA:Linear Discriminant Analysis (Doğrusal Ayırma Analizi)
17. MANOVA:Multivariate Analysis of Variance (Çok değişkenli Varyans Analizi)
18. MARS:Multiadaptive Regression Splines
19. MCMC:Markov Chain Monte Carlo
20. MDA:Mixture Discriminant Analysis (Karma Ayırma Analizi)
21. MLE:Maximum Likelihood Estimation (En büyük benzerlik tahmini)
22. PCA:Principal Component Analaysis (Anabileşenler Analizi)
23. PPR:Projection Pursuit Regression (Öngörüsel Takip Regresyonu)
24. QDA:Quadratic Discriminant Analysis (Karesel Ayırma Analizi)

1- GİRİŞ ve AMAÇ

Gauss Karma Modeli, çok değişkenli istatistiksel yöntemlerin özelliklerinden olan boyut indirgeme, ayırma ve kümeleme işlemlerini başarı ile uygular. Değişkenlerin sürekli değişimini açıklayan Normal Dağılım ve Normal Dağılımdan türetilen ve çok değişkenli problemlerin incelenmesinde yaygın olarak kullanılan Çok Değişkenli Normal Dağılım fonksiyonları, Gauss Karma regresyonunu açıklamakta kullanılmaktadır (1,21,23,24,26,29,31).

Veri modellemede parametrik modeller çok boyutlu problemlere başarı ile uygulanırken esneklik varsayımlarını sağlayamaz böylece yanlı tahminler verir. Parametrik olmayan yöntemler ise esnekliği sağlar ama yüksek boyutlarda güçlük çekerler. Böylece hem çok boyutlu veriler için doğrusal olmayan örnekleri modelleyecek hem de parametrik olmayan yöntemlerin esnekliğini koruyabilecek bir model oluşturulmak istenmiştir. Verideki heterojeniteyi belirlemek ya da aşırı yayılmayı açıklamak için regresyon modellerinin sonlu karmaları kullanılarak Gauss Karma Regresyon yöntemi oluşturulur. İlk adımda regresyon modeli hemen oluşturulmaz, yerine yoğunluk fonksiyonunu modelleme amacıyla Gauss karmaları kullanılır. Gauss karmalarından verinin bileşik yoğunluğu modellenir. Daha sonra Gauss karma modelinden regresyon fonksiyonu türetilir. Bu fonksiyona Gauss Karma Regresyonu (Gauss Mixture Regression-GMR) denir (2,10,14,22,30).

Gauss karması kullanılarak oluşturulan modellerin yüksek doğruluk verdiği bilinmektedir. Gauss karmaları ile kümeleme yöntemlerinin kullanımında bulgusal yöntemlerden çok, olasılıksal yöntemler kullanılmaktadır. Model tabanlı kümeleme uygulaması, bilgisayar donanım ve yazılım performanslarının artması ile sonuçların da kolay yorumlanabilmesi ile ilgi kazanmıştır (5,22).

Karma modeller esnekliklerinden dolayı bilinmeyen dağılımları modellemede tercih edilmektedirler. Kümeleme analizinde olduğu gibi veri setindeki

gruplaşma yapısının analiz edildiği durumlarda her bileşen olasılığının bir kümeye uyması ile kümeleme problemine istatistiksel çözüm sunmaktadır. Karma model uygulamaları sadece istatistik analizlerdeki genel konularda değil aynı zamanda diğer alanlarda kontrolsüz örnek belirleme, konuşma tanıma (speech recognition), tıbbi görüntüleme yöntemleri gibi konularda önem kazanmıştır (17).

Karma model uygulamalarında bileşen sayısında ve dağılımında farklılık gösteren modeller istatistiksel yöntemler kullanılarak karşılaştırılabilir. Araştırmacı farklı parametrizasyonlar seçebilmekte, çok değişkenli karesel veya küresel modelleri dahil edebilmekte, karma parametreleri EM (Expectation Maximization) algoritması ile tahmin edilebilmekte, Model Tabanlı Hiyerarşik Kümeleme yapılabilmektedir. En iyi model BIC (Bayesian Information Criterion)'e göre seçilmektedir (2,5,9,10,13,22,30).

Son yıllarda bilgisayar performanslarının artması ile birlikte geliştirilen ve sayıları artan analiz yöntemleri, karma analizleri uygulamada daha sıklıkla kullanılan ve performansları araştırmacıları memnun eden uygulamalar haline getirmiştir. Bu çalışma;

1- Gauss Karma Regresyon yönteminin ülkemiz literatüründe çok bilinen bir yöntem olmaması nedeniyle bu yöntemin teorik temellerini açıklamak,

2- Türetilen verilerin analizinde Doğrusal Ayırma Analizi, Karesel Ayırma Analizi, Esnek Ayırma Analizi yöntemlerinden MARS ve BRUTO ile Karma Ayırma Analizi kullanılarak bu yöntemlerin sonuçlarını birbirleri ile karşılaştırmak,

3- Gauss Karma Regresyon yönteminin ayırma (discrimination) problemlerinde kullanımını göstermek,

4- Türetilmiş veriler kullanarak GMR yönteminde parametre tahmini yapmak için kullanılan prosedürlerin değişik birim, değişken ve dağılım tiplerine sahip olan verilerdeki etkinliklerini karşılaştırmak,

5- Model Tabanlı Kümeleme Yöntemi ile en iyi modelin belirlenmesini göstermek,

6- Model Tabanlı Kümeleme Yöntemi'ne Poisson Gürültü (Poisson Noise) uygulamak ve grafikle gösterimini açıklamak,

amacıyla yapılmıştır.

2 GENEL BİLGİLER

Gauss Karma Regresyonunun çok değişkenli veri modellemesinde de Normal ve Çok Değişkenli Normal Dağılım'dan yararlanılmaktadır.

2.1 Normal Dağılım (Gauss Dağılımı)

Normal Dağılım diğer bir adıyla Gauss Dağılımı, değişkenlerin sürekli değişimini açıklayan teorik bir dağılımdır. Frekans dağılımı çan eğrisi görünümüne sahip simetrik bir dağılım olan Normal Dağılımın yoğunluk fonksiyonu;

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad -\infty < X < +\infty$$

gibidir.

Burada μ, X değişkeninin toplum ortalamasını, σ, toplum standart sapmasını, π pi (π=3.1415..) sayısını ve e (e=2.718..) Neper sayısını göstermektedir. Normal dağılan bir X değişkeni için X~N(μ,Σ) gösterimi kullanılır (21,24,26).

2.2 Çok Değişkenli Normal Dağılım (Çok Değişkenli Gauss Dağılımı)

Çok değişkenli normal dağılım; ortalama vektörü (μ) ve varyans-kovaryans matrisi (Σ) ile tanımlanan, tek değişkenli normal dağılımın değişken sayısının iki ve/veya daha fazla olduğu durumlar için oluşturulmuş dağılımdır. Çok değişkenli problemlerin incelenmesinde yaygın olarak kullanılan dağılım, çözümleme kolaylığı ile tercih edilen bir dağılım türüdür. Çok değişkenli normal dağılımın yoğunluk fonksiyonu,

$$f(x) = \frac{1}{(2\pi)^{p/2}\left|\Sigma\right|^{1/2}} e^{-(X-\mu)'\Sigma^{-1}(X-\mu)/2} \qquad -\infty < X_i < +\infty;\ \text{i=1,2,..,p}$$

gibidir.

Burada p; boyut, μ; p*1 boyutlu ortalama vektörü, Σ, p*p boyutlu kovaryans matrisini, X rastgele değişken matrisini göstermektedir. Çok değişkenli normal dağılan bir X vektörü için $X \sim N_p(\mu,\Sigma)$ gösterimi kullanılır (1,23,29,31).

2.3 Çok Değişkenli Veri Analizi Yöntemleri

Çok değişkenli veri analiz edilirken bilinmeyen bir $F_{x,y}$ dağılımından bağımsız ve aynı şekilde dağılan rastgele değişkenler X ve Y arasındaki ilişki belirlenir. $Y = X\beta + \varepsilon$ şeklinde gösterilen model parametrik ve parametrik olmayan modelin temelini oluşturur. Çok değişkenli veri analizinde birden fazla bağımsız değişken (X) ve bağımlı değişken (Y) vardır (1,11-13,22,30,33).

Çok Değişkenli Varyans Analizi (Multivariate Analysis of Variance), Çok Değişkenli Regresyon Analizi, Ayırma Analizi (Discriminant Analysis), Faktör Analizi (Factor Analysis), Ana Bileşenler Analizi (Principal Component Analysis), Çok Boyutlu Ölçekleme (Multidimensional Scaling), Uyum Analizi (Corresponding Analysis), Kümeleme Analizi (Cluster Analysis), Setler Arası Korelasyon Analizi, çok değişkenli yöntemlerden yaygın olarak uygulanan yöntemlerdendir (23).

Gauss Karma Modeli, çok değişkenli istatistiksel yöntemlerin özelliklerinden olan boyut indirgeme, ayırma ve kümeleme analizlerini uygulamaktadır.

2.4 Ayırma Analizi

Ayırma Analizi farklı toplumlardan benzer özelliklere sahip grupların, birimlerin benzer özelliklerine göre gerçek gruplarına sınıflandırılmalarını sağlayan bir atama yöntemidir. 1936'da Fisher'in Iris veri seti (3 tür ve 4 ölçümden oluşur) için doğrusal ayırma fonksiyonunu oluşturmasından sonra Ayırma Analizi ile ilgili birçok çalışma yapılmış ve farklı kurallar oluşturulmuştur. Doğrusal Ayırma Analizi bütün değişkenleri kullanan klasik yaklaşımdır (23,11,13,22,31).

Ayırma Analizinde, veri setine uygun olacak biçimde parametrik ya da parametrik olmayan Ayırma Analiz yöntemleri kullanılmaktadır.

2.4.1 Parametrik Yöntemler

Parametrik yöntemlerin uygulanabilmesi için,

- X veri matrisinin normal dağılım gösteren toplumdan çekilmiş olması,
- X veri matrisi değişkenlerinin, kovaryans matrisi ortak çok değişkenli toplumlardan çekilmiş olması,
- Değişken ortalamaları ve varyansları arasında korelasyon olmaması,
- Değişkenlerde çoklu bağımlılık (multicollinearity) olmaması
- X matrisinin gereksiz değişkenler içermemesi,

gibi koşulların yerine getirilmesi gerekir (1,23).

2.4.1.1 Doğrusal Ayırma Analizi

İki grup ve çok gruplu veriler için uygulanan Doğrusal Ayırma Analizinde amaç, yeni bir gözlemi, her grubun kendi özelliğini değerlendirerek ait olduğu gruba atamaktır. Bunun için açıklayıcı değişkenlere göre ayırma fonksiyonları oluşturulur ve bu fonksiyonlardan ortak ayırma fonksiyonu elde edilir (1,23,29,31).

Doğrusal Ayırma Analizi'nde

$$Y_i = b_{0i} + b_{1i}X_1 + b_{2i}X_2 + \ldots + b_{pi}X_p$$

Sabit değer *Kanonik değişkenler*

-Doğrusal ayırma fonksiyonu-

- Ortalama vektörler $(\overline{X}_i)$, her gruba ilişkin kovaryans matrisleri (S_i) ve ortak kovaryans matrisleri (S_p) hesaplanır.
- Her grup için sabit değer (b_o) ve kanonik değişkenler (b_i) hesaplanır.

- Hesaplanan kanonik değişkenlere göre her grup için doğrusal ayırma fonksiyonları yazılır.
- Ortak ayırma fonksiyonu oluşturulur.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p$$

- Grupları birbirinden ayırmak için kullanılan Mahalanobis uzaklığı (Bölüm 2.4.1.3) hesaplanır, önemliliği ise Hotelling T^2 ile test edilir.
- Mahalanobis uzaklığı kullanılarak oluşturulan ortak ayırma fonksiyonu Y'nin önemliliği, normal dağılım varsayımına göre F testi ile test edilir.

Doğrusal Ayırma Analizinde kovaryans matrisi bütün sınıflarda eşit olur bu da herhangi iki sınıf arasında karar sınırını doğrusal yapar (1,13,29-30,33)

Kovaryans matrisinin eşit olma durumu **homoskedastisite** olarak adlandırılır. Kovaryans matrisi sınırlandırılmıştır ve her bileşen için aynıdır.

$$X_{ij} \sim N(\mu_j, \Sigma) \qquad \Sigma_j = \Sigma \quad (j = 1.....g)$$

Karma ayırma analizi ile çok boyutlu verilerin düşük boyutlarla tahmin edilmesi ve grafiklerle gösterilmesi doğrusal ayırma analizi ile uygulanabilir. Doğrusal ayırma analizinde büyük sayıda veri setlerinde model tahmini yapılırken, çok boyutluluk istenmeyen bir durumdur. Bu anlamda Doğrusal Ayırma Analizi sınıflandırmanın yanında boyut azaltma için de kullanılan önemli bir yöntemdir (1,12,23-24,33).

2.4.1.2 Karesel Ayırma Analizi

Karesel Ayırma Analizi'nin Doğrusal Ayırma Analizi'nden farkı grupların kovaryans matrislerinin birbirine eşit olmamasıdır. Hesaplamalarda kovaryans matrislerinin farkı alınarak işlemler yapılmaktadır.

Kovaryans matrisinin eşit olmama durumu, **heteroskedastisite** olarak adlandırılır. Her toplumun farklı ortalama ve kovaryans matrisine sahip olduğu ve kovaryans matrisinin sınırlandırılmadığı durumdur (22,31,33).

$$X_{ij} \sim N(\mu_j, \Sigma_j) \qquad \Sigma_j (j = 1.....g)$$

2.4.1.3 Mahalanobis Uzaklığı ve Yeni Bir Gözlemin Atanması

Mahalanobis uzaklığı Ayırma Analizinde önemli bir kavramdır ve değişkenler arasındaki uzaklığı bularak ait olduğu gruba atama yapmak için kullanılır.

x ve μ arasındaki Mahalanobis uzaklığı,

$$D(x, \mu) = \left[(x - \mu)' \Sigma^{-1} (x - \mu) \right]$$

şeklinde hesaplanır.

μ ortalamalar vektörü, Σ pozitif sonlu simetrik kovaryans matrisidir (2,11,22).

İki grup doğrusal ayırma analizinde; İki ortalama $(\mu_1; \mu_2)$ ve iki toplum (π_1, π_2) vardır. Eğer $(\mu_1 - \mu_2)^T \Sigma^{-1} X \geq C$ koşulu sağlanıyorsa doğrusal ayırma fonksiyonu yeni bir gözlemi birinci toplum (π_1)'a atar, diğer durumda ise yeni gözlem ikinci toplum (π_2)'a atanır.

Gruplar aynı derecede ise sabit,

$$C = \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 + \mu_2)$$

şeklinde hesaplanır.

Σ ortak kovaryans matrisidir. 2 grup homoskedastik normal toplumlarda Mahalanobis uzaklığı ne kadar büyükse gruplar o kadar iyi ayrılır. Yanlış sınıflandırma hatası da o kadar küçük olur (2).

2.4.2 Parametrik Olmayan Yöntemler

Çok değişkenli istatistiksel yöntemlerde, normal dağılım varsayımı sık kullanılan bir yöntemdir. Çok değişkenli parametrik olmayan yöntemler ise; çok değişkenli dağılan toplumlardan alınan çok değişkenli normal dağılıma uymayan verilerin analizinde kullanılan modelleme yöntemleridir. Çok değişkenli parametrik olmayan regresyon problemleri için birçok değişik yaklaşım vardır.

Regresyon modelleri, istatistiksel veri analizinde çok önemli bir yer tutar. Gerçek verilerde doğrusallık sağlanmadığında tahmin ve ayırma kurallarının uygulanabilmesi için esnek modeller geliştirilmiştir. Esnek modellerin varsayımları yok denecek kadar azdır. Bu modeller Hastie ve Tibshirani (1990) tarafından doğrusal olmayan regresyon için ***Genelleştirilmiş Eklemeli Modeller*** (Generalized Additive Models) diye isimlendirilmişlerdir. Her değişkene tek değişkenli düzleştirici ekleyerek uygulanmaktadır (2,10-11,14,43).

Bir eklemeli model aşağıdaki gibi gösterilir (2,10,14):

$$Y = \alpha + \sum_{j=1}^{p} m_j(X_j) + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

Buradaki α sabiti, X_j ise düzeltme terimi ya da bir doğrusal modeldir.

Parametrik olmayan düzleştirme yöntemlerinden Kernel En Yakın Komşu, Spline düzleştirme gibi parametrik olmayan yöntemler, p-boyutlu yerel ortalamalar yöntemi ile uygulanan regresyon yöntemleridir. p-boyutlu yerel ortalamalar yöntemi, X_0 noktasındaki bir regresyon yüzeyinin tahmininin, X_0'ın komşuluğundaki gözlemlerin ortalaması alınarak tahmin edilmesidir. Yüksek boyutlarda verilerin seyrekliği (sparsity)

durumu ile karşılaşıldığında (curse of dimensonality=boyut problemi) bu yöntemleri uygulamakta güçlükler yaşanmaktadır.

Öngörüsel Takip Regresyonu (Projection Pursuit Regression) Kruskal tarafından (1969-1972) bulunmuş, Friedman ve Stuetzle tarafından (1981) geliştirilmiş çok değişkenli parametrik olmayan regresyon problemidir. Öngörüsel Takip Regresyonu'nda p-boyutlu, bileşenlerine tahminci değişken denen rassal X vektörü ve bağımlı değişken Y vardır, bu yöntem boyut indirgemeyi ve regresyon modelini uygulayan bir yöntemdir. Tek değişkenli parametrik olmayan regresyon probleminde GAM ve PPR etkili yöntemlerdir (2,6,15,43).

