T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

AÇIKLAYICI VERİ ÇÖZÜMLEMESİNDE İZDÜŞÜM ARAMA YÖNTEMİ

YÜKSEK LİSANS TEZİ
Alev BAKIR

İstatistik Anabilim Dalı
İstatistik Programı

Tez Danışmanı: Prof. Dr. Aydın ERAR

TEMMUZ, 2009

T.C.
MİMAR SİNAN GÜZEL SANATLAR ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

AÇIKLAYICI VERİ ÇÖZÜMLEMESİNDE İZDÜŞÜM ARAMA YÖNTEMİ

YÜKSEK LİSANS TEZİ

Alev BAKIR

İstatistik Anabilim Dalı
İstatistik Programı

Tez Danışmanı: Prof. Dr. Aydın ERAR

TEMMUZ, 2009

Alev BAKIR tarafından hazırlanan AÇIKLAYICI VERİ ÇÖZÜMLEMESİNDE İZDÜŞÜM ARAMA YÖNTEMİ adlı bu tezin ..... tezi olarak uygun olduğunu onaylarım.

.....

Tez Yöneticisi

Bu çalışma, jürimiz tarafından ..... Anabilim Dalında

..... tezi olarak kabul edilmiştir.

Başkan: : _____

Üye : _____

Üye : _____

Üye : _____

Üye : _____

Bu tez, Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygundur.

ÖZET

Bu çalışmanın amacı, çok boyutlu verilerin bir eksen ya da bir düzlem üzerindeki doğrusal izdüşümünü, tekrarlı bir arama işlemi ve bir endeks kriteri kullanarak bulan açıklayıcı veri çözümlemesi yöntemini incelenmektir. Bu yöntem ilk kez Friedman ve Tukey tarafından 1974 yılında İzdüşüm Arama (Projection Pursuit) Yöntemi olarak adlandırılmıştır.

Dört bölümden oluşan bu çalışmanın birinci bölümünde, neden izdüşüm arama yönteminin çok boyutlu veri kümelerinde diğer yöntemlere göre tercih edildiği; böyle bir yönteme neden gereksinim duyulduğu ve izdüşüm arama yönteminin tarihçesi yer almaktadır.

İkinci bölümde, doğrusal modelleme problemlerinde sıkça kullanılan genel tanımlamalar ve izdüşüm ile ilgili özet bir bilgi, izdüşüm arama yönteminin algoritma ve endeks hesabının gelişim süreci, farklı izdüşüm arama algoritmaları ve adımları, yönteme ait çeşitli endeks açıklamaları yer almıştır.

Üçüncü bölüm uygulama kısmından oluşup, izdüşüm arama yönteminin, gözlem ve değişken sayısının az ve çok olduğu, aykırı değerin ve çoklu bağlantının olup olmadığı durumlardaki veri kümesi, değişkenlerin dönüşümlü ve dönüşümsüz değerleri ile yapılan uygulamalarının karşılaştırmalı sonuçlarını içermektedir.

Dördüncü bölümde ise sonuç ve tartışma kısımlarına yer verilmiştir.

Anahtar Kelimeler: İzdüşüm arama yöntemi, izdüşüm arama algoritmaları, izdüşüm arama endeksleri, boyut indirgeme, yapısal uzaklaştırma.

SUMMARY

The aim of this study is a technique for the exploratory analysis of multivariate data sets, the method seeks out linear projection of the multivariate data onto a line or a plane. Projection Pursuit refers to a technique first described by Friedman and Tukey (1974).

In the first chapter of the study consisting of five chapters, we explain why prefer to projection pursuit technique instead of other reducing dimensional techniques and why we need this technique. Also, some histories about projection pursuit techniques are also included.

In the second chapter, the general definition of linear modelling, some information about projection, algorithms of projection, steps of projection algorithm, different kinds of projection pursuit indexes, comparisons of some combination of algorithm and projection index are introduced.

In the third chapter, we have the outcome of projection pursuit techniques from two different data sets.

Finally in the fourth chapter, conclusions and discussion section are presented.

Keywords: Projection pursuit technique, projection pursuit algorithms, projection pursuit index, reducing dimensional, structure removal.

TEŞEKKÜR

Bu çalışmaya başlarken verdiği motivasyondan ve yardımlarından dolayı, öğretimde esas olan olguları benimsediği ve öğrencisine değer verdiği için sayın danışmanım Prof. Dr. Aydın ERAR' a, her türlü destek ve yapıcı eleştirilerinden dolayı Prof. Dr. Gülay KIROĞLU' na, yardım ve önerilerinden dolayı Prof. Dr. Nalan CİNEMRE' ye, bu bölümde bu çalışmamı tamamlamam için ilk destekçim olan Yrd. Doç. Funda SEZGİN' e sonsuz şükranlarımı ve saygılarımı sunar, beni her zaman destekleyen AİLEME, ARKADAŞLARIMA ve yardımlarından dolayı tüm bölüm HOCALARIMA teşekkürlerimi sunarım.

İÇİNDEKİLER DİZİNİ

ŞEKİLLER DİZİNİ

TABLO DİZİNİ

BİRİNCİ BÖLÜM

GİRİŞ

Veri analizinde, çok boyutlu verileri boyut indirgeyerek düşük boyutlu veriler şeklinde grafiklemek sıkça kullanılan bir tekniktir. Bununla birlikte çok boyutlu veri analizlerinde bazı hesaplama problemleri ile karşılaşılmakta; aykırı değerlerin varlığı, çoklubağlantının (multicolinearity) bulunması gibi sorunlar da istatistiksel analizleri etkilemekte yanlış tahminleri beraberinde getirmektedir (Kıral ve Billor, 2005). Bu problemlerin ortadan kaldırılmasında veri boyutunun indirgenmesi önemli bir rol oynamaktadır.

Çok boyutlu veri kümelerinde boyut indirgemek adına çok sık başvurulan Temel Bileşenler Analizinde klasik varyans-kovaryans matrisinin kullanılması, aykırı değerlerin varlığından etkilenerek yanlış sonuçlara götürebilir (Kıral ve Billor, 2001). Bunun için önerilen sağlam (robust) çözümler ise çok boyutlu veri kümeleri karşısında etkinliğini kaybetmektedir. Bu nedenle 1970 yılında Switzer tarafından ilk çalışmalar yapılmış, Friedman ve Tukey tarafından ilk olarak yöntemin adı İzdüşüm Arama Yöntemi (Projection Pursuit) olarak önerilmiş ve geliştirilmiştir. Çalışmalarda veri kümesinin boyutunu indirgemek için önerilen bu yöntem çok boyutlu verilerde aykırı değerlere karşı avantaj sağlanmış durumdadır (Friedman ve Tukey, 1974).

İzdüşüm Arama Yöntemi, çok boyutlu verilerde bir doğrusal grafikleme biçimidir. Yöntem, yüksek boyutlu olan orijinal veriyi daha düşük boyutlu (bir ya da iki boyutlu) bir uzaya taşıyarak verinin saçılım dağılımı yardımıyla verilerdeki örüntüyü (pattern) keşfetmeyi ve yorumlamayı amaçlar.

Klasik çok boyutlu analiz yöntemlerinden çoğunun izdüşüm arama yönteminin özel bir durumu olduğu birkaç yazarca belirtilmiştir. Örneğin Temel Bileşenler Analizi, Kümeleme Analizi ve Faktör Analizindeki Quartimax Oblimax metotları (Huber, 1985). Bu yöntem, tek boyutlu veri kümelerinden yüksek boyutlu veri kümelerine kadar uygulanabilen güçlü bir grafikleme yöntemi olarak karşımıza çıkmaktadır.

Bu konuda birkaç durağan yılın ardından Friedman ve Stuetzle, İzdüşüm Arama fikrini, 1980 yılında İzdüşüm Arama Sınıflandırması' nı, 1981 yılında İzdüşüm Arama Regresyon' unu ve 1984 yılında da İzdüşüm Arama Yoğunluk Tahmini şeklinde geliştirmişlerdir.

Yöntemin tek dezavantajı, uzun hesaplama süresi gerektirmesidir. Ancak yüksek hızdaki bilgisayarlar ile izdüşüm arama yönteminin uygulanabilirliği daha kolay bir duruma geldi ve bu yöntem çeşitli uygulamalar ve bazı yöntemlerle birleştirilerek geliştirildi (Huber,1985).

İKİNCİ BÖLÜM

GENEL BİLGİLER

2.1. İZDÜŞÜM (PROJECTION)

Doğrusal modeller için altküme ve izdüşüm oldukça önemli tanımlamalardır. Bir noktanın, birbirine dik ve kesiştikleri noktayı başlangıç noktası olarak alan iki eksen çiftine Kartezyen Sistemi ya da Dik Koordinat Sistemi de denir; bunun yeri bir sayı çifti (x,y) ile ifade edilebilir. Bazen bu noktayı, düzlem üzerinde orijinle birleştiren bir doğru (r) ve yatay eksenle arasındaki açıyı (θ) vererek de belirtebiliriz. Bu şekilde bir tanımlamaya da Kutupsal (Polar) Koordinat Sistemi denir. Dairesel hareket yapan nokta çiftlerini yarıçap değişmeyeceğinden bu şekilde belirtmek daha kolaydır. Bir noktanın kutupsal koordinatlardaki konumu (r,θ) biliniyor ise bu noktanın dik koordinatlar üzerindeki izdüşümleri (x,y),

$$x = r.\cos\theta, \qquad y = r.\sin\theta \tag{2.1}$$

biçiminde bulunabilir.

Vektörel hesaplamalarda ise örneğin bir **y** vektörünün **x** vektörü üzerindeki izdüşümü $\hat{y}$ şeklinde gösterilirse, b bir sabit olmak üzere, $\hat{\mathbf{y}} = b\mathbf{x}$ olur; burada, $(\mathbf{y} - \hat{\mathbf{y}}) \perp \mathbf{x}$ 'dir (Stapleton, 1995).

$f(\mathbf{y} \mid \mathbf{x})$, bir olasılığı değil izdüşümü ifade eder, yani **x** vektörü üzerindeki **y** vektörünün izdüşümü olan $\hat{\mathbf{y}} = f(\mathbf{y} \mid \mathbf{x})$ şeklindedir. $\hat{\mathbf{y}} = b\mathbf{x}$ eşitliğinde, b sabitlerini bulabilmek için,

$$\langle \hat{\mathbf{y}}, \mathbf{x} \rangle = \langle b\mathbf{x}, \mathbf{x} \rangle = b\langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle \tag{2.2}$$

eşitliğinden **x**' in sıfır dışındaki değerleri için, $b = \langle \mathbf{y}, \mathbf{x} \rangle / \|\mathbf{x}\|^2$ kullanılır. Buradan, $\hat{\mathbf{y}} = \left(\langle \mathbf{y}, \mathbf{x} \rangle / \|\mathbf{x}\|^2 \right)\mathbf{x}$ şeklinde bulunur. Eğer $\mathbf{x} = 0$ ise $b = 0$ olur (Stapleton, 1995).

Burada, **x** vektörünün uzunluğunun ya da büyüklüğünün ölçüsüne vektörün normu denir ve $\|\mathbf{x}\|$ şeklinde gösterilir. Çeşitli norm hesaplama yöntemleri vardır. En yaygın olarak bilineni Öklid (veya l_2) normudur ve $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ biçiminde hesaplanır (Bronson,1989).

Burada, $\langle \mathbf{y}, \mathbf{x} \rangle$ iç çarpımı ifade etmektedir ve $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}.\bar{y}$ şeklinde olup $\bar{y}$, **y**' nin kompleks eşleniğini (z=a+bi ise kompleks eşleniği $\bar{z}$ = a-bi' dir) belirtir. $\langle \mathbf{x}, \mathbf{y} \rangle_w$=(w**x**).$(\overline{\mathbf{wy}})$, **x** ve **y**' nin w tekil olmayan alt matrisine göre iç çarpımlarıdır (Bronson, 1989).

En genel durumda l_p normu, aşağıdaki biçimde gösterilir;

l_p normu, (p≥1): $\|\mathbf{x}\|_p = [|x_1|^p + |x_2|^p + \cdots + |x_n|^p]^{1/p}$.

Norm koşulları ise,

(N1): $\|\mathbf{x}\| \geq 0$

(N2): Ancak ve ancak **x** =0 ise $\|\mathbf{x}\| = 0$ olur.

(N3): Bir c sabiti için $\|c\mathbf{x}\| = |c|\,\|\mathbf{x}\|$' dır.

(N4): $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

biçiminde verilir. Burada x ve y aynı boyuttadır (Bronson,1989).

Eğer Ω uzayının bir n boyutlu alt uzayında herhangi bir k vektörler grubu ile çalışılacaksa bu vektörler doğrusal bağımsız olmalıdır. n boyutlu alt vektör uzayını **V** olarak düşünürsek alt kümesindeki tüm vektörler ($\mathbf{v_1},\ldots,\mathbf{v_k}$) de aralarında doğrusal bağımsız vektörler olmalıdır. Herhangi bir vektör diğer bir vektöre dönüştürülebiliyorsa bu alt uzayın doğrusal bağımlı olduğunu gösterir. Bir **y** vektörü **V**' deki tüm vektörlere dik ise **V** alt uzayına da diktir denir ve $\mathbf{y} \perp \mathbf{V}$ şeklinde gösterilir. $\mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}$ artıklar vektörü olmak üzere $(\mathbf{y} - \hat{\mathbf{y}}) \perp \mathbf{V}$ 'dir. Örneklem uzayında **V** alt uzayının, $\mathbf{v_1},\ldots,\mathbf{v_k}$ birbirlerine de dik olan vektörleri için,

$$(\mathbf{y} \mid \mathbf{V}) = \sum_{i=1}^{k} f(\mathbf{y} \mid \mathbf{v}_i) \qquad (2.3)$$

şeklindedir ve bu eşitlik yalnızca $\mathbf{v_1},\ldots,\mathbf{v_k}$ vektörleri birbirlerine dik ise tüm **y** 'ler için doğrudur (Stapleton, 1995).

Doğrusal bağımsız vektörler kümesini ($x_1,x_2,\ldots,x_n$) dik vektörler kümesine dönüştürme işlemi için aşağıda verilen Gram-Schmidt dikleştirme yöntemi kullanılabilir:

Yöntem Algoritmasının Adımları:

ADIM1: j=1 için

$$Q_1 = \frac{1}{\sqrt{\langle x_1, x_1 \rangle_w}} x_1 \text{ alınır.} \quad (2.4)$$

ADIM2: j=n ise durulur. Değilse j, 1 arttırılarak devam edilir.

ADIM3: $y_j = x_j - \sum_{i=1}^{j-1} \langle x_i, Q_i \rangle_w Q_i$ hesaplanır.

ADIM4: $Q_j = \frac{1}{\sqrt{\langle y_j, y_j \rangle_w}} y_j$ alınır. (2.5)

ADIM5: 2. adıma dönülür (Bronson, 1989).

Dikleştirme için bu algoritmanın yuvarlama hatalarına karşı daha duyarlı olduğu QR uygulaması da kullanılabilir (Erar, 2007).

Dik matrisler kümesi düzlem döndürme matrisleri olarak da tanımlanır. Döndürme işlemi n-boyutlu uzayda bir noktanın koordinatlarını değiştirmeyi amaçlar. İlk sistemin i. ve j. eksenleri orijin çevresinde bir θ açısı ile döndürülerek bir kartezyen sistemi ortaya çıksın; bu sistemde **A** matrisinden **B** matrisine dönüşüm yapılabilir. **P** ve **Q** matrisleri tekil olmayan matrisler olmak üzere **B** = **PAQ** eşitliği ile açıklanabilecek bazı dönüşümler aşağıdadır:

1- $\mathbf{P}=\mathbf{Q}^{-1}$ ise $\mathbf{B}=\mathbf{Q}^{-1}\mathbf{AQ}$ 'dir (benzerlik dönüşümü).

2- $\mathbf{P}=\mathbf{Q}'$ ise $\mathbf{B}=\mathbf{Q}'\mathbf{AQ}$ 'dir (eşleşim (congruence) dönüşümü).

3- $\mathbf{P}=\mathbf{Q}'=\mathbf{Q}^{-1}$ ise (dik matris) $\mathbf{B}=\mathbf{Q}'\mathbf{AQ}=\mathbf{Q}^{-1}\mathbf{AQ}$ olur. Bu dönüşüme dik dönüşüm denir. Böylece **B**, **A** 'ya hem dik olarak eşdeğer hem de dik olarak eşleşimlidir (Erar, 2007).

Öte yandan, Jacobi yöntemin olarak da bilinen benzerlik dönüşümünden,

$$\mathbf{Q} = \begin{bmatrix} \cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \quad (2.6)$$

döndürme matrisi kullanılarak θ açısına göre Kartezyen düzleminin saat yönünde dönüşü gerçekleştirilebilir. Böylece **Q'AQ** den döndürülmüş matris olan **B** bulunur (Erar, 2007).

Örneklem uzayına ait bir **y** vektörünün, alt uzay **V** üzerindeki izdüşümü $\hat{\mathbf{y}}$ ise $\mathbf{P}_V: \mathbf{y} \rightarrow \hat{\mathbf{y}}$ şeklinde yazılabilir. Kestirilecek olan **y** vektörü için **y** = **θ** + **ε** şeklinde bir model tanımlanmaktadır ve burada **θ**∈**V,** bilinmemektedir; **θ** kestirilecektir. **ε** ise

rastlantı vektördür. Bu dönüşüm doğrusaldır. **P** izdüşüm matrisi, **V** doğrusal bağımsız $\mathbf{x_1},\ldots,\mathbf{x_k}$ sütun vektörleri iken,

$$f(\mathbf{y} \mid \mathbf{V}) = \mathbf{X}b = \mathbf{X(X'X)^{-1}X'y} \qquad (2.7)$$

eşitliğinden $\mathbf{P}_v=\mathbf{X(X'X)^{-1}X'}$ şeklinde bir "izdüşüm operatör"dür. Burada $\mathbf{P_v}$ simetrik $[(\mathbf{P}_x,\mathbf{y})=(\mathbf{P}_x,\mathbf{P}_y)=(\mathbf{x},\mathbf{P}_y)]$ ve idempotenttir ($\mathbf{P}^2=\mathbf{P}$) (Stapleton, 1995).

Çok boyutlu analizlerde rastlantı değişkenlerinin izdüşümleri, **X**, **Y** için iç çarpımları $\langle \mathbf{X}, \mathbf{Y} \rangle$ = E(**XY**), $\|\mathbf{X}\|^2$ = E($\mathbf{X}^2$) = VAR(**X**) şeklindedir. **V** rastlantı değişkenler uzayında bir alt uzay olan ($\mathbf{X_1}$, …,$\mathbf{X_k}$) ve **Y**'nin **V** alt uzayındaki izdüşümü $\hat{\mathbf{Y}}$ (rastlantı değişken) vektörü, **V** uzayında tüm **X**ϵ**V** için (**Y**- $\hat{\mathbf{Y}}$, **X**) = 0 şeklindedir. Ayrıca tüm **X**ϵ**V** için E[**X**(**Y**- $\hat{\mathbf{Y}}$)] = cov(**x**, **Y**- $\hat{\mathbf{Y}}$) = 0' dır.

Yukarıdaki kesitlerde belirtildiği gibi $\hat{y}$ = b**x** burada b=$\sum \mathbf{X}^{-1}\mathbf{U}$, $\sum \mathbf{X} = D(\mathbf{X})$, **U**=C[**X**,**Y**] $\sum \mathbf{X}$ tekil değildir ve Var($\hat{\mathbf{Y}}$) = $\|\hat{\mathbf{Y}}\|^2$, Var($\hat{Y}$) =b'$\sum \mathbf{X}$ b=**U**'$\sum \mathbf{X}^{-1}\mathbf{U}$ şeklindedir. Artıklar vektörü **e** = **Y**- $\hat{\mathbf{Y}}$ ' dır; buradan **Y** = **e** + $\hat{\mathbf{Y}}$ olur; tüm **u**ϵ**V** için cov(**U**,**e**) = 0 ve cov(, **e**)=0' dır. Var(**Y**) = Var(**e**) + Var($\hat{\mathbf{Y}}$)
=$(\sigma_Y^2 - \mathbf{U}'\sum \mathbf{X}^{-1}\mathbf{U}) + \mathbf{U}'\sum \mathbf{X}^{-1}\mathbf{U}$ olur.

$\hat{\mathbf{Y}}$, **Y**' nin en iyi doğrusal kestirici ise, $\sum b_j X_j$ doğrusal birleşim ve **Y**- $\hat{\mathbf{Y}}$=**e**, mümkün olan en küçük varyansa sahiptir. Herhangi bir **W**ϵ**V** için $\|\mathbf{Y} - \mathbf{W}\|^2$ = Var(**Y** – **W**) ≥ Var(**Y** - $\hat{\mathbf{Y}}$) = $\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2$ olur.

2.2. İZDÜŞÜM ARAMA KAVRAMI

Friedman (1987) izdüşüm arama algoritmasının, en uygun izdüşümü aramak için varyans gibi noktalar arası uzaklığı kullanan bir doğrusal dönüşüm (mapping) algoritması önermiştir. Doğrusal olmayan algoritmaların yüksek boyutlu nokta kümelerinin sunumunda doğrusal yöntemlere göre daha güvenilir olduğu; ancak sonucun grafiklenmesinin daha zor olduğu, birkaç parametre ile özetlenemeyeceği, hesaplamanın çok zaman alacağı gibi kusurları olduğunu dile getirmiştir.

İzdüşüm Arama Yöntemi, çok boyutlu verilerde bir doğrusal grafikleme biçimidir. Yöntem, yüksek boyutlu olan orijinal veriyi daha düşük boyutlu (bir ya da iki boyutlu) bir uzaya taşıyarak verinin saçılım dağılımı yardımıyla verinin yapısını keşfetmeyi ve yorumlamayı amaçlar.

Friedman ve Tukey 1974 yılında İzdüşüm Arama adını verdikleri bir teknik ile doğrusal olmayan yüksek boyutlu veri setlerini düşük boyutlu doğrusal yapılar ile açıklamaya çalışmışlardır. 1987 yılında Friedman istatistiksel doğruluk ve hesaplama kolaylığı açısından algoritmayı geliştirmiştir. Literatürde izdüşüm arama endeksini hesaplamaya ve algoritmaya ilişkin çok fazla seçenek vardır. Bunlardan Friedman(1987)'ın L^2-Uzaklığı Endeksi, Jones ve Sibson(1987)'ın Moment ve Entropi Endeksleri, Posse (1990)' un Ki-Kare Uzaklığı Endeksi en bilinen ve farklı türden algoritma yaklaşımları ile tamamlanmış izdüşüm arama yöntemleridir (Posse, 1995b). En geliştirilmiş, kolaylıkla hesaplanabilen ve bazı temel varsayımların sağlandığı izdüşüm arama yöntemi, Posse' un 1990 yılında önerdiği Ki-Kare Endeksinin yine kendi tarafından 1995 yılında Friedman' ın yapısal uzaklaştırma yöntemi ile birleştirerek değiştirdiği yöntemdir; bu yönteme aşağıda genişçe yer verilmiştir.

Eğer düşük boyutlu veri kümelerinde, bazı sınıflandırılmamış ya da beklenmeyen yapının bulunup bulunmadığını kontrol etmek istersek o zaman en basit ve etkili yaklaşım, veri kümesinin grafiğini çizmektir; tek ya da iki boyutlu verilerde histogram, iki ya da üç boyutlularda ise saçılım dağılım grafikleri ile veri ile ilgili bir yoruma gidilebilir.

Renk ve kabartı tekniği ile grafiklere bir kaç boyut daha eklemek olanaklıdır; ancak üç boyutludan daha fazla boyutlularda, eş zamanlı olarak tüm boyutları gösterebilen bir grafik çizimi yapılamamaktadır.

Çok boyutluluğun zor bir yanı da kodlanamamasıdır. Üç uzaylı boyutlarda, iki uzay ve bir zaman boyutu şeklinde dönüştürme kolaylık sağlar ve iki uzay bir renk boyutu şeklinde de bir dönüştürme uygulanabilir. Dört ya da daha çok boyutlularda işlem, dördüncü boyutun renkler ile kodlanmasıdır ki verilerin herhangi bir yönden bu yöntem için sınıflandırılabiliyor olması ile iyi bir çözüm sunulabilir (Huber,1985).

Genellikle tercih edilen bir yol bir, iki ya da üç boyutlu uzaydaki izdüşümlere bakmaktır. Ancak veri kümesinin boyutu yüksek ise (örneğin 10'dan daha fazla) teker teker her iki ya da üç boyutlu saçılım grafiklerini incelemek sabır ister ve zaman kaybına neden olur. Çok boyutluluğun bu dezavantajı İzdüşüm Arama

Yöntemi ile giderilmeye çalışılmıştır. Böylece eğer 4 ve daha fazla boyutlu veri kümelerinde anlaşılır ve iyi sonuçlar edinmek istenirse öncelikle boyut indirgenir; sonra indirgenen boyut üzerinden yorumlamalara gidilir.

Bu yöntemin temel iki adımı vardır:

1- Yapının ya da ayrılışın derecesinin ölçüsü olan bir izdüşüm arama endeksi bulmak,

2- Endeksin en büyük değerini verecek olan en iyi düzlemi bulmak.

2.3. İZDÜŞÜM ARAMA YÖNTEMİ

2.3.1. İzdüşüm Algoritmaları

İzdüşüm arama yöntemi, genelde orijinal verilerin standartlaştırılmış biçimi ile çalışır (Martinez W. ve Martinez A., 2005). n bağımsız gözlemli ve k değişkeni olan bir veri setinde, **X**, veri matrisi (nxk); **1**, nx1 boyutlu 1'ler vektörü; verilerin standartlaştırılmış matrisi **Z** ,

$$\mathbf{Z} = \mathbf{\Lambda}^{-1/2}\,\mathbf{Q}^{T}\,(\mathbf{X}-\mathbf{1}.\overline{\mathbf{X}})^{T} \qquad (2.8)$$

şeklinde bulunur. Burada, kovaryans matrisinden elde edilen özvektör matrisi **Q** ve özdeğerler **Λ** ile belirtilmiştir. Burada, $\overline{\boldsymbol{X}}$ örnek ortalaması (1xk),

$$\overline{X_j} = \sum_{i=1}^{n} X_{ij}/n \qquad j=1,\ldots,k \qquad (2.9)$$

$\hat{\boldsymbol{\Sigma}}$ örneklemin kovaryans matrisi (kxk),

$$\widehat{\Sigma_{ij}} = \sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ij}-\overline{X}_j)^{T}/(n-1) \qquad j=1,\ldots,k \qquad (2.10)$$

olarak verilir.

