YÜKSEK BOYUTLU KANSER SINIFLAMA PROBLEMİNDE BİLGİ KARMAŞIKLIĞI KRİTERİ İLE AYKIRI GÖZLEM TESPİTİ VE BOYUT İNDİRGEME

Esra PAMUKÇU

Doktora Tezi
İstatistik Anabilim Dalı
Danışman: Doç. Dr. Sinan ÇALIK
İkinci Danışman: Prof. Dr. Hamparsum BOZDOĞAN
MART-2015

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK BOYUTLU KANSER SINIFLAMA PROBLEMİNDE BİLGİ KARMAŞIKLIĞI KRİTERİ İLE AYKIRI GÖZLEM TESPİTİ VE BOYUT İNDİRGEME

DOKTORA TEZİ

Esra PAMUKÇU

092133201

Anabilim Dalı: İstatistik

Programı: Uygulamalı İstatistik

Danışman: Doç. Dr. Sinan ÇALIK
İkinci Danışman: Prof. Dr. Hamparsum BOZDOĞAN

Tezin Enstitüye Verildiği Tarih: 17 Şubat 2015

T.C.
FIRAT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK BOYUTLU KANSER SINIFLAMA PROBLEMİNDE BİLGİ KARMAŞIKLIĞI KRİTERİ İLE AYKIRI GÖZLEM TESPİTİ VE BOYUT İNDİRGEME

DOKTORA TEZİ

Esra PAMUKÇU

092133201

Tezin Enstitüye Verildiği Tarih: 17 Şubat 2015
Tezin Savunulduğu Tarih: 18 Mart 2015

Tez Danışmanları:	**Doç. Dr. Sinan ÇALIK (FÜ)**
	Prof. Dr. Hamparsum BOZDOĞAN (UT)
Diğer Jüri Üyeleri:	**Prof. Dr. Ergun KARAAĞAOĞLU (HÜ)**
	Prof. Dr. Olcay ARSLAN (AÜ)
	Prof. Dr. Aydın ERAR (MSÜ)
	Prof. Dr. Mehmet BEKTAŞ (FÜ)
	Doç. Dr. Mehmet GÜRCAN (FÜ)

MART-2015

ÖNSÖZ

Doktora tez çalışmasına başladığım ilk zamanlarda, içimde en iyiyi ve en güzeli doğru bir şekilde yapma arzusu varken, kendimi literatürün içinde kaybolmuş ve ne çalışmak istediğine karar verememiş, tamamen yönsüz ve haritasız olarak bulmuştum. Bu esnada, sorularıma vermiş olduğu cevaplar ile bana ışık tutan, önerdiği çalışma konusu ve problem ile tezimin temel yapısını oluşturan, Tennessee Üniversitesi'nde bulunduğum sürede problemin çözümü için bana eşsiz destekte bulunan, sağlamış olduğu kaynaklar ve imkânlar ile benden yardımını hiçbir zaman esirgemeyen tez danışmanım sayın Prof. Dr. Hamparsum BOZDOGAN'a sonsuz şükranlarımı sunmayı bir borç bilirim.

Aynı şekilde tez çalışmalarının sürdürülebilmesi için beni sürekli olarak destekleyen bölüm başkanım ve aynı zamanda tez danışmanım sayın Doç. Dr. Sinan ÇALIK' a en içten teşekkürlerimi sunarım.

Ayrıca, Tennessee Üniversitesi'nde çalıştığım süre içindeki ekip arkadaşlarım Dr. Elçin KARTAL KOÇ ve Dr. Oğuz AKBİLGİÇ'e yorum ve desteklerinden dolayı teşekkür ederim.

Doktora süresinin tamamı boyunca yurt içi doktora burs programı kapsamında bana burs veren TÜBİTAK'a ve tez çalışmalarımı tamamlayabilmek için Amerika'da kaldığım sürede bana destek veren YÖK'e teşekkürlerimi iletirim.

Meslek hayatımın her anında hem manen hem de madden sürekli yanımda olan, bu mesleğin zorluklarını da güzelliklerini de her zaman beraber yaşadığım eşim Hilmi Emrah PAMUKÇU'ya, kıymetli ailelerim EMİR ve PAMUKÇU ailelerine, ulaştığım tüm başarıları onlar sayesinde elde ettiğimi belirterek şükranlarımı sunarım.

Esra PAMUKÇU
ELAZIĞ – 2015

İÇİNDEKİLER

ÖZET

DNA mikrodizilim tekniklerinde meydana gelen son gelişmeler, muhtemel gen ifadelerinin binlercesinin aynı anda görüntülenmesine imkan sağlamıştır. Gen ifade verilerindeki bu zenginlik nedeniyle araştırmacılar, bu verileri kullanarak kanser sınıflaması yapmanın ihtimalleri üzerinde durmaya başlamışlardır. Bu konuyla alakalı olarak son yıllarda metotların bir çoğunda umut verici sonuçlar elde edilmeye başlanmıştır. Fakat hala çözülmeye ve anlaşılmaya ihtiyaç duyulan bir çok konu vardır. Bunlardan en önemlileri boyutsallık problemi ve aykırı gözlemlerin tespitidir.

Özellikle gen ifade verilerindeki aşırı derecede küçük örneklem problemi ($n << p$) boyut indirgeme ve aykırı gözlem tespiti konularında klasik istatistiksel yöntemlerin uygulanmasını imkansız kılan bir durumdur. Gözlem sayısının, değişken sayısından aşırı derecede küçük olması sonucunda örnek varyans-kovaryans matrisi olan S'in dejenere olması ve tersinin hesaplanamaması, klasik istatistiksel yöntemler açısından karşılaşılan en büyük problemdir.

Bu noktadan hareketle bu çalışmada öncelikli olarak, mikrodizilim verilerinin analizinde literatürde ilk defa, Maksimum Entropi kovaryans matrisinin ve diğer bazı sağlam veya düzgünleştirilmiş kovaryans matrislerinin kullanımı ile S'in dejenere olmasının önüne geçilmiş ve dolayısıyla boyut indirgeme ve aykırı gözlemlerin tespitleri mümkün hale getirilmiştir. İkinci olarak boyut sayısına karar verirken önemli bileşenler, klasik yöntemlerden farklı olarak yine literatürde ilk defa bilgi karmaşıklığı kriteri ICOMP yardımıyla seçilmiş ve üçüncü olarak, verinin boyutu indirgendikten sonra elde edilen alt uzay üzerinde yine literatürde ilk defa ICOMP yardımı ile verideki aykırı gözlemler tespit edilmiştir. Literatürde bu problemleri eş zamanlı olarak değerlendirip bilgi kriterlerinin yardımı ile tutarlı ve doğru bir şekilde hem boyut indirgeyen hem de aykırı gözlem tespiti yapabilen bir çalışma bulunmamaktadır.

Bilgi Karmaşıklığı Kriteri ICOMP ile önerilen bu yaklaşımların hem benzetim verilerine hem de çeşitli mikrodizilim veri setlerine uygulanması sonucunda, boyut indirgemenin ve aykırı gözlemlerinin tespitinin başarılı bir şekilde yapılabildiği görülmüştür. Sonuçların geçerliliği, bazı sınıflama prosedürlerinin doğru sınıflama yüzdesi kullanılarak da ortaya koyulmuştur.

Anahtar Kelimeler: Gen İfade Verileri Analizi, Boyut İndirgeme, Aykırı Gözlem Tespiti, Bilgi Karmaşıklığı Kriteri

SUMMARY

Dimension Reduction and Detection of Outliers in Cancer Classification Using Information Complexity for Undersized Samples

Recent developments in DNA microarray techniques has allowed simultaneously to display thousands of the potential gene expressions. Due to the wealth of gene expression data, researchers have begun to focus their attention on how to optimally classify cancer using the gene expression data. Although many of the methods used have produced promising results, there remains many problems yet to be resolved and to be understood. One of the most important of these problems is the dimension reduction and the detection of outliers.

The classical statistical techniques cannot be used to reduce the dimension and to detect the outliers because of the severity of undersized sample problem $(n << p)$ in gene expression data. When the number of observations is much smaller than the number of features (or variables), the usual sample covariance matrix S degenerates and it can not be inverted. This is one of the biggest encountered obstacle to the classical statistical methods.

In this thesis, to remedy the manifestation of the singular covariance matrices, for the first time, we introduce new robust estimators of the covariance matrix with a well-structured eigen-system. These robust (or smoothed) estimators of the covariance matrix overcome the singularity of the covariance matrix. Therefore, to reduce the dimension and to detect the outliers have been made possible. Secondly, for the first time in the literature, we derive and score the information complexity ICOMP criterion to choose the number of principal components in the data to reduce the dimension. After the dimension reduction, we carry out supervised classification and also a case deletion diagnostics is proposed to determine the outliers in the reduced subspace using ICOMP criterion.

We demonstrate our approach on both the simulation studies and the various benchmark real microarray data sets to reduce the dimension and at the same time to classify the data for cancer and detect the outliers. The results show the flexibility and utility of the new approaches presented.

Key Words: Gene Expression Data Analysis, Dimension Reduction, Detection of Outliers, Information Complexity Criteria.

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

SEMBOLLER LİSTESİ

n : Gözlem sayısı
p : Değişken sayısı
λ : Özdeğerler
μ : Ortalama vektörü
Σ : Anakütle varyans-kovaryans matrisi
S : Örnek varyans-kovaryans matrisi
$\hat{\Sigma}_{MLE}$: Varyans-kovaryans matrisinin maksimum olabilirlik tahmini
ε : Hata, gürültü
$I(X)$: Bilgi miktarı
$H(X)$: Entropi
$L(\theta \mid X)$: Olabilirlik fonksiyonu
$l(\theta \mid X)$: Olabilirlik fonksiyonun logaritması
$\mathcal{F}(\theta)$: Fisher bilgi matrisi
$\hat{\mathcal{F}}(\theta)^{-1}$: Ters Fisher bilgi matrisi
D : Duplikasyon matrisi
$\otimes$: Kronecker iç çarpım
D^{+} : D'nin Moore-Penrose tersi
R : Toplam risk
$R(M)$: Modelleme riski
$R(E)$: Kestirim riski
$f(X)$: Ortak dağılım fonksiyonu
$f_i(X_i)$: Marjinal dağılım fonksiyonu
$N_p(\mu, \Sigma)$: Çok değişkenli normal dağılım
$E(X)$: Beklenen değer
$C_0(\Sigma)$: Varyans-kovaryans matrisi için van Emden komplekslik ölçüsü
$C_1(\Sigma)$: Varyans-kovaryans matrisi için maksimal komplekslik ölçüsü
C_{1F} : Maksimal kompleksliğin Frobenius formu
$tr(\Sigma)$: Matrisin izi
$|\Sigma|$: Kovaryans matrisinin determinantı, genelleştirilmiş varyans
$rank(\Sigma)$: Matrisin rankı
λ_a : Özdeğerlerin aritmetik ortalaması
λ_g : Özdeğerlerin geometrik ortalaması
σ^2 : Varyans bileşeni
Σ^{-1} : Hassas-precison matris
$\hat{\Sigma}_R$: Ridge düzenlemesi
$\hat{D}$: Ridge düzenlemesinde büzülme hedefi
$\hat{\rho}$: Optimal büzülme katsayısı
$\kappa(\Sigma)$: Singülerlik ölçümü için hesaplanan sayı

$\hat{\Sigma}_{EB}$: Emprical Bayes tahmin edicisi

$\hat{\Sigma}_{SRE}$: Stipulated Ridge tahmin edicisi

$\hat{\Sigma}_{SDE}$: Stipulated Diagonal tahmin edicisi

$\hat{\Sigma}_{CSE}$: Convex Sum tahmin edicisi

$\hat{\Sigma}_{STA}$: Thomaz Stabilization ile elde edilen tahmin

ξ_j : Maksimum Entropi kovaryans tahmininde birincil orta noktalar

η_j : Maksimum Entropi kovaryans tahmininde ikincil orta noktalar

$\hat{\Sigma}_{ME\_STA}$: Maksimum Entropi +Thomaz Stabilization hibrit yapısı

$\hat{\Sigma}_{ME\_STA\_CSE}$: Maksimum Entropi +Thomaz Stabilization+Convex Sum hibrit yapısı

$\hat{\Sigma}_{HCE}$: Hibritleştirilmiş kovaryans tahmin edicisi

π_i : i. Sınıf

$p(x \mid \pi_i)$: Şartlı sınıf olasılığı

$d_i(x)$: Kuadratik diksriminant ölçüsü

$d^*_i(x)$: Kuadratik diskriminant kuralı

S_i : i. grup için örnek varyans-kovaryans matrisi

S_p : Örnek grup kovaryanslarının ağırlıklandırılmış ortalaması

KISALTMALAR

DNA : Deoksirübo nükleik asit
PCA : Temel bileşenler analizi
PPCA : Olasılıklı temel bileşenler analizi
SVD : Tekil değer ayrışımı
EM : Expectation-Maximization algoritması
AIC : Akaike bilgi kriteri
ICOMP : Bilgi karmaşıklığı kriteri
GAIC : Genelleştirilmiş Akaike bilgi kriteri
TIC : Takeuchi'nin bilgi kriteri
AIC_T : Takeuchi'nin bilgi kriteri
BIC : Bayes bilgi kriteri
SBC : Schwart'ın bilgi kriteri
HQN : Hannan-Quinn'in bilgi kriteri
CAIC : Tutarlı Akaike bilgi kriteri
CAICF : Fisher bilgi matrisinin kullanıldığı tutarlı Akaike bilgi kriteri
KL : Kullback-Leibler uzaklığı
IFIM : Ters Fisher bilgi matrisi
ICOMP_PEU : Sonsal beklenen faydaya bir yaklaşım olarak bilgi karmaşıklığı kriteri
ICOMP_MISS: Modelin hatalı belirlenme riskini değerlendiren bilgi karmaşıklığı kriteri
CICOMP : Tutarlı bilgi karmaşıklığı kriteri
ME : Maksimum Entropi kovaryans tahmin edicisi
CN : Koşul sayısı
HCE : Hibritleştirilmiş kovaryans tahmin edicisi
LDA : Lineer diskriminant analizi
QDA : Kuadratik diskriminant analizi
HDR : Hibrit boyut indirgeme yöntemi
SRBCT : Small Round Blue Cell Tumor
MD : Mahalanobis uzaklığı
RD : Sağlam uzaklık
MVE : Minimum hacimli elipsoid yaklaşımı
MCD : Minumum kovaryans determinantı yaklaşımı
FAST_MCD : Geliştirilmiş minumum kovaryans determinantı yaklaşımı
BACON : Bloklanmış adaptif etkin hesaplamalı aykırı gözlem belirleyicisi
$ICOMP_i$: i. gölzem hariç verinin geri kalan üyeleri için bilgi karmaşıklığı kriteri değeri
$ICOMP_{tümdata}$: Tüm veri seti için bilgi karmaşıklığı kriteri değeri
$ICOMP_r$: Aykırı gözlem tespiti için bilgi karmaşıklığı kriteri oran değeri
AIC_r : Aykırı gözlem tespiti için Akaike bilgi kriteri oran değeri
$CAIC_r$: Aykırı gözlem tespiti için tutarlı Akaike bilgi kriteri oran değeri
$CICOMP_r$: Aykırı gözlem tespiti için tutarlı bilgi karmaşıklığı kriteri oran değeri
HOD : Hibrit aykırı gözlem tespit etme yöntemi

Babamın anısına…

1. GİRİŞ

Çağımızın vebası olarak görülen ve henüz tam bir tedavisi geliştirilememiş olan kanser hastalığı, tüm dünya toplumlarını her geçen gün daha da tehdit eder düzeye gelmiştir. Türk Kanser Araştırma ve Savaş Kurumu Derneği'nin son verilerine göre, Dünya'da 20 milyondan fazla kanser hastası bulunmakta ve her yıl 10 milyondan fazla yeni hasta tespit edilmektedir. Kanser hastalığının kabul edilmiş dört evresi bulunmaktadır ve kanser ne kadar erken evredeyken fark edilirse, tedavi şansı da o derecede artmaktadır. Bu yüzden, kanser hastalığında erken teşhis çok önemlidir [URL-1].

DNA mikrodizilim tekniklerinde meydana gelen son gelişmeler, muhtemel gen ifadelerinin binlercesinin aynı anda görüntülenmesine imkan sağlamıştır. Gen ifade verilerindeki bu zenginlik nedeniyle, araştırmacılar bu verileri kullanarak kanser sınıflaması yapmanın ihtimalleri üzerinde durmaya başlamışlardır. Bu bağlamda, bir dokudan alınan tümörün iyi huylu ya da kötü huylu olduğunun anlaşılması, kötü huylu ise önce tümörün tipinin, daha sonra bu tümörün alt tipinin belirlenmesi, son olarak da kanserin evresinin belirlenmesi hem zor, hem maliyetli, hem de zaman isteyen süreçlerdir. Gen ifade verilerini kulanarak kanser sınıflaması yapmanın ihtimalleri üzerinde son yıllarda metotların bir çoğunda umut verici sonuçlar da rapor edilmiştir. Fakat hala çözülmeye ve anlaşılmaya ihtiyaç duyulan bir çok konu vardır. Bunlardan en önemlileri boyutsallık problemi ve aykırı gözlemlerin tespitidir.

Geleneksel istatistik veri analizinde, özel bir fenomenin örneğinin gözlemleri düşünülür. Bu gözlemler, kan basıncı, ağırlık, boy, başarı puanı.. vb. gibi bazı değişkenler üzerinde ölçülen değerlerin bir vektörü olur. Geleneksel istatistik metodolojisinde, iyi seçilmiş değişkenlerin birkaç tane, gözlemlerin ise daha fazla olduğu farzedilir. Günümüzde ise gözlemler çok olsa bile değişkenlerin sayısının radikal bir şekilde daha fazla olabildiği gözlenmektedir. Burada, çalışma için ulaşılabilen örnekler, onlar veya yüzlerle ifade edilirken, tek bir gözlem binlerce hatta milyonlarca boyuta sahip olabilmektedir. Klasik yöntemler bu tarz verilerle başa çıkabilecek şekilde tasarlanmış değillerdir [Donoho,2000]. İstatistikçiler bazen bu problem için "Big p, Small n" yani "büyük boyut, küçük gözlem" ifadesini kullanmaktadırlar. Bir diğer tanımlama ise "undersized sample problem", aşırı derecede küçük örneklem problemidir [Cunningham, 2007; Fiebig, 1984].

İstatiksel bakış açısına göre aslında, bir veri setindeki örneklerin sayısının, bu örnekleri açıklamak için kullanılan değişkenlerin sayısından anlamlı bir şekilde fazla olması gerekmektedir. Hatta teorik olarak, eğer veriler hakkında çıkarsama yapılacaksa örneklerin sayısının değişken sayısıyla üstel bir şekilde katlanarak artması beklenir. Aksi durumda boyutsallık problemi *(curse of dimensionality*) meydana gelebilmektedir. *Curse of dimensionality* ilk defa 1961' de Bellmann tarafından literatüre kazandırılan bir terim olup, bazı değişkenlerin bir fonksiyonunu tahmin etmek için, yani verilen bir doğruluk derecesine göre mümkün düşük varyanslı tahmini elde etmek için, gerekli örneklem büyüklüğünün, değişkenlerin sayısı ile üstel olarak artış göstermesi gerektiğini ifade eder. Uygulamada ise gerçek yüksek boyutlu veriler, giriş uzayında sadece bir manifold meydana getirdiği için durum böyle değildir ve bu yüzden verinin örtük boyutu, *p* değişken sayısından daha az olacaktır [Miguel ve Carreira,1997].

