Panel Veri Analizi Kullanılarak Zaman Serilerinin Model Tabanlı Kümelenmesi

Selim Dönmez

DOKTORA TEZİ

İstatistik Anabilim Dalı

Eylül 2020

Model-Based Clustering of Time Series Using Panel Data Analysis

Selim Dönmez

DOCTORAL DISSERTATION

Department of Statistics

September 2020

Panel Veri Analizi Kullanılarak Zaman Serilerinin Model Tabanlı Kümelenmesi

Selim Dönmez

Eskişehir Osmangazi Üniversitesi

Fen Bilimleri Enstitüsü

Lisansüstü Yönetmeliği Uyarınca

İstatistik Anabilim Dalı

İstatistik Bilgi Sistemleri Bilim Dalında

DOKTORA TEZİ

Olarak Hazırlanmıştır

Danışman: Dr. Öğr. Üyesi Özer Özaydın

İkinci Danışman: Prof.Dr. Hamza Erol

Eylül 2020

ETİK BEYAN

Eskişehir Osmangazi Üniversitesi Fen Bilimleri Enstitüsü tez yazım kılavuzuna göre, Dr. Öğr. Üy. Özer Özaydın ve Prof. Dr. Hamza Erol danışmanlığında hazırlamış olduğum "Panel Veri Analizi Kullanılarak Zaman Serilerinin Model Tabanlı Kümelenmesi" başlıklı tezimin özgün bir çalışma olduğunu; tez çalışmamın tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı; tezimde verdiğim bilgileri, verileri akademik ve bilimsel etik ilke ve kurallara uygun olarak elde ettiğimi; tez çalışmamda yararlandığım eserlerin tümüne atıf yaptığımı ve kaynak gösterdiğimi ve bilgi, belge ve sonuçları bilimsel etik ilke ve kurallara göre sunduğumu beyan ederim. 28/09/2020

Selim Dönmez

İmza

ÖZET

Bu çalışmada, model tabanlı kümelemenin panel veri analizine sunabileceği katkıları incelenmiştir. Panel veri, seçilmiş birtakım birimlerin zaman serilerinden oluşan özel bir veri türüdür. Panel verideki birimlerin sınıflandırması, birtakım panel veriler için uygulamaya açık olduğu gibi aynı zamanda gereklidir. Bu çalışmada, Borsa İstanbul(BIST) panel verisi, gayri safi yurtiçi hasıla panel verisi, yabancı yatırım panel verisi ve Dow-Jones panel verisinin model tabanlı kümelenmesi gerçekleştirilmiştir. Bu çalışmada sözkonusu veriler üzerinde gerçekleştirilen kümeleme analizinde elde edilen kümelerin geçerliliği, kümelerin gölge değişkenlerinin sözkonusu verilerin panel veri analizine dahil edilmesi ile değerlendirilmiştir. Sözkonusu verilerden model tabanlı kümeleme ile elde edilen kümeler panel veri analizine olumlu katkılar sunmuştur.

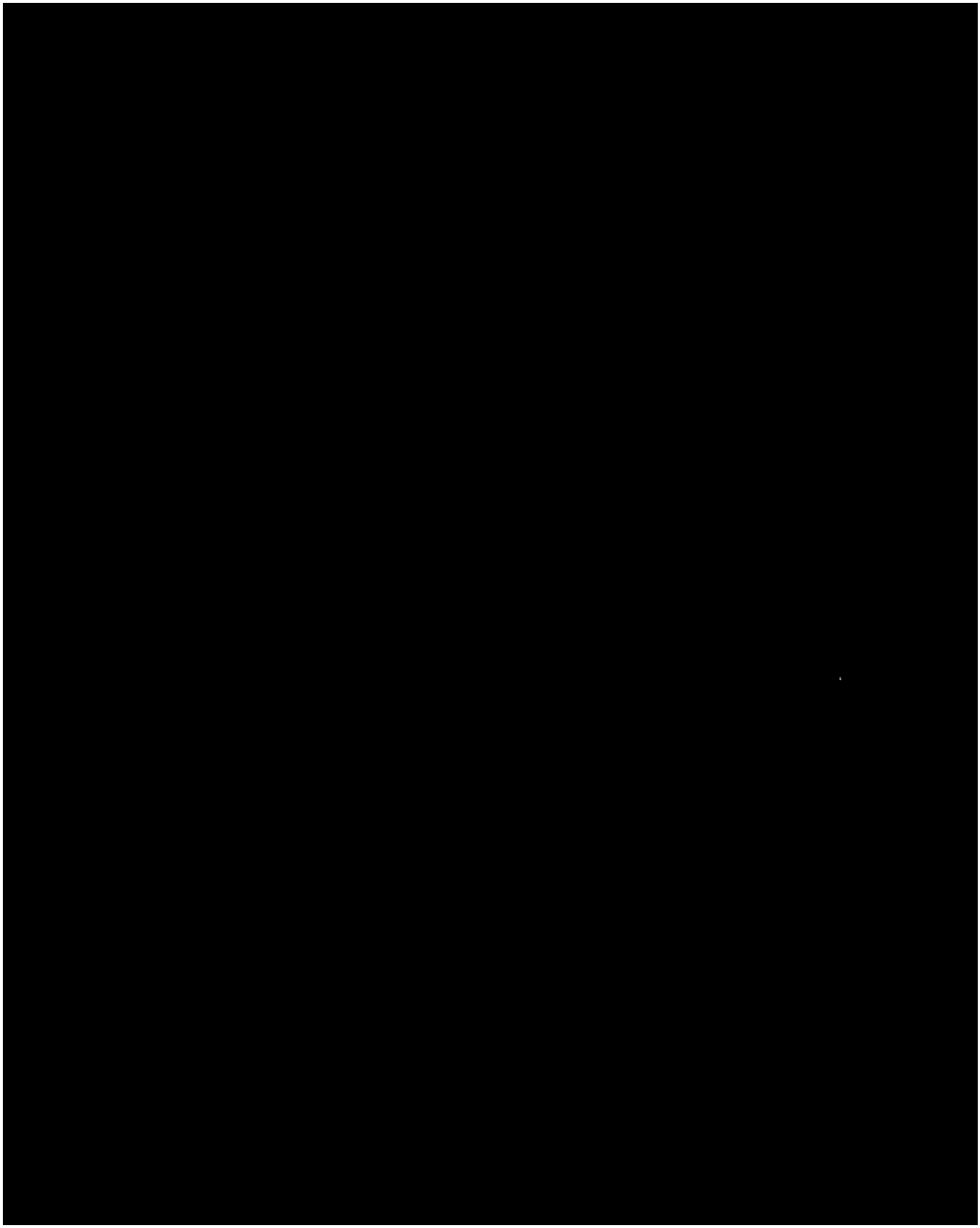
Anahtar Kelimeler: Model Tabanlı Kümeleme, Panel Veri, Zaman Serisi, Kümeleme Analizi

SUMMARY

In this study, we aimed to examine the potential contributions of model-based clustering to the panel data analysis. Panel data, is an aggregation of time series of selected cross-sectional units. The clustering of such units is not only applicable to some panel data but also necessary to them. In this study, we applied model-based clustering to data such as Borsa Istanbul(BIST) panel data, gross domestic product panel data, foreign direct investment panel data and Dow-Jones panel data. In this study, the validity of clusters obtained by the cluster analysis of aforementioned data were evaluated by introducing the dummy variables of clusters to the panel data analysis of aforementioned data. This application provided positive contributions to the panel data analysis of aforementioned data.

Keywords: Model-Based Clustering, Panel Data, Time Series, Cluster Analysis

TEŞEKKÜR



İÇİNDEKİLER

Sayfa

İÇİNDEKİLER(DEVAM)

ŞEKİLLER DİZİNİ

Şekil **Sayfa No**

ÇİZELGELER DİZİNİ

Çizelge **Sayfa No**

SİMGELER VE KISALTMALAR DİZİNİ

Simgeler	Açıklama
$N(\mu_1, \sigma_1)$	μ_1 ortalamalı σ_1 standart sapmalı normal dağılım
$\pi_j^{(A)}$	A. aşamada elde edilen j. bileşenin tahmini ağırlık katsayısı
$\log_e(.)$	Doğal logaritma
n_G	G. kümenin eleman sayısı

Kısaltmalar	Açıklama
AIC	Akaike bilgi kriteri
BIC	Bayesçi bilgi kriteri
BLUE	Doğrusal en iyi yansız tahmin edici
CEM	Sınıflayıcı beklenti(Expectation) yükseltme(Maximization) algoritması
EDDA	Özdeğer ayrıştırmalı diskriminant analizi
EKK	En küçük kareler yöntemi
EM	Beklenti(Expectation) yükseltme(Maximization) algoritması
GEKK	Genelleştirilmiş en küçük kareler
ICL	Entegre tam olabilirlik
ILO	Uluslararası işgücü organizasyonu
LCM	Gizil sınıf modeli
LDA	Basit doğrusal diskriminant analizi
QDA	Kuadratik diskriminant analizi
RDA	Düzenlenmiş diskriminant analizi
RMDA	Sağlamcı karma diskriminant analizi
SEM	Stokastik beklenti(Expectation) yükseltme(Maximization) algoritması
UGEKK	Uygun genelleştirilmiş en küçük kareler
VEM	Değişimli beklenti(Expectation) yükseltme(Maximization) algoritması

1. GİRİŞ VE AMAÇ

İstatistik, veri toplama, veri çözümleme, sonuç çıkarma ve sonuç yorumlama bilimi olarak tanımlanır. İstatistik bilim olarak devletlerin kaynaklarını verimli kullanılmasını sağlamak amacıyla ortaya çıkmıştır. Bu açıdan bakıldığında istatistiğin karar vermede bir araç olduğu söylenebilir. Bunun yanında bu aracın ne kadar faydalı olduğu hata payının ne kadar düşük ya da kontrol edilebilir olduğu ile ölçülür. İstatistiğin tanımı ile hata payı olgusu bağdaştırmaya çalıştığımız zaman, veri ile sonuç arasında bir arabulucuya ihtiyaç duyulmaktadır. Bu arabulucu, veri üzerinden kurduğumuz model olmaktadır. Her veriseti için ayrı bir model oluşturulmalıdır.

Panel veri hem zamana göre hem de birimlere göre gözlem içeren bir veri türüdür. Panel verinin toplanmasındaki amaç birimler arasındaki farklılıkları açıklayarak öngörü yapmak olduğundan değişik bir modele ihtiyaç duyulmaktadır. Panel veri için kurulan doğrusal model üç uyum endeksi üzerinden değerlendirilir. Bunlardan ikisi birimlerdeki ve birimler arasındaki değişimin açıklanması için kullanılır. Diğeri ise daha genel olarak değişimin açıklanması için kullanılır.

Bu tez çalışmasında kullanılan panel veri setleri zaman serilerinden oluşmaktadır. Panel veri'de özellikle birimler arasındaki değişimleri açıklamak için kümeleme analizi yapılabilir ve bu tip analiz, panel veriyi modellemede işe yaramaktadır. Bu noktada kümeleme analizi gerçekleştirilirken, karma modelleri kullanmak gerekmektedir. Karma modeller, veriye uygun bir şekilde teorik olasılık dağılımlarının ağırlıklandırılarak ifadesinden elde edilir. Bu nedenle karma modeller, kümeleme yapmak için elverişli yöntemler sunarlar. Bunun gerçekçi bir şekilde yapılabilmesi için geniş bir literatür oluşturacak kadar akademik çalışmalar yapılmıştır. Karma modeller ile yapılan kümelemeye kısaca model tabanlı kümeleme denmektedir. Bu tezde panel veride model tabanlı kümeleme uygulanması ile panel veri modeline katabileceği zenginlik incelenmiştir ve anlamlı sonuçlar tespit edilmiştir. Bunun sonucunda, panel veriden işe yarar bilgiler edinmenin yolu gösterilmiştir.

Tez, beş bölümden oluşmaktadır. İlk bölüm, giriş bölümüdür. Tezin ikinci bölümünde, karma modeller ve model tabanlı kümeleme anlatılmaktadır. Tezin üçüncü bölümünde, panel veri için kümeleme yöntemleri incelenecektir. Tezin dördüncü bölümünde, panel veri için

analiz yöntemleri incelenecektir. Son bölümde, dört tane gerçek hayat verisi üzerine panel veri analizi ve model tabanlı kümeleme analizi uygulanarak sonuçlara varılacaktır. Elde edilen kümeleme sonuçları yorumlanacaktır.

2. LİTERATÜR ARAŞTIRMASI

2.1 Giriş

Karma model, iki veya daha fazla dağılımın ağırlıklandırılmasıyla oluşturulan dağılıma denir. Dağılımlar, toplamı 1 olan karıştırma ağırlığı değerleriyle ağırlıklandırılır ve olasılık yoğunluk fonksiyonları buna göre yeniden düzenlenir. Karma model, (2.1)'deki eşitlikteki gibi tanımlanır:

$$\mathrm{KN}(\mu_1, \mu_2, \sigma_1, \sigma_2, \mathrm{p}) = \mathrm{pN}(\mu_1, \sigma_1) + (1 - \mathrm{p})\mathrm{N}(\mu_2, \sigma_2) \qquad (2.1)$$

Burada $\mathrm{KN}(\mu_1, \mu_2, \sigma_1, \sigma_2, \mathrm{p})$, karma modelin olasılık yoğunluk fonksiyonunu; $\mathrm{N}(\mu_1, \sigma_1)$ ve $\mathrm{N}(\mu_2, \sigma_2)$ bileşen olasılık yoğunluk fonksiyonlarını ve p birinci normal dağılım için bileşen ağırlığını göstermektedir. Birinci normal dağılım için ağırlık p ne kadar büyük olursa normal dağılımlar arasında o kadar büyük bir fark oluşur. Sözkonusu dağılımdaki diğer parametreler $\mu_1, \mu_2, \sigma_1, \sigma_2$ sırasıyla her bir normal dağılım için ortalamaları ve standart sapmaları temsil etmektedir. Sonlu karma modelin parametreleri en çok olabilirlik yöntemiyle belirlenir. İşlem sonucunda çözülmesi zor denklemler çıktığı için parametreler EM algoritmasıyla iteratif olarak hesaplanır. Sonlu karma modeller, Newcomb (1886) tarafından sapan değerleri tespit etmek için ortaya atılmıştır. Bu çalışmadan sonra Pearson (1894) tarafından kullanılan sonlu karma modeller daha sonra kullanımda yaygınlaşmaya devam etmiştir. Bugün karma model, model tabanlı kümelemenin temelini oluşturmaktadır. Model tabanlı kümeleme, çok kolay yorumlama olanağı sağlayan bir kümeleme türüdür. Model tabanlı kümelemede, normal karma modelleri sıklıkla kullanılır.

Karma modeller, kümeleme, denetimli sınıflandırma, yarı denetimli sınıflandırma ve diskriminant analizinde önemli rol oynarlar. Bu rol, yorumlanabilirlikte ve analizde kolaylık olarak ifade edilebilir haldedir. Özellikle de veriye uygunluk alanında karma modellerin esnekliği ve literatürde 60 ila 70 yıllık birikim çok önemli yardımlarda bulunmuştur. Model tabanlı kümeleme, karma modellerle yapılan kümeleme ve sınıflama analizlerinin genel adıdır.

Model tabanlı kümeleme, karma modelleri uygun bir şekilde seçerek verideki gruplaşmaları ortaya çıkarak anlamlı sonuçlar çıkarmaya yaramaktadır.

Tarihte model tabanlı kümeleme üzerine çalışılan birtakım tezler, Açıkgöz (2007) tarafından verilen bazı tezler Kim (1984), Erol (1995)'tir. Ancak tarihte ilk defa model tabanlı kümelemenin kim tarafından keşfedildiği araştırıldığında 1950 yılına gitmemiz gerekmektedir. Bu soruya yanıt aramak için Bouveyron vd. (2019), model tabanlı kümelemeyi üç özelliği ile tanımlamışlardır:

i. Karma model üzerine inşa edilmiş olması
ii. Sistematik bir istatistiksel yöntem ile karma model parametrelerinin tahmin edilmiş olması
iii. Sistematik bir yöntem ile gözlemlerin sınıflandırılması

Lazarsfeld'in "The Logical And Mathematical Foundations Of Latent Structure Analysis" başlıklı makalesi her ne kadar gizil sınıf analizi olarak adlandırılmış olsa da bu konuda yapılan ilk çalışma olarak göze çarpmıştır (Bouveyron vd, 2019). İncelenen veri kesikli bir veri olup, 2. dünya savaşı sırasında savaşan askeri personel ile ilgili çalışmalardan elde edilmiştir. Bu çalışmalarda, multinomial dağılımlardan oluşan bir karma model kullanmış ve momentler yöntemiyle parametre tahminini gerçekleştirmiştir. Bunun yanısıra, sonsal olasılıklarla her birey sınıflandırılmıştır.

Wolfe (1963), sürekli veri üzerine yapılmış model tabanlı bir kümeleme analizine bir örnek teşkil ediyordu. Buna rağmen model tabanlı kümeleme analizini daha iyi hale getirmek 2 yıl beklemek gerektirdi. Wolfe (1965), model-tabanlı kümelemeyi sürekli veri için elverişli hale getirmiş bir çalışma olarak tarihte yerini almıştır. Bu çalışmanın 3 noktayla önemini göstermiştir:

i. Başlangıç değerleri ile kümeleme yapmanın önemli olduğunu ifade etmiştir.
ii. Değişken seçiminin önemine vurgu yapmıştır.
iii. Büyük verilerde bilgisayar işlem zamanının uzun sürebileceğini düşünmüş ve küçük bir örneklemden önce kümeleri(Bazı çalışmalarda tip olarak ifade ediliyor) oluşturmayı ve başlangıç değerleri elde ettikten sonra kümeleri bütün veriye uygulamayı önermiştir.

O zamanlardan bu yana çeşitli çalışmalarda karma modeller kullanılmıştır. Punzo ve Ingrassia (2016), çalışmalarında iki boyutlu veri sınıflandırması için ağırlıklandırılmış küme modeli kullanmışlardır. Bu model,

- y yanıt değişken, x açıklayıcı değişken olmak üzere (2.2)'deki denklemi sağlarsa;

$$E(y|x;\beta_{0j},\beta_{1j}) = \beta_{0j} + \beta_{1j}x \tag{2.2}$$

- y yanıt değişken, x açıklayıcı değişken ve ε hata değişkeni olmak üzere (2.3)'teki denklem her karma model bileşeni $p(x,y;\varepsilon_j)$ için sağlanırsa;

$$p(x,y;\varepsilon_j) = p(y|x;\varepsilon_{Y|j})p(x;\varepsilon_{X|j}) \tag{2.3}$$

bir sonlu karma modeldir ve iki boyutlu veri sınıflandırması için ağırlıklandırılmış küme modeli olarak kullanılır. (2.2)'de ve (2.3)'de, x ve Y sırasıyla açıklayıcı değişken matrisi ile yanıt değişkeni gösterirken; $\varepsilon_j = (\varepsilon_{X|j}, \varepsilon_{Y|j})$ olarak temsil edilir. Bu model Monte Carlo simülasyonları ile model tabanlı sınıflandırma yöntemleriyle kıyaslanmış, regresyon katsayılarının tahmin edicileri hesaplanmış ve uygun sayıdaki karma model bileşenini hesaplamak için olabilirlik tabanlı yaklaşımlar kullanılmıştır. Son olarak gerçek yaşam verisine uygulama yapılmıştır.

Kosmidis ve Karlis (2016), çalışmalarında kopula tabanlı sonlu karma modellerin model tabanlı kümelemede uygulamalarını göstermişlerdir. Kopula tabanlı sonlu karma modellerin parametre tahminleri aşağıda adımları verilen EM algoritmasıyla gerçekleştirilir:

- E adımı: i=1,...,n ve j=1,...,k için karma model ağırlık tahminleri olan $w_{ij}^{(l+1)}$ değerleri (2.4)'teki gibi hesaplanır:

$$w_{ij}^{(l+1)} = \frac{\pi_j^{(l)} f_j(x_i; v^{(l)})}{\sum_{j=1}^{k} \pi_j^{(l)} f_j(x_i; v^{(l)})} \tag{2.4}$$

- 1. M adımı: j=1,...,k için düzeltilmiş karma model ağırlık tahminleri olan $\pi_j^{(l+1)} = \sum_{i=1}^{n} w_{ij}^{(l+1)}/n$ olarak belirlenir.
- 2. M adımı: (2.5)'teki amaç fonksiyonu Z(.)'yi $v^{(l)}$'ye göre maksimize et ve çıkan sonucu $v^{(l+1)}$olarak kaydet:

$$Z(v^{(l)}) = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}^{(l+1)} \log[f_j(x_i; v^{(l)})] \tag{2.5}$$

Burada k, bileşen sayısını; $w_{ik}^{(s)}$ s. adımda i. gözlemin k. gruba ait olma olasılığını; $f_j(.)$, j. bileşenin olasılık yoğunluk fonksiyonunu; $\pi_j^{(s-1)}$, (s-1). adımdaki j. bileşenin ağırlığını; v $f_k(.)$'ya ait bütün parametrelerin matris formunu göstermektedir. Sözkonusu çalışmada algoritmanın durdurma kriterini, Log-olabilirlik fonksiyonundaki göreceli değişimin 10^{-8}'den küçük olması olarak belirlendiği söylenmektedir.

Holzmann ve Schwaiger (2016), makalelerinde sonlu karma modelleri kullanarak saklı markov modellerinde durum sayılarını test etmeye çalışmışlardır. Makalede önerilen testlerin asimptotik dağılımı sonlu karma modeller olarak elde edilmiştir. Sözkonusu çalışmada S&P500 endeksi üzerine uygulama yapılmıştır.

Xiang vd. (2016), makalelerinde yarı parametrik sonlu karma modeller için yeni bir tahmin etme yöntemi ortaya koymuştur. Yöntemin getirdiği yenilik, olabilirlik fonksiyonunun hiçbir değişime ya da düzleştirmeye ihtiyaç duymaması ve elde edilen parametre tahminlerinin sapan değerlerden etkilenmemesidir.

Ingrassia ve Punzo (2016), makalelerinde normallik varsayımı altında sonlu karma modelli değişkenlere sahip regresyonların sınıflandırmada sınıfları belirlemede kullanılan karar bölgelerini incelemişlerdir. Ele aldıkları üç tip regresyon tipi vardır: sabit değişkenli, konkomitant değişkenli ve rassal değişkenli. Bu değişkenlere göre regresyon modelleri sırasıyla (2.6), (2.7) ve (2.8)'teki gibidir:

$$P(y|x,\psi) = \sum_{g=1}^{G} f(y|x,\theta_g)\pi_g \tag{2.6}$$

$$P(y|x,\psi) = \sum_{g=1}^{G} f(y|x,\theta_g)\frac{\exp(\alpha_{g0} + \alpha'_{g1}x)}{\sum_{j=1}^{G}\exp(\alpha_{j0} + \alpha'_{j1}x)} \tag{2.7}$$

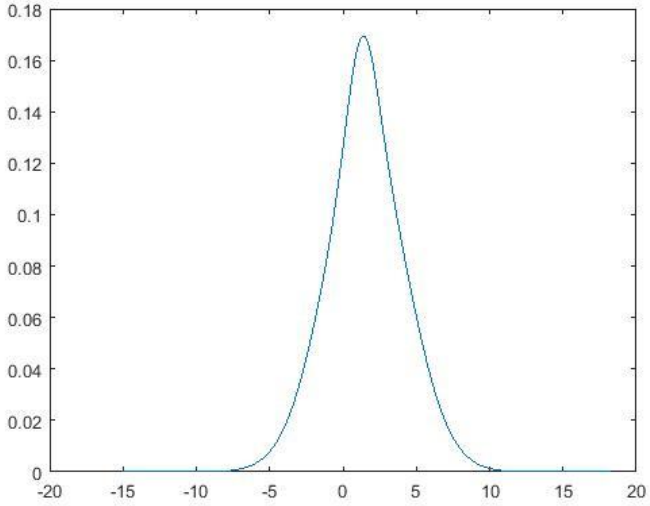
$$P(y|x,\psi) = \sum_{g=1}^{G} f(y|x,\theta_g)p(x;\xi_g)\pi_g \tag{2.8}$$

Burada G, bileşen sayısını; $w_{ik}^{(s)}$ s. adımda i. gözlemin k. gruba ait olma olasılığını; $f_k(.)$, k. bileşenin olasılık yoğunluk fonksiyonunu; $\pi_j^{(s-1)}$, (s-1). adımdaki j. bileşenin ağırlığını; v $f_k(.)$'ya ait bütün parametrelerin matris formunu; θ_g koşullu dağılım için parametreler matrisini; $f(y|x,\theta_g)$ y'nin θ_g ve x'e bağlı koşullu olasılık dağılımını; α_{j0} ile α_{j1} j. grup için katsayıları ve $p(x;\xi_g)$ x'in marjinal olasılık dağılımını göstermektedir. Tek boyutlu olarak ele alındığında, literatürde sonlu karma modellerin üzerinde çok sayıda çalışma olduğu belirtilmektedir(Everitt ve Hand, 1981). (2.1)'deki denklem için tanımlanan parametrelerle tanımlanan ikili sonlu karma modeller için Eisenberger (1964) makalesinde şu iki önermeyi kanıtlamaktadır (Everitt ve Hand, 1981):

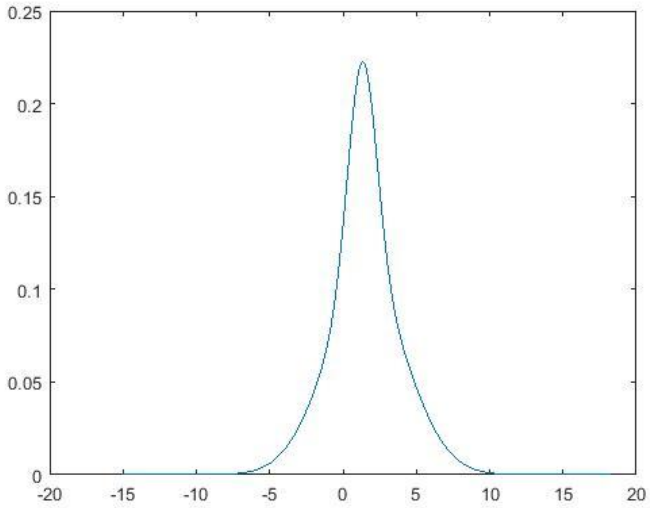
- Şayet $\mu_1 = \mu_2$ ise bu durumda bütün p değerleri için sonlu karma normal model tek modludur.
- Şayet $\mu_1 = \mu_2$ değilse, bütün p değerleri için sonlu normal karma modelin tek modlu olmasının şartı (2.9)'daki eşitsizliğin sağlanmasıdır:

$$(\mu_2 - \mu_1)^2 < \frac{27\sigma_1^2\sigma_2^2}{4(\sigma_1^2 + \sigma_2^2)} \tag{2.9}$$

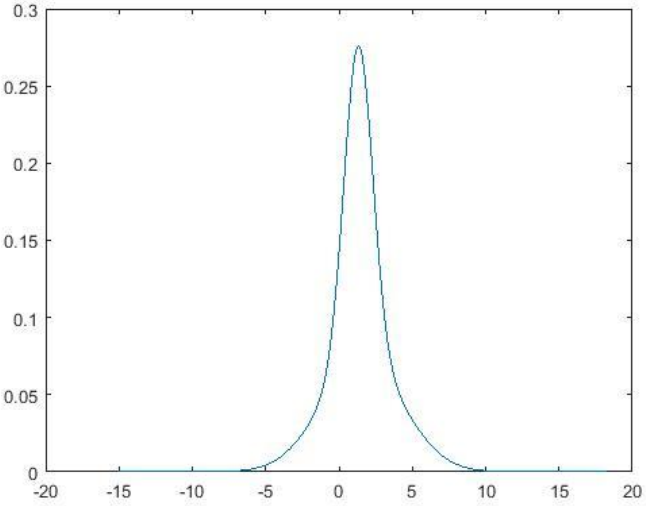
Bu eşitsizliğin sağlandığı birkaç örnek Şekil 2.1, Şekil 2.2, Şekil 2.3'te gösterilmiştir. Şekiller MATLAB R2016a yazılımı kullanılarak oluşturulmuştur:



Şekil 2.1: $\mu_1 = 1{,}3236$ $\mu_2 = 1{,}5756$ $\sigma_1 = 0{,}9754$ $\sigma_2 = 2{,}7850$ p=0,1 parametreli normal dağılımların karma modeli



Şekil 2.2: $\mu_1 = 1{,}3236$ $\mu_2 = 1{,}5756$ $\sigma_1 = 0{,}9754$ $\sigma_2 = 2{,}7850$ p=0,3 parametreli normal dağılımların karma modeli



Şekil 2.3: $\mu_1 = 1{,}3236$ $\mu_2 = 1{,}5756$ $\sigma_1 = 0{,}9754$ $\sigma_2 = 2{,}7850$ p=0,5 parametreli normal dağılımların karma modeli

Normal karma modeller için diğer özellikler şöyle ifade edilir:

i. En az bir p değeri için sonlu normal karma modelin iki modlu olmasının şartı (2.10)'daki eşitsizliğin sağlanmasıdır:

$$(\mu_2 - \mu_1)^2 < \frac{8\sigma_1^2\sigma_2^2}{(\sigma_1^2 + \sigma_2^2)} \qquad (2.10)$$

ii. Bütün $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ değerleri için sonlu normal karma modelin tek modlu olmasını sağlayacak p değerleri mevcuttur.

İkili sonlu karma modeller için Behboodian (1970) makalesinde şu iki önermeyi kanıtlamaktadır (Everitt ve Hand, 1981):

i. Tek modlu sonlu karma normal dağılım elde etmek için gerekli koşul (2.11)'deki eşitsizliğin sağlanmasıdır:

$$|\mu_2 - \mu_1| \leq 2\min(\sigma_1, \sigma_2) \tag{2.11}$$

ii. Şayet $\sigma_1 = \sigma_2 = \sigma$ ise tek modlu karma normal dağılım için gerekli koşul (2.12) ile ifade edilir:

$$|\mu_2 - \mu_1| \leq 2\sigma\sqrt{1 + |\log(p) - \log(1-p)|/2} \tag{2.12}$$

Everitt ve Hand (1981), bu çözümlerin $\mu_1 = \mu_2$ olduğu durumda geçerli olmadığını belirtmiştir. Momentler yönteminin yetmediği durumlarda en çok olabilirlik yöntemi kullanılması gerekmektedir. En çok olabilirlik yöntemine göre olabilirlik fonksiyonu (2.13)'teki gibi ifade edilir:

$$L = \sum_{i=1}^{n} \log_e \left(\sum_{j=1}^{k} p_j g_j(x_i; \mu_k, \Sigma_k) \right) \tag{2.13}$$

Böyle bir durumda $P(s|x_i) = \frac{p_s g_s(x_i;\mu_k,\Sigma_k)}{\sum_{j=1}^{k} p_j g_j(x_i;\mu_k,\Sigma_k)}$ olmak üzere çözümlerden elde edilen tahmin ediciler (2.14), (2.15) ve (2.16) 'deki eşitliklerle ifade edilir:

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{P}(k|x_i) \tag{2.14}$$

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{P}(k|x_i)x_i \tag{2.15}$$

$$\hat{\Sigma}_k = \frac{1}{n\hat{p}_k}\sum_{i=1}^{n} \hat{P}(k|x_i)(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)' \tag{2.16}$$

Burada $\hat{p}_k$ karma modeldeki k. bileşen için ağırlığın; $\hat{\mu}_k$ karma modeldeki k. bileşen için ortalama vektörünün ve $\hat{\Sigma}_k$ karma modeldeki k. bileşen için kovaryans matrisinin tahminlerini

temsil eder. Bu tahmin edicilerin asimptotik kovaryans matrisi Fisher bilgi matrisi'nin tersi ile elde edilir. Hill (1963), tek boyutlu ikili sonlu karma modelde şayet $\sigma_1 = \sigma_2 = \sigma$ ve $(\mu_1 - \mu_2)/\sigma$ değerleri 0,125 ile 0,250 arasında olacak olursa p'nin standart sapmasının 0,1'den küçük olması için en az 1600 birimlik örnekleme ihtiyaç olduğunu göstermiştir. Bunun yanında p'nin standart sapması için gerekli örneklem hacimleri Hill (1963)'te mevcuttur. Behboodian (1972), tek boyutlu ikili sonlu karma modelde bilgi matrisinin elemanlarını hesaplamak için (2.17)'deki integrali hesaplamaya çalışmıştır:

$$M_{mn}(g_i, g_j) = \int_{-\infty}^{\infty} \left(\frac{x - \mu_i}{\sigma_i}\right)^m \left(\frac{x - \mu_j}{\sigma_j}\right)^n \frac{g_i(x) g_j(x)}{f(x)} dx \qquad (2.17)$$

Burada $g_i(x)$ ile $g_j(x)$ karma modeldeki bileşenlerin olasılık yoğunluk fonksiyonlarını; f(x) karma model olasılık yoğunluk fonksiyonunu; μ_i ile σ_i normal dağılım için ortalama ve standart sapmayı göstermektedir. Behboodian (1972), ikili sonlu karma modeller için elde edilmiş bilgi matrisi değerleri yer almaktadır. Bunun yanında bileşen yoğunluk fonksiyonları birbirlerine yaklaştıkça, bilgi matrisinin elemanları sıfıra yaklaşmaktadır. Aynı şey p sıfıra ya da bire yaklaşınca da olmaktadır. Bu nedenle sonlu karma modellerin bileşenlerinden bir şekilde ayrılmadığı durumlarda ya da p sıfıra yakın olduğunda, Fisher bilgi matrisini hesaplamak için büyük örneklemlere ihtiyaç vardır. Everitt ve Hand (1981), ikiden fazla bileşene sahip sonlu karma modeller için de Fisher bilgi matrisinin elde edilebileceğini iddia etmektedir. Chang (1979), ortak bir kovaryans matrisine sahip ikili normal karma modeller için boyut ne olursa olsun Fisher bilgi matrisini hesaplamanın mümkün olduğunu belirtmiştir.

Grafiksel tahmin yöntemleri, en çok olabilirlik ve moment yöntemleriyle elde edilen sonuçları daha güvenilir kılmaya yaramaktadır. Everitt ve Hand (1981), grafiksel tahmin yöntemlerden bazılarına yer vermişlerdir. Örneğin, Harding (1949) normal dağılıma ait P-P grafiklerini kullanarak sonuca gitmeye çalışmıştır. Çalışmasında normal dağılıma ait P-P grafikleri düz bir çizgi halindeyken, sonlu normal karma modellerin P-P grafiklerinin sigmoidal bir grafik oldukları ifade edilmektedir. Makalede P-P grafiklerindeki kıvrımların oluştuğu yerleri inceleyerek bileşenlerin ortalamalarını, varyanslarını ve karma olasılıklarını hesaplamaya çalışılmıştır. Bu yöntem subjektiftir ve bu konu üzerinde daha objektif bir yaklaşım Fowlkes (1979) tarafından sunulmuştur. Bu yaklaşımda, noktalara modifiye edilmiş bir lojistik eğri uydurulmaya çalışılmıştır. Everitt ve Hand (1981), bu yaklaşımın şayet

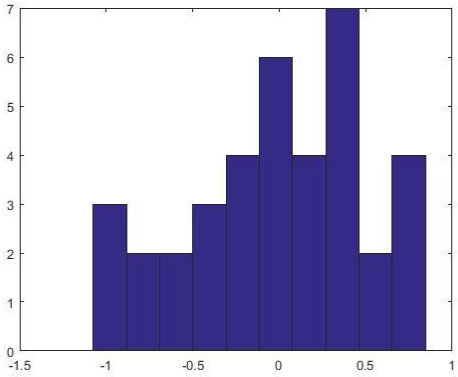
bileşenler birbirlerinden net bir biçimde ayrılmamışsa doğru sonuçlara götürmeyeceğini iddia etmekte ancak başka alanlarda faydalı olabileceğini belirtmektedir. Bir başka çalışma gruplandırılmış veri üzerinde yapılmış bir çalışma olan Bhattacharya (1967)'nındır. Bu makalede i. grup için ortanca nokta x_i ve frekansı ϕ_i olmak üzere, $\log\left(\frac{\phi_{i+1}}{\phi_i}\right)$ değerlerinin x_i değerlerine göre grafiklerini çizdirmeyi esas alınmıştır. Bu noktalara çizgi uydurularak her bir bileşenin alanı oluşturulur ve her bir bileşen için ortalama ve varyans değerleri (2.18) ve (2.19)'daki gibi ifade edilir:

$$\hat{\mu}_k = \lambda_k + w/2 \tag{2.18}$$

$$\hat{\sigma}_k^2 = w\cot(\alpha_k) - w^2/12 \tag{2.19}$$

Burada α_k, k. doğru ve x ekseninin arasındaki açıyı, w oluşturulan alanın genişliğini ve λ_k doğruların x ekseninde kestiği yeri göstermektedir. Bhattacharya (1967), p'nin değerlerini bulmak için yaklaşımlar önermiştir.

Sonlu karma modellerin bileşen sayısı grafiksel olarak ya da hipotetik olarak belirlenebilir. Tek boyutlu örneklemler ele alındığında sonlu karma modellerin bileşen sayısı histogram tarafından belirlenebilmektedir. Tek bir normal dağılımdan elde edilen 50 birimlik örneklemlerin histogramlarında tek mod, iki mod hatta üç mod görüldüğü karma model literatürüne geçmiştir(Everitt ve Hand, 1981; Murphy,1964). Bu sebepten örneklem histogramını incelemek her zaman doğru sonuçlar vermeyebilir. Şekil 2.4'deki histogram, standart normal dağılımdan simüle edilerek elde edilen 50 birimlik bir örneklemin histogramdır:



Şekil 2.4: Örneklem histogramı

Bu grafikte görülüyor ki bileşen sayısı 2 olarak belirlenebilir. Grafiksel olarak bileşen sayısını belirlemek için tek yöntem histogram değildir ve P-P grafikleri de bu işlem için kullanılabilir. P-P grafikleri, önerilen dağılımı sınamak ve konum ve ölçek parametrelerini tahmin etmek için kullanılmaktadır. P-P grafikleri kullanıldığında önerilen dağılım sonlu karma model ise S şeklinde bir grafik çizdirir ve normal dağılımın P-P grafiği ile kıyaslandığında bu daha çok göze batar. Benzer şey çok boyutlu sonlu normal karma modelde uygulanmaktadır. Everitt ve Hand (1981), çok boyutlu normal dağılım için (2.20)'deki dönüşümü yapmayı önermektedir:

$$D_i = (X_i - \overline{X})' S^{-1} (X_i - \overline{X}) \qquad (2.20)$$

Bu denklemde X_i, veriyi; $\overline{X}$, ortalamalar vektörü; S, örneklem kovaryans matrisini göstermektedir. Çok boyutlu normal dağılımdan gelen X_i verisi düz bir doğru şeklinde xy düzleminde kendini gösterir. Şayet çok boyutlu normal karma modelden geliyorsa, bu durumda S şeklinde grafik elde edilir.

Hipotez testi ile veri için ideal k sayıda bileşen kullanılacağına karar vermeye dayanan çalışmalar vardır (Everitt and Hand, 1981). Hipotez testi ile bu işlemi gerçekleştirmek için ilk akla gelenlerden biri olabilirlik oran testidir. Bu testte $H_0: k = k_0$ ve $H_1: k = k_1$ hipotezleri kurulur. Bu hipotezler ışığında olabilirlik oranı (2.21)'deki gibi ifade edilir:

$$\lambda = L_{k_0} / L_{k_1} \qquad (2.21)$$

Burada L_{k_0} ile L_{k_1} sıfır ve alternatif hipotezi kabul ettiğimizde log-olabilirlik fonksiyonlarını temsil etmektedir. Bu istatistik üzerinde orjinal çalışma yapan Wilks (1938), belli koşullar altında $-2\log_e(\lambda)$ istatistiğinin asimptotik olarak parametre sayıları arasındaki fark kadar serbestlik derecesine sahip kikare dağılımına sahip olduğunu belirtmektedir. Ancak Wolfe (1971), H_0 altında $k_0 + 1, \ldots, k_1$ değerleri için p değerleri otomatik olarak 0 olacağından asimptotik olarak bunun gerçekleşmeyeceğini iddia etmektedir. Bunun yerine Wolfe (1971), WT ile gösterilen (2.22)'deki test istatistiğini önermektedir:

$$WT = -\frac{2}{n}\left(n - 1 - d - \frac{c_1}{2}\right)\log_e(\lambda) \tag{2.22}$$

Burada n örneklem hacmini ve d boyutu göstermektedir. Bu istatistik $2d(c_1 - c_0)$ serbestlik dereceli kikare dağılımına sahip olmakla birlikte c_1 arttırılarak karma model bileşen sayısı belirlenir. Everitt (1981)'in bu testi n>10d için makul bulmakta ve iki bileşen için testin gücünün düşük olduğunu belirtmektedir. Hasselblad (1969) ise bu testin gücünün üssel, Poisson ve binom dağılımlarının sonlu karma modeller için tatmin edici olduğunu söylemiştir.