Öte yandan **α** ve **β**, ortonormal özellikli ($\boldsymbol{\alpha}^{T}\boldsymbol{\alpha}=\boldsymbol{\beta}^{T}\boldsymbol{\beta}=1$ ve $\boldsymbol{\alpha}^{T}\boldsymbol{\beta}=0$) ve k boyutlu izdüşümün gösterileceği, genelde başlangıç değerleri rastgele seçilen, vektörler olsun. P(α,β), **α** ve **β** düzlemi üzerindeki izdüşüm çizim iken, z_i^{α}, z_i^{β} standartlaştırılmış gözlemlerin **α** ve **β** vektörleri üzerindeki karşılıklarıdır ve

$$z_i^{\alpha} = z_i^{T}\,\alpha \qquad \text{ve} \qquad z_i^{\beta} = z_i^{T}\,\beta \qquad i=1,\ldots,n \qquad (2.11)$$

olarak verilir. Bu örüntüler içinde endeksin maksimum değerini veren en iyi düzlem yapısı (**α*,β***) olarak kabul edilir (Martinez W. ve Martinez A., 2005). Bu yapı göz önüne alınarak veri ile ilgili yorumlamalara gidilir.

Yukarıdaki tanımlar altında, tüm olası iki boyutlu düzlemler arasından bulunacak olan en iyi düzlemin seçim algoritması belirlenir. Çok boyutlu verilerde çok sayıda ve farklı yapıda düzlem çeşitleri bulunabilir. Ayrıca, analistler için, düzlem çıktıları ile zaman kaybetmek büyük bir sorundur. Bu nedenle en kısa zamanda ve en optimum düzlemi keşfeden algoritmayı kullanmak büyük bir avantaj sağlar. Bu konuda da farklı yaklaşımlar bulunmaktadır.

İlk algoritma Friedman ve Tukey (1974) tarafından tanımlanmış; daha sonra Jones ve Sibson (1987) bir rastgele başlangıç ya da temel bileşenler analizinden herhangi biri ile başlayarak en dik çıkış (steepest-ascent) algoritmasını kullanmayı tercih etmişlerdir. Ancak bu işlemler genelde en uygun çözümü başlangıç noktasının yakınlarında bulmaktadırlar (Posse, 1995a). Bulunan çözüm başlangıç noktasında durmasa bile sonuç ilgilenilen izdüşümü vermemektedir. Bu döngüden kaçınmak için Friedman (1987), Jones ve Sibson' nın işlemine geri dönüp bir basit adım ile en dik çıkış yöntemini iyileştirmiştir. Bu algoritma temel bileşenlerin her bir çifti üzerinden hesaplanan izdüşüm endeksinin maksimum değerini bulur. Her bir çift üzerinden hesaplanan izdüşüm arama çözümü tek olmayacağı için çok sayıda düşük boyutlu izdüşümler ve her izdüşüm için hesaplanan endeks değerleri söz konusu olacaktır. Bu yüzden izdüşüm arama algoritması çok sayıda görünüme sahip olup her bir görünüm bize endeksin maksimumunu verebilecek durumda en iyi çözüm olarak kabul edilecektir. Bu nedenle, bir k-boyutlu veri kümesinin farklı izdüşüm sonuçlarının varlığı karşısında bulunabilecek en iyi çözüm stratejisi için Huber 1984 yılındaki makalesinde bir silme işleminden bahsetmiştir (Friedman, 1987). Friedman, burada, yapısal uzaklaştırma (structure removal) adını verdiği, "her defasında bir önce bulunan çözümü silip yeni en iyi çözümü bulmak için tekrarlı bir izdüşüm arama" algoritmasını önermiştir. Posse, 1990 yılında genel en iyi çözümü bulan algoritmadan esinlenerek yeni bir izdüşüm arama endeksi olan ki-kare izdüşüm endeksini geliştirmiştir. Posse, 1995 yılında ise bu endeks hesabına Friedman 'ın yapısal uzaklaştırma tekniğini uyarlayarak yeni bir algoritma geliştirmiştir. Bu işlem, izdüşüm arama endeksi olan Ki-Kare Endeksinin maksimum değerine ulaşana kadar

karşılaştırmalı ve tekrarlı bir yapı silme algoritmasıdır. Bu silme bir sonraki endeksin daha maksimum olan değerini bulabilmek içindir (Posse, 1995b). Ayrıca geliştirilmiş bu yeni endeks hesabı, beş temel istek olan, iki değişkenli normal dağılım minimizasyonunu, afin değişmezliği, tutarlılığı, dirençliliği ve geniş veri kümeleri için çabuk hesaplama kolaylığı avantajını sağlamaktadır (Posse, 1995a). Ayrıca Crawford, genetik en iyileştirme konusunda sınırlı bir biçimde ve çok sayıda rastgele seçim ile benzer bir tutum sergileyen bir algoritmadan bahsetmiştir (Posse, 1995b).

2.3.2. Posse Algoritması

Burada, öncelikle Posse' un algoritmasına değinilecek daha sonra bu algoritma ile birleştirilen Friedman' nın yapısal uzaklaştırılması anlatılacaktır.

Posse algoritmasına göre, amaç, normal dağılımlı rastgele üretilmiş bir 1xk boyutlu başlangıç düzlemleri olan (**α**, **β**) ile başlanarak; bunu, başlangıç için en iyi düzlem kabul edilen (**α***, **β***)' a dönüştürmektir. Bu dönüştürme işlemi aslında düzlemlerin birbirine dikleştirilmesini ve normlarının alınmasını ifade etmektedir. Çeşitli dikleştirme yöntemleri olup burada Gram-Schmidt dikleştirme yöntemi kullanılmıştır.

Rastgele seçilmiş 1xk boyutlu **α** vektörü,

$$\boldsymbol{\alpha}^{*} = \alpha_i / \sqrt{\sum \alpha_i^2} \qquad (2.12)$$

normalleştirilmiş **α*** vektörüne dönüştürülür. Rastgele seçilmiş 1xk boyutlu **β** vektörü ise **α***' a göre dikleştirilir buna "$\boldsymbol{\beta}^{+}$" dersek,

$$\boldsymbol{\beta}^{+} = \boldsymbol{\beta} - (\boldsymbol{\alpha}^{*}.\ \boldsymbol{\beta}).\ \boldsymbol{\alpha}^{*} \qquad (2.13)$$

şeklinde bulunur ve $\boldsymbol{\beta}^{+}$ vektörü de,

$$\boldsymbol{\beta}^{*} = \boldsymbol{\beta}_i^{+} / \sqrt{\sum \boldsymbol{\beta}_i^{+}} \qquad (2.14)$$

normalleştirilmiş **β*** vektörüne dönüştürülmüş olur.

Rastgele seçilen (**α**, **β**) vektörlerinin dikleştirilmiş biçimleri olan ilk başlangıç düzlemi (**α***, **β***) olarak kabul edilir; endeks hesabına gidilir. Çeşitli endeks hesapları ve Posse' un burada kullandığı Ki-kare Endeks hesabı daha sonra ayrıntılı olarak yer almaktadır. Aşağıdaki formüllerle bu "başlangıç en iyi düzlemi" kullanılarak yakın komşuluğu olan (α_1,β_1) ve (α_2,β_2) gibi iki yeni birbirine dik düzlemler elde edilir ve bu düzlemlerin de endeksleri hesaplanıp bir endeks karşılaştırmasına gidilir. Maksimum endeksi veren düzlem, yeni en iyi düzlem olarak alınır. Bu tekrarlı işlem, en iyi düzlemi bulmak için endeks değeri maksimum olana kadar devam eder. Komşu düzlem bulma işlemi,

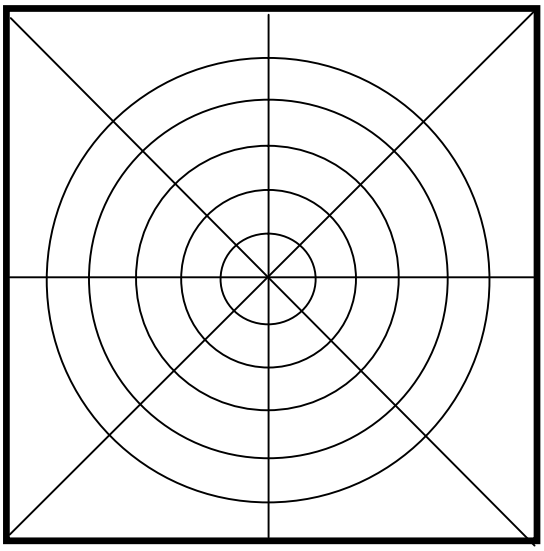
$$\boldsymbol{\alpha}_1 = (\boldsymbol{\alpha}^* + c\mathbf{v}) / \| \boldsymbol{\alpha}^* + c\mathbf{v} \| \qquad \boldsymbol{\beta}_1 = (\boldsymbol{\beta}^* - (\boldsymbol{\alpha}_1^T \boldsymbol{\beta}^*)\boldsymbol{\alpha}_1 / \| (\boldsymbol{\beta}^* - (\boldsymbol{\alpha}_1^T \boldsymbol{\beta}^*)\boldsymbol{\alpha}_1 \|$$

$$\boldsymbol{\alpha}_2 = (\boldsymbol{\alpha}^* - c\mathbf{v}) / \| \boldsymbol{\alpha}^* - c\mathbf{v} \| \qquad \boldsymbol{\beta}_2 = (\boldsymbol{\beta}^* - (\boldsymbol{\alpha}_2^T \boldsymbol{\beta}^*)\boldsymbol{\alpha}_2 / \| (\boldsymbol{\beta}^* - (\boldsymbol{\alpha}_2^T \boldsymbol{\beta}^*)\boldsymbol{\alpha}_2 \| \quad (2.15)$$

biçiminde uygulanır (Posse, 1995b). Burada c, kaç komşuluğun ziyaret edileceğini gösteren araştırmacı tarafından karar verilen bir skaladır; **v**, R^k boyutlu uzayda, k boyutlu normalleştirilmiş rastgele bir birim vektördür.

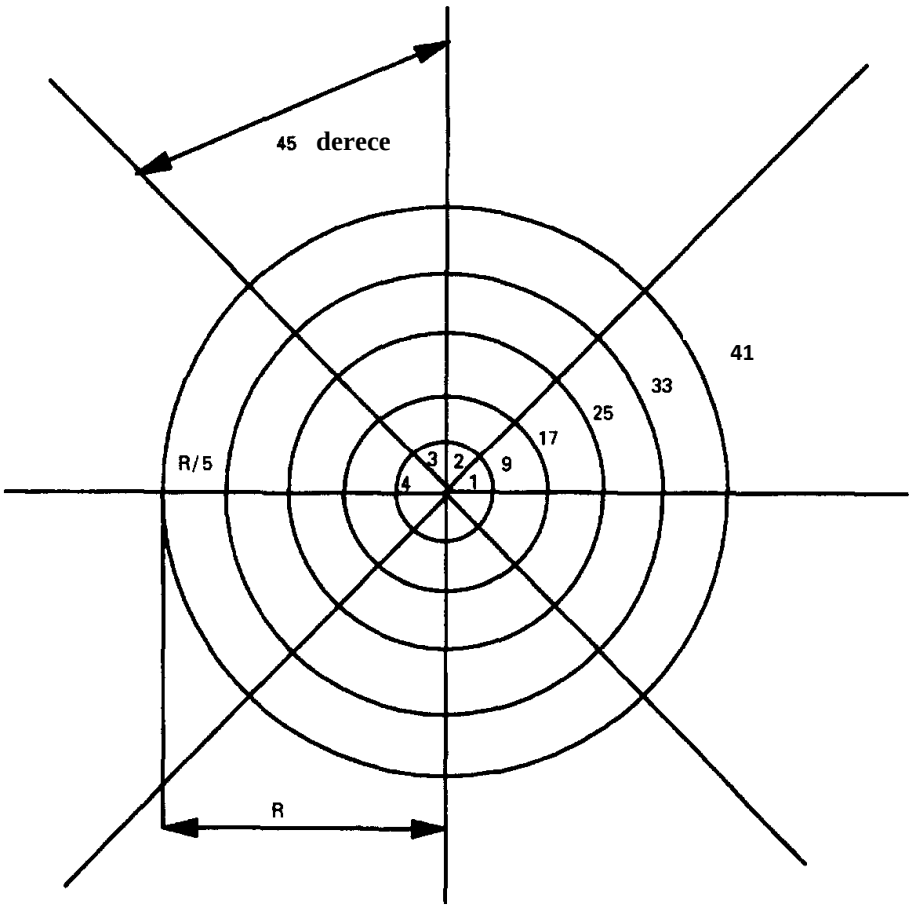
Posse algoritmasında önerilen yol, rastgele bir başlangıçtan sonra *c* değerinin azalması ile bağlantılı olarak bulunacak belli bir düzlemi aramaktır. *c* değeri, belli bir adım sayısından sonra yarıya indirilir. Bu adım sayısı bulunan endeks değerinin en az yarısı kadar iyileşme göstermemesi (artmaması) ile belirlenir. Bu c için araştırmacı tarafından verilmiş olan en küçük değere kadar ya da verilen en yüksek tekrar sayısına kadar devam eder ve sonrasında döngü durdurulur. Sonuçta en iyi düzlem arama *m* farklı başlangıç düzlemi ile tamamlanmış olur (Posse, 1995b).

En iyi endeks değerini veren en iyi düzlemin grafiksel gösterimi için Şekil 2.1' de verildiği gibi Posse bir B_k izdüşüm düzlemi kutusu kullanmıştır.



Şekil 2.1. Bk Kutusunun Çizimi.

$I_{Bk,}$ B_k kutusunun gösterge fonksiyonu biçiminde tanımlansın. Bu kutunun orta noktası orijin kabul edilerek 45 derecelik açılar ile kutu 8 eşit üçgen bölgesine ayrılır. Her bölgede iki değişkenli normal dağılımın dairesel simetrikliği hesaba katılır. Posse' a göre her bölge $(2\log(6))^{1/2}/ 5$ yay genişliği kadar bölümlenir; dolayısıyla orijinden 0.3786 uzaklığında bir ilk daire elde edilmiş olur. Daha sonra bu yay genişliği 2 katına çıkarılarak ikinci daire elde edilir. Bu işlem 5 daire elde edilene kadar devam edilir. 5. dairenin yarıçap R büyüklüğünde olup diğer bir ifade ile her daire arası aralık R/5 oranında bölümlenmiş olur. Kutular, normal dağılmış veriler için aynı ağırlığa sahip olacak biçimde R değeri belirlenir ki bu da $(2\log(6))^{1/2}$ 'ye eşittir (Posse, 1995a). Toplamda iç içe 5 dairesel ve 40 iç bölge elde edilmiş olur. 5. dairenin dış hattında kutunun kenarlarına kadar olan 6 kısım ise geri kalan 8 bölgemizi oluşturur. Toplamda 48 ayrı bölge elde edilmiş olur. Posse tarafından bu şekilde tanımlanmış olan izdüşüm düzleminin ayrıntılarını kutu üzerinde aşağıdaki gibi gösterebiliriz;



Şekil 2.2. Bk Kutusunun Ayrıntılı Çizimi.

2.3.3. Yapısal Uzaklaştırma Adımları

Friedman tarafından izdüşüm arama algoritmasına uyarlanan yapısal uzaklaştırma yönteminin adımları aşağıdaki gibidir.

1 - İlk iki satırı izdüşümün karşılık geleceği vektörler (α^*, β^*) olan bir kxk boyutlu **U*** matrisi ile başlanır. Diğer satırları birim matris şeklindedir; köşegen değerler bir ve geri kalanları 0 değerini alır. Örneğin, k = 4 iken,

$$\mathbf{U^*} = \begin{bmatrix} \alpha_1^* & \alpha_2^* & \alpha_3^* & \alpha_4^* \\ \beta_1^* & \beta_2^* & \beta_3^* & \beta_4^* \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad (2.16)$$

şeklinde olur. **U*** matrisi Gram-Schmidt yöntemi ile dikleştirilmiş olan **U** matrisine dönüştürülür.

2 - **T** matrisi kxn boyutlu olmak üzere $\mathbf{T} = \mathbf{U}\mathbf{Z}^{\mathbf{T}}$ şeklinde **Z** matrisine dönüştürülür, bu dönüşüm **T** matrisinin ilk iki satırındaki gözlemlerin (**α***, **β***) düzlemleri üzerindeki izdüşümdür. **T**' nin ilk iki satırı standart normal olan,

$$\mathbf{T_1} = (z_1^{\alpha^*}, \ldots, z_j^{\alpha^*}, \ldots, z_n^{\alpha^*})$$

ve

$$\mathbf{T_2} = (z_1^{\beta^*}, \ldots, z_j^{\beta^*}, \ldots, z_n^{\beta^*}) \qquad (2.17)$$

vektörlerine dönüşür.

Burada $z_j^{\alpha^*}$ ve $z_j^{\beta^*}$ 'lar (**α***, **β***) düzlemi üzerindeki j. gözlemin koordinatlardır.

3 - Bu iki satır vektörüne orijinden başlayarak γ açısı kadar bir döndürme işlemi gerçekleştirilir:

$$\tilde{z}_j^{1(t)} = z_j^{1(t)} cos\gamma + z_j^{2(t)} sin\gamma$$

$$\tilde{z}_j^{2(t)} = z_j^{2(t)} cos\gamma - z_j^{1(t)} sin\gamma \qquad (2.18)$$

$z_j^{1(t)}$ bu işlemde t. tekrardaki $\mathbf{T_1}$' in j.gözlemini ifade eder.

4 - Verinin büyüklük sırası ile sıralanmış durumuna karşılık gelen sıra numaraları yani rankları bir eksende ve bu ranklar kullanılarak bulunan dağılım fonksiyonu öteki eksende yer alır. Böylece, bir eksende $x_{(i)}$ sıralı değerleri bulunur ve p_i= (i/0.5) / n değerleri hesaplanır. Daha sonra $F^{-1}(p_i)$ değerleri bulunarak öteki eksende yer alır. Diğer sık kullanılan bir kestirim de, p_i = (i-3/8) / (n+1/4) şeklindedir (Blom Yöntemi).

$r\left(\tilde{z}_j^{1(t)}\right)$, $\tilde{z}_j^{1(t)}$, nin sıralı durumuna karşılık gelen rank değerlerini ifade etmektedir.

Daha sonra elde edilen izdüşüm endeksi için döndürülmüş değerlerin, standart normal birikimli dağılım fonksiyonlarının bulunması işlemi yapılır:

$$z_j^{1(t+1)} = \Phi^{-1}\left\{\frac{r\left(\tilde{z}_j^{1(t)}\right) - 0.5}{n}\right\}$$

$$z_j^{2(t+1)} = \Phi^{-1}\left\{\frac{r\left(\tilde{z}_j^{2(t)}\right) - 0.5}{n}\right\} \qquad (2.19)$$

5 - γ = 0, π/4, π/8, 3π/8 açısı ile 3 ve 4. adımlar tekrarlanır.

6 - Θ normallik dönüşümü, standart birikimli yoğunluk fonksiyonunun tersi (Φ^{-1}) kullanılarak bulunur:

$$\Theta(T1) = \Phi^{-1}[F(T_1)]$$

$$\Theta(T2) = \Phi^{-1}[F(T_2)]$$

ve diğer satırlar değişmeden kalacağından,

$$\Theta(Ti) = Ti \qquad i=3\ldots k$$

olur.

7 – Araştırmacı tarafından verilen maksimum tekrarlanma sayısına ulaşılana kadar 3. adımdan 6. Adıma tekrarlanan bir süreçtir. Bu tekrar, Friedman (1987)' ın çalışmasında ilk birkaç tekrardan sonra durdurulmuştur. Yapı silme işlem tekrarı genellikle 5-15 arasında tamamlanmaktadır.

8 - $Z' = U^T\Theta(UZ^T)$ eşitliğini kullanılarak veriler orijinal biçime dönüştürülür.

Friedman tarafından bulunan bu optimizasyon yöntemi hala geçerliliğini korumaktadır (Martinez W. ve Martinez A., 2005).

2.3.4. İzdüşüm Arama Yönteminin Adımları

Yukarıdaki kesimlerde açıklanan adımlara ilişkin genel işleyiş algoritması aşağıdaki gibidir.

1- Gözlemler (**X**), kovaryans matrisinden bulunan özvektör matrisi (**Q**) ve özdeğerler (**Λ**) kullanarak bir dönüşüm matrisi bulur.

$$\mathbf{Z_i} = \mathbf{\Lambda}^{-1/2}\,\mathbf{Q}^T(\mathbf{X_i} - \bar{X})^T \qquad i=1,\ldots,n \qquad (2.20)$$

2- Rastgele bir başlangıç düzlemini ($\boldsymbol{\alpha_0}$, $\boldsymbol{\beta_0}$) seçilir.

3- Başlangıç düzlemi için İzdüşüm Endeksi hesaplanır.

4- İki tane aday düzlem (($\boldsymbol{\alpha_1}$, $\boldsymbol{\beta_1}$), ($\boldsymbol{\alpha_2}$, $\boldsymbol{\beta_2}$)) üretir. Bu aday düzlemleri başlangıç olarak seçilen düzlemden uzak komşuluk olan değerlerden alınmaya başlanır; her tekrarda alınacak olan komşuluklar yakınlaşır.

5- Bu düzlemlerin izdüşüm endeksleri hesaplanır. Bu endeks hesaplama işlemine çeşitli yazarlar tarafından farklı yaklaşımlar vardır. Bunlar aşağıda ayrıntılı bir şekilde yer alacaktır.

6- Bu aday düzlemlerden izdüşüm arama endeksi en yüksek olan alınır ve bu yeni en iyi düzlem ($\boldsymbol{\alpha^*}$, $\boldsymbol{\beta^*}$) olarak belirlenir.

7- İzdüşüm arama endeksi ilerleme gösterdiği sürece 4. ve 6. adımlar araştırmacı tarafından verilen maksimum tekrar etme sayısına kadar tekrarlanır.

8- Endeks en az yarısı kadar iyileşme göstermezse, düzlemlerin hesaplandığı formüldeki c değeri yarıya indirilir.

9- c değerine yeni birkaç küçük sayı verilerek 4. adımdan 8. adım tekrarlanır. c' nin minimum değeri araştırmacı tarafından belirlenir. (Martinez W. ve Martinez A., 2005).

Genelde her zaman uygun olan bir izdüşümden söz edilmesi olası değildir. İzdüşüm, bağımsız dağılımlar ya da anlamlı kümelenmeler olabilir. Ayrıca uygun olan izdüşümde ayrı sınıflar arası kontur, yoğunlaşmalar, bölümler arası sınır olmayabilir. Ayrışmanın maksimumunu verecek şekilde en uygun izdüşüm düzlemi bulunduğunda işleyen sistemi durdurmak için bir endeks ölçüsü kullanılır. Var olan çeşitli endeks hesaplama yöntemlerinin formül ve açıklama kısımları aşağıdaki bölümde sunulmaktadır.

2.4. İZDÜŞÜM ARAMA YÖNTEMİ ENDEKSLERİ

Literatürde izdüşüm arama endeksini hesaplamaya ilişkin çok fazla alternatif vardır. Bunlardan ilk olanı Friedman ve Tukey Endeksi (1974) ve Friedman' ın L^2-Uzaklık Endeksi (1987), Jones ve Sibson' ın Moment ve Entropi Endeksleri (1987), Posse' un Ki-Kare Uzaklık Endeksi (1990) en bilinen ve algoritmalarıyla birlikte

tamamlanmış endekslerdir. Cook, Buja, Cabrera (1993), Hall (1989) ve Huber (1985) ise izdüşüm endekslerine özgün çeşitli eleştirisel makaleler yayınlamışlardır.

Etkili bir izdüşüm arama endeksi bazı özelliklere sahip olmalıdır. Bunlardan biri "afin" değişmezliktir; koordinat sistemi belli olan bir eğri nin göreceli olarak etkilenmeden değiştirilebilmesidir. Bu döndürüldüğü, ölçeklendirildiği ya da dönüştürüldüğü zaman eğrinin geometrisinin sabit kalması, değişmemesi, anlamına gelir. Diğer istenen özellikler ise tutarlılık, hızlı hesaplama sürecine sahip olma, dağılımın uç noktalarından ziyade normallikten sapan değerler için daha çok hassaslık gösterme şeklindedir (Friedman 1987).

2.4.1. Friedman-Tukey Endeksi:

1974 yılında Friedman ve Tukey tarafından ilk olarak izdüşüm arama endeksi adı verilen bu yöntem verinin yapısını keşfetmeyi ve yorumlamayı aramaktadır. Bu sistem, çok boyutlu veri kümesinin beklenen iki boyutlu izdüşümünü verene kadar devam eden bir veri döndürme operatörünü kullanır. Endeksin maksimum olduğu yer kontrol operatörü kabul edilir ve bu döndürmeler kontrol operatörü altında devam eder.

Bu tekrarlı süreç, her başlangıç noktası $\hat{m}$ gibi bir izdüşüm ekseninin bazı nicel özelliklerini kullanarak, hesaplanan izdüşüm endeksinin bu eksen üzerinde gösterimi ile başlar. Bulunan her düzlem bir sonraki daha büyük endeks için kullanılır ve sonunda en iyi düzleme ($\hat{m}^*$) ulaşılır. Her tekrarda hesaplanan bu endeks iki fonksiyonun çarpımından oluşur,

$$I(\hat{m}) = s(\hat{m})\, d(\hat{m}) \qquad (2.21)$$

olarak verilir. Burada, $\hat{m}$ izdüşüm eksenini (tek boyutlu), $s(\hat{m})$ verinin dağılım ölçüsünü yani verinin varyansını ve $d(\hat{m})$ normal yoğunluk fonksiyonunu ifade eder. Aykırı değerlere karşı sağlam bir sonuç elde edebilmek adına izdüşümün uçlarının her birinde bulunan noktaların p oranı kadar (2.22) ve (2.23) denklemlerindeki toplamlardan atılır.