Hastie ve ark. (1994) parametrik olmayan regresyon ve ayırma analizi arasındaki bağlantıyı uygun ölçek yaklaşımı (optimal scoring approach) kullanılarak sağlamışlardır. Bu yöntem Breiman ve Ihaka tarafından önerilmiş (1984) Hastie ve ark.(1994) tarafından Esnek Ayırma Analizi (Flexible Discriminant Analysis) adını almıştır. Konuşma tanıma ya da görüntü belirleme gibi çalışmaların büyük veri setlerinde Kısıtlanmış Ayırma Analizi (Penalised Discriminant Analysis) uygulanır. Karma Ayırma Analizinde de uzaklıklar hesaplanırken kısıtlanmış ölçümler kullanılır (12-13,22,33).

Parametrik olmayan regresyonda doğrusal olmayan örnekleri belirlemek için Esnek Açıklayıcı Veri Analizi (Flexible Exploratory Data Analysis) kullanılır. Esnek Açıklayıcı Veri Analizi'nde regresyon fonksiyonun sürekliliği varsayılır ve regresyon fonksiyonun düzgün bir tahmini oluşturulur. Bu nedenle parametrik olmayan regresyon modelleri düzgünleştirici (smoother) olarak da tanımlanmaktadır (30).

2.4.2.1 Esnek Ayırma Analizi Yöntemleri

Esnek Ayırma Analizi, Doğrusal Ayırma Analizinin parametrik olmayan yöntemidir. Uygun Ölçek Yaklaşımı, Kanonik Korelasyon Analizi ve Doğrusal Ayırma Analizinin kullanıldığı çoklu bağımlı regresyon yöntemidir (32).

Breiman ve arkadaşları (1984) çok yönlü parametrik olmayan yöntemlerden biri olan CART'ı önermişlerdir. MARS, 1991'de Friedman tarafından geliştirilmiş olup CART'ın daha genelleştirilmiş şeklidir.

Hem CART hem de MARS parametrik olmayan regresyonun genel bir durumudur, X özellik uzayını kare bloklara böler ve her bir blok için bir regresyon modeli uydurur. (3,4,7,9,10,25,30,32,43).

MARS yöntemi değişkenler arasındaki etkileşimleri modelleyen bir yöntemdir. MARS regresyon yönteminde, etkileşimler aşamalı olarak uygulanır. MARS modelinin genel formu,

$$Y = \beta_0 + \sum_{m=1}^{M} \beta_k h_m(x)$$

şeklindedir.

CART yöntemi, sabit parçalı fonksiyonlara dayalı iken MARS yöntemi parçalı ya da hinge fonksiyonlarının birleşimini oluşturur. Bu hinge fonksiyonu bir düğüm noktasına sahiptir. Modeldeki bağımsız değişkenin etkileşimleri ve doğrusal olmayan dönüşümler temel fonksiyon tarafından ifade edilmektedir. Bu anlamda bağımsız değişkenlerin birbiri ile doğrusal ya da doğrusal olmayan etkileşimlerinden her biri bir düğüm oluşturur ve buna göre tüm olası fonksiyonlar oluşturulur:

Hata kareler toplamının en küçük değeri aldığı düğüm değeri, aynı zamanda bağımsız değişkenin eğrinin eğimin değişmeden önceki son değeridir. İki ardışık düğüm değeri arasındaki doğrunun eğimik atsayısı yani regresyon eğimidir. Temel fonksiyonlar yardımı ile bulunan düğüm değerleri ile regresyon düzgün hale getirilir. Bu temel fonksiyonların hata kareler ortalamasının minimum olana kadar modelden çıkarması ile elde edilen "budama algoritması" ile gerçekleştirilir (4,25,32).

BRUTO analizi, düzgünleştirici eğriler (smoothing splines) kullanılarak uygulanan bir eklemeli model uygulamasıdır. BRUTO, temel veri seti modellendikten sonra katsayıları küçültüp analiz yapmaktadır.

Esnek veri analizi yöntemlerinden birisi Gauss Karma Regresyon yöntemidir. Çok değişkenli parametrik olmayan regresyonda Kernel fonksiyonu kullanılarak Gauss Karma regresyon modeli oluşturulur. Gauss Karma regresyon fonksiyonundan yoğunluk tahmin edilir. Gauss regresyon fonksiyonu Kernel yönteminin bir çeşidi olarak düşünüldüğü için herhangi bir boyutta uygulanabilmektedir (14,30).

2.5 Kümeleme Analizi

Kümeleme Analizi, n*p boyutlu X veri matrisinin çok değişkenli gözlemlerine ilişkin bilinmeyen gruplarının k kümeye ayrılması olarak tanımlanabilir. Kümeleme Analizinin Ayırma Analizinden ayıran fark, grup sayılarının ve grupların önceden bilinmemesidir. Kümeler ayrılırken gözlem çiftlerinin benzerlikleri (similarity) ve farklılıkları (dissimilarity) genel olarak uzaklık (distance) ölçüleri olarak isimlendirilir. Bu ölçüler, birimler ve değişkenler arasındaki uzaklıkları hesaplamak için kullanılır (23,26,31).

2.5.1 Uzaklık Ölçüleri, Uzaklık ve Benzerlik Matrisleri

2.5.1.1 Öklid Uzaklığı (Euclidean Distance)

Minkowski uzaklığının (2.5.1.2) özel bir hali olan Öklid uzaklığı iki vektör arasındaki karesel uzaklığı ölçer.

n×p boyutlu X veri matrisi aşağıdaki gibi verilmiş olsun,

$$X_{n\times p} = \begin{bmatrix} x_{11}\ x_{12}\ \ldots\ x_{1p} \\ x_{21}\ x_{22}\ \ldots\ x_{2p} \\ \ldots\ \ \ldots\ \ \ldots\ \ldots \\ x_{n1}\ x_{n2}\ \ldots\ x_{np} \end{bmatrix}$$

i. ve j. birimler arasındaki karesel uzaklık,

$$d(i,j) = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2}$$

şeklindedir. i=1,2,…,n; j=1,2,…,n ve k=1,2,…,p'dir.

Uzaklık matrisinde,

1. Bütün matris elemanları pozitiftir (Her x_i ve x_j için $d(i,j)\geq 0$)
2. Birinci köşegen elemanları 0'dır ($x_i = x_j$ ise $d(i,j)=0$)
3. Uzaklık matrisi simetrik matristir. ($d(i,j)=d(j,i)$)

Her matris elemanı bu matristeki en büyük değere bölünüp 1'den çıkarılır ve yüzde ile ifade edilir. Bu hesaplanan yeni matris *benzerlik matrisi*dir ve birimlerin birbiri ile benzerlik düzeylerini göstermektedir (23,26,31).

2.5.1.2 Minkowski uzaklığı

Minkowski uzaklığı,

$$d(x,y) = \left[\sum_{k=1}^{p} |x_k - y_k|^r \right]^{1/r}$$

şeklindedir. Burada r=2 için Öklid karesel uzaklığını elde edilir (23,26,31).

2.5.1.3 Manhattan (City block) uzaklığı

r = 1, için Minkowski uzaklığı **Manhattan (City block)** uzaklığı adını alır.

2.5.1.4 Korelasyon Uzaklığı

İki gözlem vektörü arasındaki ilişkiyi ölçen Korelasyon ve Korelasyonlar aracılığı ile hesaplanan korelasyon uzaklığı kümeleri birbirinden ayırmaya yardımcı olur (23,31).

$$r_{kj} = \frac{\sum_{i=1}^{n} X_k X_j - \sum_{i=1}^{n} X_k \sum_{i=1}^{n} X_j}{\sqrt{\left(\sum_{i=1}^{n} X_k^2 - \frac{(\sum_{i=1}^{n} X_k)^2}{n}\right)\left(\sum_{i=1}^{n} X_j^2 - \frac{(\sum_{i=1}^{n} X_j)^2}{n}\right)}}$$

Kümeleme yöntemleri uygulanırken temel olarak Aşamalı (Hierarchical) ve Aşamalı olmayan (Nonhierarchical) olmak üzere iki şekilde incelenmektedir (23,26,31).

2.5.2 Aşamalı Kümeleme Yöntemi (Hierarchical Clustering Method)

Aşamalı Kümeleme Analizi yönteminde birimler -uzaklık matrisinin elemanları kullanılarak ağaç diyagramları ve dendogramlar yardımı ile ardışık şekilde kümelenirler. Yöntem uygulanırken birimlerin başlangıçta nasıl kümeleneceği ve küme sayısının belirlenmesi için farklı yöntemler vardır.

2.5.2.1 Ayırıcı Aşamalı Kümeleme Yöntemi (Divisive Hierarchical Clustering Method)

Birimler başlangıçta tek küme oluşturuyor ise ile Ayırıcı Aşamalı Kümeleme Yöntemi uygulanır ve aşamalı olarak kümeler ayrılır.

2.5.2.2 Birleştirici Aşamalı Kümeleme Yöntemi (Agglomerative Hierarchical Clustering Method)

Başlangıçta her birim bir küme olarak kabul ediliyorsa bu kümeleme yöntemi Birleştirici Aşamalı Kümeleme Yöntemidir ve birimler aşamalı olarak kümelere yerleştirilir. Aşamalı yerleştirmede küme sayıları gittikçe küçülürken kümelerin kendisi büyür. Birleştirici Aşamalı Kümeleme Yöntemi uygulamada daha yaygın kullanılan yöntemdir (26,31-32).

2.5.3 Aşamalı Olmayan Kümeleme Yöntemi (Nonhierarchical Clustering Method)

Aşamalı olmayan kümeleme yöntemlerinde *optimizasyon, dağılım karmaları* ve *yoğunluk tahminleri* yöntemleri kullanılarak yerleştirilme yapılmaktadır (26).

Aşamalı olmayan kümeleme yöntemlerinin Aşamalı kümeleme yöntemlerinden farkı,

- Aşamalı olmayan kümeleme yöntemlerinde başlangıç küme sayısı; küme merkezleri ya da çekirdek noktalarından oluşan bir değerdir.
- Aşamalı olmayan kümeleme yöntemlerinde benzerlik ya da farklılık matrisi yerine X matrisi kullanılır.
- Aşamalı kümeleme yöntemlerinde bir kümeye atanan bir birim bir daha yer değiştirmez, Aşamalı olmayan kümeleme yöntemlerinde ise yer değiştirerek optimizasyon sağlanmaktadır (26).

2.5.3.1 Optimizasyon

Bir uzaklık ya da benzerlik matrisi kullanılmadan n birimi k kümeye belirli kriterler kullanarak yerleştirme işlemi optimizasyon ya da paylaştırma (partitioning)

işlemi olarak tanımlanmaktadır. Aşamalı olmayan kümeleme yöntemlerinde en yaygın olarak kullanılanları,

- K-ortalamalar yöntemi
- Yığma kümeleme yöntemleri'dir.

2.5.3.1.1 K-Ortalamalar Yöntemi

Birimlerin bir kümeden diğerine yer değiştirmek suretiyle uygun kümeye atanmasını sağlayan algoritmadır.

Başlangıçta çekirdek nokta olarak seçilen daha sonra ise küme merkezleri (ortalama vektörler) ile yer değiştirecek olan k değerleri seçilir. Çekirdek noktalar seçilirken çeşitli yaklaşımlar vardır. Örneğin bu noktalar rastgele, gelişigüzel ya da ilk k birimin k olarak alınması şeklinde olabilir. K çekirdek noktanın seçiminden sonra kalan birimler Öklid uzaklığına göre en yakın çekirdek noktanın bulunduğu kümeye atanır. Birden fazla birimlerin oluşturduğu noktaların küme merkezi çekirdek noktalarla yer değiştirir. Öklid uzaklığına göre kendi küme merkezinden başka bir küme merkezine daha yakın gözlem kalmayana kadar bu işlem devam ettirilir (23,26,31).

k-ortalamalar yöntemi başlangıç çekirdek noktasının seçimine duyarlıdır ve bazen bu seçime göre kümeler değişebilir (26)

2.5.3.1.2 Yığma Kümeleme Yöntemi

Bir noktanın küme merkezinden uzaklığına göre değil de küme içi ve kümeler arası kovaryans matrislerine göre hesaplanan *Yığma Kümeleme Yöntemi* (Hill Climbing Method) diğer bir paylaştırma yöntemidir (23).

2.5.3.2 Dağılım Karmaları

Diğer bir yöntem olan dağılım karmalarında; çok değişkenli normal dağılan g dağılım varsayımı ile hareket edilir ve örnekteki kümeler belirlenmeye çalışılır (26).

2.5.3.3 Yoğunluk Tahmini Yöntemi

Mod denilen yoğunluk noktaları aranarak yığılımın olduğu bölgelerin yığılım olmayan bölgelerden ayrılma işlemidir. r yarıçapı ve k noktası belirlenir. Verideki her nokta için r yarıçapının oluşturduğu küreye dahil noktalar bulunur ve küre içinde başka noktalarda varsa bu nokta yoğunluk noktası olur. Bir yoğunluk noktası diper yoğunluk noktalarından r çapından daha uzaksa o bir küme çekirdeğidir, değil ise kümeler birleştirilerek oluşturulur (26).

2.6 Karma Modeller

Karma modellerin incelenmesi çok uzun bir zaman dilimine dayanmasına rağmen, araştırma yöntemlerinin gelişmesi ve bilgisayar performanslarının artması ile kullanımındaki artış son yıllarda belirginleşmiştir. Karma modeller esnek ve üretimi kolay modeller olduğundan, uygulamada da tercih edilen modeller haline gelmişlerdir. İstatistiksel analizin geniş bir kısmına hitap eden matematik tabanlı karma modeller; kümeleme analizi, ayırma analizi, yoğunluk tahmini, sinir ağları gibi konularda, Tıp, Biyoloji, Sosyal Bilimlerin hemen her alanında kullanılmaktadır (2,22,30).

Karma modeller Pearson tarafından 1884'de incelenmiştir. 1000 tane yengecin baş uzunluklarının gövdeye oranları alınarak oluşturulan veri setinin histogram grafiğinden yola çıkılarak, grafikteki asimetrinin (çarpıklık) iki alt cinse ait olabileceği bunun da yeni bir türün evriminin göstergesi olabileceği Pearson'un model tabanlı hiyerarşik yaklaşımı ile belirlenmiştir. 1800'lü yılların sonunda Pearson tarafından analiz edilen 2 normal yoğunluk fonksiyonunun karması kullanarak uygulanan analiz zamanla gelişerek önemini artıran bir yöntem olmuştur (22).

Gauss karmaları parametrik olmayan yoğunluğu hesaplamak için kullanılan bir yöntemdir. Gauss karmaları kullanılarak hem parametrik yapı korunmuş, hem de model esnekliği sağlanmış olmaktadır. Gauss karmaları kullanılırken her bileşen orijinal bir model gibi gösterilip analiz edilmektedir. Oluşturulan regresyon modeli verideki gözlemlenmemiş heterojeniteyi belirlemek ya da aşırı yayılmayı açıklamak için kullanılmaktadır (30).

Karma modellerde, her karma bileşen G, gizli bir küme değişkeni gibi düşünülür $G \in \{1, 2, ..., K\}$.

Bu gizli değişkenlerin her bir bileşene eşit olma olasılığı π_k'dır.

$$\Pr(G = k) = \pi_k$$

π_k'a karma oranlar denir. Karma oranlar 0-1 arasında değer alır ve toplamı 1'e eşittir.

$$\pi_i \ (0 \leq \pi_i \leq 1) \qquad \sum_{i=1}^{g} \pi_i = 1$$

$$X \mid G = k \sim N(\mu_k, \Sigma_k)$$

p-boyutlu R örnek uzayında n birimden çekilmiş rasgele örnekler X_1, X_2,….,X_n olmak üzere X_i vektörünün olasılık yoğunluk fonksiyonu $f(x_i)$ aşağıdaki gibi gösterilir,

$$f(x_i) = \sum_{j=1}^{K} \pi_j \phi(x; \mu_j, \Sigma j)$$

ϕ çok değişkenli Gauss yoğunluk fonksiyonunu gösterir,

$$\phi(x; \mu, \Sigma) = \left|2\pi\Sigma\right|^{-1/2} \exp\left\{-1/2(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

Sonsal olasılık, Gauss karma regresyon yönteminin açıklanmasında önemlidir ve

$$\Pr(G = k \mid X = x) = \frac{\Pr(G = k, X = x)}{f_X(x)} = \frac{\pi_k \phi(x; \mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \phi(x; \mu_j, \Sigma_j)}$$

olarak tanımlanmaktadır (11,30).

2.7 Gauss Karma Regresyonu

Çok değişkenli parametrik olmayan regresyonda çok boyutlu verinin analizi Gauss Karma Regresyonu ile yapılmaktadır. Gauss Karma Regresyonunun temeli yoğunluk fonksiyonuna bağlı olduğundan öncelikli olarak yoğunluk fonksiyonunun tahmin edilmesi gerekir.

Yoğunluk fonksiyonu tahmin edilirken model tabanlı istatistiksel yöntemlerde kullanılan benzerlik fonksiyonu kullanılır (9,22,30).

2.7.1 Benzerlik Fonksiyonu

Benzerlik fonksiyonu,

$D = \left\{ \left(x_i, y_i \right)_{i=1}^{n} \right\}$ bağımsız ve aynı şekilde dağılan değişkenler olmak üzere

$$L(\theta; D) = \prod_{i=1}^{n} f_{X,Y}(x_i, y_i; \theta)$$

olarak tanımlanmaktadır.

Benzerlik fonksiyonu, $f_{X,Y}(x_i, y_i; \theta)$ yoğunluk modelidir.

Parametrik bir model, her (x_i, y_i) veri noktası genel bir benzerlik fonksiyonu $L(\theta; D)$ belirtir.