Boyutsallık problemi ile karşı karşıya kalındığında, değişken sayısı yani boyut arttıkça uygulamalarda kullanılan geleneksel algoritmaların performanslarında düşüş meydana geldiği de iyi bilinen bir gerçektir. Bu konuyla ilgili olarak, boyut indirgeme teknikleri, veri modelini basitleştirmek için, veri analizinin bir parçası olarak veya bir ön-hazırlık aşaması olarak sıklıkla uygulanır. Bu genel olarak, yüksek boyutlu orijinal veri seti için, düşük boyutlu uygun bir gösterimin tanımlanmasını içermektedir. Düşük boyutlu indirgenmiş veri ile çalışıldığında, hem hesaplama yükü önemli ölçüde azaltılmış olur, hem de sınıflama veya kümeleme gibi kullanılan yöntemler ile daha doğru ve kolaylıkla yorumlanabilir sonuçlar üretilmiş olur.

Boyut indirgeme sayesinde,

i. Değişkenlerin indirgenmiş bir setinin tanımlanması, bilgi keşfi açısından çok kullanışlı olabilir.
ii. Bir çok öğrenme algoritması için eğitim ve/veya sınıflama zamanı, değişkenlerin sayısının indirgenmesi ile azalabilir.
iii. Sınıflama üzerinde, tahmin edici değişkenler gibi aynı etkiye sahip olabilecek gürültülü veya ilgisiz özelliklerin ayıklanmasıyla, bunların doğruluk üzerindeki negatif etkileri arındırılabilir [Cunningham, 2007].

Bu yüzden, birçok durumda gereksiz bilgiyi söken ve verinin daha ekonomik bir gösterimini üreten bir yapı bulmak mümkün olabilmelidir.

Bu bağlamda gen verilerini inceleyecek olursak, gen ifade verileri herhangi bir veri tipinden çok farklı bir yapıya sahiptir.

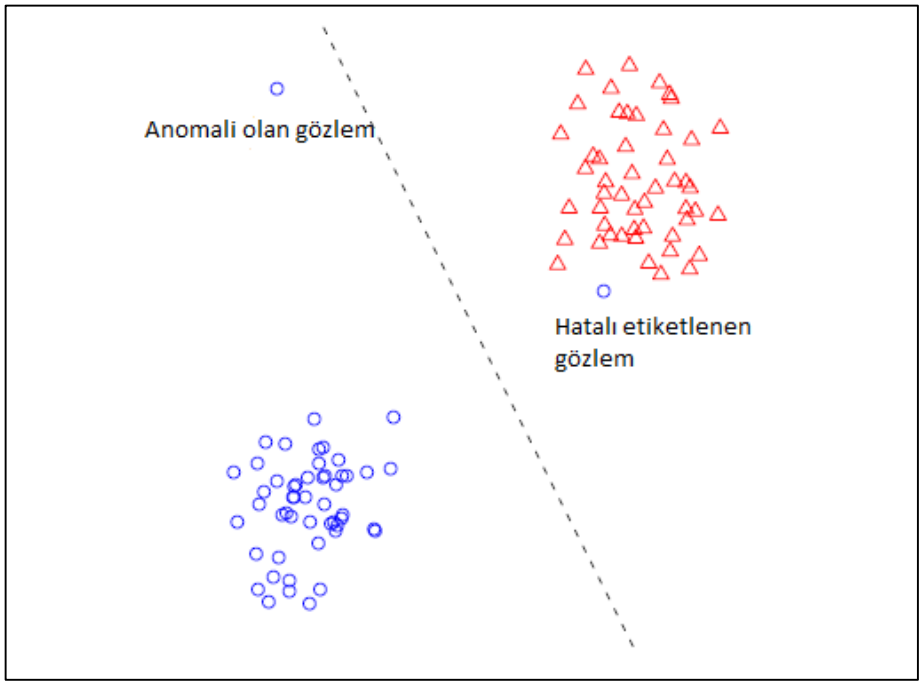
i. Gen verileri çok yüksek boyutsallığa sahiptir
ii. Örneklem sayısı çok sınırlı olabilmektedir
iii. Genlerin büyük bir çoğunluğu ilişkisiz olabilmektedir

Kullanılan veri kümesinde bulunan gen örnekleri için kaydedilmiş tüm niteliklerin, kanser teşhisi ve sınıflandırması için kullanılıp kullanılamayacağı da belirsiz bir durumdur. Bir veri niteliği tek başına performansa çok az bir etki yaratırken, diğer niteliklerle birleştiğinde tüm veri kümesini temsil edecek bir alt grup oluşturabilir [Lu ve Han, 2003]. Bu değişik ve farklı algoritmaların denenmesini gerektirecek bir durumdur. Araştırmacılardan bazıları Fisher Oranı, t-testi tabanlı gen seçim yaklaşımı gibi bazı yöntemler ile gen seçimi yaptıktan sonra sınıflandırma prosedürü kullanmışlardır [Chu ve Wang,2005]. Alternatif olarak uygun bir boyut indirgeme algoritmasının kullanılması da, gen verisindeki yüksek boyutluluk ve ilişkisiz genlerin etkisinin azaltılması amacına hizmet edebilecektir. Ancak, gen ifade verilerindeki aşırı derecede küçük örneklem problemi ($n << p$) Temel Bileşenler Analizi (Principal Component Analysis-PCA) gibi bazı geleneksel boyut indirgeme yöntemlerinde sıkıntılar ortaya çıkarmaktadır. Gözlem sayısının, değişken sayısından aşırı derecede küçük olması sonucunda örnek varyans-kovaryans matrisi olan S dejenere olur ve PCA için gerekli olan özdeğerlerin büyük bir çoğunluğu negatif veya sıfır çıkar. Bu da beraberinde özvektörlerin keyfi seçimini ve temel bileşenlerin hesabını mümkünsüz olmasını sağlayan bir durumdur. Bu durumu dikkate almadan yapılacak olan klasik bir boyut indirgeme yönteminden elde edilecek sonuçlara güvenilemez.

PCA yönteminin kullanılmasındaki bir diğer önemli sorun, veriyi en iyi temsil edebilecek önemli temel bileşen sayısına karar vermektir. Bu konuda, 1'den büyük özdeğerlere karşılık gelen özvektörlerin ele alınması veya sadece ilk iki, ilk üç tane temel bileşenin keyfi olarak seçilmesi gibi bazı geleneksel yaklaşımlar sözkonusu olmakla beraber, bu konu da literatürde tatmin edici bir şekilde çözülebilmiş bir problem değildir.

Bir diğer önemli sorun aykırı gözlemlerin tespitidir. Mikrodizilim verilerinde iki tür aykırı gözlem vardır. İlki, veride farklı bir sınıfa ait olan hatalı etiketlenmiş gözlemlerdir. Bu gözlemler normal doku olarak etiketlenen bir tümör dokusu gibi yanlış bir sınıfa atanan gözlemlerdir. Bu aykırı gözlemler kullanılan sınıflama prosedürü ile tespit edilebilirler.

İkincisi ise veride herhangi bir sınıfa ait olmayan anomali gözlemlerdir. Bu aykırı gözlemlerin kaynağı daha çok belirsizdir fakat keşfedilmemiş bir biyolojik sınıftan, zayıf sınıf tahmininden, deneysel hatalardan veya aşırı biyolojik çeşitlilikten kaynaklanabilir. Anomali gösteren bir gözlemin sınıfı belli olmadığı zaman, etiketlendiği sınıfın doğruluğuna tutarlı bir şekilde itiraz da edilemeyecektir. Örnek olarak, bir gözlem gerçekten tümör dokusu olabilir fakat onun ifade düzeyi diğer tümör dokularının ifade düzeylerinden oldukça farklı olabilir. Aykırı gözlemlerin iki çeşidi Şekil 1.1'de gösterildiği gibidir.



Şekil 1. 1: Aykırı gözlemlerin iki çeşidi

Mikrodizilim veri setlerinde aykırı gözlemlerin belirlenmesi ve ayıklanması verinin geri kalanıyla karşılaştırıldığında tutarsız davrandıkları için önemlidir. Aykırı gözlem varlığında bir model uygulamak, parametre tahminlerini çarpık yapabileceği gibi yanlış çıkarsamalara da sebep verebilirler. Literatürde mikrodizilim gen ifade verilerinin analizleri ve yorumlanması ile ilgili olarak sınıflama ve gen seçilimi hakkında çok fazla çalışma bulunmaktadır. Ancak aynı şeyi gen ifade verilerinde aykırı gözlem tespiti için söylememiz mümkün değildir [Shieh ve Hung, 2009].

Tüm bu bahsi geçen problemlerle ilgili olarak, bu tezin amacını üç başlık altında toplamak mümkündür. İlk olarak aşırı derecede küçük örneklem problemine sahip gen veri setleri için, özdeğerlerde meydana gelen sıkıntıyı, literatürde ilk defa Maksimum Entropi Kovaryans Matrisi ve onun diğer düzgünleştirilmiş kovaryans yapıları ile hibritleştirilmiş farklı formlarını kullanarak çözmek ve olasılıklı temel bileşenler analizi (Probabilistic Principal Component Analysis-PPCA) ile gen ifade verilerinin boyutunu tutarlı bir şekilde indirgemektir. İkinci olarak boyut sayısına karar verirken önemli bileşenleri, klasik yöntemlerden farklı olarak yine literatürde ilk defa bilgi kriterlerinin yardımıyla seçmek ve üçüncü olarak, bilgi kriterleri ile verinin boyutu indirgendikten sonra elde edilen alt uzay üzerinde yine bilgi kriterlerinin yardımı ile verideki aykırı gözlemleri tespit etmektir. Literatürde bu problemleri eş zamanlı olarak değerlendirip bilgi kriterlerinin yardımı ile tutarlı ve doğru bir şekilde hem boyut indirgeyen hem de aykırı gözlem tespiti yapabilen bir çalışma bulunmamaktadır. Sonuçların geçerliliğinin, bazı sınıflama prosedürlerinin doğru sınıflama yüzdesi kullanılarak ortaya koyulması amaçlanmaktadır.

Bu çalışma, yukarıda anlatılan strateji çerçevesinde, aşağıdaki bölümleri ihtiva etmektedir. Materyal ve metot bölümünün birinci kısmında, PCA ve PPCA'nın teorik yapısı, özdeğer ayrışımı ve tekil değer ayrışımı kavramları ile örneklerle birlikte sunulduktan sonra, ikinci kısımda, Entropi kavramından, Kullback-Leibler ölçüsüne kadar uzanan bir tarihsel gelişim içinde, Akaike'nin bilgi kriterinden (Akaike Information Criteria-AIC) Bozdogan'ın bilgi karmaşıklığı kriterine (Information Complexity-ICOMP) kadar uzanan bir yelpazede, model seçim kriterleri tanıtılmıştır. Aynı bölümün sonunda tezin orijinal kısımlarından olan, bilgi kriterlerinin PPCA analizi yapıldığı zaman boyut seçerken kullanılabilmesi için, gerekli teorik yapı formülasyonları ile birlikte verilmiştir.

Üçüncü kısımda, tezin ikinci orijinal kısmının ilham kaynağı olan, Maksimum Entropi Kovaryans Tahmin Edicisi yapısı ile diğer bazı sağlam kovaryans tahmin edicileri, ayrıca bunların bazı hibritlenmiş formları olan hibritleştirilmiş kovaryans tahmin edicisi tanıtılmıştır. Dördüncü kısımda, tez kapsamında önerilen boyut indirgeme yöntemi Hibrit Boyut İndirgeme (Hybrid Dimension Reduction-HDR) tanıtılmıştır. Beşinci kısımda, aykırı gözlem tespiti konusu anlatıldıktan sonra tezin üçüncü orijinal kısmı olan, bilgi kriterlerinin aykırı gözlem tespitinde nasıl kullanılabileceği tanıtılmıştır. Altıncı kısımda bu yöntem Hibrit Aykırı Gözlem Tespiti (Hybrid Outlier Detection-HOD) olarak adlandırılmış ve hesaplama adımları tanıtılmıştır. Yedinci kısımda, uygulamada son aşama olarak uygulanan

Lineer Diskriminant Analizi ve Kuadratik Diskriminant Analizi gibi bazı sınıflandırma yöntemlerinden bahsedilmiştir.

Tezin üçüncü bölümü, HDR ve HOD yöntemlerinin başarısını gösterebilmek amacıyla Uygulama ve Bulgular bölümü olarak ayrılmıştır. Bu bölümün birinci kısmında, PPCA analizinde ICOMP ile boyut sayısına karar verme yönteminin başarısı yapılan simülasyon çalışması ile ortaya koyulduktan sonra, ikinci kısımda HDR ile mikrodizilim veri analizi yapılmıştır. Üçüncü kısımda, ICOMP ile aykırı gözlem tespit edebilme yönteminin başarısı da yine bir simülasyon çalışması ile ortaya koyulduktan sonra, dördüncü kısımda HDR yöntemi ile boyutları indirgenen mikrodizilim veri setlerinin aykırı gözlemlerinin tespiti yapılmıştır. Sonuç ve tartışma bölümünde, hem HDR hem de HOD ile elde edilen sonuçlar literatürle karşılaştırmalı olarak tartışılmıştır. Son olarak öneriler bölümünde ise konuyla alakalı olarak yapılması muhtemel olan bazı çalışmalar için öneriler de bulunulmuştur.

2. MATERYAL VE METOT

2.1 Temel Bileşenler Analizinin Klasik ve Olasılıksal Yaklaşımı

2.1.1 Temel Bileşenler Analizi (Principal Component Analysis-PCA)

Bilgisayar olanaklarının çok geliştiği günümüzde işlem yükü bir sorun olarak görülmese de, çok sayıda değişkene ilişkin analiz sonuçlarının yorumlanması ve özetlenmesi gerçekten zor olabilmektedir. PCA ile çok boyutlu değişken uzayını en az bilgi kaybıyla daha az boyutlu değişken uzayına indirgemek, hem diğer çok değişkenli analiz yöntemlerine veri hazırlama bakımından hem de başlı başına kendisinin bir analiz tekniği olması açısından araştırmacılar tarafından çok başvurulan bir yöntemdir.

Tarihsel gelişimi içinde PCA'nin kökeni, Beltrami (1873) ve Jordan (1874)'te bağımsız bir şekilde elde ettikleri "Tekil Değer Ayrışımı" (Singüler Value Decomposition-SVD)'na kadar uzanmaktadır [Stewart,1993]. Bununla birlikte günümüzde temel bileşen analizi olarak bilinen tekniğin ilk adımları Pearson (1901) tarafından atılmıştır. Onun modern örneklemesini yapan ve temel bileşen terimini literatüre kazandıran ise Hotelling (1933) olmuştur [Abdi ve Williams, 2010].

Bilgisayarların oldukça yaygınlaşmasından 50 yılı aşkın bir süre önce, Pearson'ın hesaplamalarla ilgili yorumları oldukça dikkat çekicidir. Pearson, metotlarının sayısal problemlere kolayca uygulanabileceğini ifade etmekte ve hesaplamalarının dört ya da daha fazla değişken için kullanışsız olduğunu söylemesine rağmen yine de metotların uygulanabilir olduğunu ileri sürmektedir.

Pearson ve Hotelling'in çalışmaları arasında geçen 32 yıllık süre zarfında çok az çalışmanın yayınlanmış olduğu görülmektedir. Rao'nun 1964'te belirttiğine göre Frisch (1929), Pearson'ın yaklaşımına benzer bir yaklaşımı benimsediğini belirtmiştir. Ayrıca Hotelling'in (1933) çalışmasındaki bir dipnot, Thurstone (1931)'un Hotelling'le benzer alanlarda çalışıyor olduğunu göstermektedir. Ancak Bryant ve Atchley (1975)' de belirttiğine göre bu çalışma temel bileşenler analizinden çok faktör analiziyle ilişkilidir. Hotelling'in yaklaşımı da faktör analizi fikirleriyle başlamakla birlikte, Hotelling'in tanımladığı temel bileşenler analizi gerçekte faktör analizinden oldukça farklıdır. Hotelling, orijinal "*p*" değişkenin değerlerini belirleyen daha küçük temel bağımsız değişkenler setinin

olabileceği noktasından hareketle işe başlamıştır. Hotelling, böyle değişkenlerin psikoloji literatüründe "faktör" olarak adlandırıldığına dikkat çekmekte ve faktör kelimesinin matematikteki diğer kullanımlarından çıkabilecek karışıklıktan kaçınmak için başka bir alternatif terim olan, "bileşen" kavramını ileri sürmektedir. Hotelling, bileşenlerini toplam varyansa maksimum katkı yapacak düzeye getirerek seçer ve bu şekilde üretilen bileşenlere "temel bileşenler" adını verir [Joliffee,2002].

Hotelling'in çalışmalarının yayınlanmasını takip eden 35 yıl boyunca temel bileşenler analizinin farklı uygulamaları ile ilgili çok az sayıda çalışmanın yapıldığı görülmektedir. Girshick (1936) ve (1939) yıllarında yaptığı çalışmalarda, bazı alternatif temel bileşen elde etme yöntemlerini geliştirmiş ve örnek temel bileşenlerinin, anakütle temel bileşenlerinin en yüksek olası tahminleri olduğu fikrini öne sürerek, temel bileşenlerin varyans ve katsayılarının asimptotik örneklem dağılımlarını araştırmıştır.

Temel bileşenler analizi önemli bir hesaplama gücü gerektirdiği için, kullanımının yaygınlaşması elektronik bilgisayarların ortaya çıkışına rastlamıştır. 1960'ların başlarından itibaren Rao (1964), Gover (1966) ve Jeffers (1967) yayınları, ortaya çıkan ve konuyla ilgili önemli kaynaklar oluşturan çalışmalardır. Cooley ve Lohnes (1971)' de temel bileşenlerin pratikte yorumlanmasını ve tekniğin bilgisayara uyarlamasını göstermişlerdir.

Tekniğin görünüşteki basitliğine rağmen, temel bileşenler analizi alanında birçok araştırma halen yapılmakta ve yaygın bir şekilde kullanılmaktadır. Bu da "principal component analysis" ifadesini başlıklarında, düşüncelerinde ya da anahtar sözcüklerinde içeren "Web of Science" da sadece 2010-2015 yılları arasında bile 23.000'i aşkın makalenin bulunmasından anlaşılmaktadır.

Temel bileşenler analizinin kökeninin tekil değer ayrışımına dayandığı daha önce belirtilmişti. Bu nedenle, bölümün ilerleyen kısımlarında matematiksel altyapı oluşturabilmek için, öncelikle özdeğer, özvektör, öz ayrışım ve tekil değer ayrışımı kavramları verilecek, daha sonra temel bileşenlerin tekil değer ayrışımı yardımıyla nasıl hesaplandığı gösterilecektir.