Karma modellerin uygulamaları üzerine yapılan bahsigeçen temel çalışmalar, istatistik biliminin veriye bağlı metot ve metodolojileri nasıl incelediğini anlatmak amacıyla bu bölümde bahsedilmiştir. İlerleyen bölümlerde öncelikle karma modellerin parametrelerinin tahmin edilme yöntemleri, model tabanlı kümeleme örnekleri, model tabanlı kümeleme yöntemleri üzerine literatürde geliştirilen esaslar ve panel veride kümeleme yapmanın yöntemleri anlatılacaktır.

2.2 Karma Modellerde Tahmin Yöntemleri

Herhangi bir d boyutlu X rassal vektörü için sonlu normal karma model, (2.23)'teki şekilde ifade edilebilir (Everitt ve Hand, 1981):

$$f(x; [p_1 \quad \cdots \quad p_k], [\mu_1 \quad \cdots \quad \mu_k], [\Sigma_1 \quad \cdots \quad \Sigma_k]) = \sum_{i=1}^{k} p_i \frac{\exp\left(-\frac{(X-\mu_i)'\Sigma_i^{-1}(X-\mu_i)}{2}\right)}{(2\pi)^{d/2}\sqrt{|\Sigma_i|}} \tag{2.23}$$

Burada k, bileşen sayısını; p_i ağırlıkları, μ_i ortalama vektörlerini, Σ_i kovaryans matrislerini ve $|.|$ determinant alma işlemini temsil etmektedir.

Karma modeller için geliştirilen parametre tahmin yöntemleri şunlardır (Everitt ve Hand, 1981):

1. En çok olabilirlik yöntemi
2. Bayesçi tahmin yöntemi
3. Hata azaltmalı tahmin yöntemleri
    a. Kabir (1968)'in aralıkları
    b. Bartholomew (1959)'un yöntemi
    c. Momentler yöntemi
    d. Moment çıkartan fonksiyon yöntemi

2.2.1 En çok olabilirlik yöntemi

Herhangi bir karma model için en çok olabilirlik fonksiyonu (2.24)'teki şekilde ifade edilir (Everitt ve Hand, 1981):

$$L(X_1, \dots, X_n, \theta_1, \dots, \theta_k, p_1, \dots, p_k) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} p_j g_j(x_i; \theta_j) \right] \tag{2.24}$$

Burada k, bileşen sayısını; p_j karma model ağırlıklarını; $g_j(.;.)$ j. bileşen için olasılık yoğunluk fonksiyonunu ve θ_j de j. bileşen için parametreleri temsil etmektedir. Böyle bir durumda $\text{Log}[L(X_1, \dots, X_n, \theta_1, \dots, \theta_k, p_1, \dots, p_k)] - \lambda\left(\sum_{j=1}^{k} p_j - 1\right)$ fonksiyonu ile tahmin edicilerin denklemleri (2.25) ve (2.26)'daki gibi elde edilir:

$$\frac{\partial L}{\partial p_j} = \sum_{i=1}^{n} \frac{g_k(x_i; \theta_j)}{\sum_{l=1}^{k} p_l g_l(x_i; \theta_l)} - \lambda = 0 \tag{2.25}$$

$$\frac{\partial L}{\partial \theta_{jk}} = \sum_{i=1}^{n} p_k \frac{\partial g_k(x_i; \theta_j) / \partial \theta_{jk}}{\sum_{l=1}^{k} p_l g_l(x_i; \theta_l)} = 0 \tag{2.26}$$

Bu durumda $\hat{p}_k$ tahmin edicileri (2.27)'deki formülle elde edilir:

$$\hat{p}_k = \frac{1}{n}\sum_{j=1}^{n} P(k|x_j) \tag{2.27}$$

Bu formülle geri kalan parametreler ikinci denklemle hesaplanır. Bu işlem EM algoritmasında da aynı şekilde uygulanır. Önce $\hat{p}_k$ tahmin edicisi sonra θ_{jk} değeri hesaplanır.

En çok olabilirlik tahmin yönteminin karma modellerdeki uygulaması son derece karmaşıktır. Bu karmaşıklığın kaynağı olabilirlik fonksiyonunun yapısıdır. Bu sorundan kurtulmak için EM algoritması geliştirilmiştir. EM algoritması ile normal karma modeldeki parametreler tahmin edilerek model tabanlı kümeleme yapılır. EM algoritması uygulanırken öncelikle E (Beklenti-Expectation) adımında $\pi_{ik}^{(s)}$ ile gösterilen sonsal olasılıklar (2.28)'deki eşitlik ile hesaplanır:

$$\pi_{ik}^{(s)} = P(X_i \in k.\,küme|X_i; v^{s-1}) = \frac{\pi_k^{(s-1)} f_k\left(x_i; v_k^{(s-1)}\right)}{\sum_{k'=1}^{K} \pi_{k'}^{(s-1)} f_k\left(x_i; v_{k'}^{(s-1)}\right)} \tag{2.28}$$

Burada K, bileşen sayısını; $\pi_{ik}^{(s)}$ s. adımda i. gözlemin k. gruba ait olma olasılığını; $f_k(.)$, k. bileşen olasılık yoğunluk fonksiyonunu; $\pi_k^{(s-1)}$, (s-1). adımdaki k. bileşenin ağırlığını ve v $f_k(.)$'ya ait bütün parametrelerin matris formunu göstermektedir. Sözkonusu olasılıklar hesaplandıktan sonra EM algoritmasının M (en çoklama- maximization) adımında geri kalan parametreler hesaplanır. E ve M adımları yakınsama kriterleri sağlanana kadar tekrarlanır. Çok boyutlu normal karma modellerde üç tip parametre bulunmaktadır: Sonsal olasılıklar, ortalama ve kovaryans matrisleri. Bu parametrelerden kovaryans matrislerinin güncellenmesi değişebilmektedir. Ancak şayet kovaryans matrisinin yapılandırılmamış olduğu varsayılırsa, bu durumda güncelleme eşitlikleri sırasıyla (2.29), (2.30), (2.31)'deki gibi olur:

$$\pi_k^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \pi_{ik}^{(s)} \tag{2.29}$$

$$\mu_k^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}^{(s)} x_i}{\sum_{i=1}^{n} \pi_{ik}^{(s)}} \tag{2.30}$$

$$\Sigma_k^{(s)} = \frac{\sum_{i=1}^{n} \pi_{ik}^{(s)} \left(x_i - \mu_k^{(s)}\right)\left(x_i - \mu_k^{(s)}\right)'}{\sum_{i=1}^{n} \pi_{ik}^{(s)}} \tag{2.31}$$

Burada $\mu_k^{(s)}$ s. adımda k. bileşenin ortalama vektörünü; $\Sigma_k^{(s)}$, k. bileşenin kovaryans matrisini; $\pi_j^{(s-1)}$, (s-1). adımdaki j. bileşenin ağırlığını ve $\pi_{ik}^{(s)}$ s. adımdaki k. bileşenin i. kümeye ait olma olasılığını göstermektedir. EM algoritmasını durdurmak için değişik kriterler ortaya konmuştur. En olağan durdurma kriteri Böhning vd. (1994) tarafından ortaya atılmıştır. Ancak şayet normal karma modellerde heterojen yayılmalar varsa bu durumda olabilirlik fonksiyonu sınırlanmamış olabilmektedir. Bu durumda olabilirlik fonksiyonunun genel bir maksimum noktası bulunmadığından bu soruna çözüm getirilmeye çalışılmıştır. Mclachlan ve Peel (2000), çok boyutlu normal karma modeller için bu sefer herhangi bir i ve j için $|\Sigma_i|^{-1}|\Sigma_j| \geq c > 0$ olması halinde bu sorunu çözmeye çalışmıştır. Normal karma modeller söz konusu olduğunda kümeleme analizinde en çok kullanılan algoritma, EM algoritmasıdır (Wolfe, 1970). Bunun yanında, kümeleme analizinde değiştirilmiş haldeki EM algoritmaları da faydalı olmuştur. Bu konuda referans verebileceğimiz iki algoritma Stokastik EM (SEM) algoritması (Celeux ve Diebolt, 1985) ve Sınıflandırıcı EM algoritması (CEM) algoritmasıdır (Celeux ve Govaert, 1992). Bunların yanında Bayesçi kümelemenin gelişmesi, kümeleme yöntemini yeniden düşünmeye sevketmiştir.

Bayesçi kümeleme ve model tabanlı kümelemenin karşılaştığı sorun, büyük verilerde ortaya çıkmaktadır. Bayesçi kümelemede Markov Zinciri Monte Carlo yönteminde ve EM algoritmasında Log-olabilirlik fonksiyonu hesaplanırken büyük veriler kullanılacağından bu yöntemler hesaplama zamanı açısından sorun yaratacaklarından bu soruna çözüm getirmek gerekmiştir. Bunun için Banfield ve Raftery (1993)'ye, Roeder ve Wasserman (1997)'ye ve Posse (2001)'ye bakılabilir. Çok boyutlu normal karma model için örneklemenin log-olabilirlik fonksiyonu (2.32)'deki gibi ifade edilir:

$$\mathrm{Logp}(y|S,\mu,\Sigma)$$

$$= -\frac{1}{2}\sum_{k}\left(N_k(S)\log|\Sigma_k(S)| + \sum_{i:S_i=k}(y_i-\mu_k(S))'\Sigma_k(S)^{-1}(y_i-\mu_k(S))\right)+c \quad (2.32)$$

Burada $N_k(S)$ k. kümenin büyüklüğü; $\Sigma_k(S)$ k. kümenin kovaryans matrisi; $\mu_k(S)$ k. kümenin ortalama varyansı; S_i ise kümenin indisini; y_i veriyi; $|.|$ determinant işlemini ve c de işleme dahil olmayan sabiti göstermektedir.

$\mathrm{Logp}(y|S,\mu,\Sigma)$ fonksiyonu çok faydalı bir fonksiyondur. Bunun nedeni bu fonksiyon, çeşitli şekillerde kullanılabilir olmasıdır. Buna bir örnek (2.33)'teki gibidir:

$$\mathrm{Logp}(y|S,\mu,\Sigma) = -\frac{1}{2}\sum_{k}\left(N_k(S)\log|\Sigma_k(S)| + \mathrm{tr}(W_k(S)(\Sigma_k(S))^{-1})\right)+c \quad (2.33)$$

Burada $W_k(S)$ haricindeki diğer parametreler (2.32)'den hemen sonra tanımlandığı şekilde tanımlanır. $W_k(S)$ ise (2.34)'deki gibi tanımlanır:

$$W_k(S) = \sum_{i:S_i=k}\left(y_i-\bar{y}_k(S)\right)\left(y_i-\bar{y}_k(S)\right)' \quad (2.34)$$

2.2.2 Bayesçi tahmin yöntemi

Bayesçi yaklaşımda herhangi bir parametre vektörü α'ya $f(\alpha|X_1,\dots,X_n)$ koşullu olasılık yoğunluk fonksiyonunu oluşturarak tahmin ederiz. ve $f(\alpha|X_1,\dots,X_n)$ için formül (2.35)'deki sonsal olasılık fonksiyonunundan elde edilir:

$$f(\alpha|X_1,\dots,X_n)=\frac{L(X_1,\dots,X_n;\alpha)f(\alpha)}{\int L(X_1,\dots,X_n;\alpha)f(\alpha)d\alpha} \tag{2.35}$$

Burada L(.;.) olabilirlik fonksiyonunu ve f(.) parametre vektörü α'nın olasılık yoğunluk fonksiyonunun formülünü ifade eder. Everitt ve Hand (1981), sonlu karma modellerde Bayesçi yöntemi iki zorluk çıkardığından bahsetmiştir. Bunlardan biri belirlenebilirlik zorluğudur. Herhangi bir sonlu karma model, (2.36)'teki formül geçerli ise o sonlu karma model belirlenebilirdir:

$$f(x)=\sum_{j=1}^{k} p_j g_j(x;\theta_j)=\sum_{l=1}^{k'} \hat{p}_l g_l(x;\hat{\theta}_l) \tag{2.36}$$

Burada k, bileşen sayısını; p_j karma modelde j. bileşen için ağırlığı, $g_j(.;.)$ j. bileşen için olasılık yoğunluk fonksiyonunu temsil eder. (2.36)'ya göre $k=k'$ ve en az bir l için $p_j=\hat{p}_l$ ve $\theta_j=\hat{\theta}_l$ olmaktadır. İşte Bayesçi yöntemin zorluğu, bu yöntemle elde edilen tahmin edicilerin bu şartı sağlamamasıdır. İkinci zorluk hesaplamaların zorlaşmasıdır. Bunu aşmak için α'nın yeterli bir tahmin edicisi olan $\hat{\alpha}$ denkleme dahil edilebilir ancak sonlu karma model örneklerinde yeterli istatistik elde etmenin çok zor olduğu belirtilmiştir.

2.2.3 Hata azaltmalı tahmin yöntemleri

Bu bölümde hataları en aza indirecek şekilde tasarlanmış parametre tahmin yöntemleri incelenmiştir.

2.2.3.1 Kabir'in aralıkları

Kabir (1968)'in yöntemine göre $\sum_{i=1}^{G} p_i=1$ olmak üzere sonlu karma model (2.37)'deki gibi tanımlanır:

$$f(x) = \sum_{i=1}^{G} p_i g_i(x; \mu_i) \tag{2.37}$$

Burada G, bileşen sayısını; $g_i(x; .)$ üstel dağılım ailesine ait x rassal değişkenin olasılık yoğunluk fonksiyonunu, her i için μ_i x'e ait dağılımın parametrelerini ve p_i karma modeldeki ağırlıkları temsil eder. Bu durumda Kabir (1968)'in oluşturduğu denklemler j=0,...,2c-1 için (2.38)'deki gibi ifade edilir:

$$\sum_{i=1}^{k} A_i \left(\lambda_i(\mu_i)\right)^j = \phi_j \tag{2.38}$$

Burada ϕ_j veriden elde edilmesi gereken uygun değerleri, A_i sıfırdan farklı katsayıları ve $\lambda_i(\mu_i)$ ise μ_i'ye bağlı monoton artan bir fonksiyonu temsil etmektedir. Kabir (1968)'in uyguladığı yönteme göre $\lambda_i(\mu_i) = \lambda_i$ ve $\beta_0 = 1$ olmak üzere (2.39)'daki eşitlikten bulunur:

$$\sum_{i=0}^{k} \beta_i \lambda_i^{k-i} = 0 \tag{2.39}$$

Bu denklemden bulunan $\beta_0 = 1$ hariç β_i'ler l=0,...,c-1 için (2.40)'daki eşitsizliklerden bulunur:

$$\sum_{j=0}^{k-1} \phi_{j+l} \beta_{k-j} < \phi_{k+l} \tag{2.40}$$

Kabir (1968)'in aralıkları böyle oluşturulmuş olur. Bunlar bir denklem formuna getirilerek çözülür. Bu aşama tamamlandıktan sonra μ_i'lerin bulunması gerekir. Sözkonusu parametrelerin bulunması başta tanımlanan $\lambda_i(\mu_i)$ fonksiyonlarının formüllerine göre

değişmektedir. Everitt ve Hand (1981), bu yolla elde edilen tahmin edicilerin istikrarlı olduklarını ve normal dağıldıklarını belirtmektedir.

2.2.3.2 Bartholomew'in aralıkları

Bartholomew (1959)'in aralıkları elde etme biçimi Kabir (1968)'inki gibi olup tek farkı aralığın $(-\infty, b)$ şeklinde değil (a, ∞) şeklinde olmasıdır (Everitt ve Hand, 1981).

2.2.4 Momentler yöntemi

Momentler yöntemi, kitleye atfedilen dağılım parametrelerini örneklemden elde edilen momentler yardımıyla çözmeye çalışan pratik bir tahmin yöntemidir. Bu yönteme göre dağılımın parametrelerinin tamamını tahmin edecek kadar sayıda moment formüllerini örneklemdeki karşılıklarına eşitler. Bütün momentler $E(.)$ şeklinde olduğu için momentlerin karşılıkları $\bar{.}$ şeklinde gösterilen örneklemdeki gözlemlerin ortalaması olur. Bu pratik yöntemin uygulanış zamanı 19. yüzyıla kadar gitmektedir.

Everitt ve Hand (1981), kitaplarında en erken sonlu karma model çalışmalarından birinin Pearson (1894) olduğunu ve onun da incelediği sonlu normal karma model'in 5 parametresini momentler yöntemiyle bulduğunu söylemektedir. Bunun yanında sonlu karma modellerde parametre tahmin etmek için en popüler yöntemlerden birinin momentler yöntemi olduğu belirtilmiştir. Bunun yanında Kendall ve Stuart (1963), sonlu karma modellerin parametre sayısı arttıkça örneklem varyanslarının da arttığı ve örneklem büyük olsa da tahmin edicilerin varyansı çok yüksek olacağı belirtilmiştir. Bu da tahmin ediciler için bir sorundur. Buradan tek boyutlu normal karma modeller için momentler yönteminin uygulanışına geçilecektir

Momentler yönteminde örneklem için $\bar{x}$ ortalamayı, n örneklem hacmini ve x_i verideki değerleri temsil etmek üzere (2.41)'deki eşitlik tanımlanır:

$$V_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r \tag{2.41}$$

$\delta_i = \mu_i - \mu$ ve (2.1)'den sonra tanımlanan parametreler geçerli olmak üzere bu değerler ile (2.42), (2.43), (2.44), (2.45), (2.46)'daki denklemler oluşturulur:

$$p\delta_1 + (1-p)\delta_2 = 0 \tag{2.42}$$

$$p(\sigma_1^2 + \delta_1^2) + (1-p)(\sigma_2^2 + \delta_2^2) = V_2 \tag{2.43}$$

$$p(3\delta_1\sigma_1^2 + \delta_1^3) + (1-p)(3\delta_2\sigma_2^2 + \delta_2^3) = V_3 \tag{2.44}$$

$$p(3\sigma_1^4 + 6\delta_1^2\sigma_1^2 + \delta_1^4) + (1-p)(3\sigma_2^4 + 6\delta_2^2\sigma_2^2 + \delta_2^4) = V_4 \tag{2.45}$$

$$p(15\sigma_1^4\delta_1 + 10\delta_1^3\sigma_1^2 + \delta_1^5) + (1-p)(15\sigma_2^4\delta_2 + 10\delta_2^3\sigma_2^2 + \delta_2^5) = V_5 \tag{2.46}$$

Bu denklemler indirgenerek $\sum_{i=0}^{9} a_i u^i = 0$ forma dönüştürülür (a_i değerleri için Everitt ve Hand (1981)'e bakınız). Buradan elde edilen gerçel negatif u çözümü (2.47)'deki denklemde yerine konur:

$$\delta^2 - \frac{w}{u}\delta + u = 0 \tag{2.47}$$

Bu durumda (2.47)'deki w, (2.48)'deki gibi ifade edilir:

$$w = \frac{-8V_3u^3 + 3(V_5 - 10V_3V_2)u^2 + 6V_3(V_4 - 3V_2^2)u + 2V_3^3}{2u^3 + 3(V_4 - 3V_2^2)u + 4V_3^2} \tag{2.47}$$

Bu yolla elde edilen 5 parametrenin tahmin edicileri (2.49), (2.50), (2.51), (2.52) ve (2.53)'deki gibidir:

$$\hat{\mu}_1 = \bar{x} + \hat{\delta}_1 \tag{2.49}$$

$$\hat{\mu}_2 = \bar{x} + \hat{\delta}_2 \tag{2.50}$$

$$\hat{\sigma}_1^2 = \frac{1}{3}\hat{\delta}_1\left(\frac{2w}{u} - \frac{V_3}{u}\right) + V_2\hat{\delta}_1^2 \tag{2.51}$$

$$\widehat{\sigma}_2^2 = \frac{1}{3}\hat{\delta}_2\left(\frac{2w}{u} - \frac{V_3}{u}\right) + V_2\hat{\delta}_2^2 \tag{2.52}$$

$$\hat{p} = \frac{\hat{\delta}_2}{\hat{\delta}_2 - \hat{\delta}_1} \tag{2.53}$$

2.2.5 Moment çıkartan fonksiyon yöntemi

Moment çıkartan fonksiyon yöntemi momentler yönteminin farklı bir şekli olarak ifade edilmektedir. Bu yöntemde moment çıkartan fonksiyonunun (2.54)'deki $M(\beta)$ tahmin edicisi hesaplanır:

$$\widehat{M(\beta)} = E\big(\exp(\beta x)\big) \tag{2.54}$$

Burada $M(\beta)$ β'ya göre çeşitli değerler alır. Moment çıkartan fonksiyon için doğal tahmin edici, (2.55)'deki ifadedir:

$$\widehat{M(\beta)} = E\big(\widehat{\exp(\beta x)}\big) = \frac{1}{n}\sum_{i=1}^{n}\exp(\beta x_i) \tag{2.55}$$

Burada n örneklem hacmini temsil etmektedir. Hata fonksiyonu olarak da (2.56)'daki fonksiyon kullanılabilir:

$$E(\alpha) = \sum_{j=1}^{m}\left(E\left(\exp(\beta_j x)\right) - \sum_{i=1}^{n}\exp(\beta_j x_i)/n\right)^2 \tag{2.56}$$

Burada β_j'lerin seçiminin nasıl yapılacağı açık değildir. Bu noktaya kadar anlatılanlar, karma modelleri incelenen veri ile uyumlu bir şekilde oluşturmak üzerine anlatılan temel

istatistiksel bilgiden ibarettir. Bu noktadan itibaren sınıflandırma için geliştiren model tabanlı sınıflandırma üzerine yöntemleri inceleyeceğiz.

2.3 Model Tabanlı Kümelemede Sınıflandırma Kriterleri

Literatürde göze çarpan optimal sınıflandırma kriterlerinden biri (2.57)'deki eşitlikle ifade edilir:

$$c(S) = -\frac{N}{2}\log|W(S)| + c \tag{2.57}$$

Burada n örneklem hacmi, W(S) toplam grup içi varyansı ve c önemli olmayan sabit bir sayıyı temsil eder. Şayet $\Sigma_k = \sigma^2 I_{rxr}$ olsaydı, bu durumda (2.57)'deki kriter (2.58)'ye dönüşecekti:

$$Logp(y|S, \mu, \Sigma) = -\frac{1}{2}(rnlog(\sigma^2) + tr(W(S))/\sigma^2) + c \tag{2.58}$$

Bunun dışında sınıflandırmada en sık kullanılan kriter (2.59)'daki gibi ifade edilir:

$$c(S) = -\frac{rN}{2}\log\big(tr(W(S))\big) + c \tag{2.59}$$

Burada $tr(W(S))$ küçüldükçe c(S)'nin büyüyeceğini ve daha iyi sınıflandırma yapılacağı görülmektedir. Şayet kovaryans matrisleri için bir kısıtlama getirilmeseydi, bu durumda optimal sınıflandırma kriteri (2.60)'daki gibi ifade edilecekti:

$$c(S) = -\frac{1}{2}\sum_k \left(N_k(S)\log\left|\frac{W_k(S)}{N_k(S)}\right|\right) + c \tag{2.60}$$

Burada $N_k(S)$, k. küme için örneklem hacmini temsil eder. Bunun yanında $\Sigma_k = \sigma_k^2 I_{rxr}$ olduğunda, bu durumda optimal sınıflandırma kriteri (2.61)'deki şekilde ifade edilecekti:

$$c(S) = -\frac{1}{2}\sum_k \left(N_k(S)\log\left(\sigma_k^2\right) + tr\left(W_k(S)/\sigma_k^2\right)\right) + c \tag{2.61}$$

Bu ifade, $\hat{\sigma}_k^2(S) = tr(W_k(S))/N_k(S)$ için maksimize edilir. EM algoritması için başlangıç noktası seçmek çok önemlidir çünkü olabilirlik fonksiyonlarında birden fazla yerel maksimum bulunmaktadır. Bu yüzden başlangıç noktası belirlemek için birden fazla yöntem önerilmiştir. Fraley ve Raftery (2006) tarafından geliştirilen ve R paketi olan Mclust paketinde kullanılan yöntemin bileşenler iyi ayrılmışken çok işe yaradığı ancak tam tersi durumda işe yaramadığı görülmüştür. Maitra (2009)'nın yönteminin düşük boyutlu düzlemlerde işe yaradığı ancak çok boyutlu düzlemlerde zaman tüketici bir performans göstermiştir. Bunun gibi birçok örnek bulunmaktadır. Normal karma modellerin haricinde Poisson karma modeli de önemli sayılabilecek modellerdendir. Bu karma modelin olasılık yoğunluk fonksiyonu (2.62)'deki şekildedir:

$$f(y_i|\mu_i, \tau) = \sum_{g=1}^{G} \tau_g \prod_{j=1}^{d} pois\left(y_{ij}|\mu_{ijg}\right) \tag{2.62}$$

Burada G, bileşen sayısını; τ_g, karma modeldeki ağırlığı ve μ_{ijg}, g. bileşen için ortalamayı ifade etmektedir. $\mu_{ijg}^{(r-1)} = w_i^{(r-1)}\lambda_{jg}^{(r-1)}$ olacak şekilde ayrım yaptıktan EM algoritması ile parametresi hesaplanabilmektedir. Bu yönüyle ilginç bir model olarak not düşülebilecek bir modeldir. Bileşen sayısının belirlenmesinin üzerine pek çok metot geliştirilmiştir (Melnykov ve Maitra, 2017). Bu çalışmalarda geliştirilen kriterler modellerin kıyaslanmasını kolaylaştırmıyor ve hipotez testi yaklaşımları da olabilirlik oran testini gerekli kıldığı halde teorik açıdan uygulanması problemlidir. Bu problemleri aşmak adına literatürde LRT test istatistiklerinin bootstrap yöntemiyle dağılımları belirlenmeye çalışılmış. Bu bootstrap yöntemlerinin uygulanması Mclachlan (1987) ve Aitkin vd. (1981) çalışmalarında olmuştur. Maitra ve Melnykov (2010) da bu konuda çalışma yapmışlardır.

2.4 Model Tabanlı Kümeleme Üzerine Örnekler ve Literatürde Geçen Model Tabanlı Kümeleme Üzerine Esaslar

Literatürde karma modeller ortak olarak model tabanlı kümeleme analizi için kullanılsa da bu yöntemlerde parametre tahmin etme ve karma model uyumunun sınanması gibi konularda model tabanlı kümeleme örnekleri çeşitlilik göstermektedir. Bu nedenle model tabanlı kümeleme analizlerinde, veri ile ilgili önbilgi sahibi olmak önemli olabilir. Literatürde geçen örnekler model tabanlı kümeleme açısından metodolojik ortaklık içerdiği için seçilmiş ve çalışmada bahsedilmeye değer görülmüştür. Bu ortaklık, parametre tahmini hesaplama, veriye uygun karma modeli belirleme ve model uyumunu inceleme ortaklığıdır. Model tabanlı kümelemede hesaplama açısından çıkabilecek sorunlar ve model tabanlı kümelemede model seçimi konuları önümüzdeki bölümlerde işlenecektir.

2.4.1 Kesikli veri sınıflandırılmasında kullanılan gizil sınıf modeli

Gizil Sınıf Modeli (GSM/LCM-Latent Class Model), d kategori'den oluşan veri için oluşturulmuş bir modeldir. Goodman (1974) tarafından formülleştirilmiştir. Bu modelde verideki her bireyin çok terimli dağılıma sahip olduğu varsayılarak karma model (2.63)'deki gibi formülleştirilmiştir:

$$f(y_i;\theta) = \sum_{g=1}^{G} \tau_g \prod_{j,h} \left(\alpha_g^{jh}\right) y_i^{jh} = \sum_{g=1}^{G} \tau_g M_g(y_i;\alpha_g) \qquad (2.63)$$

Burada G, bileşen sayısını; τ_g karma model g. bileşen ağırlığını, y_i^{jh} i. değişkenin gerçekleşip gerçekleşmediğini 1 veya 0 değerini alarak belirleyen gölge değişkeni ve α_g^{jh} g. küme'de j. değişkeninin h seviyesine sahip olma olasılığını ifade etmektedir. (2.63)'deki bu tanımlamalar bölüm boyunca geçerli olmak üzere sözkonusu model değişkenlerin koşullu olarak birbirinden bağımsız olduğunu varsaymaktadır. Bu modelin tanımlanabilir olması için

gerekli koşul, kategorik değişkenlerin alabileceği değerlerin sayısı d'nin (2.64)'teki eşitsizliği sağlamasıdır:

$$2^{d} - 1 \geq (G - 1) + dG \tag{2.64}$$

Bunun yanında Gyllenberg vd. (1994), bu koşulun yeterli olmadığını ifade etmiştir. Carreira-Perpiñan ve Renals (2000) bu iddia'yı taşımış ve Allman vd. (2009) iddia'yı ispatlamış ve yeterlilikler için koşullar öne sürmüştür. Allman vd. (2009), m seviyeli kesikli veriler için $\lceil . \rceil$ fonksiyonu tamdeğer fonksiyonu olmak üzere yeterli koşul olarak $d \geq 2\lceil \log_m G \rceil + 1$ 'i olması gerektiğini öne sürmüştür. GSM modeli için log-olabilirlik fonksiyonu (2.65)'teki gibi ifade edilir:

$$L(\theta; y) = \sum_{i=1}^{n} \log\left(\sum_{g=1}^{n} \tau_g \prod_{j=1}^{d} \prod_{h=1}^{m_j} \left(\alpha_g^{jh}\right)^{y_i{}^{jh}} \right) \tag{2.65}$$

Burada (2.63) için yapılan tanımlamalar geçerlidir. Bu model için EM algoritmasının E ve M adımları aşağıdaki şekilde çalışır:

E adımı: r. adımda y_i için g. kümeye düşme koşullu olasılığı $z_{ig}^{(r)}$, (2.66) ile hesaplanır:

$$z_{ig}^{(r)} = \frac{\tau_g^{(r)} M_g\left(y_i; \alpha_g^{(r)}\right)}{\sum_{g=1}^{G} \tau_g^{(r)} M_g\left(y_i; \alpha_g^{(r)}\right)} \tag{2.66}$$

M adımı: $\tau_g^{(r)}$ ve $\alpha_g^{jh\,(r)}$ parametreleri küme için (2.67) ve (2.68)'daki gibi güncellenir:

$$\tau_g^{(r+1)} = \frac{n_g^{(r)}}{n} \tag{2.67}$$

$$\alpha_g^{jh\,(r+1)} = \frac{\sum_{i=1}^{n} z_{ig}^{(r)} y_i^{jh}}{\sum_{i=1}^{n} z_{ig}^{(r)}} = \frac{u_g^{jh\,(r)}}{n_g^{(r)}} \tag{2.68}$$

Burada $(.)^{(r)}$, r. iterasyonda (.) değerini belirtmektedir. $\sum_{i=1}^{n} z_{ig}^{(r)} = 0$ olduğundan $\alpha_g^{jh\,(r)}$ elde edilemez ise bu durumda güncelleme için (2.69)'daki formül kullanılır:

$$\alpha_g^{jh\,(r)} = \frac{u_g^{jh\,(r)} + c - 1}{n_g^{(r)} + m_j(c-1)} \tag{2.69}$$

Burada c, sabit bir sayı; m_j, j. değişkenin seviye sayısını belirtmektedir. Bu durumda (2.65)'daki log-olabilirlik fonksiyonu (2.70)'deki log-olabilirlik fonksiyonuna dönüştürülür:

$$L_c(\theta) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \log\left(\tau_g \prod_{j=1}^{d}\prod_{h=1}^{m_j}\left(\alpha_g^{jh}\right)^{y_i^{jh}}\right) \tag{2.70}$$

Bu fonksiyondan hareketle EM algoritmasının E ve M adımları aşağıdaki şekilde olur:

E adımı: z_{ig}, (2.71)'deki gibi hesaplanır:

$$z_{ig}^{(r)} = \frac{\tau_g^{(r)} M_g\left(y_i; \alpha_g^{(r)}\right)}{\sum_{j=1}^{G} \tau_g^{(r)} M_j\left(y_i; \alpha_j^{(r)}\right)} \tag{2.71}$$

M adımı: (2.72), (2.73), (2.74) ve (2.75)'daki eşitliklerle parametre tahmin edicileri hesaplanır:

$$\tau_g^{(r+1)} = \frac{n_g^{(r)}}{n} \tag{2.72}$$

$$\left(\alpha_g^{jh}\right)^{(r+1)} = \frac{\left(u_g^{jh}\right)^{(r)}}{n_g^{(r)}} \tag{2.73}$$

$$n_g^{(r)} = \sum_{i=1}^{n} z_{ig}^{(r)} \tag{2.74}$$

$$\left(u_g^{jh}\right)^{(r)} = \sum_{i=1}^{n} z_{ig}^{(r)} y_i^{jh} \tag{2.75}$$

Şayet α_g^{jh} hesaplanamazsa, düzeltme terimli olarak (2.76)'deki gibi ifade edilir:

$$\left(\alpha_g^{jh}\right)^{(r+1)} = \frac{\left(u_g^{jh}\right)^{(r)} + c - 1}{n_g^{(r)} + m_j(c-1)} \tag{2.76}$$

α_g^{jh} üzerinden yeniden ifade edildiğinde 5 çeşit GSM modeli oluşturulabilir. Sözkonusu parametre oluşturma yöntemi ile (2.77) ile (2.78)'deki parametreler elde edilir:

$$\varepsilon_g^{jh} = \begin{cases} 1 - \alpha_g^{jh} & h = a_g^j \\ \alpha_g^{jh} & \text{diğer durumlarda} \end{cases} \tag{2.77}$$

$$a_g^j = \text{argmax}_h \alpha_g^{jh} \tag{2.78}$$

argmax fonksiyonu, kümenin içinde olası değerler içinde azami değeri vermektedir. Buna göre ε_g^{jh} üzerinden 5 çeşit model ortaya konmaktadır:

1. durum: ε_g^{jh} hem kümelere hem değişkenlere hem de seviyelere göre değişir.

2. durum: ε_g^{jh} sadece kümelere ve değişkenlere göre değişir.

3. durum: ε_g^{jh} sadece kümelere göre değişir.

4. durum: ε_g^{jh} sadece değişkenlere göre değişir.

5. durum: ε_g^{jh} değişmez.

Burada ε_g^{jh} saçılımı sadece kümelerle ve değişkenlerle belirlendiğinde (2.79)'daki eşitlik ile ε_g^{jh} hesaplanır:

$$\varepsilon_g^{jh} = \begin{cases} \varepsilon_g^j & h = a_g^j \\ \frac{\varepsilon_g^j}{m_j - 1} & \text{diğer durumlarda} \end{cases} \tag{2.79}$$

Bu durumda karma modelin G-1+Gd parametresi varken, 1, durum için $G - 1 + G\sum_j (m_j - 1)$ parametresi vardır. Karma modelin formülü, Aitchison ve Aitken (1976) tarafından sunulmuştur. Olasılık fonksiyonu formülü (2.80)'deki gibi ifade edilir:

$$f(y;\theta) = \sum_{g=1}^{G} \tau_g \prod_{j=1}^{m_j} (1 - \varepsilon_g^j) \left(\frac{\varepsilon_g^j}{(m_j - 1)(1 - \varepsilon_g^j)} \right)^{1 - \delta\left(y_i^j, a_g^j\right)} \tag{2.80}$$

Burada $\delta(.,.)$, Kronecker çarpım fonksiyonu olarak geçmektedir. Bunun için verinin tamamı için log-olabilirlik fonksiyonu (2.81)'deki gibi ifade edilir:

$$\begin{aligned} L_C(z,\theta) = & \sum_{g=1}^{G} n_g \log(\tau_g) \\ & + \sum_{g=1}^{G} \sum_{j=1}^{d} \log\left(\left(\frac{\varepsilon_g^j}{(m_j - 1)(1 - \varepsilon_g^j)} \right) \left(n_g - \sum_{i=1}^{n} z_{ig} \delta\left(x_i^j, a_g^j\right) \right) \right) \\ & + \sum_{g=1}^{G} n_g \log(1 - \varepsilon_g^j) \end{aligned} \tag{2.81}$$

Burada n_g, g. bileşene dahil gözlem sayılarını temsil etmektedir. E adımında z_{ig} hesaplandıktan sonra M adımında, bütün durumlarda (2.82) ve (2.83)'teki eşitliklere göre güncelleme yapılır:

$$\left(a_g^j\right)^{(r+1)} = \text{argmax}_h\left(u_g^{jh}\right)^{(r)} \tag{2.82}$$

$$\left(\varepsilon_g^{jh}\right)^{(r+1)} = \frac{n_g^{(r)} - \left(v_g^j\right)^{(r)}}{n_g^{(r)}} \tag{2.83}$$

Burada h^*, a için mod seviyesi olmak üzere $v_g^j = u_g^{jh^*}$ sağlanır. 2. durum için parametrelerin tahmin edicileri, (2.82) ile (2.83)'deki gibi elde edilir. 3. durumda ε_g^{jh}, (2.84)'teki ifade edilir:

$$\left(\varepsilon_g^{jh}\right)^{(r+1)} = \frac{n_g^{(r)} d - v_g^{(r)}}{n_g^{(r)} d} \tag{2.84}$$

Burada $v_g = \sum_{j=1}^{d} v_g^j$ sağlanmaktadır. 4. durumda ε_g^{jh} (2.85)'daki ifade edilir:

$$\left(\varepsilon_g^{jh}\right)^{(r+1)} = \frac{nd - v^{(r)}}{nd} \tag{2.85}$$

Burada $v = \sum_{j=1}^{d} \sum_{g=1}^{G} v_g^j$ alınır. GSM, sınıflayıcı EM (CEM) algoritması ile hesaplamaları gerçekleştirilir. Bu durumda CEM algoritmasını E adımı, C adımı ve M adımı aşağıdaki şekilde gerçekleştirilir:

- E adımı: τ_{ig}'ler hesaplanır.
- C adımı: her gözlem azami olasılığa sahip olduğu kümeye konur.
- M adımı: (2.86) ve (2.87)'deki eşitlikler ile parametrelerin tahmin edicileri hesaplanır:

$$p_g^{(r+1)} = \frac{n_g^{(r+1)}}{n} \tag{2.86}$$

$$\left(\alpha_g^{jh}\right)^{(r+1)} = \frac{\left(u_g^{jh}\right)^{(r+1)}}{n_g^{(r+1)}} \tag{2.87}$$

CEM algoritması yanlı GSM parametreleri üretir ancak çok küçük sayıda iterasyon ile sonuca varabilmektedir (Bouveyron vd., 2019). GSM için BIC (Bayesian Information Criteria) formülü (2.89)'daki gibidir:

$$\mathrm{BIC} = \log\left(f\left(y;\hat{\theta}\right)\right) - \frac{v}{2}\log(n) \tag{2.88}$$

GSM için ICL (Integrated Complete Loglikelihood) kriteri ise (2.89) ile elde edilir:

$$\mathrm{ICL} = \log\left(P(x,\hat{z})\right) + O_p(1) \tag{2.89}$$

$O_p(1)$ azalmakta olan fonksiyonlar için kullanılır. Bu önemsenecek bir terim değildir. $P(x,\hat{z})$ için $P\left(x,\hat{z};\hat{\theta}\right)P\left(\hat{\theta}\right)$'nın $\hat{\theta}$ üzerinden integrali alınır ve formülde yerine konur. $\hat{z}$ için $\hat{z}_{ig}$ elemanları (2.90)'daki gibi hesaplanır:

$$\hat{z}_{ig} = \begin{cases} 1 & \mathrm{argmax}_l t_{il}\left(\hat{\theta}\right) = g \\ 0 & \text{diğer durumlarda} \end{cases} \tag{2.90}$$

Standart GSM modeli için belli Dirichlet önsel dağılımlar kullanılarak ICL (2.92)'deki şekilde ifade edilir:

$$\mathrm{ICL} = \log\big(\Gamma(bG)\big) - G\log\big(\Gamma(b)\big) + \left(\sum_{g=1}^{G} \log\Big(\Gamma\big(\hat{n}_g + b\big)\Big) - \log\big(\Gamma(bG + n)\big)\right)$$
$$+ g\sum_{j=1}^{d}\Big\{\log\Big(\Gamma\big(cm_j\big)\Big) - m_j\log\big(\Gamma(c)\big)\Big\}$$
$$+ \sum_{g=1}^{G}\sum_{j=1}^{d}\left\{\sum_{h=1}^{m_j} \log\Big(\Gamma\big(\hat{u}_g^{jh} + c\big)\Big) - \log\Big(\Gamma\big(\hat{n}_g + cm_j\big)\Big)\right\} \qquad (2.91)$$

Burada $\#\{.\}$ bir kümenin eleman sayısı olmak üzere, $\hat{n}_g = \#\big\{i: \hat{z}_{ig} = 1\big\}$, n örneklem hacmi ve $\hat{u}_g^{jh} = \#\Big\{i: \hat{z}_{ig} = 1, y_i^{jh} = 1\Big\}$ olmaktadır; c ve b değerleri ICL değerini hesaplamaya yardımcı olurken, diğer parametreler daha önceden belirtilen şekliyle tanımlanmıştır. Bu noktada model seçimi için AIC(Akaike Information Criterion) ve AIC3 sırasıyla (2.92) ve (2.93)'teki eşitliklerle hesaplanır:

$$\mathrm{AIC} = \log\Big(f\big(y;\hat{\theta}\big)\Big) - v \qquad (2.92)$$

$$\mathrm{AIC3} = \log\Big(f\big(y;\hat{\theta}\big)\Big) - \frac{3v}{2} \qquad (2.93)$$

Bu 2 kriterden AIC3'ün performansı AIC'a göre daha iyi olduğu durumlar bulunmaktadır (Bouveyron vd., 2019). Bu konuda değinilmesi gereken hususlardan birisi de algoritmaların başlangıç değerlerine bağımlılığıdır. Bu problem için Biernacki vd. (2003) aşağıdaki aşamaları önermiştir:

1) EM algoritması esnek bir durdurma kriteriyle, r kısa çalıştırma gerçekleştirsin.
2) Bu kısa çalıştırmalarda en yüksek olasılık değeri veren çözüm, ilk pozisyon olarak seçilir.
3) EM algoritması sıkı yakınsama kriteri ile bu başlangıç değerle çözüme ulaşır.