Doğrusal bir model için $f_{Y|X}(y_i \mid x_i) = \phi(y_i, x_i^T \beta, \sigma^2)$ koşullu yoğunluğu verildiğinde, (ϕ Gauss yoğunluk fonksiyonudur),

$f_{X,Y} = f_{Y|X} f_X$ olduğundan klasik doğrusal model için benzerlik,

$$L(\beta; D) = \prod_{i=1}^{n} \phi(y_i, x_i^T \beta, \sigma^2) f_X(x_i) = \prod_{i=1}^{n} f_X(x_i) \prod_{i=1}^{n} \phi(y_i, x_i^T \beta, \sigma^2)$$

şekilde türetilmektedir (30):

Boyut arttıkça veri seyrekliği (sparsity) durumu ile karşılaşılır. Seyreklik durumunda doğrusal olmayan örnekleri modellemek için yine benzerlik fonksiyonu yaklaşımı kullanılır ve model parametrelerinin tahmini yapılır.

Hastie ve arkadaşları (2001) yerel benzerlik fonksiyonunu Kernel fonksiyonu ile tanımlamışlardır (14),

$$l(\beta(x_0)) = \sum_{i=1}^{n} K(x_0, x_i) l(y_i, x_i^T \beta(x_0))$$

$K(x_0, x_i)$ x_o komşuluğundaki Kernel fonksiyonudur. Bu model Kernel regresyon modelinin bir çeşididir. Benzerlik fonksiyonu esnek merkezi doğrusal modellerin karmasını oluşturur, Gauss karması olasılık yoğunluk fonksiyonu da doğrusal modellerin karmasını oluşturur. Sonlu Gauss karmaları çok boyutlu verinin esnek regresyon modellemesi için kullanılan parametrik bir modeldir ve doğrusal regresyonun esnekliğini sağlamaktadır (30).

2.8 Yoğunluk Fonksiyonu

Parametrik modeller, çok boyutlu problemlerde başarılı yöntemlerdir. Ama esneklik ile ilgili varsayımlar sağlanamadığında yanlı tahminler verebilirler. Parametrik olmayan yöntemler ise esnekliği sağlar ama çok boyutlu problemlerde uygulanması sıkıntı yaratmaktadır. Bu problemi ortadan kaldırmak için bileşik yoğunluk fonksiyonu kullanılır, böylece parametrik ve parametrik olmayan yöntemlerin dezavantajları ortadan kaldırılmış olmaktadır.

Yoğunluk fonksiyonu,

$$f_{X,Y}(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial x}$$

olarak tanımlanır ve burada x ve y, $F_{x,y}$ dağılımından bağımsız ve aynı şekilde dağılan p boyutlu örneklerdir.

Regresyon fonksiyonu m(x), $f_{x,y}$ ile türetilir ve

$$m(x) = \int y f_{Y|X}(y|x) dy$$

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x,y)}{\int f_{X|Y}(x,y) dy}$$

olarak tanımlanır.

m(x) regresyon modellemede yoğunluk tahmini ile iki sınıf yoğunluğunun oranına dayalı sınıflandırıcı oluşturur. Böylece boyut artışından daha az etkilenilmektedir. Bu eşitlikten dolayı, Gauss karma regresyonu ve Gauss karma sınıflandırmasının her boyutta uygulanabilmektedir (30,43).

2.8.1 Yoğunluk Fonksiyonunun Kernel ile Tahmini

$$m(x) = E\left[Y \middle| X = x\right] = \int y f_{Y|X}(y|x) dy = \frac{\int y f_{X,Y}(x,y) dy}{\int f_{X,Y}(x,y) dy}$$

Burada m(x) regresyon fonksiyonunu ve $f_{x,y}$ yoğunluk fonksiyonunu göstermektedir. m(x) fonksiyonunu tahmin etmek için koşullu beklenti fonksiyonu tanımlanmıştır.

Ortak yoğunluk fonksiyonun tahmini iki değişkenli Kernel tarafından tahmin edilmektedir.

$$\hat{f}_{X,Y}(x,y) = \sum_{j=1}^{n} n^{-1} K_h(x - x_j) K_h(y - y_j)$$

Nadaraya Watson Kernel fonksiyonu bu eşitlikten

$$\hat{m}(x) = \frac{\int y \hat{f}_{X,Y}(x,y)dy}{\int \hat{f}_{X,Y}(x,y)dy} = \frac{\sum_{j=1}^{n} n^{-1} y_j K_h(x - x_j)}{\sum_{j=1}^{n} n^{-1} K_h(x - x_j)} = \frac{\sum_{j=1}^{n} y_j K(x, x_j)}{\sum_{j=1}^{n} K(x, x_j)}$$

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i K(x, x_i)}{\sum_{i=1}^{n} K(x, x_i)}$$

şeklinde tahmin elde edilmektedir:

Nadaraya Watson tahmincisi Kernel yoğunluk tahmincisinin formunda $f_{x,y}$ yi modelleyerek m(x) için Kernel düzgünleştirici üretir. Burada Gauss olasılık yoğunluk fonksiyonunun Kernel K tarafından tahmin edildiği görülmektedir. Karma modeller yoğunluk tahmini için kullanılırken Kernel metodunun bir çeşidi olarak düşünülmektedir (7,22,30).

$f_{x,y}$ Gauss olduğunda koşullu olasılık yoğunluk fonksiyonu $f_{Y,X}$ Gauss'dur ve regresyon modeli m(x) doğrusaldır. Gauss olasılık yoğunluk fonksiyonu $f_{x,y}$'yi modellemek için kullanılır. K bileşenli Gauss karmalarından oluşan olasılık yoğunluk fonksiyonu,

$$\hat{f}_{X,Y}(x,y) = \sum_{j=1}^{K} \pi_j \phi(x, y; \mu_j, \Sigma_j)$$

olarak gösterilmektedir (2,19,22,30):

Olasılık yoğunluk fonksiyonun sonsal regresyon fonksiyonu m(x), $m_j(x)$ doğrusal fonksiyonundan yola çıkmıştır. Gauss karma regresyon yöntemi bu fonksiyondan türetilir (14).

Gauss karma yoğunluğundan regresyon fonksiyonu oluşturulduktan sonra, karma bileşen K'lar kullanılarak, karmanın en büyük olabilirlik tahmini hesaplanır. EM (Expectation Maximization) algoritmasını kullanılarak m(x)'in olabilirlik k sayısı elde edilir.

Yoğunluk tahminin zor hesaplanabilir olmasından dolayı yoğunluk fonksiyonundan tek değişkenli regresyon hesaplanmadan, standart tek değişkenli parametrik olmayan yöntemler hesaplanabilir (22,30).

Regresyon problemini belirlemede yoğunluk tahmini problemini belirlemeden daha çok parametre vardır. Yoğunluk tahmini direkt veri modellemesinden daha zordur, çünkü yoğunluk fonksiyonunda daha çok veri hesaplaması vardır (27,30).

Veri modellemede, çok boyutlu problemler ile karşı karşıya kalındığında, yoğunluk modeli en uygun yöntemdir. Çünkü X ve Y'nin tüm istatistiksel özeti yoğunluk olabilirlik fonksiyonunda saklıdır. Çok boyutlu veri için parametrik olmayan regresyon metodu oluşturmanın uygun yolunun Gauss yoğunluk regresyonudur. Karma modeller genelde formüldeki bileşen yoğunluğunu kullanırlar. Gauss karma modeli bunların içinden en çok kullanılanlardan biridir. (14).

2.8.2 Kernel Yoğunluk Tahmincileri

Parametrik olmayan yoğunluk tahmininde, bilinmeyen f_x yoğunluğundan bir bağımsız ve aynı şekilde dağılan örnek $\{x_i\}_{i=1}^{n}$ verildiğinde Kernel tahmincisi

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

şeklinde hesaplanır.

Kernel fonksiyonu K simetrik yoğunluk fonksiyonudur. Gauss Kernel'inin özel durumu için $K(u) = \phi(u;0,1)$ dir.

Gauss Kernel yoğunluk tahmincileri,

$$\hat{f}_X(x) = \sum_{i=1}^{n} n^{-1} \phi(x; x_i, h^2)$$

gibi gösterilir.

Bu eşitliği $X \in R^p$ p boyutlu yazacak olursak,

$$\hat{f}_X(x) = \sum_{i=1}^{n} n^{-1} \phi(x; x_i, h^2 I_p)$$

şeklini alır.

Böylece eşit karma ağırlıklı n-bileşen Gauss karma yoğunluk tahmincisi elde edilmiş olur. h (bant genişliği) bu tahmincideki tek parametredir (30).

Parametrik olmayan düzgünleştiriciler merkezi yöntemlerdir. Merkezi yöntemlerde, x'in komşuluğundadır. m(x), konum özelliği olan süreklilik varsayımına dayalıdır. x_i $B(X, \delta)$ içinde her gözlemlenen y_i değeri m(x)'in bilinmeyen değerine yakındır. B(x)'e göre y_i nin ortalaması alınarak m(x) fonksiyonu,

$$\hat{m}(x) = \sum_{i:|x_i - x| < \delta} \frac{y_i}{|B(x)|} \qquad \rightarrow \qquad \hat{m}(x) = \frac{\sum_{i=1}^{n} y_i I(|x_i - x| < \delta)}{\sum_{i=1}^{n} I(|x_i - x| < \delta)}$$

şeklinde tahmin edilmiş olur (14,30).

$|B(x)|$ B(x) setinin büyüklüğüdür ve Nadaraya Watson Kernel Düzgünleştirici formunu alır,

$$\hat{m}(X) = \frac{\sum_{i=1}^{n} y_i K(x, x_i)}{\sum_{i=1}^{n} K(x, x_i)}$$

Kernel fonksiyonu, x_i gözleminin konumu belli x'e yakınlığını ölçer. Sonuçta $K(x, x_i)$ genel olarak sürekli ve x ile x_i arasındaki uzaklığa göre monoton azalan bir fonksiyondur. $K(x, x_i)$ seçmek için en çok bilinen yollardan biri K'yı tek şekilli, simetrik yoğunluk fonksiyonlu, 0 merkezli belirlemektir.

$$K(x, x_i) = K(x_i, x) = K\left(x - x_i\right)$$

Burada Gauss olasılık yoğunluk fonksiyonu Kernel fonksiyonu gibi kullanarak Nadaraya & Watson tahmincisi belirlenir.

$$\hat{m}(X) = \frac{\sum_{i=1}^{n} y_i \phi(x; x_j, h^2 Ip)}{\sum_{i=1}^{n} \phi(x; x_j, h^2 Ip)}$$

ϕ - $x \in R^p$ için çok değişkenli Gauss olasılık yoğunluk fonksiyonu.

Bant genişliği h küçük olursa, Kernel aralığını küçük olacaktır, bu da daha yakın komşuluk ve az düzgün eğri demektir. Bir boyutlu (p) için, Gauss olasılık yoğunluk fonksiyonu kolaylıkla elde edilir. Ne var ki p boyutu arttıkça algoritmaları hesaplamakta zorlaşır (30).

2.9 Karma Modelin Oluşturulması

2.9.1 Bileşen Sayısının Belirlenmesi

Karma modellerde, her karma bileşen, gizli bir küme değişkeni gibi düşünülebileceğinden küme sayısının ve küme yönteminin seçimi ayrı bir problem olarak ele alınır (25).

Bileşen sayısı belirlenirken K için rasgele ya da tahmini ve model parametrelerinin değerleri alınır. Sonra en büyük olabilirlik tahminini hesaplamak için EM algoritması kullanılır.

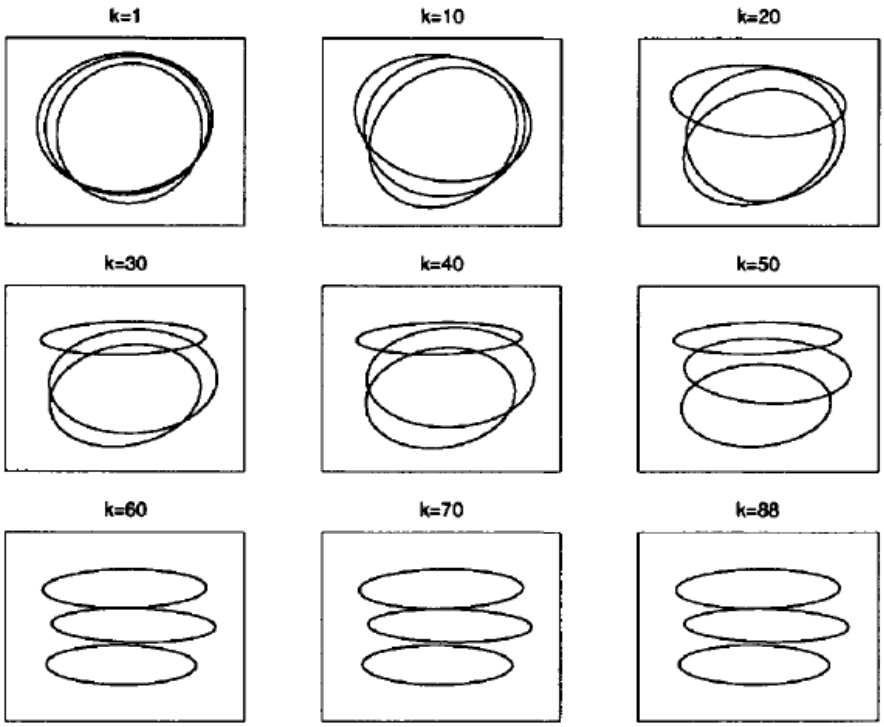
Küme sayısının seçilmesi için birçok çalışma yapılmıştır. Bu başlangıçtaki değerleri belirlemek için farklı stratejiler vardır. Parametreler için uygun başlangıç tahminleri belirlemenin en iyi yolu çok değişkenli karmaları oluştururken veriye kümeleme analizinin bazı formlarını uygulamaktır. Sonlu karma modellerde rastgele başlangıç değerlerin seçiminde rastgele alt örneklemlerinin kullanımı önerilmiştir (22).

K ortalamalar küme algoritması bu başlangıç tahminlerini elde etmek için kullanılır. Her bir j toplumu için uygun sayıda kümenin seçiminden sonra k ortalama kümeleme algoritması R_j setini tahmin için kullanılır. Daha sonra j sınıfındaki tüm gözlemler için,

$$\hat{P}(C_{jr} \left| x_{ij}, j \right.) = \begin{cases} \hat{\mu}_{jr}, \text{ } x_{ij} \text{ ye en yakın merkez ise,} & 1 \\ \text{diğer durumda} & , \quad 0 \end{cases}$$

şeklindedir.

Σ'nın başlangıç tahmini için küme içi kovaryans matrisleri birleştirilerek toplanır.



Şekil 2.1 Rastgele başlangıç değerlerini kullanan EM algoritması için farklı k'lar ile k. M adımı ile üretilen parametrelere dayalı her bileşen için asimptotik(%95) elipsoidler grafiği (22).

2.9.2 Başlangıç Değerler

Küme sayılarının belirlenmesinden sonra k ortalamalar küme algoritması çekirdek noktası, rastgele seçilir ve x_i ler gibi başlangıç değerler seçer. Hastie ve arkadaşları (1996), en iyi değerin seçilmesini önerir.

Yanlış sınıflandırma olasılığı çalışma örneğini minimize eder. Küme merkezleri gibi bir dizi rastgele farklı başlangıç dener. Başlangıç değerlerin seçiminden sonra ME tekrarlı algoritma ile π_{jr}, μ_{jr} ve Σ'nun tahminini belirlemek için kullanılır (9,22).

2.9.3 Ayırma Kuralı

J sınıfı için üyeliğinin olasılığı

$$P(G = j \mid X = x) \sim \pi_j \Pr ob(x \mid j) \sim \pi_j \sum_{r=1}^{R_j} \pi_{jr} e \exp[-D(x, \mu_{jr})]/2$$

şeklinde tahmin edilir. Ayırma kuralı $P(j|x)$ i maksimize etmek için j yi seçer, sonraki olasılıklar doğrusal olmayan eğilimlidir (2).

2.9.4 Model Seçimi

En iyi model seçimi BIC (Schwarz-1978) kriteri ile yapılır. Normal bir model parametrik olmayan bir yoğunluğu tahmin etmek için kullanıldığında BIC ile seçilen bileşen sayıları tutarlı sonuçlar verir (2,19,22,30).

Farklı serbestlik derecesinde modeller karşılaştırılırken, Bayes kriteri yorum ve hesaplama kolaylığı ile tercih nedeni olan bir yöntem haline gelmiştir. Model seçimi, en yüksek BIC değerine sahip olan model seçilerek yapılır.

$$BIC = -2 \times \log L(x; \theta^*) + d \times \log N$$

Burada $L(x; \theta^*)$ olabilirlik fonksiyonu, N örnek büyüklüğü ve d serbestlik derecesidir.

BIC kriteri kullanılarak, her bir K değeri için en büyük olabilirlik tahmini hesaplanır. En büyük olabilirlik tahmini EM algoritması ile hesaplanır. EM algoritmasının iyi bir maksimuma ulaşması için iyi bir başlangıç değerini seçilmesi önemlidir, bu yüzden farklı başlangıç değerler için yeniden başlatan değerler kullanılarak maksimuma ulaşılmaya çalışılır (5,22,30).

Gauss karma modeli benzerlik fonksiyonu sınırlı olmadığından en büyük olabilirlik tahmini hesaplanırken karma benzerliğin maksimumu yoktur. Bir bileşenin

varyansı 0'a giderse benzerlik fonksiyonu sonsuza gider. Verilen K da EM algoritması için model parametrelerinin başlangıç değerlerini oluştururken K ortalamaları kullanırlar. Bileşen sayılarının seçimi deneme yanılma şeklindedir. Her K için EM algoritmasını başlatmak üzere başlangıç parametresi sağlanır. Veri boyutu büyüdükçe bu döngüyü tekrarlamak zorlaşır (30).

2.9.5 Parametre Tahmini

2.9.5.1 EM Algoritması

Gauss Karma modelinde *bilinmeyen yer ve ölçek parametreleri* EM algoritması tarafından tahmin edilir ve modelin ayırma kuralı Bayes tarafından oluşturulur (6,10,13,16,19,25,27,33,36,43).