2.1.2 Özdeğerler ve Özvektörler

Karesel bir matrisle ilişkili olan sayılar ve vektörlerdir. Karesel bir matrisin özdeğer ve özvektör hesabının nasıl yapıldığını görebilmek için Alpar (2011) kullanılabilir. Özdeğerler ve özvektörlerin ikisi birlikte, bir matrisin yapısını analiz ederken kullanılan özayrışımı (eigen-decomposition) oluşturma imkanı sunar. Tüm karesel matrisler için, bir özayrışım var olmasa bile, korelasyon, kovaryans ve iç çarpım matrisleri gibi matrisler için özayrışım özellikle kullanışlı bir ifadeye sahiptir. Bu şekildeki matrislerin özayrışımı, bu matrisleri ihtiva eden fonksiyonların bir maksimumunu veya minimumunu bulmada kullanıldığı için önemlidir. Özel olarak PCA, bir korelasyon veya kovaryans matrisinin özayrışımından elde edilebilir.

Bir *nxn* boyutlu A matrisinin $\{u_1, u_2, ..., u_n\}$ özvektörler seti bir U matrisine yerleştirilmiş olsun. Yani U 'nun herbir kolonu, A 'nın özvektörüdür. A 'nın $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ özdeğerleri, Λ diagonal matrisine yerleştirilirse, bu matrisin köşegen elemanları özdeğerler ve diğer elemanları sıfir olur. Bu durumda $Au = \lambda u$ denklemi

$$AU = \Lambda U$$
$$A = U \Lambda U^{-1} \tag{2.1}$$

şeklinde yazılabilir. Yani bir matrisin özdeğer ve özvektörleri beraber, bu matrisin özayrışımını oluştururlar. Tüm matrislerin bir özayrışımının olamayabileceğini belirtmek önemlidir. Ayrıca bazı matrislerin özdeğer ve özvektörleri sanal da olabilir [Abdi,2007].

Örnek: Matlab programında bir `A` matrisinin öz ayrışımının elde edilmesi için `A=2+3*rand(3,3)` komutu ile `3x3` boyutunda rasgele bir `A` matrisi üretilmiştir. `[V E]=eig(A)` komutu, `A` matrisinin özdeğerlerini `E` diagonal matrisi içine, özvektörlerini ise `V` matrisi içine yerleştirir. Çıktılar aşağıdaki gibidir.

```
A =
   4.2732   3.9664   2.0955
   4.2294   2.5136   2.8308
   3.1767   4.1181   2.1385

V =
  -0.6107  -0.5981   0.3015
  -0.5659   0.1510  -0.7471
  -0.5539   0.7870   0.5924
```

```
E =
  9.8488       0         0
     0      0.5145       0
     0          0   -1.4380
```

Öz ayrışımın yukarıdaki tanımına göre `V* E *inv(V)` komutu ile `A` matrisinin tekrar elde edilmesi gerekmektedir

```
V* E *inv(V) =
  4.2732   3.9664   2.0955
  4.2294   2.5136   2.8308
  3.1767   4.1181   2.1385
```

2.1.3 Pozitif yarı tanımlı matris (Positive semi-definite matrix)

İstatistikte çok sık kullanılan bir matris çeşidi pozitif yarı tanımlı matris olarak adlandırılır. Bu matrislerin özayrışımı daima vardır. Bir matris kendi transpozu ile çarpılırsa, elde edilen yeni matrise pozitif yarı tanımlı matris denir. Açıkça bellidir ki, bu matris karesel ve simetriktir. Dolayısıyla X , gerçel sayılar içeren bir matris olmak üzere

$$A = XX^T \tag{2.2}$$

pozitif yarı tanımlı matrisdir. Özel olarak, korelasyon, kovaryans ve iç çarpım matrisleri bu gruptan olan matrislerdir.

Bir pozitif yarı tanımlı matrisin en önemli özelliği, özdeğerlerinin daima pozitif veya sıfır olmasıdır. Ayrıca özdeğerler farklı olduğu zaman, bu özdeğerlere karşılık gelen özvektörler ortogonal çift oluştururlar. Ayrıca özdeğerler gerçel değerlerden oluşur. Bu önemli özelliklerin ispatına Strang (2003)'den ulaşılabilir. Farklı özdeğerlere karşılık gelen özvektörler ortogonal olduğu için, tüm özvektörleri ortogonal bir matrise yüklemek mümkündür. Pozitif yarı tanımlı matrisler için (2.1) nolu denklem

$$A = U\Lambda U^T \quad U^T U = I \tag{2.3}$$

şeklinde yazılabilir. Burada U normalleştirilmiş özvektörlerin yüklendiği matrisdir. Eğer normalleştirme olmazsa, U ortogonal değil diagonal olur [Abdi, 2007].

Örnek: Matlab programında rassal olarak üretilen `X=rand(3,2)` komutu ile 3x2 boyutundaki bir matrisin, `A=X*X'` komutu ile transpoz çarpma işlemi yapılmıştır. Elde edilen yeni `A` matrisi için öz ayrışım yapılmış ve özdeğerlerin pozitif veya sıfır çıkması gerektiği incelenmiştir. Çıktılar aşağıdaki gibidir.

```
X =
   0.5377    0.8622
   1.8339    0.3188
  -2.2588   -1.3077

A =
   1.0324    1.2609   -2.3420
   1.2609    3.4647   -4.5593
  -2.3420   -4.5593    6.8124

V =
   0.6658    0.6981   -0.2634
   0.4938   -0.6769   -0.5459
   0.5594   -0.2334    0.7954

E =
   0.000     0.000     0.000
   0.000     0.5927    0.000
   0.000     0.000    10.7169
```

Pozitif tanımlı matris olmanın yukarıdaki tanımı gereği, özdeğerler pozitif ya da sıfır olarak elde edilmiştir. Bir önceki örnekde olduğu gibi, `V* E *inv(V)` komutu ile tekrar `A` matrisi elde edilebilmektedir. Ayrıca tanım gereği `V*V'` komutu ile birim matris elde edildiği de gösterilmiştir.

```
V* E *inv(V) =
   1.0324    1.2609   -2.3420
   1.2609    3.4647   -4.5593
  -2.3420   -4.5593    6.8124

V*V' =
    1.000    0.000     0.000
    0.000    1.000     0.000
    0.000    0.000     1.000
```

2.1.4 Tekil değer ayrışımı (Singular Value Decomposition-SVD)

Öz ayrışım, sadece karesel matrisler için tanımlanır. SVD ise, dikdörtgen biçimindeki matrisleri analiz etmek için kullanılabilen bir genelleştirilmiş özayrışımdır. Bir matrisi iki tane basit matrise ayıran özayrışım ile benzerlik kuran SVD'nin ana düşüncesi, dikdörtgen biçimindeki matrisi üç tane basit matrise ayırmaktır: iki ortogonal ve bir diagonal. Bir pozitif yarı tanımlı matrise uygulandığı zaman, SVD, özayrışıma denktir. Formül olarak eğer A bir dikdörtgensel matris ise, onun SVD ayrışımı

$$A = USV^T \tag{2.4}$$

ile gösterilir. Burada;

U: AA^T matrisinin özvektörleridir. Yani $U^TU = I$'dır. U'nun kolonları, A'nın sol tekil vektörleri olarak adlandırılır.

V: A^TA matrisinin özvektörleridir. Yani $V^TV = I$'dır. V'nin kolonları A'nın sağ tekil vektörleri olarak adlandırılır.

S: Tekil değerlerin diagonal matrisidir. AA^T veya A^TA matrisinin özdeğerleri olan Λ ile S arasında $S = \Lambda^{1/2}$ bağıntısı vardır [Abdi, 2007]. A matrisinin rankı, sıfır olmayan tekil değer sayısına eşittir.

Yani eğer $rank(A) = r \leq n$ ise, $\sigma_1, \dots, \sigma_n$ tekil değerleri için

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r \geq 0, \sigma_{r+1} = 0, \dots, \sigma_n = 0 \tag{2.5}$$

olur. Eğer (2.4) nolu denklemin her iki tarafı sağdan A^T ile matris çarpımına tabi tutulursa,

$$\begin{aligned} A^TA &= (USV^T)^T(USV^T) \\ &= VS^TU^TUSV^T \\ &= VS^TSV^T \\ &= VS^2V^T \end{aligned} \tag{2.6}$$

elde edilir. Bunun anlamı, V'nin v_i kolonlarının (sağ tekil vektörler) A^TA matrisinin özvektörleri olmasıdır. S'nin elemanlarının karesi ise, A^TA matrisinin özdeğerleridir. Benzer şekilde (2.4) nolu denklemin her iki tarafı soldan A^T ile matris çarpımına tabi tutulursa

$$
\begin{aligned}
AA^T &= (USV^T)(USV^T)^T \\
&= USV^TVS^TU^T \\
&= US^TSU^T \\
&= US^2U^T
\end{aligned} \tag{2.7}
$$

elde edilir. Bunun anlamı, U'nun u_i kolonlarının (sol tekil vektörler) AA^T matrisinin özvektörleri olmasıdır. S'nin elemanlarının karesi ise, AA^T matrisinin özdeğerleridir [Kumar vd., 2008].

SVD'nin önemli sonuçlarından birisi, bir matrisin aynı veya daha düşük ranklı en küçük kareler tahminini üretmesidir. Yani

$$
A^{(l)} = \sum_{k=1}^{l} u_k s_k v_k^T \tag{2.8}
$$

A'ya en yakın $l-ranklı$ matrisdir. Burada u_k, s_k, v_k^T değerleri sırasıyla, A'nın k. sol tekil vektörü, k. tekil değeri ve k. sağ tekil vektörüdür. Bu A matrisinin $l-ranklı$ bir matrisler topluluğu olarak kurulduğunu gösterir. Bu matrislerden ilki, bir ranklı bir matrisle A'nın en iyi kurulumunu, ilk ikisinin toplamı; iki ranklı bir matrisle A'nın en iyi kurulumunu ve böyle devam edilirse genel olarak, ilk l matrisin toplamı, $l-ranklı$ bir matrisle A'nın en iyi kurulumunu verir. O halde SVD ile

$$
\sum_{i,j} | a_{ij} - a_{ij}^{(l)} |^2 \tag{2.9}
$$

minimize edilir [Abdi, 2007; Wall vd., 2003].

Örnek: Matlab programında**,** dikdörtgen biçimindeki matrislerinin SVD ayrışımını yapmak için tekrar `A=rand(3,2)` komutu ile `3x2` boyutunda rasgele matris üretilmiştir. `[U,S,V]=svd(A)` komutu ile X matrisinin, bir önceki örnekte olduğu gibi transpozu ile çarpma işlemi yapılmadan ayrışımı oluşturulacaktır. Burada `S,` `A` ile aynı boyutlarda tekil değerleri içeren matrisdir. Çıktılar aşağıdaki gibidir.

```
A =
   1.4090   -1.2075
   1.4172    0.7172
   0.6715    1.6302
```

```
[U,S,V]=svd(A)

U =
   0.0286    0.9140    0.4048
  -0.6490    0.3250   -0.6879
  -0.7603   -0.2430    0.6025

S =
   2.2266         0
        0    2.0291
        0         0

V =
  -0.6242    0.7812
  -0.7812   -0.6242
```

Yukarıdaki tanımlara göre `A*A'` matrisinin özvektörlerinin `U` matrisini, `A'*A` matrisinin özvektörlerinin `V` matrisini oluşturması gerekmektedir. Bunu göstermek amacıyla `[V1 E1]=eig(A*A')` ve `[V2 E2]=eig(A'*A)` komutları ile sırasıyla `A*A'` ve `A'*A` Matrislerinin özvektörleri `V1` ve `V2` matrislerine atanmıştır.

```
V1 =
   0.4048    0.9140   -0.0286
  -0.6879    0.3250    0.6490
   0.6025   -0.2430    0.7603

V2 =
  -0.7812    0.6242
   0.6242    0.7812
```

2.1.5 PCA'nın Özayrışım ve SVD ile Hesabı:

X, nxp boyutlu merkezileştirilmiş veri seti olsun. Burada n örneklem sayısı, p değişken ya da özellik sayısıdır. C_x ile X veri setinin varyans-kovaryans matrisi gösterilmektedir. PCA'nin amacı aşağıdaki gibi özetlenebilir.

"$C_y = \frac{1}{n} YY^T$ bir diagonal matris olacak şekilde $Y = PX$ dönüşümünde öyle bir ortonormal bir P matrisi bulunmalıdır ki, P'nin satırları X'in temel bileşenleri olsun."

Bu amaçla C_y bilinmeyen değişkenlere göre tekrar yazılırsa,

$$\begin{aligned} C_y &= \frac{1}{n} YY^T \\ &= \frac{1}{n} (PX)(PX)^T \\ &= \frac{1}{n} PXX^T P^T \\ &= P(\frac{1}{n} XX^T)P^T \\ &= PC_x P^T \end{aligned} \tag{2.10}$$

elde edilir. C_x , özayrışım tanımı gereğince $C_x = V\Lambda V^T$ olacak şekilde yazılabilir ki, burada C_x'in sıralanmış özvektörler seti V matrisine, C_x'in özdeğerleri Λ diagonal matrisine yerleştirilmiştir. C_x'in tanımı yerine yazılırsa,

$$C_y = P(V\Lambda V^T)P^T \tag{2.11}$$

olur. Amaç $C_y = \Lambda$ olacak, yani C_y'yi diagonelleştirecek bir P seçmektir. $P \equiv V^T$ alınırsa,

$$\begin{aligned} C_y &= P(V\Lambda V^T)P^T \\ &= (PP^T)\Lambda(PP^T) \\ &= (PP^{-1})\Lambda(PP^{-1}) \\ C_y &= \Lambda \end{aligned} \tag{2.12}$$

elde edilir. Yani $P \equiv V^T$ seçilmesi C_y'yi diagonelleştirmekte bu da yeni elde edilen değişkenlerin kovaryanslarının sıfır olduğu anlamına gelmektedir ki bu da PCA'nin amacıdır. O halde X'in temel bileşenleri C_x'in özvektörleridir [Shlens, 2009].

$Y = V^T X$ için Y'nin varyansı

$$\begin{aligned} YY^T &= (V^T X)(V^T X)^T \\ &= V^T XX^T V \\ &= V^T C_x V \\ &= V^T (V\Lambda V^T)V \end{aligned}$$

$$= (V^T V)\Lambda(V^T V)$$
$$= \Lambda \qquad (2.13)$$

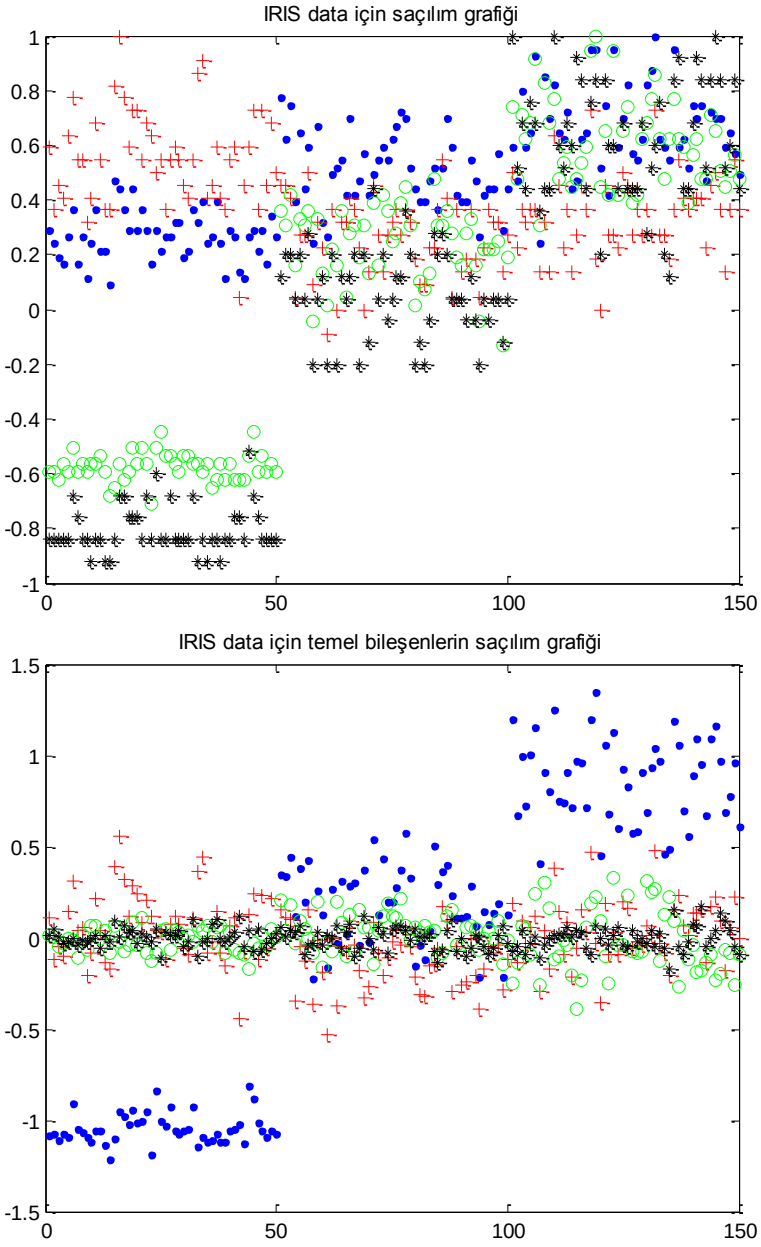
elde edilir. Temel bileşen skorlarının varyansı, özdeğerlere eşittir. Özdeğerlerin toplamı, XX^T matrisinin izidir. Bu, ilk temel bileşen skorunun orijinal verideki varyansı mümkün olduğu kadar çok açıkladığı ve ikinci temel bileşen skorunun, birinci temel bileşen tarafından açıklanamayan geride kalan varyansı mümkün olduğu kadar çok açıkladığı anlamına gelir ve bu böyle devam eder. Temel bileşenlerin standart sapmaları olan $\Lambda^{1/2}$ matrisinin köşegen elemanları, X matrisinin tekil değerleridir [Abdi, 2007]

PCA'nin tüm düşüncesinin merkezi, verinin kovaryans matrisidir. Kovaryans matrisinin köşegen elemanları, belirli özelliklerdeki varyansı yakalar ve köşegen dışı elemanlar, karşılık geldiği özellik çiftleri arasındaki kovaryansı belirler. PCA ile amaçlanan, kovaryans terimlerini sıfırlayacak şekilde dönüşüm yapmaktır [Hotelling, 1933].

Bu şekilde elde edilen temel bileşenler, verinin yeni bir dönüşümüdür. Tüm temel bileşenler orijinal değişkenlerin bir lineer kombinasyonu olup, her biri diğerleri ile ortogonal bir yapıya sahiptir. Böylece değişkenler arasındaki bağımlılık yapısının yok edilmesi sağlanmış olmaktadır. Ancak boyut indirgemenin gerçekleşmesi için elde edilen temel bileşenler arasından önemli olanların sayısına karar vermek gerekmektedir. Temel bileşenler analizi yapılırken, özdeğerler azalacak şekilde sıralanır. Temel bileşen skorlarının varyansının özdeğerlere eşit olması nedeniyle, ilk özdeğere ait olan temel bileşen, maksimum varyansa sahip olacaktır. İkinci bileşen ise en büyük ikinci varyansa sahiptir. Bu yapısından dolayı, ilk birkaç bileşenin maksimum varyansa sahip olması gerçeğinden hareketle, ilk iki veya üç bileşenin önemli bileşen olarak seçilmesi geleneksel bir yaklaşımdır. Bir başka yöntem ise, standartlaştırılmış veri matrislerinin kullanıldığı durumlarda birden büyük değerli özdeğerlerin sayısını, bileşen sayısı olarak almaktır [Alpar, 2011]. Bu konuyla alakalı farklı yöntemler olmasına rağmen halen daha literatürde tatmin edici bir çözüme ulaşılamamıştır. Bu bağlamda tezin bir amacı olarak, bilgi kriterleri yardımıyla boyut sayısına nasıl karar verileceğine üçüncü Bölüm 2.2.8'de değinilecektir.