Bu aşamalara, Bouveyron vd. (2019) em-EM algoritması demektedir. Buradan GSM için Bayesçi analizin nasıl gerçekleştirileceğine geçilecektir.

GSM için Bayesçi analiz Dirichlet (a,...,a) önsel dağılımıyla mümkündür. Bu dağılımın olasılık yoğunluk fonksiyonu (2.94)'tür:

$$f(a_1, \dots, a_r; a, \dots, a) = \frac{\Gamma(ra)}{\left(\Gamma(a)\right)^r} \prod_{j=1}^{r} a_j^{a-1} \tag{2.94}$$

GSM modelde çoklu dağılımlar bir karma model oluşturduğu için her çoklu dağılıma iyi tahmin ediciler saptamak için Dirichlet fonksiyonu uygundur. Bunun haricinde yeniden parametrizasyon durumunda $\left(\alpha_g^{j1}, \dots, \alpha_g^{jm_j}\right)$ için önsel dağılım da Dirichlet$(c + u_g^{j1}, \dots, c + u_g^{jm_j})$ olarak seçilir. Bunun yanında α_g^j ile z'ye bağlı olarak, ε_g^{jh}'lerin koşullu dağılımları aşağıdaki gibidir:

1. durum: v_g^j'nin dağılımı yoktur.
2. durum: $v_g^j \sim B_i\left(n_g, 1 - \varepsilon_g^j\right)$
3. durum: $v_g \sim B_i\left(n_g d, 1 - \varepsilon_g\right)$
4. durum: $v^j \sim B_i\left(n_g, 1 - \varepsilon^j\right)$
5. durum: $v \sim B_i(nd, 1 - \varepsilon)$

Bunun yanında Bayesçi analiz için ε_g^{jh}'lerin önsel dağılımları kullanılabilir. $Bt_{[.]}$ [.,.] aralığında sınırlandırılmış beta dağılımı olmak üzere bunlar aşağıda açıklandığı gibi örneklenir:

2. durum: Önsel dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}\left((m_j - 1)(c - 1) + 1, c\right)$ iken, sonsal dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}\left(n_g - v_g^j + (m_j - 1)(c - 1) + 1, v_g^j + c\right)$ şeklindedir.

3. durum: Önsel dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}(d(m-1)(c-1)+1, d(c-1)+1)$ iken, sonsal dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}(n_g d - v_g + d(m-1)(c-1)+1, v_g + d(c-1)+1)$ şeklindedir.
4. durum: Önsel dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}(g(m_j-1)(c-1)+1, G(c-1)+1)$ iken, sonsal dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}(n - v^j + G(m_j-1)(c-1)+1, v^j + G(c-1)+1)$ şeklindedir.
5. durum: Önsel dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}(Gd(m-1)(c-1)+1, d(c-1)+1)$ iken, sonsal dağılım $\varepsilon_g^j \sim Bt_{[0,(m_j-1)/(m_j)]}(nd - v + Gd(m-1)(c-1)+1, v + d(c-1)+1)$ şeklindedir.

Bayesçi analiz, en çok olabilirlik yöntemine bir alternatif olmaktadır ve başka sorunları çözmede faydası vardır. Örneğin en çok olabilirlik yaklaşımını düzenlemekte kullanılabilmektedir. Bu düzenleme EM algoritmasını değiştirir. Bütün diğer parametreler için güncellemeler aynı kalırken, $p_g^{(r+1)}$ ile $\left(\alpha_g^{jh}\right)^{(r+1)}$ sırayla (2.95) ve (2.96)'daki gibi ifade edilir:

$$p_g^{(r+1)} = \frac{n_g^{(r)} + b - 1}{n + G(b-1)} \tag{2.95}$$

$$\left(\alpha_g^{jh}\right)^{(r+1)} = \frac{\left(u_g^{jh}\right)^{(r)} + c - 1}{n_g^{(r)} + m_j(c-1)} \tag{2.96}$$

ε_g^j'ler için geçerli olan durumlarda güncellemeler aşağıdaki gibi yapılır:

2.durum: (2.97)'deki gibi güncellenir:

$$\varepsilon_g^j = \frac{n_g - v_g^j + (m_j-1)(c-1)}{n_g + m_j(c-1)} \tag{2.97}$$

3.durum: (2.98)'deki gibi güncellenir:

$$\varepsilon^{j} = \frac{n_g^d - v_g + d(m-1)(c-1)}{n_g d + dm(c-1)} \tag{2.98}$$

4.durum: (2.99)'daki gibi güncellenir:

$$\varepsilon^{j} = \frac{n - v^{j} + G(m_j - 1)(c-1)}{n + Gm_j(c-1)} \tag{2.99}$$

5. durum: (2.100)'deki gibi güncellenir:

$$\varepsilon = \frac{nd - v + Gd(m-1)(c-1)}{nd + Gdm(c-1)} \tag{2.100}$$

n ve n_g sırayla verinin büyüklüğü ve küme büyüklüğü olarak ifade edilirken, d değeri veri boyutunu temsil eder.

2.4.2 Model tabanlı kümeleme gerçekleştirilirken gözetilmesi gereken kriterler

Bayesçi bilgi kriteri'ni (BIC) ifade etmek için öncelikle $P(D|M_k)$ ifadesini formülleştirmek gerekir. M_k, modeli ve D, veriyi temsil etsin. Bu durumda $P(D|M_k)$ (2.101)'deki gibi ifade edilir (Bouveyron vd., 2019):

$$P(D|M_k) = \int P(D|\theta_{M_k}, M_k) P(\theta_{M_k}|M_k) d\theta_{M_k} \tag{2.101}$$

Burada θ_{M_k}, M_k modeli için parametre vektörünü; $P(.|.)$ koşullu olasılığı ifade eder. $P(M_k)$'lar hakkında önsel bilgi yoksa onlar eşit olarak alınabilir. Bu durumda iki farklı model M_i ve M_j için Bayesçi faktör (2.102)'deki gibi ifade edilir:

$$B_{ij} = P(D|M_i)/P(D|M_j) \tag{2.102}$$

Burada $P(D|M_i)$, M_i karma modeli verili iken D verisi için uygunluk kriterini temsil eder. Bu faktör, birden fazla modele uygulanabilir olduğundan çok faydalıdır. Ayrıca elimizde hiyerarşik bir biçimde inşa edilen modeller varsa bu yöntem genel hatayı azaltmayı da sağlar (Jeffreys, 1961). Bouveyron vd. (2019) bu kriter için en önemli problemin integral hesaplanması olduğunu belirtmiştir. Sıradan modeller için BIC kriteri (2.103)'teki eşitlik ile ifade edilir:

$$2\log(P(D|M_k)) \cong 2\log\left(P(D|\hat{\theta}_{M_k}, M_k)\right) - v_{M_k}\log(n) = BIC_{M_k} \tag{2.103}$$

Burada n örneklem hacmini; M_k uygun olabilecek modellerden birini; v_{M_k} M_k için parametre sayısını ve M_k için $\hat{\theta}_{M_k}$ parametre vektörünü temsil etmektedir. BIC kriteri uygulandığı alanda elde edilen dış bilgi faydalı olabilmektedir ama bunun okuyuculara anlatılması gerekebilir (Bouveyron vd., 2019). Örneklem hacmi arttıkça BIC kriteri, kümelerin sayısını gereksiz yere fazla gösterebilmektedir. Normal dağılımlardan oluşan karma model için olabilirlik fonksiyonu sınırlı değildir ancak kovaryans matrislerinin en düşük özdeğeri üzerine bir alt sınır koymak bu işi çözebilmektedir (Bouveyron vd., 2019). Model tabanlı kümelemede sıkça kullanılan programlarda bu iş çözülmüştür. Roeder ve Wasserman (1997), tek değişkenli normal dağılımların parametrik olmayan yoğunluk fonksiyonunu tahmin etmek için BIC kriteri kullanılacak olursa iyi bir tahmin elde edilebileceğini göstermişlerdir. Buradaki fark bileşenler aslında normal dağılıma sahip olmayan kümeleri de temsil edebilmesinden kaynaklanmaktadır. Şayet kümeleme veriye normal karma model uydurmaktan daha önemli ise entegre tam olabilirlik (ICL) kullanmak daha iyi sonuç verir. Bu kriteri elde etmek için (2.104)'teki integralin hesaplanması gerekir:

$$\int P(y, z|\theta_{M_k}, M_k)P(\theta_{M_k}|M_k)d\theta_{M_k} \tag{2.104}$$

BIC'in hesaplanmasına benzer bir şekilde ICL, (2.105)'deki gibi elde edilir:

$$2\log(P(y, z|M_k)) \cong 2\log\left(P(y, Z^*|\hat{\theta}_{M_k}, M_k)\right) - v_{M_k}\log(n) = ICL_{M_k} \tag{2.105}$$

Burada y, kümelenmek için kullanılan veriyi; Z^*, $\hat{Z}_{ig}'$in en yüksek değerini aldığı i ve g değerleri için 1 değerini alır. Tam tersi için ise 0 değerli küme üyeliklerini $\left(\hat{Z}_{ig} = \frac{\hat{\tau}_g f_g(y_i|\theta_h)}{\sum_{h=1}^{G_{M_k}} \hat{\tau}_h f_h(y_i|\theta_h)}\right)$ alır. Bunun yanında v_{M_k}, tahmin edilecek parametre sayısını; log(.), doğal logaritmayı ve n, örneklem hacmini göstermektedir. ICL_{M_k}, aynı zamanda Bayesçi bilgi kriteri BIC_{M_k}'ye göre (2.106)'daki gibi ifade edilir (Bouveyron vd., 2019):

$$ICL_{M_k} = BIC_{M_k} - E(M_k) \tag{2.106}$$

Burada $E(M_k)$, M_k modelinin beklenen entropisini gösterir. Bu, (2.107)'deki gibi ifade edilir:

$$E(M_k) = -\sum_{i=1}^{n}\sum_{g=1}^{G_{M_k}} \hat{Z}_{ig}\log(\hat{Z}_{ig}) \tag{2.107}$$

Böylelikle ICL, BIC'ın sınırlandırılmış durumu olarak geçmektedir. ICL, uygun sayıda küme sayısı seçmeye yardımcı olur.

Gürültülü (hatalı) verilerin kümelenmesi için yapılan çalışmalardan birisi Garcia-Escudero vd. (2008) tarafından gerçekleştirilmiş ve R'da tclust paketi haline dönüştürülmüştür. Bu çalışmalarda olabilirlik fonksiyonu (2.108)'deki gibi tanımlanmıştır:

$$L(Y_i, \mu_g, \Sigma_g, \tau_g) = \sum_{y_i \notin S} \sum_{g=1}^{G} \log\left(\tau_g \phi(y_i; \mu_g, \Sigma_g)\right) \tag{2.108}$$

Bu fonksiyon, EM algoritması ile aşağıdaki şekilde hesaplanır:

E adımı: $D_k(y_i; \Theta) = \tau_k \phi(y_i; \mu_k, \Sigma_k)$ olmak üzere d_i değerleri (2.109)'daki ifadeye göre hesaplanır:

$$d_i = \max\left(D_1(y_i; \Theta), \dots, D_g(y_i; \Theta)\right) \tag{2.109}$$

Bu değerlerden en küçük $n.\alpha$ tanesi S kümesine alınır.

M adımı: S kümesine alınmayanlar önceden oluşturulmuş kümelere aynı formül üzerinden atanarak güncelleme yapılır.

Garcia-Escudero vd. (2008) klasik model seçim kriterlerinin bunun için yeterli olmadığını belirtmiş ve onun yerine log-olabilirlik eğrilerini α, G+1 ve G değerlerine göre çizmeyi önermiştir.

2.4.3 Model tabanlı kümelemede ortaya çıkan olabilirlik fonksiyonundan kaynaklanan eksiklikler ve onların çözümü

Normal karma modeller için olabilirlik fonksiyonu maksimize edilirken herhangi bir sınırlama olmazsa çözüm bulmak imkansız olmaktadır (Bouveyron vd., 2019). Algoritma sahte çözümler üretir ve tahmin edilen parametrelerle alakası olmayan sonuçlar verir. Redner ve Walker (1984), olabilirlik fonksiyonunun belli bölgelerinde yerel maksimumlara ulaşılabileceğini ve örneklem hacmi büyüdükçe bunların gerçek değerlere daha yakın çözümler bulacaklarını ifade etmiştir. Ancak herhangi bir model için en iyi çözümlerin bulunması bu problemi çözmemektedir çünkü karma modellerin seçimi bile sahte çözüm oluşturabilir ve iş zorlaşmaktadır. Bunun nedeni model seçiminde en iyi model yerine veriye uygun daha çok bileşenli modellerin tercih edilmesidir (Bouveyron vd., 2019). Örneğin veri 2 küme ile açıklanabilecekken karma modele 3 bileşen katılırsa, model seçim kriterinin 2 küme yerine 3 küme ile oluşturan bir model seçme şansı vardır. Daha önceden ICL'in daha BIC'a göre daha küçük kümeler seçebileceğini belirtilmişti ancak ICL, küme olasılık değerleri 0 ya da 1 olduğunda ya da bu değerlere çok yakın olduğunda BIC ile aynı sayıda küme seçmektedir. Bu da yanlış sonuçlara yönlendirmeye yol açar. Bu yüzden sahte çözümlerin varlığını aşmak için çalışmalar yapılmıştır. Bouveyron vd. (2019) aşağıdaki stratejiyi önermiştir:

i. İlk önce her model bileşeninin kovaryans matrisi için özdeğerleri hesaplayıp en büyük özdeğeri en küçüğe bölerek bir oran elde edilir.
ii. Belli bir eşik değeri altındaki oranlar için çözüm aranmaz.

Bu eşik değeri göreceli makine hassasiyeti olup IEEE uyumlu bilgisayarlarda eşik değeri olarak $2x10^{-16}$ seçilmiştir. Bu yöntemle bazı iyi modellerin atılması düşük bir ihtimalle mümkün olup böyle birşey olsa bile veri tarafından destek görmeyen bileşenler üretir. Bouveyron vd. (2019) d boyutlu bir veride bir bileşen d noktadan daha az sayıda gözlem içeriyorsa, bunu çözmek için geometrik kısıt kullanılabileceğini ama bunun veri tarafından desteklenmeyebileceği belirtmektedir. Fraley ve Raftery (2007), çalışmalarında standart normal-ters wishart önsel dağılımını kullanarak sonsal dağılımı elde etmiş ve EM algoritmasını kullanarak bu sefer sonsal dağılımın mod'unu tespit etmiştir. Bu Bayesçi yaklaşım sayesinde:

i. $\mu|\Sigma$, çoklu normal dağılıma sahip olur.
ii. Σ dağılımı, invers wishart dağılıma sahip olur.
iii. $Y|\mu,\Sigma$ sonsal dağılımı, sahte çözümleri büyük oranda eler.

Bu yaklaşım kümeleme analizinde sahte çözüm elde etme ihtimalini yok etmede başka bir yol sunmaktadır. Şayet kümeler normal dağılıma sahip olmuyorsa bazı kümeler birleştirerek uygun bir çözüm bulunabilir (Bouveyron vd., 2019). Böyle durumda BIC, ICL kriterine göre daha iyi bir seçenek sunmaktadır çünkü ICL normal olmayan dağılımları normal dağılımla ifade edebilir ve BIC normal olmayan dağılımlara fazladan bir bileşen daha ekleyip bu konuda ICL'den daha avantajlı olmaktadır. Normal karma model bileşenlerini birleştirmek için Hennig (2010) karma model bileşenlerinden birleşmeye en uygun kümeleri seçip bir durdurma kriteri sağlanıncaya kadar kümeleri birleştirmeyi önermiştir. Bunun için (2.110)'daki kriter ortaya atılmıştır (Baudry vd., 2010):

$$\mathrm{Ent}(G) = -\sum_{g=1}^{G}\sum_{i=1}^{n} \hat{z}_{i,g}^{[g]} \log\left(\hat{z}_{i,g}^{[g]}\right) \tag{2.110}$$

Burada $\hat{z}_{i,g}^{[g]}$ kümelerin birleştirilmiş olasılıklarını temsil etmektedir. Model tabanlı sınıflandırma'da, sınıflar belli olduğundan kümelemenin kendisi denetimli bir şekilde yürütülür. Karma modeller bu noktada başka kriterlerle denetlenebilecek şekle dönüşüyor ve yeni yöntemler bu şekilde çıkıyor. Örneğin özdeğer ayrıştırmalı diskriminant analizi(EDDA) bunlardan biridir. Bu yöntemde, kovaryans matrisi dxd biçimindeki Σ_g (2.111)'deki eşitlikle ifade edilir:

$$\Sigma_g = \left|\Sigma_g\right|^{\frac{1}{d}} D_g A_g D_g' \tag{2.111}$$

Bu formülde D_g özvektör matrisini, A_g normalleşmiş özdeğerler matrisini (Bu matris, köşegen elemanları azalan sırada olan köşegensel bir matristir) temsil eder. $|\Sigma_g|^{\frac{1}{d}}$ ile g. kümenin hacmi, A_g ile g. kümenin şekli belirlenirken, D_g ile yönelim belirlenir (Bouveyron vd., 2019). EDDA modeli en küçük çapraz onaylama hata oranını oluşturan bir sınıflayıcıdır (Bouveyron vd., 2019). EDDA'nın iki özel durumu bulunmaktadır. Bunlar doğrusal diskriminant analizi (LDA) ve kuadratik diskriminant analizidir (QDA). Bunların etkilerini birleştirerek ortaya çıkan diskriminant analizine düzenlenmiş diskriminant analizi (RDA) denilmektedir.

RDA, veri küçük olduğunda birtakım tahmin edicilerle düşürülmüş varyanslarla sınıflayıcı tahminleri oluşturmaktadır. Bunu yaparken (2.112), (2.113), (2.114) ve (2.115)'teki formüllerden faydalanır:

$$\Sigma_g(\lambda, \Upsilon) = (1-\Upsilon)\hat{\Sigma}_g + \Upsilon\left(\frac{\mathrm{tr}(\hat{\Sigma}_g)}{d}\right) I_d \tag{2.112}$$

$$\hat{\Sigma} = \frac{\sum_{g=1}^{G}\sum_{i=1}^{n} z_{ig}(y_i - \mu_g)(y_i - \mu_g)'}{n} \tag{2.113}$$

$$\mu_g = \frac{\sum_{i=1}^{n} z_{ig} y_i}{n_g} \tag{2.114}$$

$$\hat{\Sigma}_g = \frac{\sum_{i=1}^{n} z_{ig}(y_i - \mu_g)(y_i - \mu_g)'}{n_g} \tag{2.115}$$

Burada G, bileşen sayısını; n toplam örneklem hacmini ve n_g g. bileşen için örneklem hacmini gösterirken, μ_g g. bileşen için ortalama vektörünün; $\hat{\Sigma}_g$ g. bileşen için kovaryans matrisinin tahminini temsil eder. Bunun yanında $\Sigma_g(\lambda, \Upsilon)$ düzeltilmiş tahmini göstermektedir. Bu formüllerde λ parametresi, QDA ve LDA katkısını, ve Υ parametresi özdeğerlerin birbirlerine yakınlığını denetlemektedir. Şayet kesikli bir veride bu uygulanacak olursa bu durumda boyut problemi ortaya çıkabilmektedir ve yeni formüller gerekmektedir. Bu formülleri oluşturmak için α_g^{jh} , y_i^j ve y_i^{jh}'lerin tanımlanması gerekir. y_i^{jh} i. gözlemin j. değişkeninin h. seviyesini tanımlamaktadır. Aynı şekilde y_i^j, i. gözlemin j. değişkenininin seviyesini tanımlamaktadır. y_i^j şayet h değerini alıyorsa, bu durumda y_i^{jh} gölge değişkeni 1 değerini aksi

taktirde 0 değerini alır. α_g^{jh} ise g. küme için y_i^{jh} 'nin ortaya çıkma olasılığı olarak tanımlanır ve sonuç olarak kesikli verideki bileşenler için olasılık fonksiyonu (2.116)'daki gibi tanımlanır:

$$f_g(y_i|\alpha_g) = \prod_{j,h}\left(\alpha_g^{jh}\right)^{y_i^{jh}} \tag{2.116}$$

$n_g = \sum_{i=1}^{n} z_{ig}$ ve $u_g^{jh} = \sum_{i=1}^{n} z_{ig} y_i^{jh}$ olmak üzere her bir bileşen için olasılık fonksiyonları (2.117)'deki gibi ifade edilir:

$$\hat{\alpha}_g^{jh} = \frac{u_g^{jh}}{n_g} \tag{2.117}$$

ancak algoritmaya konulduğunda n_g'yi 0 elde etme riski olmaktadır. Bundan dolayı c sabit alınıp ve en yüksek seviyeyi m_j varsayıp bu risk (2.118)'deki eşitlik kullanılarak çözülebilir:

$$\hat{\alpha}_g^{jh} = \frac{u_g^{jh} + c - 1}{n_g^{(r)} + m_j(c-1)} \tag{2.118}$$

c'nin seçimi bu algoritmanın gidişatını etkilediğinden bu çözüm de çok ideal durmamaktadır. $\hat{\alpha}_g^{jh}$ parametresi iki farklı parametreyi tek başına tanımlayabilmektedir ve bunlardan $\hat{\alpha}_g^{j}$ hangi seviyenin en sık rastlandığını (ya da seviyelerden hangisinin mod olduğunu), $\hat{\varepsilon}_g^{j}$ modal seviyeden sapılma sıklıklarını (bu parametre dolaylı olarak seviyelerin saçılma büyüklüğünü de gösterir) (2.119) ve (2.120) ile ifade eder:

$$\hat{\alpha}_g^{j} = \operatorname{argmax}\left(u_g^{jh}\right) \tag{2.119}$$

$$\hat{\varepsilon}_g^j = \frac{n_g - u_g^{jh^*}}{n_g} \tag{2.120}$$

Bu tahmin ediciler kullanılarak çapraz onaylama işlemi aşağıdaki şekilde gerçekleştirilir:

- Deneme kümesi v parçaya ayrılır ve bunlardan her birine A_i denir. $1 \leq j \leq v$ olacak şekilde herhangi bir $\bigcup_{l=1}^{j} A_l - A_j$ kümesinde M^j modeli seçilir.
- A_j'yi M^j modeli ile sınıflanır ve yanlış sınıflandırma oranını E_j olarak ifade edilir. Böylece yanlış sınıflandırma oranları ortalaması hesaplanır ve çapraz onaylamadan doğan yanlış sınıflandırma tahmini (2.121)'deki eşitlik ile elde edilir:

$$E_{CV} = \frac{1}{V}\sum_{j=1}^{v} E_j \tag{2.121}$$

Şayet bu yöntem modeller arasında uygulansaydı bu durumda modeller, bu yönteme tabi tutularak aralarında en küçük yanlış sınıflandırma oranlı model seçilerek yapılırdı.

2.4.4 Yarı denetimli model tabanlı kümeleme

Bu sınıflandırma türünde ilk n gözlem etiketlenmiş iken sonraki m gözlem etiketlenmemiş durumdadır. Bu durumda tam veri olabilirlik fonksiyonu (2.122)'deki gibi yazılır:

$$L(\theta|y,z) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\left(\tau_g P\left(y_i|\theta_g\right)\right) + \sum_{i=n+1}^{n+m}\sum_{g=1}^{G} z_{ig}\log\left(\tau_g P\left(y_i|\theta_g\right)\right) \tag{2.122}$$

Burada τ_g g. bileşen için karma model ağırlığını; θ_g y için olasılık yoğunluk fonksiyonunun parametrelerini; $P(.|.)$ olasılık yoğunluk fonksiyonunu temsil etmektedir. z_{ig} burada i. gözlem g. sınıfa giriyorsa 1 değerini alan aksi taktirde 0 değerini alan bir parametredir (Bouveyron vd., 2019). Burada m gözlem henüz etiketlenmediği için bu formülün iki toplam şeklinde ifade edilmesi daha uygundur. Buradan EM algoritmasında kullanılacak beklenen tam log-olabilirlik fonksiyonu (2.123)'teki gibi ifade edilir.

$$Q(\theta|\theta^*) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\log\left(\tau_g P(y_i|\theta_g)\right) + \sum_{i=n+1}^{n+m}\sum_{g=1}^{G} \tau_g(y_i|\eta)\log\left(\tau_g P(y_i|\theta_g)\right) \quad (2.123)$$

Burada (2.122)'den sonraki tanımlamamalar geçerlidir. $\tau_g(y_i|\eta)$ ifadesi etiketlenmemiş gözlemler için etiketlere dahil edilme olasılıklarını ifade etmektedir. Bu durumda EM algoritması (2.124), (2.125), (2.126), (2.127), (2.128), (2.129), (2.130) ve (2.131)'deki formüllerle işler:

E adımı:

$$\tau_g\left(y_i|\eta^{(s)}\right) = \frac{\hat{\tau}_g^{(s-1)} P\left(Y_i|\hat{\eta}_g^{(s-1)}\right)}{P(Y_i|\hat{\eta}^{(s-1)})} \quad (2.124)$$

M adımı:

$$\hat{\tau}_g^{(s)} = \frac{n_g + m_g^{(s)}}{n+m} \quad (2.125)$$

$$\hat{\mu}_g^{(s)} = \frac{1}{n}\left(\sum_{i=1}^{n} z_{ig} y_i + \sum_{i=1}^{n} \tau_g\left(y_i|\theta^{(s)}\right) y_i\right) \quad (2.126)$$

$$\hat{\Sigma}_g = \frac{1}{n}\left(S_g + S_g^{(s)}\right) \quad (2.127)$$

$$S_g = \sum_{i=1}^{n} z_{ig}\left(y_i - \hat{\mu}_g^{(s)}\right)'\left(y_i - \hat{\mu}_g^{(s)}\right) \tag{2.128}$$

$$S_g^{(s)} = \sum_{i=1}^{n} \hat{\tau}_g^{(s)}\left(y_i - \hat{\mu}_g^{(s)}\right)'\left(y_i - \hat{\mu}_g^{(s)}\right) \tag{2.129}$$

$$n_g = \sum_{i=1}^{n} z_{ig} \tag{2.130}$$

$$m_g^{(s)} = \sum_{i=n+1}^{n+m} \tau_g\left(y_i|\theta^{(s)}\right) \tag{2.131}$$

Burada $S_g^{(s)}$ ile S_g hariç bütün parametreler daha önceden belirtildiği şekillerde tanımlanırken, $S_g^{(s)}$ ile S_g karma modeldeki g. bileşen için kovaryans matrisi ile örneklem için tanımlanan g. bileşen için kovaryans matrisi gösterir. Model seçimi için kriter olarak AIC, BIC ve BEC kriterleri kullanılabilir. BEC kriteri, (2.132)'deki eşitlik ile ifade edilir:

$$\mathrm{BEC}(m) = \log\left(P\left(y, z|\hat{\theta}_{y,z}^{m}\right)\right) - \log\left(P\left(y|\hat{\theta}_{y}^{m}\right)\right) \tag{2.132}$$

Burada $\hat{\theta}_{y,z}^{m}$ ve $\hat{\theta}_{y}^{m}$ sırayla y ve z'den elde edilen ve y'den elde edilen $\hat{\theta}^{m}$ parametre tahmini demektir. Bu kriterin kümeleme yazılımı için hesaplama maliyeti, AIC ve BIC'ın iki katı ama kümeleme yazılımı için çapraz onaylama maliyeti daha düşüktür (Bouveyron vd., 2019). Bunun yanında BEC asimptotik olarak tek bir modele yakınlaştırır (Bouchard ve Celeux, 2006). Pratikte BEC'in AIC ve BIC'tan daha iyi iş gördüğü görülmüştür. Buna rağmen (2.133)'teki BEC'den daha az kompleks bir model seçimi için $\mathrm{AIC}_{\mathrm{cond}}$ geliştirilmiştir:

$$\mathrm{AIC}_{\mathrm{cond}}(m) = 2\mathrm{logp}\left(z|y, \hat{\theta}_{y,z}^{m},\right) - 4\log\left(\frac{P\left(y|\hat{\theta}_{y}^{m}\right)}{P\left(y|\hat{\theta}_{y,z}^{m}\right)}\right) \tag{2.133}$$

(2.133)'teki eşitlik, (2.134)'teki eşitliğe dönüştürülür:

$$\mathrm{AIC_{cond}}(\mathrm{m}) = 2\mathrm{BEC}(\mathrm{m}) - 4\log\left(\frac{\mathrm{P}\left(\mathrm{y}|\hat{\theta}_{\mathrm{y}}^{\mathrm{m}}\right)}{\mathrm{P}\left(\mathrm{y}|\hat{\theta}_{\mathrm{y,z}}^{\mathrm{m}}\right)}\right) \tag{2.134}$$

Bir modelin tahmin ediciliğinden sapması nasıl en aza indirileceğinin araştırılmasından doğmuştur (Bouveyron vd., 2019). Bu kriter, z_i ile $z_i|y_i$ arasındaki beklenen Kullback-Leibler uzaklığını en aza indirir. $\mathrm{AIC_{cond}}$, bazı verilerde BEC'ten daha iyi sonuçlar elde ettiği ve bazı verilerde onunla aynı sonuçlara ulaştığı belirtilmiştir (Bouveyron vd., 2019).

Bazı durumlarda gözlem çiftleri aynı anda bir kümeye girmesi gerekmektedir. Bunun için ayrı kümeleme analizi stratejileri geliştirilmek zorunda kalınır. Buna bir örnek biyoenformatik alanından çıkmıştır. Genler kümelenirken proteinler arası ilişkiler göz önüne alınarak kümeleme yapılır. Bu yarı denetimli kümeleme olarak karşımıza çıkmaktadır. Örneğin, proteinler arası ilişkiyi bilindiğinde, hangi gözlemlerin aynı kümeye düşüp düşmeyeceği önceden bilinebilir. Böylece tam olabilirlik fonksiyonu (2.135)'teki eşitliğe dönüşür:

$$\mathrm{Q}(\theta|\theta^*) = \sum_{\mathrm{i}=1}^{\mathrm{n}}\sum_{\mathrm{g}=1}^{\mathrm{G}} \mathrm{P}(\mathrm{Z}|\mathrm{Y},\theta^*,\mathrm{c})\log\big(\mathrm{P}(\mathrm{Y},\mathrm{Z}|\theta^*,\mathrm{c})\big) \tag{2.135}$$

Burada θ^* $\mathrm{P}(\mathrm{Z}|\mathrm{Y},\theta^*,\mathrm{c})$ için parametreleri ve $\mathrm{P}(.|.)$ koşullu olasılığı temsil ediyor. Burada c, c_l kümelerini içeriyor ve c_l $z_i = z_j$ olacak şekilde, y_i ile y_j noktalarını barındırır (Bouveyron vd., 2019). EM algoritmasında kümelere atama yapılırken c_l'deki elemanlar aynı kümeye konurlar ve ilişkisi belirtilmemiş gözlemler ayrı birer küme olarak ele alınır. Burada bunu kolayca ifade edebilmek adına bunlara alt küme olarak geçecektir. Bu durumda EM algoritması aşağıdaki şekilde işler:

- E adımı: $\xi = \sum_{\mathrm{j}=1}^{\mathrm{G}} \tau_{\mathrm{j}}^{(\mathrm{s}-1)} \prod_{\mathrm{y_i}\in\mathrm{c_l}} \mathrm{P}\left(\mathrm{y_i}|\mathrm{Z_l}=\mathrm{j},\theta_{\mathrm{j}}^{(\mathrm{s}-1)}\right)$ ve τ_{jg} model bileşen katsayısı olmak üzere (2.136)'daki eşitlik hesaplanır:

$$\tau_{\mathrm{lg}} = \frac{1}{\xi}\tau_{\mathrm{g}}^{(\mathrm{s}-1)} \prod_{\mathrm{y_i}\in\mathrm{c_l}} \mathrm{P}\left(\mathrm{y_i}|\mathrm{Z_l}=\mathrm{g},\theta_{\mathrm{g}}^{(\mathrm{s}-1)}\right) \tag{2.136}$$

Bu formülde $P\left(Y_i|\hat{\eta}^{(s-1)}\right)$, (s-1). iterasyonda karma modeldeki olasılığı ifade eder.

- M adımı: n_g c_g kümesinin büyüklüğü, $\bar{y}_{jg}$ c_g kümesinin j. alt kümesinin ortalaması ve Σ_{jg} c_g kümesinin j. alt kümesinin kovaryans matrisi olmak üzere model parametreleri (2.137), (2.138), (2.139)'taki eşitliklerle güncellenir:

$$\tau_g^{(s)} = \frac{1}{L}\sum_{j=1}^{L}\tau_{jg} \tag{2.137}$$

$$\hat{\mu}_g^{(s)} = \frac{1}{n_g}\sum_{j=1}^{L}\tau_{jg}\bar{y}_{jg} \tag{2.138}$$

$$\hat{\Sigma}_g^{(s)} = \frac{1}{ng}\sum_{j=1}^{L}\Sigma_{jg} \tag{2.139}$$

İstatistiksel analizde etiketlemede oluşan hatalar, kümelemeyi zora sokabilmektedir. Doğrusal diskriminant analizi (McLachlan, 1992) ve karma diskriminant analizi (Hastie ve Tibshirani, 1996) bu gürültüden etkilenir ve buna çözüm olarak:

- Veriyi temizlemek
- Model parametrelerini sağlamcı tahmin edicilerler hesaplamak
- Etiket hatasını modellemek

stratejileri önerilmiştir. Bunun haricinde EM algoritması ile aşağıda açıklanan bir strateji geliştirilmiştir:

- E-adımı: $P\left(y_i|\tilde{Z}_i = g, \hat{\theta}\right)P\left(Z = g|\tilde{Z}_i\right)$hesaplanır.