$s(\hat{m})$ için, verinin düzeltilmiş standart hatası aşağıdaki gibi hesaplanır:

$$s(\hat{m}) = \left[\sum_{i=pN}^{(1-p)N} (\vec{X}_i \hat{m} - \bar{X}_m)^2 / (1-2p)N\right]^{1/2} \quad (2.22)$$

Burada

$$\bar{X}_m = \sum_{i=pN}^{(1-p)N} \vec{X}_i \hat{m} / (1-2p)N \quad (2.23)$$

şeklindedir.

N, toplam veri sayısı; $\vec{X}_i$ (i=1,…,N) verinin çok değişkenli vektörel gösterimidir. d($\hat{m}$) için, en yakın komşuluk fonksiyonunu kullanırız:

$$d(\hat{m}) = \sum_{i=1}^{N} \sum_{j=1}^{N} f(r_{ij})\, \gamma(R - r_{ij}) \quad (2.24)$$

Burada,

$$r_{ij} = \left| X_i . \hat{m} - X_j . \hat{m} \right| \quad (2.25)$$

şeklindedir. $\gamma(\eta)$, negatif değerler için 0, pozitif değerler için 1 dir. f(r) fonksiyonu r'nin r ≤ R sınırına yaklaştıkça monoton azalan olmalı, r = R' de ise 0'a indirgenmelidir. Bu, maksimum düzleştirme için gereklidir.

İki boyutluda izdüşüm için $\hat{m}$ ve $\hat{l}$ gibi iki doğrultu gerekmektedir:

$$s(\hat{m}, \hat{l}) = s(\hat{m})\, s(\hat{l}) \quad (2.26)$$

r_{ij} ise $r_{ij} = \left[(\vec{X}_i \hat{m} - \vec{X}_j \hat{m})^2 + (\vec{X}_i \hat{l} - \vec{X}_j \hat{l})^2\right]^{1/2}$ biçiminde yazılır. Algoritmanın tekrar edilen başvuruları, f(r) açık fonksiyon formunun duyarsız olduğunu ve karakteristik genişlik üzerinde büyük bir bağlılık gösterir. Bu bağlılık,

tek boyutlu için,

$$\bar{r} = \int_0^K r f(r) d_r \Big/ \int_0^K f(r) d_r \quad (2.27)$$

iki boyutlu için,

$$\bar{r} = \int_0^K r f(r) r d_r \Big/ \int_0^K f(r) r d_r \quad (2.28)$$

şeklindedir. Bu karakteristik genişlik maksimum yoğunluk dağılımı için aranan genel bir tanımlamadır. Bu değer duyarlı algoritmanın izdüşüm alt kümelerinde uzaklığı verir.

İki boyutlu izdüşüm arama algoritması tek boyutluya göre daha yavaş ve hafifçe daha az kararlıdır, bununla birlikte iki boyutlu olan gösterim veri hakkında daha fazla bilgi içermektedir. Deneyimler gösteriyor ki en kullanışlı strateji ilk olarak tek boyutlu birkaç izdüşüm arama çözümleri bulmak ve iki boyutlu izdüşüm arama için bunlardan her birini başlangıç koordinatları olarak almak (Friedman, 1974).

2.4.2. Entropi Endeksi

Jones ve Sibson 1987 yılında yayınladıkları makalelerinde Friedman ve Tukey' in 1974 yılındaki önerdikleri izdüşüm arama yöntemini biraz daha geliştirmişlerdir. Yöntemi bazı öneri ve farklı metotlar ile değiştirmişler ve bu yöntemlere etkili yaklaşımlar da bulunmuşlardır.

Jones ve Sibson tarafından 1987 yılında standartlaştırılmış birinci derece entropiden esinlenerek iki boyutlu normallikten sapmayı ölçen entropi izdüşüm arama endeksini yayınlamışlardır. Bu endeksin formülü aşağıdaki gibidir:

$$\widehat{PI}_E(\alpha,\beta) = \frac{1}{n}\sum_{i=1}^{n} log\left\{\frac{1}{nh_\alpha h_\beta}\sum_{j=1}^{n}\Phi_2\left(\left(z_i^\alpha - z_j^\alpha\right)/h_\alpha, \left(z_i^\beta - z_j^\beta\right)/h_\beta\right)\right\} + \log(2\pi e) \quad (2.29)$$

Ve burada ki h_γ , her koordinatta γ = α, β için bir bant genişliğidir. Φ_2 standart iki değişkenli normal yoğunluktur. Burada,

$$h_\gamma = 1.06\, n^{-1/5}\left(\sum_{i=1}^{n}\left(z_i^\gamma - \sum_{i=1}^{n} z_i^\gamma/n\right)^2/(n-1)\right)^{1/2} \quad (2.30)$$

şeklindedir (Posse,1995b).

2.4.3. Moment Endeksi

Jones ve Sibson tarafından 1987 yılında iki değişkenli üçüncü ve dördüncü momentten esinlenerek ve iki boyutlu Entropi Endeksine alternatif olarak geliştirilen

izdüşüm arama endeksidir. Bu endeks geniş veri kümeleri için hızlı hesaplama kolaylığı sağladığı için çok kullanışlıdır. Bununla birlikte bu endeksin dezavantajı dağılımın uç noktalarındaki normallikten sapan değerlerden fazlasıyla etkilenmesidir. Formülü aşağıdaki gibidir (Posse, 1995b):

$$\widetilde{PI}_M(\alpha,\beta) = \frac{1}{12}\left\{K_{30}^2 + 3K_{21}^2 + 3K_{12}^2 + K_{03}^2 + \frac{1}{4}(K_{40}^2 + 4K_{31}^2 + 6K_{22}^2 + 4K_{13}^2 + K_{04}^2)\right\} \quad (2.31)$$

Burada,

$$K_{21} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^{\alpha})^2 z_i^{\beta} \quad (2.32)$$

$$K_{12} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(z_i^{\beta}\right)^2 z_i^{\alpha} \quad (2.33)$$

$$K_{30} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}(z_i^{\alpha})^3 \quad (2.34)$$

$$K_{03} = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(z_i^{\beta}\right)^3 \quad (2.35)$$

$$K_{40} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}(z_i^{\alpha})^4 - \frac{3(n-1)^3}{n}\right\} \quad (2.36)$$

$$K_{04} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}\left(z_i^{\beta}\right)^4 - \frac{3(n-1)^3}{n}\right\} \quad (2.37)$$

$$K_{31} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(z_i^{\alpha})^3 z_i^{\beta} \quad (2.38)$$

$$K_{13} = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(z_i^{\beta}\right)^3 z_i^{\alpha} \quad (2.39)$$

$$K_{22} = \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)\sum_{i=1}^{n}(z_i^{\alpha})^2(z_i^{\beta})^2 - \frac{(n-1)^3}{n}\right\} \quad (2.40)$$

şeklinde hesaplanan momentler endeks formülünde yerine konularak hesaplama yapılır.

2.4.4. L^2- Uzaklığı:

Friedman 1987 yılında, daha önce 1974 yılında yayınladığı açıklayıcı izdüşüm arama yöntemini istatistiksel doğruluk ve hesaplama kolaylığı açısından yeni bir algoritma yardımı ile geliştirmiştir. Standart iki değişkenli normal dağılım ve veri dağılımının yoğunluğu ile L^2 uzaklıkları ile hesaplanan bir endeks tasarlamıştır. Ancak bu endeks çeşitli yazalar tarafından eleştirilmiş ve uzaklık tahmin edici formül için birçok seçenek sunulmuştur. Uzaklık hesabı için sunulan eşitlik,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(f_{\propto\beta}\left(z^{\propto},z^{\beta}\right)-\Phi_2\left(z^{\propto},z^{\beta}\right)\right)^2\omega\left(z^{\propto},z^{\beta}\right)dz^{\propto}z^{\beta} \qquad (2.41)$$

Biçimindedir. Burada, $\omega\left(z^{\propto},z^{\beta}\right)$ için değişik eşitlikler verilmiştir:

Friedman' a göre, $\omega\left(z^{\propto},z^{\beta}\right)=\Phi_2^{-1}\left(z^{\propto},z^{\beta}\right)$.

Hall(1989) ve Morton (1989)' a göre, $\omega\left(z^{\propto},z^{\beta}\right)=1$.

Cook(1993)' a göre, $\omega\left(z^{\propto},z^{\beta}\right)=\Phi_2\left(z^{\propto},z^{\beta}\right)$.

Firedman, Hall ve Morton ağırlıklandırılmamış L^2 uzaklıklarını kullanmışlar; ayrıca, Friedman ve Morton verileri dönüştürerek uygulamışlardır. Cook ağırlıklandırılmış L^2 uzaklıklarını kullanmıştır. Bu uzaklık ölçer dağılımdan sapan değerlere karşı daha hassastır. Verilerde dönüştürme yapılmaksızın $\omega=\Phi_2$ ile bu uzaklıklar hesaplanmıştır (Posse, 1995b).

Bu tartışmalar sadece kuramsal uzaklığı ölçebilmek adına kullanılan farklı yaklaşımlar içindir; onların kestirileceği izdüşüm endeksi için değildir. Sonuçlar birbirinin yerini tutabilen kuramsal uzaklıkların benzer tutum sergilediğini göstermektedir.

Friedman, $y^{\alpha}=2\Phi(z^{\alpha})-1$ ve $y^{\beta}=2\Phi\left(z^{\beta}\right)-1$ dönüşümü ile $[-1,1]\times[-1,1]$ birim kare matris kullanmıştır; burada, Φ standart Gaussian dağılımdır ve Legendre Polinomunun genişletilmişidir.

Buna göre,

$$\widehat{PI}_{Leg}(\alpha,\beta) = \frac{1}{4}\left\{\sum_{j=1}^{J}(2j+1)\left(\frac{1}{n}\sum_{i=1}^{n}P_j(y_i^{\alpha})\right)^2 + \sum_{k=1}^{J}(2k+1)\left(\frac{1}{n}\sum_{i=1}^{n}P_k\left(y_i^{\beta}\right)\right)^2 + \sum_{j=1}^{J}\sum_{k=1}^{J-j}(2j+1)(2k+1)\left(\frac{1}{n}\sum_{i=1}^{n}P_j(y_i^{\alpha})P_k(y_i^{\beta})\right)^2\right\} \quad (2.42)$$

şeklindedir ve burada $P_a(.)$ a. sıradan Legendre Polinomunu ifade eder.

Ancak bu önerisi de 1989 yılında Morton tarafından afine değişmez olmadığı için eleştiriye uğramıştır. Morton, L^2 uzaklığını $\rho = (z^{\alpha})^2 + (z^{\beta})^2$ ve $\theta = \arctan(z^{\beta}/z^{\alpha})$ kutupsal koordinatlar ile yeniden yapılandırmıştır. Bu endeks Laguerre Polinomunun ve bir Fourier serisinin genişletilmişidir.

$$\widehat{PI}_{LF}(\alpha,\beta) = \frac{1}{\pi}\sum_{l=0}^{L}\sum_{k=1}^{K}\left(\left(\frac{1}{n}\sum_{i=1}^{n}L_l(\rho_i)\exp\left(-\frac{\rho_i}{2}\right)\cos(k\theta_i)\right)^2 + \left(\frac{1}{n}\sum_{i=1}^{n}L_l(\rho_i)\exp\left(-\frac{\rho_i}{2}\right)\sin(k\theta_i)\right)^2\right) + \frac{1}{2\pi}\sum_{l=0}^{L}\left(\frac{1}{n}\sum_{i=1}^{n}L_l(\rho_i)\exp\left(-\frac{\rho_i}{2}\right)\right)^2 - \frac{1}{2\pi n}\sum_{i=1}^{n}\exp\left(-\frac{\rho_i}{2}\right) + \frac{1}{8\pi} \quad (2.43)$$

biçimindedir ve burada $L_a(.)$ a. sıradaki Laguerre Polinomunu ifade eder.

Hall, 1989 yılında verileri dönüştürmeksizin ve Morton (1989) ile aynı uzaklık yöntemini kullanarak ve θ_1 için ortogonal olan Hermit Polinomunu genişletmiştir. Bu endeks,

$$\widehat{PI}_{Her}(\alpha,\beta) = \sum_{j=0}^{H}\sum_{k=0}^{H-j}\frac{2^{-(j+k)}}{j!k!}\left(\frac{1}{n}\sum_{i=1}^{n}H_j(z_i^{\alpha})\Phi_1(z_i^{\alpha})\right)^2\left(\frac{1}{n}\sum_{i=1}^{n}H_k\left(z_i^{\beta}\right)\Phi_1(z_i^{\beta})\right)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\Phi_1(z_i^{\alpha})\sum_{i=1}^{n}\Phi_1(z_i^{\beta}) + \frac{1}{4\pi} \quad (2.44)$$

şeklindedir ve burada $H_a(.)$ a. sıradaki Hermite Polinomunu ifade eder.

Ağırlıklı θ_2 için Cook 1993 yılında θ_1^2 için ortogonal olan Natural Hermit Polinomunu geliştirmiştir ve bu endeks,

$$\widehat{PI}_{Nat}(\alpha,\beta) = \sum_{j=0}^{N}\sum_{k=0}^{N-j}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{j!k!}}N_j(z_i^{\alpha})N_k(z_i^{\beta})\Phi_2(z_i^{\alpha},z_i^{\beta}) - b_j b_k\right)^2 \quad (2.45)$$

şeklindedir ve burada $N_a(.)$ a. sıradan Natural Hermite Polinomunu ve $b_{2i+1} = 0$ ve $b_{2i} = (-1)^i\sqrt{(2i)!}/(\sqrt{\pi}i!\,2^{2i+1})$ i=1,2,3… ifade eder.

2.4.5. Ki - Kare Endeksi:

Posse, 1990 yılındaki rastgele seçim algoritması ile uygun olan izdüşüm arama endeksini, 1995 yılında Friedman' ın yapısal uzaklaştırma yöntemi ile kombine ederek kullanmıştır. Her bulunan yapı bir öncekinden daha küçük etkinlik göstermelidir (Martinez W. ve Martinez A., 2005). Posse' un endeks hesaplama yöntemin de düzlem bulma algoritmasını daha önce değindiğimiz yapısal uzaklaştırma yöntemi ile birleştirerek kullanılmasına yer verilecektir.

P(**α**, **β**) düzlemi üzerindeki teorik ve ampirik marjinal dağılımı $F_{\alpha\beta}$ ve $\hat{F}_{\alpha\beta}$ şeklinde gösterildiğini ve P(**α**, **β**)' nın η açısı kadar saat yönünün tersine döndürülmesinin sonuncunda marjinal dağılımın $F_{\alpha\beta}(\eta)$ ve $\hat{F}_{\alpha\beta}(\eta)$ şeklinde olduğunu kabul edelim. $F_{\alpha\beta}(\eta)$, (z^{α}, z^{β}) veri dağılımının dağılım fonksiyonunu ifade eder;

$$PI(\alpha,\beta) = \frac{4}{\pi}\int_0^{\pi/4} PI(\alpha,\beta,\eta)\,d\eta \quad (2.46)$$

Buradan,

$$\begin{aligned} PI(\alpha,\beta,\eta) &= \sum_k^{48}\left[\left(\iint_{B_k}\{dF_{\alpha\beta}(\eta) - d\Phi_2\}\right)^2 / \iint_{B_k} d\Phi_2\right] \\ &= \sum_k^{48}\left[\left(\iint_{B_k}\{dF_{\alpha\beta}(\eta)\}\right)^2 / \iint_{B_k} d\Phi_2\right] - 1 \end{aligned} \quad (2.47)$$

şeklini alır.

Burada $\iint_{B_k} d\Phi_2 = \iint_{B_k} \Phi_2$' dir. $PI(\alpha,\beta,\eta)$ formülündeki integral hesabının $\widehat{PI}(\alpha,\beta,\eta)$ da sınırlı toplam ile yer değiştirmesi daha hızlı, kolay ve varyanssız bir hesaplama sunmaktadır;

$$\widehat{PI}(\alpha,\beta) = \frac{1}{l}\sum_{j=0}^{l-1}\sum_{k=1}^{48}\frac{1}{c_k}\left(\frac{1}{n}\sum_{i=1}^{n} I_{B_k}\left(z^{\alpha(\eta_j)}, z^{\beta(\eta_j)}\right) - c_k\right)^2 \qquad (2.48)$$

Eğer F fonksiyonu tamamıyla sürekli ise $\widehat{PI}(\alpha,\beta) \rightarrow PI(\alpha,\beta)$ yakınlaşır.

Burada

l=9,

$c_k = \iint_{B_k} \Phi_2 dz_1 dz_2$ c_k standart iki değişkenli normal yoğunluk kullanılarak k. derece olasılığıdır.

$\eta_j = \pi j/4l$ yani l=9 ise $\eta_j = \pi_j/36$ olur j=0,…,8 için.

$I_{B_k}(\ldots)$ B_k kutusu için gösterge fonksiyonudur ve

$$\alpha(\eta_j) = \alpha\cos(\eta_j) - \beta\sin(\eta_j)$$

$$\beta(\eta_j) = \alpha\sin(\eta_j) - \beta\cos(\eta_j) \qquad (2.49)$$

şeklindedir ve bu tanımlamalarla birlikte endeks;

$$\widehat{PI}_{\chi^2}(\alpha,\beta) = \frac{1}{9}\sum_{j=1}^{8}\sum_{k=1}^{48}\frac{1}{c_k}\left(\frac{1}{n}\sum_{i=1}^{n} I_{B_k}\left(z^{\alpha(\eta_j)}, z^{\beta(\eta_j)}\right) - c_k\right)^2 \qquad (2.50)$$

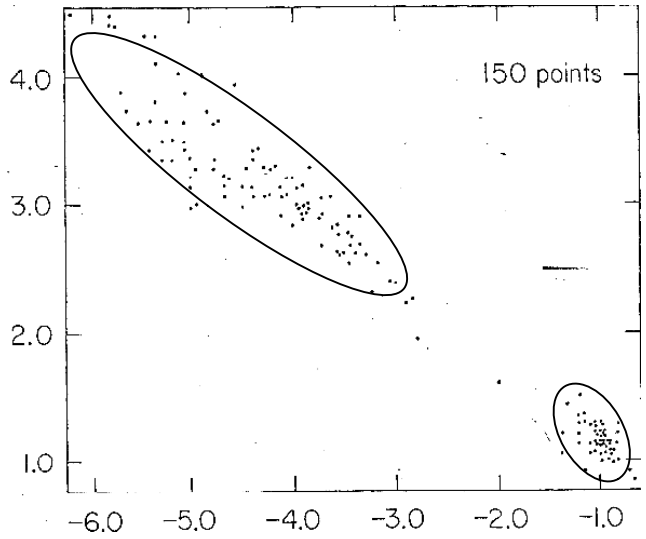
şeklini alır (Martinez W. ve Martinez A., 2005).

İlgilenilen izdüşümün bulunmasından sonraki adım yapının bu çözümde gösterilip gösterilmediğine bağlıdır. Eğer düzlem kümelenmeleri açıkça gösteriliyorsa, biri onlardan ayrılmış olabilir ve onlardan ayrıca araştırılmalıdır. Diğer durumda ise 3 alternatif düşünülebilir. Öncelikle çözüme dik olan alt uzaylarda tarama yapılır. İkinci olarak art arda gelen iki çözüm arasındaki uzaklığı dikkate alan izdüşüm arama endeksine hata fonksiyonu eklenmelidir. Son olarak endeks daha karmaşık ve hesaplanması daha zor hale gelir.

2.5. İZDÜŞÜM ARAMA YÖNTEMİ İLE İLGİLİ BAZI AÇIKLAYICI UYGULAMALARI

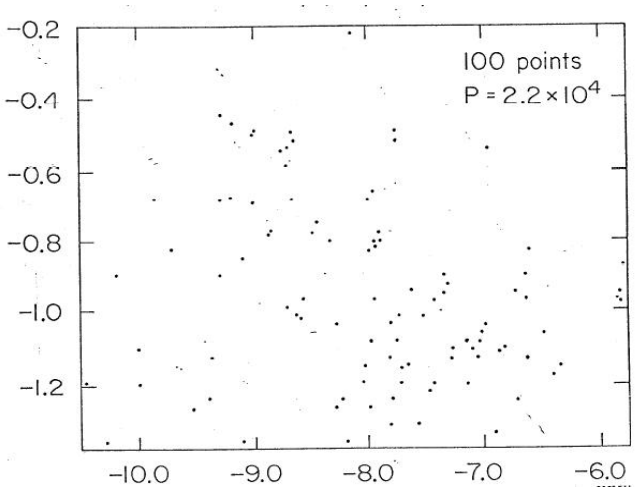
2.5.1. Friedman ve Tukey (1974) İzdüşüm Arama Yöntemi Uygulaması

Friedman ve Tukey tarafından 1974 yılında savunulan izdüşüm arama endeksi klasik bir veri kümesi olan ilk kez Fisher tarafından kullanılan iris verisi ile çalışılmıştır. Bu yöntem ile çözümü burada ve Posse 'un Ki-Kare Endeksi ile olan çözümü ise uygulamalar kısmında sunulmuştur. İris verisi 3 farklı çiçek türü (Setosa, Versicolor, Virginica) için 4 ayrı özellikten ve her çiçek türü için 50 gözlemden oluşan bir veri kümesidir. Bunlar çiçeğin sap genişliği, sap uzunluğu ve yaprak uzunluğu, yaprak genişliğidir. Önceden farklı tür olduğu bilinen bu üç çiçek türü olan iris veri kümesinin iki boyutlu izdüşüm çizimi aşağıdaki gibidir:



Şekil 2.3. İris Veri Kümesinin İzdüşüm Arama Çizimi.

Şekilden de görülüğü gibi 150 gözlemli veri açıkça iki kümeye ayrılmıştır. Tek boyutlu izdüşüm arama yönteminin de ayrıca bu ayrışmayı açıkça verdiği gözlemlenmiştir (Friedman, 1974). Bu kümelenme bir türün diğerlerinden oldukça farklı bir dağılıma sahip olduğu sonucunu vermektedir. Belirgin bir şekilde ayrılan küme çıkarılıp diğer kümelenmeye tekrardan izdüşüm arama yöntemi uygulanırsa aşağıdaki Şekil 2.4. ile karşılaşılır.



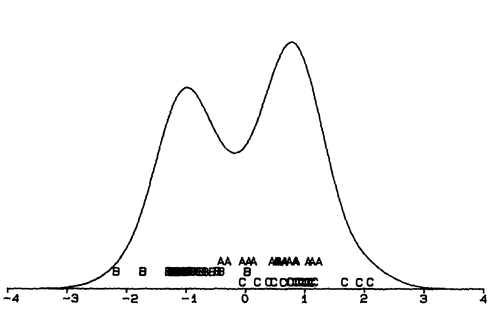
Şekil 2.4. İki Farklı Çiçek Türünün Dağılımı.

Şekilden de görüldüğü gibi biri sağ alt kısımda diğeri sol yukarıya yayılmış bir şekilde iki kümelenme olduğu görülebilir.

Daha sonra Friedman' ın bu algoritma ve endeks ölçümü üzerine yapılan eleştiriler değişik yaklaşımları beraberinde getirmiştir. Bu yaklaşım ise izdüşüm arama yönteminin başlangıcı olarak kalmış, yeni algoritma ve endeks hesaplama yolları ile izdüşüm arama yöntemi geliştirilmiştir.

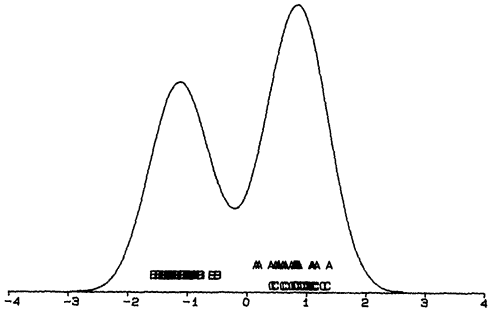
2.5.2. Başlangıç Noktası Temel Bileşenler Olan İzdüşüm Arama Yöntemi Uygulaması

Lubischew (1962)' in 3 farklı erkek pire böceği (Chaetocnema) türünden elde ettiği ve 6 değişkenden oluşan 74 gözlemli veri kümesi üzerinde Jones ve Sibson (1987) birinci ikinci ve üçüncü temel bileşenler analizini başlangıç noktası alarak izdüşüm arama yönteminin çizim sonuçlarını karşılaştırmışlardır. Temel bileşenler analizinden sonra entropi endeksi kullanılarak bulunan izdüşüm arama yöntemi sonuçları aşağıda verildiği gibidir;



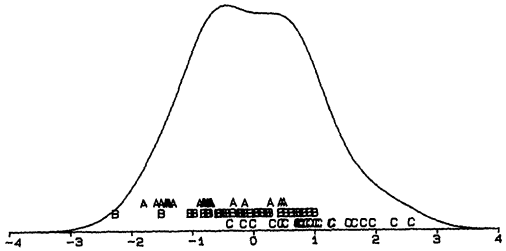
(a). Birinci Temel Bileşenler,

e= -1.44.



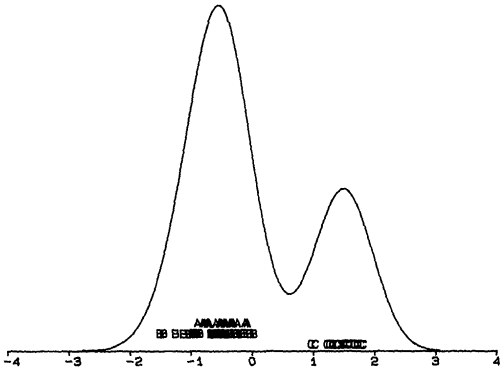
(b). İzdüşüm Arama Çözümü,

e = -1.35.



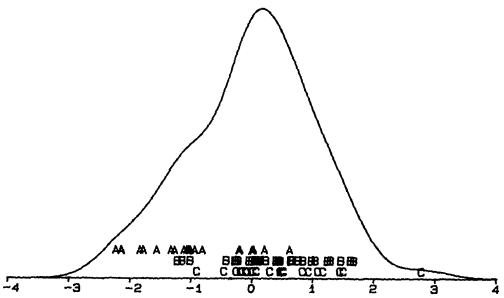
(c). İkinci Temel Bileşenler,

e= -1.49.