Örnek: Temel bileşenler analiziyle, veri noktalarının temel bileşenler üzerine izdüşümünün nasıl yapıldığını görselleştirmek amacıyla, IRIS veri setine temel bileşenler analizi yapılmıştır. Bu veri seti Sir Ronald Fisher tarafından (1936) yılında diskriminant analizinin bir örneği olarak tanıtılmıştır. IRIS çiçeğinin üç farklı türü için çanak ve taç yapraklarının uzunlukları ve genişlikleri olmak üzere dört farklı değişken tanımlanmıştır. Sonuç olarak IRIS veri seti $150x4$ boyutunda bir veri setidir. Analiz için Matlab programında `pca` fonksiyonu kullanılmıştır. Bu fonksiyon, varsayılan olarak tekil değer ayrışımı ile temel bileşenler analizini yapmaktadır. Eğer öz ayrışım ile yapılması isteniyorsa, giriş parametrelerinde bunun için bir seçenek mevcuttur. Özdeğerler ve bileşenlerin açıklayıcılık oranları dikkate alındığında, 4 boyuta sahip olan bu veri setinin %7.76 oranında bir bilgi kaybıyla tek boyuta indirgenebileceği görülmektedir. Bu sonuç Şekil 2.1' den de anlaşılmaktadır.

Özdeğerler	Açıklayıcılık(%)
0.6608	92.1446
0.0368	5.1336
0.0160	2.2377
0.0035	0.4842



Şekil 2. 1: IRIS verisi için orijinal gözlem noktaları ile temel eksenler üzerindeki gözlem noktalarının dağılımı.

Üst panel, IRIS data için gözlem noktalarının konumunu belirtirken, alt panel temel bileşenler üzerinde bu gözlem noktalarının büyük çoğunluğunun tek boyut üzerine iz düştüğünü göstermektedir. Her bir değişken farklı renklerle işaretlenmiştir.

2.1.6 Olasılıklı Temel Bileşenler Analizi (Probabilistic Principal Component Analysis-PPCA)

PCA, her ne kadar uygulamada sıklıkla kullanılan bir yapı olsa da, aşağıdaki dezavantajlara sahiptir.

i. Bayescil karar ve karma modelleme gibi birçok alanda önemli olan bir olasılık modelinden yoksundur.
ii. Sahip olduğu lineer yapıdan dolayı, yüksek dereceden istatistiksel bilgi edinmek için kısıtlı bir yapıya sahiptir [Zhou, 2003; Zhou vd, URL-2]
iii. Veri girişlerinde eksiklik, kayıp gözlem varsa PCA nasıl uygulanır?
iv. Eğer verinin boyutunu oluşturan değişken sayısı, veri noktalarının sayısından çok büyük ise PCA nasıl uygulanır?
v. Global olarak değil de, lokal olarak PCA uygulamak mümkün müdür?

PCA gibi deterministik bir metodun, bu sorunlarla direkt olarak ilgili olması zordur. Diğer taraftan bir olasılık modeli, bu sorunların üstesinden kolaylıkla gelebilir. Eksik girişler öğrenmek için entegre edilebilir, bir EM algoritması iteratif olarak sorunu çözmek için kullanılabilir, karma modelleme ile lokalleştirilmiş PCA uygulanabilir [Yu, 2006].

İşte olasılıklı Temel Bileşenler Analizi (PPCA) bu amaçları gerçekleştirmek için geliştirilmiş bir yöntemdir [Tipping ve Bishop, 1997, 1999] Ayrıca, lineerlik dezavantajını ortadan kaldırmak için geliştirilen Kernel Temel Bileşenler Analizi gibi lineer olmayan bazı metotlar da vardır. Ancak bu metotlar tez kapsamının dışındadır.

İlk olarak, Lawley (1953) ve Anderson (1956)'da, temel bileşenler analizinin gözlenen değişkenlerin bir olasılık modeline dayalı olarak bir maksimum olabilirlik prosedürünü göstermişlerdir. Michael E. Tipping ve Christopher M.Bishop (1997)'de bu düşünceyi yineleyip geliştirerek, faktör analiziyle daha yakından ilişkili olan bir gizli (*latent*) değişken modelinde parametrelerin maksimum olabilirlik tahminleri sayesinde, gözlenen veri vektörlerinin bir setinin temel eksenlerinin oluşturulabileceğini göstermişlerdir. Aynı çalışmada, PCA'nin aslında hassas bir tahmin çerçevesi üzerinden üretilebileceği belirtilmiştir. Bunun için faktör analizinde olduğu gibi bir gizli değişken modeli ele alınmıştır. Bir gizli değişken modeli, m boyutlu gözlenemeyen veya gizli t değişkenleri ile p boyutlu gözlenen X değişkenleri arasındaki ilişkiyi inceler.

$$X = \sum_{j=1}^{m} w_j t_j + \mu + \varepsilon$$

$$= Wt + \mu + \varepsilon \tag{2.14}$$

Geleneksel olarak, gizli değişkenler $t \sim N(0, I)$ şeklinde birim varyanslı bir Gaussian dağılımına sahip ve bağımsız olarak tanımlanır. İlave olarak hata terimleri $\varepsilon \sim N(0, \psi)$ şeklinde dağılıma sahiptir. ψ ; diagonaldir ve W; pxm boyutlu parametre matrisi faktör yüklerini içerir. μ, verinin ortalamasının maksimum olabilirlik tahmin edicisi olup sabittir. Verilen bu formülasyona göre gözlenen değişkenler $X \sim N(\mu, \mathrm{WW}' + \psi)$ şeklinde bir dağılıma sahip olacaktır. Burada ψ'nin diagonalliğinden dolayı gözlenen X değişkenleri, t gizli değişkenlerine bağlı ve şartlı olarak bağımsızdır. ψ gürültü modeli diagonal olduğu için, W faktör yükleri genel olarak temel eksenlerden farklıdır. Aslında faktör yükleri ve temel eksenler arasındaki benzerlik, ψ'nin elemanlarının yaklaşık olarak eşit olduğu durumda gözlenir. Eğer WW' modeli tam olursa ve bu yüzden örnek kovaryans matrisi S'e eşit olursa faktör yükleri teşhis edilebilir ve iterasyona başvurmaksızın S'in öz ayrışımı sayesinde analitik olarak belirlenebilir. Aksi durumda, W, μ ve σ^2 parametrelerinin tahmini için maksimum olabilirlik tahmin edicisinin tanımlanmasına ve bunun için de X'in marjinal dağılımına ihtiyaç vardır.

İzotropik bir gürültü modeli $\varepsilon \sim N(0, \sigma^2 I)$ için (2.14) denklemi t verildiği zaman $x -$ uzayı üzerinde aşağıdaki gibi bir olasılık dağılımı ima eder.

$$p(X \mid t, W, \mu, \sigma^2) \sim N(Wt + \mu, \sigma^2 I_p)$$

$$p(X \mid t, W, \mu, \sigma^2) = (2\pi\sigma^2)^{-p/2} \exp\left\{-\frac{1}{2}\left\|x - Wt - \mu\right\|^2\right\} \tag{2.15}$$

Gizli değişkenler üzerinde bir Gaussian dağılımı tanımlandığında, $p(x, t)$ ortak dağılımı, $p(x, t) = p(x|t).p(t)$ olarak yazılabileceğinden x'in marjinal dağılımı elde edilebilir. Yani $p(t) = 2\pi^{-m/2} exp\left\{-\frac{1}{2}t't\right\}$ için,

$$p(X \mid W, \mu, \sigma^2) = \int p(x \mid t).p(t)dt$$

$$= (2\pi)^{-p/2} \mid \Sigma \mid^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right\} \tag{2.16}$$

Burada,

$$\mu_x = E(Wt + \mu + \varepsilon) = \mu$$

$$Cov(X) = \Sigma = E\left[(Wt + \varepsilon)(Wt + \varepsilon)^T\right]$$

$$= E(Wtt^T W^T) + E(\varepsilon\varepsilon^T)$$

$$\Sigma = WW^T + \sigma^2 I_p \tag{2.17}$$

şeklindedir. Bu model altında gözlenen değerlerin olabilirlik fonksiyonu,

$$L(W, \mu, \sigma^2 \mid X) = \prod_{i=1}^{n} p(x_i \mid W, \mu, \sigma^2)$$

$$= (2\pi)^{-np/2} \mid WW^T + \sigma^2 I_p \mid^{-n/2} \exp\left\{-\frac{1}{2} tr\left[(WW^T + \sigma^2 I_p)^{-1} A\right]\right\} \tag{2.18}$$

olur. Burada $A = \sum(x_i - \mu)(x_i - \mu)'$ olduğu belirtilmelidir. Denklem (2.18)'den log-olabilirlik fonksiyonu

$$\log L(W, \mu, \sigma^2 \mid X) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log \mid \Sigma \mid -\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^T (\Sigma)^{-1}(x_i - \mu) \tag{2.19}$$

olarak elde edilir. W ve σ^2'ye bakılmaksızın μ'nün maksimum olabilirlik tahmin edicisi log-olabilirlik fonksiyonunun μ'ye göre kısmi türevinin sıfıra eşitlenmesi ile bulunur.
Yani,

$$\frac{\partial \log(L(W, \mu, \sigma^2 \mid X))}{\partial \mu} = \sum_{i=1}^{n}(\Sigma)^{-1}(x_i - \mu) = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.20}$$

şeklindedir. W ve σ^2'nin maksimizasyonu ise daha komplekstir, ancak yine de kapalı formda çözümleri mevcuttur. Tekrar, log-olabilirliğin W'ya göre kısmi türevi alınıp sıfıra eşitlenirse, S örnek varyans kovaryans matrisi olmak üzere,

$$\frac{\partial \log L(W, \mu, \sigma^2 \mid X)}{\partial W} = n(\Sigma^{-1} S \Sigma^{-1} W - \Sigma^{-1} W) = 0$$

$$S\Sigma^{-1} W = W \tag{2.21}$$

elde edilir. Bu denklemin üç muhtemel çözümü vardır,

i. $W = 0$

ii. $C = S$ için $W = U(L - \sigma^2 I)^{1/2} R$

iii. $C \neq S$ ve $W \neq 0$ için $W = U_m(L_m - \sigma^2 I_m)^{1/2} R$

ile verilir. Burada U_m, tüm kolonları S örnek varyans kovaryans matrisinin özvektörlerinden oluşan pxm boyutlu bir matris, L_m köşegen elemanları S'in λ_j, özdeğerlerinden oluşan mxm boyutlu diagonal bir matris, R ise keyfi bir ortogonal matrisdir. Kolaylık olması açısından R sıklıkla birim matris olarak alınır. Yani $R = I$'dır. Üçüncü çözümde karekök operasyonu, σ^2 tahmini yapıldığı zaman, pozitifliği garanti edeceği için, W'nın bu seçimi ile, Tipping ve Bishop'un (1997) ve (1999)'da belirttiği gibi, W üzerinde maksimum olabilirlik fonksiyonu, S örnek varyans kovaryans matrisinin özdeğerlerinde meydana gelir. Özel olarak W'nin olabilirlik fonksiyonu

$$\hat{W} = U_m(L_m - \sigma^2 I_m)^{1/2} R \tag{2.22}$$

alınabilir. Gürültü varyansının maksimum olabilirlik tahmin edicisi ise, dışarıda kalan özdeğerlerin ortalamasından başka bir şey değildir.

$$\hat{\sigma}^2 = \frac{1}{p-m} \sum_{j=m+1}^{p} \lambda_j \tag{2.23}$$

$\hat{\mu}, \hat{W}$ ve $\hat{\sigma}^2$ değerlerinin olabilirlik denkleminde yerlerine yazılması ile

$$L(\hat{W}, \hat{\mu}, \hat{\sigma}^2 \mid X) = (2\pi)^{-np/2} \left(\prod_{j=1}^{m} \lambda_j\right)^{-n/2} (\sigma^2)^{-\frac{n(p-m)}{2}} \exp\left(-\frac{n}{2\sigma^2} \sum_{j=m+1}^{p} \lambda_j\right) \exp\left(-\frac{nm}{2}\right) \tag{2.24}$$

şeklinde yazılabilir. Bu yüzden, log-olabilirlik fonksiyonu

$$\log L(\hat{W}, \hat{\mu}, \hat{\sigma}^2 \mid X) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\left(\prod_{j=1}^{m} \lambda_j\right) - \frac{n(p-m)}{2}\log(\hat{\sigma}^2) - \frac{nm}{2} \tag{2.25}$$

olarak elde edilir. Esasen, $-\frac{np}{2}\log(2\pi)$ ve $-\frac{nm}{2}$ ifadeleri sabit olduğu için, (2.25) denkleminin maksimize edilmesi $-\frac{n}{2}\log(\prod_{j=1}^{m}\lambda_j) - \frac{n(p-m)}{2}\log(\hat{\sigma}^2)$ ifadesinin minimize edilmesine eşdeğerdir [Tipping ve Bishop, 1997, 1999; Zhau, URL-3; Bozdogan ikili görüşme, 2013].

Bir olasılık çerçevesi içinde çalışmanın en kritik ve en önemli avantajı, istatistik tabanlı model seçim araçlarının kullanılmasına imkan vermesidir. Yani, olasılık yoğunluğuna dayanan bir yaklaşım, farklı PPCA modellerinin karşılaştırılmasını ve veri için en iyi modelin belirlenmesini kolaylaştırır.

2.1.7 PPCA için EM algoritması

EM (Expectation-Maximization) algoritması gizli değişkenlere sahip olan olasılıksal modeller için, maksimum olabilirlik çözümlerinin bulunmasında kullanılan genel bir tekniktir. Algoritma, yakınsamaya kadar iki adım arasında dönüşümlü olarak devam eder, Expectation (E-adımı) ve Maximization (M-adımı). E-adımında, gizli değişkenlerin beklenen değerleri, gözlenen verilerden ve model parametrelerinin mevcut tahminleri kullanılarak tahmin edilir. M-adımında, model parametreleri bir önceki E-adımında türetilen gizli değişkenlerin beklenen değerlerini kullanarak ve log-olabilirlik fonksiyonunu maksimize ederek tekrar tahmin edilir. Bu iki adım yakınsama olana kadar devam eder ve veri olabilirliğinin lokal minumumunu bulmayı garanti eder. Yakınsama için bir çok değerlendirme kriteri mevcuttur. Bazı kriterler, iterasyonlar arasındaki log-olabilirlik kazancını ölçerken, diğer bazı kriterler algoritmayı durdurmak için birleşik log-olabilirlik değerinin bir tahminini kullanır [Nyamundada, 2010].

PPCA'nde model parametrelerinin maksimum olabilirlik tahminlerini bulmak için bir EM algoritması kullanılabilir. Maksimum olabilirliğin yukarıda verilen kapalı formdaki çözümlerinden ziyade bir EM algoritmasının kullanılması aşağıdaki avantajlara sahiptir.

i. Yüksek boyutlu uzaylarda EM algoritması daha hızlı çalışır
ii. EM algoritması sayesinde, eksik gözlemlerin olduğu veri setleri üzerinde çalışılabilir [Zhau, URL-3]

E-adımı:

Mevcut parametreler ve X gözlem değerleri verildiği zaman, $p(t,x)$ ortak olasılığın log-olabilirlik fonksiyonunun, t gizli değişkenlerin sonsal şartlı dağılımı $p(t|x)$'e göre beklentisi alınır. O halde bu adım için $p(t,x)$ ve $p(t|x)$ tanımlanmalıdır.

$p(t)$ ile t'lerin ve bağımsız normal dağılımlı gürültü teriminin bir lineer fonksiyonu olan X'in dağılımı normal olduğu için, $p(t|x)$'in dağılımı da normaldir. $X = Wt + \mu + \varepsilon$ modeli altında,

$$E(t) = M^{-1}W^T(x-\mu) \tag{2.26}$$

$$E(tt^T) = \sigma^2 M^{-1} + E(t)E(t^T) \tag{2.27}$$

olur. Burada, $M = \mathrm{W}'\mathrm{W} + \sigma^2 I_m$'dir. Denklem (2.17)'de tanımlanan kovaryans modeli, pxp boyutunda iken, bu modelin mxm boyutunda olduğunu belirtmek gerekir. (2.26) ve (2.27) denklemlerinden $Cov(t) = \sigma^2 M^{-1}$ olarak elde edilir. Dolayısıyla

$$p(t \mid x, W, \sigma^2) \sim N(M^{-1}W^T(x-\mu), \sigma^2 M^{-1}) \tag{2.28}$$

$$p(t \mid x, W, \sigma^2) = (2\pi)^{-m/2} \mid \sigma^2 M^{-1} \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(t - M^{-1}W^T(x-\mu)^T)^T(\sigma^2 M^{-1})^{-1}(t - M^{-1}W^T(x-\mu))\right\} \tag{2.29}$$

yazılabilir.

$p(x,t)$ ortak olasılık dağılımı için $p(x|t).p(t)$ tanımından hareketle

$$p(x,t) = (2\pi)^{-p/2} \mid \Sigma \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}(2\pi)^{-\frac{m}{2}} \exp\left\{-\frac{1}{2}t^T t\right\} \tag{2.30}$$

elde edilir. Buradan hareketle, ortak olasılığın log- olabilirlik fonksiyonu

$$\log L_c(W, \mu, \hat{\sigma}^2 \mid X, t) = \sum_{i=1}^{n}\{\log p(x,t)\} \tag{2.31}$$

$$\sum_{i=1}^{n}\{\log p(x\,|\,t)+\log p(t)\} \tag{2.32}$$

$$=\sum_{i=1}^{n}\left\{-\frac{p}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\|x_i-\mu\|^2-\frac{1}{\sigma^2}(x_i-\mu)^T Wt-\frac{1}{2\sigma^2}t^T W^T Wt-\frac{m}{2}\log(2\pi)-\frac{1}{2}t^T t\right\} \tag{2.33}$$

şeklinde olur. E-adımı için (2.31) denkleminin $p(t|x)'$e göre beklentisi alınmalıdır.

$$E\left[\log L_c(W,\mu,\hat{\sigma}^2\,|\,X,t)\right]=$$
$$\sum_{i=1}^{n}\left\{\begin{array}{l}-\frac{p}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\|x_i-\mu\|^2-\frac{1}{\sigma^2}(x_i-\mu)^T Wt-\frac{1}{2\sigma^2}t^T W^T Wt\\ -\frac{m}{2}\log(2\pi)-\frac{1}{2}t^T t-\frac{m}{2}\log(2\pi)-\frac{1}{2}tr\left[E(t^T t)\right]\end{array}\right\} \tag{2.34}$$

M-adımı:

M-adımında ise, (2.31) nolu denklemin W ve σ^2'ye göre kısmi türevleri alınıp sıfıra eşitlenerek $logL_c(W,\mu,\hat{\sigma}^2|X,t)$ maksimize edilir. Bu işlemler için, (2.26) ve (2.27) nolu denklemlerin gözönüne alınması ve gerekli işlemlerin yapılması ile,

$$\tilde{W}=SW(\sigma^2 I+M^{-1}W^T SW)^{-1} \tag{2.35}$$

$$\tilde{\sigma}^2=\frac{1}{p}tr(S-SWM^{-1}\tilde{W}^T) \tag{2.36}$$

elde edilir. Burada tekrar, $S=1/n\sum(x-\mu)\,(x-\mu)'$ ve $M=\mathrm{W'W}+\sigma^2 I_m$ 'dir.