Sonlu karmalar içinde en çok kullanılan parametrik ifade EM algoritmasıdır. EM algoritması, benzerlik fonksiyonun en büyük yerel değerini bulur ve ilk defa Dempster ve ark. (1977) tarafından bulunmuştur. En büyük yerel değeri bulan eşitlikten hangi köklerin seçileceği, EM algoritması için alınacak başlangıç değer, bileşen sayılarının seçimi, EM algoritmasının başlıca konularıdır (9,22,43).

EM algoritması, Gauss karmalarının en büyük olabilirlik tahmincilerini (MLE) hesaplar, yani amaç benzerlik fonksiyonunu maksimize eden parametreleri bulmaktır.

EM algoritmasının işleyişi,

- Öncelikle parametreler için başlangıç tahminleri belirlenir. E adımı; ile sonraki olasılıklar hesaplanır.
- M adımı; ise ağırlıklı ortalamalar ve varyansların hesaplandığı adımdır.
- Uygun değer elde edilinceye kadar E ve M adımları tekrar eder.

şeklindedir (9,22).

2.9.5.2 İki Bileşenli Karma Model İçin EM Algoritması

Y iki normal dağılımın karması olarak modellenir.

$$Y=(1-p).Y_1+pY_2 \qquad Y_1 \sim N(\mu_1, \sigma_1^2) \;\; Y_2 = N(\mu_2, \sigma_2^2)$$

Başlangıç değerler: Rastgele başlangıç değeri belirlenir ya da karma oran tahminleri başta .5 olarak alınabilir (14).

İki bileşen olduğu için 5 tane parametre olacaktır:

$$\theta^T = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)$$

2.9.5.2.1 E adımı

EM algoritmasının E adımında her sınıf için o sınıftan örnekler alınır ve bileşenlerin sonraki olasılıkları hesaplanır.

$$\tilde{p} = \frac{p f_1(y_i|\theta)}{p f_1(y_i|\theta) \; + \; (1-p) f_2(y_i|\theta)}$$

$$Q(\theta;\theta') = \sum_i \left\{ \left[\tilde{p}_i \left(\log p' - \log\sqrt{2\pi} - \log\sigma_1' - \frac{(Y_i - \mu_1')^2}{2\sigma_1^2} \right) \right] + \left[(1-\tilde{p}_i) \left(\log(1-p') - \log\sqrt{2\pi} \right) - \log\sigma_2' - \frac{(Y_i - \mu_2')^2}{2\sigma_2^2} \right] \right\}$$

2.9.5.2.2 M adımı

Bütün parametreler için ağırlıklı en büyük olabilirlik tahmincileri,

$$p' = \frac{\sum_i \tilde{p}_{\hat{i}}}{n}$$

$$\mu_1' = \frac{\sum_i \tilde{p}_{\hat{i}} Y_i}{\sum_i \tilde{p}_{\hat{i}}}$$

$$\sigma_1' = \sqrt{\frac{\sum_i \tilde{p}_i (Y_i - \mu_1')^2}{\sum_i (\tilde{p}_i)}}$$

$$\mu_2' = \frac{\sum_i (1 - \tilde{p}_i) Y_i}{\sum_i (1 - \tilde{p}_i)}$$

$$\sigma_2' = \sqrt{\frac{\sum_i (1 - \tilde{p}_i)(Y_i - \mu_2')^2}{\sum_i (1 - \tilde{p}_i)}}$$

şeklinde hesaplanır.

2.9.5.3 k Bileşenli Karma Model İçin EM Algoritması

EM algoritmasında başlangıç değeri rastgele alınır (9,22).

2.9.5.3.1 *E Adımı*

Koşulluk olasılıklar; z_i bileşenlerinin tahminini, gözlenen x_i değerlerini ve k. tekrarda parametre tahminlerinin hesaplanmasını sağlar. z_{ir} parametresi π_r' ye bağlıdır. x_i değerleri her bir r_i grubuna aittir. Böylece E adımında, x_i değerlerine bağlı z_{ir} koşullu beklentisi hesaplanır. Bu yapılırken de k tekrarda parametreler yer değiştirir. E adımında $E_\psi^{(k)}\left[\log L_c(\psi \left| x_{obs}\right.)\right]$'nin hesaplanır.

$$E_\psi^{(k)}\left[\log L_c(\psi \left| x_{obs}\right.)\right] = E_\psi^{(k)}\left[z_{ir} \left| x_i \right.\right]$$

$$= P(z_{ir} = 1 \left| x_i \right.)$$

$$= P(x_i \in r \left| \psi^k \right.)$$

$$= \frac{\pi_r \Phi(x_{i,}\mu_r, \Sigma)}{\sum_{k=1}^{R} \pi_k \Phi(x_{i,}\mu_k, \Sigma)}$$

$P(x_i \in r|\psi^k)$ sonsal olasılığın tahminidir (22).

2.9.5.3.2 M Adımı

Bütün parametreler için ağırlıklı en büyük olabilirlik tahmincileri hesaplanır. Amaç $E_{\psi}^{(k)}\left[\log L_c(\psi|x_{obs})\right]$ maksimize eden ψ değerini seçmektir.

$$\hat{\pi}_r^{k+1} = \frac{\sum_{i=1}^{n} P(x_i \in r|\psi^k)}{\sum_{i=1}^{n}\sum_{r=1}^{R} P(x_i \in r|\psi^k)}$$

$$\hat{\mu}_r^{k+1} = \frac{\sum_{i=1}^{n} x_i P(x_i \in r|\psi^k)}{\sum_{i=1}^{n} P(x_i \in r|\psi^k)}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} P(x_i \in r|\psi^k)(x_i - \mu_r)(x_i - \mu_r)'$$

2.9.6 Aykırı Değerler

Veri setinde *aykırı değerler* olduğu durumlarda ortalama vektörü ve kovaryans matrisinin en büyük olabilirlik tahmincileri etkilenir ve ayırma kuralında gözlemlerin yanlış sınıflandırmasına neden olur (2).

2.9.7 Model Oluşturma

Gauss Karma Regresyon modelinde verinin ortak yoğunluğunu elde etmek için Gauss karma modelleri kullanılır ve her modelden koşullu yoğunluk ve regresyon fonksiyonu türetir.

Verinin ortak yoğunluk fonksiyonu kullanılarak Gauss Kernel Yoğunluk Modeli oluşturulur. Kernel Yoğunluk modelini uygulamak için tekrarlı çiftler yer değiştirme algoritması kullanılır. Böylece karma bileşenlerden elde edilen Gauss karma yoğunluk modeli oluşturulur. Her bir yoğunluk modeline karşı bir regresyon fonksiyonu denk gelir. Klasik lineer model K=1 ile gösterilir. Parametrik olmayan Kernel regresyon tahmincisi ise (K=n) şeklinde gösterilirse K= 1 den n'e değişen bir dizi regresyon modeli gösterilmiş olunur.

Tekrarlı çiftler arasındaki oran tahmin edilerek her bir yoğunluk modelinde karmaların sayıları ile indekslenen lojistik model ailesi elde edilmiş olur. Bu yönteme *Gauss Karma Sınıflandırması* (Gaussian Mixture Classification) denir (2).

Gauss karma regresyon ya da sınıflandırma modelinde, alt uzayların belirlenmesi için ileri ve geri işlem algoritmaları uygulanır. Model tabanlı yöntemler bu algoritmaları uygulamak için kullanılırlar. Gauss karma sınıflandırması karma ayırma analizine alternatif ya da tamamlayıcı bir yöntemdir. Gauss karma regresyon ve sınıflandırması, olasılık yoğunluk fonksiyonunu modellemek için kullanılır, bileşenlerin sayısını belirlemeye yarar. Gauss karma regresyonda model belirlemek için değişik yöntemler vardır (30).

2.10 Model Tabanlı Kümeleme Analizi

Model tabanlı kümeleme yöntemleri çok değişkenli verilerde grupları birbirinden ayırmada sık kullanılan uygulamalardandır. Sonlu karma modellerde her bileşen olasılık dağılımı bir kümeye karşılık gelmektedir. Kümelere dayalı ayırma probleminde bir uzaklık ya da benzerlik fonksiyonu belirlenerek veriler gruplanır. Model tabanlı ayırmada ise, her kümeye ait bir model belirlenir (5).

Kümeleme analizi olasılık modellerine dayalı olduğundan iyi uygulanan bir kümeleme yöntemi olasılık modellerinin doğru çözümlenmesini ve gelişmesini sağlar. Sonlu karma modeller kümeleme yöntemlerini uygularken her bileşen dağılımı bir kümeye karşılık gelir. Bundan sonra küme sayısının belirlenmesi problemi ve uygun kümeleme yönteminin seçimi istatistiksel model seçim problemi gibi düşünülebilir ve

bileşen sayıları ya da bileşen dağılımları karşılaştırılabilir. Eğer aykırı değer oluşturulmak isteniyorsa, başka dağılımı temsil eden bir ya da birden fazla bileşen eklenerek elde edilir. Ayırma problemlerinde, uzaklık ölçütleri olarak Mahalanobis ya da Öklit uzaklığı kullanılır. Öklit uzaklığı çok boyutlu model tipi olarak Gauss modelini ya da HMM'i kullanır. Model seçim tekniği için ise EM algoritması gibi yöntemler kullanılır (34).

Model tabanlı yöntemler içinde en çok kullanılanlar Gauss Karma Modelleridir. Model tabanlı kümeleme yönteminde algoritmanın her aşamasında birleştirme yapılırken bazı kriterler kullanılır:

Aşamalı kümeleme yöntemi (Hierarchical agglomerative clustering) da grup içi kareler toplamı ve gruplar arası tek bağlantı kümeleme yöntemi gibi (single link method) gibi bazı ölçütlere göre iki grup seçilir ve algoritmanın her aşamasında birleştirme yapılır.

Tekrarlı parçalama (Iterative partitioning) da veri noktaları bir gruptan diğerine, bir kritere göre düzeltme kalmayana kadar hareket ettirilir. Kareler ortalaması kriteri ile tekrarlı yer değiştirme yöntemi uygulanır. Bu yöntem kullanılırken çoğunlukla k-ortalamalar algoritması kullanılır (5,22).

2.10.1 En İyi Model Seçimi

Model tabanlı hiyerarşik kümeleme, gruplama ile ilgili herhangi bir bilgi olmadan bile başlangıçta iyi gruplar üretebilir. EM algoritması ile parametreler tahmin edilir. Böylece model tabanlı hiyerarşik kümeleme başlangıcı iyi kullanarak BIC ile en iyi model belirlenerek tahmin edilir(5,22,30).

2.10.2 Kovaryans Matrisinin Kümelerin Geometrik Özelliklerine Göre Belirlenmesi

Çok değişkenli normal yoğunlukların karması ile oluşturulan veriler μ_k ortalamaları etrafında merkezlenen gruplar ve kümeler ile karakterize edilir Bu ortalamalar etrafındaki noktaların yoğunluğunun arttığı bir grafik gösterir.

Bu kümelerin geometrik özellikleri; şekil, hacim, yönlendirme şeklinde ifade edilir ve Σ_k kovaryansları ile belirlenir.

- Genel durum $\Sigma_k = \lambda I$ 'dır. Bu durumda kümeler küreseldir ve eşit büyüklüktedirler.
- $\Sigma_k = \Sigma$ ise bütün kümeler küreseldir.
- Σ_k sınırlandırılmamış ise her küme farklı geometriye sahip olabilir.

$\Sigma_k = \lambda I$ için sadece bir parametre karmanın kovaryans yapısını açıklar. Veri d boyutlu ise sabit Σ_k ve sınırlandırılmamış Σ_k için d(d+1)/2, ve G(d(d+1)/2) parametreleri gereklidir.

Çok değişkenli normal karmalarda kovaryans matrisi özdeğer ayrışımı ile düzenlenir ve aşağıdaki eşitlik oluşturulur.

$$\Sigma_k = \lambda_k D_k A_k D_k^T$$

Burada D_k , k bileşenin yönünü, A_k şeklini ve λ_k hacmini gösterir. λ_k , D_k ve A_k parametrelerin bağımsız setleridir, ya her küme için ayrıdır ya da kümeler arası değişmesine izin verilmektedir.

2.10.3 Kümelerin Oluşturulması

Parametreler belli ise, kümeler kesin geometrik özellikleri paylaşırlar. Σ_k 'nın en büyük özdeğeri diğer özdeğerlerden daha büyük ise, k. küme d uzayında yakın bir çizgiye yoğunlaşır, bu k. grubun dağılımının birinci ana bileşeni olur. Benzer olarak 2 en büyük özdeğer aynı büyüklüğün ise k. küme d-uzayında bir düzleme yoğunlaşır. k. küme kabaca Σ_k 'nın en büyük ve en küçük değerleri aynı büyüklükte ise küresel olacaktır.

Bu yaklaşım genelde 3 bilinen modeli içerir.

λI eşit varyans, sınırlandırılmamış varyans

$\Sigma_k = \lambda_k I$ kümeler küresel farklı hacimli

$\Sigma_k = \lambda_k I_k$ bütün kovaryanslar karesel ama şekil büyüklük ve yönleri farklılık gösterir. Kovaryans matrisinin önerilen diğer parametrizasyonları kümeleme analizinin içeriğinde vardır. Bu sınıf içi korelasyon ya da bir faktör modelini içerir (5,9,22,33).

2.11 Yüksek Bozulma Tahmini

İstatistiksel çıkarsamalar, verilerdeki gözlemlere ve varsayımlara dayalıdır. Bu varsayımlar verilerin geldiği gerçek ortalama vektörler ve ortak kovaryans matrisine göre hesaplanır. Sınıflandırma kuralları oluşturulurken çalışma verisi tarafından tahmin edilen bilinmeyen parametreler kullanılır. Eğer veride aykırı gözlemler varsa, ortalama vektörlerinin ve kovaryans matrisinin en büyük benzerlik tahmincileri etkilenir. Bilinmeyen parametrelerin tahmini, bu düzensiz gözlemlerin aşırı etkisine bağlı olarak istikrarsız olabilir ve bozulmalara neden olur.

Yüksek bozulma tahmini yöntemi ile aykırı değerler tarafından meydana gelen çarpıklığa sağlam (robust) tahminciler üretilir. Böylece bu tip düzensiz gözlemlerin etkisini azaltılmış olur (2).

Bozulma noktası, tahminin elde edilebileceği aykırı değerlerin oranıdır, tahmincilerin sağlamlığını (robustness) belirlemek için kullanılır. Tahminin çökmeye başlaması için örnekteki en küçük kontaminasyonun oranı bozulma noktasını oluşturur. Bu durumda yerel tahmin sınırsız olabilir ya da yayılma matris tahmincisinin öz değeri keyfi küçük ya da büyük olabilir.

Doğrusal Ayırma Analizinde çökme noktaları incelenirken ortalama değiştirme yöntemi uygulanır (Mean-shift outlier model). Grup içi kovaryans determinant değerinin çökme noktası belirlenir ve bozulmanın oluşturduğu yanlılık ortadan kaldırılır.

Karma Ayırma Analizinde ise aykırı değerler varlığında EM algoritmasının M adımında, sağlam tahminciler sağlam olmayan tahmincilerle yer değiştirir.

Çökme noktası değerlendirilirken, ½ civarındaki çökme noktası yüksek çökme noktası olarak adlandırılır. Çok değişkenli yer vektörünün en büyük olabilirlik tahmini 1/n'lik bir çökme noktasına sahiptir. Sadece bir noktayı bozuk bir noktayla değiştirmek yer vektörünü sınırsız yapar. Çok değişkenli yer vektörünün ve şekil matrisinin sağlam tahmini zordur, çünkü bilinen yöntemlerin çoğu aykırı değerler 1/(1+p) den çok ise bozulacaktır (p veri boyutunu göstermektedir).

Bir p–1 boyutsal alt uzayına düşen p noktadan fazla yoksa p özellik vektörünün n büyüklüğündeki bir örneklemi, genel pozisyondadır denir. Genel pozisyon bir yüksek bozulma tahmini varsayımıdır. Çok grup ayırma analizinde bu varsayım ortak kovaryans matrisinin tekil olmadığını gösterir (2).

3 GEREÇ VE YÖNTEM

3.1- Veri Türetimi

Araştırmada türetilmiş verilerden yararlanılmıştır. Veri türetimi için R 2.7.0 istatistik paket programı (The R Project for Statistical Computing) kullanılmıştır. R programı ***http://www.r-project.org*** adresinden ücretsiz elde edilmiştir. R programının türetim ve analiz işlemlerinde kullanılabilmesi için R programlama dilinde makro programlar (paket) yazılmıştır ya da hazır R paketleri (MASS, MDA, MCLUST) kullanılarak sonuçlar alınmıştır (35-40).

Veri türetimi ve analizlerde 3 farklı adım uygulanmıştır. Birinci adımda ortalama vektörleri, kovaryans matrisleri ve grup gözlem sayıları değiştikçe ayırma yöntemlerinin doğruluk oranları arasında nasıl bir değişiklik olduğu görülmektir. İkinci adımda aynı yöntemlere veri sayısının artışı ile birlikte büyük sayıda grup gözlem sayılarının elde edilerek, ayırma yöntemlerinin doğruluk oranları arasında nasıl bir değişiklik olduğu görülmektir. Üçüncü adımda, Gauss Karma modeline Model Tabanlı Kümeleme yöntemi uygulanarak en iyi modelin belirlenmesi ve Poisson gürültü uygulandığında ayırma yönteminin nasıl işlediği görülmektedir.

3.1.1 Veri Türetiminde Birinci Adım

Veri türetim algoritması

1. Başlangıçta Karesel, Lineer, Karma, Mars ve Bruto değerleri 0 olarak atanır.

2. Ortalama vektörleri, kovaryans matrisleri ve n sayıları belli setler normal dağılımdan türetilir.

3. Alt alta birleştirilerek veri setleri oluşturulur.

4. Doğrusal ve karesel ayırma analizinin sınıflara göre tahmin tablosu oluşturulur doğruluk oranı hesaplanır.

5. Karma, MARS ve BRUTO analizlerinin ayırma analizi özet istatistiklerini veren konfüzyon matrisi hesaplanır ve hata oranları belirlenir.