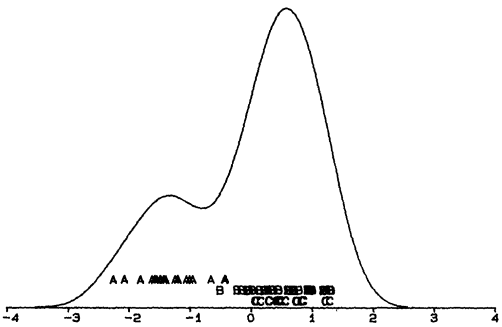
(d). İzdüşüm Arama Çözümü,

e = -1.32.



(e). Üçüncü Temel Bileşenler,

e= -1.48.



(f). İzdüşüm Arama Çözümü,

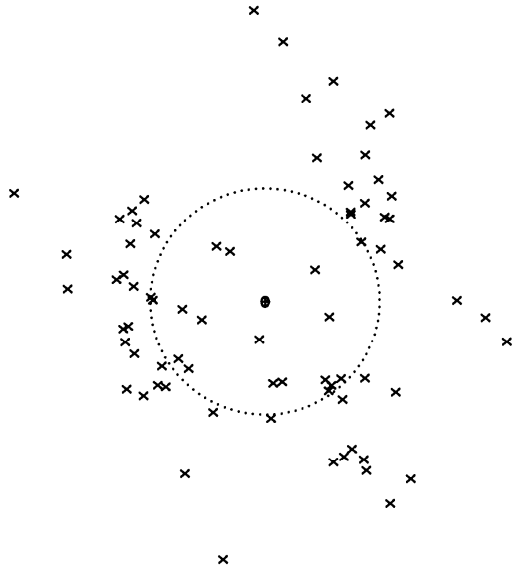
e = -1.41.

Şekil 2.5. Temel Bileşenler ile İzdüşüm Arama Yöntemi Çizimleri.

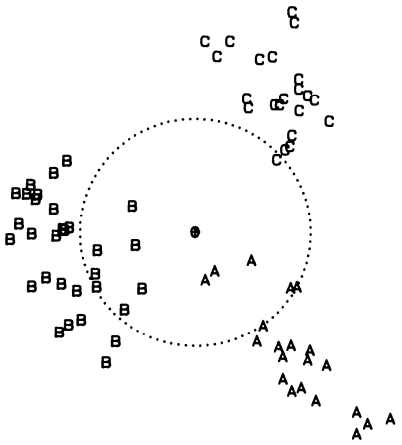
Şekil 2.5.'de (a), (c) ve (e) sırasıyla birinci, ikinci ve üçüncü temel bileşenleri göstermektedir. (b), (d) ve (f) şıklarında ise bu çizimler başlangıç noktası alınarak bulunan izdüşüm arama yöntemi çizimlerini ifade etmektedir. (b), (d) ve (f)'de diğer çizimlere göre A, B ve C uzaylarının gayet iyi ayrıştığı açıkça görülmektedir. Ayrıca

bu başlangıç noktalarını göre izdüşüm arama çizimlerini yorumlanırsa en yüksek endeks değerinin şekil 2.5. (d)' de ikinci temel bileşenlerde olduğu görülmektedir. Lubischew' in bu verisinin oldukça düzgün dağıldığı (Jones ve Sibson, 1987).

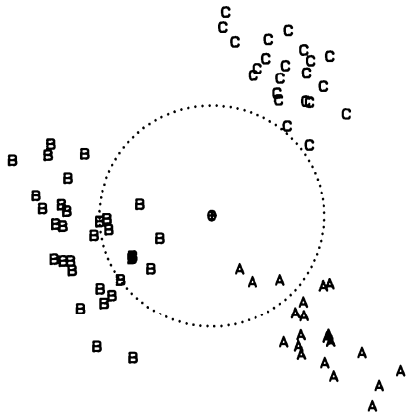
Aynı verileri saçılım dağılım grafiğini ve moment endeks hesabını kullanılarak bulunan grafik çizimleri aşağıda verilmiştir.



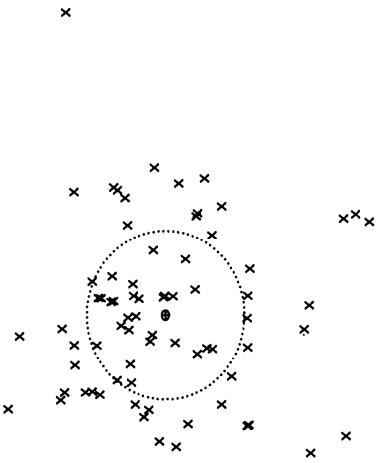
(a). İlk İki Temel Bileşenler, *m*= 0.90.



(b). İzdüşüm Arama Çözümü, *m* = 2.53.



(c). İlk İki Kanonik, *m*= 2.36.



(d). İzdüşüm Arama Çözümü, *m* = 2.40.

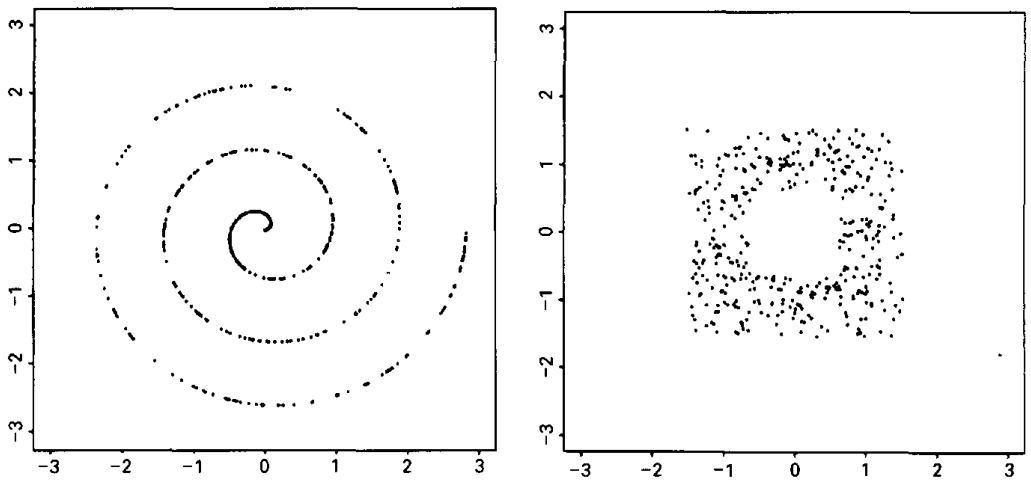
Şekil 2.6. Temel Bileşenler ile İzdüşüm Arama Yöntemi Çizimleri.

Şekil 2.6. (a) başlangıç noktası alınarak bulunan moment endeks grafiği (b)' de gösterilmektedir. Aynı durum (c) ve (d) içinde geçerlidir. Şekil 2.6. (d) daha yüksek endeks değerine sahip olup (b)' ye göre daha açıklayıcı bir dağılıma sahiptir.

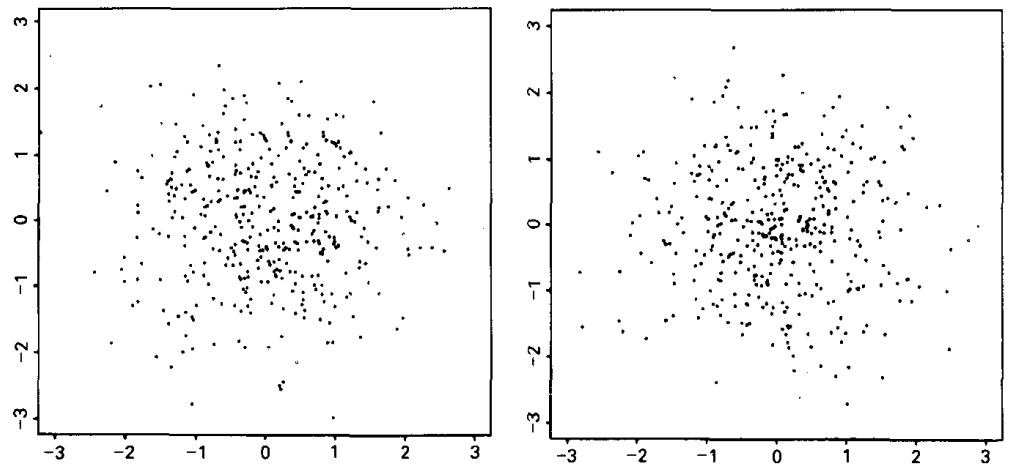
2.5.3. Farklı Optimizasyon Algoritmaları ve İzdüşüm Arama Endekslerinin her kombinasyonunun karşılaştırılması:

Farklı optimizasyon algoritmaları ve izdüşüm arama endekslerinden hangi kombinasyonların daha iyi sonuç verdiğini görmek için yayınladığı makalesinde Posse (1995b) farklı dağılım şekli üzerinden karşılaştırmalar yapmıştır. İki boyutlu olan bazı dağılım şekilleri üzerinde denenen kombinasyonlar da güçlü bir optimizasyon algoritmasını ihtiyaç duyulduğu ve bununla birlikte iyi bir endeks değerinin istenilen sonucu verdiğini gösterilmiştir.

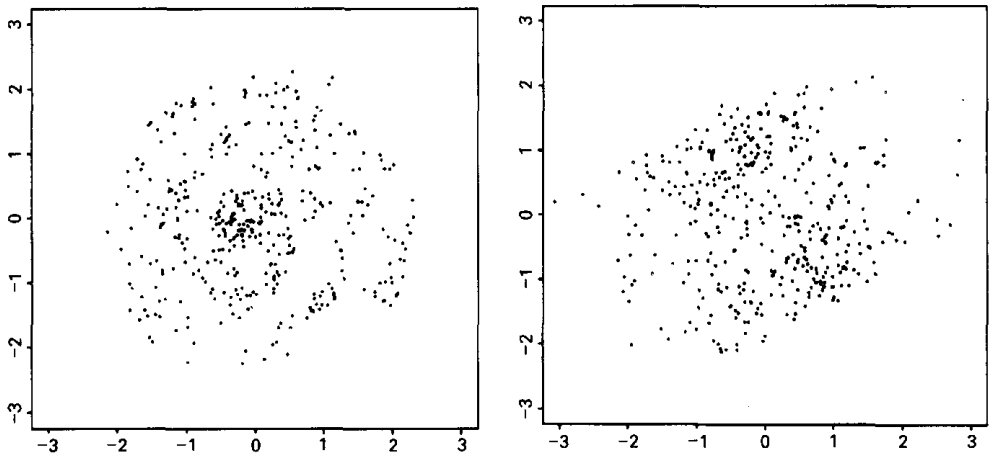
Jones ve Sibson (1987), Friedman (1987) ve Posse (1995a) tarafından savunulan teknikler kullanılarak iki boyutlu spiral ve donut şekilleri üzerinden bir karşılaştırma yapılmak istenirse;



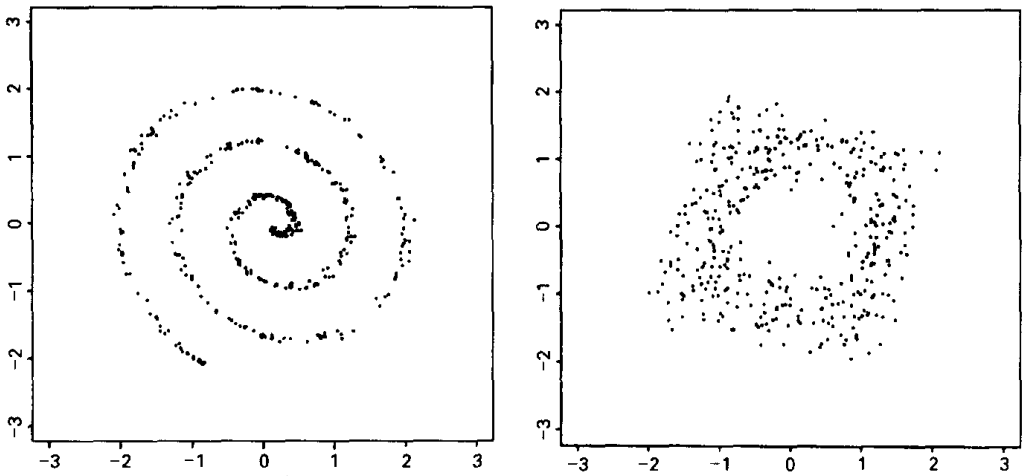
(a) Orijinal verilerin örüntüleri



(b) Jones ve Sibson Endeksi kullanılarak çizim



(c) Friedman Endeksi kullanılarak çizim



(d) Posse Ki-Kare Endeksi kullanılarak çizim

Şekil 2.7. Jones ve Sibson, Friedman ve Posse yöntemlerinin karşılaştırılması.

Şekil 2.7. (a) ’da spiral ve donut şeklini veren veri kümemizin Şekil 2.7. (b) şıkında Jones ve Sibson tekniği kullanılarak çizimi yapılmıştır. Şekil 2.7. (c) Friedman izdüşüm arama yöntemi kullanılmıştır. Şekil 2.7. (d) ’de ise Posse tarafından savunulan modifiye edilmiş ki-kare yönteminin çizimini göstermektedir. Şekillerden de anlaşılacağı üzere en yakın çizim Şekil 2.7. (d) şıkkı olan Posse’ un tekniği kullanılarak yapılmış olandır. Jones ve Sibson tekniğinin tamamen başarısız olduğu, Friedman tekniğinin ise daha iyi bir sonuç verdiği görülmektedir.

Legendre endeksi kısmen de olsa en kötü sınıflandırılmış veri dağılımına bile iyi uyum sağlayan endeks olarak, ki-kare endeksi de çok etkili ve oldukça hızlı hesaplamaya sahip olduğu görülmüştür.

ÜÇÜNCÜ BÖLÜM

UYGULAMA

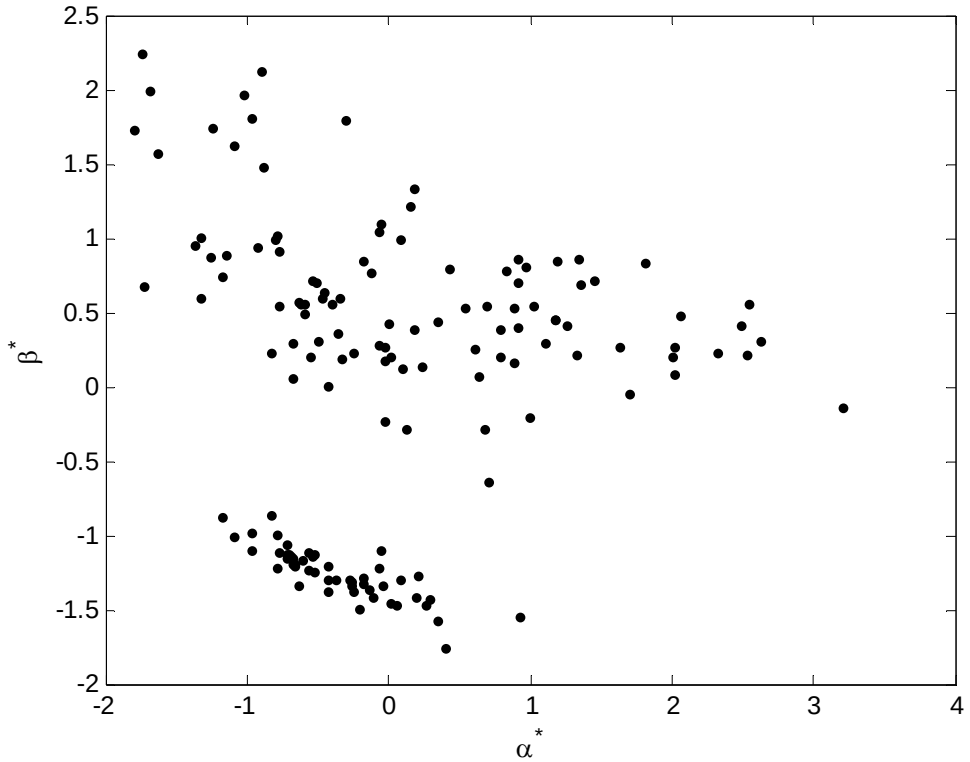
Bu bölüm iki farklı veri kümesi üzerinden izdüşüm arama yöntemin uygulamasını ve veri kümelerinin bazı özelliklere sahip olduğu durumlardaki karşılaştırılmalı yorumları içermektedir.

İlk kısım Friedman ve Tukey' in de 1974 yılındaki makalelerinde kullandıkları İris veri kümesine Posse' un Ki-Kare Endeksinin uygulanması ve paralel koordinatlar çizimi ile karşılaştırılmasıdır.

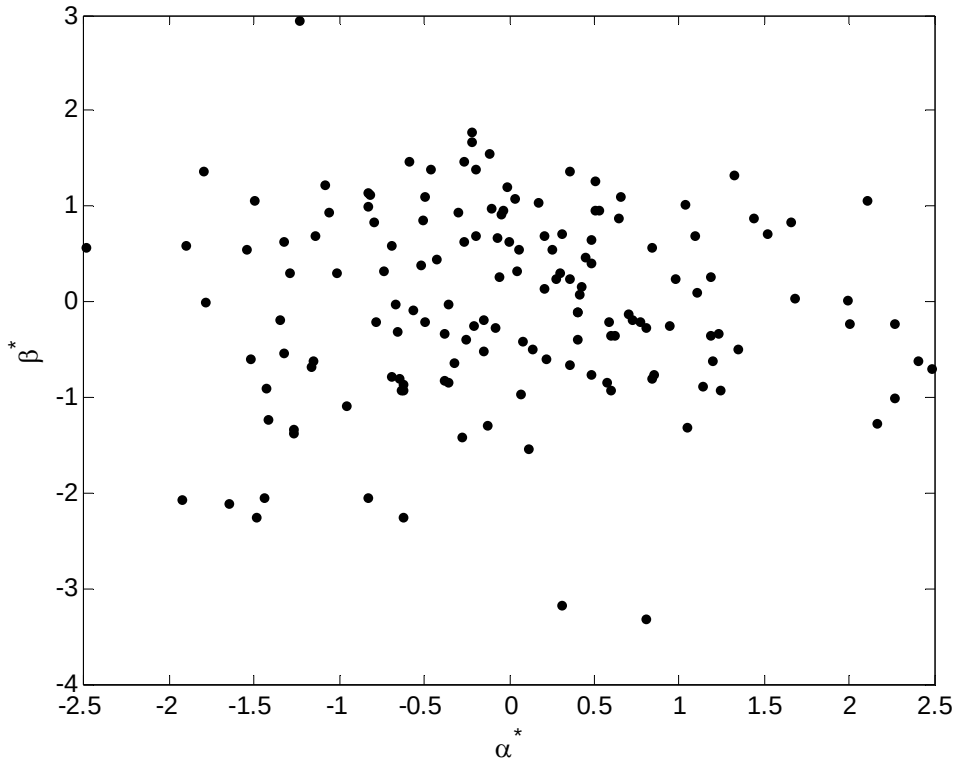
İkinci kısım ise Auto Katalog 2008' in Auto Motor özel sayısından alınan Aralık 2007 araba verilerine Posse' un Ki-Kare Endeksinin farklı gözlem ve değişken sayısı altında, çoklu bağlantının ve aykırı değerin olduğu durumlarda, çarpık dağılıma sahip değişkenlerin dönüştürülmüş ve dönüştürülmemiş durumlarda ki çalışmaların sonucu yer almaktadır.

3.1. UYGULAMA 1

Daha önce Friedman ve Tukey Endeksi için kullanılan 3 farklı çiçek türü olan Setosa, Versicolor, Virginica isimli çiçek türleri için 4 ayrı özellik gözlemlenen veri kümesi, bunlar çiçeğin sap genişliği, sap uzunluğu ve yaprak uzunluğu, yaprak genişliğidir, Posse 'un ki-kare izdüşüm arama endeks çizimi şekildeki gibidir.

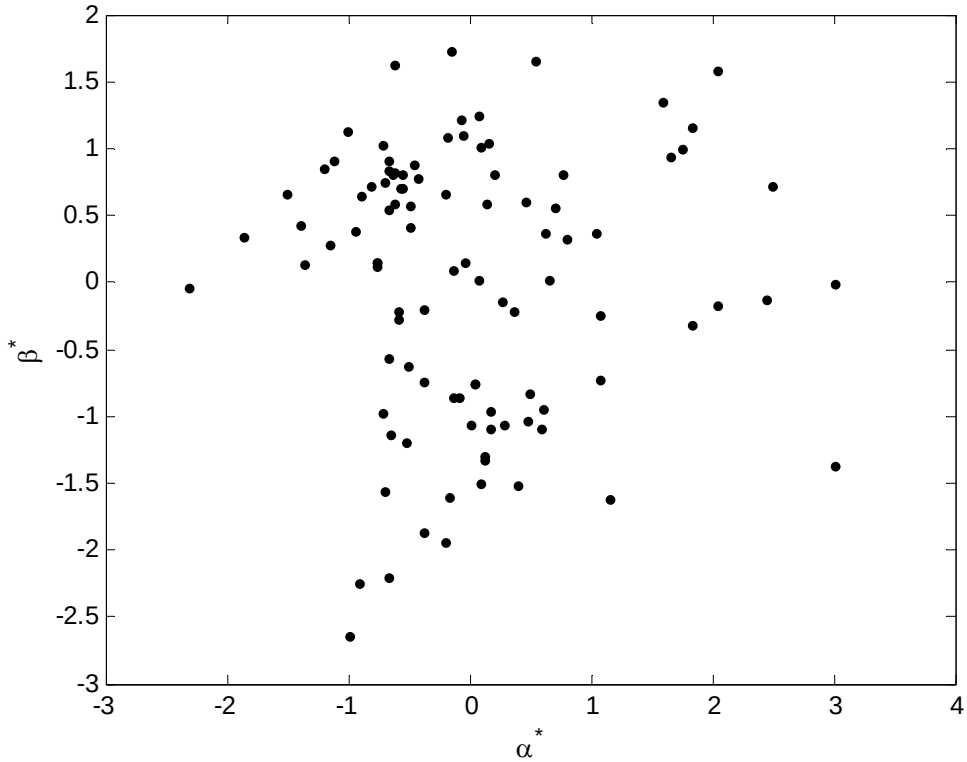


Şekil 3.1. (a) Ki-Kare Endeks Değerinin 2.33 olduğu İzdüşüm Arama Çizimi.

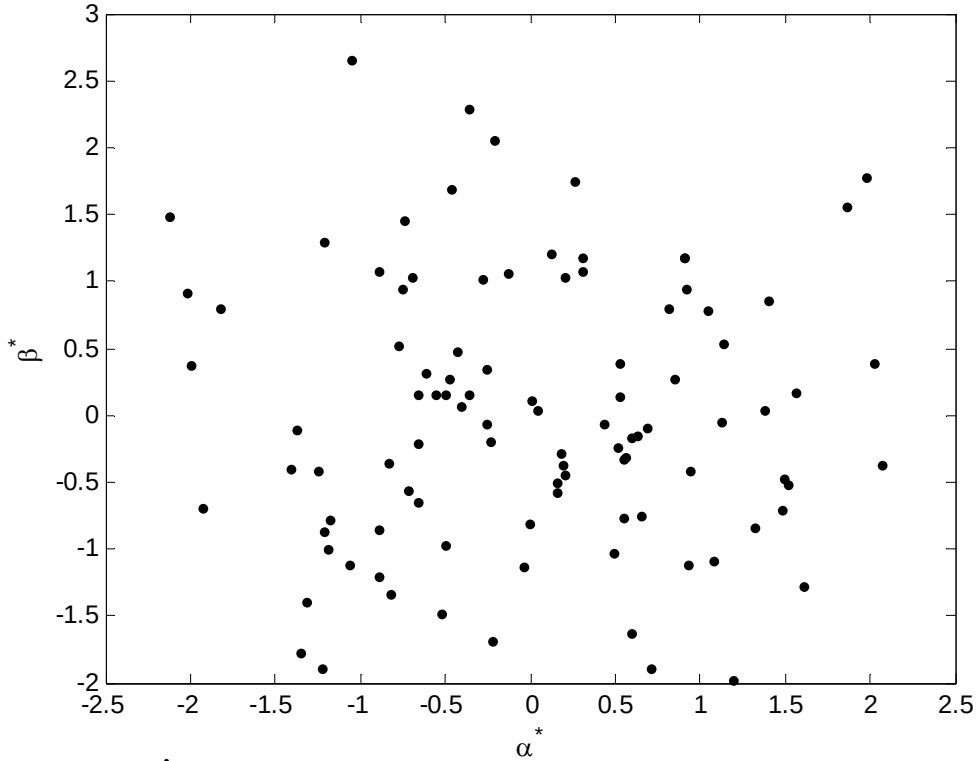


Şekil 3.1. (b) Ki-Kare Endeks Değerinin 0.48 olduğu İzdüşüm Arama Çizimi.

İlk çizimdeki ki-kare endeksi 2.33 ikinci çizimde ise ki-kare endeksi 0.48 olduğu gözlemlenmiştir. Yüksek endekse sahip üstteki çizimde verilerin daha net ayrıştığı açıkça görülmektedir. Bir türün diğerlerinden oldukça farklı dağıldığı söylenebilir ve diğer dağılımın için ayrı bir inceleme söz konusu olabilir. Dağılımı diğerlerinden farklılık gösteren tür yani birinci tür çıkartılıp geri kalan ikinci ve üçüncü türün oluşturduğu veri kümesine İris(1) adı verilip izdüşüm arama yöntemi uygulanırsa sonuç Şekil 3.2. gibi olur.