E-adımında gerekli olan istatistikler, (2.26) ve (2.27) denklemleri ve *S* ile *M* kullanılarak hesaplanır, M-adımında ise bu değerler ve (2.35) ile (2.36) denklemleri yardımıyla yeni parametreler hesaplanır. Bu prosedür, algoritma için belirlenmiş yakınsama kriterine göre, yakınsama sağlanana kadar devam eder [Tipping ve Bishop, 1997,1999; Zhau URL-3].

Sonuç olarak, PPCA hem kovaryans matrisinin özdeğer ve özvektörlerinin hesabını gerektiren açık formda, hem de özellikle kayıp gözlemlerin olduğu veriler veya yüksek boyutlu veriler üzerinde daha etkili olabilecek bir EM algoritmasına göre tanımlanmıştır.

Örnek: PPCA analizi, Matlab programında `ppca(data,k)` komutu ile uygulanmaktadır. Analiz sonucunda, bileşen skorlarına, her bir bileşen üzerindeki orijinal değişkenlerin katsayılarına ve özdeğerlere ulaşılabilmektedir. Giriş komutlarında `data` yerine analiz edilecek veri, `k` yerine ise, hesaplanması istenen bileşen sayısı yazılmaktadır. Bu bileşen sayısı maksimum değişken sayısının bir eksiği kadar olmaktadır. Yani PPCA analizi ile değişken sayısı kadar değil, bir eksiği kadar özdeğere ve temel eksene ulaşılır. Daha önceden kullanılan IRIS data için PPCA analizi uygulanması sonucunda, klasik PCA'den farklı bir sonuca ulaşılmamıştır. Çünkü IRIS data içerisinde kayıp gözlem bulunmamaktadır.

PPCA'in faydasını daha iyi anlayabilmek için, kayıp gözlemlerin bulunduğu bir veri üzerinde sonuçları klasik PCA ile karşılaştırmak anlamlı olacaktır. Bu nedenle IRIS data üzerinde, `data(x,y)=NaN` komutu ile keyfi olarak bazı gözlemler silinmiş ve analizler yapılmıştır. PCA sonucunda, kayıp gözlemlerin olduğu satırların tamamen yok sayılarak analiz yapıldığı, PPCA sonucunda ise EM algoritması sayesinde, kayıp gözlemlerin beklentisi hesaplanarak analiz yapıldığı ve bu nedenle bilgi kaybının yaşanmadığı tespit edilmiştir. Sonuç olarak, kayıp gözlemi olmayan veri setleri için PCA ve PPCA sonuçları aynı olmakla beraber, bu noktada PPCA'nin tek üstünlüğünün bir olabilirlik fonksiyonuna sahip olması söylenebilir.

2.2 Model Seçimi ve Bilgi Kriterleri

Model seçimi, temelde açıklayıcı değişkenlerden hangisi veya hangilerinin cevap değişkeni üzerinde etkili olduğunu ortaya koyarak mevcut modeller altkümesi içinden en iyi modelin bulunması ile ilgili bir süreçtir. Açıklayıcı değişken sayısının çok olduğu ve model hakkında önsel bilginin bulunmadığı veya yetersiz olduğu durumlarda en iyi modelin belirlenmesi araştırmacıların önündeki en büyük problemlerden biridir. Özellikle açıklayıcı değişkenlere ilişkin alternatif alt küme sayısının milyonları bulabildiği bazı durumlarda ($p = 20$ için $2^{20} = 1048576$ adet değişken kombinasyonunun olması gibi) geniş çözüm uzayını açıklamak için, model seçiminde nümerik optimizasyon teknikleri ve stratejilerine ihtiyaç duyulur. Genel olarak bir alt küme seçim probleminde nümerik teknikleri kullanmak için iki tane bileşene ihtiyaç duyulur.

i. Çözüm uzayını etkili bir şekilde tarayacak bir algoritma
ii. Rakip modellerin karşılaştırılması için bir kriter veya ölçü

1970'lerin başından beri, geçtiğimiz birkaç on yılda, model seçim algoritması ve kriterleri ile ilgili birçok çalışmaya rastlamak mümkündür. Bunlar arasında klasik model seçim metotları ve bilgi kriterlerine dayanan model seçim metotları yer almaktadır. Klasik seçim metotları genellikle hipotez testleri ile gerçekleştirilir. Açıklayıcı değişkenlerin sonuç modelinde yer alıp almayacağına karar vermek için araştırmacı tarafından önceden keyfi bir güven düzeyi seçilir. Bununla beraber istatistikçilerin ve diğer bilim adamlarının çoğu, klasik model seçiminde kullanılan güven düzeyinin temelsiz olduğunu vurgulamaktadırlar [Linhart ve Zucchini,1986]. Bazı bilim adamları ise hipotez testi yaklaşımının teorik bir doğruluğa sahip olmadığını ve genellikle yetersiz ölçüde geçerli olduklarını ifade etmişlerdir [Burnham ve Anderson, 2002].

Popüler istatistik paketlerinin hemen hemen tamamında ileriye doğru seçim (Forward Selection), geriye doğru eleme (Backward Elemination) ve adımsal teknikler (Stepwise) gibi birçok klasik model seçim prosedürleri vardır. Ancak bu seçim metotları değişkenler arasındaki bağımlılık yapısıyla ilgilenmezler. Özellikle ileriye doğru seçim ve geriye doğru eleme yöntemlerini, Boyce vd. (1974)'de şu sözlerle kritize etmişlerdir.

"Algoritmaya giren ve çıkan değişkenlerin sırası ile ilgili teorik bir doğrulama yoktur veya çok azdır."

Dolayısıyla doğru değişkenlerin seçimi rastlantısallık içermektedir. Adımsal tekniklerle ilgili önemli bir eleştiri ise Mantel (1970), Hocking (1983) ve Sokal & Rohlf (1981) tarafından şu şekilde yapılmıştır.

"Adımsal teknikler, tamamıyla doğru modeli nadiren belirleyebilmektedir. Adımsal teknikler en iyi ihtimalle uygun bir model belirleyebilir."

Farklı klasik model seçim prosedürlerindeki tüm bu eksiklikler, en iyi veya en iyi olmaya yakın model alt kümelerinin seçiminde sınırlamalar ortaya koymaktadır. Bazı istatistikçiler ve araştırmacılar tüm muhtemel alt kümeler üzerinden seçim yapmayı önerseler de birçok durumda bu metot, hesaplama açısından uygun olmamakla beraber zaman ve maliyet açısından da mümkün değildir [Bozdogan, 2004]

Bilgi kriterlerine dayanan model seçim prosedürleri, klasik yaklaşımlara bir alternatif oluşturmaktadır. Bu türdeki kriterlerin temel fikri, gerçek model ve en uygun model altında cevap değişkenlerinin dağılımları arasındaki Kullback-Leibler uzaklığını minimize etmektir. Literatürde bilgi kriterlerine dayanan kriterler tanıtılmış ve geliştirilmiştir. İlk olarak Akaike (1973), model seçiminde yeni bir paradigma sağlayan ve Akaike Bilgi Kriteri (Akaike Information Criterion-AIC) olarak bilinen kendi kriterini geliştirmiştir. O zamandan beri, Akaike'nin kriterine dayanan bir çok kriter ortaya atılmıştır. Takeuchi (1976)'da önerdiği Genelleştirilmiş Akaike Bilgi Kriteri (GAIC) olarak da bilinen Takeuchi Bilgi Kriteri (Takeuchi's Information Criterion- TIC or AIC_T), Sugiura (1978)'de geliştirdiği Düzeltilmiş Bilgi Kriteri (Corrected Information Criteria- AIC_c), Bayes Bilgi Kriteri (Bayesian Information Criterion-BIC) olarak da bilinen ve Schwartz (1978) tarafından geliştirilen Schwartz Bayesci Bilgi Kriteri (Schwart's Bayesian Information Criterion-SBC), Hannan ve Quinn (1979) tarafından geliştirilen Hannan-Quinn Bilgi Kriteri (Hannan & Quinn's Criterion-HQC), Bozdogan (1987) tarafından geliştirilen Tutarlı Akaike Bilgi Kriteri (Consistent Akaike Information Criterion-CAIC) ve yine aynı yıl Fisher bilgi matrisinin kullanımını önerdiği Tutarlı Akaike Bilgi Kriteri CAICF, bu kriterlerden bazılarıdır.

Bu kriterler, modellerden birinin seçiminde uyum iyiliğini ve model yalınlığını dikkate almaktadırlar. Yalınlık için genel bir kural, daha basit (yalın) bir modelin daima daha karmaşık bir modele tercih edilmesidir. Dolayısıyla en iyi model, en yalın ve en yüksek bilgiye sahip olan modeldir [Deniz, 2007]. Kullback-Leibler ölçüsüne dayanan bu metotlar, günümüze kadar sıklıkla kullanılan R^2 gibi metotlardan kesinlikle daha iyi sonuç vermektedirler [Liu, 2007]. Dolayısıyla, bu kriterlerin yapısının daha iyi anlaşılabilmesi için öncelikle bilgi teorisinde önemli yerlere sahip olan Entropi, Kullback-Leibler ölçüsü ve Fisher Bilgi Matrisi kavramları tanıtılacaktır.

2.2.1 Entropi

Evrendeki herşey olasılığı daha az olan bir durumdan, olasılığı daha yüksek olan bir duruma doğru sürekli bir artış içindedir. Bu durum tüm kapalı sistemler için geçerlidir. Bu olasılık artışının bir ölçüsü olarak ortaya çıkan entropi basit olarak, bir şeyin herhangi bir durumda bulunabilme olasılığı olarak tanımlanabilir [Erbaş, 2010].

Entropi kelimesi termodinamik literatüründe ortaya çıkmış olmasına rağmen, Claude E. Shannon (1948) tarafından tasarlanan bilgi kavramı ile olan ilişkisi nedeniyle, kullanımı hemen hemen tüm disiplinlerin içine girmiştir. Shannon'ın bir bilgi kaynağındaki belirsizliği ölçen formülasyonu, matematiksel benzerliği ve fizikteki önemi nedeniyle, termodinamikteki ve istatistiksel mekanikteki entropi ile aynı adla anılır. Ancak ayırt edilebilmesi için bilgi teorik entropi (information theoretical entropy) veya Shannon'ın entropisi olarak adlandırılır.

Bir araştırmadan elde edilebilecek veri seti içinde, her türlü ölçme düzeyinde sorulmuş soruların/gözlemlerin sonuçları mevcut olabilir. Hangi ölçme düzeyinde ölçülmüş olursa olsun, elde edilen sonuçlar olasılıklar biçiminde ifade edilebiliyorsa, değişkenler için entropi değerleri hesaplanabilir. En düzenli durum olasılığı en az olan durum, en düzensiz durum olasılığı en fazla olan durumdur. Bu nedenle entropi, sistemde var olan düzensizliğin derecesinin bir ölçüsüdür.

p tane muhtemel değere sahip olan bir X olayını ele alalım. X'in kesikli değerleri x_j $(j = 1,2, \ldots . p)$ olsun. x_j'nin her bir değeri, olasılık şartlarını sağlayan $p(x_j)$ olasılıklarına sahip olsunlar. Yani; $0 \leq p(x_j) \leq 1$ ve $\sum_{j=1}^{N} p(x_j) = 1$ olsun. $p(x_j)$ olasılığı ile $X = x_j$ olayı gözlendikten sonra elde edilen bilgi miktarı,

$$I(x_j) = \log\left(\frac{1}{p(x_j)}\right) = -\log p(x_j) \tag{2.37}$$

olarak tanımlanır. Olasılık birden küçük olduğunda, logaritması negatif bir sayı çıkacağı için, formüldeki eksi işareti bilgi miktarını pozitif yapmayı amaçlar. (2.37) nolu denklemde verilen x_j değerleri tarafından tanımlanan bilgi miktarı, onların meydana gelme olasılıklarının tersi ile ilişkilidir. Yani, eğer X olayı için p adet muhtemel değer farklı olasılıklarda meydana gelirse, özellikle de $p(x_j)$ olasılığı düşükse, o zaman çok daha ilginç bir durum olur ve bu yüzden daha yüksek olasılıklarda meydana gelebilecek diğer x_k değerlerinden ziyade X'in x_j değerini alması daha fazla bilgi içerir. X olayının entropisi ise bilginin ortalaması veya beklenen değeri olarak tanımlanır. Yani:

$$H(X) = E\{I(X)\} = \sum_{j=1}^{p} p(x_j) I(x_j) = -\sum_{j=1}^{p} p(x_j) \log p(x_j) \tag{2.38}$$

şeklindedir.

Sürekli durum için

$$H(x) = E\{I(X)\} = -\int_{-\infty}^{+\infty} p(x_j) I(x_j) dx \tag{2.39}$$

şeklindedir.

X rastlantı değişkeninin *H*(*X*) ile gösterilen entropisi, dağılımın düzgünlüğünü ölçer ve dağılıma ilişkin belirsizliğin bir ölçüsüdür. Entropi, olasılıksal modelin bilinmeyen parametresine dayalıdır ve örneklemden elde edilir. Entropi standart sapma gibi bir değişim ölçüsüdür. Örneğin μ ortalamalı, σ^2 varyanslı normal dağılımın entropi fonksiyonu $H(X) = log(\sqrt{2\pi}\sigma)$'dir [Thomaz,2004].

2.2.2 Kullback-Leibler Ölçüsü

Kullback ve Leibler (1951)'de yapmış oldukları bir çalışmada, iki olasılık dağılımı arasındaki uzaklığı bilgi cinsinden ölçmüşlerdir [Bozdogan, 1987]. Bilgi kuramı içerisinde yer alan temel kavramlardan biri olan bu uzaklık, Kullback-Leibler (KL) uzaklığı olarak bilinir. KL uzaklığı; göreli entropi, ıraksama, ayırım için bilgi gibi farklı isimlerle de adlandırılmaktadır. $p(x)$ ve $q(x)$, iki kesikli değişkenin olasılık dağılımı ise, bu iki dağılım arasındaki KL uzaklığı,

$$D(p \| q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{2.40}$$

ve sürekli durum için KL uzaklığı,

$$D(p \| q) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{2.41}$$

şeklindedir [Bozdogan,1990].

2.2.3 Fisher Bilgi Matrisi

p bağımsız bileşenden oluşan $X = (X_1, X_2, \dots, X_p)$ rassal değişkeninin olasılık yoğunluk fonksiyonu $f(X|\theta)$ olmak üzere, X rassal değişkenine ait olabilirlik fonksiyonu

$$L(\theta \mid X) = \prod_{i=1}^{p} f_i(X_i \mid \theta) \tag{2.42}$$

ve buradan log-olabilirlik fonksiyonu

$$l(\theta) = l(\theta \mid X) = \sum_{i=1}^{p} \log f_i(X_i \mid \theta) \tag{2.43}$$

olarak tanımlansın. Fisher bilgisinin beklenen değeri, log-olabilirlik fonksiyonun kullanımıyla aşağıdaki gibi tanımlanır.

$$\mathcal{F}(\theta) = -E\left(\frac{\partial^2 l(\theta)}{\partial \theta^2}\right) \tag{2.44}$$

Negatif olmayan değerler alan bu bilgi etkinlik ve yeterlilik kavramları ile yakından ilgilidir. θ parametresine ilişkin bilginin miktarını ölçer ve θ parametresinin yansız kestiricisinin doğruluğu ile orantılıdır. Fisher Bilgi Matrisi'nin tersi, Ters Fisher Bilgi Matrisi (Inverse Fisher Information Matrix-IFIM) model kovaryans matrisi ile yakından ilişkili olup bazı bilgi kriterlerinde önemli rol oynar. Gerçek θ parametresinin kestirimi olan $\hat{\theta}$'nın yardımı ile hesaplanan $\hat{\Sigma}$'nın kullanılmasıyla normal dağılım için IFIM,

$$\hat{\mathcal{F}}^{-1} = \begin{bmatrix} \hat{\Sigma}(\hat{\theta}) & 0 \\ 0 & \left(\frac{2}{n} D_{p+q}^{+}\left[\hat{\Sigma}(\hat{\theta}) \otimes \hat{\Sigma}(\hat{\theta})\right] D_{p+q}^{+}\right) \end{bmatrix} \tag{2.45}$$

şeklinde hesaplanır. Burada, D ile duplikasyon matrisi, D^{+} ile $(D_p{}^T D_p)^{-1} D_p{}^T$ şeklinde tanımlanan Moore-Penrose tersi, $\otimes$ ile kronecker çarpım ifade edilmektedir [Williams ve Bozdogan, 1995; Akbilgic,2011].

2.2.4 Akaike Tipi Bilgi Kriterleri

İstatistiksel model değerlendirme sürecinde iki tip hata ile karşılaşılabilir. Bunlardan birincisi modelleme hatası, ikincisi ise yan ve varyans olarak da adlandırılan tahmin hatasıdır. Başka bir ifade ile θ parametre vektörlerinin tahmini esnasında yapılan kestirim hatalarıdır. O halde, R toplam riski, $R(M)$ modelleme riskini, $R(E)$ kestirim riskini göstermek üzere bu durum,

$$\underbrace{Toplam\ Risk}_{R} = \underbrace{Modelleme\ Riski}_{R(M)} + \underbrace{Kestirim\ Riski}_{R(E)}$$

şeklinde ifade edilir. Model değerlendirme sürecinde, değişkenlerin seçimindeki amaç toplam risk R'nin en küçüklenmesidir. Bu anlamda bir model seçim kriteri, maksimum olabilirlik tahmini altında bir modelin toplam riskinin tahmin edicisidir [Bozdogan, 2000].

Akaike (1973) ve (1974) yılları arasında ard arda yayınladığı makaleler ile istatistiksel modelleme ve istatistiksel model tespiti veya değerlendirmesi alanlarındaki gelişmelere ön ayak olmuştur. Bu nedenle bu alandaki ilk araştırmacılardan biri olarak kabul edilmektedir. Akaike tipi bilgi kriterleri ise, Akaike'nin AIC kriterini temel alan kriterlere verilen genel bir addır. Bu kriterlerin uyum eksikliği bileşenleri aynı olmakla beraber, uyum eksikliğindeki yanlılığı telafi etmek için kullanılan ceza terimleri farklıdır [Akbilgic, 2011]. Araştırmacı için karar kuralı, minumum bilgi kriteri (hangi kriter kullanılmışsa) değerine sahip olan modeli seçmektir.