- M-adımı: n_g c_g kümesinin büyüklüğü, $\bar{y}_{jg}$ c_g kümesinin j. alt kümesinin ortalaması ve Σ_{jg} c_g kümesinin j. alt kümesinin kovaryans matrisi olmak üzere model parametreleri (2.140), (2.141), (2.142) ve (2.143)'deki formüllerle güncellenir:

$$v_g = \frac{1}{L}\sum_{j=1}^{L} \tau_{jg} \tag{2.140}$$

$$\hat{\mu}_g = \frac{1}{v_g}\sum_{j=1}^{L} \tau_{jg} y_j \tag{2.141}$$

$$\hat{\Sigma}_g = \frac{1}{v_g}\sum_{j=1}^{L} \tau_{jg}\left(y_j - \hat{\mu}_g\right)\left(y_j - \hat{\mu}_g\right)' \tag{2.142}$$

$$\hat{\gamma}_g = \frac{1}{v_g}\sum_{j=1}^{L} \tau_{jg} 1_{\{\widetilde{Z}_l = g\}} \tag{2.143}$$

Burada $1_{\{.\}}$, parantez içersindeki koşul sağlandığında 1 değeri veren aksi taktirde 0 değerini veren bir fonksiyondur. Şayet ikiden fazla sınıf olsaydı $\xi = \sum_{c=1}^{C}\sum_{y\in S_c} \log(P(Y))$ olmak üzere, log-olabilirlik fonksiyonu (2.144)'teki gibi ifade edilirdi:

$$L(R) = \sum_{c=1}^{C}\sum_{y\in S_c} \log\left(\sum_{g=1}^{G} r_{cg} P(Z = g|Y = y)\right) + \xi \tag{2.144}$$

R_c, R'nin c. satırı ve $\Psi(Y) = \left(P(Z = 1|Y = y), \ldots, P(Z = G|Y = y)\right)'$ olmak üzere bu formül matris formunda (2.145)'teki gibi ifade edilir:

$$L(R) = \sum_{c=1}^{C}\sum_{y\in S_c} \log(R_c \Psi(Y)) + \xi \tag{2.145}$$

Bu durumda bir optimizasyon problemi çıkmaktadır. Amaç $\sum_{c=1}^{C} r_{cg} = 1$ kısıtı altında, $\sum_{c=1}^{C} \sum_{y \in S_c} \log(R_c \Psi(Y))$ fonksiyonunu r_{cg}'lere göre maksimize etmektir (Burada ξ r_{cg}'nin fonksiyonu olmadığından ötürü dışlanabilir durumdadır). Bu problemin kapalı çözümü yoktur. İteratif olarak çözüm hesaplanması gerekmektedir. Bu yönteme sağlamcı karma diskriminant analizi (RMDA) denir. Bu, karma diskriminant analizinin genelleştirilmiş halidir (Bouveyron vd., 2019). Hastie ve Tibshirani (1996), kümeleme analizinin zorlandığı alanlardan birinin yenilik seçimi olduğunu belirtir. Yenilik seçimi, yeni ya da bilinmeyen bir yapının ortaya çıkışının belirlenme yoludur (Bouveyron vd., 2019). Kümeleme yapılırken deneme fazında, genelde bütün kümelerin ortaya çıktığı varsayılır ancak bazı örneklerde bu geçerli olamamaktadır (Bouveyron vd., 2019). Miller ve Browning (2003) gözlenmemiş sınıf tespiti konusunda ilk çalışmayı gerçekleştirmiştir (Bouveyron vd., 2019). Bu çalışmada G bileşenli bir karma modelin C tane bilinen sınıf ortaya koyduğu varsayılır. Burada θ_g g. bileşen için parametre vektörü; $\sum_{g=1}^{G} P(C|\theta_g) = \sum_{g=1}^{G} \beta_{cg} = 1$ ve $\sum_{g=1}^{G} \tau_g = 1$ olacak şekilde log-olabilirlik fonksiyonu (2.146)'daki gibi tanımlanır:

$$L(y_1, \dots, y_n; \Theta) = \sum_{i=1}^{n} \log\left(\sum_{g=1}^{G} \tau_g \beta_{c_i g} f(y_i; \theta_g) \right) \tag{2.146}$$

Bu durumda EM algoritması aşağıdaki şekilde işler:

E adımı: $c \in \{1, \dots, C\}$ ve $g \in \{1, \dots, G\}$ kabul edilir ve formülle (2.147)'deki gibi hesaplanır:

$$t_{ig} = \begin{cases} \dfrac{\hat{\tau}_g^{(s-1)} \hat{\beta}_{c_i g}^{(s-1)} f\left(y_i; \hat{\theta}_g^{(s-1)}\right)}{\sum_{j=1}^{c} \hat{\tau}_j^{(s-1)} \hat{\beta}_{c_i g}^{(s-1)} f\left(y_i; \hat{\theta}_g^{(s-1)}\right)} & g \in \{1, \dots, C\} \text{ ise} \\ 0 & \text{diğer durumlarda} \end{cases} \tag{2.147}$$

etiketlenmemiş gözlemler için t_{ig} (2.148)'deki şekilde hesaplanır:

$$t_{ig} = \begin{cases} \dfrac{\hat{\tau}_g^{(s-1)} \hat{\beta}_{ug}^{(s-1)} f\left(y_i; \hat{\theta}_g^{(s-1)}\right)}{\sum_{j=1}^{C} \hat{\tau}_j^{(s-1)} \hat{\beta}_{uj}^{(s-1)} f\left(y_i; \hat{\theta}_j^{(s-1)}\right) + \sum_{j=C+1}^{G} \hat{\tau}_j^{(s-1)} f\left(y_i; \hat{\theta}_j^{(s-1)}\right)} & g \in \{1, \dots, C\} \text{ ise} \\ \dfrac{\hat{\tau}_g^{(s-1)} f\left(y_i; \hat{\theta}_g^{(s-1)}\right)}{\sum_{j=1}^{C} \hat{\tau}_j^{(s-1)} \hat{\beta}_{uj}^{(s-1)} f\left(y_i; \hat{\theta}_j^{(s-1)}\right) + \sum_{j=C+1}^{G} \hat{\tau}_j^{(s-1)} f\left(y_i; \hat{\theta}_j^{(s-1)}\right)} & \text{diğer durumlarda} \end{cases} \quad (2.148)$$

- M adımı: formülle $\hat{\beta}_{c_i g}^{(s)}$ (2.149)'daki eşitlik ile hesaplanır:

$$\hat{\beta}_{c_i g}^{(s)} = \frac{\sum_{\forall i \in \{i | c_i = c\}} t_{ig}}{\sum_{i=1}^{n} t_{ig}} \quad (2.149)$$

Burada sınıflama işlemi (2.150)'deki $P(C = c|y_i; \theta)$ olasılık yoğunluk fonksiyonu formülüne göre yapılır:

$$P(C = c|y_i; \theta) = \frac{\sum_{g=1}^{C} \tau_g \beta_{cg} f(y_i; \theta)}{\sum_{g=1}^{C} \tau_g f(y_i; \theta)} \quad (2.150)$$

Şayet yeni bir sınıf ortaya çıkmışsa bu durumda (2.151)'deki eşitlikle kıyaslama yapılır:

$$P(C = u|y_i) = 1 - \sum_{g=1}^{C} P(G = g|y_i; \theta) \quad (2.151)$$

$P(G = g|y_i; \theta)$, (2.130)'daki gibi hesaplanır. Bu konuda çalışmalardan bir tanesi de Bouveyron (2014)'e aittir. Bu çalışmada kısıtlandırılmış EM algoritması önerilmiştir:

E adımı: Burada $f\left(y_i^*; \hat{\Theta}^{(s-1)}\right)$ karma model formülü ve $f_g\left(y_i^*; \hat{\theta}_g^{(s-1)}\right)$ g. bileşen için olasılık formülü olmak üzere $t_{ig}^{*(s)}$ (2.152)'deki gibi hesaplanır:

$$t_{ig}^{*(s)} = \frac{\hat{\tau}_g^{(s-1)} f_g\left(y_i^*; \hat{\theta}_g^{(s-1)}\right)}{f\left(y_i^*; \widehat{\theta}^{(s-1)}\right)} \tag{2.152}$$

M adımı: (2.153), (2.154), (2.155) ve (2.156)'daki eşitliklerle güncelleme yapılır:

$$\hat{\tau}_g^{(s)} = \begin{cases} \left(1 - \sum_{j=c+1}^{G} \frac{n_l^{*(s)}}{n^*}\right) \frac{n_g}{n} & g = 1, \dots, C \\ \frac{n_g^{*(s)}}{n^*} & g = C+1, \dots, G \end{cases} \tag{2.153}$$

$$n_g^{*(s)} = \sum_{i=1}^{n^*} t_{ig}^{*(s)} \tag{2.154}$$

$$\hat{\mu}_g^{(s)} = \frac{1}{n_g^{*(s)}} \sum_{i=1}^{n^*} t_{ig}^{*(s)} y_i \tag{2.155}$$

$$\hat{\Sigma}_g^{(s)} = \frac{1}{n_g^{*(s)}} \sum_{i=1}^{n^*} t_{ig}^{*(s)} \left(y_i - \hat{\mu}_g^{(s)}\right)' \left(y_i - \hat{\mu}_g^{(s)}\right) \tag{2.156}$$

Bouveyron (2014) tarafından bu yöntem uyarlamalı karma diskriminant analizi olarak adlandırılmıştır.

2.4.5 Model tabanlı kümelemede model seçimi

Model seçimi, Law vd. (2004) ve Tadesse vd. (2005) tarafından ilk defa çalışılmıştır (Bouveyron vd., 2019). Law vd. (2004), değişkenleri gerekliliğine göre ayırdıktan sonra her birinin bağımsız değişkenler olarak saymış ve (2.157)'deki gibi olasılık yoğunluk fonksiyonlarınının karma modeline göre düzenlemiştir:

$$P(y_i|\theta, P) = \sum_g \tau_g \prod_{j=1}^{d} \left[p_j P\left(y_{ij}|\theta_{ij}\right) + \left(1 - p_j\right) \varphi\left(y_{ij}|\mu_j, \sigma_j^2\right)\right] \tag{2.157}$$

Bu modelde $P(y_{ij}|\theta_{ij})$, kümelemede kullanılan gereksiz değişkenlere ait θ_{ij} parametreli olasılık yoğunluk fonksiyonunu; $\varphi(y_{ij}|\mu_j, \sigma_j^2)$, orijinal değişkenlere ait μ_j ortalamalı σ_j^2 varyanslı normal dağılımın olasılık yoğunluk fonksiyonlarını; τ_g, karma model ağırlıklarını ve p_j tahmin edilmesi gereken değişkenleri temsil etmektedir. Burada p_j tahmin edilerek edilerek gereksiz değişkenler atılır (Bouveyron vd., 2019). Tadesse vd. (2005), Bayesçi yaklaşımı kullanarak değişken seçimini gerçekleştirmiştir. Bu değişkenlerin gerekliliğine göre ayrım yaparak ve değişkenlerin bağımsız olduğunu varsayarak bu iş gerçekleştirilir ancak Raftery ve Dean (2006), gereksiz değişkenlerin sıklıkla gereklilerle ilişkili olduklarını farkettiler ve kendi seçim algoritmalarını tanımladılar. Bu çalışmaları, daha sonra Maugis vd. (2009) tarafından RD-MCM yöntemine dönüştürülmüştür. Bu yöntemde değişkenlere aşağıdaki olası 4 rol biçilmiştir:

i. Gerekli kümeleme değişkenleri: Bu değişkenlerin kümesi S ile temsil edilmiştir.
ii. Gereksiz kümeleme değişkenleri: Bu değişkenlerin kümesi U ile temsil edilmiştir.
iii. Gereksiz ama gerekli kümeleme değişkenlerini açıklayan kümeleme değişkenleri: Bu değişkenlerin kümesi R ile temsil edilmiştir.
iv. Bağımsız kümeleme değişkenleri: Bu değişkenlerin kümesi W ile temsil edilmiştir.

Böyle veri için olasılık yoğunluk fonksiyonu (2.158)'deki şekilde ifade edilir:

$$f(y_i|G, m, r, l, V, \theta) = \sum_{g=1}^{G} \tau_g \varphi(y_i^S|\mu_g, \Sigma_{g(m)}) \varphi(y_i^U|a + y_i^R b, \Omega_{(r)}) \varphi(y_i^W|\gamma, \phi_{(l)}) \qquad (2.158)$$

Bu denklemde y_i^S, y_i^U, y_i^R, y_i^W sırayla S, U, R ve W kümelerine düşen y_i gözlem değerlerini ifade eder ve. $\varphi(.|a, b)$ fonksiyonunda a ortalama vektörünü ve b kovaryans matrisini ifade ediyor. RD-MCM yöntemiyle (2.159)'daki kriter optimize edilir (Bouveyron vd., 2019):

$$CRIT(G, m, r) = BIC_{clust}(y^S|G, m) + BIC_{reg}(y^U|r, y^R) + BIC_{ind}(y^W) \qquad (2.159)$$

Burada G, çok boyutlu karma modeldeki bileşen sayısını, m eldeki normal karma model, BIC_{clust}, BIC_{reg}, BIC_{ind} sırayla karma modelde yer alan kümeleme değişkenleriyle oluşturulan BIC kriterlerini, gereksiz değişkenlerin R kümesine düşen değişkenler tarafından modellenmesiyle elde edilen BIC kriterini ve W kümesine düşen değişkenlerle elde edilen normal karma modelin BIC kriterini ifade etmektedir. Bu kriteri optimize etmek için karma model, küme sayısını ve S, U, R, W kümelerini eş zamanlı olarak düzenlemek gerekir. Bu düzenlemenin hızlı bir biçimde yapılması için Celeux vd. (2019), aşağıdaki adımları önermiştir:

1. adım: Zhou vd. (2009) tarafından geliştirilen yöntem ile değişkenlerin önemliliğinin sıralanması
2. adım: 1. adımda elde edilen bilgiden yola çıkılarak değişkenlerin rollerinin RD-MCM tarafından belirlenmesi

2.4.6 Açıklayıcı değişkenlerle kurulan karma modeller üzerinden kümeleme

Jacobs vd. (1991) karma modeli açıklayıcı değişkenlerle tanımlayan çalışmalardan biridir. Çalışmasında olasılık yoğunluk fonksiyonunu (2.160)'daki gibi tanımlamıştır:

$$f(y_i|x_i) = \sum_{g=1}^{G} \tau_g(x_i) f_y\left(\psi\left(\gamma_g' x_i\right)\right) \tag{2.160}$$

Burada $\tau_g(x_i) = \frac{\exp(\beta_g' x_i)}{\sum_{h=1}^{G} \exp(\beta_h' x_i)}$ açıklayıcı değişkenlerle tanımlanan karma model ağırlıkları olurken, $\psi\left(\gamma_g' x_i\right) = \theta_g(x_i)$ açıklayıcı değişkenler modeli olarak ifade edilebilmektedir. Karma model ile ilgili 4 varsayım bulunmaktadır. Bunların her biri başka model oluşturmaktadır:

1. y_i ile z_i ilişkili ancak z_i x_i'den bağımsız

2. y_i hem x_i hem z_i ile ilişkili ancak z_i x_i'den bağımsız
3. y_i hem x_i hem z_i ile ilişkili ancak y_i, $x_i|z_i$'den bağımsız
4. y_i hem x_i hem z_i ile ilişkili ayrıca z_i ile x_i ilişkili

Bu veri için olabilirlik fonksiyonu (2.161)'deki gibi ifade edilmiştir:

$$L(\beta,\gamma) = \prod_{i=1}^{N}\sum_{g=1}^{G} \tau_g(x_i) f\left(y_i|\theta_g(x_i)\right) \tag{2.161}$$

Tam veri log-olabilirlik fonksiyonu (2.162)'deki gibi ifade edilmiştir:

$$\begin{aligned} L_c(\beta,\gamma) &= \prod_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\left(\left[\tau_g(x_i) f\left(y_i|\theta_g(x_i)\right)\right]\right) \\ &= \prod_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\left(\tau_g(x_i)\right) + \prod_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\left(f\left(y_i|\theta_g(x_i)\right)\right) \end{aligned} \tag{2.162}$$

Bu formüle göre EM algoritması aşağıda açıklandığı gibi çalışmaktadır:

E adımı: z_{ig}'ler (2.163) ile hesaplanır:

$$\hat{z}_{ig}^{(t+1)} = \frac{\hat{\tau}_g^{(t)}(x_i) f\left(y_i|\theta_g(x_i)\right)}{\sum_{h=1}^{G}\hat{\tau}_h^{(t)}(x_i) f\left(y_i|\theta_h(x_i)\right)} \tag{2.163}$$

M adımı: Bu adımda tam veri log-olabilirlik fonksiyonuna göre optimizasyon yapılır. $\prod_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\left(\tau_g(x_i)\right)$ kısmını optimize etmek çoklu lojistik regresyon modelini optimize etmek iken, $\prod_{i=1}^{N}\sum_{g=1}^{G} z_{ig}\log\left(f\left(y_i|\theta_g(x_i)\right)\right)$'yi optimize etmek G farklı genelleştirilmiş lineer modellerin optimize edilmesiyle aynıdır.

2.4.7 Görüntü analizi için kullanılan özel kümeleme analizi

Bir resmi gürültüden arındırmak için 8x8 piksellik kümeler seçilir (Bouveyron vd., 2019). Burada model kurulurken, X verisinde bozulmanın doğrusal olduğu, X'in G bileşenli normal karma modele sahip olduğu ve ε varyansı $\sigma^2 I_d$ olacak şekilde normal dağıldığı varsayımında bulunulur. Böyle durumda model (2.164)'teki eşitlik ile ifade edilir:

$$Y = \phi(X) + \varepsilon = UX + \varepsilon \tag{2.164}$$

Burada $\phi(X) = UX$, orjinal resimde bozulmayı gösterir; ve ε gürültüyü göstermektedir. Burada ϕ fonksiyonu ile ε biliniyor olarak kabul edilmektedir. Böyle bir durumda beklenen değer, (2.165)'teki formülle hesaplanır:

$$E(X|Y = y) = \sum_{g=1}^{G} P(Z = g|Y = y)\left[\mu_g + \Sigma_g U'\left(U\Sigma_g U' + \sigma^2 I_d\right)^{-1}\left(y - U\mu_g\right)\right] \tag{2.165}$$

2.4.8 Model tabanlı ortak kümeleme

Bu analiz, veriyi bir matris olarak alındığında satır ve sütunlara göre kümeleyerek saklı bir yapıyı ortaya çıkarmak için yapılır (Bouveyron vd., 2019). Bu tarz kümelemeye blok kümeleme de denir. Bu kümeleme, deterministik ve model tabanlı diye ikiye ayrılır. Bu konuda geliştirilmiş algoritmalar, değişimli EM (VEM) ve stokastik EM (SEM)'dir. Burada z_{ig} satır küme değerleri ile w_{jl} sütun küme değerleri, birbirinden bağımsız kabul edilmektedir ve satır ve sütunlardan bloğa dahil edilip edilmeyeceğini belirler. Bunun yanında 1 veya 0 değerlerini alırlar. VEM, Bayesçi mantık ile hareket eder ancak başlangıç değerlerine bağımlıdır (Bouveyron vd., 2019). İşleyişi şu şekildedir:

1.adım: $P\left(z|y, w^{(c)}; \theta^{(c)}\right)$'den $z^{(c+1)}$ seçilir.

2.adım: $P\left(w|y, z^{(c+1)}; \theta^{(c)}\right)$'den $w^{(c+1)}$ seçilir.

3.adım: $z_{.g}^{(c+1)} = \sum_i z_{ig}$ olacak şekilde $D\left(a + z_{.1}^{(c+1)}, \dots, a + z_{.g}^{(c+1)}\right)$'den $\pi^{(c+1)}$ seçilir.

4.adım: $w_{.l}^{(c+1)} = \sum_i w_{jl}$ olacak şekilde $D\left(a + w_{.1}^{(c+1)}, \dots, a + w_{.L}^{(c+1)}\right)$'den $p^{(c+1)}$ seçilir.

5.adım: $N_{gl}^{h(c+1)} = \sum_{i,j} z_{ig}^{(c+1)} w_{jl}^{(c+1)} y_{ij}^{h}$ olmak üzere $D\left(b + N_{gl}^{1(c+1)}, \dots, b + N_{gl}^{r(c+1)}\right)$'den $\alpha_{gl}^{(c+1)}$ seçilir.

a ve b parametreleri düzenleme için EM algoritmasına dahil edilir. Bu nedenle bu parametrelerin seçimi çok önemlidir. Küme sayısı 8 'den küçük ise a=4 olması önerilmektedir ve daha fazla sayıda küme gerekirse, bu durumda a=16 olması önerilir (Frühwirth-Schnatter, 2011a; Keribin vd., 2015). Bu durumda EM algoritması, (2.167), (2.168) ve (2.169) eşitlikleri ile çalışır:

- E adımı: $E\left(\log\left(P(y, z, w|\theta)\right)|y, \theta^{(c)}\right)$ hesaplanır:
- M adımı: $s_{ik}^{(c+1)}$ ve $t_{jl}^{(c+1)}$ satır ve sütun etiketleri ve $\sum_i s_{ik}^{(c+1)} = s_{.k}^{(c+1)}$ ve $\sum_j t_{jl}^{(c+1)} = t_{.l}^{(c+1)}$ olmak üzere $\pi_k^{(c+1)}$, $p_l^{(c+1)}$, $\alpha_{gl}^{h(c+1)}$ (2.166), (2.167) ve (2.168) ile ifade edilir:

$$\pi_k^{(c+1)} = \frac{a - 1 + s_{.k}^{(c+1)}}{n + G(a-1)} \tag{2.166}$$

$$p_l^{(c+1)} = \frac{a - 1 + t_{.l}^{(c+1)}}{d + L(a-1)} \tag{2.167}$$

$$\alpha_{gl}^{h(c+1)} = \frac{b - 1 + \sum_{i,j} s_{ig}^{(c+1)} t_{jl}^{(c+1)} y_{ij}^{h}}{r(b-1) + s_{.k}^{(c+1)} t_{.l}^{(c+1)}} \tag{2.168}$$

Böylece model seçiminde kullanılan g ve m değerlerine göre değişen ICL kriterinin formülü (2.169)'daki gibidir:

$$\begin{aligned}\mathrm{ICL}(g,m) = {} & \log\left(\Gamma(Ga)\right) + \log\left(\Gamma(La)\right) - (L+G)\log\left(\Gamma(a)\right) \\ & + LG\left(\log\left(\Gamma(rb)\right) - r\log\left(\Gamma(b)\right)\right) - \log\left(\Gamma(n+Ga)\right) \\ & - \log\left(\Gamma(d+La)\right) + \sum_{g}\log\left(\Gamma\left(z_{.g}+a\right)\right) + \sum_{l}\log\left(\Gamma(w_{.l}+a)\right) \\ & + \sum_{g,l}\left[\left(\sum_{h}\log\left(\Gamma\left(N_{gl}^{h}+b\right)\right)\right) - \log\left(\Gamma\left(z_{.g}w_{.l}+rb\right)\right)\right]\end{aligned} \tag{2.169}$$

Burada z ve w değerleri iteratif olarak sırasıyla (2.170) ve (2.171)'deki gibi hesaplanır:

$$z^{(c)} = \mathrm{argmax}_{z}\left(P\left(z|y, w^{(c-1)}; G, L, \hat{\theta}\right)\right) \tag{2.170}$$

$$w^{(c)} = \mathrm{argmax}_{w}\left(P\left(w|z^{(c)}; G, L, \hat{\theta}\right)\right) \tag{2.171}$$

2.4.9 Mcnicholas panel veri kümeleme yöntemi

Mcnicholas panel veri kümeleme yöntemi, Cholesky ayrıştırmasına dayanan bir kümeleme yöntemidir. Cholesky ayrıştırması, Benoit (1924) tarafından duyurulmuş bir matrisi bir alt üçgensel matrisi ile transpozunun çarpımı şeklinde ifade etmektir. Pourahmadi (1999), Cholesky ayrıştırmasını değiştirerek panel veride uygulanmasının yolunu açmıştır. Bu çalışmada, bir rassal vektörün kovaryans matrisi Σ için (2.172)'deki gibi formül geliştirilmiştir:

$$T\Sigma T' = D \tag{2.172}$$

Burada T tekil birim (köşegen elemanları 1 olan) alt üçgensel matris ve D tekil köşegensel bir matristir. Bu formül, (2.173)'teki eşitliğe dönüştürülebilir:

$$\Sigma^{-1} = TD^{-1}T' \tag{2.173}$$

Bu formülle T ve D'nin elemanları, (2.174)'teki eşitlikte belirtilen genelleştirilmiş otoregresif modeli tanımlamaktadır:

$$\widehat{X}_t = \mu_t + \sum_{s=1}^{t-1}(-\varphi_{ts})(X_s - \mu_s) + \sqrt{d_t}\varepsilon_t \qquad (2.174)$$

Burada φ_{ts} T'nin t. satır ve s. sütundaki elemanı, ε_t 0 ortalamalı 1 standart sapmalı normal dağılımdan gelmekte ve d_t ise D matrisinin t. satırdaki köşegen elemanıdır. Bu şekilde ifade edildiğinde panel veride Cholesky ayrıştırmalı karma modelleme gerçekleştirilir. Buna göre bu modelleme her T_g ve D_g için başka sonuçlar elde etmektedir. Bunun sonucunda ortaya çıkan model tipleri(Cholesky ayrıştırmasından elde edilen T_g ve D_g'nin özelliklerine göre belirlenmiştir) Çizelge 2.1'de gösterilmiştir:

Çizelge 2.1: McNicholas (2017) yönteminde model sınıflarına karşılık gelen özellikler

Model	T_g	D_g	D_g	boş kovaryans parametreleri
EEA	eşit	eşit	anizotropik	p(p-1)/2+p
DDA	değişken	değişken	anizotropik	G[p(p-1)/2]+Gp
DEA	değişken	eşit	anizotropik	G[p(p-1)/2]+p
EDA	eşit	değişken	anizotropik	p(p-1)/2+Gp
DDİ	değişken	değişken	izotropik	G[p(p-1)/2]+G
DEİ	değişken	eşit	izotropik	G[p(p-1)/2]+1
EDİ	eşit	değişken	izotropik	p(p-1)/2+G
EEİ	eşit	eşit	izotropik	p(p-1)/2+1

McNicholas (2017), bu durumlardan DEA ve EDİ'ye değinmiştir. EDA haricindeki diğer durumlarda hepsinde hesaplamaların benzer olduğunu belirtmiştir. Ayrıca EDA durumu için McNicholas ve Murphy (2010)'a bakılmasını tavsiye etmiştir. EDA durumu için belirli bir EM algoritması oluşturmak mümkündür ve başlangıç noktası çalışması yapılabilir.

2.4.9.1 Mcnicholas'ın kümeleme yönteminde DEA durumu

Cholesky ayrıştırması yapıldıktan sonra her bir bileşenin (X) olasılık yoğunluk fonksiyonu (2.175)'teki şekilde ifade edilir:

$$\phi\left(X|\mu_g,\left(T_g'D_gT_g\right)^{-1}\right)=\frac{1}{\sqrt{(2\pi)^p|D_g|}}\exp\left(-\frac{1}{2}(X-\mu_g)'(T_g'D_gT_g)(X-\mu_g)\right) \quad (2.175)$$

DEA durumunda tam veri log-olabilirlik fonksiyonu (2.176)'daki gibi ifade edilir:

$$L(\mu_g,\pi_g,z_{ig})=\sum_{i=1}^{n}\sum_{g=1}^{G}z_{ig}\left[\log(\pi_g)+\log\left(\phi\left(x_i|\mu_g,\left(T_g'D_gT_g\right)^{-1}\right)\right)\right] \quad (2.176)$$

Burada z_{ig}, i. gözlemin g. gruba girip girmediğini belirten bir gölge değişkendir ve i. gözlem g. gruba ait ise 1; değilse 0 değerini alır. π_g ise g. grup için ağırlık parametresidir. Bunun yanında $D_g = D$ olarak ele alınır. Bu durumda tahminler (2.177), (2.178), (2.179), (2.180) ve (2.181)'deki gibi ifade edilir:

$$\widehat{z_{ig}}=\frac{\widehat{\pi_g}\phi\left(x_i|\mu_g,\left(T_g'D_gT_g\right)^{-1}\right)}{\sum_{h=1}^{G}\widehat{\pi_h}\phi(x_i|\mu_h,(T_h'D_hT_h)^{-1})} \quad (2.177)$$

$$\widehat{S_g}=\frac{\sum_{i=1}^{n}\widehat{z_{ig}}(X_i-\mu_g)(X_i-\mu_g)'}{\sum_{i=1}^{n}\widehat{z_{ig}}} \quad (2.178)$$

$$\widehat{\pi_g}=\frac{\sum_{i=1}^{n}\widehat{z_{ig}}}{n} \quad (2.179)$$

$$\widehat{\mu_g}=\frac{\sum_{i=1}^{n}\widehat{z_{ig}}X_i}{\sum_{i=1}^{n}\widehat{z_{ig}}} \quad (2.180)$$

$$\widehat{D}=\sum_{g=1}^{G}\widehat{\pi_g}\,\mathrm{diag}\left(\widehat{T_g}\widehat{S_g}\widehat{T_g}'\right) \quad (2.181)$$

Burada $\widehat{T_g}$ köşegen elemanları 1 olan alt üçgensel bir matris olduğundan i>j için her elemanı $\varphi_{ij}^{(g)}$'den oluşan bir matristir. $\varphi_{ij}^{(g)}$ tahmini için (2.182)'teki eşitlikten faydalanılır:

$$\begin{bmatrix} \widehat{\varphi}_{r1}^{(g)} \\ \vdots \\ \widehat{\varphi}_{r,r-1}^{(g)} \end{bmatrix} = -\begin{bmatrix} s_{11}^{(g)} & \cdots & s_{1,r-1}^{(g)} \\ \vdots & \ddots & \vdots \\ s_{r-1,1}^{(g)} & \cdots & s_{r-1,r-1}^{(g)} \end{bmatrix}^{-1} \begin{bmatrix} s_{r1}^{(g)} \\ \vdots \\ s_{r,r-1}^{(g)} \end{bmatrix} \tag{2.182}$$

Burada $s_{ij}^{(g)}$ S_g 'nın i. satır ve j. sütundaki elemanıdır.

2.4.9.2 Mcnicholas'ın kümeleme yönteminde EDİ durumu

EDİ durumunda, $T_g = T$ ve $D_g = \lambda_g I_p$ olmaktadır. Bu durumda tam veri(verinin tamamını kattığımızda) log-olabilirlik fonksiyonu (2.183)'teki gibi ifade edilir:

$$L = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left[\log(\pi_g) + \log\left(\phi\left(x_i|\mu_g, \left(\lambda_g^{-1}T_g'T_g\right)^{-1}\right)\right)\right] \tag{2.183}$$

Burada z_{ig}, i. gözlemin g. gruba girip girmediğini belirten bir gölge değişkendir ve i. gözlem g. gruba ait ise 1; değilse 0 değerini alır. π_g ise g. grup için ağırlık parametresidir. Bunun yanında $D_g = D$ olarak ele alınır. Bu durumda tahminler (2.184), (2.185), (2.186), (2.187) ve (2.188)'deki gibi ifade edilir:

$$\widehat{z_{ig}} = \frac{\widehat{\pi_g}\phi\left(x_i|\mu_g, \left(\lambda_g^{-1}T_g'T_g\right)^{-1}\right)}{\sum_{h=1}^{G}\widehat{\pi_h}\phi(x_i|\mu_h, (\lambda_h^{-1}T_h'T_h)^{-1})} \tag{2.184}$$

$$\widehat{S_g} = \frac{\sum_{i=1}^{n} \widehat{z_{ig}} (X_i - \mu_g)(X_i - \mu_g)'}{\sum_{i=1}^{n} \widehat{z_{ig}}} \tag{2.185}$$

$$\widehat{\pi_g} = \frac{\sum_{i=1}^{n} \widehat{z_{ig}}}{n} \tag{2.186}$$

$$\widehat{\mu_g} = \frac{\sum_{i=1}^{n} \widehat{z_{ig}} X_i}{\sum_{i=1}^{n} \widehat{z_{ig}}} \tag{2.187}$$

$$\widehat{\lambda_g} = \frac{1}{p} \mathrm{tr}\{\widehat{T} S_g \widehat{T}'\} \tag{2.188}$$

Burada $\widehat{T_g}$ köşegen elemanları 1 olan alt üçgensel bir matris olduğundan i>j için her elemanı $\widehat{\varphi}_{ij}$'den oluşan bir matristir. $\widehat{\varphi}_{ij}$ tahmini için (2.189)'daki eşitlikten faydalanılır:

$$\begin{bmatrix} \widehat{\varphi}_{r1} \\ \vdots \\ \widehat{\varphi}_{r,r-1} \end{bmatrix} = - \begin{bmatrix} K^{11} & \cdots & K^{r-1,1} \\ \vdots & \ddots & \vdots \\ K^{1,r-1} & \cdots & K^{r-1,r-1} \end{bmatrix}^{-1} \begin{bmatrix} K^{r1} \\ \vdots \\ K^{r,r-1} \end{bmatrix} \tag{2.189}$$

Burada $K^{ij} = \sum_{g=1}^{G} \frac{s_{ij}^{(g)}}{\lambda_g}$ olmaktadır. Bu durumların dışında köşegen altı elemanlarını sınırlandıran durumlar bulunmaktadır.

2.4.10 Panel veri'de kümeleme analizine Frühwirth-Schnatter'in bayesçi bakışı

Frühwirth-Schnatter (2011a), çalışmasında model tabanlı kümeleme üzerine ortaya çıkan literatürü değerlendirmiştir. Frühwirth-Schnatter (2011a), Markov zincirlerinin model tabanlı kümelemede nasıl kullandığını anlatmıştır. Panel verideki her bir zaman serisi bir Markov zinciri olarak değerlendirdiğinde, (2.190) ile (2.191)'deki gibi ayrı ayrı ifade edilmektedir:

$$P(y_i | v_h) = \prod_t P(y_{it} | y_{i,t-1}, v_h) \tag{2.190}$$

$$P(y_i|v_h) = \prod_t P(y_{it}|y_{i,t-1}, x_{it}, v_h) \tag{2.191}$$

Bu formüllerde y_{it}, yanıt değişkenini; x_{it}, açıklayıcı değişkenleri ve v_h sahip olunan parametreleri ifade etmektedir. $P(y_i|v_h)$'lerin hepsi karma modeldeki bileşenleri ifade etmektedir. Bu yöntemle analiz yapabilmek için küme sayısını ve kümeleme çekirdek fonksiyonu $P(y_i|v_h)$'yi seçmek gerekir.

Bayesçi yaklaşım, bir dağılımın parametresini bir rasgele değişken olarak ele alınmasını esas alır ve böylelikle elde edilen sonsal dağılımdan çıkarımlar gerçekleştirir. Bayesçi kümelemede ise veri farklı bir biçimde ele alınabilmektedir. Pamminger (2007), çalışmasında Bayesçi kümelemeyi markov zincirlerinin üzerine kurmuştur. Bu yapıyı işlemek için kategorik bir veri seçen Pamminger (2007), Markov zincir olasılıklarını (ε) Dirichlet dağılımına$\left(D(e_{N,j1}, \dots, e_{N,jK})\right)$ sahip olarak ele almıştır (K durumların sayısıdır). Bu noktada iteratif olarak parametrelerin güncellenmesi için markov zincirindeki fazlar arası geçişlerin sayısından(N_{jk}) faydalanılmaktadır ve olasılıklar için varyans ve sonsal mod sırasıyla (2.192) ve (2.193)'teki eşitliklerle ifade edilir:

$$\mathrm{Var}(\varepsilon_{jk}|y) = \frac{\varepsilon_{jk}(1-\varepsilon_{jk})}{\sum_l N_{jl}} \tag{2.192}$$

$$\varepsilon_{mod,jk} = \frac{N_{jk} + e_{0,jK} - 1}{\sum_l N_{jl} + \sum_l e_{0,jl} - K} \tag{2.193}$$

Buna ek olarak, işlem yapılırken aşağıdaki hususlar dikkate alınmalıdır:

i. Her birim için sınıf çoklu dağılımdan seçilir.
ii. Her sınıf için ağırlıklar seçilir ki bu ağırlıklar Dirichlet dağılımına sahiptir.
iii. Örneklem bileşen parametreleri $v_1, \dots, v_H$ seçilir (burada karma modeldeki dağılımlar önemlidir.)

Bu noktada grup sayısını belirlemek için geri atlamalı MCMC yöntemleri kullanılmaktadır. Model yeterliliğini test etmek için AIC ve BIC kriterleri kullanılmaktadır. Sonsal tahmin yeterliliğini test etmek de analizin bir parçasıdır ve p-değeriyle yapılmaktadır.

Kümeleme analizi adından anlaşılabileceği gibi veride küme oluştularak anlamlı yapılar elde etme çabasıdır. Genellikle kümeleme kriterleri, (2.194)'teki toplam varyansın (T) iki ayrı parçaya ayrıştırılması ile elde edilir:

$$T = \sum_{i=1}^{N} (y_i - \bar{y})(y_i - \bar{y})' \tag{2.194}$$

Burada T varyansı, (2.195)'teki W(S) ve (2.196)'daki toplam gruplararası varyans B (S) olarak ayrıştırılır:

$$W(S) = \sum_{k=1}^{K} W_k(S) \tag{2.195}$$

$$B(S) = \sum_{k=1}^{K} N_k(S)(\bar{y}_k(S) - \bar{y})(\bar{y}_k(S) - \bar{y})' \tag{2.196}$$

Burada $N_k(S), \bar{y}_k(S)$ ve $W_k(S)$ sırasıyla grup büyüklüğü, grup ortalaması ve grup içi varyansı ifade etmektedir. Bu ifadeler kullanılarak, (2.197) ve (2.198)'deki iki tür uyumluluk kriteri oluşturulur:

$$A = 1 - \frac{tr(W(S))}{tr(T)} \tag{2.197}$$

$$B = 1 - \frac{|W(S)|}{|T|} \tag{2.198}$$

B test istatistiği, Friedman ve Rubin (1967) tarafından önerilmiştir. Frühwirth-Schnatter (2006), bunların çok boyutlu normal karma modelleriyle B'nin ilişkili olduklarını ve parametre

tahmininde kullanıldığından bahsetmiştir. Bunun yanında model tabanlı kümelemenin çok pratik uygulamaları olmuştur. Normal karma modelleri kullanarak model tabanlı kümeleme yapıldığında karakter tanımlama (Murtagh ve Raftery,1984), doku segmentasyonu (Banfield ve Raftery, 1993), sismik fay ve madenyeri belirleme (Dasgupta ve Raftery, 1998), genetik veri kümelemesi(Yeung et al., 2001) ve astronomik veri sınıflandırması (Celeux and Govaert, 1995) yapılabilir. Frühwirth-Schnatter (2006), model-tabanlı kümelemenin Bayesçi yaklaşımla klasik yaklaşıma göre üç avantajı olduğunu belirtmektedir: Birincisi EM algoritmasının dejenere sonuçlar vermesi ki Bayesçi yaklaşım düzgün önsel dağılımlarla dejenere sonuç vermesini önleyebilir, ikincisi Bayesçi yaklaşımın daha prensipli bir yaklaşım olması ve son olarak güçlü Markov Zincir Monte Karlo yöntemlerinin teorik sonsal dağılıma son derece yaklaşmasıdır. Son iddia, kümelemenin başarıya ulaşabilmesi için çok önemlidir. Sınıflandırma için kullanılan klasik olabilirlik fonksiyonu (2.199)'daki eşitlikle ifade edilmiştir:

$$P(y|S,v)P(S|v) = \prod_{i=1}^{N} P\left(y_i|\theta_{S_i}\right) \prod_{k=1}^{N} {n_k}^{N_k(S)} \tag{2.199}$$

Burada S_i, y verisinden oluşturulan kümeleri; θ_{S_i}, S_i'ye karşılık gelen parametre vektörünü ve n_k k. bileşenin ağırlığını tahmin etmek için gerekli değişkeni temsil etmektedir. Buna karşılık Bayesçi yöntemde MAP(Maximum A Posteriori) yöntemiyle sınıflandırma yapılır ki burada olabilirlik fonksiyonu (Frühwirth-Schnatter (2006) bunu birleşik sonsal olasılık yoğunluk fonksiyonu olarak tanımlıyor) (2.200)'deki eşitlikle ifade edilir:

$$P(v,S|y) = P(y|v,S)P(S|v)P(v) \tag{2.200}$$

3. TEORİK BİLGİ

Panel veri hem birim etkisinin hem de zamana bağlı etkinin cereyan ettiği veri türüdür. Birim'den kastımız, zaman serisini elde ettiğimiz yapıdır. Bunlara örnek olarak, ülkeler, iller, ilçeler vs. verilebilir. Panel veriye örnek olarak, Türkiye'deki ilçelerin aylara göre elektrik tüketimi gösterilebilir.

Panel veriyi toplamanın yollarından biri panel anketleridir. Panel anketleri, Amerika'da 1980'lerin ortalarından itibaren toplanmaya başlanan veri elde etme yolu olmuştur.

Türkiye'de 15.06.2017 tarihine kadar panel veri analizi üzerine yazılmış bütün tezler 141 tanedir. Bunlardan ilki Cihan Yalçın'ın 1995 tarihli Türkiye imalat sanayiinde fiyat-maliyet marjları ve dış ticaret: Bir panel veri analizi'dir.

Bu bölümde panel verinin yararlarını, panel veride analiz gerçekleştirmeyi, panel veriye uygun klasik doğrusal regresyon modeli oluşturmayı, doğrusal model için parametre tahmini gerçekleştirmeyi ve alternatif model türlerini inceleyeceğiz.

3.1 Panel Verinin Yararları

Panel veri analizi, tahmin yanlılığını düşürmeyi, daha doğru öngörülerde bulunmayı, veri birleşimini daha uygun seviyede nasıl oluşturabileceğini ve parametre tahmin elde etme ile istatistiksel çıkarsamayı basitleştirmeyi sağlar. Bu özellikler birkaç cümle ile özetlenecek kadar basittir. Örneğin, tahmin yanlılığı model seçiminde dışladığımız değişkenlerden kaynaklanabilmektedir. Doğru öngörülerde bulunmak için modelde kategorik olarak önemli sayılan faktörleri barındırmak gerekir. Her model gerçeklikten sınırlı anlamda kopuşu ifade etse de(Bu modelin doğası gereği böyledir), bu öngörüleri etkileyecek kadar önemli olmamalıdır. Panel veri analizi, birimler arası benzerliği hesaba katarak öngörüler üzerinde yanlılığı azaltır. Bunun yanında panel veri analizi, model seçiminde faydalı değişkenleri katarak tahmin yanlılığını azaltır. Panel veride, her birim zaman serisi barındırır. Bu zaman serilerinin birbirlerine benzerliği homojenliği ortaya çıkarttığı gibi benzersizliği heterojenliği ortaya çıkarır. Heterojenliğin varlığı, ki tez çalışmasında da ilgilenilen husustur, veri birleşimini ve

model seçimini etkiler. Sözkonusu sorunu aşmak için teorik çalışmalar mevcut olmakla birlikte tez çalışmasının ilerleyen bölümlerinde pratik bir yol sunarak uygulamada basitçe kullanılacak bilgiyi verecektir.

3.1.1 Serbestlik derecesini yükseltmek ve çoklu bağlantı probleminin etkisini azaltmak

Hsiao (2014), çoklu bağlantı problemi ile düşük serbestlik derecesinin sıkça karşılaştığını belirtmektedir. Hsiao (2014) bunu açıklarken, belirtilen modeldeki gereksinimleri karşılamak için yeterli veri zenginliği olmaması olarak görmektedir. Panel veri bu zenginliği taşıyan bir veridir.

3.1.2 Hipotezler arasında ayrıma gitmek ve ayrımı belirleme

Veri üzerinde istatistiksel analiz gerçekleştirdikten sonra sonuçları yanlış yorumlamaya neden olan faktörlerden bir tanesi, modelde değişkenlere etki eden bir gizli değişkendir. İşte bu gizli değişken faktörü yüzünden, bir hipotezin reddedilmesi diğer bir hipotezin reddedilmesine etki edebilir. Panel veri birden fazla birim için regresyon modeli kurduğu için gizli değişken etkisi daha kolay tespit edilebilir ve çözülebilir durumdadır. Bunun nedenleri şu şekildedir:

i. Bireysel ve zamansal etkilerini, gözlem değer farklarını alarak etkisizleştirme
ii. Gölge değişken kullanarak bireyden ve zamandan bağımsız etkileri elde etmek
iii. Gözlenmeyen etkileri dışlayarak gözlenen verilerle uygun model oluşturmak

Panel veri analizi yapılırken, birimler arası heterojenlik sözkonusu analizi etkileyebilmektedir. Bu etki, birimlerin zaman içinde kendilerine has seyirlerinden kaynaklanmaktadır. Bu da birimler arası kıyaslamayı etkilemektedir (Bu konuda bilgi edinmek için Baltagi ve Pesaran(2007)'ye bakılabilir).