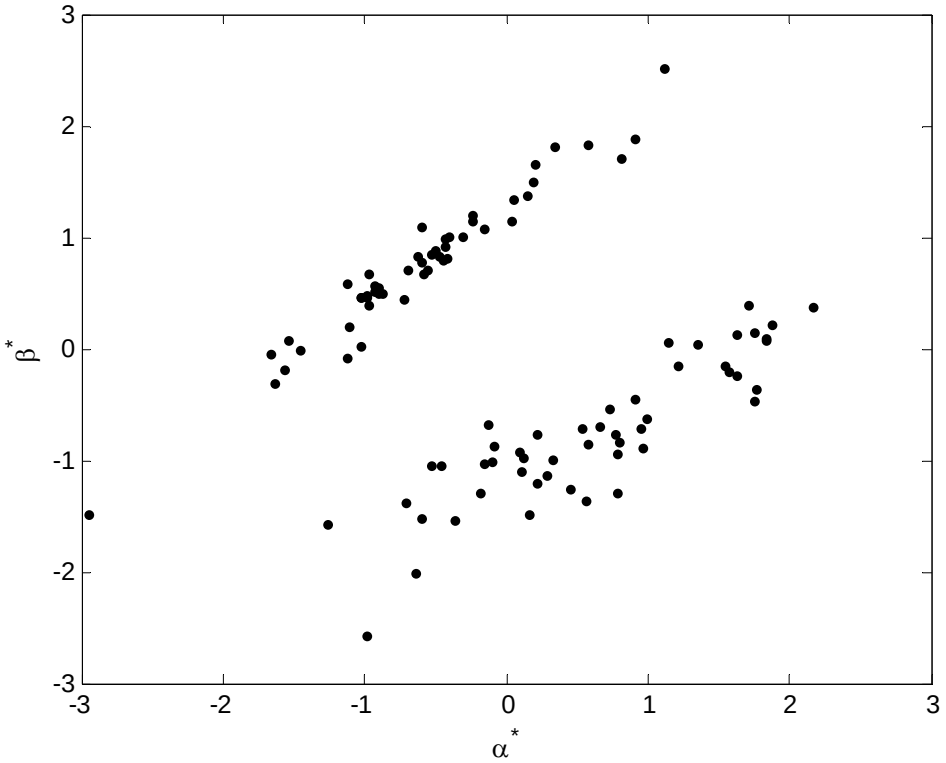
Şekil 3.2. (a) İris(1) veri kümesi için Ki-Kare Endeks Değerinin 1.13 olduğu İzdüşüm Arama Çizimi.



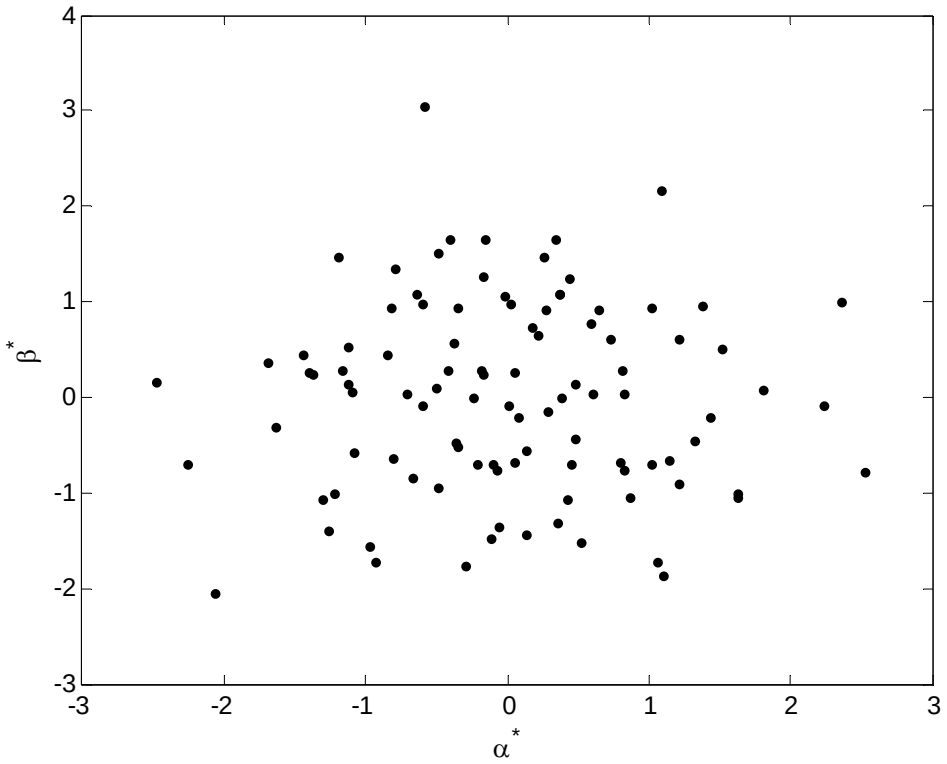
Şekil 3.2. (b) İris(1) veri kümesi için Ki-Kare Endeks Değerinin 0.77 olduğu İzdüşüm Arama Çizimi.

İlk çizimindeki ki-kare endeksi 1,13 ikinci çizimde ise ki-kare endeksi 0,77 olduğu görülmüştür. Endeks değerinin büyük olduğu çizime bakılırsa çok net olmayan fakat iki ayrı kümelenmenin olduğundan bahsedilebilir.

İlk iki çiçek türü alınarak İris (2) adı verilen veri kümesi ile yapılan uygulama sonucu izdüşüm çizimi aşağıdaki gibidir.

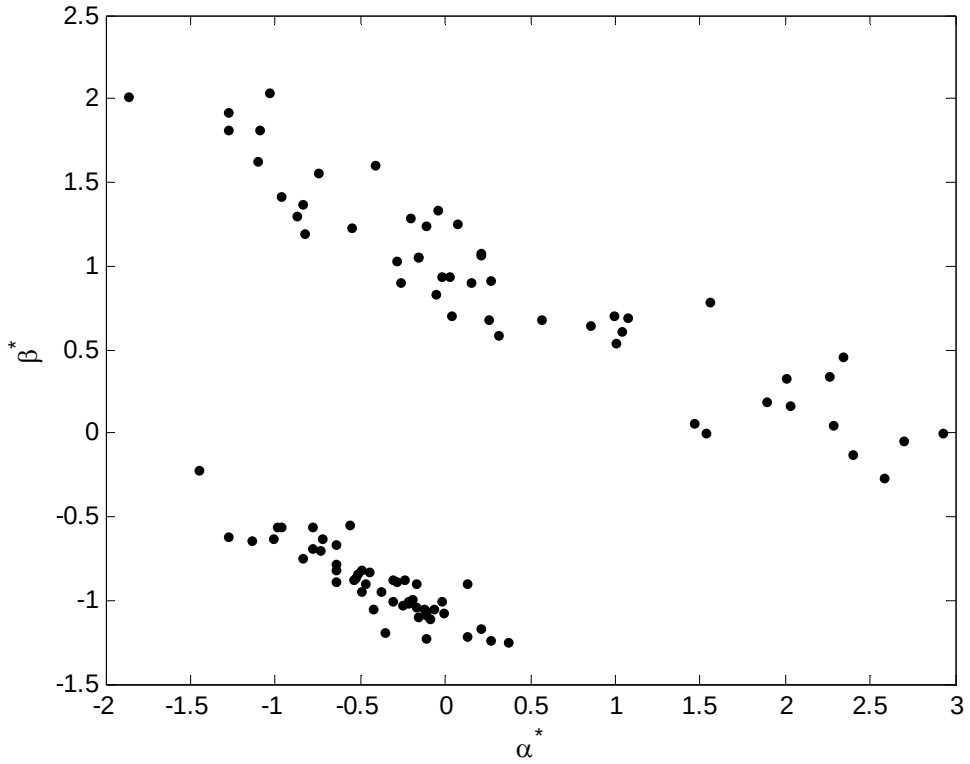


Şekil 3.3. (a). İris (2) veri kümesi için Ki-Kare Endeks Değerinin 2.10 olduğu İzdüşüm Arama Çizimi.

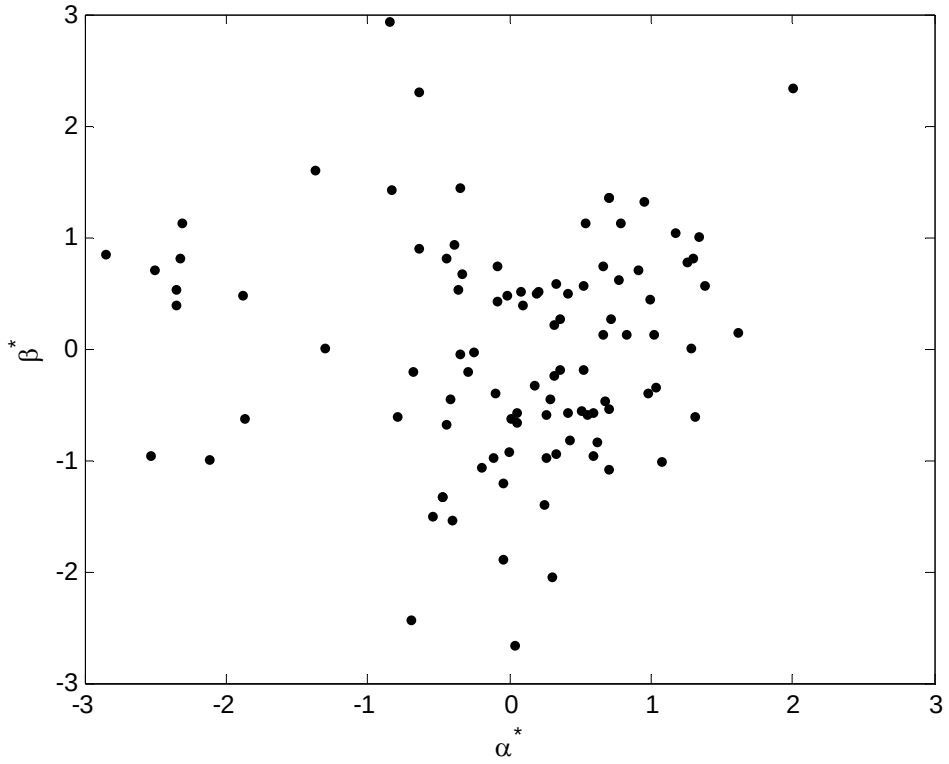


Şekil 3.3. (b). İris (2) veri kümesi için Ki-Kare Endeks Değerinin 0.83 olduğu İzdüşüm Arama Çizimi.

Endeks değerinin büyük olduğu (a)'da belirgin bir ayrışma olduğu görülmektedir. Son olarak ise birinci ve üçüncü çiçek türlerine ait verilerin oluşturduğu İris (3) veri kümesi ile yapılan uygulama sonucu şekildeki gibidir.

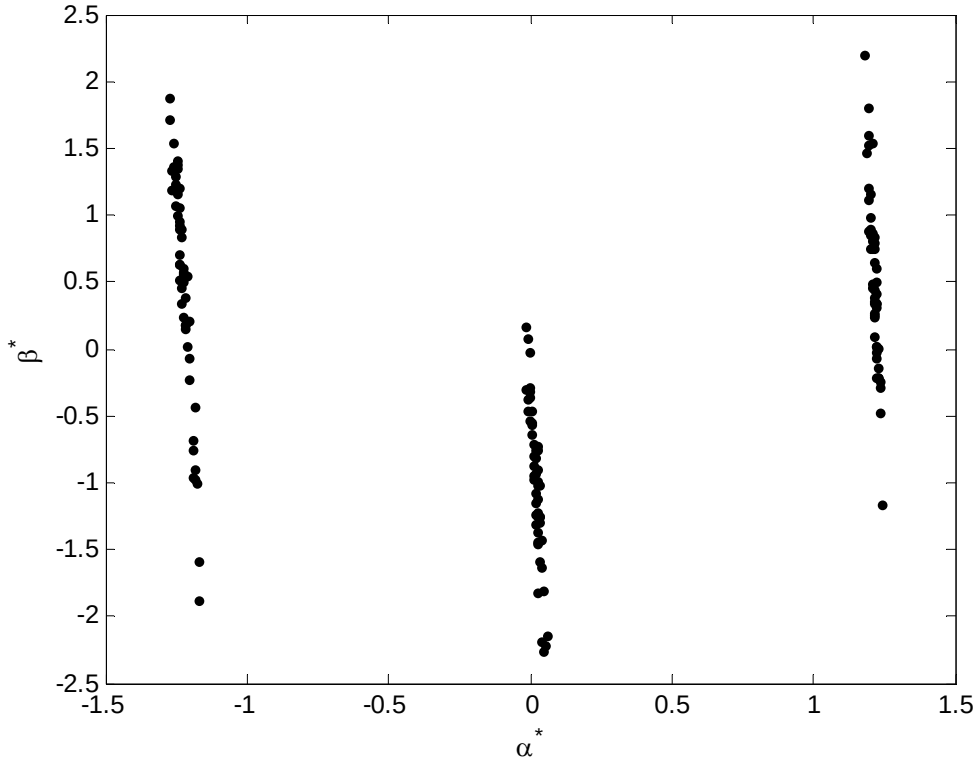


Şekil 3.4. (a) İris (3) veri kümesi için Ki-Kare Endeks Değerinin 3.57 olduğu İzdüşüm Arama Çizimi.

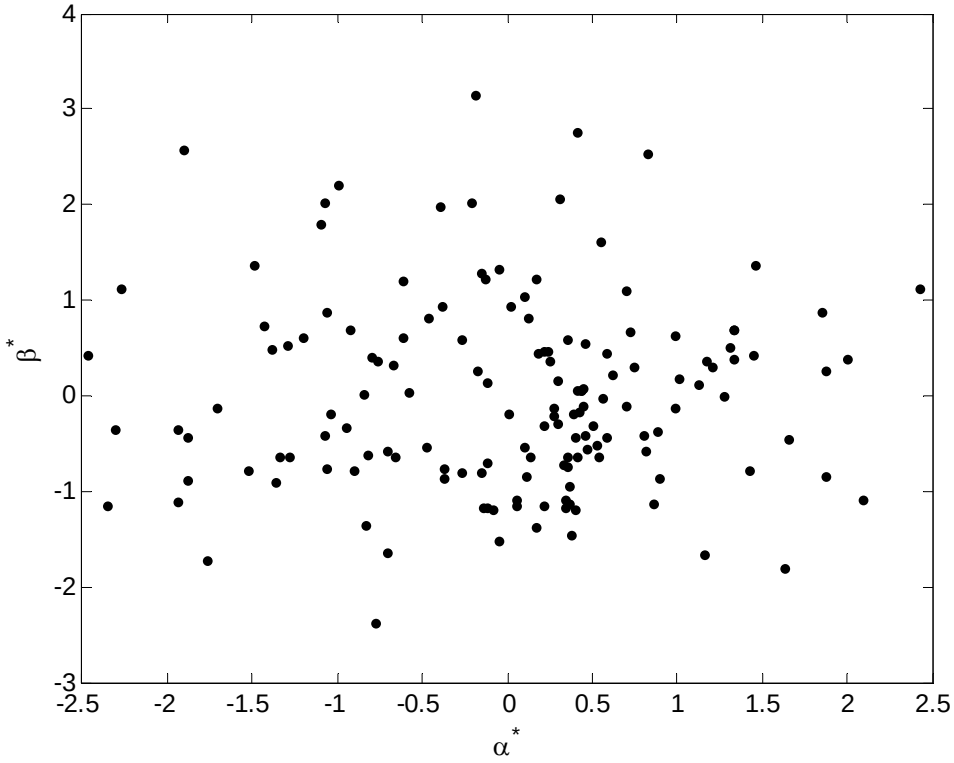


Şekil 3.4. (b) İris (3) veri kümesi için Ki-Kare Endeks Değerinin 0.80 olduğu İzdüşüm Arama Çizimi.

Endeks değerinin düşük olduğu çizimde bile ayrışmanın belirgin olduğu, endeks değerinin yüksek olduğu çizimde ise ayrışma gayet açık bir biçimde olduğu görülmektedir. İzdüşüm Arama Yöntemine göre değişkenlerin ne olduğu bilinmese de veriler hakkında herhangi bir bilgi olmasa da uygulama sonrası bir yorum yapılabilir. Bununla birlikte türlere ait çok belirleyici bir özelliğin değişken olarak alındığı İris (4) adı verilen veri kümesinde uygulama yapıldığında sonuç aşağıdaki gibi gözlemlenmiştir.



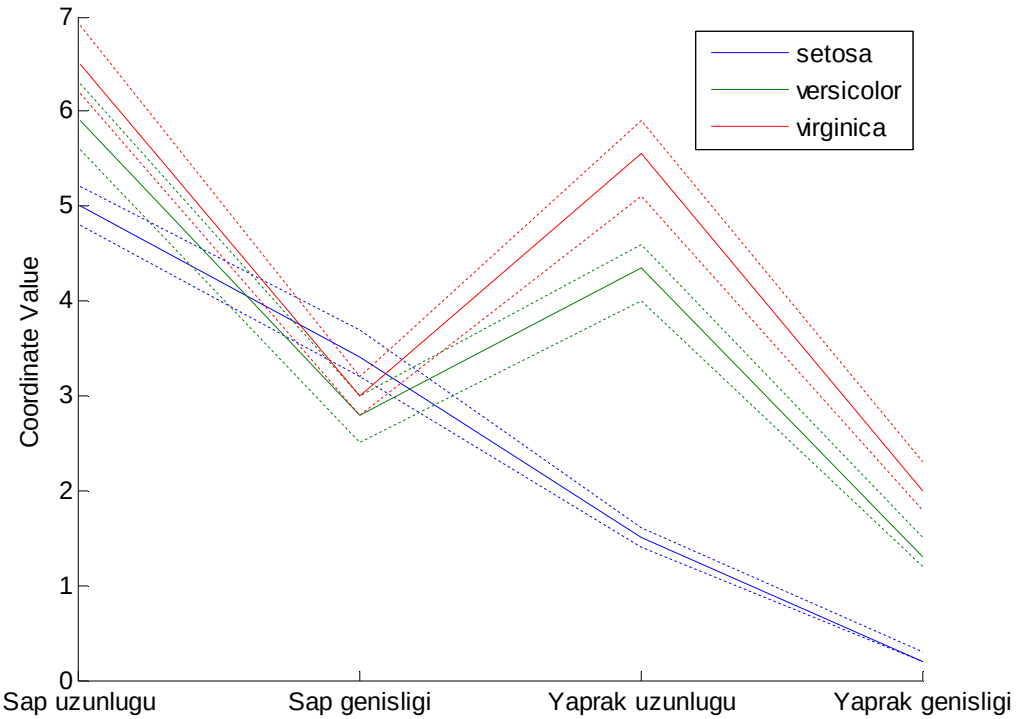
Şekil 3.5. (a) İris (4) veri kümesi için Ki-Kare Endeks Değerinin 2.85 olduğu İzdüşüm Arama Çizimi.



Şekil 3.5. (b) İris (4) veri kümesi için Ki-Kare Endeks Değerinin 0.54 olduğu İzdüşüm Arama Çizimi.

Üç farklı türden kolaylıkla bahsedebileceğimiz endeksin büyük olduğu Şekil3.5. (a)'da veri kümesinde yer alan değişkenin, açıklayıcı gücüne bağlı olarak grafik çizimi daha anlaşılır hale getirilebilir.

Aynı veri kümesine paralel koordinatlar grafik yöntemi uygulanarak yukarıdaki dağılımların benzer bir çizimi burada da açıkça görülmektedir.



Şekil 3.6. İris Verisinin Paralel Koordinatlar Grafiği.

Sap uzunluğu ile sap genişliği arasında Virginica ve Versicolor çiçek türlerinde ilişkili bir dağılım gözükmektedir. Setosa türü çiçeğin sap uzunluğu ve sap genişliği arasındaki ilişki katsayısı bir olmasa da bire yakın gözükmektedir. Yani Setosa türü çiçeğin sap uzunluğu ile sap genişliği arasında kuvvetli bir ilişki vardır. Sap genişlikleri arasında çok fark olmayan üç çiçek türünün yaprak uzunluğunda kümelendiği ve bu kümelerin birini Versicolor ve Virginica diğerini ise Setosa türü çiçeğin oluşturduğu söylenebilir. Sap genişliğinden bağımsız olarak çiçekler yaprak uzunluğuna göre farklılıklar göstermektedir. Yaprak uzunluğu ve yaprak genişliği arasında ise normal bir dağılım olduğu fakat Setosa da yine daha kuvveti bir ilişki olduğu gözlemlenmektedir. Sonuç olarak Setosa türü çiçeğin özelliklerinin diğer türlerden daha farklı olduğu söylenebilir.

3.2. UYGULAMA 2

Bu uygulama da Auto Katalog 2008' in Auto Motor özel sayısından alınan Aralık 2007 araba verileri ile çalışılmıştır. Bu veri kümesi Audi (Almanya), Bmw (Almanya), Fiat (İtalya), Ford (Almanya), Hyundai (Kore), Jaguar (İngiltere), Kia (Kore), Mercedes (Almanya), Opel (Almanya), Peugeot (Fransa), Renault (Fransa), Toyota (Japonya), Volkswagen (Almanya), Volvo (İsveç) araba markalarına ait bazı değişkenlerin olduğu 192 gözlemli bir veri kümesidir. Posse' un izdüşüm arama yöntemi için önerdiği, yapısal uzaklaştırma ile kombine ettiği algoritmasının kullanıldığı ki-kare endeks yöntemi uygulanmış ve bazı yorumlamalara gidilmiştir.

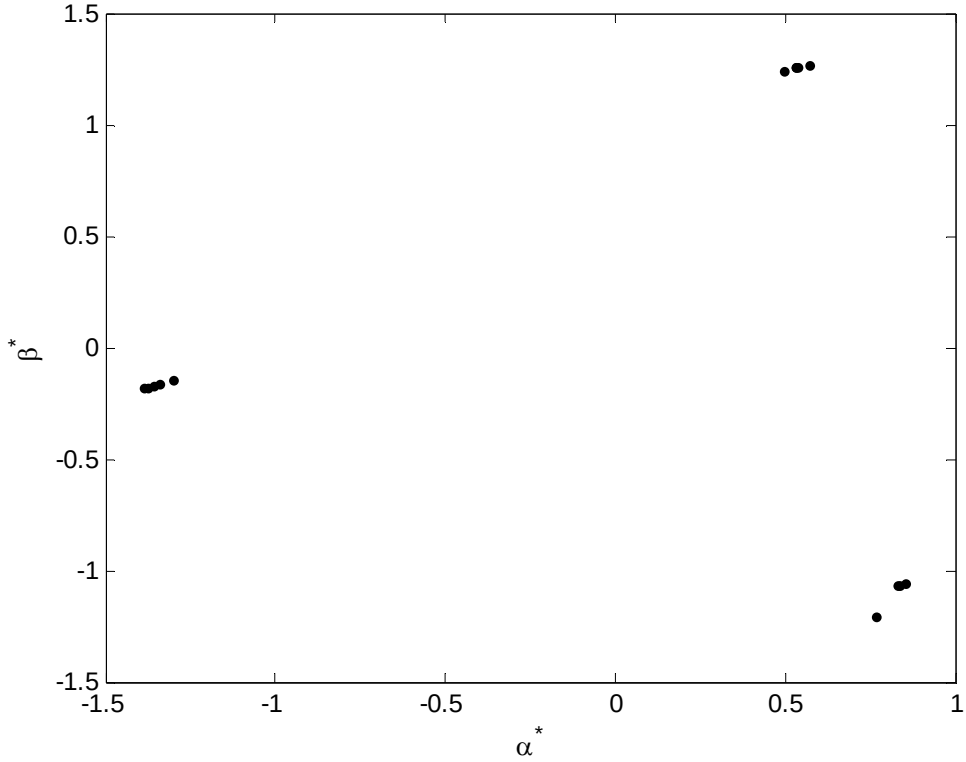
Bağımlı değişkenin arabaların yakıt tüketim miktarı olarak alındığı bu veri kümesinde, araba markalarının hangi ülkelerde üretildiği markaların modellerine göre alınan silindir tipi, silindir sayısı, silindir hacmi, çap, sıkıştırma oranı (kompresyon), beygir gücü, tork, boş ağırlığı gibi bağımsız değişkenlerden elde edilen veriler ile çalışılmıştır. Değişkenlerden hacim, beygir gücü, tork ve boş ağırlık dağılımları çarpık olduğundan dolayı "log dönüşümü" uygulanmıştır. Ayrıca iki farklı türe sahip olan silindir tipi değişkeni 0 ve 1 göstermelik değişken olarak eklenmiştir.

Başlangıç olarak 3 değişken (Hacim, Güç ve Boş Ağırlık) ele alınarak Otomobil (1) verisi olarak adlandırdığımız alt küme aşağıdaki gibidir.

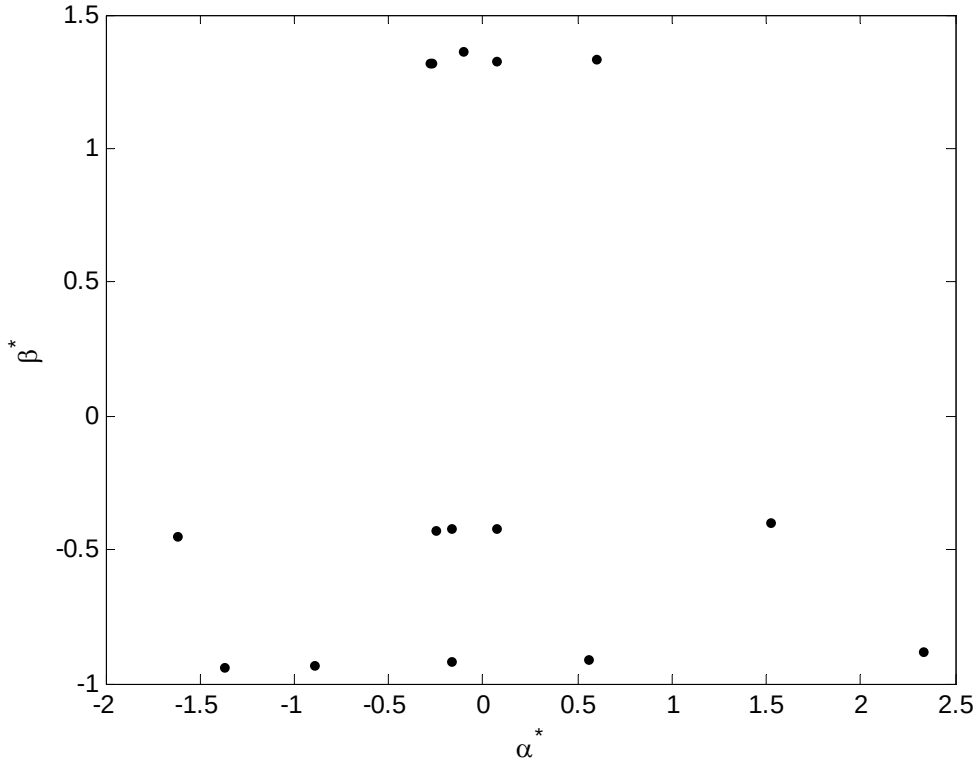
Marka	Log Hacim	Log Güç	Boş ağırlık
Hyundai	0,34	4,57	1.133,00
kia	0,34	4,57	1.154,00
Hyundai	0,34	4,57	1.155,00
Hyundai	0,34	4,57	1.100,00
Fiat	0,34	4,55	1.135,00
Mercedes	1,7	5,96	1.785,00
Mercedes	1,7	5,96	1.895,00
Mercedes	1,7	5,96	1.940,00
Mercedes	1,7	5,96	1.985,00
Mercedes	1,7	5,96	2.015,00
BMW	1,1	5,61	1.605,00
BMW	1,1	5,61	1.715,00
BMW	1,1	5,61	1.695,00
BMW	1,1	5,61	1.710,00
BMW	1,1	5,61	1.800,00

Tablo 3.1. Otomobil (1) veri tablosu.