6. Döngü 2'den itibaren 1000 kez tekrarlanır.

7. Döngünün her tekrarında doğrusal ve karesel ayırma analizlerini toplanarak döngü sonunda ortalamaları alınır doğruluk oranları belirlenir.

8. Karma Ayırma Analizi, MARS ve BRUTO yöntemlerinin sonuçları ise konfüzyon matrisi oluşturularak hata oranları belirlenir ve bu hata oranlarından doğruluk oranları elde edilir.

Veri türetim parametreleri

Bir bağımlı değişken ve 4 bağımsız değişken içeren ortalama vektörleri ve kovaryans matrisleri farklı çok değişkenli normal dağılımdan veri matrisleri türetilerek veri türetimi yapılmıştır. Türetimler oluşturulurken grup n sayıları 50, 100, 250, 500, 1000, 3000 şeklinde seçilmiştir. Veri setleri ise bu grupların birleştirilmesi ile elde edilmiştir ve veri setlerinde her bir değişkenin toplam sayısında (3 grup olduğundan) 9000'e kadar çıkarılmıştır. Böylece toplam bağımsız değişken gözlem sayısında 36000'e kadar ulaşılmıştır. Veri setlerinin oluşturulmasında kullanılan 12 farklı kovaryans matrisi ve değişik aralıklarda seçilen ortalama vektörleri sırası ile aşağıdaki gibidir. 1. Adımda 1000 tekrar kullanılmıştır.

Birinci adım veri türetiminde kullanılan kovaryans matrisleri aşağıdaki gibidir:

$$\Sigma 1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \Sigma 4 = \begin{bmatrix} 1 & .25 & .25 & .25 \\ .25 & 1 & .25 & .25 \\ .25 & .25 & 1 & .25 \\ .25 & .25 & .25 & 1 \end{bmatrix} \quad \Sigma 7 = \begin{bmatrix} 1 & .75 & .75 & .75 \\ .75 & 1 & .75 & .75 \\ .75 & .75 & 1 & .75 \\ .75 & .75 & .75 & 1 \end{bmatrix} \quad \Sigma 10 = \begin{bmatrix} 1 & .90 & .90 & .90 \\ .90 & 1 & .90 & .90 \\ .90 & .90 & 1 & .90 \\ .90 & .90 & .90 & 1 \end{bmatrix}$$

$$\Sigma 2 = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \quad \Sigma 5 = \begin{bmatrix} 3 & .25 & .25 & .25 \\ .25 & 3 & .25 & .25 \\ .25 & .25 & 3 & .25 \\ .25 & .25 & .25 & 3 \end{bmatrix} \quad \Sigma 8 = \begin{bmatrix} 3 & .75 & .75 & .75 \\ .75 & 3 & .75 & .75 \\ .75 & .75 & 3 & .75 \\ .75 & .75 & .75 & 3 \end{bmatrix} \quad \Sigma 11 = \begin{bmatrix} 3 & .90 & .90 & .90 \\ .90 & 3 & .90 & .90 \\ .90 & .90 & 3 & .90 \\ .90 & .90 & .90 & 3 \end{bmatrix}$$

$$\Sigma 3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \Sigma 6 = \begin{bmatrix} 10 & .25 & .25 & .25 \\ .25 & 10 & .25 & .25 \\ .25 & .25 & 10 & .25 \\ .25 & .25 & .25 & 10 \end{bmatrix} \quad \Sigma 9 = \begin{bmatrix} 10 & .75 & .75 & .75 \\ .75 & 10 & .75 & .75 \\ .75 & .75 & 10 & .75 \\ .75 & .75 & .75 & 10 \end{bmatrix} \quad \Sigma 12 = \begin{bmatrix} 10 & .90 & .90 & .90 \\ .90 & 10 & .90 & .90 \\ .90 & .90 & 10 & .90 \\ .90 & .90 & .90 & 10 \end{bmatrix}$$

Birinci adım veri türetiminde kullanılan ortalama vektörleri aşağıdaki gibidir:

μ1: (0,0,0,0)
μ2: (0.5,0.5,0.5,0.5)
μ3:(1,1,1,1)

μ4:(0,0,0,0)
μ5:(1,1,1,1)
μ6:(2,2,2,2)

μ7: (0,0,0,0)
μ8: (2,2,2,2)
μ9:(4,4,4,4)

μ10:(0,0,0,0)
μ11:(0.5,1,1.5,2)
μ12:(2.5,3,3.5,4)

Veri Analizi Yöntemleri

Y tahmin, X_i açıklayıcı çok değişkenli normal dağılan değişkenler olmak üzere, Analizlerde Doğrusal Ayırma Analizi, Karesel Ayırma Analizi, Esnek Ayırma Analizi yöntemlerinden MARS ve BRUTO ile Karma Ayırma Analizi kullanılmıştır. Doğrusal ve Karesel ayırma analizinde MASS paketleri, Esnek ve Karma Ayırma Analizinde MDA paketleri kullanılmıştır. Şekil 3.3'de oluşturulan veri setlerinden bir örnek görülmektedir.

MARS yönteminde maksimum etkileşim derecesini belirleyen 2 tamsayı, 1 (varsayılan derece) ve 2 seçilmiştir. Trevor Hastie and Robert Tibshirani'nin MARS algoritması Friedman'ın MARS kodlaması ile benzer sonuçlar vermesine rağmen aynı kodlama kullanılmamıştır. BRUTO yönteminde ise model seçimi için tekrarlı işlem algoritmasının her adımında kullanılan GCV kriteri yaklaşımı kullanılmıştır (39).

Karma Ayrıma Analizinde model tahmini,

$$P(X = x, Z = k) = a_k f_k(x) = a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(x | \mu_{kr}, \Sigma)$$

a_k , k sınıfının başlangıç olasılığıdır. Bilinmeyen parametreler EM algoritması ile tahmin edilir.

MARS, BRUTO ve MDA konfüzyon matrisi hesaplayarak ayırma analizindeki hata oranlarını verir.

Veri setlerinden bir örnek:

	I	II	III	IV	V	grup
[1,]	-0.595442470	1.28732083	-0.80643971	1.459721679	1
[2,]	-0.535203680	0.90607354	-0.70158405	0.612204831	1
[3,]	-1.825022159	-0.12739726	-0.43665782	0.520707999	1
.	.	.	.	.	.
.	.	.	.	.	.
[51,]	0.554526363	-2.03203957	-0.23437572	-0.122780971	2
[52,]	0.155689482	0.21035346	-1.49241905	-0.076156679	2
[53,]	0.343583085	1.06212973	1.64951534	0.321043430	2
.	.	.	.	.	.
.	.	.	.	.	.
[101,]	2.734219887	1.17918380	1.37555951	1.331494481	3
[102,]	0.018123559	-0.07998244	0.94583222	1.278495210	3
[103,]	0.584033621	1.53333590	-1.07160360	-0.819955151	3
.	.	.	.	.	.
.	.	.	.	.	.
[150,]	-0.17897636	2.992459680	0.115194141	0.236130550	3

3.1.2 Veri Türetiminde İkinci Adım

Veri türetim algoritması 1. Adımdaki gibidir.

Türetim parametreleri, 3 ayrı (I, II, III) şekilde oluşturulmuştur. Bunlarda kendi içlerinde 10'lu 15'li ve 20'li olmak üzere yine veri setlerinin alt alta eklenmesi ile gruplandırılmıştır.

İkinci adım veri türetiminde kullanılan ortalama vektörleri ve kovaryans matrisleri:

μ1:(0,0,0,0)	μ1:(0,0,0,0)	μ1:(0,0)
μ2:(0.5,0.5,0.5,0.5)	μ2:(2,2,2,2)	μ2:(2,2)
μ3:(1,1,1,1)	μ3:(4,4,4,4)	μ3:(4,4)
μ4:(1.5,1.5,1.5,1.5)	μ4:(6,6,6,6)	μ4:(6,6)
μ5:(2,2,2,2)	μ5:(8,8,8,8)	μ5:(8,8)
μ6:(2.5,2.5,2.5,2.5)	μ6:(10,10,10,10)	μ6:(10,10)
μ7:(3,3,3,3)	μ7:(12,12,12,12)	μ7:(12,12)
μ8:(3.5,3.5,3.5,3.5)	μ8:(14,14,14,14)	μ8:(14,14)
μ9:(4,4,4,4)	μ9:(16,16,16,16)	μ9:(16,16)
μ10:(4.5,4.5,4.5,4.5)	μ10:(18,18,18,18)	μ10:(18,18)
μ11:(5,5,5,5)	μ11:(20,20,20,20)	μ11:(20,20)
μ12:(5.5,5.5,5.5,5.5)	μ12:(22,22,22,22)	μ12:(22,22)
μ13:(6,6,6,6)	μ13:(24,24,24,24)	μ13:(24,24)
μ14:(6.5,6.5,6.5,6.5)	μ14:(26,26,26,26)	μ14:(26,26)
μ15:(7,7,7,7)	μ15:(28,28,28,28)	μ15:(28,28)
μ16:(7.5,7.5,7.5,7.5)	μ16:(30,30,30,30)	μ16:(30,30)
μ17:(8,8,8,8)	μ17:(32,32,32,32)	μ17:(32,32)
μ18:(8.5,8.5,8.5,8.5)	μ18:(34,34,34,34)	μ18:(34,34)
μ19:(9,9,9,9)	μ19:(36,36,36,36)	μ19:(36,36)
μ20:(9.5,9.5,9.5,9.5)	μ20:(38,38,38,38)	μ20:(38,38)
(I)	(II)	(III)

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{(a)}$$

Kovaryans matrisleri varyansların birer birim artırılması ile elde edildi (b)

Çok boyutlu gözlem türetimleri oluşturulurken grup n sayıları 50 ve 500 şeklinde seçilmiştir. Veri setleri ise bu grupların birleştirilmesi ile elde edilmiştir ve veri setlerinde her bir değişkenin toplam sayısında (50×10, 50×15, 50×20) ve (500×10, 500×15, 500×20)'e kadar çıkılmıştır. Bu türetimler I, II, III ortalama vektörleri için ayrı ayrı yapılmıştır. 4 değişken ve 2 değişken kullanılmıştır. Böylece gözlem sayısında en fazla 40000 (500×20×4)'e kadar çıkılmıştır. Veri setlerinin oluşturulmasında aynı kovaryans matrisleri ve farklı kovaryans matrisleri (varyansların 1'er br artırılması ile elde edildi) kullanılmıştır. Ortalamalar .5 br aralıklı ve 2 br aralıklı seçilmiştir. Burada R programının MASS, MDA paketi kullanılmıştır.

3.1.3 Veri Türetiminde Üçüncü Adım

μ1=(0,0,0,0), μ2=(3,3,3,3), μ3=(6,6,6,6) ortalamalı, 4 değişkenli 50'şer verili 3 gruplu veri setine Model tabanlı kümeleme analizi uygulanmıştır.

En iyi model BIC değerine bağlı olarak seçilmiştir.

$BIC = -2 \times \log L(x;\theta) + d \times \log N$

$L(x;\theta^*)$ benzerlik fonksiyonudur. N örnek büyüklüğü d serbestlik derecesidir. Gauss karma modeli için

$d = K \times (1 + p + p \times (p+1)/2) - 1 = 0(Kp^2)$ (13,14,27).

Veriye 500 verilik Poisson Noise uygulanmıştır. Sonuçlar grafikle (Şekil 4.4) gösterilmiştir. Şekil 3.5.'de Model Tabanlı Veri analizinde kullanılan algoritmadan bir örnek gösterilmiştir. Burada R programının MASS (veri türetiminde) ve MCLUST (sınıflandırma tablosunun elde edilmesinde, en iyi BIC değerlerinin elde edilmesinde, en iyi modelin seçilmesinde ve veriye Poisson gürültü uygulanmasında) paketi kullanılmıştır. Veri modelinin kovaryans yapısına göre en iyi modeli şekil 3.6'da oluşturulmuş belirleyicilere (EEE, EEI, EII vb.) göre kendimiz belirleyebiliriz ve böylece en iyi modelin Bayes kriteri değerini elde edebiliriz.

Model Tabanlı Veri analizinde kullanılan algoritmadan bir örnek

```
mu<-c(0,0,0,0)
Sigma <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
x1<-mvrnorm(n =50 , mu, Sigma, tol = 1e-6, empirical = FALSE)
mu<-c(3,3,3,3)
Sigma <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
x2<-mvrnorm(n =50, mu, Sigma, tol = 1e-6, empirical = FALSE)
mu<-c(6,6,6,6)
Sigma <- matrix(c(1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1),4,4)
x3<-mvrnorm(n =50, mu, Sigma, tol = 1e-6, empirical = FALSE)
g<-rep(1:3, c(50,50,50))
y1<-matrix(c(x1[,1], x2[,1], x3[,1]))
y2<-matrix(c(x1[,2], x2[,2], x3[,2]))
y3<-matrix(c(x1[,3], x2[,3], x3[,3]))
y4<-matrix(c(x1[,4], x2[,4], x3[,4]))
y<-matrix(c(y1, y2, y3, y4, g),nrow=150)
colnames(y)<- c("v1", "v2", "v3", "v4", "grup")

yBIC <- mclustBIC(y[,-5])
ySummary3 <- summary(yBIC, y[,-5], G = 1:6, modelNames = c("EII", "EEI", "EEE"))
ySummary3

coordProj( data = y[,-5], dimens = c(2,4), what = "classification",
parameters = ySummary3$parameters, z = ySummary3$z)
 coordProj( data = y[,-5], dimens = c(2,4), what = "uncertainty",
parameters = ySummary3$parameters, z = ySummary3$z)
 coordProj( data = y[,-5], dimens = c(2,4), what = "errors",
parameters = ySummary3$parameters, z = ySummary3$z, truth = y[,5])

yy<- y[,-5]

b <- apply( yy, 2, range)
nNoise <- 500
set.seed(0)
poissonNoise <- apply(b, 2, function(x, n)
runif(n, min = min(x)-.1, max = max(x)+.1), n = nNoise)
yyNdata <- rbind(yy, poissonNoise)
set.seed(0)

yyNoiseInit <- sample(c(TRUE,FALSE),size=nrow(yy)+nNoise,
replace=TRUE,prob=c(3,1))

yyNbic <- mclustBIC(yyNdata,
initialization = list(noise = yyNoiseInit))

yyNsummary <- summary(yyNbic, yyNdata)
yyNsummary
coordProj( data = yyNdata, dimens = c(2,4), what = "classification",
parameters = yyNsummary$parameters, z = yyNsummary $z)
```

Belirleyici	Model	HC	EM	Dağılım	Hacim	Şekil	Oryantasyon
E		•	•	Tek	Eşit		

				değişkenli			
V		•	•	Tek değişkenli	Değişken		
EII	λI	•	•	Küresel	Eşit	Eşit	NA
VII	$\lambda_k I$	•	•	Küresel	Değişken	Eşit	NA
EEI	λA		•	Köşegen	Eşit	Eşit	Koordinat ekseni
VEI	$\lambda_k A$		•	Köşegen	Değişken	Eşit	Koordinat ekseni
EVI	λA_k		•	Köşegen	Eşit	Değişken	Koordinat ekseni
VVI	$\lambda_k A_k$		•	Köşegen	Değişken	Değişken	Koordinat ekseni
EEE	λDAD^T	•	•	Elipsoidal	Eşit	Eşit	Eşit
EEV	$\lambda D_k A D_k^T$		•	Elipsoidal	Eşit	Eşit	Değişken
VEV	$\lambda_k D_k A D_k^T$		•	Elipsoidal	Değişken	Eşit	Değişken
VVV	$\lambda_k D_k A_k D_k^T$	•	•	Elipsoidal	Değişken	Değişken	Değişken

Çok boyutlu veride MCLUST ve EM algoritması için kullanılan kovaryans parametrizasyonu (5)

4-BULGULAR

4.1. Birinci adım

Ortalaması ve kovaryans matrisleri belirtilen 50'şer değişkenden oluşturulmuş 3 gruplu 50*3*4 verili gözlem setinin analiz sonuçları çizelge 4.1'de verilmiştir.