Panel veri analizi yapılırken, birimler arasında heterojenliğe doğrudan bir çözüm getirmek için panel veri kümelenebilir ve kümeleri temsil eden gölge değişkenler analize dahil edilebilir.

Panel veri analizi, panel veri ile ilgili model oluşturma için kullanılan istatistiksel bir yöntem olarak tanımlanır. Bu yöntemin uygulanması için kullanılan modelleri ve sözkonusu modellerin parametrelerini tahmin etme yollarını bilmek gerekir. Panel veriye uygulanabilecek modellerden birisi doğrusal regresyon modelidir.

Panel veri için doğrusal regresyon modellerden iki çeşidi kullanılabilmektedir. Söz konusu modeller sabit etkiler modeli ve rassal etkiler modeli olarak adlandırılır. Bu ikisi arasındaki fark, panel veri için uygun görülen doğrusal modelin ifade edilişi ile ilgilidir. Sözkonusu model (3.1)'deki gibi ifade edilebilir:

$$Y = X\beta + \alpha_i + \lambda_t + u_{it} \qquad (3.1)$$

Bu modelde α_i birim etkileri, λ_t zamana ilişkin etkileri, u_{it} hataları temsil etmektedir. (3.1)'de α_i sabit kabul edileceği gibi, rassal değişken olarak da kabul edilebilir. İşte bu sebeple panel veriye uygulanan doğrusal model, sabit etkiler ve rassal etkiler olarak ikiye ayrılmaktadır. Sözkonusu iki modelden rassal etkiler modelinin avantajları, parametre sayısının örneklem büyüklüğü arttıkça sabit kalması, grup içi ve gruplar arası etkilerinin tahmin edicilerinin etkin olması ve zaman değişmezliğine sahip değişkenlerin etkilerinin tahmin edilmesini sağlamasıdır.

Rassal etkiler modelinin dezavantajı, bireysel zaman değişmezliğine sahip değişkenlerin etkisinin dağılımının bilinmemesidir. Bu duruma karşılık, sabit etkiler modelinin avantajları zaman etkileri ile birim etkilerinin birbirleriyle ilişkili olmasının sağlanması ile korelasyon yapılarının incelenmesi gerekmemesi iken, dezavantajları örneklem büyüklüğü arttıkça bilinmeyen parametre sayısnın artması ile zaman değişmezliğine sahip değişkenlerin etkilerinin tahmin edilemez oluşudur.

Panel veri analizinde modelin birimlere göre değişip değişmediğine göre ek analiz prosedürü geliştirilmiştir. Bu analiz prosedürü ANCOVA'dır ve önümüzdeki bölümde incelenecektir.

Panel veri analizi için model parametre tahmini yapmak, matematiksel olarak ciddi bir altyapı gerektiren bir süreçtir. Bu zorluğa rağmen panel veri analizinde bugüne kadar geliştirilen yöntemlerin eksikliğini aşabilmek için teori geliştirmek istatistik biliminin görevidir. Bu bölümde panel veri analizi adına geliştirilen model oluşturma ve model parametre tahmin yöntemleri incelenecektir.

3.2 Panel Veriye Uygun Doğrusal Model Oluşturma

Bu bölümde panel veri analizinde kullanılan teorik modeller, modellerin tahmin edilişi incelenecektir.

3.2.1 Değişken sabit terimlerle panel regresyon modeli

Panel veri için geliştirilen regresyon modelleri, genel bir homojenlik varsayımının sağlanmaması ile oluşturulan modellerdir (Hsiao,2014). Bu modeller için temel varsayım, bütün açıklayıcı değişkenlerin zaman boyunca değişmeyen, birim boyunca değişmeyen ve son olarak hem zaman hem birim boyunca değişen değişkenlerden herhangi biri olmasıdır. Buna ek olarak ihmal edilmiş değişkenlerin tek tek ele alındığında herbirinin önemsiz olmasını ama toplu halde ele alındığında önemli bir rol oynamasını ve modele konulabilecek bütün değişkenlerle ilişkisiz olan rasgele değişkenler olmasını varsayar. Bu tip modellerde, birim boyunca değişmeyen ve zaman boyunca değişmeyen değişkenlerin etkilerindeki hatalar sabit terime eklenir. Bu tip modellere örnek olarak verilebilecek örneklerden bir tanesi, Cobb-Douglas üretim fonksiyonudur ve (3.2)'deki eşitlikle ifade edilmiştir:

$$y_{it} = \mu + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + v_{it} \qquad (3.2)$$

Burada y toplam üretimin logaritmasını, j=1,....,k için x_{jit} üretime giren faktörlerin logaritmalarını ifade etmektedir. Hsiao (2014), bu modelin yönetimsel araçları hesaba katmadığından dolayı eleştirildiğini belirtmektedir. Bu durumda v_{it} (3.3)'teki gibi ifade edilir:

$$v_{it} = \alpha M_i + \lambda P_t + u_{it} \tag{3.3}$$

Bu modelde M_i ve P_t sırasıyla firmaya özel ve zamana özel değişkenler olarak ifade edilir. Bu noktada λ ve α parametreleri için tahmin geliştirmek istendiğinde $\alpha M_i = \alpha_i$ ve $\lambda P_t = \lambda_t$ alınır. Değişken sabitli modelimiz ortaya çıkar. Hoch (1962) tarafından kullanılan bu model başarılı olmuştur. Düzeltilmiş R^2 0,75 iken α_i ve λ_t parametreleri tanıtılınca 0,88 olmuş ve parametre tahminleri değişmiştir. Bu model aynı zamanda üretimin artışının üretimin kalitesiyle doğru orantılı olduğunu göstermiştir. (3.4)'deki eşitlikle ifade edilen model, tipik bir regresyon modeli olarak görülse de aynı zamanda bir ANCOVA modelidir:

$$y_{it} = \alpha_i^* + X_{it}'\beta + u_{it} \tag{3.4}$$

ANCOVA modeli olarak adlandırılmasının sebebini anlamak için ANOVA ile regresyon modeli arasındaki farkı açıklamak gerekir. Hsiao (2014) regresyon modelinde y değişkeninde sadece X değişkeninin etkisi var iken ANOVA modelinde X değişkeninin etkisi yerine y'nin sınıflarının etkisinin var olduğunu belirtmiştir. y'nin sınıflarının etkisi yetersizse, y'deki değişimi açıklamaya ANOVA ile regresyon modelinin uzlaşması olarak ANCOVA modeli girmektedir. Bu eşitlik, $e' = [1, \dots, 1]_{1xT}$ olmak üzere (3.5)'deki gibi ifade edilebilir:

$$Y = \begin{bmatrix} e \\ 0 \\ \vdots \\ 0 \end{bmatrix} \alpha_1^* + \begin{bmatrix} 0 \\ e \\ \vdots \\ 0 \end{bmatrix} \alpha_2^* + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ e \end{bmatrix} \alpha_N^* + X\beta + u \tag{3.5}$$

Burada $Q = I_T - \frac{ee'}{T}$ olmak üzere β'nın LSDV (gölge değişkenli en küçük kareler) tahmin edicisi (3.6)'daki gibi ifade edilir:

$$\hat{\beta} = \left[\sum_{i=1}^{N} X_i' Q X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' Q y_i\right] \tag{3.6}$$

Bu tahmin ediciye grup içi tahmin edicisi ve kovaryans tahmin edicisi denir. Grup içi tahmin edicisi denmesinin nedeni sadece grup içi değişimleri esas alarak tahmin üretmesiyken, kovaryans tahmin edicisi denmesinin nedeni bulunduğu model ANCOVA modeli olmasıdır (Hsiao, 2014). Yansız olmasının yanında N ya da T, veya her ikisi de, büyüdükçe etkin bir tahmin edici olmaktadır ve kovaryans matrisi (3.7)'deki gibi ifade edilir:

$$V(\hat{\beta}) = \sigma_u^2 \left[\sum_{i=1}^{N} X_i' Q X_i\right]^{-1} \tag{3.7}$$

Bu durum $\alpha_i^* = \bar{y}_ı - \bar{x}_ı'\hat{\beta}$ için geçerli değildir. Her ne kadar bunun tahmin edicisi yansız da olsa etkin olması için T'nin büyük olması gerekir. (3.7)'deki eşitlik, (3.8)'deki gibi ifade edilebilir:

$$y_{it} = \mu + \alpha_i + X_{it}'\beta + u_{it} \tag{3.8}$$

Hsiao (2014), bu modelde $\sum_i \alpha_i = 0$ kısıtıyla hem μ hem de α_i'yi tespit edilebilir hale dönüştürülebileceğini göstermiştir. Böyle bir durumda α_i, ortalamadan sapmaları ifade eder hale gelmektedir. Şayet $var(u_{it}) = \sigma_i^2$ olursa, bu durumda gölge değişkenli en küçük kareler tahmin edici BLUE (En iyi doğrusal yansız tahmin edici) olmaktan çıkar. Ancak yine de etkin bir tahmin edicidir. Böyle bir durumda uygulanabilecek bir başka tahmin edici, ağırlıklandırılmış en küçük kareler tahmin edicisidir. Bu durumda $(y_{it}, x_{it}', 1)$ üçlüsüne (3.9)'daki tahmin edicinin tersiyle ağırlıklandırılarak ortaya çıkar (Hsiao, 2014):

$$\widehat{\sigma_ı^2} = \frac{1}{T}\sum_{i=1}^{T}\left(y_{it} - \widehat{\alpha_ı^*} - x_{it}'\hat{\beta}\right)^2 \tag{3.9}$$

Rassal Etkiler Modeli, α_i parametreleri rassal değişken olarak kabul edildiğinde ortaya çıkar. Bu durumda artıklar bu modelde $v_{it} = \alpha_i + \lambda_t + u_{it}$ şeklinde ifade edilir. Bu modelin temel varsayımı (3.10) ile ifade edilir:

$$f(\alpha_i, \lambda_t|X) = f(\alpha_i, \lambda_t) = f(\alpha_i)f(\lambda_t) \tag{3.10}$$

Sözkonusu bağımsızlık varsayımından ötürü rassal etkiler modelinde $V(y_{it}|X_{it}) = \sigma_y^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_u^2$ olmaktadır. σ_y^2 ile gösterilen y'nin varyansı bu şekilde ifade edildiğinden modele varyans bileşenleri modeli de denmektedir (Hsiao, 2014). $\lambda_t = 0$ kabul edilirse bu durumda (3.11)'teki eşitlik, (3.12)'deki eşitlik ile ifade edilir:

$$y_i = \widetilde{X}_\imath \delta + v_i \tag{3.12}$$

Burada $\widetilde{X}_\imath = (e, X_i)$ ve $\delta' = (\mu, \beta')$ olmaktadır. Bu durumda $E(v_i v_i')$, $(3.13)'$teki eşitlik ile ifade edilir:

$$E(v_i v_i') = \sigma_u^2 I_T + \sigma_\alpha^2 ee' = V \tag{3.13}$$

V matrisinin tersi, (3.14)'deki eşitlik ile bulunur (Hsiao, 2014; Graybill, 1969; Nerlove, 1971; Wallace and Hussain, 1969):

$$V^{-1} = \frac{1}{\sigma_u^2}\left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + \sigma_\alpha^2} ee'\right] \tag{3.14}$$

Bu formül genelleştirilmiş en küçük kareler tahmin edicisini elde etmede kolaylık sağlamaktadır. Rassal etkiler modelinde (3.42)'deki tahmin edici kullanıldığında, etkin ve yansız tahmin edici olur. En iyi doğrusal yansız tahmin edici rassal etkiler modeli söz konusu olduğunda genelleştirilmiş en küçük kareler tahmin edicisidir (Hsiao, 2014). (3.15)'deki eşitlik ile genelleştirilmiş en küçük kareler tahmin edicisi elde edilir:

$$\left[\sum_{i=1}^{N} \widetilde{X}_i' V^{-1} \widetilde{X}_i\right] \hat{\delta}_{GLS} = \left[\sum_{i=1}^{N} \widetilde{X}_i' V^{-1} y_i\right] \tag{3.15}$$

(3.15)'deki eşitlikte V^{-1}, (3.16)'daki gibi ifade edilir:

$$V^{-1} = \frac{1}{\sigma_u^2}\left[\left(I_T - \frac{1}{T} ee'\right) + \frac{\psi}{T} ee'\right] = \frac{1}{\sigma_u^2}\left[Q + \frac{\psi}{T} ee'\right] \tag{3.16}$$

Burada $\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$ olmaktadır. Bu durumda (3.15)'deki eşitlik, (3.17)'deki eşitliğe dönüştürülür:

$$[W_{\tilde{x}\tilde{x}} + \psi B_{\tilde{x}\tilde{x}}] \begin{bmatrix} \hat{\mu} \\ \hat{\beta} \end{bmatrix}_{GLS} = \left[W_{\tilde{x}\tilde{y}} + \psi B_{\tilde{x}\tilde{y}}\right] \tag{3.17}$$

Burada $T_{\tilde{x}\tilde{x}} = \sum_{i=1}^{N} \widetilde{X}_i' \widetilde{X}_i$ ve $T_{\tilde{X}Y} = \sum_{i=1}^{N} \widetilde{X}_i' y_i$ olmak üzere (3.17)'deki $W_{\tilde{x}\tilde{x}}, B_{\tilde{x}\tilde{x}}, W_{\tilde{x}\tilde{y}}, B_{\tilde{x}\tilde{y}}$, (3.18), (3.19), (3.20) ve (3.21)'deki eşitliklerle hesaplanır:

$$B_{\tilde{x}\tilde{x}} = \frac{1}{T}\sum_{i=1}^{N} \widetilde{X}_i' ee' \widetilde{X}_i \tag{3.18}$$

$$B_{\tilde{x}y} = \frac{1}{T}\sum_{i=1}^{N} \widetilde{X}_i' ee' y_i \tag{3.19}$$

$$W_{\tilde{x}\tilde{x}} = T_{\tilde{x}\tilde{x}} - B_{\tilde{x}\tilde{x}} \tag{3.20}$$

$$W_{\tilde{x}y} = T_{\tilde{x}y} - B_{\tilde{x}y} \tag{3.21}$$

Bu durumda ise genelleştirilmiş en küçük kareler tahmin edicisi (3.22)'deki gibi ifade edilir:

$$\hat{\beta}_{GLS} = \left[\frac{1}{T}\sum_{i=1}^{N} X_i' Q X_i + \psi \sum_{i=1}^{N} (\overline{x_i} - \overline{x})(\overline{x_i} - \overline{x})'\right]^{-1} \cdot \left[\frac{1}{T}\sum_{i=1}^{N} X_i' Q y_i + \psi \sum_{i=1}^{N} (\overline{x_i} - \overline{x})(\overline{y_i} - \overline{y})\right] \quad (3.22)$$

Genelleştirilmiş en küçük kareler tahmin edicisi parametrize edilerek grup içi ve gruplar arası tahmin edicilerinin ağırlıklandırılmış haline dönüştürülebilir (Hsiao, 2014). Genelleştirilmiş en küçük kareler tahmin edicisinin dezavantajı, hesaplanabilmesi için varyans bileşenlerinin hesaplanması gerekmektedir. Varyans bileşenleri bilinmediğinden ötürü, uygun en küçük kareler yöntemi ile hesaplanır (Detaylar için Hsiao (2014)'nun). Bu tahmin edici örneklem büyüdükçe genelleştirilmiş en küçük kareler tahmin edicisiyle aynı etkinliğe sahip olur.

v_{it} ayrıştırıldığında α_i elde edilir. α_i sabit ya da rassal olmasıyla modeller değişkenlik gösterir. α_i sabit ise sabit etkiler modeli; rassal ise rassal etkiler modeli oluşur. Bu ayrım da N sabit ve T büyükken bir şeyi değiştirmemektedir. Çünkü gölge değişkenler en küçük kareler tahmin edicisi ile genelleştirilmiş en küçük kareler tahmin edicisi aynı olmaktadır. Bunun tam tersi olursa, hangi modelin kullanılacağı önem kazanmaktadır. Bu nedenle, rassal etkiler ile sabit etkiler modeli ayrımı yapılabilmesi için Hausman testinin gerçekleştirilmesi gerekiyor.

Hausman (1978), Hausman testini geliştirirken $y = X\beta + \varepsilon$ modeli üzerinden sırayla Teorem 1, Sonuç Teorem 1 ve Teorem 2'yi ispat ederek yapar:

Teorem 1: β'nın iki normal dağılan etkin tahmin edicisi β_0 ile β_1'i düşünelim. Şayet β_0 asimptotik olarak Cramer-Rao sınırına sahip ise bu durumda T sabit olmak üzere $\sqrt{T}(\beta_0 - \beta)$ ile $\sqrt{T}(\beta_1 - \beta_0)$ aralarındaki kovaryans matrisi 0 matrisidir.

Sonuç Teorem 1: $V(\beta_1 - \beta_0)$ negatif olmayan tanımlı bir matristir.

Teorem 2: Elimizde iki tahmin edici β_0 ile β_1 olsun. H_0 hipotezimiz $\sqrt{T}(\beta_0 - \beta)$ ile $\sqrt{T}(\beta_1 - \beta)$ asimptotik olarak normal dağılmakta ve 0 matris ortalamalı sabit bir kovaryans matrislerine sahip olduğunu belirtirken; H_1 hipotezimiz $\sqrt{T}(\beta_0 - \text{plim}\beta_0)$ ile $\sqrt{T}(\beta_1 - \beta)$ asimptotik olarak normal dağılmakta ve 0 matris ortalamalı ve β'nın fonksiyonu olan bir

kovaryans matrislerine sahip olduğunu belirtmektedir ve bu iki hipotezlerden hangisinin doğru olduğuna $q = \beta_1 - \beta_0$ olmak üzere $m = Tq'\widehat{V}(q)^{-1}q$ test istatistiği kullanılır ve bu da kikare dağılmaktadır.

$\widehat{V}(q)$ β_1 ile β_0'ın fonksiyonu olup gerçek V (q)'nun istikrarlı tahmin edicisi olup m test istatistiği H_1 hipotezi altında asimptotik olarak merkezi olmayan kikare dağılımına sahiptir.

Panel veri analizinde uygulanması önemli olan yöntemlerden birisi de ANCOVA prosedürüdür. Bu prosedürün uygulanmasının temel amacı doğrusal modeller için değişkenliğin varlığını tespit etmektir.

3.2.2 Panel veri analizinde ANCOVA prosedürü

ANCOVA prosedürü uygulanacağı vakit, denetlenmesi gereken temel varsayım (3.1)'deki u_{it}'nin 0 ortalamalı σ_u^2 varyanslı normal dağılıma sahip olmasıdır. Bunun yanında panel veri için ANCOVA prosedürünün hangi koşullar altında uygulandığını öğrenmek için Kuh (1963)'e bakılabilir. Hsiao (2014), panel veri için ANCOVA prosedürü uygulanma durumlarını incelemiştir:

1. Eğim ve sabit terimlerin aynı anda birbirlerine eşit olması: Böyle bir durumda, panel model (3.2)'deki gibi ifade edilmektedir:

$$y_{it} = \alpha + \beta' x_{it} + u_{it} \tag{3.23}$$

2. Eğimlerin eşit olması durumunda, model (3.3)'teki gibi ifade edilmektedir:

$$y_{it} = \alpha_i + \beta' x_{it} + u_{it} \tag{3.24}$$

3. Sabit terimlerin eşit olması durumunda, model (3.4)'teki gibi ifade edilmektedir:

$$y_{it} = \alpha + \beta_i' x_{it} + u_{it} \tag{3.25}$$

Bunların hepsi ayrı bir hipotez halinde değerlendirilir. Böyle olduğunda işlemler için gerekli eşitlikler (3.26), (3.27), (3.28), (3.29), (3.30), (3.31), (3.32), (3.33), (3.34), (3.35), (3.36), (3.37), (3.38), (3.39), (3.40), (3.41) ve (3.42)'deki gibi ifade edilmektedir:

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \tag{3.26}$$

$$\bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it} \tag{3.27}$$

$$\bar{y} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} y_{it} \tag{3.28}$$

$$\bar{x} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it} \tag{3.29}$$

$$W_{xx,i} = \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \tag{3.30}$$

$$W_{xy,i} = \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)' \tag{3.31}$$

$$W_{yy,i} = \sum_{t=1}^{T} (y_{it} - \bar{y}_i)(y_{it} - \bar{y}_i)' \tag{3.32}$$

$$T_{xx,i} = \sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \bar{x})(x_{it} - \bar{x})' \tag{3.33}$$

$$T_{xy,i} = \sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \bar{x})(y_{it} - \bar{y})' \tag{3.34}$$

$$T_{yy,i} = \sum_{i=1}^{N}\sum_{t=1}^{T} (y_{it} - \bar{y})(y_{it} - \bar{y})' \tag{3.35}$$

$$W_{xx} = \sum_{i=1}^{N} W_{xx,i} \tag{3.36}$$

$$W_{xy} = \sum_{i=1}^{N} W_{xy,i} \tag{3.37}$$

$$W_{yy} = \sum_{i=1}^{N} W_{yy,i} \tag{3.38}$$

$$RSS_i = W_{yy,i} - W'_{xy,i} W_{xx,i}^{-1} W_{xy,i} \tag{3.39}$$

$$S_1 = \sum_{i=1}^{N} RSS_i \tag{3.40}$$

$$S_2 = W_{yy} - W'_{xy} W_{xx}^{-1} W_{xy} \tag{3.41}$$

$$S_3 = T_{yy} - T'_{xy} T_{xx}^{-1} T_{xy} \tag{3.42}$$

K, katsayı sayısı olmak üzere ilk hipotez için test istatistiği (3.43)'deki eşitlikle hesaplanır:

$$F_3 = \frac{(S_3 - S_1)/(N-1)(K+1)}{S_1/(NT - N(K+1))} \tag{3.43}$$

Test istatistiği reddedilirse heterojen eğimler mi yoksa heterojen sabit terimler mi olup olmadığının test edilmesi gerekir. Bunun için heterojen eğimlerden başlayarak (3.44)'deki test istatistiği kullanılır.

$$F_1 = \frac{(S_2 - S_1)/[(N-1)K]}{S_1/(NT - N(K+1))} \tag{3.44}$$

Homojen eğim olduğundan yola çıkarak sabit terimlerin eşitliğini test etmek için (3.45)'teki test istatistiği kullanılır:

$$F_4 = \frac{(S_3 - S_2)/(N-1)}{S_2/[N(T-1) - K]} \tag{3.45}$$

ANCOVA ile parametrelerin zaman içerisinde değişip değişmediğini de test edebiliriz. Bu modellerin test edilmesi için (3.46), (3.47), (3.48), (3.49), (3.50), (3.51) ve (3.52)'deki eşitlikler kullanılır:

$$\bar{y_t} = \frac{1}{N}\sum_{i=1}^{N} y_{it} \tag{3.46}$$

$$\bar{x_t} = \frac{1}{N}\sum_{i=1}^{N} x_{it} \tag{3.47}$$

$$W_{xx,t} = \sum_{t=1}^{T} (x_{it} - \bar{x_t})(x_{it} - \bar{x_t})' \tag{3.48}$$

$$W_{xy,t} = \sum_{t=1}^{T} (x_{it} - \bar{x_t})(y_{it} - \bar{y_t})' \tag{3.49}$$

$$W_{yy,t} = \sum_{t=1}^{T} (y_{it} - \bar{y_t})(y_{it} - \bar{y_t})' \tag{3.50}$$

$$S_1' = \sum_{t=1}^{T} W_{yy,t} - W_{xy,t}' W_{xx,t}^{-1} W_{xy,t} \tag{3.51}$$

$$S_2' = \sum_{t=1}^{T} W_{yy,t} - \left(\sum_{t=1}^{T} W_{xy,t}'\right)\left(\sum_{t=1}^{T} W_{xx,t}\right)^{-1}\left(\sum_{t=1}^{T} W_{xy,t}\right) \tag{3.52}$$

İlk işlem olarak bütün parametrelerin homojenliği (3.53)'deki test istatistiği ile test edilir:

$$F_3' = \frac{(S_3 - S_1')/(T-1)(K+1)}{S_1'/(NT - T(K+1))} \tag{3.53}$$

Heterojen sabit terimler (homojen eğimler verildiğinde) için ise (3.54)'deki test istatistiği kullanılır:

$$F_1' = \frac{(S_2' - S_1')/(T-1)K}{S_1'/(NT - T(K+1))} \tag{3.54}$$

Homojen sabit terimler (homojen eğimler verildiğinde) için (3.55)'deki test istatistiği kullanılır:

$$F_4' = \frac{(S_3 - S_2')/(T-1)}{S_2'/(NT - T - K)} \tag{3.55}$$

Bu testler birbirinden bağımsız değildir (Hsiao, 2014). Bu nedenle birbirleriyle çelişen sonuçlar çıkarılabilir. Buna ek olarak başka test etme yöntemleri de bu testlerden yola çıkarak oluşturulmaktadır. Hsiao (2014), bunun için Stock ve Watson (2008)'ı örnek olarak vermiş. Son olarak bu testlerin geçerliliği, artıkların açıklayıcı değişkenlerden bağımsız, birbirinden bağımsız aynı dağılıma sahip olması; açıklayıcı değişkenlerin dışsal değişkenler olması gerekmektedir.

Doğrusal regresyon modellerinde tespit edilmesi güç problemler doğmasına neden olabilecek faktörlerden birisi, otoregresif zaman seriye sahip yanıt değişkenlerdir. Sözkonusu yanıt değişkenleri incelemek için dinamik panel modeli gereklidir.

3.2.3 Dinamik panel modeli

Dinamik panel modelinin özel bir dikkate ihtiyacı vardır. Nedenleri şunlardır:

i. Ekzojen değişkenlerle ilişkili değişkenler çıkınca genelleştirilmiş en küçük karelerle elde edilen tahminler yanlı hale gelmektedir.
ii. Başlangıç değerleri üzerinde yapılan varsayımlar önem kazanmaktadır.

Söz konusu anlatılanlar, ilerleyen konularda açıklanacaktır. Buradan dinamik panel modellerinin parametrelerinin nasıl tahmin edildiği açıklanacaktır.

Hsiao (2014)'nun dinamik rassal etkiler modeli, (3.56)'daki gibi ifade edilebilir:

$$y_{i,t} = \gamma y_{i,t-1} + \beta' x_{it} + v_{it} \tag{3.56}$$

Bu durumda γ'nın en küçük kareler tahmin edicisi (3.57)'deki gibi ifade edilir:

$$\hat{\gamma}_{ekk} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t}y_{i,t-1}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2} \tag{3.57}$$

Bu formül (3.58)'deki gibi ifade edilebilmektedir:

$$\frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t}y_{i,t-1}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2} = \gamma + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} (\alpha_i + u_{it})y_{i,t-1}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2} \tag{3.58}$$

(3.58) için (3.59)'deki formül geçerli olmakta ve bu durumda zaman serisi uzunluğu T arttıkça artan bir yanlılık, model parametrelerinin hesabını etkilemektedir:

$$\begin{aligned} &\operatorname{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\alpha_i + u_{it})y_{i,t-1} \\ &\qquad = \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\operatorname{cov}(y_{i0},\alpha_i) + \frac{1}{T}\frac{\sigma_a^2}{(1-\gamma)^2}[(T-1) - T\gamma + \gamma^T] \end{aligned} \tag{3.59}$$

Hsiao (2014), parametre tahmini yapmak için dinamik modeli (3.60)'daki ve (3.61)'deki gibi formülleştiriyor:

$$y_{it} = \gamma y_{i,t-1} + p'z_i + \beta' x_{it} + v_{it} \quad (3.60)$$

$$w_{it} = \gamma w_{i,t-1} + p'z_i + \beta' x_{it} + u_{it} \quad (3.61)$$

(3.60)'daki formülde z_i zamanla değişmeyen ekzojen değişkenler, x_{it} zamanla değişen ekzojen değişkenler, $|\gamma| < 1$ ve $v_{it} = \alpha_i + u_{it}$ olmaktadır. (3.60)'daki eşitlikte (3.62), (3.63), (3.64), (3.65), (3.66), (3.67)'da ifade edilen varsayımlar geçerlidir:

$$E(\alpha_i) = 0 \quad (3.62)$$

$$E(\alpha_i \mathbf{z_i'}) = E(\alpha_i \mathbf{x_{it}'}) = 0' \quad (3.63)$$

$$E(\alpha_i u_{it}) = 0 \quad (3.64)$$

$$E(\alpha_{it}\alpha_{js}) = \begin{cases} \sigma_a^2 & , i = t \text{ ve } j = s \\ 0 & , \text{d. y.} \end{cases} \quad (3.65)$$

$$E(u_{it}) = 0 \quad (3.66)$$

$$E(u_{it}u_{js}) = \begin{cases} \sigma_u^2 & , i = t \text{ ve } j = s \\ 0 & , \text{d. y.} \end{cases} \quad (3.67)$$

(3.61)'deki formülde $y_{it} = w_{it} + \eta_i$,z_i zamanla değişmeyen ekzojen değişkenler, x_{it} zamanla değişen ekzojen değişkenler, $|\gamma| < 1$ ve $v_{it} = \alpha_i + u_{it}$ olmaktadır. (3.64)'deki ifadede $\alpha_i = (1-\gamma)\eta_i$ şeklinde temsil edilmekte ve η_i 0 ortalamalı ve $\sigma_a^2/(1-\gamma^2)$ varyanslı dağılımdan gelmektedir. Hsiao (2014), bu modelin parametrelerini tahmin edebilmek için başlangıç değerleri y_{i0}'lar için belli durumları ele almıştır:

Durum 1: y_{i0} sabit

Durum 2: y_{i0} rassal bir değişken ve bu durumda alt başlıkları var:(bu durumda $y_{i0} = \mu_{y0} + \epsilon_i$ olarak ele alınmıştır ve sonlu bir ortalamaya (μ_{y0}) ve varyansa (σ_{y0}^2) sahiptir. Ayrıca modelde ϵ_i ortalamadan düşüldüğünde başlangıçtaki bireysel farklılıkları ifade etmekle işlev görür)

Durum 2a: y_{i0} rassal bir değişken ve α_i'den bağımsız (başka bir şekilde söylenecek olursa $\mathrm{cov}(\epsilon_i, \alpha_i) = 0$). Hsiao (2014)'ya göre böyle bir durumda ϵ_i 'lerin etkisi modelde düşmesi bekleniyor.

Durum 2b: y_{i0} rassal bir değişken ve α_i ile ilişkili. Böyle bir durumda ϵ_i 'lerin etkisi, α_i'lerle olan ilişkisiyle birlikte iş görür hale gelmiştir. $\epsilon_i = \alpha_i$ olduğu durumda bireysel etkiler zaman boyunca başlangıç değerleriyle ifade edilir hale gelmiştir. Bunun yanında $\frac{\alpha_i}{1-\gamma} = \eta_i$ olarak ifade edilmektedir. Buradan w_{it} ile alakalı durumlara geçilmektedir.

Durum 3: w_{i0} sabit

Durum 4: w_{i0} rassal

Durum 4a: w_{i0} rassal ve μ_w değerli ortak bir ortalama ile $\frac{\sigma_u^2}{(1-\gamma^2)}$ değerli varyansa sahiptir.

Durum 4b: w_{i0} rassal ve μ_w değerli ortak bir ortalama ile σ_{w0}^2 değerli varyansa sahiptir.

Durum 4c: w_{i0} rassal ve θ_{i0} değerli ortalama ile $\frac{\sigma_u^2}{(1-\gamma^2)}$ değerli varyansa sahiptir.

Durum 4d: w_{i0} rassal ve θ_{i0} değerli ortalama ile σ_{w0}^2 değerli varyansa sahiptir.

Durum 1, durum 2a, durum 2b, durum 3 ve durum 4a için olabilirlik fonksiyonları sırayla (3.68), (3.69), (3.70), (3.71) ve (3.72)'deki gibi ifade edilmiştir:

$$L_1 = (2\pi)^{-\frac{NT}{2}}|V|^{-\frac{N}{2}}\exp\left\{\frac{-1}{2}\sum_{i=1}^{N}\left(y_i - \gamma y_{i,-1} - Z_i\mathbf{p} - X_i\boldsymbol{\beta}\right)'V^{-1}\left(\mathbf{y}_i - \gamma\mathbf{y}_{i,-1} - Z_i\mathbf{p} - X_i\boldsymbol{\beta}\right)\right\} \tag{3.68}$$

$$L_{2a} = L_1 x(2\pi)^{-\frac{N}{2}}\left(\sigma_{y0}^2\right)^{-\frac{N}{2}}\exp\left\{\frac{-1}{2\sigma_{y0}^2}\sum_{i=1}^{N}\left(y_{i0} - \mu_{y0}\right)^2\right\} \tag{3.69}$$

$$L_{2b} = (2\pi)^{-\frac{NT}{2}}(\sigma_u^2)^{-\frac{N(T-1)}{2}}\Big(\sigma_u^2 + T\big(\sigma_a^2 - \phi^2\sigma_{y0}^2\big)\Big)^{-\frac{N}{2}} \exp\left\{\frac{-1}{2\sigma_u^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\big[y_{it} - \gamma y_{i,t-1} - \mathbf{p}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it} - \phi\big(y_{i0} - \mu_{y0}\big)\big]^2 + \frac{\big(\sigma_a^2 - \phi^2\sigma_{y0}^2\big)}{2\sigma_u^2\Big(\sigma_u^2 + T\big(\sigma_a^2 - \phi^2\sigma_{y0}^2\big)\Big)}\sum_{i=1}^{N}\left\{\sum_{t=1}^{T}\Big(y_{it} - \gamma y_{i,t-1} - \mathbf{p}'z_i - \boldsymbol{\beta}'\mathbf{x}_{it} - \phi\big(y_{i0} - \mu_{y0}\big)\Big)\right\}^2\right\}.(2\pi)^{-\frac{N}{2}}\big(\sigma_{y0}^2\big)^{-\frac{N}{2}}\exp\left\{\frac{-1}{2\sigma_{y0}^2}\sum_{i=1}^{N}\big(y_{i0} - \mu_{y0}\big)^2\right\} \tag{3.70}$$

$$L_3 = (2\pi)^{-\frac{NT}{2}}(\sigma_u^2)^{-\frac{NT}{2}}\exp\left\{\frac{-1}{2\sigma_u^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\big[(y_{it} - y_{i0} + w_{i0}) - \gamma\big(y_{i,t-1} - y_{i0} + w_{i0}\big) - \mathbf{p}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it}\big]^2\right\}.(2\pi)^{-\frac{N}{2}}\big(\sigma_\eta^2\big)^{-\frac{N}{2}}\exp\left\{\frac{-1}{2\sigma_\eta^2}\sum_{i=1}^{N}(y_{i0} - w_{i0})^2\right\} \tag{3.71}$$

$$L_{4a} = (2\pi)^{-\frac{N(T+1)}{2}}|\Omega|^{-\frac{N}{2}}\exp\left\{\frac{-1}{2}\sum_{i=1}^{N}\big(y_{i0} - \mu_w, y_{i1} - \gamma y_{i,0} - \mathbf{p}'\mathbf{z}_i - \boldsymbol{\beta}'x_{i1}, \dots, y_{iT} - \gamma y_{i,T-1} - \mathbf{p}'\mathbf{z}_i - \boldsymbol{\beta}'x_{iT}\big)\Omega^{-1}\big(y_{i0} - \mu_w, y_{i1} - \gamma y_{i,0} - \mathbf{p}'\mathbf{z}_i - \boldsymbol{\beta}'x_{i1}, \dots, y_{iT} - \gamma y_{i,T-1} - \mathbf{p}'\mathbf{z}_i - \boldsymbol{\beta}'x_{iT}\big)'\right\} \tag{3.72}$$

L_{4a} için Ω formülü (3.73)'deki gibidir:

$$\Omega_{(T+1)x(T+1)} = \sigma_u^2\begin{bmatrix}\frac{1}{1-\gamma^2} & 0\\ 0 & I_t\end{bmatrix} + \sigma_a^2\begin{bmatrix}\frac{1}{1-\gamma}\\ \mathbf{e}\end{bmatrix}\begin{bmatrix}\frac{1}{1-\gamma}\\ \mathbf{e}\end{bmatrix}' \tag{3.73}$$

Burada **e** birlerden oluşan bir Tx1 matrisi temsil etmektedir. Durum 4b için olabilirlik fonksiyonu aynı formda olup Ω formülündeki toplamaya giren ilk elemanda $\frac{1}{1-\gamma^2}$ yerine σ_{w0}^2/σ_u^2 olacak. Durum 4c için olabilirlik fonksiyonunda sadece ortalamalar değişecek yani μ_w'lar θ_{i0} olacak. Durum 4d için olabilirlik fonksiyonu durum 2b'ye benzeyecek ordaki μ_{y0}, ϕ ve σ_0^2 terimleri θ_{i0} , $(1-\gamma)\sigma_\eta^2/(\sigma_\eta^2+\sigma_{w0}^2)$, $(\sigma_\eta^2+\sigma_{w0}^2)$ terimlerle değişecek. Maalesef durum 3 ve durum 4d'de olabilirlik fonksiyonlarında içerisindeki exp(.) fonksiyonunun içerisindeki eleman, σ_η^2 ve $(\sigma_\eta^2+\sigma_{w0}^2)$ 0'a yaklaştığında fonksiyonlar sınırlandırılamamaktadır. Bunu çözmek için kısmi diferansiyel denklemler kullanılmaktadır. Bunun yanında Hsiao (2014) birim sayısı N'nin ve zaman serisi uzunluğu T'nin artışı olduğu durumda, en çok olabilirlik tahminlerinin gerçek değerlere yaklaşmasında problemler olduğunu ifade etmiştir. Anderson ve Hsiao (1982) N'yi arttırdığında durum 3, durum 4c ve durum 4d hariç bütün durumlarda tahmin edicileri gerçek değerlere yaklaştırmayı başarmıştır. Durum 4c ve Durum 4d için Bhargava ve Sargan (1983) bir çözüm önermiş ve durum 4c ve durum 4d'yi dönüştürmek için x_{it} için (3.74)'deki modeli varsaymıştır:

$$x_{it} = c + \sum_{j=0}^{\infty} b_j \xi_{i,t-j} \tag{3.74}$$

Burada $\xi_{i,t-j}$ birbirinden bağımsız, aynı dağılıma sahip en küçük kareler ortalamalı tahmin ediciler olarak temsil edilmiştir. Böylelikle başlangıç değer y_{i0} ile ilgili (3.75)'deki modeli varsaymışlardır:

$$y_{i0} = \sum_{t=1}^{T} \pi_{0t}' x_{it} + \mathbf{p}^{*\prime} z_i + v_{i0} \tag{3.75}$$

Burada $v_{i0} = \epsilon_{i0} + u_{i0}^* + \eta_i$ olmaktadır. ϵ_{i0} θ_{i0}'ın tahmin edilme hatasını, u_{i0}^* başlangıç noktasından önce biriken şokların etkisi ve son olarak η_i bireysel etkileri göstermektedir. Durum 4c ile durum 4d böyle bir dönüşümden geçince N veya T'nin biri ya da ikisi de artsa parametreler gerçek değerlerine erişirler. Hsiao (2014), aynı zamanda durum 3, durum 4c ve

durum 4d hariç bütün durumlarda Genelleştirilmiş En Küçük Kareler (GEKK) tahmin edicisinin Ω, σ_u^2, $\sigma_a^2, \sigma_{y0}^2$ ve ϕ'ye bağlı olarak en çok olabilirlik tahmin edicisine eşdeğer olduğunu ifade etmiştir. $\delta' = (\pi', p^{*\prime}, \gamma, \beta', p')$ için tahmin edici (3.76)'daki gibi ifade edilmektedir:

$$\hat{\delta}_{gekk} = \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{V}^{-1} \tilde{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \tilde{X}_i' \tilde{V}^{-1} \tilde{y}_i\right) \tag{3.76}$$

V matrislerini hesaplamak için Uygun Genelleştirilmiş En Küçük Kareler (UGEKK) prosedürü kullanılabilir (Hsiao, 2014). UGEKK prosedürü, GEKK'ya göre asimptotik olarak daha az etkindir. Ayrıca $\mathrm{Cov}(y_{i0}, \alpha_i) \neq 0$ olduğunda, T sabit kaldığında ve N arttıkça GEKK tahmin edicileri gerçek değerlerine erişmiyor.