Otomobil (1) veri kümesine izdüşüm arama yöntemi uygulanırsa sonuç Şekil 3.7. (a) ve (b) gibidir.

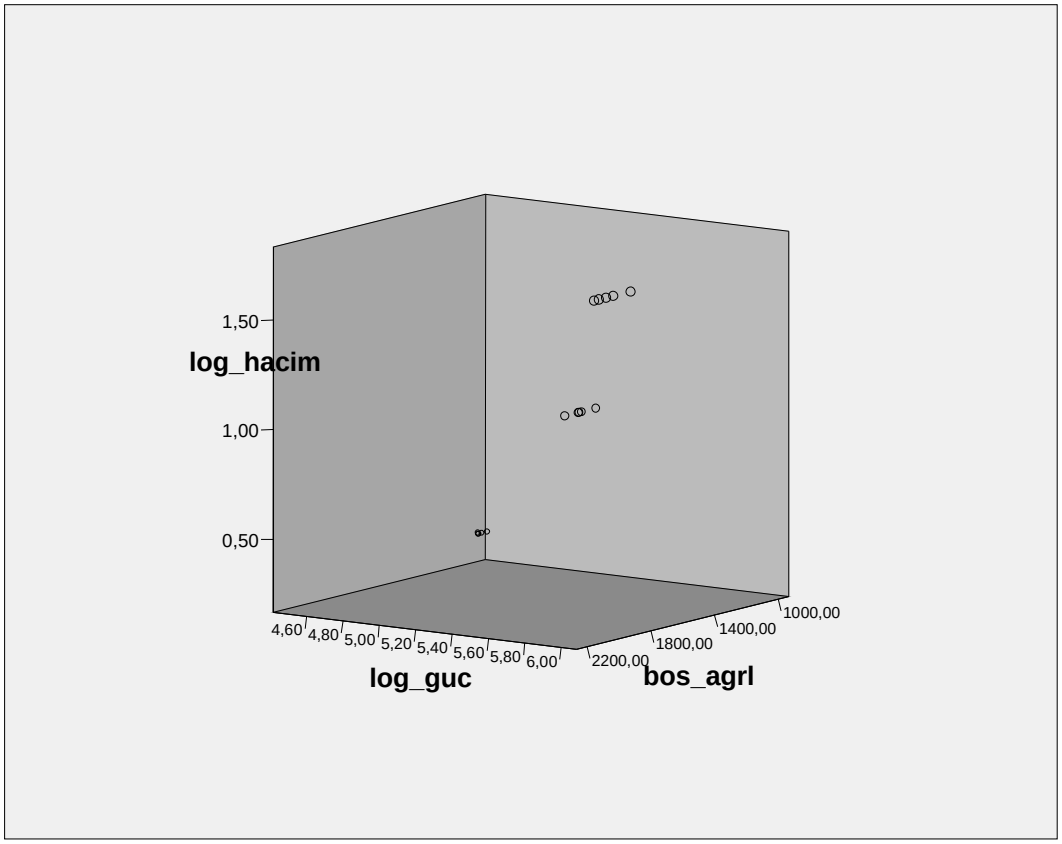


Şekil 3.7. (a) Otomobil (1) veri kümesi için Ki-Kare Endeks Değerinin 11.73 olduğu İzdüşüm Arama Çizimi.



Şekil 3.7. (b) Otomobil (1) veri kümesi için Ki-Kare Endeks Değerinin 4.95 olduğu İzdüşüm Arama Çizimi.

Şekil 3.7 (a)’ da ki-kare endeks değerinin daha büyük ve kümelenmenin Şekil 3.7 (b)’ ye göre daha net olduğu görülmektedir. Değişken ve gözlem sayımız az olduğu için veri tablosuna dönüp baktığımızda hacim ve güç değişkenlerine göre 3 kümelenmenin olması gerektiği açıkça söylenebilir bu sebeple şekil 3.7. (b)’ de endeksin düşük olmasına rağmen yatay 3 farklı küme olduğu şeklinde bir yorumlama yapılabilir. Aynı veri kümesi kullanılarak SPSS paket programında çizilen saçılım dağılım grafiği aşağıdaki gibidir.



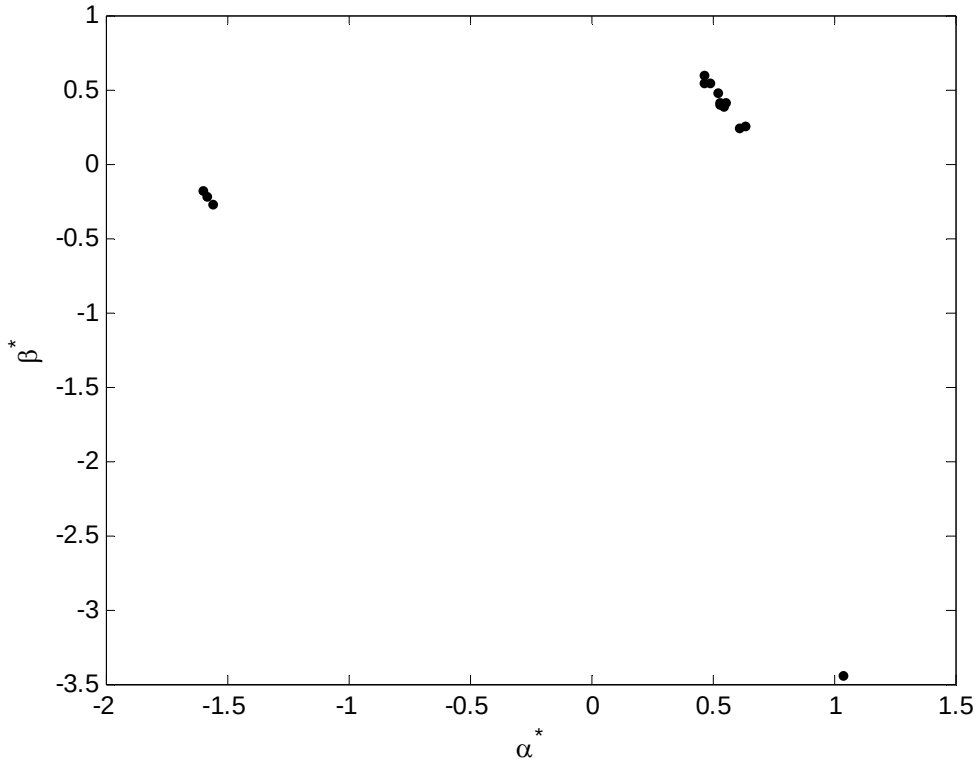
Şekil 3.8. Otomobil (1) Verileri Üzerinden Saçılım Grafiği.

Şekil 3.8’de de kümelenmenin oldukça net olduğu görülmektedir. Otomobil (1) veri kümesine silindir tipi değişkeni eklenirse veri kümesi otomobil (2) tablo 3.2. gibi olur.

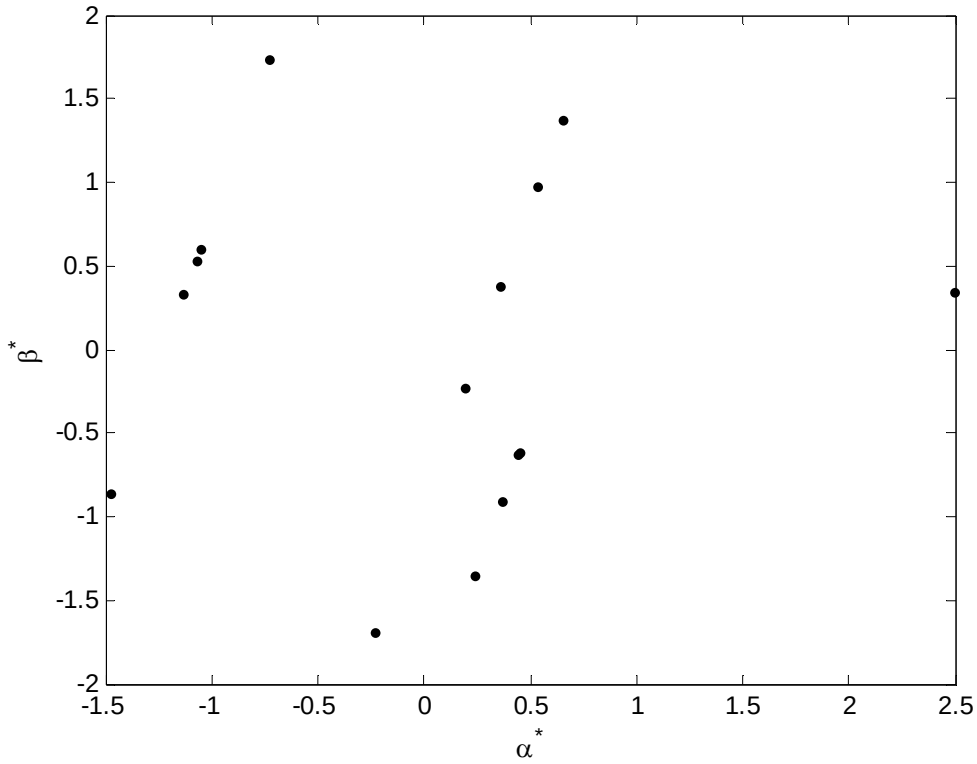
Marka	Silindir Tipi	Log Hacim	Log Güç	Boş ağırlık
Hyundai	1	0,34	4,57	1.133,00
Kia	1	0,34	4,57	1.154,00
Hyundai	1	0,34	4,57	1.155,00
Hyundai	1	0,34	4,57	1.100,00
Fiat	1	0,34	4,55	1.135,00
Mercedes	0	1,7	5,96	1.785,00
Mercedes	0	1,7	5,96	1.895,00
Mercedes	0	1,7	5,96	1.940,00
Mercedes	0	1,7	5,96	1.985,00
Mercedes	0	1,7	5,96	2.015,00
BMW	1	1,1	5,61	1.605,00
BMW	1	1,1	5,61	1.715,00
BMW	1	1,1	5,61	1.695,00
BMW	1	1,1	5,61	1.710,00
BMW	1	1,1	5,61	1.800,00

Tablo 3.2. Otomobil(2) veri tablosu.

Bu veri kümesini izdüşüm arama yöntemini uygulasak sonuç aşağıdaki gibi olur.

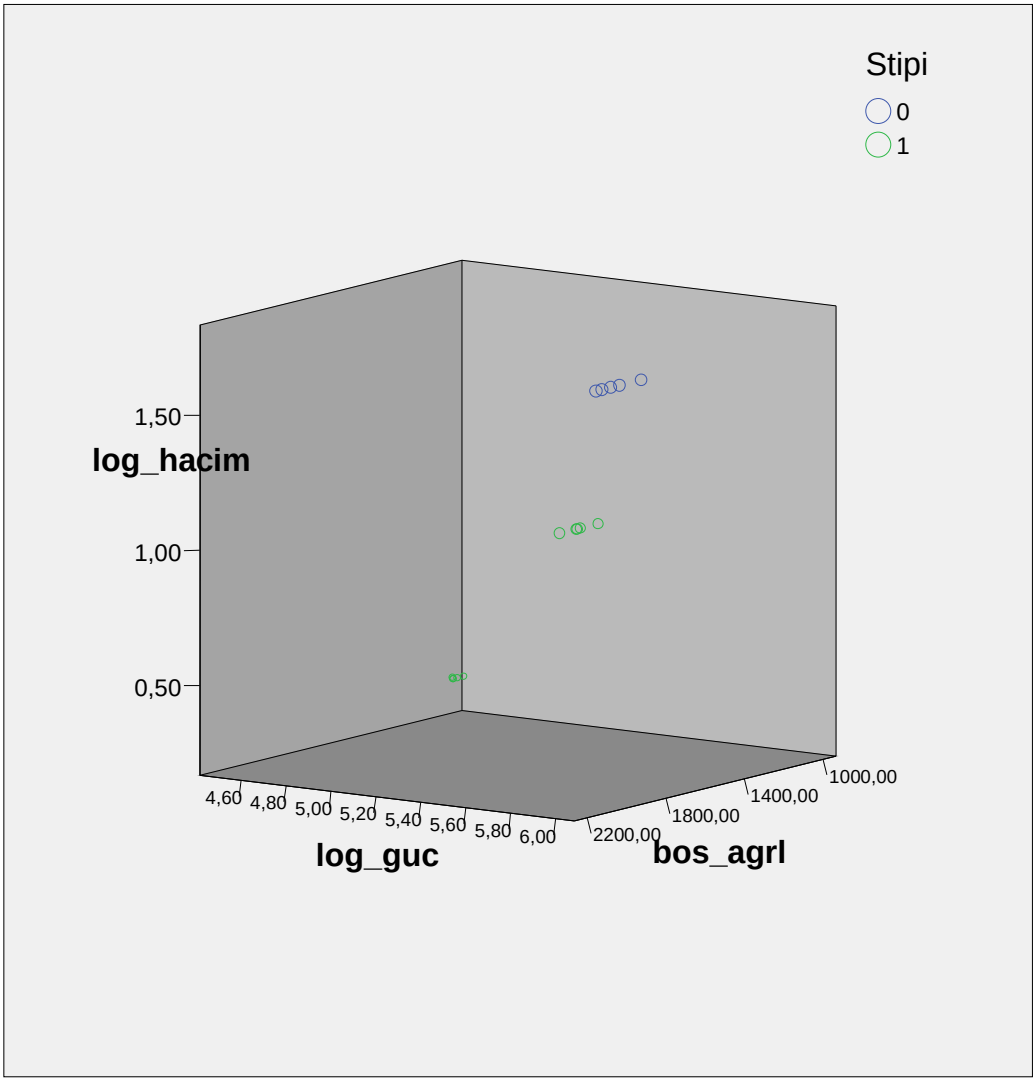


Şekil 3.9. (a) Otomobil (2) veri kümesi için Ki-Kare Endeks Değerinin 18.24 olduğu İzdüşüm Arama Çizimi.



Şekil 3.9. (b) Otomobil (2) veri kümesi için Ki-Kare Endeks Değerinin 6.45 olduğu İzdüşüm Arama Çizimi.

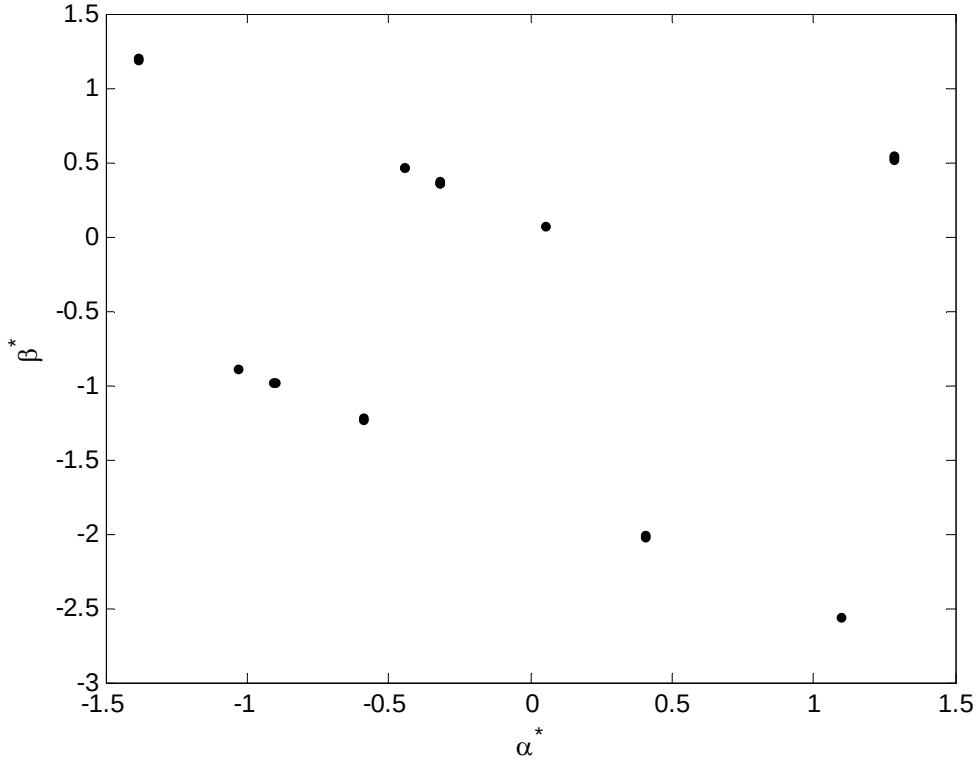
Kümelenmenin daha net olduğu şekil 3.9. (a)' da aykırı bir değerin varlığından ya da aynı gözlem değerlerinin üst üste çakışmasından kaynaklanan bir yığıntıdan bahsedilebilir, endeksin daha düşük olduğu şekil 3.9.(b)' de de bir aykırı değer olarak görülmektedir. Bu uygulama da verilerimize bakıp inceleyebildiğimiz için o noktanın üst üste çakışan değişken değerlerinden kaynaklandığını söyleyebiliriz. Aynı veri kümesinin SPSS paket programında çizilmiş saçılım dağılımına bakacak olursak,



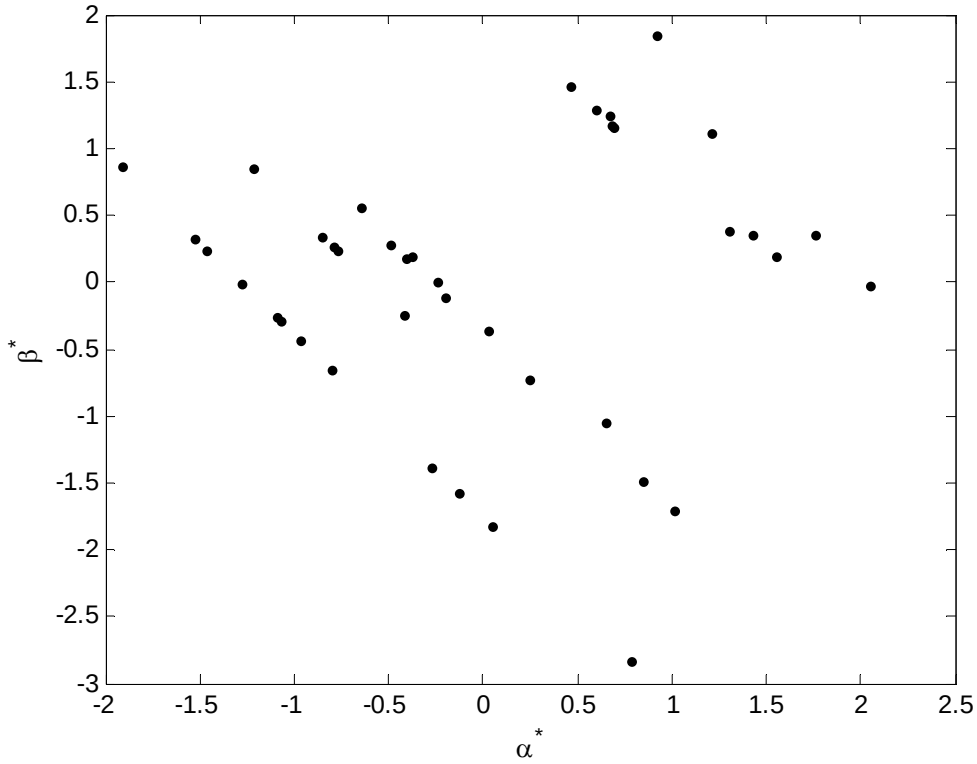
Şekil 3.10. Otomobil (2) Verileri Üzerinden Saçılım Grafiği.

şekil 3.10’ da kümelenmenin gayet net olduğu ve aykırı değerin olmadığı görülmektedir. Veri kümesine baktığımızda aykırı değer ile karşılaşılmamıştır. Sayıca az veri kümesi ile SPSS paket programının daha iyi sonuç verdiği, izdüşüm arama yönteminin ise daha ayırt edici özelliğinden dolayı daha karmaşık bir yapı sunduğu sonucuna varılmıştır ve izdüşüm arama yönteminin çok boyutlu veri kümelerinde daha iyi sonuç verdiği gözlemlenmiştir. Şekil 3.7. (a) ve şekil 3.9. (a) karşılaştırılarak, çok belirleyici bir değişken eklendiği zaman (silindir tipi) izdüşüm arama sonucunun çok fazla değişmediği söylenebilir.

Kümeleme yapılarak seçilen 40 gözlem ile oluşturulan otomobil (3) veri kümesine izdüşüm arama yöntemi 3 değişken (hacim, güç ve boş ağırlık) üzerinden uygulandığında sonuç Şekil 3.11. (a) ve (b) gibi olur.

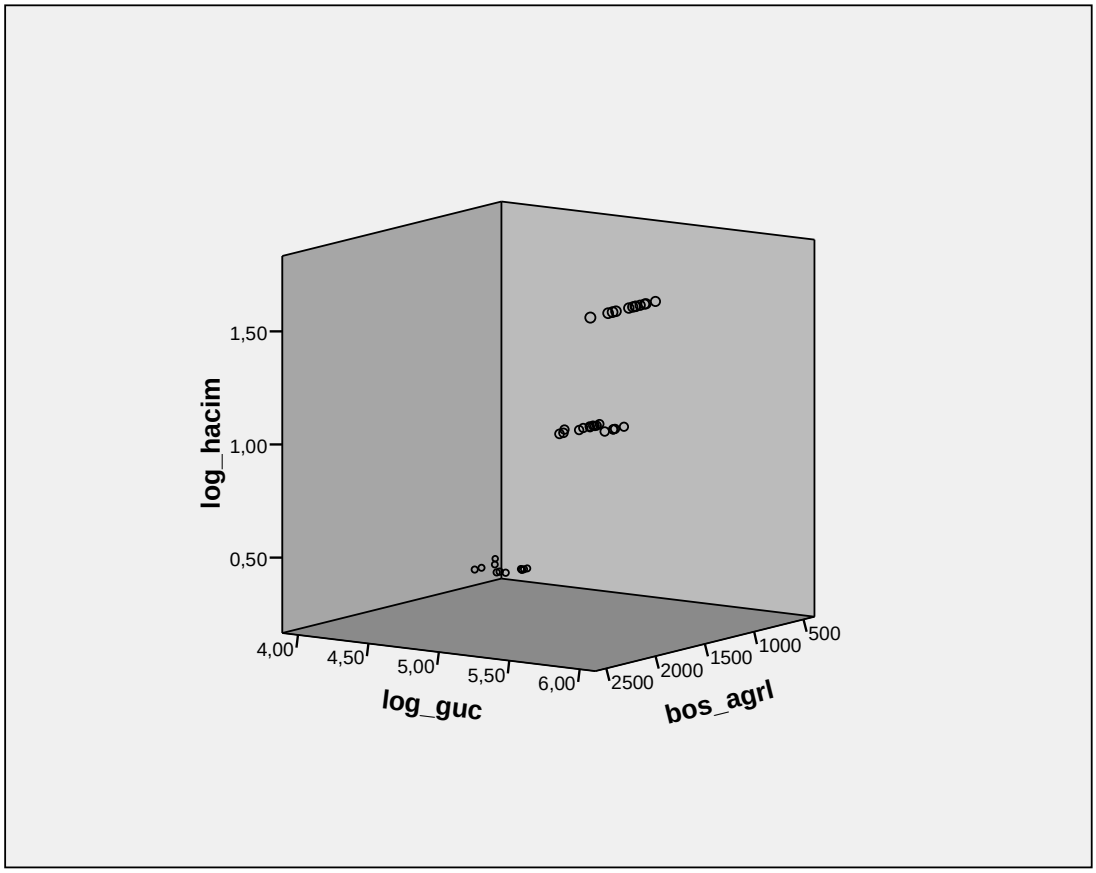


Şekil 3.11. (a) Otomobil (3) veri kümesi için Ki-Kare Endeks Değerinin 1.58 olduğu İzdüşüm Arama Çizimi.