Kovaryans matrisi ve ortalama vektörüne göre doğruluk oranının değişmediği çizelge 4.1'de görülmektedir. Farklı ortalama vektörleri ile aynı kovaryans matrisine sahip ilk sütun verileri için çoğunlukla en yüksek doğruluk oranını karma ayırma analizi vermektedir (ilk sütun verileri LDA:.61 QDA:.63 FDAM1:.59 FDAM2:.58 FDAB:.57 MDA:.65). Ortalama vektörleri aynı iken Kovaryans matrisleri arasındaki sayısal fark arttıkça doğruluk oranları çoğunlukla düşmektedir. Kovaryans matrislerinin aynı ya da farklı olduğu durumlarda ortalama vektörleri arasındaki sayısal farklılık arttıkça doğruluk oranları çoğunlukla da artış göstermektedir ($MDA_{1,2,3}$=.65, .73, .93)

Çizelge 4.1- Ortalaması ve kovaryans matrisleri belirtilen 50'şer değişkenden oluşturulmuş 3 gruplu 50*3*4 verili gözlem seti analiz sonuçları

	μ1: (0,0,0,0) μ2: .5,.5,.5,.5) μ3:(1,1,1,1)			μ1:(0,0,0,0) μ2:(1,1,1,1) μ3:(2,2,2,2)			μ1: (0,0,0,0) μ2: (2,2,2,2) μ3:(4,4,4,4)			μ1:(0,0,0,0) μ2:(.5,1,1.5,2) μ3:(2.5,3,3.5,4)		
	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
LDA	.61	.54	.51	.80	.68	.60	.97	.87	.74	.86	.75	.70
QDA	.63	.66	.73	.81	.74	.75	.98	.89	.82	.86	.84	.81
FDAM1	.59	.66	.72	.83	.74	.81	.96	.89	.77	.82	.73	.78
FDAM2	.58	.65	.75	.83	.70	.76	.97	.90	.82	.79	.84	.78
FDAB	.57	.67	.63	.81	.75	.75	.96	.90	.78	.82	.76	.76
MDA	.65	.67	.65	.87	.75	.71	.96	.91	.75	.83	.78	.73
	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4
	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4
	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6
LDA	.55	.52	.50	.71	.64	.58	.92	.84	.72	.88	.79	.68
QDA	.58	.65	.73	.73	.73	.76	.92	.87	.82	.89	.84	.81
FDAM1	.59	.62	.65	.69	.74	.71	.92	.78	.75	.89	.78	.75
FDAM2	.59	.61	.63	.71	.74	.75	.93	.83	.73	.89	.82	.75
FDAB	.61	.54	.67	.74	.68	.78	.92	.81	.76	.87	.77	.71
MDA	.65	.62	.64	.73	.71	.70	.93	.86	.77	.88	.83	.75
	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
LDA	.01	.00	.50	.63	.59	.55	.83	.78	.69	.86	.75	.66
QDA	.01	.00	.75	.65	.75	.77	.84	.85	.82	.86	.85	.82
FDAM1	.47	.49	.62	.64	.65	.75	.85	.77	.81	.85	.80	.76
FDAM2	.47	.49	.61	.65	.70	.69	.87	.77	.82	.89	.81	.79
FDAB	.49	.47	.62	.64	.63	.72	.86	.76	.81	.85	.79	.76
MDA	.52	.60	.65	.67	.73	.67	.85	.79	.75	.86	.82	.76
	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
LDA	.00	.00	.50	.61	.59	.55	.81	.76	.68	.89	.75	.65
QDA	.01	.00	.76	.64	.78	.77	.82	.86	.83	.89	.86	.83
FDAM1	.49	.41	.67	.63	.63	.69	.73	.71	.78	.87	.71	.76
FDAM2	.45	.44	.67	.62	.75	.65	.71	.82	.77	.91	.80	.77
FDAB	.48	.42	.67	.64	.65	.68	.75	.74	.78	.88	.69	.75
MDA	.54	.53	.65	.65	.71	.71	.79	.80	.76	.85	.85	.75

Çizelgelerde μ'ler örnek ortalama vektörlerini, Σ'lar ise kovaryans matrislerini göstermektedir. İlk sütunda zigzag şeklinde kenarlıkları oluşturulmuş olan (.61, .63, .59, .58, .57, .65) değerleri μ1: (0,0,0,0), μ2: (.5,.5,.5,.5), μ3:(1,1,1,1) ortalama vektörlü; (Σ1, Σ1, Σ1) kovaryans matrisli, (50*3*4 birimlik) veri setinin 6 farklı karma analizine göre (LDA, QDA, FDAM1, FDAM2, FDAB, MDA) doğruluk oranlarını göstermektedir. İlk doğruluk oranları tablosunun MDA satırında çift çizgili kenarlıkla gösterilmiş hücredeki (.87, .75. .71) değeri, μ1: (0,0,0,0), μ2: (.1,.1,.1,.1), μ3:(2,2,2,2) ortalama vektörlü (Σ1, Σ1, Σ1) kovaryans matrisine sahip 50*3*4 verili gözlem setinin, karma ayırma analizine göre MDA doğruluk oranı sonucunu vermektedir.

İkinci doğruluk oranları tablosunun MDA satırında çift zigzaglı kenarlıkla gösterilmiş ayrı hücrelerdeki 3 değer ise; karma ayırma analizi doğruluk oranlarının kovaryans matrisi aynı olan (Σ1, Σ1, Σ1) ama ortalama vektörleri değişen 3 veri setine göre nasıl bir değişim gösterdiğini açıklamak üzere işaretlenmiştir.

Ortalaması ve kovaryans matrisleri belirtilen 100'er değişkenden oluşturulmuş 3 gruplu 100*3*4 verili gözlem setinin analiz sonuçları çizelge 4.2'de gösterilmiştir. μ'ler örnek ortalamaları, ise kovaryans matrislerini göstermektedir.

Kovaryans matrisi ve ortalama vektörüne göre doğruluk oranı değiştiği çizelge 4.2 de görülmektedir. Farklı ortalama vektörleri ile aynı kovaryans matrisine sahip veriler için çoğunlukla en yüksek doğruluk oranını karma ayırma analizi vermektedir. Ortalama vektörleri aynı iken Kovaryans matrisleri arasındaki sayısal fark arttıkça doğruluk oranları genelde düşmektedir. Kovaryans matrislerinin aynı ya da farklı olduğu durumlarda ortalama vektörleri arasındaki sayısal farklılık arttıkça doğruluk oranları artış göstermektedir.

Çizelge 4.2. Ortalaması ve kovaryans matrisleri belirtilen 100'er değişkenden oluşturulmuş 3 gruplu 100*3*4 verili gözlem seti analiz sonuçları

	μ1: (0,0,0,0) μ2: .5,.5,.5,.5) μ3:(1,1,1,1)			μ1:(0,0,0,0) μ2:(1,1,1,1) μ3:(2,2,2,2)			μ1: (0,0,0,0) μ2: (2,2,2,2) μ3:(4,4,4,4)			μ1:(0,0,0,0) μ2:(.5,1,1.5,2) μ3:(2.5,3,3.5,4)		
	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
LDA	.60	.52	.49	.79	.67	.58	.97	.87	.73	.93	.82	.69
QDA	.61	.63	.70	.80	.73	.73	.97	.89	.81	.93	.85	.80
FDAM1	.63	.60	.73	.83	.74	.76	.98	.87	.78	.91	.85	.81
FDAM2	.61	.63	.68	.83	.72	.78	.97	.90	.80	.90	.88	.78
FDAB	.64	.60	.71	.81	.69	.75	.98	.87	.79	.92	.83	.81
MDA	.64	.60	.65	.83	.68	.68	.98	.86	.74	.93	.87	.72
	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4
	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4
	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6
LDA	.54	.50	.48	.71	.63	.56	.92	.83	.71	.88	.78	.67
QDA	.56	.63	.70	.72	.71	.73	.92	.86	.81	.88	.83	.80
FDAM1	.52	.62	.71	.69	.70	.68	.94	.81	.77	.89	.81	.79
FDAM2	.53	.60	.71	.71	.71	.78	.94	.83	.79	.91	.83	.80
FDAB	.51	.61	.73	.69	.72	.71	.95	.80	.79	.88	.77	.79
MDA	.53	.59	.66	.73	.68	.67	.94	.81	.75	.89	.81	.71
	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
LDA	.49	.02	.47	.62	.58	.54	.83	.77	.68	.85	.74	.64
QDA	.51	.03	.72	.63	.73	.75	.83	.84	.81	.86	.84	.81
FDAM1	.48	.41	.66	.62	.59	.69	.79	.79	.79	.88	.76	.74
FDAM2	.48	.39	.66	.61	.69	.71	.78	.82	.78	.88	.81	.77
FDAB	.46	.44	.66	.62	.60	.67	.77	.79	.79	.87	.77	.74
MDA	.52	.46	.63	.61	.64	.65	.78	.82	.76	.88	.81	.73
	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
LDA	.48	.48	.47	.61	.57	.61	.81	.76	.67	.88	.74	.64
QDA	.51	.72	.73	.62	.76	.74	.81	.85	.82	.89	.85	.82
FDAM1	.47	.57	.70	.58	.60	.65	.82	.79	.77	.88	.79	.75
FDAM2	.45	.66	.69	.57	.69	.68	.80	.81	.79	.89	.85	.77
FDAB	.47	.54	.66	.58	.59	.66	.81	.80	.81	.88	.79	.77
MDA	.50	.62	.61	.60	.65	.65	.79	.81	.73	.88	.80	.72

Ortalaması ve kovaryans matrisleri belirtilen 250'şer değişkenden oluşturulmuş 3 gruplu 250*3*4 verili gözlem seti analiz sonuçları çizelge 4.3.'de verilmiştir.

Çizelge 4.3'de görüldüğü gibi kovaryans matrisi ve ortalama vektörüne göre doğruluk oranı değişmektedir. Farklı ortalama vektörleri ile aynı kovaryans matrisine sahip verilerin doğruluk oranlarının birbirine yaklaştığı görülmektedir. Ortalama vektörleri aynı iken Kovaryans matrisleri arasındaki sayısal fark arttıkça doğruluk oranları genelde düşmektedir. Kovaryans matrislerinin aynı ya da farklı olduğu durumlarda ortalama vektörleri arasındaki sayısal farklılık arttıkça doğruluk oranları artış göstermektedir.

Ortalaması ve kovaryans matrisleri belirtilen 500'er değişkenden oluşturulmuş 3 gruplu 500*3*4 verili gözlem seti analiz sonuçları Çizelge 4.4'de verilmiştir.

Çizelge 4.4'de görüldüğü gibi kovaryans matrisi ve ortalama vektörüne göre doğruluk oranı değişmektedir. Farklı ortalama vektörleri ile aynı kovaryans matrisine sahip verilerin doğruluk oranlarının birbirine yaklaştığı görülmektedir. Ortalama vektörleri aynı iken Kovaryans matrisleri arasındaki sayısal fark arttıkça doğruluk oranları çoğunlukla düşmektedir. Kovaryans matrislerinin aynı ya da farklı olduğu durumlarda ortalama vektörleri arasındaki sayısal farklılık arttıkça doğruluk oranları artış göstermektedir.

Çizelge 4.3. Ortalaması ve kovaryans matrisleri belirtilen 250'şer değişkenden oluşturulmuş 3 gruplu 250*3*4verili gözlem seti analiz sonuçları

	μ1: (0,0,0,0) μ2: .5,.5,.5,.5) μ3:(1,1,1,1)			μ1:(0,0,0,0) μ2:(1,1,1,1) μ3:(2,2,2,2)			μ1: (0,0,0,0) μ2: (2,2,2,2) μ3:(4,4,4,4)			μ1:(0,0,0,0) μ2:(.5,1,1.5,2) μ3:(2.5,3,3.5,4)		
	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
LDA	.59	.52	.48	.79	.67	.58	.97	.87	.73	.93	.82	.69
QDA	.60	.62	.68	.79	.72	.72	.97	.89	.80	.93	.85	.80
FDAM1	.62	.63	.68	.79	.73	.73	.97	.86	.78	.92	.83	.77
FDAM2	.61	.63	.67	.81	.70	.72	.97	.87	.80	.93	.83	.79
FDAB	.62	.63	.67	.79	.73	.73	.97	.86	.79	.93	.82	.78
MDA	.61	.60	.61	.80	.71	.68	.97	.87	.75	.93	.83	.72
	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4
	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4
	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6
LDA	.54	.49	.46	.70	.63	.69	.91	.83	.71	.87	.78	.66
QDA	.54	.61	.68	.71	.70	.74	.91	.85	.80	.87	.83	.79
FDAM1	.52	.59	.68	.70	.67	.75	.93	.84	.78	.88	.81	.76
FDAM2	.52	.59	.66	.71	.69	.76	.93	.87	.79	.89	.80	.77
FDAB	.55	.59	.70	.71	.67	.75	.93	.85	.78	.88	.79	.77
MDA	.55	.56	.61	.71	.65	.75	.93	.85	.73	.89	.79	.71
	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
LDA	.48	.45	.45	.62	.58	.52	.82	.77	.67	.85	.74	.64
QDA	.49	.47	.71	.62	.72	.74	.83	.84	.80	.85	.83	.80
FDAM1	.48	.46	.62	.62	.61	.67	.85	.78	.76	.85	.75	.74
FDAM2	.49	.45	.65	.64	.67	.68	.85	.80	.77	.86	.77	.77
FDAB	.50	.46	.63	.62	.60	.68	.84	.77	.77	.85	.74	.75
MDA	.48	.49	.59	.62	.64	.61	.85	.78	.73	.86	.73	.71
	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
LDA	.48	.47	.45	.60	.57	.60	.80	.75	.67	.88	.73	.63
QDA	.49	.70	.72	.61	.75	.73	.81	.85	.81	.88	.85	.81
FDAM1	.48	.54	.63	.59	.62	.69	.80	.78	.73	.86	.75	.74
FDAM2	.45	.62	.65	.61	.70	.70	.80	.80	.75	.86	.79	.77
FDAB	.48	.53	.65	.60	.61	.69	.80	.78	.73	.86	.75	.76
MDA	.49	.58	.58	.60	.63	.67	.80	.80	.69	.87	.73	.72

Çizelge 4.4. Ortalaması ve kovaryans matrisleri belirtilen 500'er değişkenden oluşturulmuş 3 gruplu 500*3*4 verili gözlem seti analiz sonuçları

	μ1: (0,0,0,0) μ2: .5,.5,.5,.5) μ3:(1,1,1,1)			μ1:(0,0,0,0) μ2:(1,1,1,1) μ3:(2,2,2,2)			μ1: (0,0,0,0) μ2: (2,2,2,2) μ3:(4,4,4,4)			μ1:(0,0,0,0) μ2:(.5,1,1.5,2) μ3:(2.5,3,3.5,4)		
	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
LDA	.59	.51	.47	.79	.67	.57	.97	.87	.73	.93	.82	.69
QDA	.59	.61	.67	.79	.71	.72	.97	.88	.80	.93	.85	.79
FDAM1	.60	.60	.67	.77	.70	.69	.97	.85	.75	.92	.83	.78
FDAM2	.61	.59	.68	.78	.71	.69	.96	.86	.77	.92	.83	.79
FDAB	.60	.61	.68	.76	.71	.70	.97	.86	.77	.92	.83	.79
MDA	.60	.58	.61	.77	.69	.61	.97	.86	.71	.93	.82	.73
	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4
	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4
	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6
LDA	.53	.49	.45	.70	.62	.68	.91	.83	.71	.87	.78	.66
QDA	.54	.61	.68	.70	.70	.74	.91	.85	.80	.87	.82	.79
FDAM1	.53	.58	.65	.67	.68	.72	.91	.84	.79	.86	.80	.77
FDAM2	.53	.60	.67	.68	.69	.73	.92	.85	.79	.86	.81	.79
FDAB	.54	.60	.66	.68	.69	.73	.91	.84	.79	.86	.80	.77
MDA	.54	.56	.58	.68	.66	.73	.92	.84	.74	.87	.81	.70
	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
LDA	.48	.45	.44	.62	.57	.52	.82	.76	.67	.85	.74	.64
QDA	.49	.46	.70	.62	.72	.73	.82	.84	.80	.85	.83	.80
FDAM1	.49	.44	.61	.63	.61	.66	.81	.77	.75	.86	.78	.74
FDAM2	.48	.44	.65	.63	.67	.68	.81	.78	.76	.85	.80	.76
FDAB	.48	.45	.62	.63	.62	.66	.81	.77	.76	.86	.78	.75
MDA	.49	.44	.57	.63	.62	.60	.81	.77	.72	.87	.77	.70
	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
LDA	.47	.46	.44	.60	.56	.60	.80	.75	.66	.88	.73	.63
QDA	.48	.70	.71	.60	.75	.73	.80	.85	.81	.88	.85	.81
FDAM1	.46	.56	.63	.63	.65	.66	.80	.77	.75	.88	.76	.75
FDAM2	.46	.61	.65	.62	.71	.69	.81	.80	.76	.88	.80	.77
FDAB	.46	.56	.64	.61	.64	.68	.80	.77	.75	.88	.76	.76
MDA	.46	.57	.57	.62	.65	.67	.80	.76	.70	.89	.77	.68

Ortalaması ve kovaryans matrisleri belirtilen 1000'er değişkenden oluşturulmuş 3 gruplu 1000*3*4 verili gözlem seti analiz sonuçları Çizelge 4.5'de verilmiştir.

Çizelge 4.5'de görüldüğü gibi kovaryans matrisi ve ortalama vektörüne göre doğruluk oranı değişmektedir. Farklı ortalama vektörleri ile aynı kovaryans matrisine sahip verilerin doğruluk oranlarının birbirine yaklaştığı görülmektedir. Ortalama vektörleri aynı iken Kovaryans matrisleri arasındaki sayısal fark arttıkça doğruluk oranları çoğunlukla düşmektedir. Kovaryans matrislerinin aynı ya da farklı olduğu durumlarda ortalama vektörleri arasındaki sayısal farklılık arttıkça doğruluk oranları artış göstermektedir.

Ortalaması ve kovaryans matrisleri belirtilen 3000'er değişkenden oluşturulmuş 3 gruplu 3000*3*4 verili gözlem seti analiz sonuçları çizelge 4.6.'da verilmiştir.

Çizelge 4.6'de görüldüğü gibi kovaryans matrisi ve ortalama vektörüne göre doğruluk oranı değişmektedir. Farklı ortalama vektörleri ile aynı kovaryans matrisine sahip verilerin doğruluk oranlarının birbirine yaklaştığı görülmektedir. Ortalama vektörleri aynı iken Kovaryans matrisleri arasındaki sayısal fark arttıkça doğruluk oranları çoğunlukla düşmektedir. Kovaryans matrislerinin aynı ya da farklı olduğu durumlarda ortalama vektörleri arasındaki sayısal farklılık arttıkça doğruluk oranları artış göstermektedir.