Araçsal değişken tahmin edicisi ile V matrisi hesaplanabilir. Panel veri modelinden yola çıkılarak şu adımlar izlenir:

Panel veri modeli (3.77)'deki gibi ifade edilir:

$$y_{i,t} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + \beta'(x_{it} - x_{i,t-1}) + u_{it} - u_{i,t-1} \tag{3.77}$$

(3.77)'ye dönüştürdüğümüz zaman, $u_{it} - u_{i,t-1}$ ifadesi $y_{i,t-2}$ ile ya da $y_{i,t-2} - y_{i,t-3}$ ile ilişkisi kalmamaktadır. β ile γ'nın tahmin edicileri (3.78)'deki gibi hesaplanabilir:

$$\begin{pmatrix} \hat{\gamma}_{iv} \\ \hat{\beta}_{iv} \end{pmatrix} = \left[\sum_{i=1}^{N} \sum_{t=3}^{T} \begin{pmatrix} (y_{i,t-1} - y_{i,t-2})(y_{i,t-2} - y_{i,t-3}) & (y_{i,t-2} - y_{i,t-3})(\mathbf{x}_{i,t} - \mathbf{x}_{i,t-1})' \\ (\mathbf{x}_{i,t} - \mathbf{x}_{i,t-1})(y_{i,t-1} - y_{i,t-2}) & (\mathbf{x}_{i,t} - \mathbf{x}_{i,t-1})(\mathbf{x}_{i,t} - \mathbf{x}_{i,t-1})' \end{pmatrix}\right]^{-1}$$

$$\cdot\left[\sum_{i=1}^{N}\sum_{t=3}^{T}\begin{pmatrix}(y_{i,t-2}-y_{i,t-3})\\(x_{i,t}-x_{i,t-1})\end{pmatrix}\left((y_{i,t}-y_{i,t-1})\right)\right] \quad (3.78)$$

$q_{it}=\left[(y_{i,t-2}-y_{i,t-3}),(x_{i,t}-x_{i,t-1})'\right]$ olmak üzere yukarıdaki formüller $\frac{1}{N\,(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}q_{it}(u_{it}-u_{i,t-1})=0$'dan çıkarılabilir (Hsiao, 2014). Bu noktada ikinci aşamaya geçilir. Bu sefer En Küçük Kareler (EKK) yöntemiyle **p** matrisi (3.79)'daki gibi hesaplanır.

$$\bar{y}_ı-\gamma\overline{y_{ı,-1}}-\beta'\bar{x}_i=\mathbf{p}'z_{\mathbf{i}}+\alpha_i+\bar{u}_i \quad (3.79)$$

Burada $\bar{y}_ı=\frac{\sum_{t=1}^{T}y_{i,t}}{T}$, $\overline{y_{ı,-1}}=\frac{\sum_{t=1}^{T}y_{i,t-1}}{T}$, $\bar{u}_ı=\frac{\sum_{t=1}^{T}u_{i,t}}{T}$ ve $\bar{x}_ı=\frac{\sum_{t=1}^{T}x_{i,t}}{T}$ olarak temsil edilmektedir.

σ_u^2 ile σ_a^2 (3.80) ve (3.81)'deki eşitliklerle tahmin edilir:

$$\hat{\sigma}_u^2=\frac{\sum_{i=1}^{N}\sum_{t=2}^{T}\left[y_{i,t}-y_{i,t-1}=\hat{\gamma}(y_{i,t-1}-y_{i,t-2})+\hat{\beta}'(x_{it}-x_{i,t-1})\right]^2}{2N(T-1)} \quad (3.80)$$

$$\hat{\sigma}_a^2=\frac{\sum_{i=1}^{N}\left(\bar{y}_ı-\hat{\gamma}\overline{y_{ı,-1}}-\hat{\beta}'\bar{x}_i-\hat{\mathbf{p}}'z_{\mathbf{i}}\right)}{N}-\frac{\hat{\sigma}_u^2}{T} \quad (3.81)$$

Bu tahmin ediciler başlangıç koşullarından bağımsız olarak gerçek değerlere yakınsarlar. Aynı durum **p** tahmin edicisi için de geçerlidir.

Bu noktaya kadar anlatılan model türleri doğrusal ilişki varsayılarak gerçekleştirilen doğrusal modellerdi. Bu model türleri dışında alternatif modeller incelenecektir.

3.2.4 Panel veri için kullanılabilecek alternatif model türleri

Panel veri toplanırken zaman zaman, tekrar eden birimli veriler, sürece bağlı kategorik olarak işlenmiş veriler, sayımlardan oluşan veriler, parametrik olmayan yöntemlerle analizi yapılabilecek veriler, beklenmedik bir olay yüzünden toplanması kesintiye uğrayan veriler ve yüzdelik regresyona ihtiyaç duyan veriler toplanabiliyor. Bu tip verilerin analizinin yapılması için gereken teorik çalışmalar, panel veri analizi literatüründe mevcuttur.

Bu anlatılanlar ilerleyen başlıklarda incelenecektir.

3.2.4.1 Tekrar eden birimler üzerinden panel veri analizi

Panel veri elde edilirken, birimlerin takip edilememesi bir soruna yol açmaktadır. Gözlemleyemediğimiz bireyler, bireysel etkiyi belirlemede sorun çıkarır. Bu nedenle bu sorunu halletmek için özel bir panel veri analizi gerçekleştirmek gerekir. Deaton (1985) bu soruna yanıt ararken cohort'lara dayalı bir yaklaşım önermiştir. Panel veri modelinin (3.81)'deki eşitlikteki gibi ortaya konduğu varsayılsın:

$$Y_{it} = X_{it}'\beta + \alpha_i + u_{it} \tag{3.82}$$

Bu noktada α_i ile X_{it} ilişkili olmadığında, tekrarlı panel verisinden elde edilen tahmin edici hiçbir şekilde gerçek değere yakınsamama sorunu yaşamaz (Hsiao, 2014). Deaton (1985) α_i'ler için (3.83)'deki eşitliği varsaymıştır:

$$\alpha_i = \sum_{c=1}^{C} \alpha_c d_{ic} \tag{3.83}$$

Burada d_{ic} c. kohorta ait olduğunda 1 olmadığında 0 değerini alan bir gölge değişkendir. Bu cohortlar modele tanımlandıktan sonra panel modelini (3.84)'deki gibi ifade edebiliriz:

$$\overline{Y}_{ct} = \overline{X}'_{ct}\beta + \alpha_c + \bar{u}_{ct} \tag{3.84}$$

Burada $\overline{Y}_{ct}$, $\overline{X}'_{ct}$ ve $\bar{u}_{ct}$ hepsi birimlere ait kohort ortalamalarını ifade eder. Şayet α_i ile X_{it} ilişkili olmazsa ve $N \to \infty$, $C \to \infty$, $C/N \to 0$ olursa, panel içi tahmin edici gerçek değerine yakınsar. Bu yöntemin de maalesef sıkıntıları bulunmaktadır. Bunlar şu şekilde ifade edilebilir:

- Kohortlar içinde homojenlik olmasına rağmen, kohortların seçimi keyfidir.
- Kohort seçiminin bir sınırı olabilir.
- Veriden elde edilecek bilgi kaybı, bu kohort seçimini sınırlandırır.

3.2.4.2 Süreç modelleri

Kategorik işlenmiş bir veride, bir durumdan bir başka duruma geçişin zamanını inceleyen modellere süreç modelleri denir. μ_{it} sabit kaldığı varsayıldığında A_{its}, t. zaman ile t+s. zaman arasında bir olay gerçekleşmemesini ifade etsin. Bunun yanında t. zaman ile $t + \Delta$. zamanda olay gerçekleşme değişkeni $A_{it,t+\Delta}$ ihtimali (3.85)'deki gibi verilmiş olsun (Hsiao, 2014):

$$P\left(A_{it,t+\Delta}\right) = \mu_{it}\Delta t \tag{3.85}$$

Burada μ_{it} t. zaman ile t+s. zaman arasında sabit kaldığı varsayılırsa, bu durumda t ile t+s. arası zaman aralığını $\frac{s}{\Delta t}$ tane noktaya bölünür. Sonrasında her birinin olasılığını hesaplanır. Sözkonusu olasılıklar, (3.86)'daki gibi elde edilir:

$$P(A_{its}) = (1 - \mu_{it}\Delta t)^{\frac{s}{\Delta t}} P\left(A_{it,t+\Delta}\right) \tag{3.86}$$

(3.85)'in bu şekilde ifade edilmesinin nedeni, t. zaman ile t+s. zaman arasında bir olay gerçekleşmemesi demek t. zamandan sonraki ilk olayın t+s. zamanda gerçekleşmesi demektir. Elimizde $[t, t+s)$ aralığında gerçekleşmeyen olaylar var ve t+s. zaman sonunda bir olay gerçekleşiyor. Bu nedenle $P(A_{its})$, (3.87)'deki gibi ifade edilir:

$$P(A_{its}) = (1 - \mu_{it}\Delta)^{\frac{s}{\Delta}} \mu_{it}\Delta \tag{3.87}$$

Burada $\Delta \to 0$ olunca bu durumda $a = \frac{1}{\Delta} \to \infty$ olmaktadır. D_i(i. birim için bekleme süresi)'ye ait olasılık ile $A_{it,t+\Delta}$ arasındaki ilişkinin formülü, (3.88)'deki gibi elde edilir:

$$P\left(A_{it,t+\Delta}\right)\left(\lim_{\Delta \to 0} P(A_{its})\right) = \exp(-\mu_{it}s)\mu_{it}\Delta = P(D_i = t)\Delta \tag{3.88}$$

Burada t. zaman ile $t + \Delta$. zaman'da gerçekleşen olaylarla ilgilenmemiz gerekmediği için limit alınan kısım ile ayırmamız gerekir. Bu noktada $P(D_i = t) = f_i(t)$ olarak ifade edilirse, (3.89) ve (3.90)'daki eşitliğe ulaşılır:

$$P(D_i < t) = \int_0^t f_i(s)ds = 1 - \exp(-\mu_{it}t) \tag{3.89}$$

$$S_i(t) = P(D_i \geq t) = \exp(-\mu_{it}t) \tag{3.90}$$

Şayet $\mu_{it} = \mu_i$ olarak kabul edilirse, en sonunda $\mu_i = -\frac{d\log(S_i(t))}{dt}$ için bir denklem oluşturulabilir ve (3.91)'deki ifade ile regresyon modeli ifade edilir:

$$\mu_i = \exp(x_i'\beta) \tag{3.91}$$

Burada Hsiao (2014) μ_i ya da μ_{it} için (3.93)'ü oluşturmanın süreç değişkenine negatif değer atfetmemek açısından faydalı olduğunu söylüyor. Alternatif bir model olarak t_i uzunlukta süreç değişkenleri için (3.92)'deki eşitlik ortaya konulur:

$$E\big(\log(t_i)\big) = \int_0^\infty (\log t) f_i(t) dt = \mu_i \int_0^\infty (\log t) \exp(-\mu_i t) dt = -c - \log(\mu_i) \tag{3.92}$$

Burada c, Euler sabiti olup yaklaşık olarak -0,577'dir ve $\log(t_i)$ için regresyon modeli kurulduğunda β için tahmin edici oluşturmak zorlaşıyor çünkü en küçük kareler tahmin edicisinin kovaryans matrisi orjinal denklemdeki β için en çok olabilirlik tahmin edicisinin kovaryans matrisinden daha büyüktür (Hsiao, 2014). Bu nedenle Cox (1972), μ_{it} için (3.93) eşitliğini varsaymıştır:

$$\mu_{it} = \mu(t) g(x_i) \tag{3.93}$$

Burada $\mu(t)$ t'ye dayalı bir fonksiyonu ve $g(x_i)$ x_i'ye dayalı bir fonksiyonnu temsil etmektedir. Daha önce benzeri yapıldığı gibi $g(x_i) = \exp(x_i'\beta)$ alınabilir. Cox (1972), log-olabilirlik fonksiyonunu (3.94)'deki gibi bulmuştur:

$$L(\beta) = \left(\prod_{i=1}^{n} \exp(x_i'\beta) \mu(t_i) \right) \left(\exp\left[- \int_0^\infty \left[\sum_{h \in R(t_l)} \exp(x_h'\beta) \right] \mu(t) dt \right] \right) \tag{3.94}$$

Burada $R(t_l) = \{i | S_i(t) \geq t_l\}$ ve n örneklem hacmi olmaktadır. Şayet (3.95) ve (3.96) tanımlanırsa, $L(\beta) = L_1(\beta) L_2(\beta)$ olmaktadır:

$$L_1(\beta) = \prod_{i=1}^{n} \frac{\exp(x_i'\beta)}{\sum_{h\in R(t_l)} \exp(x_h'\beta)} \tag{3.95}$$

$$L_2(\beta) = \sum_{i=1}^{n} \left[\sum_{h\in R(t_l)} \exp(x_h'\beta)\mu(t_i) \right] . \exp\left\{ - \int_0^{\infty} \left[\sum_{h\in R(t_l)} \exp(x_h'\beta) \right] \mu(s)ds \right\} \tag{3.96}$$

Cox (1975), $L_1(\beta)$ fonksiyonu burada maksimize edilerek kısmi en çok olabilirlik fonksiyonu elde edilebilir (Hsiao, 2014). Tsiatis (1981), kısmi en çok olabilirlik fonksiyonunun hem istikrarlı hem de asimptotik olarak normal dağılıma yakınsadığı belirtir. Bunun yanında Tsiatis (1981), kovaryans matrisini (3.97)'deki şekilde ifade etmiştir:

$$\mathrm{Cov}(\widehat{\beta_p}) = -\left[E\left(\frac{\partial^2 \log(L_1)}{\partial\beta\, \partial\beta'} \right) \right]^{-1} \tag{3.97}$$

3.2.4.3 Sayım verisi modelleri

Sayım veri modeli, belirli bir zaman aralığında gerçekleşen olaylarla ilgilenir ve dolayısıyla süreç modelleriyle ikili ilişki içerisindedir (Hsiao, 2014). Bu model için log-olabilirlik fonksiyonu (3.98)'deki gibi ifade edilir:

$$\log L = \sum_{i=1}^{N} \sum_{t=1}^{T} [y_{it} \log(\mu_{it}) - \mu_{it} - \log(y_{it})] \tag{3.98}$$

Burada $\mu_{it} = \exp[x_{it}'\beta + \alpha_i]$ olmaktadır. α_i rassal kabul edilirse, $(y_{i1}, \dots, y_{iT})$ dağılımı (3.99)'daki gibi bulunur:

$$f(y_{i1}, \dots, y_{iT}) = \int \prod_{t=1}^{T} \left[\frac{(\mu_{it})^{y_{it}} e^{-\mu_{it}}}{y_{it}!} \right] g(\alpha) d_{\alpha} \tag{3.99}$$

β'nın en çok olabilirlik tahmin edicisi istikrarlı ve asimptotik olarak normal dağılır. Burada N veya T'den herhangi birisi veya ikisi de arttıkça, bu durum geçerlidir. Ancak bazı durumlarda hesaplama yapmak zordur. Bu durumlardan birisi $g(\alpha)$ gama dağıldığı zamanlardan biridir. Bu durumu basitleştirmek adına $\mu_{it} = \alpha_i \exp[x'_{it}\beta]$ olarak ele alınabilir ve y_{it}'nin önceki değer ve değerlerle ilişkisini gözardı edebiliriz. Böyle bir durumda y_{it}'nin dağılımı, (3.100)'deki gibi elde edilir:

$$f(y_i) = \frac{[\exp(x'_i\beta)]^{y_i} \Gamma(y_i + v)}{y_i! \, \Gamma(v)} \left(\frac{1}{\exp(x'_i\beta) + v} \right)^{y_i + v} \tag{3.100}$$

Bu denklemden faydalanılarak Log-olabilirlik fonksiyonu hesaplanır. $\frac{\partial \log L}{\partial \eta_i} = 0$ ve $\frac{\partial \log L}{\partial \beta} = 0$ denklemleri çözülür. Bu denklemlerde $\eta_i = \exp(\alpha_i)$ veya $\eta_i = \alpha_i$ olarak alınabilmektedir. Her iki durumda da $\widehat{\eta_i} = \frac{\overline{y_i}}{\overline{\mu_i}}$ olarak bulunur. Burada $\overline{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ olarak ifade edilirken, $\overline{\mu_i} = \frac{1}{T}\sum_{t=1}^{T} \exp(x'_{it}\beta)$ olarak ifade edilir. $\hat{\beta}$'yı elde etmek için, (3.101)'deki eşitlik çözülür:

$$\sum_{i=1}^{N} \sum_{t=1}^{T} \left[y_{it} - \frac{\overline{y_i}}{\overline{\mu_i}} \exp(x'_{it}\beta) \right] x_{it} = 0 \tag{3.101}$$

3.2.4.4 Panel yüzdelik regresyonu

Yüzdelik regresyon yöntemi, yanıt değişkenin belli yüzdelik dilimdeki değerlerinin açıklayıcı değişkenlerle ilişkisini ortaya koymak için yapılan bir regresyon yöntemidir. Yanıt değişken için t. yüzdelik olarak y_t kullanılırsa (3.102)'deki gibi basitçe ifade ederiz:

$$P(y \leq y_t) = \int_{-\infty}^{y_t} f(y)dy = t \tag{3.102}$$

Panel regresyon modeli için amaç fonksiyonu (3.103)'deki gibi ifade edilir (Hsiao, 2014):

$$f(c) = \sum_{i \in \Psi_c} \tau|y_i - c| + \sum_{i \in \overline{\Psi}_c} (1-\tau)|y_i - c| \tag{3.103}$$

Burada $\Psi_c = \{i|y_i \geq c\}$ ve $\overline{\Psi}_c = \{i|y_i \geq c\}$ olmaktadır. N büyüdükçe $f(c)/N$ (3.104)'e yakınsamaktadır:

$$S(c) = (1-\tau)\int_{-\infty}^{c} |y-c|f(y)dy + \tau\int_{c}^{\infty} |y-c|f(y)dy \tag{3.104}$$

S(c) değeri, doğal olarak $S(y_t)$ değerinden büyüktür. Bu büyüklük, (3.105)'deki eşitsizlik ile ispatlanır:

$$S(c) = (1-\tau)\int_{-\infty}^{c}|y-c|f(y)dy + \tau\int_{c}^{y_t}|y-c|f(y)dy + \tau\int_{y_t}^{\infty}|y-c|f(y)dy$$
$$= S(y_t) + |y_t - c|\big(\tau - F(c)\big) - \int_{c}^{y_t}|y-y_t|f(y)dy \geq S(y_t) \qquad (3.105)$$

Böylelikle c tahmin edicisi, gerçek parametre y_t için bir tahmin edicidir. Panel yüzdelik modeli için yüzdelik tahmin edici, (3.106)'daki fonksiyon küçültülerek elde edilir:

$$f\big(b(\tau), \alpha_i(\tau)\big) = \sum_{i=1}^{N}\sum_{t=1}^{T}\rho_\tau\big(y_{it} - x_{it}'b(\tau) - \alpha_i(\tau)\big) \qquad (3.106)$$

$\rho_\tau(.)$ fonksiyonu burada (3.107)'deki gibi tanımlanır (Hsiao, 2014):

$$\rho_\tau(u) = [\tau - 1(u \leq 0)]u \qquad (3.107)$$

Burada u, klasik panel regresyon modelindeki hata terimi ve $1(u \leq 0)$ fonksiyonu $u \leq 0$ koşulu sağlandığında 1 değerini; sağlanmadığında 0 değerini verir. Belli koşullar altında $b(\tau)$ ile $\alpha_i(\tau)$ tahmin değerleri istikrarlı ve asimptotik olarak normal dağılır. Bunun için $\frac{N^2(\log N)^3}{T} \rightarrow 0$ ve T büyük olmalıdır.

3.2.4.5 Çok seviyeli panel veride regresyon yöntemleri

Verilerin hiyerarşik olarak seviyeli olması bazı durumlarda rastlanabilecek bir olgudur. Örneğin Eskişehir'de ilçelerdeki parsellere göre baz istasyonlarında sinyal seviyesi 3 seviyeli olarak ifade edilir. Burada Eskişehir en üst seviye olurken, ilçe ve parsel bir alt seviyeler olmaktadır. Seviye sayısı yapay olarak arttırılabilir. Örneğin, ülke bir seviye olarak devreye girebilir. Bu nedenle bu konuda analiz adına bir şey söylemek gerekmektedir.

Wansbeek ve Kapteyn (1978), 4 seviyeli veride regresyon modelini (3.108)'deki gibi tanımlamışlardır (Hsiao, 2014):

$$y_{ijlt} = x'_{ijlt}\beta + v_{ijlt} \tag{3.108}$$

Burada v_{ijlt} için (3.109)'daki gibi bir eşitlik varsayılmaktadır:

$$v_{ijlt} = \alpha_i + \lambda_{ij} + v_{ijl} + \epsilon_{ijlt} \tag{3.109}$$

Burada $\alpha_i, \lambda_{ij}, v_{ijl}, \epsilon_{ijlt}$ terimleri için standart sapmalar sırayla $\sigma_a, \sigma_\lambda, \sigma_v, \sigma_e$ olmak üzere (3.110), (3.111) ve (3.112)'deki eşitlikler tanımlanır:

$$\sigma_1^2 = T\sigma_v^2 + \sigma_\epsilon^2 \tag{3.110}$$

$$\sigma_2^2 = LT\sigma_\lambda^2 + T\sigma_v^2 + \sigma_\epsilon^2 \tag{3.111}$$

$$\sigma_3^2 = MLT\sigma_a^2 + LT\sigma_\lambda^2 + T\sigma_v^2 + \sigma_\epsilon^2 \tag{3.112}$$

Burada M,L,T değerleri her bir seviyenin alabileceği toplam değeri ifade eder. (3.111)'daki ifade için genelleştirilmiş denklem y^*_{ijlt} ve x^*_{ijlt} üzerinden (3.113) ve (3.114) ile bulunur (Hsiao, 2014):

$$y^*_{ijlt} = y_{ijlt} - \left(1 - \frac{\sigma_e}{\sigma_1}\right)\overline{y_{ijl.}} - \left(\frac{\sigma_e}{\sigma_1} - \frac{\sigma_e}{\sigma_2}\right)\overline{y_{ij..}} - \left(\frac{\sigma_e}{\sigma_2} - \frac{\sigma_e}{\sigma_3}\right)\overline{y_{i...}} \tag{3.113}$$

$$x^*_{ijlt} = x_{ijlt} - \left(1 - \frac{\sigma_e}{\sigma_1}\right)\overline{x_{ijl.}} - \left(\frac{\sigma_e}{\sigma_1} - \frac{\sigma_e}{\sigma_2}\right)\overline{x_{ij..}} - \left(\frac{\sigma_e}{\sigma_2} - \frac{\sigma_e}{\sigma_3}\right)\overline{x_{i...}} \tag{3.114}$$

Burada ū, u'nun ortalamasını gösterir ve bütün ortalamalar seviyeler üzerinden tanımlanır.

3.2.4.6 Parametrik olmayan panel veri analiz yöntemleri

Parametrik olmayan yöntemler, açıklayıcı değişkenlerle yanıt değişkenin ilişkisini başka şekilde ele almak için ortaya konur. Temel olarak (3.115)'deki gibi ele alınır:

$$y_{it} = m(x_{it}) + v_{it} \tag{3.115}$$

Burada $v_{it} = \alpha_i + u_{it}$ olarak geçmektedir. Bunun yanında, Ai ve Li (2008), alternatif olarak (3.116)'yı oluşturmuşlardır:

$$y_{it} = v_0(x_{it}, \theta_0) + \sum_{j=1}^{m} h_{j0}\left(v_j(x_{it}, \theta_0)\right) + \alpha_i + u_{it} \tag{3.116}$$

Bu modelde birbirinden ayırt edilmesi gereken fonksiyonlar ve başka parametreler bulunmaktadır. Örneğin $\left(h_{j0}(.), \alpha_i\right)$ ikililerinin ayırt edilebilir olması için $h_{j0}(0) = 0$ alınır. $\left(\theta_0, h_0(.)\right)$ ikililerinin ayrılması için $\theta_0 = (1, \theta_{20}, \dots, \theta_{k0})'$ olmak üzere $\theta_0'\theta_0 = 1$ alınır. Bunun haricinde x_{it} kategorileri arasında gizli ortaklık durumu varsa, (Örneğin, x_{1it} ile x_{2it}'nin ortak olarak x_{3it}'nin bir fonksiyonu olarak ifade edilmesinde olduğu gibi) bu durumda model etkileri gene ayırtedilemez olur. Böylece varsayımlar tamamlanır ve model analizine geçilebilir. Model analizinde öncelikli olarak ilgilenmemiz gereken sorun, $h_{j0}(.)$ parametrelerinin sonsuz sayıda olması durumudur (Bu gerçekçi gelmeyebilir ama parametreler hakkında bir şey bilmediğimiz zaman, analizde yer alacak bir fonksiyonun parametrelerinin sonlu olması modelinin genellenebilirliğine ters düşmektedir). Bu durumda bunu tahmin edecek veri bulunamamaktadır. Bunu aşmak için Ai ve Li(2008) (3.116)'daki gibi ifade edilen fonksiyonları kullanmayı önerir:

$$h_0(a) = \rho_j'(a)\pi_j \tag{3.117}$$

Bu tip fonksiyonlarda a değerlerinin ayarlanması gerekir. Çünkü bunu yapmazsak sapan değerler (3.117)'daki ifadeyi etkilemektedir (Ai ve Li, 2008). Bu nedenle a değerlerini oluşturmak için (3.118)'deki B_r tipi fonksiyonlar kullanılır (Chui (1992); Matyas ve Sevestre, 2008):

$$B_r(x|t_0, \dots, t_r) = \frac{1}{(r-1)!}\sum_{j=0}^{r}(-1)^j \binom{r}{j}\left[\max\left(0, (x - t_j)\right)\right]^{r-1} \tag{3.118}$$

(3.117) ile (3.118)'deki eşitlikler, tahmin edici hesaplamaları için gerekli varsayımları ifade eder. Bunlardan yola çıkılarak parametrelerin tahmin edicileri, panel veri kullanılarak hesaplanır.

3.2.4.7 <u>Örneklem'de kesinti olduğu durumda panel veri analizi</u>

Örneklem'de kesim yapmak için Tobin (1958), (3.119) ile (3.120) ile ifade edilen iki eşitliği önermiştir:

$$y = \begin{cases} y^* & y^* > 0 \\ 0 & , dy \end{cases} \tag{3.119}$$

$$y^* = x'\beta + u \tag{3.120}$$

Burada u tek bir dağılımdan gelen 0 ortalamalı ve σ_u^2 varyanslı bir rassal değişkendir. Bu durumda (3.121) ile hesaplanır:

$$E(y|x, y > 0) = x'\beta + E(u|u > x'\beta) \tag{3.121}$$

Böylece $E(y|x) = P(u > x'\beta)[x'\beta + E(u|u > x'\beta)]$ olarak elde edilir. Y değişkeninin formülü (3.122)'deki gibi ifade edilsin:

$$y = x'\beta + \epsilon \tag{3.122}$$

Örneklemde kesim yapıldığı durumlarda $E(\epsilon|x) \neq 0$ olmaktadır (Hsiao, 2014) ve gerçek değerler yakınsamak için (3.123)'deki L2 fonksiyonunu en çoklamak gerekir:

$$L2 = \left(\prod_{y_i^* \leq 0} P(y_i = 0|x_i)\right) x \left(\prod_{y_i^* > 0} f(y_i)\right) \tag{3.123}$$

Burada $P(y_i = 0|x_i)$ ile $f(y_i)$ u'nun normal dağıldığı varsayılırsa, (3.124) ile (3.125) hesaplanır:

$$P(y_i = 0|x_i) = P\left(u \leq \frac{-x_i'\beta}{\sigma_u}\right) \tag{3.124}$$

$$f(y_i) = \frac{1}{\sigma_u\sqrt{2\pi}} \exp\left(-\frac{(y_i - x_i'\beta)^2}{2\sigma_u^2}\right) \tag{3.125}$$

Hem β hem σ hesaplamak için Heckman (1976a) şunu önermiştir(Hsiao, 2014):

i. $\prod_{y_i^* \leq 0} P(y_i = 0|x_i)$'yi en çoklayacak şekilde $\delta = \frac{\beta}{\sigma_u}$ hesaplayın

ii. $s(.)$ İle $\theta(.)$ standart normal dağılımın olasılık yoğunluk ve birikimli olasılık yoğunluk fonksiyonu olmak üzere (3.126)'daki eşitlik hesaplanır:

$$y_i = x_i'\beta + \sigma_u \frac{s(x_i'\hat{\delta})}{\theta(x_i'\hat{\delta})} + \eta_i \qquad (3.126)$$

(3.126)'daki eşitlikte σ_u ile β'nın en küçük kareler tahmin edicileri elde edilir. Heckman tahmin edicisi gerçek değere yakınsar ama en çok olabilirlik fonksiyonu kadar etkin değildir. Ancak gene de faydalı bir tahmin edicidir ve $y_i \geq 2x_i'\beta$ olanları atarak işe yarayabilir.

Bu noktaya kadar anlatılanlar, panel veri analiz edilirken kullanılacak modellerin teorik arka planını vermektedir. Bu modeller, anlaması amaçlanan gerçekliği mümkün olan en yansız şekilde analiz etmek için gerekli araçlardır. Bu araçlar aynı zamanda incelenen veri ile ilgili güvenilir bilgiler üretmek için de kullanılabilmektedir. Sözkonusu bilgiler, incelenen gerçeklikten elde edilebilecek ve belirli bir faydayı gözeten bir şekilde elde edilebilmektedir. Bu bilgileri elde edebilmek için tez çalışmasında kullanılan çeşitli panel verilere uygulanan model tabanlı kümelemeden elde edilen sonuçlar, sözkonusu panel verilerin analizinde kullanılmıştır. Bir sonraki bölümde bu bilgilerin elde edilişi ve bunların yorumlanması incelenecektir.

4. YÖNTEM

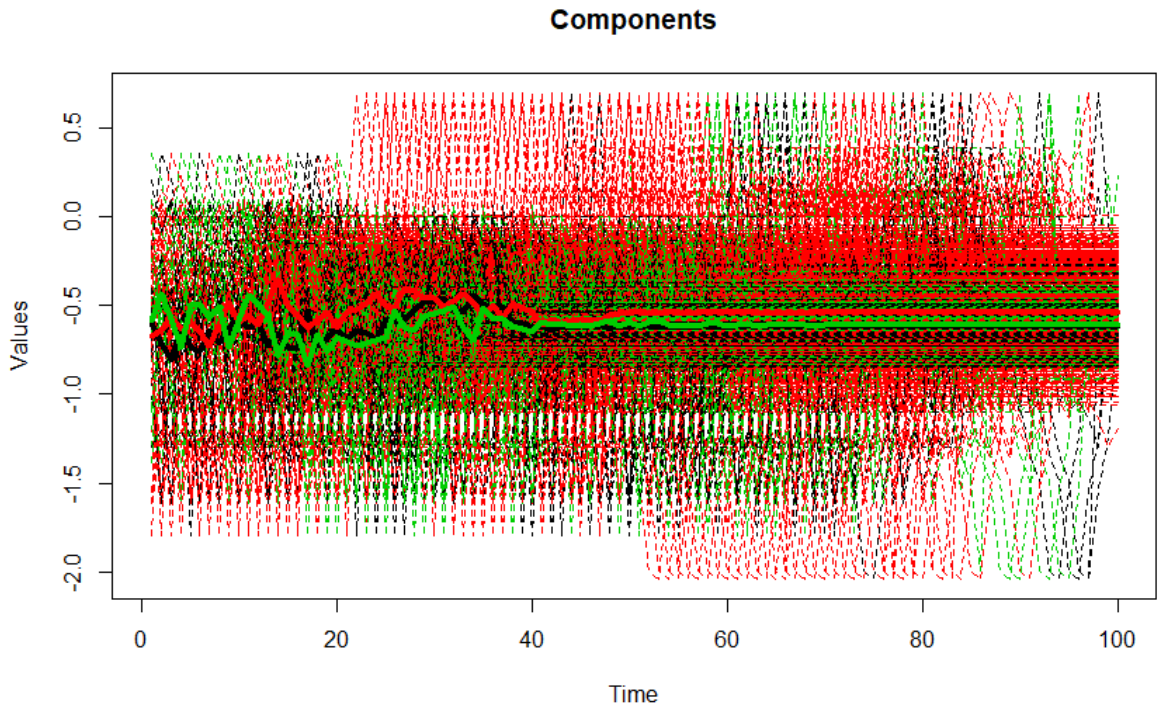
Bu bölümde kullanılan panel veriler, birimler ait zaman serilerinden oluşmaktadır. Veriler üzerinde yapılan uygulamalarda verilerin hepsi, rassal etkiler modelleriyle analiz edilmiştir. Bölümde Borsa İstanbul'dan elde edilen BİST verisi, dünya bankasından elde edilen veri, Maddison projesi kapsamında toplanmış gayrisafi hasıla verisi ile UCI'dan elde edilen Dow-Jones verisi üzerine çalışmalar bulunmaktadır. BİST ve Dow-Jones verisinde sadece hisse senetlerinden hangilerine yatırım yapılabileceği üzerine çalışıldığı için karşılaştırılmalı analize gerek görülmemiştir. Bunun sebebi, bu veriler için kullanılan model tabanlı kümeleme yönteminin piyasalarda yatırım yapması makul olan hisse senetleri seçiminde ön bilgi vermesidir. Söz konusu verilerde, model tabanlı kümeleme yapılmaksızın hisse senetleri seçiminde ön bilgi sağlamak mümkün değildir. Genel olarak bütün verilerde kümeleme analizi gerçekleştirilirken, veriye uygulanan doğrusal modellerde açıklama oranına bağlı bir bilgi artışı hedeflenmiştir. Bu hedef için seçilen verilerde otoregresif zaman serileri bulunmasına özen gösterilmiştir. Gayri safi hasıla verisi haricinde tüm verilerde sözkonusu bilgi artışı sağlanmıştır. Gayri safi hasıla verisi ise ülke gelişim benzerliği konusunda bilgi verdiği için beklenmeyen bir fayda sağlamıştır.

Bütün uygulamalarda, $\alpha = 0{,}10$ olarak alınmıştır. Bütün veriler kümelenirken, R Studio Version 1.2.1335'ten faydalanılmıştır. Bu veriler kümelenirken, R içerisindeki longclust paketi kullanılmıştır. Longclust paketi kullanılırken, longclustEM ile kümeler oluşturulmuş. longclustEM fonksiyonu her biri 2 değer alan 4 parametre üzerinden çalışır. Bütün verilerde $2^4 = 16$ longclustEM fonksiyonu çalıştırılmış ve en yüksek BIC'ye sahip kümeleme stratejisi kullanılmıştır. Bu kümeleme sonuçlarından elde edilen kümelerin gölge değişkenleri, panel veri analizine dahil edilmiştir. Son olarak uygulamalarda model tabanlı kümelemenin yapısı gereği, kümeleri ifade eden gölge değişkenleri içeren doğrusal modellerle öngörü yapılmamıştır.