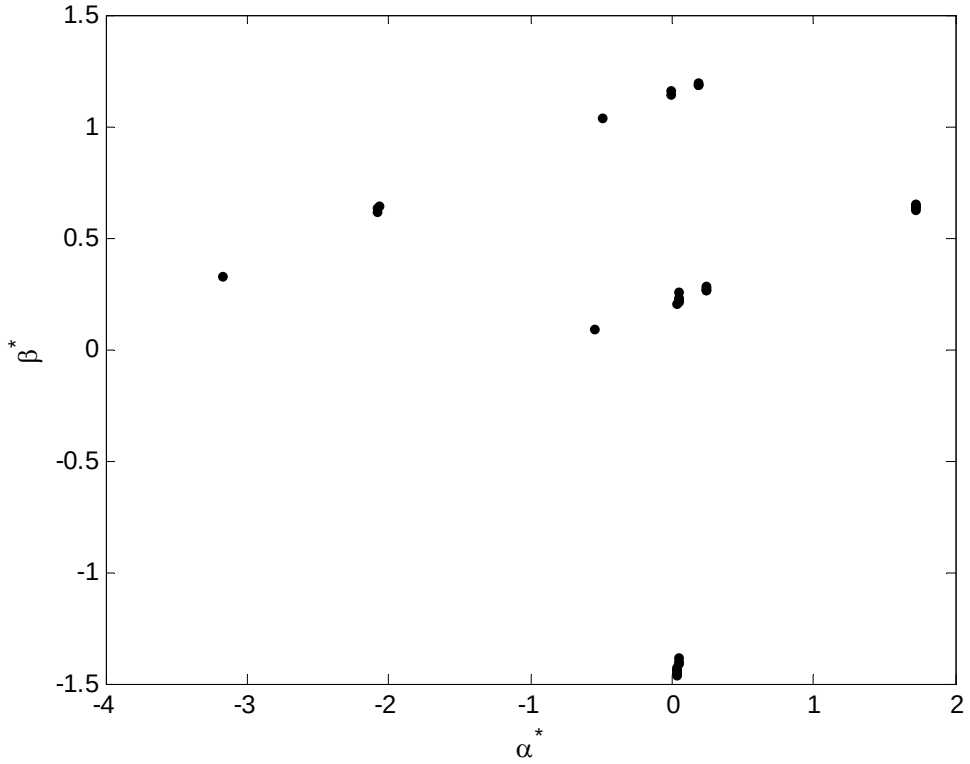
Şekil 3.11. (b) Otomobil (3) veri kümesi için Ki-Kare Endeks Değerinin 7.86 olduğu İzdüşüm Arama Çizimi.

Aynı veri kümesinin bir de SPSS paket programında saçılım grafiğini çizdirirsek şekil 3.12.' de görülen kümelenmenin izdüşüm arma sonucuna göre çok daha belirgindir.

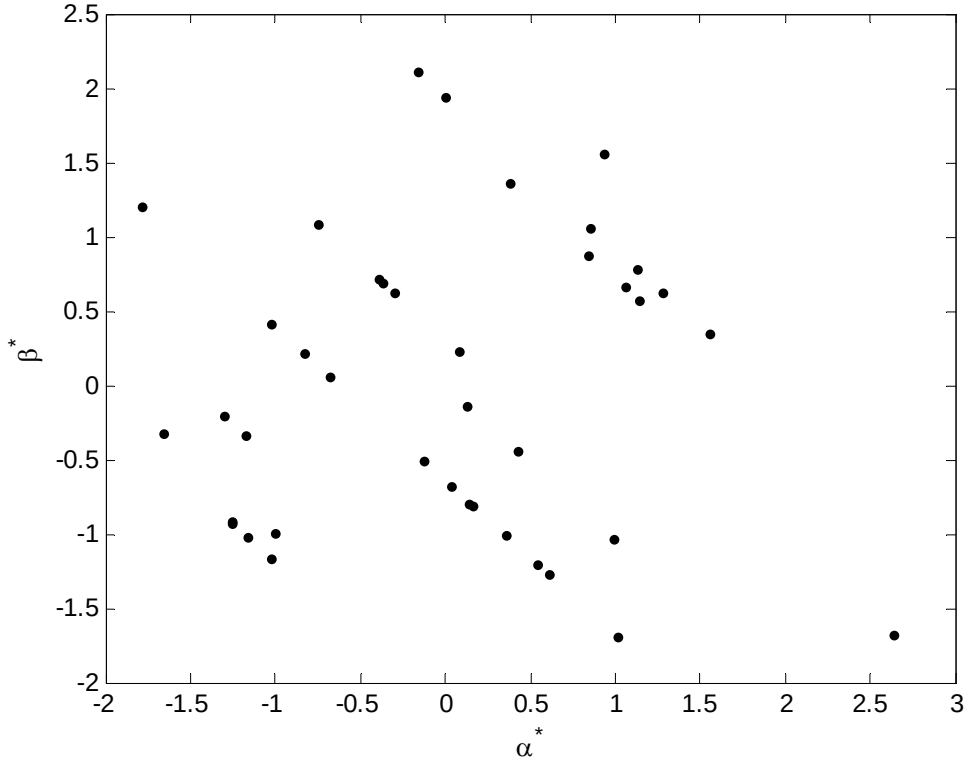


Şekil 3.12. Otomobil (3) Verileri Üzerinden Saçılım Grafiği.

Bu veri kümesine silindir tipi gibi çok belirleyici bir değişken eklendiğinde otomobil (4) veri kümesi için ise sonuç aşağıdaki gibi olur.

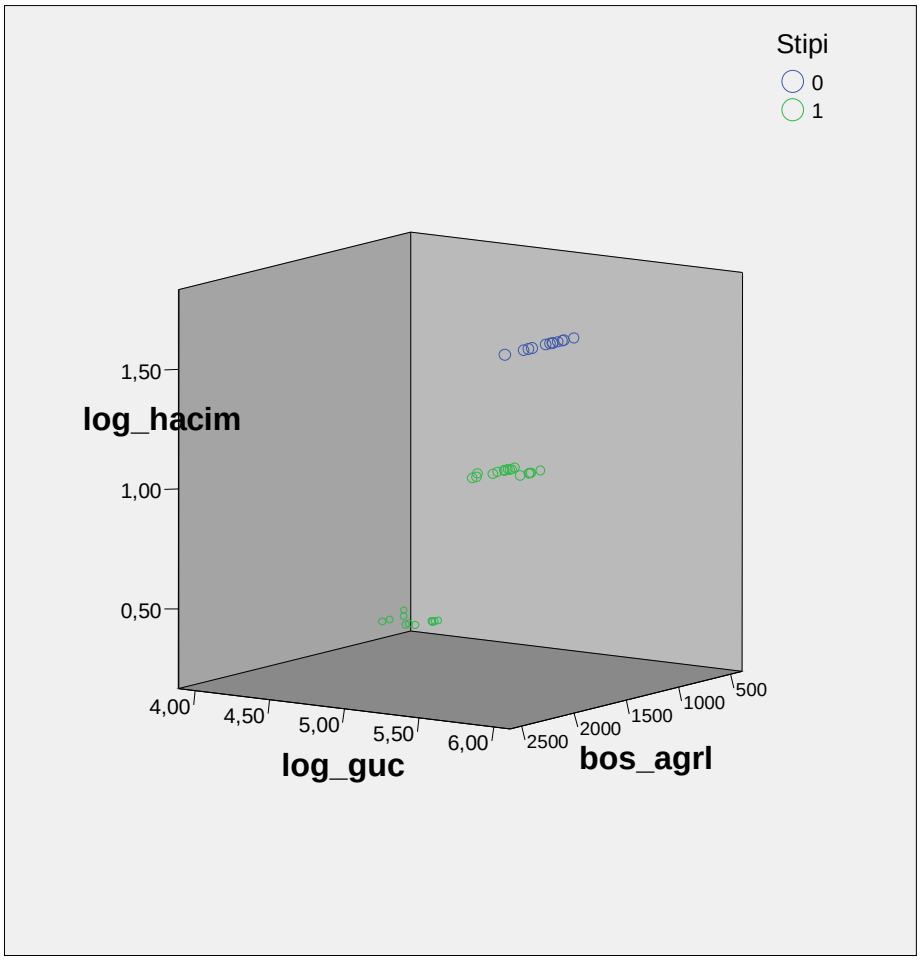


Şekil 3.13. (a) Otomobil (4) veri kümesi için Ki-Kare Endeks Değerinin 9.79 olduğu İzdüşüm Arama Çizimi.



Şekil 3.13. (b) Otomobil (4) veri kümesi için Ki-Kare Endeks Değerinin 2.19 olduğu İzdüşüm Arama Çizimi.

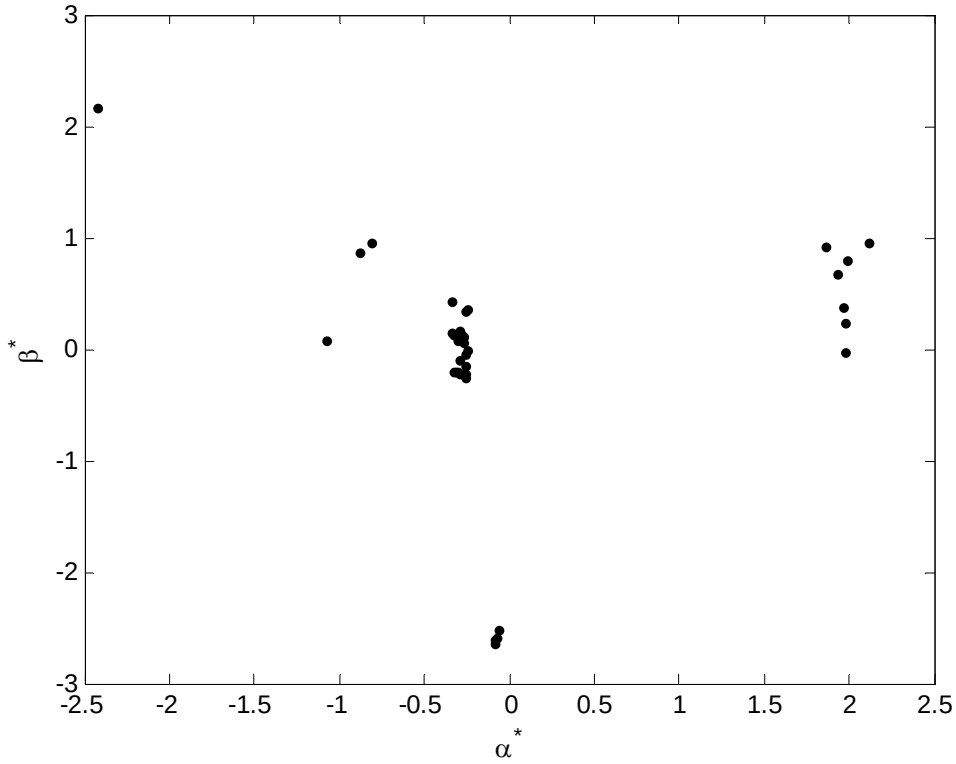
Şekil 3.13 (a)’ da kümelenmenin olduğundan bahsedilebilirken Şekil 3.13 (b)’ de daha karmaşık bir yapı bulunmaktadır. Otomobil(4) veri kümesi aynı değişkenler alınarak SPSS paket programındaki saçılım grafiği aşağıdaki gibidir.



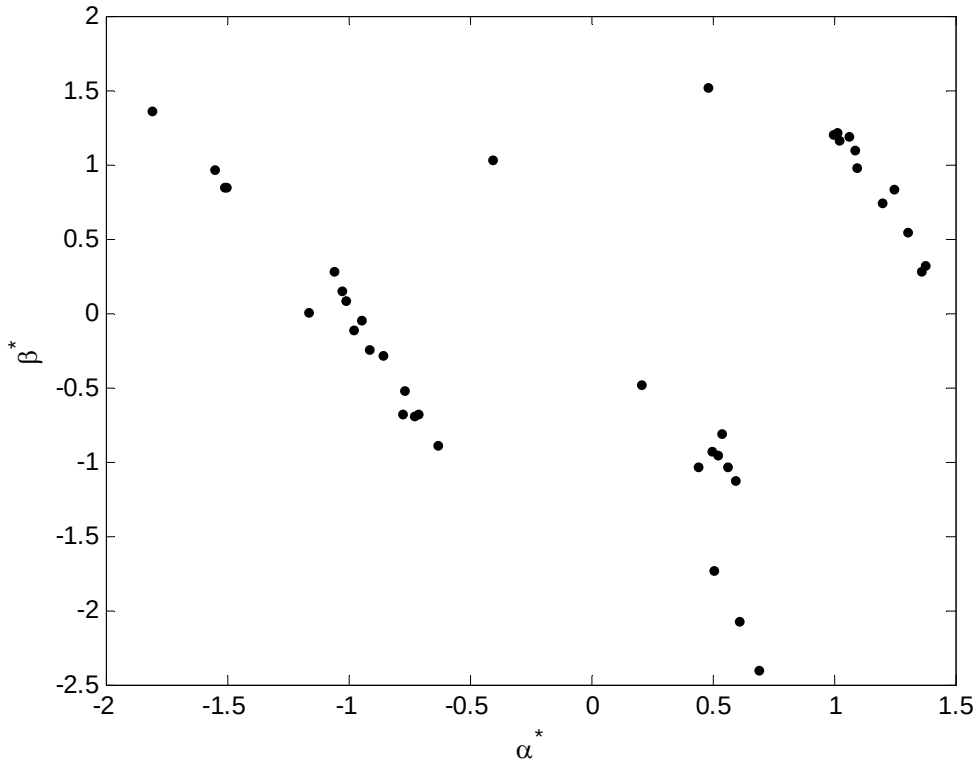
Şekil 3.14. Otomobil (4) Verileri Üzerinden Saçılım Grafiği.

Kümelenmenin net olduğu ve farklı renkler ile belirtilen 4. değişkeninde (silindir tipi) gayet kümelenmenin içinde yer aldığı gözlemlenmiştir.

Bu yorumlamalardan izdüşüm arama yönteminin düşük boyutlu veri kümelerinde çok iyi sonuç vermediği gözlemlenmiştir. 40 gözlem üzerinden hacim, güç, boş ağırlık, silindir sayısı, hızlanma, çap ve sıkıştırma oranı değişkenleri eklenerek elde edilen otomobil (5) veri kümesi ile uygulama yapıldığında sonuç Şekil 3.15. (a) ve (b) gibidir.

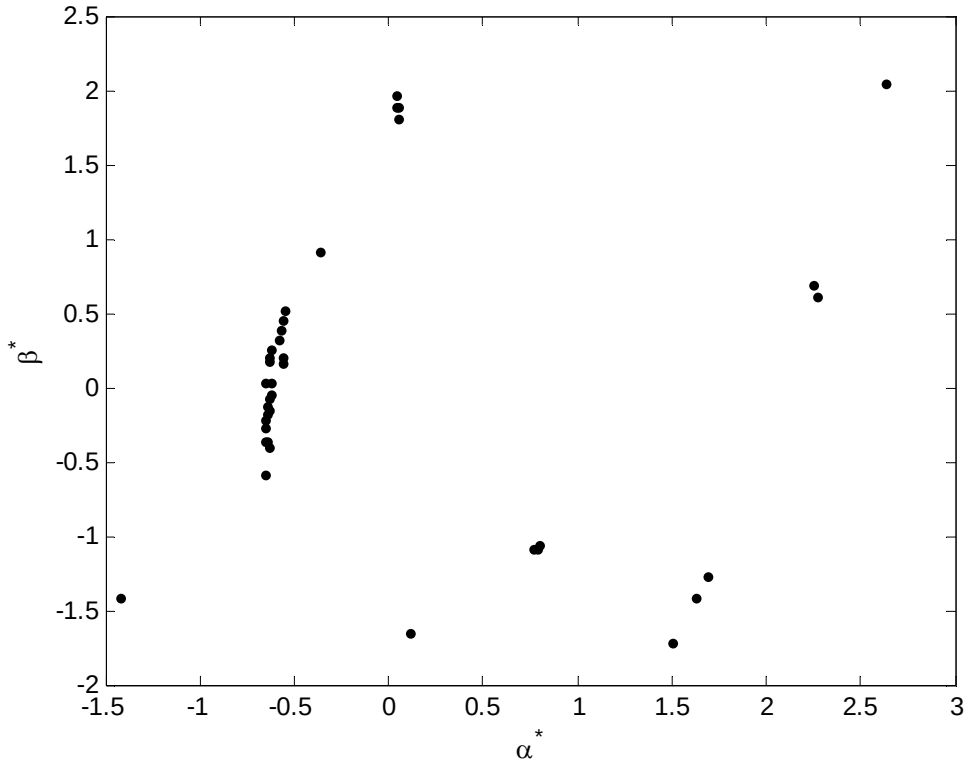


Şekil 3.15. (a) Otomobil (5) veri kümesi için Ki-Kare Endeks Değerinin 14.08 olduğu İzdüşüm Arama Çizimi.

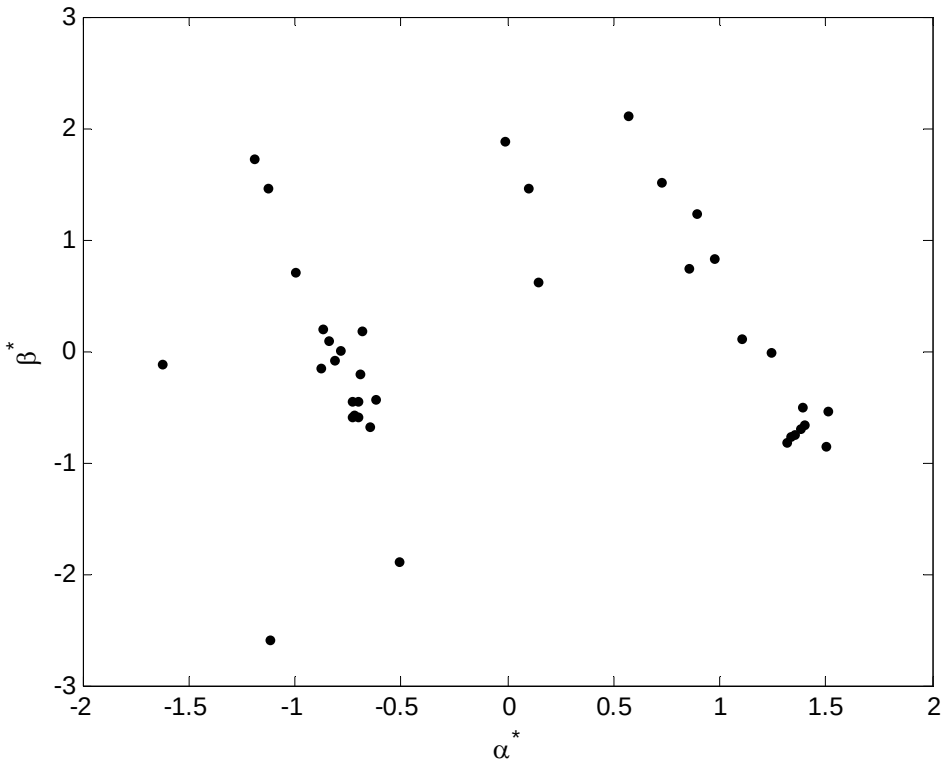


Şekil 3.15. (b) Otomobil (5) veri kümesi için Ki-Kare Endeks Değerinin 2.86 olduğu İzdüşüm Arama Çizimi.

Beklenen üçlü kümelenmenin şekil 3.15. (a)' da görüldüğü fakat bir aykırı değerin olduğu gözlenebilmektedir. Regresyon Analizi ile elde edilen Cook uzaklıklarına bakıldığında bu değerin 4.gözlemden kaynaklandığı sonucuna varılmıştır bu gözlem çıkartılarak uygulama yapıldığında sonuç aşağıdaki gibi olur.



Şekil 3.16. (a) Otomobil (5) veri kümesi için Ki-Kare Endeks Değerinin 6.38 olduğu İzdüşüm Arama Çizimi.



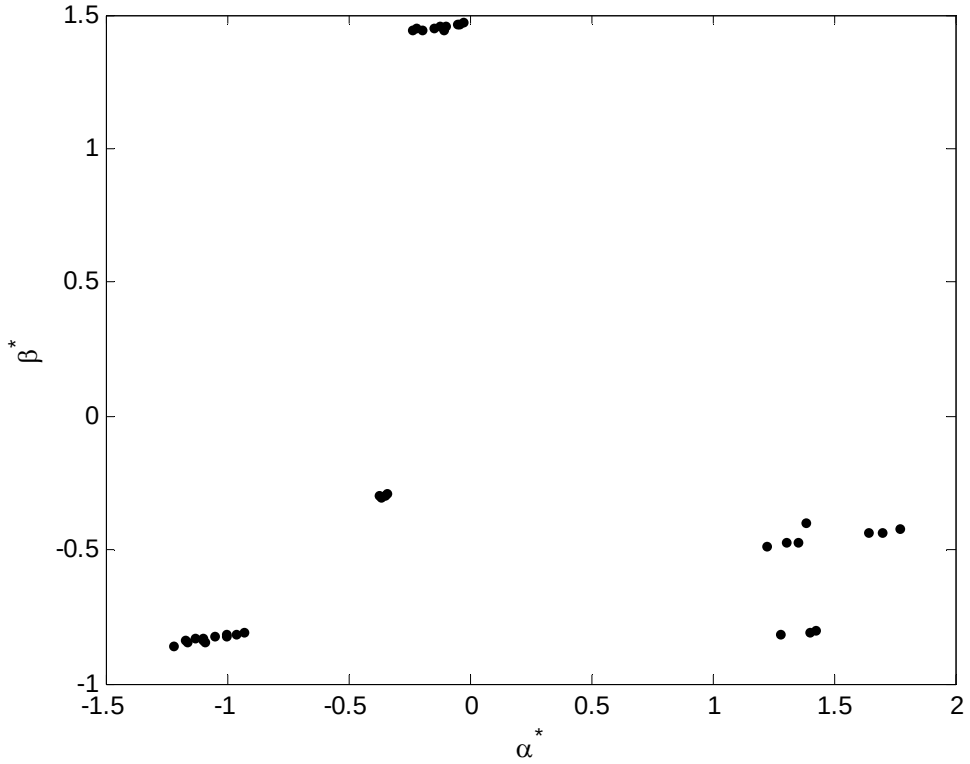
Şekil 3.16. (b) Otomobil (5) veri kümesi için Ki-Kare Endeks Değerinin 2.55 olduğu İzdüşüm Arama Çizimi.

Şekil 3.15. (a) ve Şekil 3.16 (a)' ya bakıldığında yöntemin aykırı değerin varlığı durumunda büyük bir oranda olmasa da yine de etkilendiği söylenebilir.
Veri kümemizde güçlü bir çoklu bağlantı olduğu Tablo 3.2' de görülmektedir.

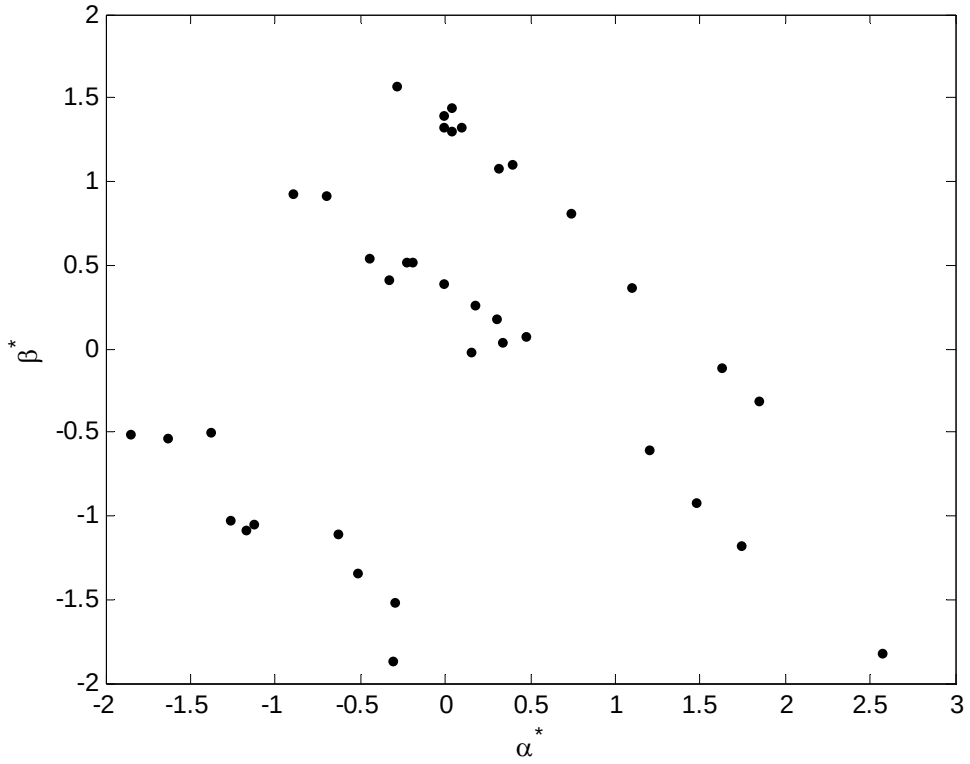
Değişken İsimleri	Varyans Şişme Oranı
Log_Hacim	4197,561
Log_Güç	268,941
Boş Ağırlık	29,934
Silindir Sayısı	3778,401
Log_Hızlanma	128,946
Çap	136,470
Sıkıştırma Oranı	8,489

Tablo 3.2. Otomobil veri kümesi varyans şişme oranları.

Çoklu bağlantı giderilerek boş ağırlık, hızlanma ve sıkıştırma oranı değişkenlerinden oluşan otomobil(6) veri kümesine izdüşüm arama yöntemi uygulandığında sonuç aşağıdaki gibi çıkmıştır.



Şekil 3.17. (a) Otomobil (6) veri kümesi için Ki-Kare Endeks Değerinin 8.28 olduğu İzdüşüm Arama Çizimi.

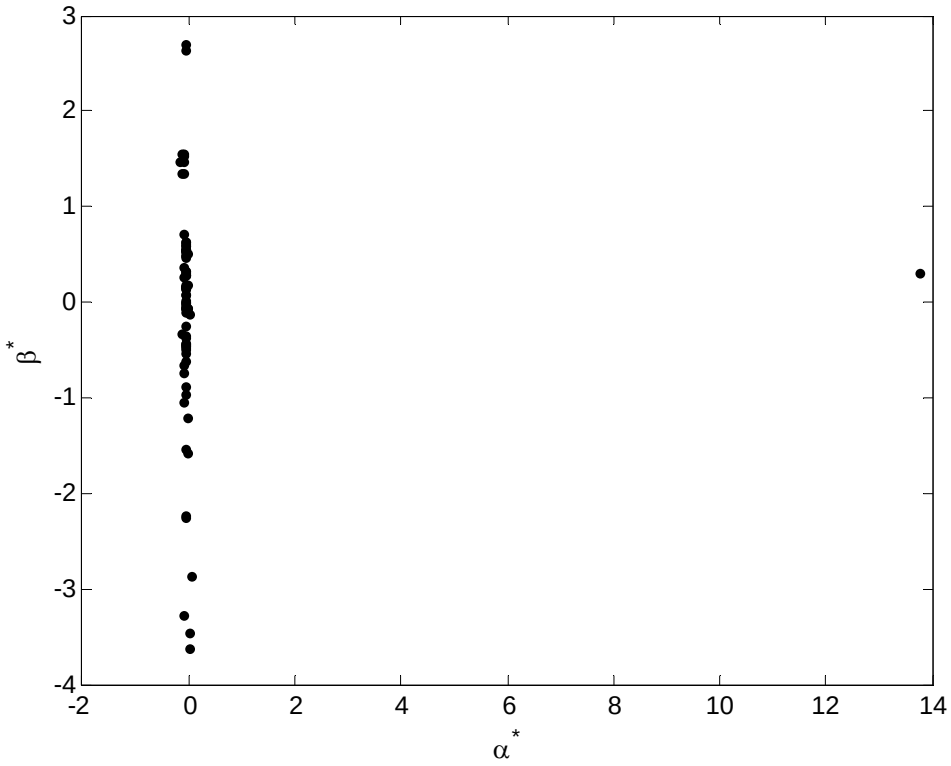


Şekil 3.17. (b) Otomobil (6) veri kümesi için Ki-Kare Endeks Değerinin 1.50 olduğu İzdüşüm Arama Çizimi.