Çizelge 4.5. Ortalaması ve kovaryans matrisleri belirtilen 1000'er değişkenden oluşturulmuş 3 gruplu 1000*3*4 verili gözlem seti analiz sonuçları

	μ1: (0,0,0,0) μ2: .5,.5,.5,.5) μ3:(1,1,1,1)			μ1:(0,0,0,0) μ2:(1,1,1,1) μ3:(2,2,2,2)			μ1: (0,0,0,0) μ2: (2,2,2,2) μ3:(4,4,4,4)			μ1:(0,0,0,0) μ2:(.5,1,1.5,2) μ3:(2.5,3,3.5,4)		
	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
LDA	.59	.51	.47	.79	.67	.57	.97	.87	.73	.93	.82	.69
QDA	.59	.61	.67	.79	.71	.71	.97	.88	.80	.93	.85	.79
FDAM1	.59	.59	.66	.78	.69	.70	.97	.88	.78	.91	.82	.78
FDAM2	.60	.59	.67	.78	.70	.71	.97	.88	.79	.92	.84	.78
FDAB	.60	.60	.67	.78	.69	.71	.97	.88	.79	.92	.83	.79
MDA	.59	.56	.59	.79	.69	.63	.97	.88	.74	.92	.83	.73
	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4
	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4
	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6
LDA	.53	.49	.45	.70	.62	.70	.91	.83	.71	.87	.78	.66
QDA	.53	.60	.67	.70	.69	.70	.91	.85	.80	.87	.82	.79
FDAM1	.53	.59	.64	.69	.67	.70	.92	.83	.77	.87	.80	.76
FDAM2	.53	.59	.66	.70	.68	.71	.92	.86	.78	.88	.82	.77
FDAB	.52	.59	.65	.69	.67	.70	.92	.84	.78	.88	.81	.77
MDA	.53	.54	.59	.70	.65	.70	.92	.84	.74	.88	.80	.71
	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
LDA	.48	.45	.43	.61	.57	.52	.82	.76	.67	.85	.74	.63
QDA	.48	.45	.69	.62	.72	.73	.82	.84	.80	.85	.83	.80
FDAM1	.48	.44	.63	.61	.62	.67	.82	.78	.73	.84	.76	.73
FDAM2	.48	.46	.64	.61	.67	.69	.82	.81	.75	.84	.79	.74
FDAB	.48	.45	.64	.61	.62	.67	.82	.79	.73	.83	.76	.73
MDA	.48	.46	.56	.61	.63	.61	.82	.79	.70	.84	.76	.68
	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
LDA	.47	.45	.43	.60	.56	.60	.80	.75	.66	.88	.73	.63
QDA	.47	.69	.70	.60	.75	.73	.80	.85	.81	.88	.85	.80
FDAM1	.47	.55	.61	.60	.63	.68	.81	.76	.73	.88	.75	.73
FDAM2	.46	.61	.64	.60	.68	.70	.81	.79	.74	.88	.77	.75
FDAB	.46	.55	.63	.59	.63	.68	.81	.76	.74	.88	.75	.73
MDA	.46	.53	.56	.60	.58	.66	.81	.76	.72	.88	.73	.68

Çizelge 4.6. Ortalaması ve kovaryans matrisleri belirtilen 3000'er değişkenden oluşturulmuş 3 gruplu 3000*3*4 verili gözlem seti analiz sonuçları

	μ1: (0,0,0,0) μ2: .5,.5,.5,.5) μ3:(1,1,1,1)			μ1:(0,0,0,0) μ2:(1,1,1,1) μ3:(2,2,2,2)			μ1: (0,0,0,0) μ2: (2,2,2,2) μ3:(4,4,4,4)			μ1:(0,0,0,0) μ2:(.5,1,1.5,2) μ3:(2.5,3,3.5,4)		
	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1	Σ1
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3	Σ1	Σ2	Σ3
LDA	.59	.51	.46	.79	.67	.57	.97	.87	.73	.93	.82	.69
QDA	.59	.61	.67	.79	.71	.71	.97	.88	.80	.93	.85	.79
FDAM1	.58	.60	.64	.78	.69	.70	.96	.87	.78	.92	.82	.77
FDAM2	.58	.60	.65	.78	.70	.71	.96	.88	.79	.92	.83	.77
FDAB	.58	.60	.65	.78	.69	.70	.97	.87	.78	.92	.83	.77
MDA	.58	.57	.58	.78	.69	.63	.97	.87	.74	.93	.83	.71
	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4	Σ4
	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4	Σ4	Σ5	Σ4
	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6	Σ4	Σ5	Σ6
LDA	.53	.49	.44	.70	.62	.70	.91	.83	.71	.87	.78	.66
QDA	.53	.60	.67	.70	.69	.70	.91	.85	.80	.87	.82	.79
FDAM1	.51	.59	.64	.70	.68	.69	.91	.83	.77	.87	.80	.76
FDAM2	.52	.59	.65	.70	.67	.70	.91	.84	.78	.87	.81	.77
FDAB	.52	.59	.65	.70	.67	.70	.91	.83	.78	.87	.80	.76
MDA	.53	.53	.56	.70	.66	.70	.91	.84	.72	.87	.79	.69
	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7	Σ7
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9	Σ7	Σ8	Σ9
LDA	.48	.45	.43	.61	.57	.52	.82	.76	.67	.85	.74	.63
QDA	.48	.45	.69	.61	.72	.73	.82	.84	.80	.85	.83	.80
FDAM1	.48	.44	.63	.62	.62	.66	.82	.78	.75	.85	.76	.73
FDAM2	.48	.44	.64	.62	.67	.67	.82	.81	.75	.85	.78	.75
FDAB	.48	.45	.63	.62	.63	.67	.82	.79	.75	.85	.76	.74
MDA	.48	.45	.55	.62	.62	.59	.82	.79	.70	.86	.75	.66
	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10	Σ10
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12	Σ10	Σ11	Σ12
LDA	.47	.45	.43	.60	.56	.52	.80	.75	.67	.88	.73	.63
QDA	.47	.69	.70	.60	.75	.73	.80	.85	.80	.88	.85	.80
FDAM1	.48	.54	.62	.61	.62	.66	.80	.77	.75	.87	.76	.73
FDAM2	.48	.62	.63	.61	.68	.67	.80	.81	.75	.87	.79	.75
FDAB	.48	.55	.63	.61	.62	.67	.80	.78	.75	.87	.76	.73
MDA	.48	.48	.54	.61	.57	.59	.80	.77	.70	.88	.76	.68

Veri büyüklükleri artırıldıkça benzer sonuçlar elde edilmekle birlikte farklı olarak veri setlerinin her alt setlerdeki gözlem sayısının artırılması ile oluşturulan türetilmiş verilerin analizinde bütün yöntemlerin doğruluk oranlarının birbirine yaklaştığı sadece kovaryans matrisleri farklı alınan setlerde karesel ayırma analizinin öne çıktığı görülmüştür. Çizelgeler arasındaki doğruluk oranlarında ise bir farklılık görülmemiştir yani alt setlerin büyüklüklerinin artması ile doğruluk oranları değişmemektedir.

4.2 İkinci adım

Çok gruplu türetim sonuçları:

Çok gruplu türetimlerde, 50'şer alt set büyüklüğüne sahip verilerin analiz sonuçlarında grup sayısında artış oldukça doğruluk oranlarının azaldığı çizelge 4.7'de görülmektedir. Ayırma yöntemleri içinde en yüksek doğruluk oranına sahip olan Karma Ayırma Analizidir. 500'er alt set büyüklüğüne sahip verilerin analiz sonuçlarında grup artışı oldukça doğruluk oranlarının azaldığı görülmektedir. Ayırma yöntemleri benzer doğruluk oranları vermektedir.

Çizelge 4.7 Ortalama vektörleri (I) ve kovaryans matrisleri (a) belirtilen 10,15,20 gruplu 50'şer ve 500'er alt setlerden oluşturulmuş 4 değişkenli veri seti analiz sonuçları

	50 veri 4 değişken		
	10grup	15 grup	20 grup
LDA	.47	.45	.35
QDA	.49	.47	.39
FDAM1	.44	.42	.32
FDAM2	.41	.37	.31
FDAB	.46	.45	.37
MDA	.53	.50	.40
	500 veri 4 değişken		
	10grup	15grup	20grup
LDA	.45	.43	.32
QDA	.45	.43	.33
FDAM1	.44	.42	.31
FDAM2	.44	.42	.31
FDAB	.44	.42	.31
MDA	.45	.43	.33

Çok gruplu türetimlerde 50'şer alt set büyüklüğüne sahip ikinci grup ortalama vektörleri ve kovaryans matrisine sahip verilerin analiz sonuçlarında grup artışı oldukça doğruluk oranlarının azaldığı Çizelge 4.8'de görülmektedir. Ayırma yöntemleri içinde en yüksek doğruluk oranına sahip olan Karma Ayırma Analizidir. 500'er alt set büyüklüğüne sahip verilerin analiz sonuçlarında grup artışı oldukça doğruluk oranlarının genelde azaldığı görülmektedir. Ayırma yöntemlerinden en yüksek doğruluk oranına sahip olan yöntem çoğunlukla Karma Ayırma Analizi olmuştur.

Çizelge 4.8. Ortalama vektörleri (II) ve Kovaryans matrisleri (a) belirtilen 10,15,20 gruplu 50'şer ve 500'er alt setlerden oluşturulmuş 4 değişkenli veri seti analiz sonuçları

	50 veri 4 değişken		
	10grup	15grup	20grup
LDA	.96	.96	.96
QDA	.97	.96	.97
FDAM1	.91	.89	.74
FDAM2	.95	.85	.83
FDAB	.98	.97	.95
MDA	.99	.97	.97
	500 veri 4 değişken		
	10grup	15grup	20grup
LDA	.96	.96	.96
QDA	.96	.96	.96
FDAM1	.89	.84	.82
FDAM2	.92	.85	.84
FDAB	.92	.81	.95
MDA	.97	.96	.96

Çok gruplu verilerde, 50'şer alt set büyüklüğüne sahip ikinci grup ortalama vektörleri ve kovaryans matrisine sahip verilerin analiz sonuçlarında grup artışı oldukça doğruluk oranlarının azaldığı Çizelge 4.9'da görülmektedir. Ayırma yöntemlerinin doğruluk oranları benzerdir. 500'şer alt set büyüklüğüne sahip verilerin analiz sonuçlarında grup artışı oldukça doğruluk oranlarının genelde azaldığı görülmektedir. Ayırma yöntemlerinin doğruluk oranları benzerdir.

Çizelge 4.9. 10-15-20 grup için Ortalama vektörleri (II), Kovaryans matrisleri farklı (b) 50'şer ve 500'er alt setlerden oluşturulmuş 4 değişkenli veri seti analiz sonuçları

	50 veri 4 değişken		
	10grup	15grup	20grup
LDA	.69	.61	.55
QDA	.71	.64	.58
FDAM1	.66	.54	.48
FDAM2	.67	.56	.53
FDAB	.67	.61	.55
MDA	.71	.63	.61
	500 veri 4 değişken		
	10grup	15grup	20grup
LDA	.68	.60	.54
QDA	.69	.60	.55
FDAM1	.61	.52	.45
FDAM2	.68	.55	.49
FDAB	.68	.58	.53
MDA	.69	.60	.54

Çok gruplu verilerde, 50'şer alt set büyüklüğüne sahip ikinci grup ortalama vektörleri ve kovaryans matrisine sahip verilerin analiz sonuçlarında grup artışı oldukça doğruluk oranlarının azaldığı Çizelge 4.10'da görülmektedir. Ayırma yöntemlerinin doğruluk oranları benzerdir. 500'şer alt set büyüklüğüne sahip verilerin analiz sonuçlarında grup artışı oldukça doğruluk oranlarının genelde azaldığı görülmektedir. Ayırma yöntemlerinin doğruluk oranları benzerdir.

Çizelge 4.10 Çizelge 2.1- Ortalama vektörleri (III) ve Kovaryans matrisleri (b) belirtilen 10,15,20 gruplu 50'şer ve 500'er alt setlerden oluşturulmuş 2 değişkenli veri seti analiz sonuçları

	50 veri 2 değişken		
	10grup	15grup	20grup
LDA	.55	.47	42
QDA	.56	.49	.44
FDAM1	.54	.47	.42
FDAM2	.56	.47	.41
FDAB	.55	.49	.44
MDA	.55	.52	.46
	500 veri 2 değişken		
	10grup	15grup	20grup
LDA	.42	.47	.42
QDA	.42	.47	.42
FDAM1	.40	.45	.41
FDAM2	.41	.46	.42
FDAB	.42	.47	.42
MDA	.41	.47	.42

4.3.3 Üçüncü Adım:

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$ kovaryans matrisli

μ1=(0,0,0,0), μ2=(3,3,3,3), μ3=(6,6,6,6) ortalama vektörlü, 4 değişkenli 50'şer verili 3 gruplu veri setinin ayırma analizi sonuçları aşağıda verilmiştir. Sınıflandırma tablosuna bakıldığında en iyi BIC değerinin 2123.958 değeri ile EII olduğu görülmektedir. En iyi model olarak EII seçilmiştir.

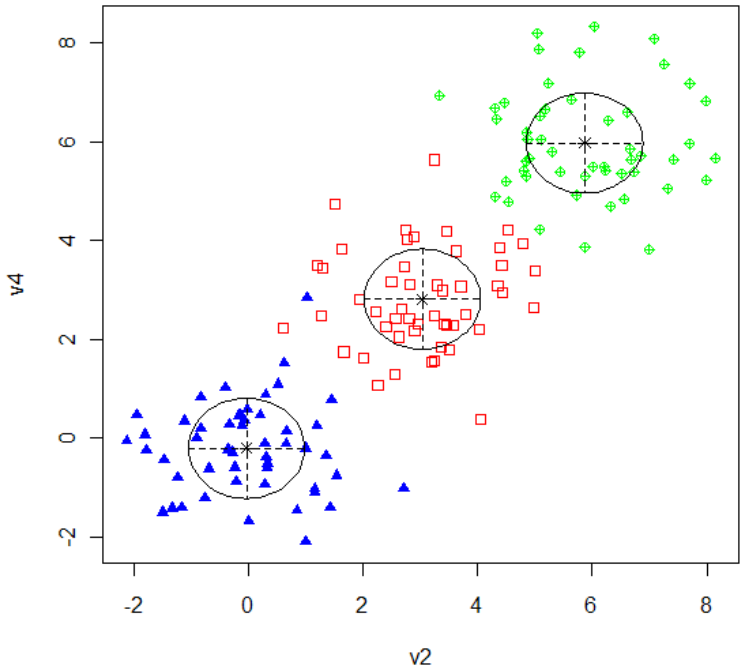
BIC değerlerine dayalı en iyi modelin sınıflandırma tablosu,

Çizelge 4.11 Sınıflandırma Tablosu

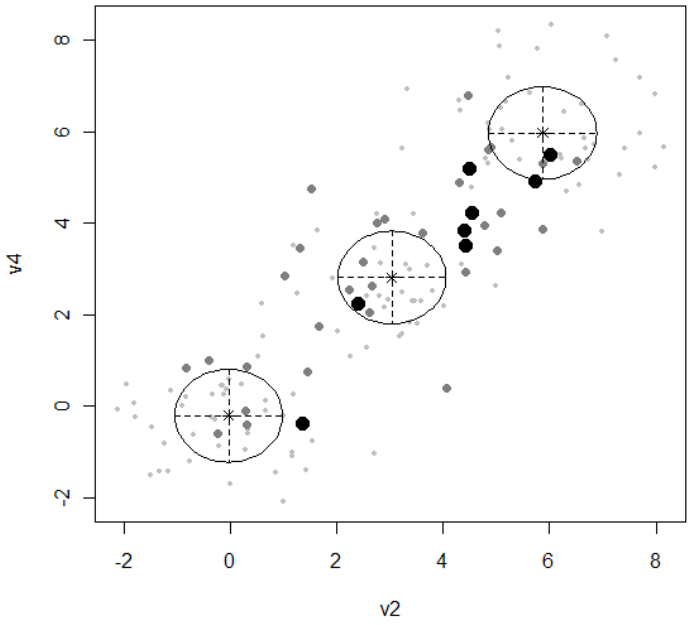
Gruplar		
I	II	III
50	50	50
En iyi BIC değerleri		
EII,3	EEI,3	EII,4
-2123.958	-2136.659	-2141.552

şeklindedir.

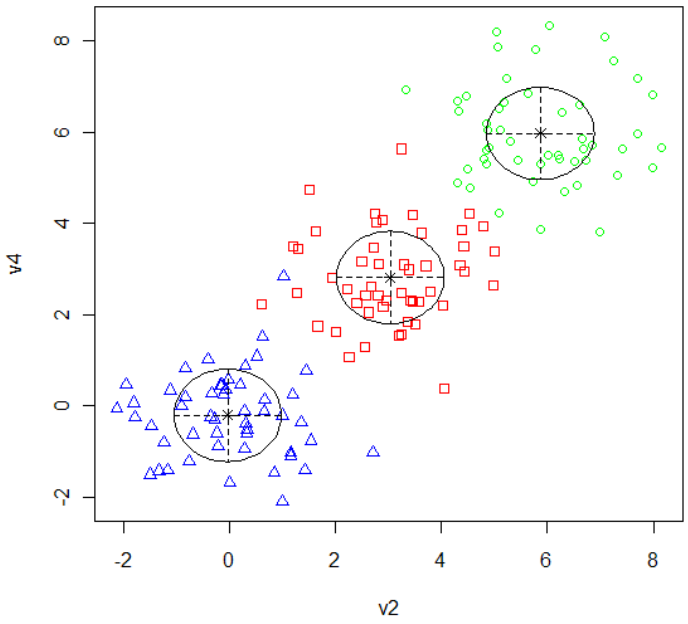
Herhangi iki değişken için (v2, v4) sınıflandırma, belirsizlik ve hata grafikleri şekil 4.1-4.2-4.3'de verilmiştir.