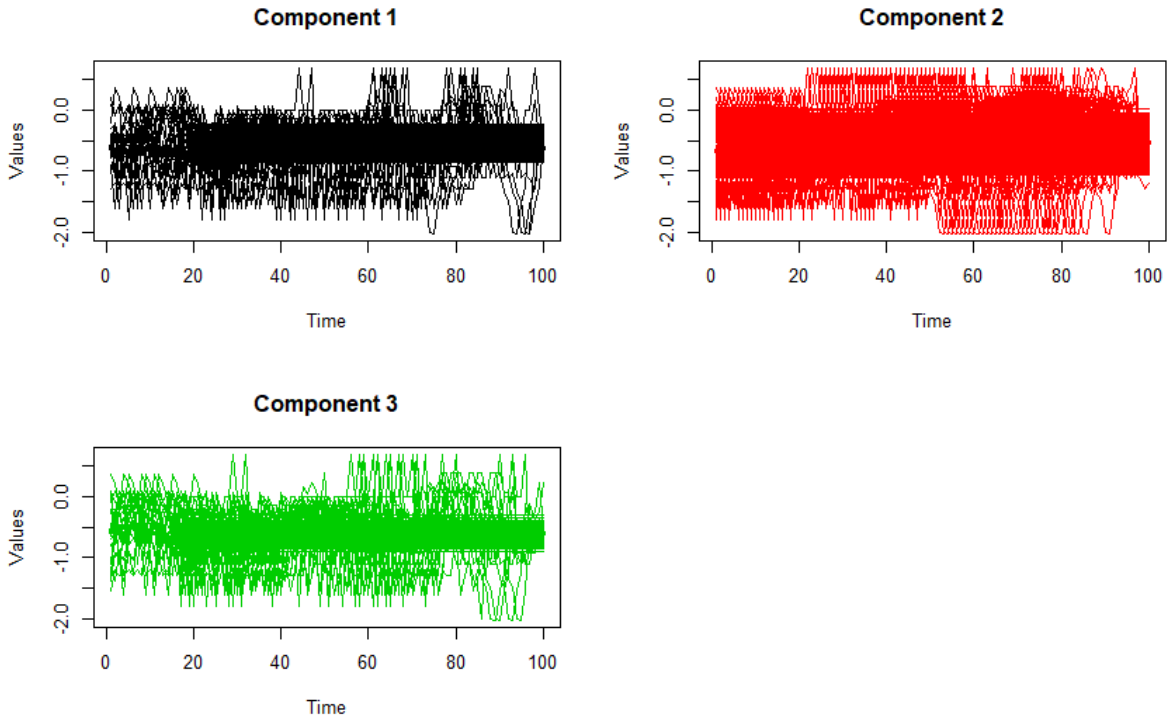
4.1 BİST Verisi ile Panel Veri Analizi ve Panel Veri Kümelemesi

BİST verisi, 2010 ile 2019 yılları arasında günlük olarak toplanmış bir veridir. Kapanış fiyatlarının ve adet olarak ölçülen işlem hacminin değişimleriyle tanımlanmıştır. Veriler

BORSA İstanbul'dan temin edilmiştir ve uzunlukları 31 ile 107 arasında değişen 7799 hisse senedi üzerinde işlem yapılmıştır. Bu hisse senetleri üzerinden 7 günlük veya 30 günlük getiri tahmin edilmiştir. En uzun işlem zamanı bu veri kümelenirken elde edilmiştir. Bu işlem zamanı 1-2 gün sürmüştür. Hisse senetlerinin kümelerinin grafikleri Şekil 4.1a ve Şekil 4.1b'deki gibi elde edilmiştir:



Şekil 4.1a: Hisse senetlerinin kümelerinin birleştirilmiş şekli



Şekil 4.1b: Hisse senetlerinin kümelerinin ayrı ayrı oluşturulmuş şekilleri

Kapanış fiyatlarını k fonksiyonu ile; bir önceki değeri 30 günlük getiri için lagy30 ve 7 günlük getiri için lagy7 ile; sınıf üyeliklerini d_1, d_2, d_3 ile; sınıf üyelikleri ile lagy30 arasındaki etkileşimleri sırayla d1#lagy30 ile; d2#lagy30 ile; d3#lagy30 ile; sınıf üyelikleri ile lagy30'un bir önceki değeri arasındaki etkileşimleri sırayla d1#L.lagy30 ile; d2#L.lagy30 ile; d3#L.lagy30 ile; sınıf üyelikleri ile lagy7 arasındaki etkileşimleri sırayla d1#lagy7 ile; d2#lagy7 ile; d3#lagy7 ile; sınıf üyelikleri ile lagy7'nin bir önceki değeri arasındaki etkileşimleri sırayla d1#L.lagy7 ile; d2#L.lagy7 ile; d3#L.lagy7 ile ve işlem hacmini de h fonksiyonu ile gösterirsek 7 günlük ve 30 günlük getiri modellerimiz (5.1), (5.2), (5.3), (5.4), (5.5), (5.6) formülleriyle ifade edilir:

$$\log\left(\frac{k(t+30)}{k(t)}\right) = \beta_0 + \beta_1 \log\left(\frac{h(t+30)}{h(t)}\right) + \beta_2 \text{lagy30} + \beta_3 d_1 + \text{d1\#lagy30} + \text{d1\#L.lagy30} + u_{it} \qquad (5.1)$$

$$\log\left(\frac{k(t+30)}{k(t)}\right) = \beta_0 + \beta_1 \log\left(\frac{h(t+30)}{h(t)}\right) + \beta_2 lagy30 + \beta_3 d_2 + d2\#lagy30 + d2\#L.\,lagy30 + u_{it} \quad (5.2)$$

$$\log\left(\frac{k(t+30)}{k(t)}\right) = \beta_0 + \beta_1 \log\left(\frac{h(t+30)}{h(t)}\right) + \beta_2 lagy30 + \beta_3 d_3 + d3\#lagy30 + d3\#L.\,lagy30 + u_{it} \quad (5.3)$$

$$\log\left(\frac{k(t+7)}{k(t)}\right) = \beta_0 + \beta_1 \log\left(\frac{h(t+7)}{h(t)}\right) + \beta_2 lagy7 + \beta_3 d_1 + d1\#lagy7 + d1\#L.\,lagy7 + u_{it} \quad (5.4)$$

$$\log\left(\frac{k(t+7)}{k(t)}\right) = \beta_0 + \beta_1 \log\left(\frac{h(t+7)}{h(t)}\right) + \beta_2 lagy7 + \beta_3 d_2 + d2\#lagy7 + d2\#L.\,lagy7 + u_{it} \quad (5.5)$$

$$\log\left(\frac{k(t+7)}{k(t)}\right) = \beta_0 + \beta_1 \log\left(\frac{h(t+7)}{h(t)}\right) + \beta_2 lagy7 + \beta_3 d_3 + d3\#lagy7 + d3\#L.\,lagy7 + u_{it} \quad (5.6)$$

Bu modeller için sonuçlar hem en yüksek p-değerliler çıkarılarak hem de çıkarılmayarak elde edilmiştir. Parantez içindeki değerler en yüksek p-değerliler çıkarıldıklarında kalan değişkenlerin parametre tahminleri olup Çizelge 4.1, Çizelge 4.2, Çizelge 4.3, Çizelge 4.4, Çizelge 4.5, Çizelge 4.6'da ifade edilmiştir:

Çizelge 4.1: 30 günlük getiri için 1. sınıfa dahil olma değişkeni d_1 ile oluşturulan rassal etkiler modeliyle elde edilen sonuçlar

Birimler içi $R^2 = 0{,}7049$ Birimler arası $R^2 = 0{,}9802$ Genel $R^2 = 0{,}8909$						
	Katsayı	Standart Hata	z-değeri	p-değeri	%90 güven aralığı	
logfarkvol30	-0,0018 (-0,0018)	0,0002 (0,0002)	-6,9 (-6,9)	<0,1	-0,0022 (-0,0022)	-0,0013 (-0,0013)
lagy30	0,8719 (0,8720)	0,0031 (0,0031)	278,66 (278,67)	<0,1	0,8668 (0,8668)	0,8771 (0,8771)
d1	-0,0014	0,00408	-0,35	0,728	-0,00813	0,005293
d1#lagy30	0,1006 (0,1005)	0,0033 (0,0033)	30,05 (30,27)	<0,1	0,0951 (0,0950)	0,1062 (0,1059)
d1#L.lagy30	0,0987 (0,0989)	0,0036 (0,0035)	26,96 (27,76)	<0,1	0,0926 (0,0931)	0,1047 (0,1048)
sabit terim	-0,0657 (-0,0662)	0,0023 (0,0019)	-27,64 (-34,04)	<0,1	-0,0696 (-0,0694)	-0,0618 (-0,0630)

Çizelge 4.2: 30 günlük getiri için 2. sınıfa dahil olma değişkeni d_2 ile oluşturulan rassal etkiler modeliyle elde edilen sonuçlar

Birimler içi R^2=0,7049 Birimler arası R^2=0,9802 Genel R^2=0,8909						
	katsayı	standart hata	z-değeri	p-değeri	%90'lık güven aralığı	
logfarkvol30	-0,0018 (-0,0018)	0,0002 (0,0002)	-6,9 (-6,89)	<0,1	-0,0022 (-0,0022)	-0,0013 (-0,0013)
lagy30	0,8719 (0,8719)	0,0031 (0,0031)	278,64 (278,66)	<0,1	0,8668 (0,8668)	0,8771 (0,8771)
d2	0,0051	0,0039	1,3	0,195	-0,0013	0,0116
d2#lagy30	0,0999 (0,1004)	0,0033 (0,0033)	29,66 (30,11)	<0,1	0,0943 (0,0949)	0,1054 (0,1059)
d2#L.lagy30	0,1002 (0,0992)	0,0036 (0,0035)	27,84 (28,22)	<0,1	0,0943 (0,0934)	0,1061 (0,1050)
sabit terim	-0,0681 (-0,0662)	0,0024 (0,0019)	-27,69 (-34,04)	<0,1	-0,0722 (-0,0694)	-0,0641 (-0,0630)

Çizelge 4.3: 30 günlük getiri için 3. sınıfa dahil olma değişkeni d_3 ile oluşturulan rassal etkiler modeliyle elde edilen sonuçlar

Birimler içi R^2=0,7049 Birimler arası R^2=0,9802 Genel R^2=0,8909						
	Katsayı	standart hata	z-değeri	p-değeri	%90 güven aralığı	
logfarkvol30	-0,0018 (-0,0018)	0,0002 (0,0002)	-6,91 (-6,9)	<0,1	-0,0022 (-0,0022)	-0,0013 (-0,0013)
lagy30	0,8719 (0,8719)	0,0031 (0,0031)	278,66 (278,67)	<0,1	0,8668 (0,8668)	0,877134 (0,8771)
d3	-0,0043	0,0042	-1,02	0,31	-0,01136	0,002687
d3#lagy30	0,0995 (0,0991)	0,0033 (0,0033)	29,9 (29,99)	<0,1	0,0940 (0,0937)	0,1050 (0,1045)
d3#L.lagy30	0,1011 (0,1021)	0,0037 (0,0036)	27 (28,21)	<0,1	0,0949 (0,0961)	0,1073 (0,1080)
sabit terim	-0,0649 (-0,0661)	0,0022 (0,0019)	-28,33 (-34,03)	<0,1	-0,0687 (-0,0693)	-0,0611 (-0,0629)

Çizelge 4.4: 7 günlük getiri için 1. sınıfa dahil olma değişkeni d_1 ile oluşturulan rassal etkiler modeliyle elde edilen sonuçlar

Birimler içi R^2 =0,5527 Birim arası R^2 = 0,9205 Genel R^2 = 0,6708						
	Katsayı	standart hata	z-değeri	p-değeri	%90 güven aralığı	
logfarkvol7	-0,0007 (-0,0007)	0,0003 (0,0003)	-2,14 (-2,13)	<0,1	-0,0012 (-0,0012)	-0,0001 (-0,0001)
lagy7	0,8236 (0,8236)	0,0031 (0,0031)	260,37 (260,38)	<0,1	0,8184 (0,8184)	0,8288 (0,8288)
d1	-0,0025	0,0040	-0,63	0,53	-0,0092	0,0041
d1#lagy7	0,0325 (0,0322)	0,0035 (0,0035)	9,03 (9,02)	<0,1	0,0265 (0,0263)	0,0384 (0,0381)
d1#L.lagy7	0,0224 (0,0229)	0,0043 (0,0042)	5,2 (5.42)	<0,1	0,0153 (0,0159)	0,0294 (0,0298)
sabit terim	-0,0624 (-0,0632)	0,0023 (0,0019)	-26,18 (-32,39)	<0,1	-0,0663 (-0,0665)	-0,0585 (-0,0600)

Çizelge 4.5: 7 günlük getiri için 2. sınıfa dahil olma değişkeni d_2 ile oluşturulan rassal etkiler modeliyle elde edilen sonuçlar

Birimler içi R^2=0,5527 Birimler arası R^2=0,9205 Genel R^2=0,6708						
	katsayı	standart hata	Z-değeri	p-değeri	%90'lık güven aralığı	
logfarkvol7	-0,0007	0,0003	-2,13	<0,1	-0,0012	-0,0001
lagy7	0,8237	0,0031	260,39	<0,1	0,8185	0,8289
d2	0,0066	0,0039	1,68	<0,1	0,0001	0,0132
d2#lagy7	0,0276	0,0036	7,6	<0,1	0,0216	0,0335
d2#L.lagy7	0,0315	0,0042	7,48	<0,1	0,0245	0,0384
sabit terim	-0,0657	0,0024	-26,75	<0,1	-0,0698	-0,0617

Çizelge 4.6: 7 günlük getiri için 3. sınıfa dahil olma değişkeni d_3 ile oluşturulan rassal etkiler modeliyle elde edilen sonuçlar

Birimler içi R^2=0,5527 Birimler arası R^2=0,9205 Genel R^2=0,6708						
	Katsayı	standart hata	z-değeri	p-değeri	%90 güven aralığı	
logfarkvol7	-0,0007 (-0,0007)	0,0003 (0,0003)	-2,13 (-2,13)	<0,1	-0,0012 (-0,0012)	-0,0001 (-0,0001)
lagy7	0,8236 (0,8236)	0,0031 (0,0031)	260,35 (260,36)	<0,1	0,8184 (0,8184)	0,8288 (0,8288)
d3	-0,0047	0,0042	-1,13	0,26	-0,01178	0,002201
d3#lagy7	0,0271 (0,0267)	0,0035 (0,0035)	7,69 (7,61)	<0,1	0,0213 (0,0209)	0,0329 (0,0325)
d3#L.lagy7	0,0335 (0,0346)	0,0045 (0,0044)	7,41 (7,85)	<0,1	0,0261 (0,0273)	0,040984 (0,0419)
sabit terim	-0,0618 (-0,0632)	0,0023 (0,0019)	-26,78 (-32,37)	<0,1	-0,06568 (-0,0664)	-0,05808 (-0,0600)

Bu modellerdeki bütün parametrelerin tahminlerini kullanılarak geriye dönük tahminlerimizi gerçekleştirildi. Bu tahminler ile sonra modellerin fiyat küçülmesini ya da büyümesini tahmin etme başarısını ele aldık ve aynı tahmin ile gerçek değerin aynı işarette (pozitif ya da negatif işaret olabilir) çıkmasını işaret30 ve işaret7 gölge değişkenleriyle ifade

ettik. Bunun yanında fiyat düşme beklenirken yükselme ortaya çıkmasını pr30 ve pr7 değişkenleriyle ifade edilmiştir. Bunlar da sırayla 30 günlük getiri ve 7 günlük getiri modellerinden elde edilmişlerdir. Mevzubahis bütün değişkenler 1 ve 0 olarak kodlanmış ve bunun neticesinde, bunların ortalamaları doğal olarak modellerimizin başarısını ortaya koymuşlardır. Bütün değişkenlerin ortalaması, Çizelge 4.7 ile gösterilmiştir:

Çizelge 4.7: 2018 yılında 30 günlük ve 7 günlük model sonucunda oluşturulan isaret30, isaret7, pr30 ve pr7 ortalamaları

	Ortalamalar
isaret30	0,940495
isaret7	0,819434
pr30	0,045964
pr7	0,145573

Bu çizelgedeki değerlere göre, 30 günlük ve 7 günlük getiri modelleri toplamda %98,85 ile %96,92 oranında yatırımcının faydalanabileceği durumları sağlamaktadır. Bu da bize 30 günlük tahminlerin daha başarılı olduklarını göstermektedir. Bunun yanında fiyatlarda düşüş beklendiğinde bunun olması, nn30 ve nn7 ile ifade edilmiştir. En önemli risk fiyatlarda artış beklendiğinde fiyatların düşmesi ise mutlakrisk30 ve mutlakrisk7 ile ifade edilmiştir. Çizelge 4.8 ile bu değişkenlerin ortalaması gösterilmiştir:

Çizelge 4.8: 2018 yılında 30 günlük ve 7 günlük model sonucunda oluşturulan nn30 ve nn7'nin ortalamaları

	Ortalamalar
nn30	0,657064
nn7	0,581738
mutlakrisk30	0,013542
mutlakrisk7	0,034994

Bu değerler düşüldüğünde mutlak kazanç sağlama olasılıkları olan kazanma30 ile kazanma7 bulunur. Bu değerler, 30 günlük öngörülerin ya da 7 günlük öngörülerin sonuçlarının bulunmasını sağlar. Bu sonuçlar kazanma/mutlak(30) ile kazanma/mutlak(7) oranlarıyla

belirlenmiştir: 2018 yılı için bu değerler sırayla 20,93 ile 6,79'dur. 2017'den 2015'e kadar için bütün bu kriterler Çizelge 4.9'da belirtilmiştir:

Çizelge 4.9: 30 günlük ve 7 günlük model sonucunda oluşturulan isaret30, isaret7, pr30, pr7, nn30, nn7, kazanma30, kazanma7, mutlakrisk30, mutlakrisk7 ortalamaları ve kazanma/mutlak(30) ile kazanma/mutlak(7) oranları

Değişkenler	2017 yılına kadar ortalamalar	2016 yılına kadar ortalamalar	2015 yılına kadar ortalamalar
isaret30	0,9375	0,9282	0,9437
isaret7	0,8138	0,8285	0,8314
pr30	0,0525	0,0594	0,0440
pr7	0,1590	0,1390	0,1351
nn30	0,6614	0,6793	0,7161
nn7	0,5569	0,5688	0,5938
kazanma30	0,2760	0,2489	0,2276
kazanma7	0,2568	0,2597	0,2375
mutlakrisk30	0,0098	0,0122	0,0122
mutlakrisk7	0,0271	0,0323	0,0334
kazanma/mutlak (30)	27,9792	20,2999	18,6017
kazanma/mutlak (7)	9,4485	8,0412	7,1077

kazanma/mutlak(30) ile kazanma/mutlak(7) aralarındaki farklardan ötürü 30 günlük öngörüler, 7 günlük öngörülere göre daha başarılıdır. Bunun nedeni kayıp yaşanana kadar para kazanmayı ifade eden kazanma/mutlak(30) değişkeni daha yüksek çıkmaktadır. p-değerleri çıkarıp aynı sonuçları elde ettiğimizde kazanma/mutlak(30) ile kazanma/mutlak(7) değerlerinin sırayla 15,87761 ile 6,676987 olduğu tespit edilmiştir. p-değerlerini yüksek bulduğumuz değişkenleri atarak yeni modeller oluşturduğumuzda bu sonucun değişmediği görülmüştür. Çizelge 4.10, p-değerlerine göre yüksek çıkan değişkenleri attığımızda sonuçların ne olacağını göstermektedir:

Çizelge 4.10: 30 günlük ve 7 günlük model sonucunda oluşturulan isaret30, isaret7, pr30, pr7, nn30, nn7 ortalamaları ile kazanma30, kazanma7, mutlakrisk30, mutlakrisk7 ortalamaları ve kazanma/mutlak(30) ile kazanma/mutlak(7) oranları

Değişkenler	2017 yılına kadar ortalamalar	2016 yılına kadar ortalamalar	2015 yılına kadar ortalamalar
isaret30	0,9341	0,923389	0,935961
isaret7	0,8132	0,828491	0,8314
pr30	0,0501	0,056246	0,041104
pr7	0,1580	0,138082	0,134753
nn30	0,6492	0,661353	0,7074
nn7	0,5557	0,567895	0,593358
kazanma30	0,2848	0,262036	0,228561
kazanma7	0,2575	0,260596	0,238042
mutlakrisk30	0,0157	0,020366	0,022936
mutlakrisk7	0,0287	0,033427	0,033847
kazanma/mutlak(30)	18,0887	12,86651	9,965372
kazanma/mutlak(7)	8,9645	7,795857	7,032871

Bu değerler p-değerlerinin atılarak yeni modeller elde etmenin 7 günlük öngörülere yaradığını gösteriyor. Buna rağmen 30 günlük öngörüler yine daha güvenilir durumdadır. 30 günlük ve 7 günlük kazançlar, p-değeri yüksek olan değişkenler çıkarıldığı için azalmıştır. Buna rağmen, 30 günlük kazançlar daha büyük düşüş göstermiştir.

4.2 Gayri Safi Yurtiçi Hasıla Verisi ile Panel Veri Analizi ve Panel Veri Kümelemesi

Gayrisafi yurt içi hasıla, gayrisafi milli hasıla'dan net dış faktör gelirlerinin düşülmesiyle hesaplanan bir değerdir ve üretim değerleri, harcamalar, gelir değerleri ve son olarak para arzı ve dolaşım hızlarının bütüncül değerlendirilmesi ile bulunur (Eğilmez, 2018). Web of Knowledge sitesine bakıldığında 11 haziran 2019 itibariyle "panel data" ve "gdp" anahtar kelimeleri yazıldığında, 2320 makale bulunmakta ve bunlardan en eskisi 1994 yılında yayınlanmıştır. Gayrisafi hasıla verisi, bu bilgiler ışığında sıkça çalışılan bir veri olarak değerlendirilebilir. Çalışmada kullanılan bu veri, Maddison projesi kapsamında Groningen Growth and Development Centre'dan indirilmiştir. Bu veri üzerinde yapılan ilk çalışmalardan biri Bolt (2018)'e aittir.

Bu veride kişi başına düşen gayrisafi yurtiçi hasıla'nın 5 yıllık oranları modellenmiştir. Gayrisafi yurtiçi hasılayı gdp(.) ile; gdp'nin değişiminin bir önceki değerini lagy5 ile; küme üyeliklerini d1, d2, d3, d4, d5, d6 ile gösterince, veriden elde edilen model (5.7), (5.8), (5.9), (5.10), (5.11), (5.12)'deki gibi ifade edilmiştir:

$$\log\left(\frac{\mathrm{gdp(t+5)}}{\mathrm{gdp(t)}}\right) = \beta_0 + \beta_1 \mathrm{lagy5} + \beta_3 \mathrm{d}_1 + \mathrm{u}_{\mathrm{it}} \tag{5.7}$$

$$\log\left(\frac{\mathrm{gdp(t+5)}}{\mathrm{gdp(t)}}\right) = \beta_0 + \beta_1 \mathrm{lagy5} + \beta_3 \mathrm{d}_2 + \mathrm{u}_{\mathrm{it}} \tag{5.8}$$

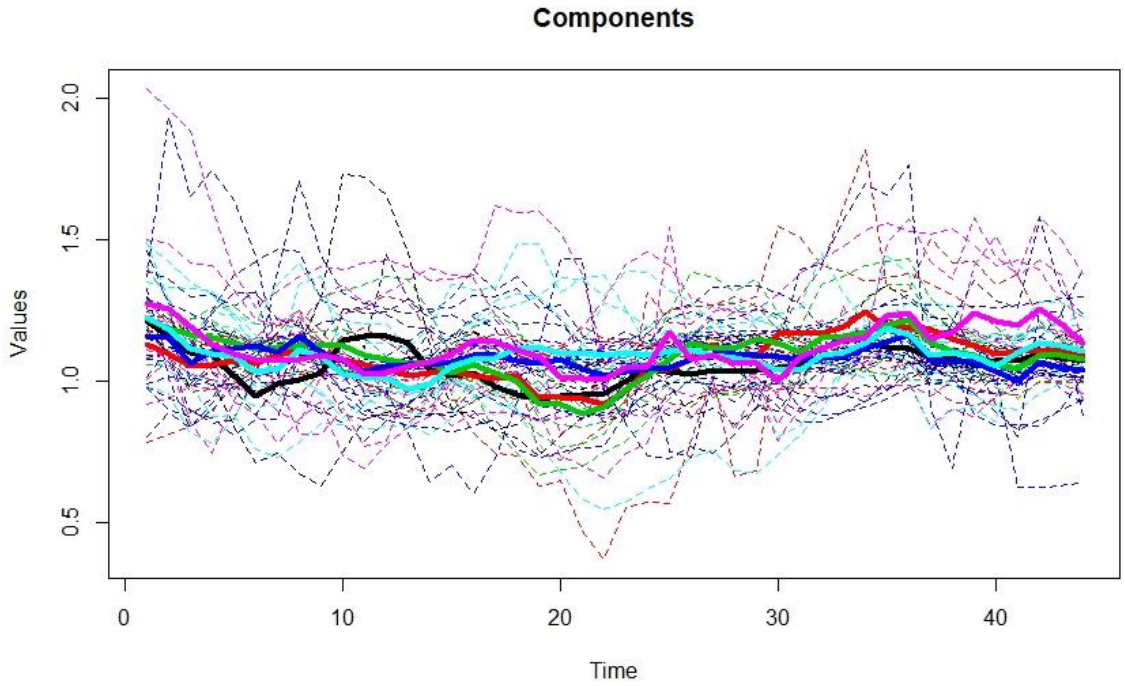
$$\log\left(\frac{\mathrm{gdp(t+5)}}{\mathrm{gdp(t)}}\right) = \beta_0 + \beta_1 \mathrm{lagy5} + \beta_3 \mathrm{d}_3 + \mathrm{u}_{\mathrm{it}} \tag{5.9}$$

$$\log\left(\frac{\mathrm{gdp(t+5)}}{\mathrm{gdp(t)}}\right) = \beta_0 + \beta_1 \mathrm{lagy5} + \beta_3 \mathrm{d}_4 + \mathrm{u}_{\mathrm{it}} \tag{5.10}$$

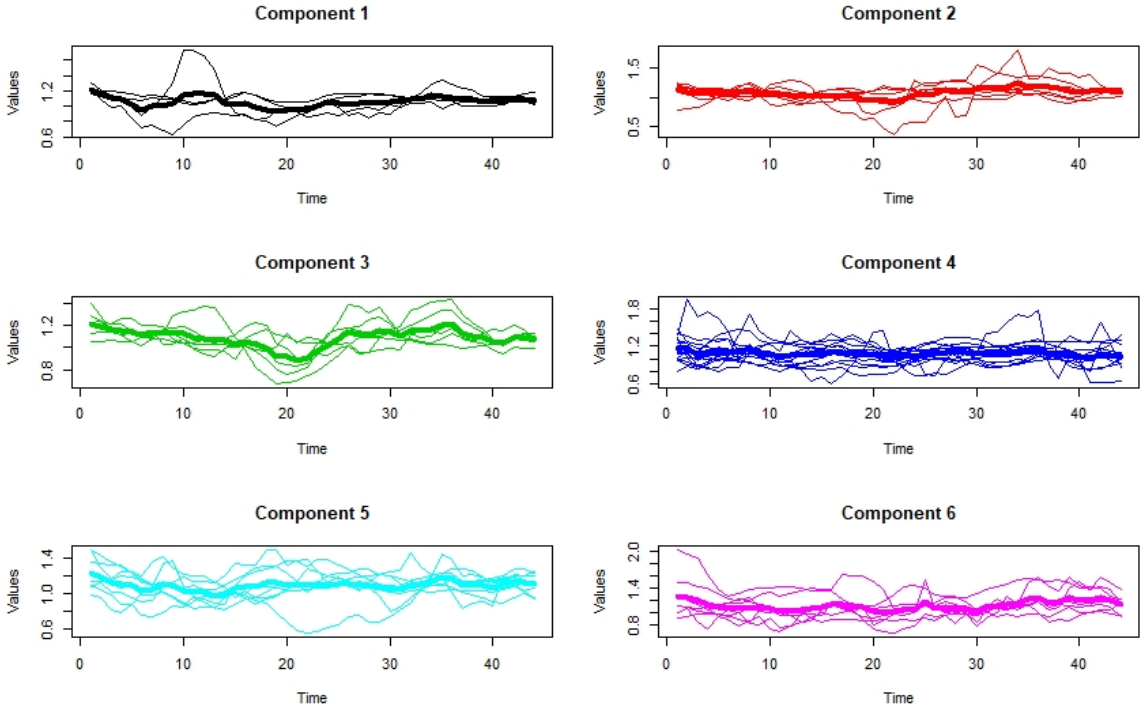
$$\log\left(\frac{\mathrm{gdp(t+5)}}{\mathrm{gdp(t)}}\right) = \beta_0 + \beta_1 \mathrm{lagy5} + \beta_3 \mathrm{d}_5 + \mathrm{u}_{\mathrm{it}} \tag{5.11}$$

$$\log\left(\frac{\mathrm{gdp(t+5)}}{\mathrm{gdp(t)}}\right) = \beta_0 + \beta_1 \mathrm{lagy5} + \beta_3 \mathrm{d}_6 + \mathrm{u}_{\mathrm{it}} \tag{5.12}$$

Kümeleme sonucunda elde edilen kümelerin birleştirilmiş ve ayrı ayrı grafikleri Şekil 4.2a ve Şekil 4.2b ile gösterilmiştir:



Şekil 4.2a: Gayrisafi hasıla oranlarının kümelerinin birleştirilmiş şekli



Şekil 4.2b: Gayrisafi hasıla oranlarının kümelerinin ayrı ayrı şekli

Ülkelerin üyelik dereceleri Çizelge 4.11 ile gösterilmiştir:

Çizelge 4.11: Gayrisafi hasıla oranlarına göre küme üyelikleri

Ülke İsmi	Küme 1	Küme 2	Küme 3	Küme 4	Küme 5	Küme 6
Afganistan	2,36E-71	**1,00E+00**	2,74E-83	4,12E-74	3,88E-76	6,56E-132
Arnavutluk	1,42E-71	**1,00E+00**	2,84E-83	2,74E-74	2,19E-76	5,88E-132
Arjantin	2,44E-38	7,68E-76	2,07E-70	1,93E-42	**1,00E+00**	1,61E-120
Avustralya	**1,00E+00**	1,74E-73	4,28E-65	4,30E-32	2,29E-43	3,03E-125
Avusturya	1,07E-27	3,58E-72	1,52E-51	**1,00E+00**	7,86E-41	4,87E-119
Burundi	8,83E-22	2,21E-69	1,66E-52	**1,00E+00**	2,92E-40	6,15E-120
Belçika	1,30E-39	2,76E-76	2,25E-71	3,93E-43	**1,00E+00**	1,80E-120
Benin	3,34E-26	2,32E-72	3,19E-52	**1,00E+00**	8,35E-39	6,53E-119
Burkina Faso	4,94E-30	5,16E-72	2,10E-49	**1,00E+00**	2,65E-41	4,78E-119
Bangladeş	2,96E-65	1,04E-84	**1,00E+00**	1,31E-54	7,98E-73	2,86E-122
Bulgaristan	2,69E-37	3,07E-75	2,02E-71	8,25E-44	**1,00E+00**	9,24E-121
Bolivya	3,81E-122	1,81E-131	3,11E-121	1,16E-120	1,66E-120	**1,00E+00**
Brezilya	6,09E-72	**1,00E+00**	1,37E-83	1,94E-74	9,93E-77	5,26E-132
Barbados	3,23E-122	1,62E-131	2,70E-121	1,02E-120	1,43E-120	**1,00E+00**
Botsvana	1,81E-29	6,72E-72	4,39E-51	**1,00E+00**	1,19E-41	3,84E-119
Merkezi Afrika cumhuriyeti	**1,00E+00**	1,30E-73	1,30E-66	5,95E-31	6,26E-38	6,23E-125
Kanada	2,89E-29	6,30E-74	7,43E-47	**1,00E+00**	3,44E-42	1,55E-118
İsviçre	2,08E-37	9,79E-76	1,92E-70	1,12E-41	**1,00E+00**	1,44E-120
Şili	3,94E-122	1,89E-131	3,19E-121	1,25E-120	1,73E-120	**1,00E+00**
Çin	2,40E-31	4,30E-73	6,04E-50	**1,00E+00**	8,53E-43	8,76E-119
Fildişi	6,49E-65	7,80E-84	**1,00E+00**	2,52E-54	1,68E-72	1,43E-122
Kamerun	2,06E-38	3,61E-75	4,14E-71	2,46E-42	**1,00E+00**	5,95E-121
Kongo Demokratik Cumhuriyeti	**1,00E+00**	2,26E-74	1,64E-66	8,68E-33	1,42E-39	6,99E-125
Kongo	3,26E-122	1,55E-131	2,19E-121	9,95E-121	1,57E-120	**1,00E+00**
Komoros	5,69E-72	**1,00E+00**	8,25E-84	6,02E-75	5,26E-77	4,66E-132

Çizelge 4.11(devam)

Cabo verde	5,06E-29	1,07E-71	4,88E-52	**1,00E+00**	3,10E-42	2,25E-119
Kosta Rika	6,67E-70	**1,00E+00**	4,12E-83	3,82E-73	6,20E-75	9,91E-132
Çek cumhuriyeti	1,54E-61	2,62E-83	**1,00E+00**	7,28E-50	1,66E-69	5,68E-122
Kıbrıs	2,90E-27	4,50E-72	1,48E-50	**1,00E+00**	1,00E-40	5,09E-119
Almanya	9,86E-29	2,60E-73	1,26E-51	**1,00E+00**	4,08E-37	1,09E-118
Cibuti	**1,00E+00**	8,09E-74	1,02E-65	5,84E-31	5,34E-40	5,19E-125
Dominik	1,53E-29	1,39E-72	1,59E-50	**1,00E+00**	1,06E-41	5,67E-119
Danimarka	7,34E-28	8,39E-74	1,03E-51	**1,00E+00**	3,32E-43	1,15E-118
Dominik cumhuriyeti	2,74E-37	6,89E-76	1,59E-71	8,21E-44	**1,00E+00**	1,42E-120
Ekvador	1,15E-64	3,97E-84	**1,00E+00**	1,14E-53	2,17E-72	2,12E-122
Mısır	5,29E-31	1,25E-72	3,64E-51	**1,00E+00**	1,82E-43	3,00E-119
İspanya	3,08E-37	1,38E-76	5,45E-70	1,02E-40	**1,00E+00**	3,70E-120
Etyopya	2,96E-122	1,49E-131	2,22E-121	8,91E-121	1,29E-120	**1,00E+00**
Finlandiya	3,00E-64	5,41E-84	**1,00E+00**	1,78E-53	4,69E-72	2,02E-122
Fransa	7,22E-65	4,06E-84	**1,00E+00**	3,51E-53	3,56E-72	2,18E-122
Gabon	4,07E-29	3,64E-72	6,21E-51	**1,00E+00**	1,44E-41	4,29E-119
Birleşik krallık	5,47E-72	**1,00E+00**	2,02E-83	1,45E-74	2,59E-76	4,33E-132
Gana	2,75E-122	1,40E-131	2,03E-121	8,24E-121	1,19E-120	**1,00E+00**
Gine	4,37E-29	7,02E-72	2,17E-50	**1,00E+00**	2,83E-42	3,56E-119
Gambiya	3,89E-122	1,79E-131	2,89E-121	1,17E-120	1,68E-120	**1,00E+00**
Gine-Bissau	2,79E-122	1,39E-131	2,18E-121	8,51E-121	1,21E-120	**1,00E+00**
Yunanistan	1,03E-28	3,26E-71	7,44E-51	**1,00E+00**	2,76E-40	2,88E-119
Guatemala	6,06E-72	**1,00E+00**	1,02E-83	1,95E-74	2,54E-76	5,26E-132
Hong Kong	1,90E-38	6,06E-76	2,23E-71	3,01E-43	**1,00E+00**	1,59E-120
Honduras	2,96E-28	1,05E-71	1,67E-50	**1,00E+00**	6,41E-41	3,68E-119

Bu veri incelenirken, geçmişe dönük olarak gayrisafi yurtiçi hasılanın hangi ülkeler tarafından daha iyi geliştirildiği bulunmaya çalışılmıştır. Bunu yaparken ilk rassal etkili model, Çizelge 4.12'deki gibi oluşturulmuştur:

Çizelge 4.12: Rassal etkiler modeli sonuçları

Birimler içi R^2=0,7626 Birimler arası R^2=0,9953 Genel R^2 =0,8174						
	katsayı	standart hata	z-değeri	p-değeri	%90'lık Güven Aralığı	
lagy5	0,8874	0,0091	96,83	<0,1	0,8723	0,9025
sabit terim	0,1195	0,0100	11,88	<0,1	0,1029	0,1360

Son model oluşturulurken: d1 haricindeki küme üyeliklerinin lagy5 ile etkileşim değişkenleri (d2#c.lagy5 gibi) STATA 12.0 tarafından atılmıştır. Söz konusu etkileşim değişkenleri ayrı ayrı değerlendirilmiştir. Bütün etkileşim değişkenleri, sonuçları ayrı ayrı değerlendirildikleri vurgulanmak için parantez içinde ifade edilmiştir. Sonuçlar Çizelge 4.13'de gösterilmiştir.

Çizelge 4.13: Küme gölge değişkenlerini içeren rassal etkili model

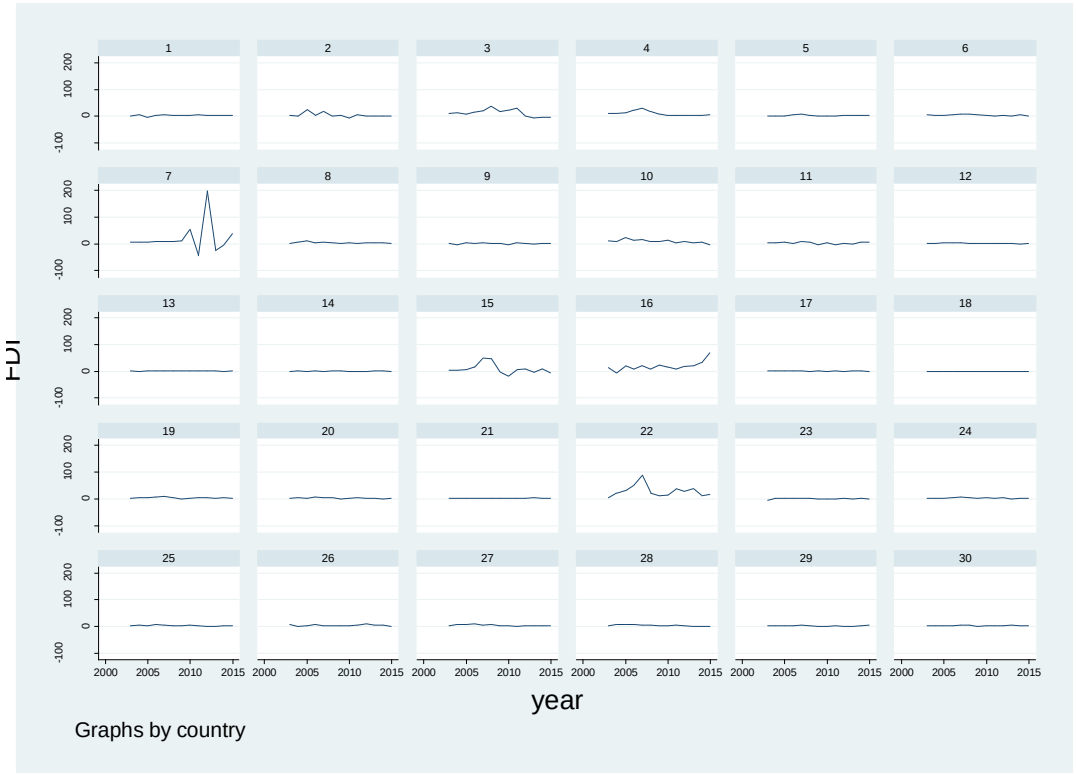
Birimler içi R^2=0,7626 Birimler arası R^2=0,9953 Genel R^2 =0,8174						
	katsayı	standart hata	z-değeri	p-değeri	%90'lık Güven Aralığı	
(d1#c.lagy5)	0,8874	0,0091	96,83	<0,1	0,8723	0,9025
(d2#c.lagy5)	0,8874	0,0091	96,83	<0,1	0,8723	0,9025
(d3#c.lagy5)	0,8874	0,0091	96,83	<0,1	0,8723	0,9025
(d4#c.lagy5)	0,8874	0,0091	96,83	<0,1	0,8723	0,9025
(d5#c.lagy5)	0,8874	0,0091	96,83	<0,1	0,8723	0,9025
(d6#c.lagy5)	0,8874	0,0091	96,83	<0,1	0,8723	0,9025
sabit terim	0,1195	0,0100	11,88	<0,1	0,0998	0,1392

Bu kümeler ülkelerdeki kişi başına düşen gayri safi hasılaların 5 yıllık oranların, bir önceki yıldaki oransal değişimlerle birlikte etkili olduklarını göstermiştir.

4.3 Yabancı Yatırım Verisi ile Panel Veri Analizi ve Panel Veri Kümelemesi

Yabancı yatırım verisi, dünya bankasından elde edilmiştir. Ülkeler arasındaki yabancı yatırımın gayrisafi yurtiçi hasılaları içindeki yüzdeleriyle ilgilenilmiştir. Ülkelerin işgüçleri ile yabancı yatırım yüzdeleri arasında ilişki olduğu literatürde belirtilmektedir. 1975-2015 yılları arasında web of knowledge sitesinde "Labour" ile "Foreign Direct Investment" anahtar kelimeleri ile 1065 makale bulunmuştur. Maza ve Villaverde (2015), avrupa birliğindeki ülkeler için yabancı yatırımı çekme endeksini işgücü ile ilgili kavramları kullanarak oluşturmuştur. Bu nedenle işgücü ile yabancı yatırımdan yola çıkılarak ülkelerin kümelenmesi incelenmeye çalışılmıştır.