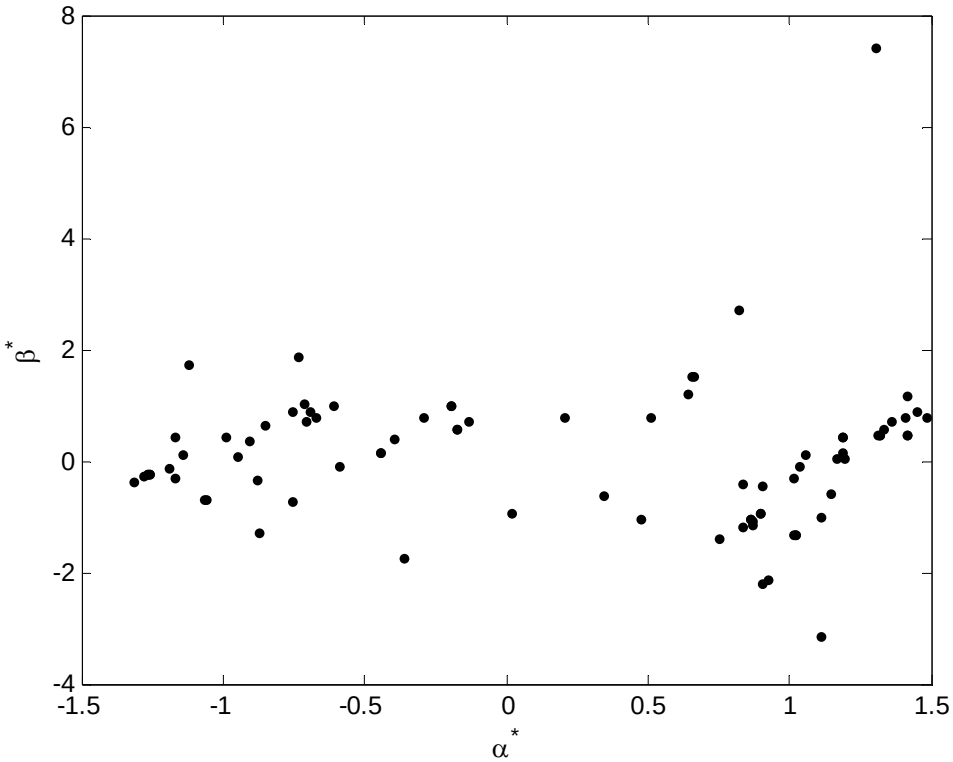
Endeks değerinin yüksek olduğu şekil 3.17 (a) beklenen kümelenmenin gayet açık olduğu görülmektedir.

Aykırı değer ve çoklu bağlantının varlığı durumunda az gözlem ve değişken üzerinde yapılan izdüşüm arama sonucu etkilenmektedir. Diğer veri analizi grafiklerinin daha açıklayıcı sonuçlara götürdüğü söylenebilir.

Bir üçüncü yaklaşım ise 192 gözlem ve 7 değişkenli (silindir sayısı, çap, sıkıştırma oranı, hacim, beygir gücü, tork ve boş ağırlık) orijinal veri kümesi ile yapılan uygulamada Ki-Kare Endeksi aşağıdaki gibi bulunmuştur.

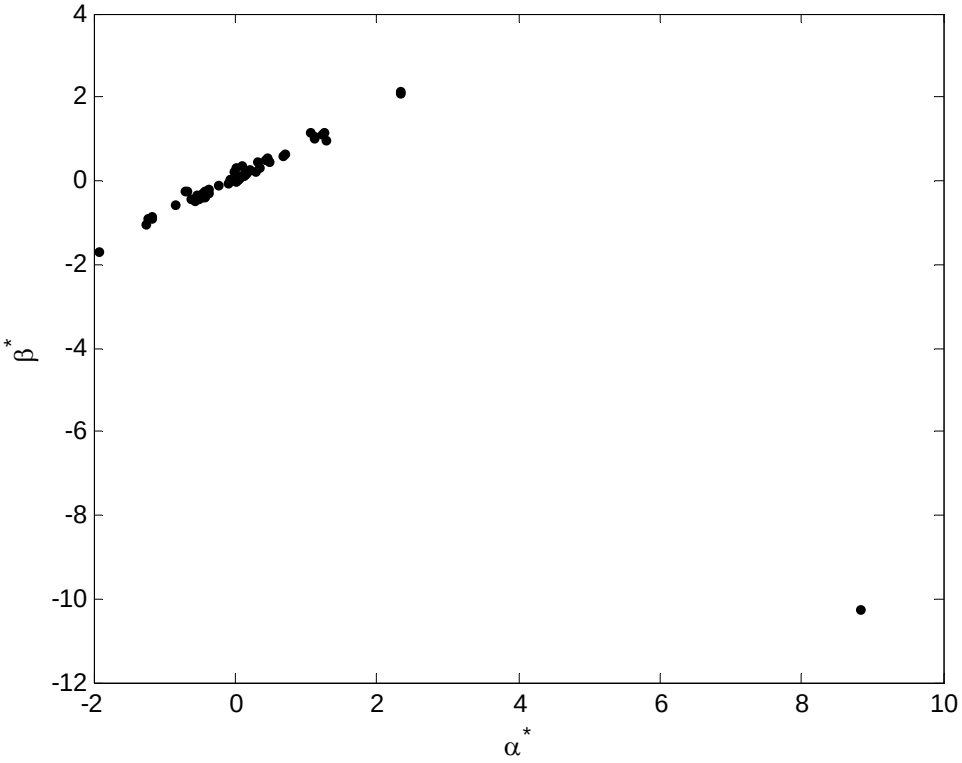


Şekil 3.18. (a) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 8.29 olduğu İzdüşüm Arama Çizimi.

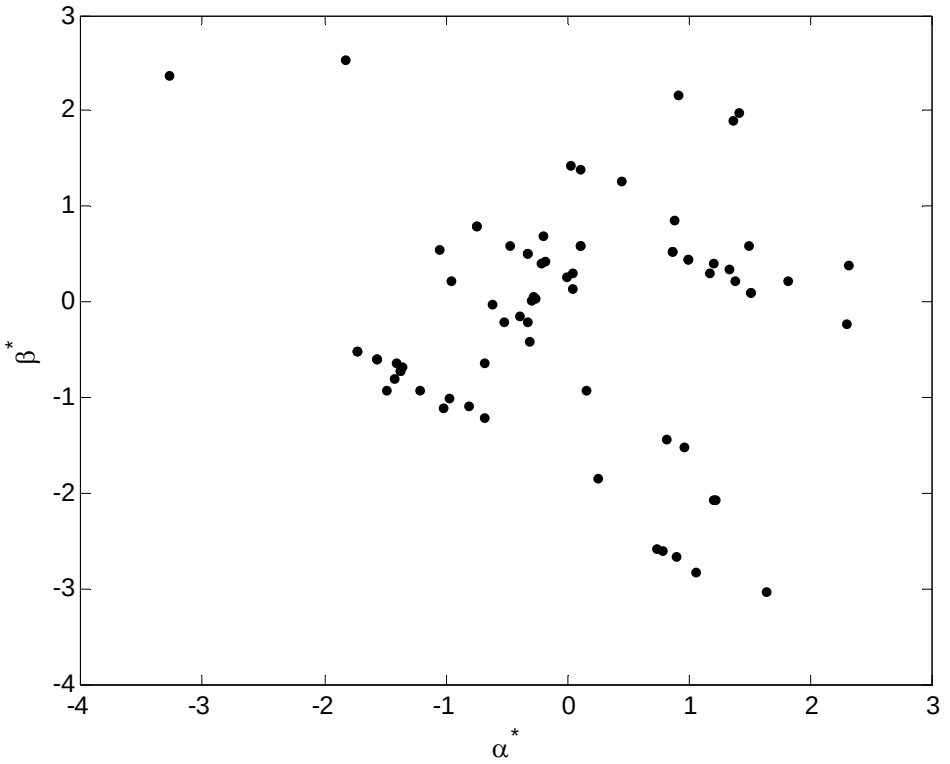


Şekil 3.18. (b) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 1.74 olduğu İzdüşüm Arama Çizimi.

İlk grafikte oldukça aykırı görünen değer ya da değerlerden arındırılarak tekrardan uygulama yapılması sonucunda aşağıdaki gibidir.



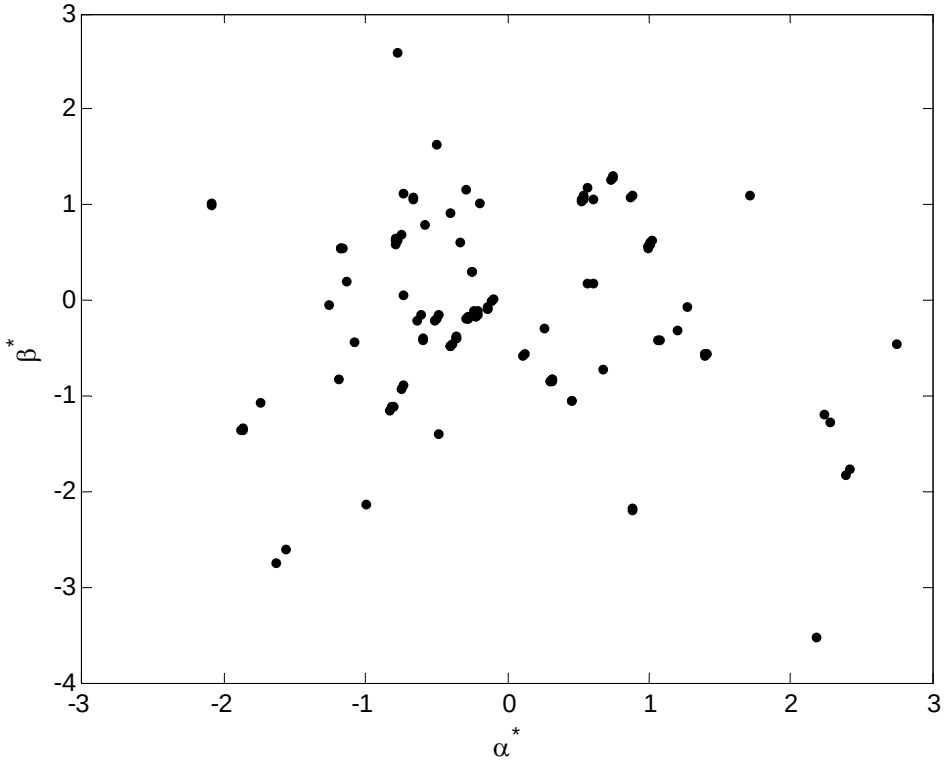
Şekil 3.19. (a) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 11.39 olduğu İzdüşüm Arama Çizimi.



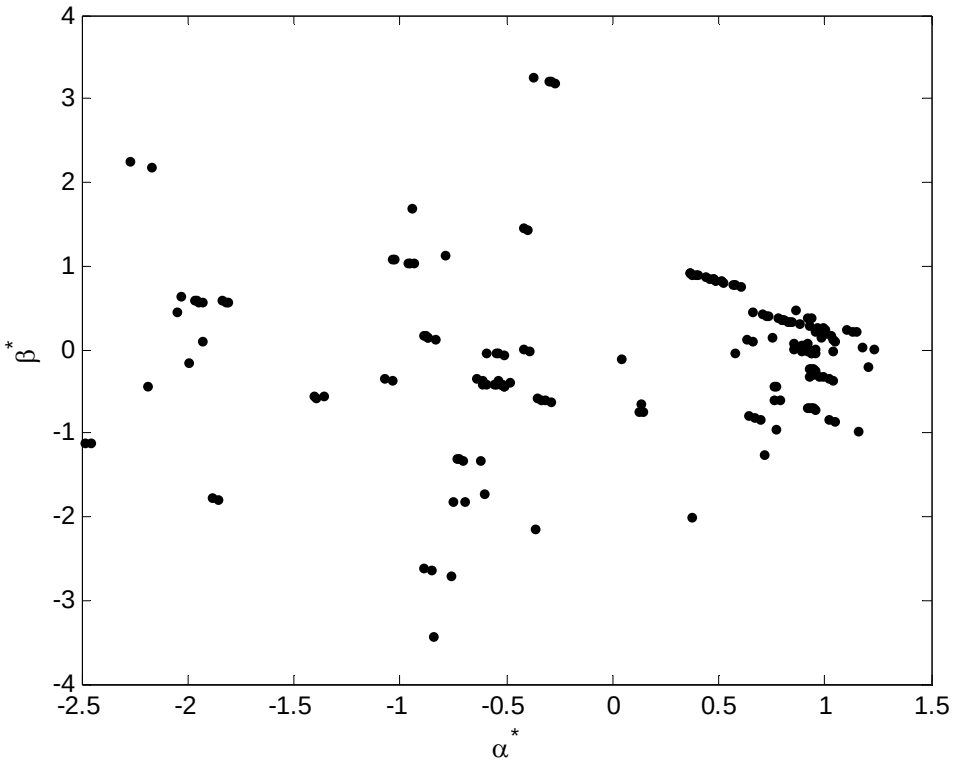
Şekil 3.19. (b) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 1.07 olduğu İzdüşüm Arama Çizimi.

Şekil 3.19’ da ki gibi olup sonucun çok da değişmediği hala bir aykırı değer ya da değerlerin varlığı gözlemlenmektedir. Aykırı ya da uç değerlerin varlığı durumunda izdüşüm arama yönteminin yeterli olmadığı görülmektedir.

Silindir tipi ve değişkenlerden hacim, beygir gücü, tork ve boş ağırlık dağılımları çarpık olduğundan dolayı “log dönüşümü” alınarak uygulanan Ki-Kare Endeks’inin sonucu aşağıdaki gibidir:



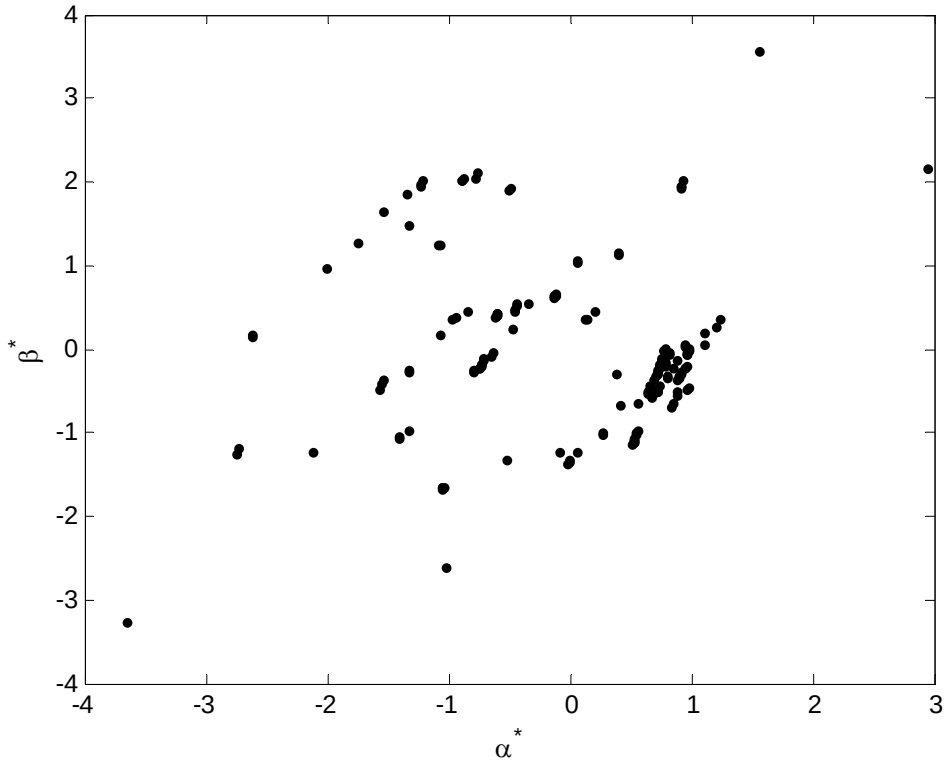
Şekil 3.20. (a) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 1.22 olduğu İzdüşüm Arama Çizimi.



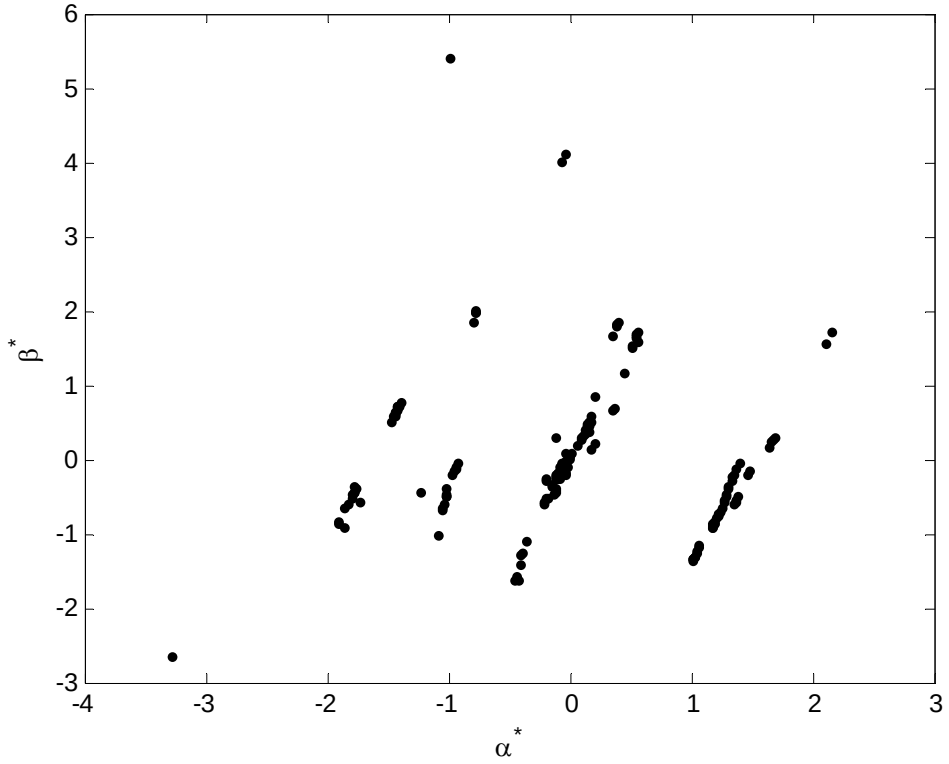
Şekil 3.20. (b) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 2.89 olduğu İzdüşüm Arama Çizimi.

Verilerin dönüştürülmüş hali ile bulunan sonuç dönüştürülmeden önce bulunan sonuçtan daha açıklayıcı sonuç vermektedir. Her iki izdüşüm düzleminde de aykırı değerlerin varlığından bahsedilirken dönüştürülmüş verilerde aykırı değerlerin daha az belli olduğu açıkça görülmektedir.

İki farklı silindir tipinin yer aldığı veri kümesine uygulanan Ki-Kare Endeks yönteminin sonucu aşağıdaki gibidir;



Şekil 3.21. (a) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 1.11 olduğu İzdüşüm Arama Çizimi.



Şekil 3.21. (b) Otomobil veri kümesi için Ki-Kare Endeks Değerinin 3.65 olduğu İzdüşüm Arama Çizimi.

Silindir tipinin yer aldığı veri kümesinde iki farklı silindir tipi olmasına rağmen bu değişkenin dağılım üzerinde önemli bir rol oynamadığı söylenebilir.

DÖRDÜNCÜ BÖLÜM

SONUÇ VE TARTIŞMA

Çalışmanın birinci bölümünde, çok boyutlu veri kümelerinde diğer boyut indirgeme yöntemlerine göre neden izdüşüm arama yönteminin tercih edildiği, böyle bir yönteme neden gereksinim duyulduğu ve izdüşüm arama yönteminin tarihçesi yer almaktadır.

İkinci bölümde ise doğrusal modelleme problemlerinde kullanılan bazı tanımlamalar ve izdüşüm ile ilgili genel bir bilgi, izdüşüm arama kavramı ve yöntemin adımları, yönteme ait çeşitli endeksler ve bunların sunumu, izdüşüm arama yöntemine ait bazı farklı algoritma hesapları ile bunların farklı endeks hesapları ile kombinasyonlarının oluşturulması ve örnek şekiller üzerinden açıklanması yer almıştır.

Üçüncü bölüm, uygulama kısmından oluşup, Fisher' in iris verisi üzerinden ve orjinal otomobil verilerinden elde edilmiş izdüşüm arama yönteminin dönüştürülmüş veri kümesi ile daha iyi sonuç verdiği, gözlem ve değişken sayısı az iken ve aykırı değerin varlığında ise iyi sonuç vermediğinin açıklamasını içermektedir.

Yapılan çalışmalar sonucunda, çok boyutlu veri analizinde boyut indirgeme yöntemlerine göre izdüşüm arama endeksinin daha hızlı hesaplama ve daha doğru istatistiksel sonuçlar verdiği anlaşılmaktadır.

Veriler hakkında hiçbir bilgi olmadan da bizi veri hakkında bir yoruma götürebildiğinden dolayı tercih edilen bir yöntemdir.

Çeşitli endeks ve algoritma kombinasyonlarının olmasına rağmen her türlü veri dağılımında en iyi sonucu Posse' un Ki-Kare Endeksinin verdiği görülmüştür.

Martinez ve Martinez (2005)' in verdiği izdüşüm arma yöntemi Matlab Program, aykırı ve uç değerlerin varlığında ya da bire bir bir değişken incelenmek istendiği zaman çok açıklayıcı olamamasından dolayı daha gelişmiş yardımcı programlara gereksinim duyulmaktadır. Ayrıca merak edilen bir değerin hangisi olduğunu öğrenebilmek imkansız olduğunda bu yöntemin çok kullanışlı olmadığı görülmüştür.

Az boyutlu veri kümelerinde sağlıklı sonuçlar vermediği, az değişken ve gözlem söz konusu olduğunda diğer grafik yöntemlerinin tercih edilmesinin daha uygun olduğu görülmüştür.

Bu yöntemin dönüştürülmüş veriler ve çoklu bağlantının olmadığı veri kümelerinde uygulamaya daha uygun olduğu anlaşılmıştır.

Genel bir sonuç olarak çok boyutlu veri analizinde ilk adım olarak izdüşüm arama yönteminin kullanılabilirliği fakat detaylı bir veri inceleme söz konusu olduğu zaman yetersiz olabileceği ve yardımcı yöntemlere gereksinim duyulduğu anlaşılmaktadır.

KAYNAKLAR

Bronson, R., 1989. Matris İşlemleri, Nobel Yayın Dağıtım, Ankara.

Erar, A., 2007. Matlab ile Sayısal Çözümleme, Ders Notları, MSGSÜ.

Friedman, J.H., 1987. Exploratory Projection Pursuit, J. American Statist. Assoc., 82, 249-266.

Friedman, J.H., Stuetzle, W. And Schroeder, A.,1984. Projection Pursuit Density Estimation, J. Am. Stat. Assoc., 79, 599-608.

Friedman, J. H. ve **Stuetzle, W.,** 1981. Projection Pursuit Regression, J. Am. Stat. Assoc., 76, 817-823.

Friedman, J. H. ve **Tukey, J. W.,** 1974. A Projection Pursuit Algorithm For Exploratory Data Analysis, IEEE Trans. Comput., **C**-23 881-889.

Huber, P.J., 1985. Projection Pursuit (with discussion), Ann. Statist., 13, 435-475.

Jones, M.C. ve Sibson, R., 1987. What is Projection Pursuit?, Royal Stat. Soc., 150, 1-37.

Kıral, G. ve Billor, N.. 2005. Yüksek Boyutlu Veri Kümeleri İçin Robust BACON Temel Bileşenler Analizi, http://www.ekonometridernegi.org/bildiriler/o19s2.pdf

Kıral, G. ve Billor, N.. 2001. Bacon Temel Bileşenler Analizi İle Sapan Değerlerin Belirlenmesi, http://idari.cu.edu.tr/sempozyum/bil27.htm

Koç, Y. S., 2007. Robust Tahmin Edicileri Ve Özellikleri, http://library.cu.edu.tr/tezler/6441.pdf , 25 Eylül 2007.

Martinez, W.L. ve Martinez, A.R., 2005. Exploratory Data Analysis with Matlab, CRC Press.

Posse, C., 1995(a). Projection Pursuit Exploratory Data Analysis, Compt. Statist. and Data Analysis, 20, 669-687.

Posse, C., 1995(b). Tools for Two-Dimensional Exploratory Projection Pursuit, J. Compt. Graph. Statist., 4, 83-100.

Stapleton, J. H., 1995. Linear Statistical Models, A Wiley-Interscience Publication, New York.

ÖZGEÇMİŞ

Adı Soyad : Alev BAKIR

Doğum Yeri : İstanbul

Doğum Tarihi : 25.05.1983

Eğitim Durumu

Lise : Bahçelievler Lisesi (1997-1999)

Lisans : Yıldız Teknik Üniversitesi Fen Fakültesi İstatistik Bölümü (2000-2005)

Yüksek Lisans : Mimar Sinan Üniversitesi Fen Bilimleri Enstitüsü İstatistik A.B.D. (2007-)

İş Tecrübesi : Erenköy Ruh ve Sinir Hastalıkları Hastanesi- Biyo-istatistikçi