Şekil 4.1. 2. ve 4. Boyut için sınıflandırma grafiği



Şekil 4.2. 2. ve 4. Boyut için belirsizlik grafiği



Şekil 4.3. 2. ve 4. Boyut için hata grafiği

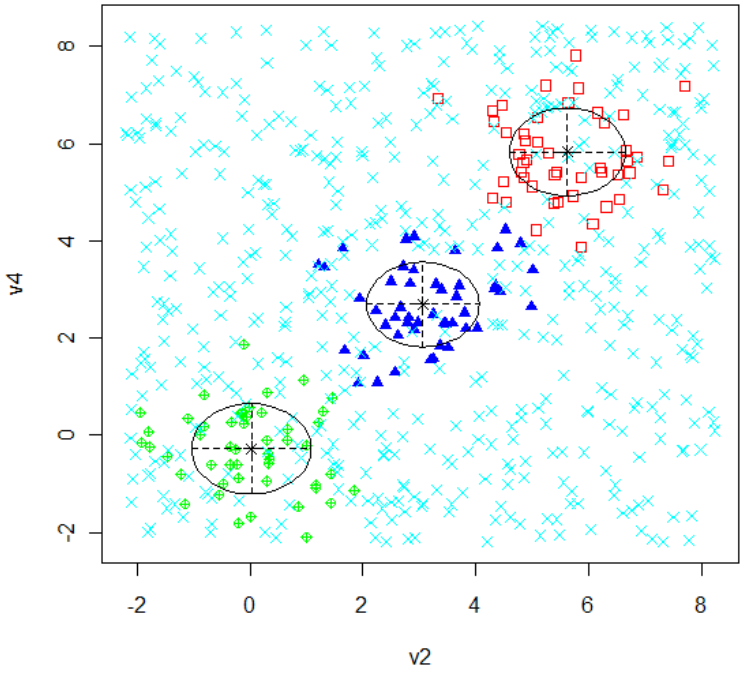
Aynı sete 500 verilik poisson noise uygulanmıştır. Sınıflandırma tablosunda en iyi BIC değerinin VEI=12072.27 olduğu görülmektedir. En iyi Bayes modeli VEI olarak belirlenmiştir. Şekil 4.4 de Poisson gürültü eklendiği durumdaki herhangi iki değişkenin (v2,v4) sınıflandırma grafiği gösterilmektedir.

BIC değerlerine dayalı en iyi modelin sınıflandırma tablosu,

Çizelge 4.12 Poisson gürültü eklendiği durumdaki v2,v4 değişkenlerinin sınıflandırma tablosu

Gruplar			
0	I	II	III
506	48	47	49
En iyi BIC değerleri			
VEI,3	VII,4	EEI,4	
-12072.27	-12086.09	-12091.95	

şeklindedir



Şekil 4.4. 2. ve 4. Boyut için Poisson gürültü eklendiği durumda sınıflandırma grafiği

Bu çalışmada yapılan analiz sonuçlarında;

Birinci aşamada,

1. Kovaryans matrisinin parametrizasyonuna göre doğruluk oranı değişmektedir;
   i. Ortalama vektörleri sabit ama grup varyansları büyüdükçe doğruluklar azalmaktadır
   ii. Ortalama vektörleri sabit ama kovaryans değerleri büyüdükçe doğruluklar azalmaktadır
2. Grup Ortalama vektörleri arasındaki farka göre doğruluk oranı değişmektedir;
   i. Varyanslar sabit ama grup Ortalama vektörleri arasındaki fark arttıkça doğruluk oranlarının arttığını gözlemlenmiştir.

3. Grup gözlem sayılarına (50-100-250-500-1000-3000) göre doğruluk oranları değişmemektedir.

İkinci aşamada,

1. Grup ve gözlem sayılarına göre doğruluk oranı değişmektedir.
    i. Grup sayısı arttıkça doğruluk oranları azalmaktadır.
    ii. Gözlem sayıları arttıkça doğruluk oranları azalmaktadır.
    iii. Grup sayısı arttıkça diğer yöntemlere göre Karma Ayırma Analizi doğruluk oranları çoğunlukla daha yüksektir.

Üçüncü aşamada,

Poisson gürültü uygulanan verilerde oldukça başarılı doğruluk oranı gözlemlenmiştir.

5- TARTIŞMA VE SONUÇ

Model tabanlı kümeleme analizi ve gizli sınıf regresyonu, bireyleri gözlemlenmeyen gruplamada popüler yöntemlerdendir. Birçok uygulamada bu gözlemler arasındaki ilişkiye dikkat çekilmekte özellikle hangi gözleminin diğerine yakın olduğu ve diğerlerinin de birbirinden ne kadar farklı olduğu konularında çalışılmaktadır (8).

İstatistiksel hesaplama yöntemlerinde karma modellerin bileşen yapılarını açıklamak için yeni yöntemler üretilmekte ve model sunumu grafiksel hesaplamalarla desteklenmektedir.

1990'ların sonlu karma modeller, standart karma regresyon modelleri ve genelleştirilmiş lineer modeller kullanarak geliştirilmiştir. Sabit sayıda bileşenli sonlu karma modeller EM algoritması, En Çok Olabilirlik ve MCMC (Markov Chain Monte Carlo) metodu ve Bayes uygulaması ile çözümlenmektedir (2,8,30).

Karma modellerin son yıllarda başarı ile uygulanmasına rağmen model tanımlama ve genel görüntüleme teknikleri uygulamada zorluklarla karşılaşmaktadır. Örneğin düşük boyutlu Gauss'ları görüntülemede kullanılan güven elipsleri, regresyon modelleri için kullanılamamaktadır (8).

Bu araştırmada; farklı büyüklükte veri setleri, farklı ortalama vektörleri ve kovaryans matrisleri ile türetilmiş veri setleri kullanılarak yöntem ve modellerin performansları incelenmiş model seçme yöntemi anlatılmış, grafikle gösterilmiştir.

Patolojik ses tanıma için yapılan bir sınıflandırma analizinde GMM ve Yapay Sinir Ağları karşılaştırılmıştır. Normal insanlardan ve hastalardan seçilen konuşmalar değerlendirmeye alınmış, 6 karakteristik parametre seçilmiştir. Yöntemler veriyi normal ve patolojik olmak üzere iki kategoriye ayırmak üzere uygulanmıştır.

GMM metodu %98.4 çalışma verisi doğru sınıflandırma oranına ve %95.2 test verisi doğru sınıflandırma oranına ulaşmış. Sınıflandırma için uygun koşulları bulmak amacıyla 3-15 arası farklı karma sayıları kullanılmış. Daha önce uygulanan sinir ağları yönteminden daha iyi sonuçlar verdiği gözlemlenmiştir. Ses patolojisi konusunda GMM'nin sınıflandırma prosedürü olarak uygulanması tavsiye edilmektedir (41).

Bu araştırmada yapılan karşılaştırmalarda GMM' nin düşük ve yüksek boyutlarda doğruluk oranlarının ya yüksek olduğu ya da diğer yöntemlerle aynı doğruluk oranına sahip olduğu görülmüştür. Yüksek boyutlar için uygulanması tavsiye edilmektedir (çizelge 4.7-4.10).

Uygulamalarda 10 tekrar ve 5 tekrar kullanılmıştır (25).

Bu araştırmada yüksek boyutlu analizlerde 10 diğerlerinde 1000 tekrar kullanılmıştır.

MDA LDA ve QDA'dan daha iyi sonuçlar vermektedir (42).

Bu araştırmada MDA'nın doğruluk oranı ya diğer ayırma analizlerden iyi ya da genelde çok yakın değerler arasındadır.

Hastie ve ark. (2008) MARS'ı türetilmiş 3 veri setine uygulamışlar. Her birinin örnek büyüklüğü n=100'dür. Birinci örnek X_1, X_2 değişkenleri ile normal dağılımdan türetilmiştir. İkinci örnekte birincinin benzeri olup 18 Gaussian gürültü (noise) eklenmiştir. Bunlara p=2 ve p=2+18=20 denmiştir. Üçüncü örnek ise sinir ağları yapısındadır. Bu 3 örnek için MARS ve GMR karşılaştırılmıştır. GMR p=2 ve p=20 de MARS'dan daha yüksek hata vermekler birlikte arada önemli fark yoktur. Sinir ağları verisinde ise GMR performansı MARS'dan daha iyidir (2).

	MARS		GMR	
	Ortalama	SE	Ortalama	SE
p=2	.97	.01	.93	.04
p=20	.96	.01	.83	.07
Sinir Ağları	.79	.01	.94	.01

Bu çalışmada Gauss karma ayırma analizi metotları diğer esnek yöntemlere göre daha çoğunlukla yüksek doğruluk oranına sahiptir.

Shi ve ark. Paraplegia veri seti ile çalışmışlar ve ayakta hastalardan birkaç yüz veri toplanmıştır. Her ayakta hasta için benzer olmakla birlikte aynı hasta için bile aynı olmayan prosedür birkaç kez her sekiz hastaya tekrarlanmış. Bu da replikasyonlar arasındaki heterojeniteyi doğurmaktadır. Burada tartışılan tekrarlı ölçümler durumunda hiyerarşik yapıda bir model tanımlanmış: verinin temel yapısını modellemek için her gruba ayrı düşük seviyede model uygulanmış böylece düşük seviyeli modeller benzer yapıda ama müşterek heterojeniteye sahip olmayan veri setleri elde edilmiş ve yüksek seviyedeki model gruplar arasındaki heterojeniteyi modellemek için kullanılmıştır. Bayes MCMC'e göre doğruluk oranı=0.984, hiyerarşik olmayan Gauss modelinde ise doğruluk oranı=0.99 elde edilmiştir (28).

Bu araştırmada 100*3*4'er değişkenden oluşturulmuş, ortalama vektörleri µ1: (0,0,0,0), µ2: (.2,.2,.2,.2), µ3:(4,4,4,4) ve kovaryans matrisleri Σ olan veri setiçin MDA doğruluk oranı 0.98'dir.

Martin-Magniette ve ark. insan geninde destekleyici DNA metilasyonu üzerine, veri Weber ve ark (2007). 15609 insan geninin destekleyici bölgede sıralanması ile ilgili çalışılmıştır. Karma ayırma Analizi ile % 55 doğruluk oranı elde etmişlerdir (20).

Angela Montanari ve arkadaşlarının 5 ölçüm içeren thyorid verisi için 215 kişilik hasta grubunda yaptıkları çalışmada, yöntemlerin karşılaştırmaları görülmektedir: IFDA= Bağımsız faktör ayırma analizini göstermektedir . Angela ve arkadaşları LDA, MDA, FDA yöntemlerini karşılaştırmışlar ve çalışma verisi doğruluk oranları LDA=0.909, MDA=0.972, FDA=0.951 bulmuşlardır (18).

Bu araştırmada, 50*3*4 4'er değişkenden oluşturulmuş, ortalama vektörleri μ1: (0,0,0,0), μ2: (.2,.2,.2,.2), μ3:(4,4,4,4) ve kovaryans matrisleri Σ olan veri setiçin MDA hata oranı 1-0.96=0.04 (çizelge 4.2)dür.

Jianling Wang ve arkadaşlarının yaptığı çalışmada Karma Ayırma Analizinin performansının karma sayıları ve döngü sayıları arttıkça doğruluk oranları gösterilmiştir (41).

Bu çalışmada döngü sayıları artırıldığında doğruluk oranı yüksek sonuçlar elde edilmiştir.

Sonuç olarak Karma Ayırma Analizi yüksek doğruluk oranları vermektedir ve parametrik olmayan regresyon metodu oluşturmak üzere Gauss yoğunluk regresyonu kullanılması önerilmektedir.

KAYNAKLAR

1. Alpar, R., 2003, Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş 1, Nobel Ankara. 412 s.

2. Bashir, S., 2003, High Breakdown Mixture Discriminant Analysis, Ph. D. thesis, GUELPH UNIVERSITY Canada 111 p.

3. Breiman, L., J- H. Friedman, R. A. Olshen, and C. J. Stone, 1984, Classification and Regression Tree. Pacific Grove, California: Wadsworth & Brooks 358 p.

4. Falk, T., H., Shatkay, H., and Chan, W., Y., 2003, Breast Cancer Prognosis via Gaussian Mixture Regression, Int. J. Adapt. Control Signal Process. 17, 149-161 p.

5. Fraley, C., and Raftery, A. E., 2006, Some Applications of Model Based Clustering in Chemistry, R News, Vol 6/3 17-23 p.

6. Friedman, J., and Stuetzle, W. 1981, Projection Pursuit Regression, Journal of the American Statistical Association, 76, 817-823 p.

7. Friedman, J., 1991, Multivariate adaptive regression splines (with discussion). An-nals of Statistics 19(1), 1-141, 16 p.

8. Grün, B., and Leisch, F., 2007, Fitting finite mixtures of generalized linear regressions in R, Computational Statistics & Data Analysis, 51(11): 5247-5252 p.

9. Halbe, Z., and Aladjem, M., 2005, Model Based Mixture Discriminant Analysis-An Experimantal Study, Science Direct Volume 38, Issue 3, 437-440 p.

10. Hastie, T., and Tibshirani, R., 1990, Generalized Additive Models, New York, Chap-man and Hall, New York, 335 p.

KAYNAKLAR DİZİNİ (devam ediyor)

11. Hastie, T., Tibshirani, R., and Buja, A., 1994, Flexible Discriminant Analysis by Optimal Scoring. Journal of the American Statistical Association 89(428), 1255-1270 p.

12. Hastie, T., Buja A., and Tibshirani, R., 1995, Penalized Discriminant Analysis. The Annals of Statistics Vol 23, No. 1, 73-102 p.

13. Hastie ,T., and Tibshirani, R,. 1996 Discriminant Analysis by Gauss Mixtures, J. R. Statist. Soc. B, 58(1):155-176 p.

14. Hastie, T., Tibshirani, R., and Friedman, J., 2001, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York: Springer, 763 p.

15. Klinke, S., and Grassman, J., 1998, Projection Pursuit Regression and Neural Networks. Humbolt University Berlin, Germany, DP 9817, SFB 373, 47 p.

16. Leisch, F., 2004, Exploring the structure of mixture model components. In Jaromir Antoch, editor, Compstat 2004 - Proceedings in Computational Statistics, pages 1405-1412. Physika Verlag, Heidelberg, Germany, ISBN 3-7908-1554-3 p.

17. Lemm, J, C., 2003, Bayesian Field Theory Nonparametric Approaches to Density Estimation, Regression, Classification, and Inverse Quantum Problems Baltimore: Johns Hopkins University Press, 14-15 p.

18. Mantanari, A., and Calo, D., G., 2008, Independent factor discriminant analysis, Econpapers, vol. 52, issue 6, 3246-3254 p.

19. Marco, D., Ugo, G., and Orietta L., 2006, Use of Finite Mixture Models in Editing and Imputation of Survey Data European Conference on Quality in Survey Statistics, 8p.

KAYNAKLAR (DEVAM EDİYOR)

20. Martin-Magniette, M.,L., Mary-Huard T., C., and Berard C, Robin, S., 2008, Regression Mixture Model for ChIP-chip Analysis, Bioinformatics, 24(16),181-186 p.

21. Matthews, D., and E., Farewell V., T., 1988, Using and Understanding Medical Statistics, Karger Basel, 2nd edition, 228 p.

22. McLachlan, G., and Peel, D., 2000 Finite Mixture Models, New York,John Wiley&Sons, Inc, 419 p.

23. Özdamar, K., 2002, Paket Programlar ile İstatistiksel Veri Analizi II, Kaan Kitabevi, 4. Baskı, Eskişehir, 649 s.

24. Özdamar, K., 2003, SPSS ile Biyoistatistik, Kaan Kitabevi, 5. Baskı, Eskişehir, 506 s.

25. Öztürk A., 2006, Esnek Ayırma Analizi ve Bir Uygulama. Doktora tezi. Eskişehir Osmangazi Üniversitesi 75 s.

26. Rencher, A., C., 2002, Methods of Multivariate Analysis, 2nd edition, John Wiley & Sons Inc., USA, 708 p.

27. Scott, D., W., 1992, Multivariate Density Estimation: Theory, Practice, and Vi-sualization. New York: Wiley, 317 p.

28. Shi, J. Q., Murray-Smith R., and Titterington, D., M., 2002, Hierarchical Gaussian Process Mixtures for Regression, Statistics and Computing, 15 (1), 31-41 p.

29. Srivastava, M. S., and Carter, M., 1983, An Introduction to Applied Multivariate Statistics New York : North-Holland, 394 p.

KAYNAKLAR DİZİNİ (devam ediyor)

30. Sung, H. G., 2004, Gaussian Mixture Regression and Classification, Ph. D. thesis, RICE UNIVERSITY, Houston Texas 157 p.

31. Tim, N.H., 2002, Applied Multivariate Analysis, Springer-Verlag New York, Inc., 693 p.

32. Türe, M., Kurt İ., Kürüm A. T., and Özdamar K., 2005, Comparing Classification Techniques for Predicting Essential Hypertension, Expert Systems with Applications 29(3): 583-588 s.

33. Zhu, M., 2001, Feature Extraction and Dimension Reduction with Applications to Classification and the Analysis of Co-Occurrence Data, Ph. D. thesis, Stanford University California 137 p.

34. Zong, S., Ghosh, J., 2003, A Unified Framework for Model Based Clustering, Journal of Machine Learning Research, 4, 1001-1037p.

35. http://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf, (2009-9-2)

36. http://cran.r-project.org/doc/contrib/Epicalc_Book.pdf, (2009-9-2)

37. http://cran.r-project.org/doc/contrib/Seefeld_StatsRBio.pdf, (2009-9-2)

38. http://cran.r-project.org/doc/contrib/usingR.pdf, (2009-9-2)

39. http://cran.r-project.org/web/packages/mda/mda.pdf, (2009-9-2)

40. http://cran.r-project.orgdoccontribFaraway-PRA_pdf.txt, (2009-9-2)

41. http://www.assta.org/sst/2006/sst2006-84.pdf (2009-9-2)

42. http://www.stat.psu.edu/~jiali/course/stat597e/notes2/mda.pdf (2009-9-2)

43. http://www-stat.stanford.edu/~hastie/Thesis/Gill_Ward.pdf (2009-9-2)

ÖZGEÇMİŞ

Doğum tarihi/yeri	24.11.1975/ Kayseri
Lisans	Anadolu Üniversitesi Fen Fakültesi İstatistik 1994-1999
Yüksek Lisans	Erciyes Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı 2000-2003 Aynı dönemde Halk Sağlığı Anabilim Dalında, Araştırma Görevlisi olarak çalışmıştır.
Doktora	Eskişehir Osmangazi Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı 2003- 2009