Ortalama beklenen olabilirliğin logaritmasının -2 katının yansız kestiricisi olan AIC, modelin uyum eksikliğinin değerlendirilmesi ve parametre sayısının cezalandırılması esasına dayalı bir kriterdir. Parametre sayısının ceza terimi olarak kritere eklenmesi, AIC'ni farklı boyutlu modellerin karşılaştırılmasına imkan sağlayan bir hale getirmektedir [Deniz, 2007]. $L(\hat{\theta})$, maksimize edilmiş olabilirlik fonksiyonu, $\hat{\theta}$, model altında θ parametre vektörünün maksimum olabilirlik tahmini ve k modeldeki bağımsız parametre sayısını göstermek üzere,

$$AIC = -2\log L(\hat{\theta}) + 2k \tag{2.46}$$

şeklinde tanımlanır. (2.46) eşitliğinde ilk terim, parametre kestirimi için maksimum olabilirlik yöntemi kullanıldığında, uyum kötülüğü veya yanlılığın bir ölçümü olduğu için

uyum eksikliği terimi, ikinci terim ise birinci terimdeki yanlılığı telafi etmenin bir ölçümü olduğu için ceza (penalty) terimi olarak adlandırılır. AIC, θ parametre vektörünün dağılımının bilindiği ve maksimum olabilirlik kestiricisi ile tahmin edilebildiği durumlarda kullanılır. Takeuchi (1976) ve daha sonradan Shibata (1989)'da iç içe olmayan veya farklı parametrelere sahip modeller arasından bir model seçme konusuyla ilgilenmişler ve bunun için Akaike'nin varsayımında bir esneklik düşünerek θ parametre vektörünün dağılımının bilinmediği ve maksimum olabilirlik kestiricisi ile tahmin edilemediği durumları incelemişlerdir. Bu durumda artık, maksimize edilmiş olabilirlik fonksiyonlarının ortalaması, log-olabilirlik fonksiyonlarının beklenen değerine yakınsamayacaktır.
Yani,

$$\frac{1}{n}\log L(x\,|\,\hat{\theta}) = \frac{1}{n}\sum \log f(x\,|\,\hat{\theta}) \rightarrow E_x\left[\log f(x\,|\,\hat{\theta})\right] \tag{2.47}$$

yazılabilir. Bu durumda maksimize edilmiş olabilirlik fonksiyonlarının ortalaması, log-olabilirlik fonksiyonlarının beklenen değeri arasında bir yanlılık söz konusu olacaktır. Bu yanlılık,

$$b = Bias = \frac{1}{n} tr(\hat{\mathfrak{F}}^{-1}\hat{\mathfrak{R}}) + O(n^{-1}) \tag{2.48}$$

şeklindedir. Burada $\hat{\mathfrak{F}}$, Fisher bilgi matrisinin iç çarpım veya Hessian formu, $\hat{\mathfrak{R}}$ ise Fisher bilgi matrisinin dış çarpım formudur. AIC formülünde yanlılığı telafi etmek için kullanılan $2k$ ceza terimi yerine (2.48) ifadesinin kullanılmasıyla Genelleştirilmiş Akaike Bilgi Kriteri ortaya çıkmıştır.

$$GAIC = -2\log L(\hat{\theta}) + 2tr(\hat{\mathfrak{F}}^{-1}\hat{\mathfrak{R}}) \tag{2.49}$$

GAIC, model seçim litaretüründe aynı zamanda Takeuchi'nin Bilgi Kriteri (TIC) veya AIC_T olarak da bilinmektedir [Bozdogan 2000]. Bayesci Bilgi Kriteri (BIC) olarak da bilinen ve Schwartz (1978) tarafından AIC'ın bir versiyonu olarak önerilen bir başka kriterdir. Bayes yaklaşımını temel alan ve Schwartz Bilgi Kriteri-SBC olarak da bilinen bu kriter

$$SBC = -2\log L(\hat{\theta}) + k\log(n) \tag{2.50}$$

şeklinde tanımlanmaktadır. AIC ile kıyaslandığı zaman daha büyük bir ceza terimine sahip olduğu için SBC ile seçilecek modelin, AIC ile seçilecek modelden daha küçük veya en azından eşit boyutta olması beklenir [Akbilgic, 2011].

Bilindiği üzere, örneklem büyüklüğü arttıkça, AIC'nin ilk terimi artar fakat ceza terimi olan $2k$, değişmez. Bunun anlamı ceza teriminin yanlılığı telafi etmedeki etkisinin, n örneklem büyüklüğü arttıkça azalmasıdır. O halde, AIC bilgi kriterinin tutarlılığının asimptotik durumda şüpheli olduğu söylenebilir. Ayrıca, ceza teriminin önündeki 2 sayısının kullanımı da literatürde eleştirilmiş ve Bhansali ve Downham (1977)'de 2'nin yerine bir α sabiti koyarak bu sabitin 1 ile 4 arasında değerler alabileceğini söylemiştir. Rissanen ise (1978)'de bu sayının rasgele seçildiğini bildirmiştir. Bu noktadan hareketle Bozdogan (1987)'de AIC kriterini geliştirmek, genişletmek ve onu asimptotik durumda tutarlı yapmak için önce Tutarlı Akaike Bilgi Kriteri CAIC'ni daha sonra Fisher Bilgi Matrisinin kullanıldığı tutarlı bilgi kriteri CAICF'i önermiştir.

$$CAIC = -2\log L(\hat{\theta}) + k\left[\log(n) + 1\right] \tag{2.51}$$

$$CAICF = -2\log L(\hat{\theta}) + k\left[\log(n) + 2\right] + \log|\mathcal{F}(\hat{\theta})| \tag{2.52}$$

CAIC formülü BIC formülüne benzemekle beraber, BIC'ın ceza teriminden parametre sayısı kadar daha fazladır. Dolayısıyla CAIC, şimdiye kadar bahsi geçen tüm kriterlere göre daha fazla cezalandırma yaparak daha yalın modeli seçebilecektir [Bozdogan, 2000; Deniz, 2007]. CAIC ve CAICF'in kullanılması ile aşırı uyum (overfitting) olarak da adlandırılan veriye aşırı bağımlı ya da veriye aşırı uyum gösteren model elde edilmesi olasılığı azaltılarak daha yalın modellerin elde edilmesi sağlanmış olur [Bozdogan, 1987].

Buraya kadar bahsi geçen bilgi kriterlerinin en büyük eksikliği değişkenler arasındaki ilişkileri göz ardı ederek model seçimi yapmalarıdır. İstatistiksel modellerde en iyi modelin seçimi, araştırmacının değişkenler arasındaki ilişkiler hakkında kesin bilgisinin olmadığı durumlarda ayrıca önemli bir sorundur. Bu noktadan hareketle, Bozdogan sadece uyum iyiliği ve model yalınlığını değil, aynı zamanda modelin karmaşıklığını da göz önüne alarak ICOMP tipi kriterleri geliştirmiştir [Bozdogan 1988, 1990, 1994, 1998, 2000, 2004, 2010,...]. Bu nedenle ICOMP tipi kriterlerin daha iyi anlaşılabilmesi için bilgi ölçümü ve bilgi karmaşıklığı kavramlarını tanımlamak faydalı olacaktır. Bu sebeple, daha önceki

paragraflarda tanımlanan Kullback-Leibler ölçüsünün çok değişkenli dağılımlar için tanımı verilecektir.

2.2.5 Bilgi Ölçümü ve Bilgi Karmaşıklığı

Rasgele bir vektör için komplekslik aşağıdaki gibi tanımlanır. "Rasgele bir vektörün kompleksliği, onun bileşenleri arasındaki bağımlılığın veya etkileşimin bir ölçüsüdür."[Bozdogan, 2004]. $f(X) = f(X_1, X_2, \dots, X_p)$ ortak dağılım fonksiyonu ve $f_j(X_j),\ j = 1,2, \dots, p$ marjinal dağılım fonksiyonları olmak üzere p değişkenli sürekli bir dağılım düşünelim. $X_1, X_2, \dots, X_p$ rasgele değişkenleri arasındaki bağımlılığın ölçüsü,

$$I(X) = I(X_1, X_2, ..., X_p) = E_f\left[\log \frac{f(X_1, X_2, ..., X_p)}{f_1(X_1).f_2(X_2)....f_p(X_p)}\right] \tag{2.53}$$

veya buna denk olarak,

$$I(X) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f(X_1, X_2, ..., X_p) \log \frac{f(X_1, X_2, ..., X_p)}{f_1(X_1).f_2(X_2)....f_p(X_p)} dX_1 ... dX_p \tag{2.54}$$

şeklinde tanımlanır [Kullback, 1968; Harris, 1978]. Burada $I(X)$ bağımsızlığa karşı sapan bilgi miktarı, Kullback-Leibler (KL) ölçüsüdür. $I(X)$ aynı zamanda beklenen karşılıklı bilgi veya uygun bilgi olarak da bilinen, değişken bileşenleri arasındaki beklenen bağımlılığın bir ölçüsüdür.

$I(X)$ aşağıdaki özelliklere sahiptir.

i. $I(X) = I(X_1, X_2, \dots, X_p) \geq 0$'dır. Yani beklenen karşılıklı bilgi negatif değildir.
ii. Eğer her p çifti için $X_1, X_2, \dots, X_p$ rasgele değişkenleri karşılıklı ve istatistiksel olarak bağımsız ise $f(X_1, X_2, \dots, X_p) = f_1(X_1).f_2(X_2) \dots f_p(X_p)$'dir. Bu durumda (2.54) eşitliği sıfıra eşit olur. O halde (2.54) eşitliği sıfırdan farklı ise bağımlılık vardır [Bozdogan, 2004].

(2.54)'deki KL sapması, Shannon (1948)'un entropisi ile ilişkilidir. Shannon'un Entropisi, herhangi bir X rastlantı değişkeninin dağılım fonksiyonu $f_X(X)$ olmak üzere,

$$H(X) = -E\left[\log f_X(X)\right] \tag{2.55}$$

olarak tanımlanır ve bu ifade X rastlantı değişkeninin $f_X(X)$ dağılımına ne kadar uyduğu bilgisini vermektedir [Akbilgiç, 2010]. Bu tanımı kullanarak, $H(X_j)\ j = 1,2,\dots,p$ marjinal entropiyi ve $H(X_1, X_2, \dots, X_p)$ ortak entropiyi göstermek üzere,

$$I(X) = I(X_1, X_2, ..., X_p) = \sum_{j=1}^{p} H(X_j) - H(X_1, X_2, ..., X_p) \tag{2.56}$$

şeklinde yazılabilir. (2.56) denklemi Watanabe (1985) tarafından, yapısal ilginçlik ve bağımlılık ölçüsü olarak adlandırılmaktadır. Bu ifade aynı zamanda Shannon Kompleksliği olarak da ifade edilmektedir. Yapı içinde ne kadar çok bağımlılık olursa, marjinal entropilerin toplamı da o kadar büyük olacaktır. Daha az ve daha önemli değişkenlerin ayıklanması isteniyorsa, bağımlılık varlığında bu değişkenlerin içerdiği bilginin karşılıklı olarak çoğaltılması anlamına geldiği için, onların istatistiksel olarak bağımsız olması istenecektir.

Bozdogan (1990, 2000, 2004) çalışmalarını takip ederek bir çok değişkenli dağılımın kompleksliğinin kuramsal bilgi ölçütünü tanımlamak için, $f(X) = f(X_1, X_2, \dots, X_p)$'i çok değişkenli normal dağılım fonksiyonu olarak ele alalım.Yani

$$f(X) = f(X_1, X_2, ..., X_p) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\} \tag{2.57}$$

olsun. Burada $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$, $-\infty < \mu_j < \infty$ $j = 1,2,\dots,p$ ve $\Sigma > 0$ olup $X \sim N_p(\mu, \Sigma)$'dır. (2.56) denkleminden ortak entropi

$$H(X) = \int f(x)\left[\frac{p}{2}\log(2\pi)|\Sigma| + \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]dx$$

$$= \frac{p}{2}\log(2\pi)|\Sigma| + \frac{1}{2}tr\left\{\int f(x)\Sigma^{-1}(x-\mu)^T(x-\mu)dx\right\} \tag{2.58}$$

yazılabilir. $E[(x-\mu)'(x-\mu)] = \Sigma$ olduğu gözönüne alınırsa

$$H(X) = \frac{p}{2}\log(2\pi) + \frac{p}{2} + \frac{1}{2}\log|\Sigma|$$

$$= \frac{p}{2}\left[\log(2\pi) + 1\right] + \frac{1}{2}\log|\Sigma| \qquad (2.59)$$

elde edilir. (2.59) denkleminden marjinal entropi

$$H(X_j) = -\int f(X_j)\log f(X_j)dx_j$$

$$= \frac{1}{2}\log(2\pi) + \frac{1}{2} + \frac{1}{2}\log|\sigma_{jj}| \quad j = 1, 2, ..., p \qquad (2.60)$$

olarak elde edilir.

Marjinal entropi ve ortak entropinin, çok değişkenli normal dağılım için değerleri elde edildikten sonra van Emden (1971, s.61) tarafından tanımlanan kovaryans kompleksliğinin başlangıç tanımı verilebilir. Bu tanım daha sonradan Bozdogan'ın kriteri ICOMP'ın formülünde kullanılan maksimal kovaryans kompleksliğinin tanımı için ilham kaynağı olacaktır. van Emden (1971)'de çok değişkenli normal dağılım için Σ kovaryans matrisinin bilgi kompleksliği için bir ölçü vermiştir. Bu ölçü,

$$I(X) = I(X_1, X_2, ..., X_p) = \sum_{j=1}^{p} H(X_j) - H(X_1, X_2, ..., X_p) \qquad (2.61)$$

ifadesinde marjinal ve ortak entropi denklemlerinin yerlerine yazılması ile,

$$I(X) = \sum_{j=1}^{p}\left[\frac{1}{2}\log(2\pi) + \frac{1}{2} + \frac{1}{2}\log|\sigma_{jj}|\right] - \frac{p}{2}\log(2\pi) - \frac{p}{2} - \frac{1}{2}\log|\Sigma| \qquad (2.62)$$

elde edilir ve bu ifade

$$C_0(\Sigma) = \frac{1}{2}\sum_{j=1}^{p}\log(\sigma_{jj}) - \frac{1}{2}\log|\Sigma| \qquad (2.63)$$

şeklinde indirgenebilir. Anlaşıldığı üzere σ_{jj}, Σ'nın j. köşegen elemanıdır ve p, Σ'nın boyutudur. Eğer değişkenler lineer olarak bağımsız ise, yani Σ bir diagonal matris ise $C_0(\Sigma) = 0$'dır. Eğer değişkenlerden bir tanesi, diğer değişkenlerin lineer bir fonksiyonu

olarak yazılabilirse, $|\Sigma| = 0$ olduğu için $C_0(\Sigma) = \infty$'dur. Ancak van Emden'in (1971)'de işaret ettiği gibi $C_0(\Sigma)$, Σ kovaryans matrisindeki komplekslik miktarının etkili bir ölçüsü değildir. Çünkü

i. $C_0(\Sigma), X_1, X_2, \dots, X_p$ rassal değişkenlerin ortak ve marjinal dağılımlarına bağlıdır.

ii. $C_0(\Sigma)$'nın ilk terimi, ortonormal dönüşüm altında değişebilir [Bozdogan,2004].

Σ'nın kompleksliğinin maksimal miktarını karakterize etmek için, Bozdogan (1988, 1990) tarafından $C_0(\Sigma)$'nın herhangi bir ortonormal dönüşüm altında maksimumu hesaplanmış ve bu yeni ölçü $C_1(\Sigma)$ olarak tanımlanmıştır. T herhangi bir ortonormal dönüşüm olmak üzere,

$$C_1(\Sigma) = \max_T \left\{ H(X_1) + \dots + H(X_p) - H(X_1, X_2, \dots, X_p) \right\}$$

$$C_1(\Sigma) = \max_T C_o(\Sigma) = \max_T \left\{ \frac{1}{2} \sum_{j=1}^{p} \log(\sigma_{jj}) - \frac{1}{2} \log |\Sigma| \right\} \tag{2.64}$$

şeklindedir. Ortonormal dönüşümler $tr(\Sigma) = \sigma_{11} + \cdots + \sigma_{pp}$ 'yi değişmez bıraktığı için

$$\max \sum_{j=1}^{p} \log(\sigma_{jj}) = \max \log \prod_{j=1}^{p} \sigma_{jj} \tag{2.65}$$

elde edilir. Bu maksimizasyonu yapmak için $\left(\prod \sigma_{jj}\right)^{1/p} \leq \frac{1}{p} \sum \sigma_{jj}$ olduğu gözönüne alınırsa,

$$= p \log tr(\Sigma) - p \log p$$

$$= p \log \left[\frac{tr(\Sigma)}{p} \right] \tag{2.66}$$

elde edilir. Bu değer, $C_0(\Sigma)$'da yerine yazılırsa

$$C_1(\Sigma) = \frac{p}{2} \log \left[\frac{tr(\Sigma)}{p} \right] - \frac{1}{2} \log |\Sigma| \tag{2.67}$$

yazılabilir [Bozdogan 1988, 1990]. $C_1(\Sigma)$, $C_0(\Sigma)$ için bir üst sınırdır ve hem varyanslar arasındaki eşitsizliği hem de Σ'da kovaryansların katkısını ölçer. Böyle bir ölçü model içindeki yüksek dereceden korelasyonları, farklılıkları, benzerlikleri ve model yapısının

ilginç yönlerini tanımlamak için, değerlendirme ve model seçim problemlerinde çok önemlidir. (2.67) denklemi için gerekli düzenlemeler yapılırsa,

$$C_1(\Sigma) = \frac{1}{2}\log\frac{\left[\frac{tr(\Sigma)}{p}\right]}{|\Sigma|} \tag{2.68}$$

ifadesi de yazılabilir. $|\Sigma|$, genelleştirilmiş varyans ve $\frac{tr(\Sigma)}{p}$ toplam varyans ortalaması olduğu için kompleksliği, genelleştirilmiş varyans ile toplam varyans ortalamasının geometrik ortalaması arasındaki logaritmik oran olarak da yorumlayabiliriz. Dahası, eğer Σ 'nın özdeğerlerini $\lambda_1, \lambda_2, \dots, \lambda_p$ olarak alırsak, o zaman $\frac{tr(\Sigma)}{p} = \bar{\lambda}_a = \frac{1}{p}\sum \lambda_j$ özdeğerlerin aritmetik ortalaması ve $|\Sigma|^{1/p} = \bar{\lambda}_g = \left(\prod \lambda_j\right)^{1/p}$ özdeğerlerin geometrik ortalaması olmak üzere Σ'nın kompleksliğini

$$C_1(\Sigma) = \frac{p}{2}\log\frac{\bar{\lambda}_a}{\bar{\lambda}_g} \tag{2.69}$$

olarak da yorumlayabiliriz. $C_1(\Sigma)$, Σ'nın özdeğerlerinin nasıl eşit olmadığını ölçer ve tek bir fonksiyon içinde determinant ve iz olarak adlandırılan çok değişkenli saçılımın en basit iki ölçeğini içerir. Σ'nın tam ranklı olamadığı durumlarda p değeri genellikle $s = rank(\Sigma)$ olacak şekilde değiştirilir.

Genel olarak, kompleksliğin büyük değerleri değişkenler arasında yüksek bir etkileşim olduğunu, küçük değerleri ise etkileşimin zayıf olduğunu gösterir. $C_1(\Sigma)$'nın minumum değeri yapıdaki en az kompleksliğe karşılık gelir. Başka bir ifade ile $\Sigma \to I$ olduğu zaman $C_1(\Sigma) \to 0$'dır. Bu, kuramsal bilgi karmaşıklığı ile hesaplama çabası arasında makul bir ilişki kurar. Dahası en az kompleksliğe sahip olan matrisin birim matris olduğu anlamına gelir. İstatistikte buna karşılık gelen yapılar, ortogonal tasarımlar veya ilişkilerin, bağlantıların olmadığı lineer modellerdir. Aksi durumlarda $C_1(\Sigma) > 0$'dır. $C_1(\Sigma)$'nın bir başka önemli özelliği ise $C_0(\Sigma)$'nın tersine ortogonal dönüşümler altında değişmez olmasıdır. Bu konuyla alakalı daha detaylı bilgi için Bozdogan (1988, 1990, 2004) kaynaklarına başvurulabilir.