Çalışmada kullanılan ülkeler kod sırasıyla Avustralya, Avusturya, Belçika, Bulgaristan, Kanada, Hırvatistan, Kıbrıs Cumhuriyeti, Çekya, Danimarka, Estonya, Finlandiya, Fransa, Almanya, Yunanistan, Macaristan, İrlanda, İtalya, Japonya, Letonya, Litvanya, Meksika, Hollanda, Yeni Zelanda, Norveç, Polonya, Portekiz, Romanya, Slovakya, Slovenya ve İspanyadır. Ülkeler için yabancı yatırım açısından grafikleri çizildiğinde Şekil 4.3'e ulaşılır:



Şekil 4.3: Yabancı yatırım yüzdelerinin ülkelere göre grafikleri

Bu çalışmada kullanılan değişkenler CFWF, CFWM, ETPR1, ETPR2, ETPR3, ETPR4, ETPR5, ETPR6, ETPR7, ETPR8, EI1, EI2, ES1, ES2, FDIe, year'dir. FDIe, bir önceki yıla ait yabancı yatırımın gayrisafi yurtiçi hasıladaki yüzdesidir. year değişkeni, yabancı yatırımın yıla ait etkisidir. CFWF, yüzdelik olarak kadın işgücüne katılan kadın işçi oranıdır. CFWM, yüzdelik olarak erkek işgücüne katılan erkek işçi oranıdır. ETPR1 ve ETPR3, yüzdelik olarak ulusal 15-24 yaş arası kadın işgücünün nüfusa oran tahminleridir. ETPR5 ve ETPR7, yüzdelik olarak ulusal 15 yaş üstü kadın işgücünün nüfusa oran tahminleridir. ETPR2, ETPR4, ETPR6, ETPR8 ise ETPR1, ETPR3, ETPR5, ETPR7'deki yaş ve cinsiyet kategorilere göre tahminlerinin Uluslararası İşgücü Organizasyonuna (ILO) göre yüzde olarak karşılığını ifade etmektedir. EI1 ve EI2 değişkenleri, endüstri sektöründe çalışan kadınların kadın işgücüne oranı ile endüstri sektöründe çalışan erkeklerin erkek işgücüne oranıdır. ES1 ve ES2 değişkenleri, hizmet sektöründe çalışan kadınların kadın işgücüne oranı ile endüstri sektöründe çalışan erkeklerin erkek işgücüne oranıdır. Elde edilen kümeleme sonuçlarından 2. kümeye düşen ülkeler için dson2 değişkeni oluşturulmuştur.

$$\begin{aligned} \mathrm{FDI} = {} & \beta_0 + \beta_1 \mathrm{ETPR5} + \beta_2 \mathrm{ETPR6} + \beta_3 \mathrm{ETPR7} + \beta_4 \mathrm{ETPR8} + \beta_5 \mathrm{dson2} \\ & + \beta_6 \mathrm{dson2\#ETPR6} + \beta_7 \mathrm{year} + \beta_8 \mathrm{year\#dson2} + \beta_9 \mathrm{year\#FDIe} \\ & + u_{it} \end{aligned} \qquad (5.13)$$

Kümeleme yapıldıktan sonra elde edilen sonuçlar, Çizelge 4.14'te gösterilmiştir:

Çizelge 4.14: Ülkelerin ait olduğu kümeler

Ülke Adı	Küme 1	Küme 2	Ülke Adı	Küme 1	Küme 2
Avustralya	**0,999998**	1,75E-06	Almanya	**1**	1,18E-10
Avusturya	**0,646585**	3,53E-01	Yunanistan	**1**	1,99E-12
Belçika	0,072746	**9,27E-01**	Macaristan	0,030176	**9,70E-01**
Bulgaristan	0,39306	**6,07E-01**	İrlanda	0,009805	**9,90E-01**
Kanada	**1**	2,18E-10	İtalya	**1**	2,59E-12
Hırvatistan	**1**	2,14E-08	Japonya	**1**	1,57E-12
Kıbrıs Cumhuriyeti	0,010144	**9,90E-01**	Letonya	**1**	4,67E-10

Çizelge 4.14(devam)

Çekoslovakya	**1**	2,60E-07	Litvanya	**1**	3,22E-11
Danimarka	**1**	1,97E-07	Meksika	**1**	1,60E-13
Estonya	**0,969147**	3,09E-02	Hollanda	0,008446	**9,92E-01**
Finlandiya	**0,999999**	1,09E-06	Yeni Zelanda	**1**	5,17E-08
Fransa	**1**	5,63E-13	Norveç	**1**	3,96E-11
			Polonya	**1**	2,04E-10
			Portekiz	**0,999996**	4,12E-06
			Romanya	**1**	7,50E-08
			Slovakya	**1**	3,45E-09
			Slovenya	**1**	3,36E-12
			İspanya	**1**	9,62E-12

Sabit etkiler modeli yıllık etkileri ölçmek için gerçekleştirildiğinde, Çizelge 4.15'teki sonuçlar elde edilmiştir. Bunun yanında Çizelge 4.16'daki sonuçlar son modelde dson2 değişkeni eklenmeden önceki sonuçları bulunmaktadır:

Çizelge 4.15: Sabit etkiler modeli sonuçları

Birimler içi R^2=0,1625

Birimler arası R^2=0,0952

GenelR^2 <0,0001

	Katsayı	Standart Hata	z-değeri	p-değeri	%90 güven aralığı	
ETPR5	5,005596	1,895929	2,64	<0,1	1,878575	8,132616
ETPR6	-4,96243	2,108702	-2,35	<0,1	-8,44038	-1,48447
ETPR7	-4,50884	1,635789	-2,76	<0,1	-7,2068	-1,81087
ETPR8	4,375728	1,755904	2,49	<0,1	1,479655	7,271801
year						
2004	-0,37822	3,158402	-0,12	0,905	-5,58749	4,831038
2005	2,980069	3,184518	0,94	0,35	-2,27227	8,232404
2006	5,043026	3,287948	1,53	0,126	-0,3799	10,46595
2007	10,38078	3,415276	3,04	<0,1	4,747851	16,01371
2008	7,009545	3,567975	1,96	<0,1	1,124761	12,89433
2009	0,13509	3,652923	0,04	0,971	-5,8898	6,159981
2010	0,086901	3,71582	0,02	0,981	-6,04173	6,215531
2011	-1,69733	3,734864	-0,45	0,65	-7,85737	4,462705
2012	4,69995	3,799006	1,24	0,217	-1,56588	10,96578

Çizelge 4.15(devam)						
2013	-1,64437	3,834079	-0,43	0,668	-7,96804	4,679312
2014	-2,60101	3,917143	-0,66	0,507	-9,06169	3,859671
2015	0,273888	3,890139	0,07	0,944	-6,14225	6,690027
sabit terim	24,33405	23,26677	1,05	0,296	-14,0406	62,70874

Çizelge 4.16: Rassal etkiler modeli sonuçları

Birimler içi R^2=0,0796 Birimler arası R^2=0,0088 GenelR^2 = 0,0234						
	Katsayı	Standart Hata	z-değeri	p-değeri	%90 güven aralığı	
ETPR5	1,874763	0,870936	2,15	<0,1	0,442201	3,307325
ETPR6	-1,72229	0,890715	-1,93	<0,1	-3,18739	-0,2572
ETPR7	-2,87215	1,099738	-2,61	<0,1	-4,68106	-1,06325
ETPR8	2,770015	1,091677	2,54	<0,1	0,974366	4,565664
year						
2004	1,300016	3,930619	0,33	0,741	-5,16528	7,765308
2005	2,637178	3,764792	0,7	0,484	-3,55535	8,829709
2006	0,630353	3,817804	0,17	0,869	-5,64938	6,910081
2007	-2,33936	3,774367	-0,62	0,535	-8,54764	3,868919
2008	1,029755	3,690782	0,28	0,78	-5,04104	7,100551
2009	0,443535	3,706002	0,12	0,905	-5,6523	6,539366
2010	-1,479	3,665164	-0,4	0,687	-7,50765	4,549662
2011	3,379869	3,55114	0,95	0,341	-2,46124	9,220974
2012	15,6575	3,494367	4,48	<0,1	9,909779	21,40522
2013	1,834013	3,50119	0,52	0,6	-3,92493	7,592959
2014	0,329053	3,497431	0,09	0,925	-5,42371	6,081816
2015	-2,46412	3,628997	-0,68	0,497	-8,43329	3,505049
year#c.FDIe						
2003	0,450124	0,432062	1,04	0,298	-0,26055	1,160802
2004	0,269242	0,542329	0,5	0,62	-0,62281	1,161293
2005	0,654076	0,385207	1,7	<0,1	0,020467	1,287686
2006	0,771463	0,242871	3,18	<0,1	0,371975	1,17095
2007	1,657326	0,202581	8,18	<0,1	1,32411	1,990541
2008	0,388589	0,111658	3,48	<0,1	0,204928	0,57225

Çizelge 4.16(devam)

2009	0,160289	0,185292	0,87	0,387	-0,14449	0,465067
2010	1,288133	0,343847	3,75	<0,1	0,722555	1,853711
2011	-0,3865	0,172724	-2,24	<0,1	-0,6706	-0,1024
2012	-1,93662	0,163627	-11,84	<0,1	-2,20576	-1,66748
2013	-0,12244	0,05496	-2,23	<0,1	-0,21284	-0,03203
2014	0,443773	0,210877	2,1	<0,1	0,096912	0,790634
2015	1,608306	0,300515	5,35	<0,1	1,114004	2,102609
sabit terim	0,31611	6,040171	0,05	0,958	-9,61909	10,25131

Yıllık etkiler için parametre testi yapıldığında p değeri 0,0369 çıkmıştır ve yıllık etkilerin var olduğu sonucuna varılmıştır. Son model kurulduğunda Çizelge 4.17'daki sonuçlar elde edilmiştir:

Çizelge 4.17: Rassal etkiler modeli sonuçları

Birimler içi R^2=0,6407

Birimler arası R^2=0,9218

GenelR^2 =0,7033

	katsayı	standart hata	z-değeri	p-değeri	%90'lık güven aralığı	
ETPR5	1,66839	0,6980457	2,39	<0,1	0,520207	2,816573
ETPR6	-1,57182	0,7130658	-2,2	<0,1	-2,74471	-0,39893
ETPR7	-2,52536	0,8854239	-2,85	<0,1	-3,98175	-1,06897
ETPR8	2,38037	0,8749743	2,72	<0,1	0,941166	3,819575
dson2	-38,9933	10,64832	-3,66	<0,1	-56,5082	-21,4784
dson2#ETPR6						
1	1,033698	0,2074166	4,98	<0,1	0,692528	1,374868
year						
2004	0,975166	3,048072	0,32	0,749	-4,03847	5,988797
2005	1,808494	2,915557	0,62	0,535	-2,98717	6,604159
2006	0,578049	2,949812	0,2	0,845	-4,27396	5,430058
2007	-1,90149	2,917419	-0,65	0,515	-6,70022	2,897236
2008	0,977315	2,849091	0,34	0,732	-3,70902	5,663652

Çizelge 4.17(devam)

2009	-0,72042	2,87271	-0,25	0,802	-5,44561	4,004768
2010	-1,69754	2,830567	-0,6	0,549	-6,35341	2,958326
2011	1,719965	2,782506	0,62	0,536	-2,85685	6,296781
2012	5,995722	2,782011	2,16	<0,1	1,419721	10,57172
2013	-0,06166	2,769481	-0,02	0,982	-4,61706	4,493727
2014	-0,78461	2,789043	-0,28	0,778	-5,37218	3,802955
2015	-3,86797	2,837067	-1,36	0,173	-8,53453	0,798589
year#dson2						
2003 1	-1,56039	6,021243	-0,26	0,796	-11,4645	8,343675
2004 1	-1,05942	6,463099	-0,16	0,87	-11,6903	9,57143
2005 1	0,331681	5,863928	0,06	0,955	-9,31362	9,976985
2006 1	3,592921	5,970038	0,6	0,547	-6,22692	13,41276
2007 1	-0,41905	6,538003	-0,06	0,949	-11,1731	10,33501
2008 1	6,29678	6,761026	0,93	0,352	-4,82412	17,41768
2009 1	-2,42437	7,188647	-0,34	0,736	-14,2486	9,399902
2010 1	-6,63001	6,522277	-1,02	0,309	-17,3582	4,098182
2011 1	5,550279	5,905493	0,94	0,347	-4,16339	15,26395
2012 1	41,60149	5,535951	7,51	<0,1	32,49566	50,70732
2013 1	0,992696	5,844149	0,17	0,865	-8,62007	10,60547
2014 1	-4,97576	5,530239	-0,9	0,368	-14,0722	4,120676
year#FDIe						
2003	0,069802	0,3866192	0,18	0,857	-0,56613	0,705734
2004	-0,17805	0,5501538	-0,32	0,746	-1,08298	0,726868
2005	0,301953	0,3303019	0,91	0,361	-0,24135	0,845251
2006	0,346946	0,218106	1,59	0,112	-0,01181	0,705699
2007	1,291538	0,2146465	6,02	<0,1	0,938476	1,6446
2008	0,077934	0,1242333	0,63	0,53	-0,12641	0,28228
2009	0,04112	0,2285153	0,18	0,857	-0,33475	0,416994
2010	1,021384	0,3657809	2,79	<0,1	0,419728	1,62304
2011	-0,75804	0,154947	-4,89	<0,1	-1,0129	-0,50317
2012	-2,23168	0,1280739	-17,42	<0,1	-2,44234	-2,02102
2013	-0,18314	0,0477326	-3,84	<0,1	-0,26165	-0,10462

Çizelge 4.17(devam)

2014	0,330908	0,1650026	2,01	<0,1	0,059503	0,602313
2015	1,32344	0,2513158	5,27	<0,1	0,910063	1,736818
sabit terim	6,147994	5,179928	1,19	0,235	-2,37223	14,66822

Bu model ülkelerin ait olduğu kümelerin etkileşimli olarak ETPR6 değişkeni ile yıllık etkilerle iyi bir şekilde açıkladığını gösteriyor. Ülkeler arasındaki yabancı yatırım farkını iyi açıklamaktadır.

4.4 Dow-Jones Verisi ile Panel Veri Analizi ve Panel Veri Kümelemesi

SCI veritabanında Dow-Jones endeksini çalışan makalelerin sayısı 2019 yılı nisan ayı itibariyle 1066 tanedir. Yakın zamanda yapılan çalışmalar, bu veriden sonuç çıkartmak için ciddi çabaların olduğunu ortaya koydu.

Novotny and Urga (2018), anlamlı fiyat sıçrayışlarını modellemek için yeni bir metot önermişlerdir. Kullandıkları Dow-Jones verisi 5 dakikalık sıklıkta 1 ocak 2010 ile 30 haziran 2012 tarihleri arasında toplanmıştır. Geliştirdikleri metot, anlamlı fiyat sıçrayışlarını tanımlamakta başarılı olmuş ve gelecek çalışmalarda birbirleriyle ilişkili olan fiyat sıçramalarını modellemenin faydalı olacağını belirtmişlerdir.

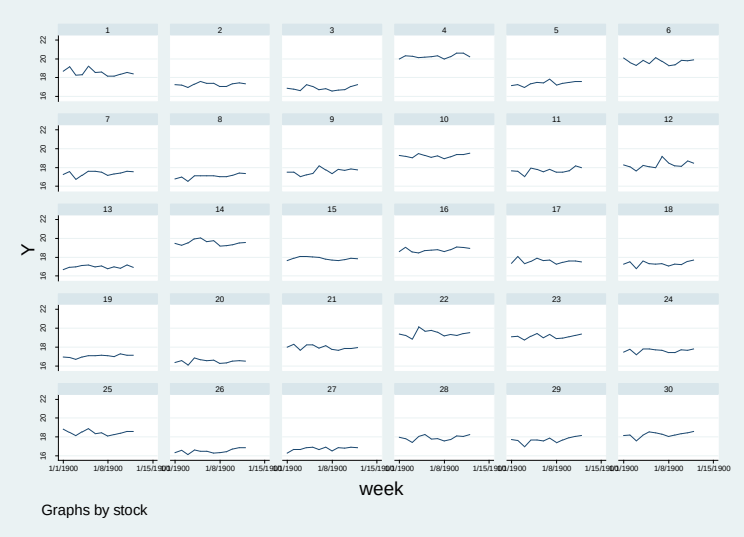
Bir başka çalışmada Eckernkemper (2018), marjinal beklenen geri düşmeleri kopula tabanlı modelle analiz etmeye çalışmıştır. Kullanılan copula dinamik karma bir kopula olup zamanla değişen doğrusal olmayan bağımlılığı modellemekte başarılı olmuştur. Bu model çeşitli varyasyonları düşünülmüş ve literatürdeki metotlarla bu metot kıyaslanmıştır. Bu kıyaslama sonucunda geliştirilen metot başarılı bulunmuştur. Veri, UCI veritabanından alınmıştır ve Brown (2013) tarafından kullanılmıştır. Veri, haftalık olarak ocak ayından mart'a ve nisan ayından haziran'a kadar olmak üzere iki bölümden oluşmaktadır Verideki kullanılan değişkenler quarter, week, stock, logvol ve Y olarak adlandırılmıştır. quarter değişkeni verinin hangi zaman aralığına ait olduğunu belirtmektedir. week değişkeni verideki gözlemin hangi tarihte alındığını belirtmektedir. stock değişkeni verideki haftanın doğal logaritması alınmış işlem hacmini belirtmektedir. Y değişkeni yanıt değişken olup, veride haftanın doğal logaritması alınmış işlem hacmini belirtmektedir ve açıklayıcı değişken, Y'nin 1 hafta önce

aldığı değerlerden oluşturulmuştur. d2, week5, week8, week10, week11, week12 değişkenleri sırayla 2. kümeye düşen hisse senetlerin gölge değişkeni, 5, hafta, 8, hafta, 10, hafta, 11, hafta ve 12, haftanın etkisini Model (5.14)'teki eşitlikle ifade edilmiştir:

$$Y = \beta_0 + \beta_1 \text{lagy2} + \beta_2 \text{logfarkvol} + \beta_3 \text{d2} + \beta_4 \text{week5} + \beta_5 \text{week8} + \beta_6 \text{week10} + \beta_7 \text{week11} + \beta_8 \text{week12} + u_{it} \quad (5.14)$$

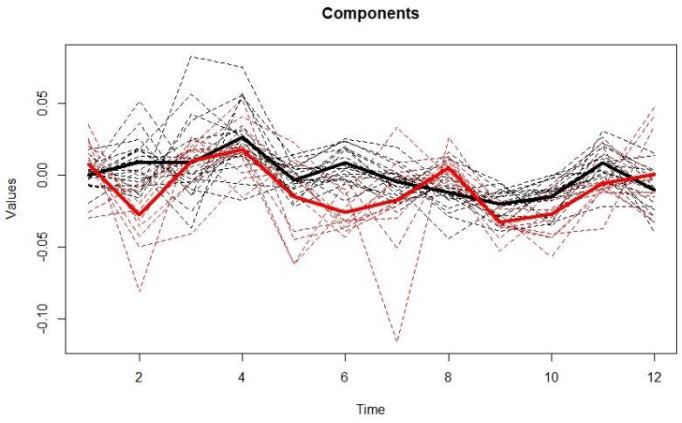
Bu zaman aralıklarına göre işlem hacimlerinin grafikleri Şekil 4.4 ile Şekil 4.5 gibidir:

Şekil 4.4: Ocak ayından mart'a kadarki dönemde işlem hacimlerinin grafikleri

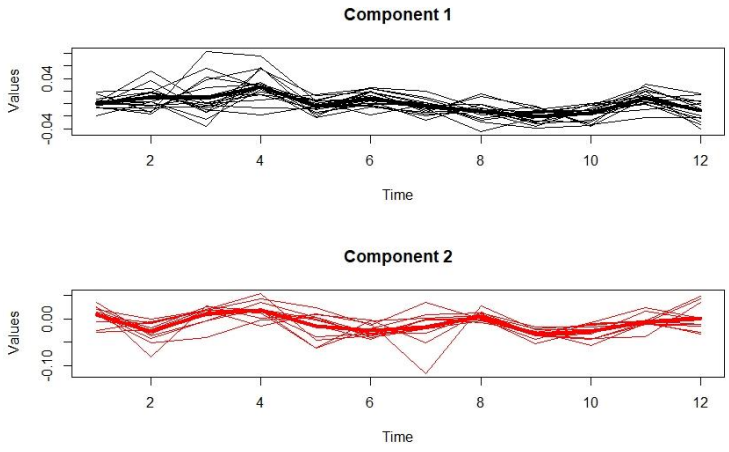


Şekil 4.5: Nisan ayından haziran'a kadarki dönemde işlem hacimlerinin grafikleri

Panel verideki kümelemeyi Mcnicholas (2017)'nin yöntemiyle gerçekleştirdik. Veriden elde edilen şekiller Şekil 4.6a ile Şekil 4.6b'teki gibidir:



Şekil 4.6a: Kümelerin birleştirilmiş şekli



Şekil 4.6b: Kümelerin bireysel şekilleri

Kümelere üyelik çizelgesi, Çizelge 4.18'deki gibidir:

Çizelge 4.18: Hisse senetlerinin kümeleri

hisse senedi	küme 1	küme 2	hisse senedi	küme 1	küme 2	hisse senedi	küme 1	küme 2
AA	2,05E-18	1,00E+06	IBM	1,00E+06	2,15E-05	T	1,00E+06	1,10E+01
AXP	3,06E-06	1,00E+06	INTC	1,00E+06	2,70E-14	TRV	1,00E+06	3,59E-02
BA	1,00E+06	7,86E+00	JNJ	1,00E+06	2,76E+01	UTX	1,00E+06	2,85E-02
BAC	4,99E-05	1,00E+06	JPM	7,38E+00	1,00E+06	VZ	1,00E+06	7,10E-03
CAT	9,77E-02	1,00E+06	KRFT	1,00E+06	7,44E-07	WMT	1,00E+06	4,34E-04
CSCO	1,57E-03	1,00E+06	KO	1,00E+06	2,35E-12	XOM	6,02E-12	1,00E+06
CVX	4,16E-09	1,00E+06	MCD	1,00E+06	3,53E-03			
DD	7,41E-07	1,00E+06	MMM	1,00E+06	4,29E+01			
DIS	2,57E-09	1,00E+06	MRK	1,00E+06	2,89E-09			
GE	1,00E+06	1,45E+00	MSFT	7,51E-01	1,00E+06			
HD	1,00E+06	1,10E-08	PFE	1,00E+06	1,25E+01			
HPQ	1,06E-23	1,00E+06	PG	1,00E+06	6,57E-09			

Elde ettiğimiz son model, rassal etkiler modeli olup Çizelge 4.19'daki gibidir:

Çizelge 4.19: Rassal etkiler modeli

Birimler $R^2 = 0{,}8309$ Birimler Arası $R^2 > 0{,}9999$ Genel $R^2 = 0{,}9987$						
	Katsayı	Standart hata	z-değeri	p-değeri	%90'lık güven aralığı	
lagy2	1,00357	0,00139	722,78	<0,1	1,00129	1,00586
logfarkvol	-0,0133	0,00264	-5,02	<0,1	-0,0176	-0,0089
d2	-0,0035	0,00193	-1,8	<0,1	-0,0067	-0,0003
week5	0,02526	0,00314	8,05	<0,1	0,0201	0,03042
week8	-0,0139	0,00318	-4,36	<0,1	-0,0191	-0,0086
week10	-0,0166	0,00311	-5,33	<0,1	-0,0217	-0,0114
week11	-0,0154	0,00319	-4,82	<0,1	-0,0206	-0,0101
week12	0,01292	0,0031	4,16	<0,1	0,00781	0,01802
sabit terim	-0,0131	0,00549	-2,38	<0,1	-0,0221	-0,0041

Bu sonuçlar ışığında kümeleme analizinin hisse senetleri arasındaki farklılıkları açıkladığını söyleyebiliriz.

5. BULGULAR VE TARTIŞMA

Uygulama yaptığımız gayri safi hasıla verisi haricindeki verilerde, kümeleme analizinden doğan gölge değişkenler modelleri geliştirmiştir. Bunun anlamı, kümeler verideki ölçümleri sağlayan süreçleri doğru olarak yansıtmış ve modelden elde ettiğimiz bilgiyi geliştirmiştir. Kümeleme analiziyle yapmayı esas hedef haline getirmediğimiz bir husus olarak geriye dönük tahminleri tutarlı bir şekilde açıklamak, bütün verilerin analizinde gerçekleştirilmiştir. Bu nedenle tezdeki bütün veri analizlerinin sonuçları, üzerinde bahsedilmeye değer konumdadır.

BİST verisinde kümeler çoğunlukla diğer değişkenlerle ilişkili bir biçimde etkilerini açığa çıkarmış ve p-değerleri değişken seçimlerini olumsuz anlamda etkilemiştir. p-değerleri 30 günlük tahminlerle 7 günlük tahminler arasındaki çatışmayı, 7 günlük tahminler lehine çevirse de gene de en güvenilir tahminler 30 günlük olanlar olarak kalmıştır. 2018 yılına kadar anlamlı sonuçlar veren küme gölge değişkenleri, 2015-2017 yılları arasında da önemini göstermiştir. 2015-2017 yılları arasında kümelerin oynadığı rolün önemi Çizelge 4.9'dan anlaşılabilir.

Gayrisafi hasıla verisi, geçmiş değerlere bağımlılık yüzünden başarısı kısıtlı bir veri olmuştur. Bu veride ülkeler arasında tutarlı kümeler doğal olarak vardı ve kümeleme analizi yapılarak elde edilen gölge değişkenler, ülkeler arasındaki farkı onayladı. Son elde edilen modelde, gölge değişkensiz modelin başarısı bütün gölge değişkenli modellerde tekrarlanmıştır ve bu kümeleme analizi ile panel veride regresyonun aynı sonuçlar doğurabileceğini göstermiştir.

Yabancı yatırım verisi, kümeleme analizinden elde edilen bilgilerin ilk modeli çok yönlü geliştirdiği veri olmuştur. Bu kümelerin değişkenlerle etkileşim içerisinde yabancı yatırımın GSYİH'ye oranını açıklamada çok faydası olmuştur. 2012 yılındaki oranların değişimi açısından anlamlı bir yıl olarak modelde ortaya çıkmıştır.

Dow-Jones verisi, haftalık şoklardan çok etkilenen bir veri olarak kümeleme analizi bulgularını olumsuz etkilemiştir. Bu nedenle, BİST verisinde uygulanan model aynı şekilde uygulanamamıştır. BİST verisinde uygulanan model büyük bilgi kaybına yol açmıştır. Benzeri uygulanan modelde verinin ilk kısmından kümeleme analizinden elde edilen küme gölge değişkenleri, bütün veri için geçerli sonuçları bulmuştur. Böylece kümeleme analizi başarıya ulaşmıştır.

6. SONUÇ VE ÖNERİLER

Verilerden elde ettiğimiz sonuçlarda, panel verinin kümelenmesinin çok büyük katkısı olmuştur. Panel veri kümelemesinin tutarlı sonuçlar üretmesi ve faydalı bilgi sağlaması kuşku götürmeyecek derecede önemli bir analiz olduğunu göstermiştir. Bunun yanında BİST verisinde kümelemenin yavaş çalışması, iyileştirilebilir gözükmektedir. İlerideki çalışmalarda veriye uygun sonuca daha hızlı gidecek kümeleme algoritmalarının geliştirilmesi önemli bir katkı sunar.

KAYNAKLAR DİZİNİ

Açıkgöz, İ., 2007, Sonlu karma dağılımlarda parametre tahmini, Ankara üniversitesi Fen bilimleri enstitüsü, s.152

Anderson, T.W., Hsiao, C., 1982, Formulation and estimation of dynamic models using panel data, Journal Of Econometrics, 18, 47-82,

Ai,C., Li, Q., Semi-parametric and non-parametric methods in panel data models, Matyas, L., Sevestre, P., 2008, The econometrics of panel data: fundamentals and recent developments in theory and practice (eds.), 451-478

Aitchison, J., Aitken, C.G.G., 1976, Multivariate binary discrimination by the kernel method, Biometrika, 63, 413-420,

Aitkin, M., Anderson, D., and Hinde, J., 1981, Statistical modelling of data on teaching styles (with discussion). Journal of the Royal Statistical Society B, 144, 419–461,

Akaike, H., 1973, Information theory and an extension of the maximum likelihood principle. In Second international symposium on information theory. 267–281,

Allman, E.S., Matias, C., Rhoades, J. A., 2009, Identifiability of parameters in latent structure models with many observed variables, The Annals of Statistics, 37, 3099-3132

Baltagi, B.H., Pesaran, M.H., 2007, Heterogeneity and crosssection dependence in panel data models: theory and applications, Journal of applied econometrics, 22, 229-232

Banfield, J.D., Raftery, A.E., 1993, Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, 803-821,

Baudry, J.P., Raftery, A.E., Celeux, G., Lo, K., Gottardo, R., 2010, Combining mixture components for clustering, Journal of computational and graphical statistics, 19, 332-353,

Bartholomew, D.J., 1959, Note on the measurement and prediction of labour turnover, Journal of Royal Statistical Society Series A, 122, 232-239

Behboodian, J., 1972, Information matrix for a mixture of two normal distributions, Journal of Statistical Computer simulation, 1, 295-314

Behboodian, J., 1970, On a mixture of normal distribuions, Biometrika, 57, 215-217,

Benoit, E., 1924, note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés a un systéme d'equations linéaires en nombre inférieur celui inconnues (Procédé du commandant Cholesky), Bulletin Géodésique, 2, 67-77,

KAYNAKLAR DİZİNİ(devam)

Bhargava, A., Sargan, J.D., 1983, Estimating dynamic random effects models from panel data covering short time periods, Econometrica, 51, 1635-1659

Bhattacharya, C.G., 1967, A simple method of resolution of a distribution into gaussian components, Biometrics, 23, 115-135

Biernacki, C., Celeux, G., and Govaert, G., 2003, Choosing starting values fort he EM algorithm for getting the highest likelihood in multivariate gaussian mixture models, Computational statistics and data analysis, 41, 561-575,

Bolt, J., Inklaar, R., Jong,H.D., Zanden, J.L.V., 2018, "Rebasing 'Maddison': new income comparisons and the shape of long-run economic development", Maddison Proje Raporu, Groningen Growth and Research Center.

Bouchard, G., Celeux, G., 2006, Selection of generative model in classification, IEEE transactions on Pattern Analysis and Machine Intelligence, 28, 544-564,

Bouveyron, C., 2014, Adaptive mixture discriminant analysis for supervised learning with unobserved classes, Journal of Classification, 31, 1, 49-84

Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E., 2019, Model-based clustering and classification for data science with applications in R, Cambridge series in statistical and probabilistic mathematics, 446 p.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B., 1994, The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. Annals of the Institute of Statistical Mathematics, 46, 2, 373–388,

Brown, M. S., Pelosi, M. & Dirska, H. (2013). Dynamic-radius Species-conserving Genetic Algorithm for the Financial Forecasting of Dow Jones Index Stocks. Machine Learning and Data Mining in Pattern Recognition, 7988, 27-41,

Carreira-Perpinan, M. A., Renals, S., 2000, Practical identifiability of finite mixtures of multivariate bernoulli distributions, Neural Computation, 12, 1, 141-152,

Celeux, G., Diebolt, G., 1985, The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly, 2, 73-82,

Celeux, G., Govaert, G., 1992, A classification EM algorithm and two stochastic versions. Computational Statistics and Data Analysis, 14, 315-332,

Celeux, G., Govaert, G., 1995, Gaussian parsimonious clustering models. Pattern Recognition, 28, 781-793,

KAYNAKLAR DİZİNİ(devam)

Celeux ,G. Maugis, C., Sedki, M., 2019, Variable selection in model-based clustering and discriminant analysis with a regularization approach, advances in Data Analysis and Classification,13, 1, 259-278

Chang, W.C., 1979, Confidence interval estimation and transformation of data in a mixture of two multivariate normal distributions with any given large dimension, Technometrics, 21, 351-355

Chui, C.K., 1992, An introduction to wavelets, San Diego, CA: Academic Press, 266 p.

Cox, D.R., 1972, The analysis of multivariate binary data, Journal of the royal statistical society C: Applied statistics, 21, 113-120

Cox, D.R., 1975, Partial likelihood, Biometrika, 62, 269-276,

Dasgupta, A., Raftery, A.E., 1998, Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American Statistical Association, 93, 294-302,

Deaton, A., 1985, Panel data, from time series of cross-sections, journal of econometrics, 30, 109-126

Eckernkemper, T. (2018). Modelling systemic risk: Time-Varying tail dependence when forecasting marginal expected shortfall. Journal of financial economics, 16, 1, 63-117,

Eisenberger, I., 1964, Genesis of bimodal distributions, Technometrics, 6, 357-363,

Erol, H., 1995, Sonlu karma dağılım modelleri (sürekli tip-tek değişkenli), Doktora tezi ,Çukurova Üniversitesi, Fen Bilimleri Enstitüsü, 210 s.

Everitt, B.S., 1981, A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions, Multivariate Behav. Res., 2, 171-180

Everitt, B., Hand, D.J., 1981, Finite mixture distributions, Chapman and Hall, 149 p.

Fowlkes, E.B., 1979, Some methods for studying the mixture of two normal(lognormal) distributions, Journal of American Statistical Association, 74, 561-575

Fraley, C., Raftery, A. E., 2006, MCLUST version 3 for R:Normal mixture modeling and model-based clustering. Teknik rapor 504, University of Washington, Department of Statistics. (unpublished)

Fraley, C., Raftery, A.E., 2007, Bayesian regularization for normal mixture estimation and model-based clustering, Teknik rapor 486, University of Washington, Department of Statistics. (unpublished)

KAYNAKLAR DİZİNİ(devam)

Friedman, H.P., Rubin, J., 1967, On some invariant criteria for grouping data. Journal of the American Statistical Association, 63, 1159-1178,

Frühwirth-Schnatter, S., 2006, Finite Mixture and Markov Switching Models. Springer Series in Statistics. New York: Springer, 514 s.

Frühwirth-Schnatter,S., 2011a,Dealing with label switching under model uncertainty. Mixtures:Estimation and Applications Mengersen, K.L., Robert, C., Titterington, D.M. (eds.) ,p. 213-240

Frühwirth-Schnatter,S., 2011b, Panel data analysis: a survey on model-based clustering of time series, Advances in Data Analysis and Classification, 5, 251-280,

Garcia-Escudero, L.A., Gordaliza, A., Matran, C., Mayo-Iscar, A., 2008, A general trimming approach to robust cluster analysis, Annals of Statistics, 36, 1324-1345,

Goodman, L.A., 1974, Exploratory latent structure models using both identifiable and unidentifiable models, Biometrika, 61, 215-231

Graybill, F.A., 1969, Introduction to matrices with applications in statistics, Belmont CA:Wadsworth, 372 s.

Gyllenberg, M., Koski, T., Reilink, E., Verlaan, M., 1994, Nonuniqueness in probabilistic numerical identification of bacteria, Journal of Applied Probability, 31, 2, 542-568

Harding, J.P., 1949, The use of probability paper for the graphical analysis of polymodal frequency distributions, Journal of the Marine Biological Association of the UK, 28, 141-153

Hasselblad, V., 1969, Estimation of finite mixtures of distributions from the exponential family, 64, 1459-1471

Hastie, T., Tibshirani, R., 1996, Discriminant analysis by gaussian mixtures,Journal of the royal statistical society series B, 155-176,

Hausman, J.A., 1978, Specification tests in econometrics, Econometrica, 46, 1251-1271

Hennig, C., 2010, Methods for merging gaussian mixture components, Advances in data analysis and classification, 4, 3-34,

Hill, B.M., 1963, Information for estimating the proportions in mixtures of exponential and normal distributions, Journal of American Statistical Association, 58, 918-932

Hoch, I.,1962, Estimation of production function parameters combining time-series and cross-sectional data, Econometrica, 69, 1645-1660

Holzmann, H., Schwaiger, F., 2016, Testing for the number of states in hidden Markov models. Computational Statistics and Data Analysis, 100, 318–330

KAYNAKLAR DİZİNİ(devam)

Hsiao, C., 2014, Analysis of panel data, cambridge university press, 564 s.

Ingrassia, S., Punzo, A., 2016, Decision boundaries for mixtures of regressions, Journal of the Korean Statistical Society, 45, 295–306,

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E., 1991, Adaptive mixture of local experts, Neural Computation, 3, 1, 79-87,

Jeffreys, H., 1961, Theory of Probability, Clarendon third edition, 459 p.

Kabir, A.B.M.L., 1968, Estimation of parameters of a finite mixture of distributions, Journal of Royal Statistical Society Series B, 30, 472-482

Kendall, M.G., Stuart, A., 1963, The Advanced Theory of Statistics III, Griffin London, 585 p.

Keribin, C., Brault, V., Celeux, G., Govaert, G., 2015, Estimation and selection for the latent block model on categorical data, Statistics And Computing, 25, 1201-1216,

Kim, B. S., 1984, Studies of multinomial mixture models, Doktora tezi, North Carolina Üniversitesi İstatistik enstitüsü, 157 p.

Kosmidis, I., Karlis, D., 2016, Model-based clustering using copulas with applications, Stat Comput, 26, 1079–1099

Kuh, E., 1963, Capital stock growth: A micro-econometric approach. Amsterdam: North-Holland, 341 s.

Law, M.H., Figueiredo, M.A.T., Jain, A. K., 2004, Simultaneous feature selection and clustering using mixture models, IEEE transactions on pattern analysis and machine intelligence,

Maitra, R., 2009, Initializing partition-optimization algorithms. IEEE/ACM transactions on computational biology and bioinformatics, 6, 144–157,

Maitra, R. and Melnykov, V., 2010, Assessing significance in finite mixture models. Tech. Rep. 10-01, Department of Statistics, Iowa State University.

Matyas, L., Sevestre, P., 2008, The econometrics of panel data: fundamentals and recent developments in theory and practice, Berlin:Springer-Verlag third edition, 980p.

Maugis, C.,Celeux, G., Martin-Magniette, M.-L., 2009, Variable selection in model-based clustering: a general variable role modeling, Computational Statistics And Data Analysis, 53, 3872-3882,

Maza, A., Villaverde, J., 2015, A new FDI potential index: Design and Application to the EU regions, European Planning Studies.

KAYNAKLAR DİZİNİ(devam)

McLachlan, G., 1987, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. Applied statistics, 36, 318–324,

Mclachlan, G., 1992, Discriminant analysis and statistical patterns recognition, John Wiley and Sons, 526 p.

McLachlan, G. and Peel, D., 2000, Finite Mixture Models. John Wiley and Sons, Inc., New York, 438 p.

McNicholas, P.D., 2017, Mixture model-based classification, CRC press, New York, 240 p.

McNicholas, P.D. and Murphy, T.B., 2010, Model-based clustering of longitudinal data. The Canadian Journal of Statistics, 38, 1, 153-168,

Melnykov, V., Maitra, R., 2017, Finite mixture models and model-based clustering. Statistics Surveys, 4, 80-116,

Miller, D., Browning, J., 2003, A mixture model and em-based algorithm for class discovery, robust classification and outlier rejection in mixed labeled/unlabeled data sets. IEEE transactions on pattern analysis and machine intelligence, 11, 25, 1468-1483,

Mundlak, Y., 1961, Empirical production free of management bias, journal of farm economics, 43, 44-56

Murphy, E.A., 1964, One Cause? Many Causes? The argument from the bimodal distribution, Journal of Chron. Dis., 17, 301-324,

Murtagh, F. and Raftery, A.E., 1984, Fitting straight lines to point patterns. Pattern Recognition, 17, 479-483,

Nerlove, M., 1965, Estimation and identification of Cobb-Douglas Production functions, Chicago: Rand McNally, s.193

Nerlove, M., 1971, A note on error components models, Econometrica, 50, 703-708

Newcomb, S., 1886, A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics, 8, 343–366,

Novotny, J., Urga, G., 2018, Testing for Co-jumps in Financial Markets, Journal Of Financial Econometrics, 16, 1, 118-128,

Pamminger, C., 2007, Bayesian Clustering of Categorical Time Series Using Finite Mixtures of Markov Chain Models, Doktora Tezi, Johannes Kepler University.

Pearson, K., 1894, Contribution to the mathematical theory of evolution. Philosophical Transactions of the Royal Society, 185, 71–110,

KAYNAKLAR DİZİNİ(devam)

Posse, C., 2001, Hierarchical model-based clustering for large data sets. Journal of Computational and Graphical Statistics, 10, 464-486,

Pourahmadi, M., 1999, Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. Biometrika, 86, 3, 425-435,

Punzo, A., Ingrassia, S., 2016, Clustering bivariate mixed-type data via the cluster-weighted model. Computational Statistics, 31, 989–1013

Raftery, A.E., Dean, N., 2006, Variable selection for model-based clustering, Journal Of The American Statistical Association, 101, 168-178,

Ray, S. and Lindsay, B., 2008, Model selection in high dimensions: a quadratic-risk-based approach. Journal of Royal Statistical Society (B), 70, 95–118,

Redner, R. A., Walker, H.F., 1984, Mixture densities, maximum likelihood and the EM algorithm, SIAM review, 26, 195-239,

Roeder, K. and Wasserman, L., 1997, Practical bayesian density estimation using mixtures of normals. Journal of the American Statistical Association, 92, 894-902,

Schwarz, G., 1978, Estimating the dimensions of a model. Annals Of Statistics, 6, 461–464,

Stock, J., Watson, M.W., 2008, Heteroskedasticity-robust standard errors for fixed effects regression, Econometrica, 76, 155-174,

Tadesse, M.G., Sha, N., Vannucci, M., 2005, Bayesian variable selection in clustering high-dimensional data, Journal Of American Statistical Association, 100, 602-617,

Tsiatis, A.A., 1981, A large sample study of Cox's regression model, The Annals Of Statistics, 9, 93-108

Xiang, S., Yao, W., Seo, B., 2016, Semiparametric mixture: Continuous scale mixture approach, Computational Statistics And Data Analysis, 103, 413–425

Wallace, T.D., Hussain, A., 1969, The use of error-components models in combining cross-section with time-series data, Econometrica, 37, 55-72,

Wansbeek, T.J., Kapteyn, A., 1978, The separation of individual variation and systematic change in the analysis of panel data, Annales De I'nsee, 30-31, 659-680,

Windham, M. P. and Cutler, A., 1992, Information ratios for validating mixture analyses. Journal of the American Statistical Association, 87, 1188–1192,

Wilks, S.S., 1938, The large sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat., 9, 60-62

KAYNAKLAR DİZİNİ(devam)

Wolfe, J.H., 1963, Object cluster analysis of social areas, Master tezi, University of California, Berkeley, 3, p.73

Wolfe, J.H., 1965, A computer program for maximum-likelihood analysis of types, USNPRA teknik rapor. (unpublished)

Wolfe, J.H., 1970, Pattern clustering by multivariate mixture analysis. Multivariate Behavioural Research, 5, 329-350,

Wolfe, J.H., 1971, A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions, Naval Personnel and Training Research Laboratory, 15 p. (unpublished)

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L., 2001, Model-based clustering and data transformations for gene expression data. Bioinformatics, 17, 977-987,

Zhou, H., Pan, W., Shen, X., 2009, Penalized model-based clustering with unconstrained covariance matrices, Electronic Journal Of Statistics, 3, 1473-1496