2.2.6 ICOMP Tipi Bilgi Kriterleri

ICOMP, (**I; I**nformation-**COMP; Comp**lexity) Bozdogan tarafından (1988) yılında, çok değişkenli doğrusal ve doğrusal olmayan modellerde model seçimi için geliştirilen bir kriterdir. Her ne kadar AIC temeline dayanan bir ölçü olsa da, AIC'dan farklı olarak

ICOMP;

i. Sonlu örneklem dağılımlarından başlayarak bir modelin parametre kestirimlerinin kovaryans matris özelliklerinin bilgiye dayalı belirlenmesini
ii. Bir modelin doğruluğunun belirlenmesinde yeni bir yaklaşım olarak ters Fisher bilgi matrisi-IFIM özelliklerinin bilgiye dayalı belirlenmesini

kullanır. Model karmaşıklığının bir ölçümü olmaksızın, model davranışını kestirmek ve model kalitesini değerlendirmek zordur. Bu nedenle ICOMP tipi kriterlerin amacı; bir modelin karmaşıklığı ve uyum iyiliği arasındaki en uygun dengeyi sağlamaktır. ICOMP, aşağıdaki kayıp fonksiyonunu tahmin etmek için tasarlanmıştır.

$$Kayıp = \underbrace{\overbrace{Uyum\ eksikliği + Parametrelerin\ ceza\ terimi}^{AİC} + Komplekslik\ \text{ölçüsü}}_{İCOMP}$$

AIC, sadece uyum eksikliği ile ceza terimi arasında bir denge kurmayı amaçlarken; ICOMP, modeldeki parametrelerin birbirleriyle nasıl ilişkili olduklarını ölçen bir komplekslik ölçüsü de göz önüne alarak bu dengeyi kurmayı amaçlar. Bu nedenle, bağımsız parametre sayısını direkt olarak cezalandırmak yerine, modelin kovaryans kompleksliğini cezalandırır. Buradan ICOMP,

$$ICOMP = -2\log L(\hat{\theta}) + 2C_1(\Sigma) \tag{2.70}$$

şeklinde tanımlanır. Burada $L(\hat{\theta})$, maksimize edilmiş olabilirlik fonksiyonu, $\hat{\theta}$, model altında θ parametre vektörünün maksimum olabilirlik tahmini, C_1; (2.67) nolu denklemde tanımlanan gerçel değerli komplekslik ölçüsü ve Σ, modelin parametre vektörlerinin kovaryans matrisidir. Bu kovaryans matrisi çeşitli şekillerde tahmin edilebilir. Bunlardan biri (2.45) nolu denklemle tanımlanan kestirilmiş ters Fisher bilgi matrisini kullanmaktır. Bu durumda ICOMP'ın en genel formu,

$$ICOMP(IFIM) = -2\log L(\hat{\theta}) + 2C_1(\hat{\mathcal{F}}^{-1}) \tag{2.71}$$

olarak tanımlanır. (2.71) nolu denklemdeki ilk bileşen modelin uyum eksikliğini ölçerken ikinci bileşen ise tahmin edilen parametrelerin doğruluğunu göz önünde bulunduran kestirilen ters Fisher bilgi matrisinin kompleksliğini ölçer. (2.67) nolu denkleme dayanarak,

$$C_1(\hat{\mathcal{F}}^{-1}) = \frac{p}{2}\log\left[\frac{tr(\hat{\mathcal{F}}^{-1})}{p}\right] - \frac{1}{2}\log|\hat{\mathcal{F}}^{-1}| \tag{2.72}$$

ile hesaplanabilir. Yine $\hat{\mathcal{F}}^{-1}$'in tam ranklı olmaması durumunda p, $s = rank(\hat{\mathcal{F}}^{-1})$ olacak şekilde değiştirilir [Bozdogan 1990, 2000, 2004].

ICOMP tipi kriterlere farklı bir yaklaşım sonsal beklenen faydayı (Posterior Expected Utility_PEU) maksimize etmek amacıyla ICOMP'ın Bayes uyarlaması

$$ICOMP\_PEU = -2\log L(\hat{\theta}) + k + 2C_1(\hat{\mathcal{F}}^{-1}) \tag{2.73}$$

şeklinde tanımlanmıştır. Bu yaklaşım, model altında θ parametrelerinin önsel (prior) dağılımı $f_{prior}(\theta|y)$ ve sonsal (posterior) dağılımı $f_{post}(\theta|y)$ arasındaki KL uzaklığına bağlı olarak ortaya atılmıştır.

Bir başka yaklaşım ise $ICOMP_{MISS}$ kriteridir. Bozdogan, modellerin genelde yanlış belirlendiği varsayımından yola çıkarak bu kriteri

$$ICOMP\_MISS = -2\log L(\hat{\theta}) + 2\left[tr(\hat{\mathcal{F}}^{-1}\hat{\Re}) + C_1(\hat{\mathcal{F}}^{-1})\right] \tag{2.74}$$

olarak tanımlamıştır. Burada $\mathcal{F}$, Fisher bilgi matrisinin iç çarpım veya Hessian formu, $\mathcal{R}$ ise Fisher bilgi matrisinin dış çarpım formudur. Eğer model doğru belirlenmişse bu iki form birbirine eşit olacak ve $tr(\hat{\mathcal{F}}^{-1}\hat{\mathcal{R}}) = iz(I) = k$ olacaktır. Bu durumda,

$$\begin{aligned} ICOMP(IFIM)_k &= -2\log L(\hat{\theta}) + 2k + 2C_1(\hat{\mathcal{F}}^{-1}) \\ &= AIC + 2C_1(\hat{\mathcal{F}}^{-1}) \end{aligned} \tag{2.75}$$

elde edilir [Bozdogan ve Howe, 2012].

Bunların yanısıra ICOMP'ın farklı uygulama sahalarında tanımlanan ve uygulanan başka yapıları da söz konusudur. Daha detaylı bilgi için Bozdogan (1987,…, 2015) kaynaklarından faydalanılabilir. Çalışma kapsamında (2.76) denklemiyle verilen en yeni ICOMP tipi kriteri Consistent ICOMP- CICOMP kullanılmıştır.

$$CICOMP = CAIC + 2C_1(\hat{\mathcal{F}}^{-1}) \tag{2.76}$$

2.2.7 Bilgi Kriterlerinin Karşılaştırılması

İstatistikte karmaşıklığın tek bir tanımı yoktur. Deneysel olarak karmaşıklık rasgele bir vektörün bileşenleri arasındaki karşılıklı bağımlılığın bir ölçümü olarak tanımlanabilir. Karmaşıklığın büyük değeri, değişkenler arasındaki etkileşimin de büyük olduğunu göstermektedir. Bu nedenle model değerlendirmede karmaşıklık kavramı önemli bir rol oynamaktadır. Kapsamlı model, karmaşıklığının bir ölçümü olmaksızın, model davranışının kestirimi ve verilen bir modelin kalitesinin değerlendirilmesi zordur [Bozdogan, 1988]. Literatürde bir çok araştırmacı, AIC tipi kriterlerde yer alan ceza teriminin işte bu karmaşıklığı önlemek için yetersiz olup olmadığını sorgulamıştır. Bu noktada yöneltilen en büyük eleştiri model bileşenleri arasındaki ilişkileri hesaba katmamasıdır. Bu bağlamda, model bileşenleri arasındaki karşılıklı bağımlılığın derecesi göz önüne alınarak tanımlanan ICOMP tipi kriterler, AIC tipi kriterlere göre daha doğru karar vermeyi sağlar. En basit örneği ile, AIC tipi kriterler ile çoklu doğrusal bağlantı problemine sahip bir model seçmek olası iken, ICOMP tipi kriterler ile en düşük seviyede çoklu doğrusal bağlantıya ve mümkün olacak en az sayıdaki değişkene sahip bir model seçmek olasıdır.

ICOMP tipi kriterlerin, AIC tipi kriterlere göre bir diğer üstün özelliği farklı parametre kestirim yöntemlerinin sağlamlık özelliklerini incelemeye imkan veren kovaryansların karmaşıklık ölçüsüne sahip olmasıdır. Maksimum olabilirlik kestirimlerine dayalı AIC-tipi kriterler genelde yanlıdır ve model uyumluluğu en iyi modelin belirlenmesi süreçlerinde parametreler arası bağımlılığı, parametrelerin gereksiz veya doğru olması kavramlarını dikkate almazlar. Oysa ICOMP tipi kriterler, Fisher bilgi matrisinin rolünü ön plana çıkarmaktadır. Ters Fisher bilgi matrisinin izi ve determinantı, sırasıyla parametre duyarlılığının etkisi ve parametrelerin korelasyonunu dikkate alan karmaşık bir fonksiyonu temsil etmektedirler [Deniz, 2007; Akbilgic, 2011; Bozdogan, 1988].

2.2.8 PPCA'de Bilgi Kriterleri ile Boyut Seçmek

Bir önceki bölümde detaylı bir şekilde anlatılan bilgi kriterlerinin yapısından anlaşıldığı üzere, bir model için bilgi kriteri uygulayabilmenin en önemli zorunluluğu, maksimum olabilirlik tahmin edicisinin var olmasıdır. Bölüm (2.1.6)'da detaylı bir şekilde anlatıldığı gibi, olasılıklı temel bileşenler analizi bize bu tahmin ediciyi sağlamaktadır. Denklem (2.25) ile verilen olabilirlik tahmin edicisinin -2 ile çarpılması sonucunda, PPCA bilgi kriterlerinin uyum eksikliği terimi,

$$-2\log L(\hat{\Lambda}, \mu, \hat{\sigma}^2 \mid X) = np\log(2\pi) + n\log\left(\prod_{j=1}^{m}\lambda_j\right) + n(p-m)\log(\hat{\sigma}^2) + nm \quad (2.77)$$

şeklinde elde edilir. Buradan, PPCA modeli için Zhao tarafından tanımlanan serbest parametre sayısının $k = pm + 1 - \frac{m(m-1)}{2}$ olduğu da göz önüne alınırsa, aşağıdaki bilgi kriteri denklemleri yazılabilir [URL-3]. Çalışma kapsamında AIC ve CAIC'nin yanısıra ICOMP tipi kriterler arasından CICOMP'ın kullanımı önerildiği için sadece bu kritere yer verilmiştir.

$$AIC = np\log(2\pi) + n\log\left(\prod_{j=1}^{m}\lambda_j\right) + n(p-m)\log(\hat{\sigma}^2) + nm + 2k \quad (2.78)$$

$$CAIC = np\log(2\pi) + n\log\left(\prod_{j=1}^{m}\lambda_j\right) + n(p-m)\log(\hat{\sigma}^2) + nm + k\left[\log(n)+1\right] \quad (2.79)$$

$$CICOMP = np\log(2\pi) + n\log\left(\prod_{j=1}^{m}\lambda_j\right) + n(p-m)\log(\hat{\sigma}^2) + nm + k\left[\log(n)+1\right] + 2C_{1F} \quad (2.80)$$

Burada, C_{1F} değeri, C_1 maksimal kovaryans kompleksiliğinin Frobenius formu olup $\hat{\Sigma}_{model}$ modelin kovaryans matrisi ve λ_j bu kovaryans matrisine ait i. özdeğer ve $\bar{\lambda}$ özdeğerlerin ortalaması olmak üzere, aşağıdaki şekilde tanımlanır,

$$C_{1F}(\hat{\Sigma}_{model}) = \frac{1}{p}tr(\hat{\Sigma}_{model}{}^{T}\hat{\Sigma}_{model}) - \left[\frac{tr(\hat{\Sigma}_{model})}{p}\right]$$

$$= \frac{1}{4\bar{\lambda}^2} \sum_{j=1}^{p} (\lambda_j - \bar{\lambda})^2 \tag{2.81}$$

PPCA analizi uygulandıktan sonra, önemli bileşen sayısına karar vermek için izlenecek yol şöyledir. Her bir bileşen modele birer birer ilave edilerek her durum için bilgi kriterleri hesaplanır. Bilgi kriteri için en düşük değerin olduğu modeldeki bileşen sayısı en iyi bileşen sayısı olarak belirlenir. Bilgi kriterlerinin verideki gerçek boyutu tespit etmedeki başarısı, Bölüm 3-1'de gerçek boyutun bilindiği bir simülasyon çalışması ile gösterilmiştir.

2.3 Maksimum Entropi ve Bazı Düzgünleştirilmiş Kovaryans Matrisleri

Matematik, fizik, mühendislik ve diğer alanlarda, problemler çoğunlukla ölçülmüş verilere, verilen koşullara ya da varsayımlara dayanarak çözümlenir. Bir problemin çözümü ile ilgili olarak Jaques Hadamard'ın (1902)'de ortaya koyduğu tanıma göre,

i. Problemin bir çözümü olmalıdır (existince)
ii. Var olan çözüm tek olmalıdır (uniqueness)
iii. Çözüm problemin verilerine bağlı olmamalıdır (stability)

Bir problem bu üç özelliği aynı anda taşıyorsa iyi şartlandırılmış (well-conditioned), en az bir tanesini taşımıyorsa kötü şartlandırılmış (ill-conditioned) olarak tanımlanır [Erbaş, 2010].

X , (pxp) boyutlu bir veri matrisi olsun. Veri, Gauss veya diğer eliptik konturlu olan çok değişkenli-t, çok değişkenli üstel, çok değişkenli Cauchy, çok değişkenli Laplace gibi dağılımlar ile olasılıksal olarak modellendiği zaman, Σ kovaryans matrisi tahmin edilmelidir. Örnek boyutu n, değişkenlerin boyutu p 'den küçük olduğu zaman ($n << p$), Σ'nın klasik örnek maksimum olabilirlik tahmin edicisi $\hat{\Sigma}_{MLE}$ durağan olmayan, kötü şartlandırılmış, pozitif tanımlı olmayan ve hatta singüler bir yapıya sahiptir. Böyle bir durumda, pratik olarak tüm çok değişkenli analizlerde ihtiyaç duyulan ve hassas matris (precision matrix) olarak da ifade edilen kovaryans matrisinin tersi hesaplanamaz. Bu durum özellikle bu tez çalışmasının da ilgilendiği problem olan gen verileri gibi aşırı derecede küçük örneklem probleminin olduğu uygulamalarda gerçektir.

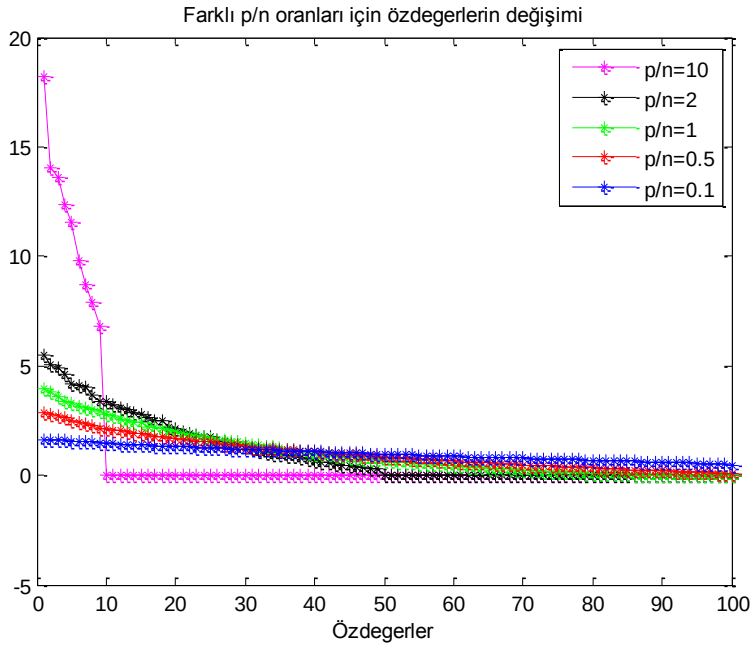
Hassas matris yani $\hat{\Sigma}_{MLE}^{-1}$, $\hat{\Sigma}_{MLE}$ 'nin determinantına bağlıdır ve tahmin edilen kovaryansı düzeltmek için azaltılmaya ihtiyaç duyulan denklem (2.82)'deki yanlılığa sahiptir [Shurygin, 1983].

$$E\left(\left|\hat{\Sigma}_{MLE}\right|\right) = |\Sigma|\left[1 - \frac{p(p+1)}{2n} + O(n^{-2})\right] \tag{2.82}$$

2.3.1 Düzgünleştirilmiş Kovaryans Yapıları

Temel bileşenler analizi, faktör analizi, sınıflama ve kümeleme analizleri, regresyon katsayılarının çıkarımı ve tahmini gibi çok değişkenli istatistiksel tekniklerinin bir çoğu, verinin kovaryans matrisinin ve/veya onun tersinin tahminini gerektirdiği ve kovaryans matrisi tahminindeki en büyük iki zorluğun, pozitif tanımlılık kısıtı ve p değişken sayısının n gözlem sayısından fazla olduğu yüksek boyutluluk problemi olduğu yukarıdaki paragraflarda belirtilmişti. Bu durumlarda amaç, örnek kovaryans matrisinden daha iyi şartlandırılmış ve daha doğru olan alternatif bir kovaryans matrisi bulmaktır. Gerçek kovaryansın doğru bir tahminine dayanan çok değişkenli teknikler için, son zamanlarda küçük örneklem boyutu altında yüksek boyutlu bir pxp kovaryans matrisi tahmin etmek önemli bir problem haline gelmiştir. Büyük p, küçük n (Big p, Small n) probleminin olduğu yerlerde, klasik örnek kovaryans matrisi sistematik olarak bozulan bir öz-yapıya sahip olduğu için bunun üstesinden gelecek farklı tahmin ediciler geliştirmek gereklidir [Chen, 2010].

Stein (1956, 1975)'de çok öncelerden de rapor edildiği gibi Σ kovaryans yapılı ve sıfır ortalamalı normal dağılan bir anakütleden gelen n boyutlu bir örneğin kovaryans matrisinin maksimum olabilirlik tahmini, p/n büyük olduğu zaman yansız ve pozitif tanımlı olmasına rağmen, kovaryans matrisinin doğru bir tahmin edicisi değildir. Bu durumda kovaryans matrisinin yapısı, en büyük özdeğerlerin yukarı yönde yanlı, en küçük özdeğerlerin aşağı yönde yanlı olması şeklinde bir bozulmaya uğrar. Bu durumu gösterebilmek amacıyla farklı p/n oranlarına sahip veri setleri üretilmiş ve örnek kovaryans matrislerinin özdeğerleri hesaplanmıştır. Sonuçlar Şekil 2.2'deki gibidir.



Şekil 2. 2: $1000x100, 200x100, 100x100, 50x100, 10x100$ boyutlarında üretilen veri setleri için özdeğerlerin değişimi.

Büyük sayılardaki p değerlerinde, p/n oranını ihmal edilebilir yapabilmek için gerekli gözlem sayısına ulaşmak zordur. Bu yüzden yüksek boyutlu kovaryans matrisleri için iyi şartlandırılmış bir tahmin edici bulmak önemlidir [Ledoit ve Wolf, 2004].

Singüler veya kötü şartlandırılmış kovaryans matris problemi için genel başlangıç çözümü ridge düzenlileştirmesidir. Ridge düzenlileştirmesinde

$$\hat{\Sigma}_R = \hat{\Sigma}_{MLE} + \alpha I_p \tag{2.83}$$

şeklinde özdeğerlerin ayarlanması ile kötü şartlandırılmışlığın üstesinden gelinmeye çalışılır. Genel olarak ridge parametresi α çok küçük seçilir. Ancak α,

i. Ne kadar büyük olabilir?

ii. Ne kadar küçük olmalıdır?

Bir kovaryans matrisi için en büyük özdeğerin, en küçük özdeğere bölümü olarak tanımlanan koşul sayısı $CN = \lambda_{max}/\lambda_{min}$ şeklindedir. Koşul sayısının tersi olarak kullanılan

$$\kappa(\Sigma) = \frac{1}{CN} \tag{2.84}$$

sayısı singülerliğin tanımı için kullanılabilir [Bozdogan, 2015]. $\kappa(\Sigma)$, sayısı sıfıra ne kadar yakın olursa, Σ kovaryans matrisi de singülerliğe o kadar yakındır demektir. Buradan hareketle bir kovaryans matrisine ne zaman ridge düzenlileştirmesi uygulanmalıdır sorusuna cevap verilecektir. Σ, kovaryans matrisi için

i. $\kappa(\Sigma) < 1e-10$ ise
ii. Pozitif tanımlı değilse

kovaryans matrisine ridge düzenlileştirmesi uygulanabilir [Bozdogan ve Howe, 2012].

Ridge düzenlileştirmesine alternatif olarak $\hat{\Sigma}_{MLE}$'nin özdeğerlerini merkezi bir değere doğru büzecek başka tahmin ediciler geliştirilmiştir. Bu yaklaşımların tamamı büzülme (shrinkage) tahmin edicilerine dayanır. Bu tahminlerin ana fikri, veya bu çalışmada düzgünleştirilmiş kovaryans tahmin edicileri (Smoothed Covariance Estimators-SCEs) dediğimiz şey, Σ'nın örnek kovaryansı $\hat{\Sigma}_{MLE}$ ile uygun seçilen bir hedef köşegen matrisi $\hat{D}$'nın konveks kombinasyonunu almaktır (yani ağırlıklandırılmış ortalama). O zaman kovaryans tahmin edicisinin büzülmüş veya düzgünleştirilmiş tahmin edicisi

$$\hat{\Sigma}_S = (1-\hat{\rho})\hat{\Sigma}_{MLE} + \hat{\rho}\hat{D} \tag{2.85}$$

şeklinde olur. Burada $\hat{\rho}$ optimal büzülme katsayısı (veya yoğunluk) olup 0 ile 1 arasında değer alır. Bu değer aynı zamanda gözlemlerin bir fonksiyonu da olabilir. $\hat{D}$ matrisi ise büzülme hedefi olarak adlandırılır. $\hat{D}$'nin naive formu,

$$\hat{D} = \frac{tr(\hat{\Sigma}_{MLE})}{p} I_p = (\frac{1}{p}\sum_{j=1}^{p}\lambda_j) I_p = \bar{\lambda} I_p \tag{2.86}$$

şeklindedir. Burada $tr(.)$ matrisin izini, λ_j , $j = 1,...,p$ için tahmin edilen örnek kovaryans matrisinin özdeğerlerini ve $\bar{\lambda}$ özdeğerlerin aritmetik ortalamasını gösterir. Denklem (2.85)'deki düzgünleştirilmiş kovaryans matrisinin tahmininde genel formdaki ağırlıklı ortalamanın kullanılması ile $\hat{\Sigma}_{MLE}$ tahmin edilen kovaryans matrisi için, aşırı derecede büyük

veya aşırı derecede küçük özdeğerler üzerinde daha düşük ağırlık koyulur. Böylece aşırı derecede büyük veya küçük özdeğerlerin etkisi azaltılır ve daha sağlam bir tahmin edici elde edilmiş olur.

Şuana kadar özdeğerlerin büzülmesine dayanan Stein tipi tahmin edicilerin daha küçük risk tahminlerine sahip olduğu ve genel olarak örnek kovaryansından daha doğru oldukları iyi bilinmektedir. Ancak Bickel and Li (2006) ve Warton (2008)'de belirttiği gibi $n \ll p$ olduğu zaman örnek kovaryans matrisinin kendisi de singüler olduğu için ridge düzenlileştirmesi, kovaryans matrisinin daha doğru ve iyi şartlandırılmış olmasına yol açacaktır [Porahmadi, 2012].

Bu tezde ilgilenilen problem kapsamında, denklem (2.85) ile belirtilen formun bazı lineer ve kuadratik kayıp fonksiyonları altında aşağıdaki düzenlenmiş veya düzgünleştirilmiş kovaryans tahmin edicileri kullanılmıştır. Bunlar literatürdeki orijinal isimleri ile:

i. Empirical Bayes tahmin edicisi [Haff, 1980]

$$\hat{\Sigma}_{EB} = \hat{\Sigma} + \frac{p-1}{n.tr(\hat{\Sigma})} I_p \tag{2.87}$$

ii. Stipulated Ridge tahmin edicisi [Shurygin, 1983]

$$\hat{\Sigma}_{SRE} = \hat{\Sigma} + p(p-1)\left[2n.tr(\hat{\Sigma})\right]^{-1} I_p \tag{2.88}$$

iii. Stipulated Diagonal tahmin edicisi [Shurygin, 1983]

$$\hat{\Sigma}_{SDE} = (1-\pi)\hat{\Sigma} + \pi diag(\hat{\Sigma})$$
$$\pi = p(p-1)\left[2n.tr(\hat{\Sigma}^{-1}) - p\right]^{-1} \tag{2.89}$$

iv. Convex Sum tahmin edicisi [Press, 1975; Chen, 1976]

$$\hat{\Sigma}_{CSE} = \frac{n}{n+m}\hat{\Sigma} + \left(1 - \frac{n}{n+m}\right)\left[\frac{tr(\hat{\Sigma})}{p}\right] I_p$$

$$0 < m < \frac{2\left[p(1+\beta) - 2\right]}{p-\beta}$$

$$\beta = \frac{tr(\hat{\Sigma})^2}{tr(\hat{\Sigma}^2)} \tag{2.90}$$

v. Thomaz Stabilization: λ_i, örnek kovaryans matrisinin i. özdeğeri, $\bar{\lambda}$; özdeğerlerin ortalaması, V; özdeğerlere karşılık gelen özvektörler olmak üzere [Thomaz, 2004],

$$\Lambda^* = \begin{pmatrix} \max(\lambda_1, \bar{\lambda}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \max(\lambda_p, \bar{\lambda}) \end{pmatrix}$$

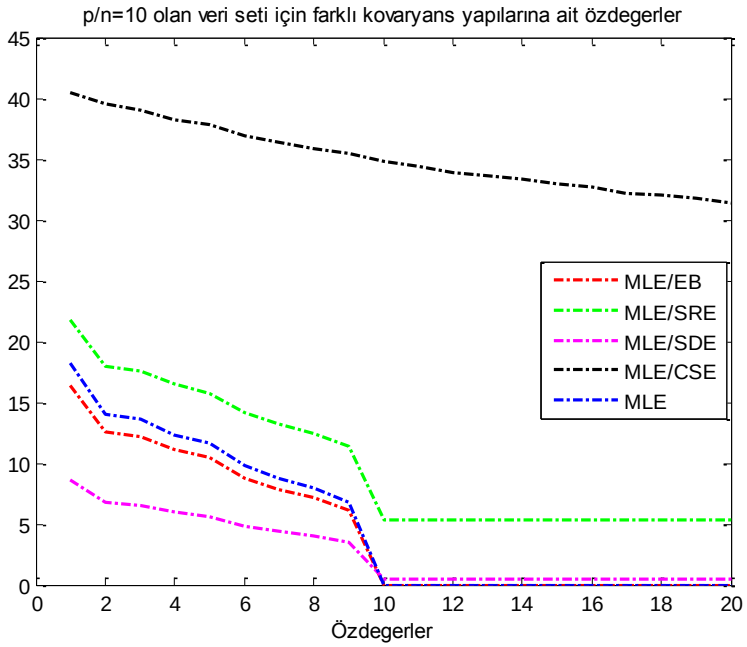
$$\hat{\Sigma}_{STA} = V\Lambda^* V \tag{2.91}$$

vi. Bozdoğan'ın Convex Sum tahmin edicisi [Bozdogan, 2010]

$$\hat{\Sigma}_{BCSE} = \hat{\rho}\hat{\Sigma} + (1-\hat{\rho})\hat{D} \tag{2.92}$$

olup burada $\hat{\rho} = \frac{1}{\alpha}$ ve α ise her boyuttaki sapmanın kareler toplamıdır. Yani $\alpha = \frac{1}{n-1}\sum_{j=1}^{p} Var(x_j)$ şeklindedir.

Yukarıdaki paragraflarda Şekil 2.2 ile, bir veri setinde p değişken sayısının artması ve giderek $n << p$ şeklinde küçük örneklem problemine ulaşılması durumunda özdeğerlerin bu durumdan nasıl etkilendikleri gösterilmişti. Şimdi ise $n << p$ problemine sahip olan p/n=10 olan veri seti için, yukarıda bahsi geçen düzgünleştirme işlemlerinin özdeğerler üzerine etkisi hesaplanmış ve Şekil 2.3'de gösterilmiştir.



Şekil 2. 3: Farklı kovaryans yapılarının özdeğerler üzerindeki etkisi.

2.3.2 Maksimum Entropi Kovaryans Tahmin Edicisi

Özdeğerlerin büyük çoğunluğunun negatif veya sıfır olarak elde edildiği durumlar için muhtemel bir çözüm Maksimum Entropi Kovaryans Matrisi'dir. Bu kovaryans yapısı, örnek kovaryans matrisine göre daha fazla farklılığı ortaya koyar ve temel bileşenler analizi gibi veri girişinin kovaryans matrisi şeklinde olduğu çok değişkenli yöntemlere yazılabilecek bir program sayesinde kolaylıkla uygulanabilir. Aşağıdaki paragrafta, maksimum entropi kovaryans matrisinin hesaplama adımları tanıtılacaktır.

İki değişkenli durum için aşağıdaki gibi bir Z matrisi ele alınsın.

$$Z = \begin{bmatrix} x_1 & y_1 \\ . & . \\ . & . \\ . & . \\ x_n & y_n \end{bmatrix} \tag{2.93}$$

Adım-1: Sıra istatistikleri oluşturulur

$$x^0 \equiv x^1 < x^2 < \ldots < x^n \equiv x^{n+1}$$
$$y^0 \equiv y^1 < y^2 < \ldots < y^n \equiv y^{n+1} \tag{2.94}$$

Adım-2: Birincil orta noktalar hesaplanır

$$\xi_j = \frac{1}{2}(x^j + x^{j+1}) \quad j = 0,1,...,n$$

$$\eta_j = \frac{1}{2}(y^j + y^{j+1}) \quad j = 0,1,...,n \tag{2.95}$$

Adım-3: Birincil orta noktaların yardımı ile ikincil orta noktalar hesaplanır.

$$\bar{x}^j = \frac{1}{2}\left(\xi_{j-1} + \xi_j\right) \quad j = 0,1,...,n$$

$$\bar{y}^j = \frac{1}{2}\left(\eta_{j-1} + \eta_j\right) \quad j = 0,1,...,n \tag{2.96}$$

Adım-4: Her bir gözlem çifti $P_k = (x_k, y_k)$için, bazı i ve j değerleri vardır ki $(x_k, y_k) = \left(x^i, y^j\right)$'dir. Buradan ikincil orta noktalarla ilişkili olan $\left(\bar{x}^i, \bar{y}^j\right)$ çifti vardır. Bu noktayı $\bar{P}_k = (\bar{x}_k, \bar{y}_k) = (\bar{x}^i, \bar{y}^j)$ olarak tanımlayalım. $\bar{P}_k$, orijinal P_k gözlem noktasının bir dönüşümüdür.

Adım-5*:* İkincil orta noktaların iç çarpım matrisi C oluşturulur.

$$C = \begin{bmatrix} \frac{1}{n}\sum(\bar{x}_k - \bar{x})^2 & \frac{1}{n}\sum(\bar{y}_k - \bar{y})(\bar{x}_k - \bar{x}) \\ \frac{1}{n}\sum(\bar{y}_k - \bar{y})(\bar{x}_k - \bar{x}) & \frac{1}{n}\sum(\bar{y}_k - \bar{y})^2 \end{bmatrix} \tag{2.97}$$

Adım-6: D ridge matrisi hesaplanır

$$D = \begin{bmatrix} \frac{1}{12n}\sum\left(\xi_j - \xi_{j-1}\right)^2 + \frac{1}{6n}\left\{\left(\xi_1 - \xi_0\right)^2 + \left(\xi_n - \xi_{n-1}\right)^2\right\} & 0 \\ 0 & \frac{1}{12n}\sum\left(\eta_j - \eta_{j-1}\right)^2 + \frac{1}{6n}\left\{\left(\eta_1 - \eta_0\right)^2 + \left(\eta_n - \eta_{n-1}\right)^2\right\} \end{bmatrix} \tag{2.98}$$

Adım-7*:* Maksimum entropi kovaryans matrisi $C + D$'dir [Fiebig, 1984].

Burada C pozitif tanımlı genel yayılım matrisi, D köşegenlerinde pozitif elemanların olduğu pozitif tanımlı köşegen matrisidir. Bu pozitif elementler, değişkenlerin ardışık birincil ve ikincil orta noktalarının arasındaki farkların ağırlıklandırılmış bir kareler toplamı

formundadır ve bu elementler maksimum entropi kovaryans matrisinde bir ridge olarak hizmet eder. Diğer bir ifadeyle, genel ridge tipi tahmin edicilerinde olduğu gibi ridge parametresinin nasıl seçileceğinden endişe etmeksizin, direkt ve otomatik olarak veriden üretir. İstatistik literatüründe ihmal edilmiş maksimum entropi kovaryans matrisinin tanıtılmasının ana fikri, gen ifade verilerinde olduğu gibi aşırı derecede küçük örneklem problemine sahip olduğumuz zaman, singüler ve kötü şartlandırılmış kovaryans matrisini pozitif tanımlı yapmasıdır. Ayrıca maksimum entropi kovaryans matrisi hakkında ilginç olan bir başka şey, veri setindeki bilgiyi lineer ve lineer olmayan sıra istatistiklerini kullanarak tamamen patlatmasıdır. Bu sıra istatistiklerinin kullanımı ile veri setinde ortalama, mod ve medyan gibi istatistikler etkilenmez. Bu da verinin sıralanması ile aslında verinin değiştirilmediği anlamına gelmektedir. Maksimum entropi kovaryans matrisinin hesabı hızlı ve yüksek boyutlu veri setleri için etkilidir ve ağır değildir. Bunun için yukarıda iki değişkenli durum için tanımlanan maksimum entropi kovaryans matrisi, p- değişkenli duruma genelleştirilmiş ve Matlab modülü yazılmıştır.

2.3.3 Hibritleştirilmiş Kovaryans Tahmin Edicisi (Hybridized Covariance Estimator-HCE)

Tez kapsamında ilgilenilen problem olan kanser sınıflamasında kullanılan gen ifade verileri, yapı olarak gözlem sayısının az, değişken sayısının fazla olduğu aşırı derecede küçük örnekleme sahip yüksek boyutlu veri setleridir. Kovaryans yapıları için buraya kadar anlatılan bilgiler göz önüne alındığında, bu verilerin kovaryans matrislerinin maksimum olabilirlik tahminlerinin singüler yapıya sahip olup, pozitif tanımlı olmayacakları açıktır. Bu nedenle, maksimum olabilirlik tahminleri $\hat{\Sigma}_{MLE}$ kullanılarak yapılacak klasik analizlerin doğruluğu ve geçerliliği şüphelidir. Bu çalışmada, öncelikle veri setinde var olan negatif özdeğer probleminin üstesinden gelebilmek amacıyla, maksimum olabilirlik tahminleri $\hat{\Sigma}_{MLE}$ yerine, maksimum entropi kovaryans matrisi $\hat{\Sigma}_{ME}$ hesaplanmıştır. Bir sonraki bölümde de gösterileceği üzere, pxp yüksek boyutlu bu kovaryans matrisi için, sıfırdan farklı ve pozitif olan p tane özdeğer elde edilebilmiştir. Bundan sonraki aşamada, herhangi bir düzgünleştirilmiş kovaryans tahmin edicisi veya onun denklem (2.91) ile belirtilen stabilizasyon algoritması ile stabil edilmiş özdeğerleri seçilebilir. Bu noktada, ME kovaryans matrisinin stabilizasyonu ile elde edilen $\hat{\Sigma}_{ME\_STA}$ kovaryansının özdeğerleri

önerilmektedir. Kovaryans matrisinin singülerlik yapısı denklem (2.84)'de verilen $\kappa(\Sigma) = \frac{1}{CN}$ [Bozdogan, 2015] ile incelenmiştir. $\kappa(\Sigma)$, sayısı sıfıra ne kadar yakın olursa, Σ kovaryans matrisi de singülerliğe o kadar yakındır demektir. $\kappa(\Sigma)$ sayısını sıfırdan daha fazla uzaklaştırmak ve singülerliği daha iyi çözebilmek için $\hat{\Sigma}_{ME\_STA}$ kovaryans yapısının Bölüm 2.3.1'de maddeler halinde verilen diğer düzgünleştirilmiş kovaryans yapıları ile bazı hibritleştirilmiş formları oluşturulmuştur.

Örnek olarak, λ_i, maksimum entropi kovaryans matrisinin i. özdeğeri, $\bar{\lambda}$; özdeğerlerin ortalaması, V; özdeğerlere karşılık gelen özvektörler olmak üzere,

$$\Lambda^* = \begin{pmatrix} \max(\lambda_1, \bar{\lambda}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \max(\lambda_p, \bar{\lambda}) \end{pmatrix}$$

$$\hat{\Sigma}_{ME\_STA} = V \Lambda^* V \tag{2.99}$$

Thomaz (2004) stabilizasyonudur. Bu kovaryansın Convex Sum yapısında kullanılması ile

$$\hat{\Sigma}_{ME\_STA\_CSE} = \frac{n}{n+m}\hat{\Sigma}_{ME\_STA} + \left(1 - \frac{n}{n+m}\right)\left[\frac{tr(\hat{\Sigma}_{ME\_STA})}{p}\right] I_p$$

$$0 < m < \frac{2\left[p(1+\beta) - 2\right]}{p - \beta}, \quad \beta = \frac{tr(\hat{\Sigma})^2}{tr(\hat{\Sigma}^2)} \tag{2.100}$$

ME_STA_CSE hibrit yapısı elde edilmiş olur.

Benzer şekilde diğer kovaryans yapılarının da hibrit formları elde edilebilir. Stabilizasyon+Hibridizasyon kullanmanın mantığı ve matematiksel teması, büzülmeye göre pozitif tanımlılığı başarmak ve düzgünleştirilmiş yapılar ile hibrit yapmadan önce, stabilizasyon ile kovaryans matrisinin daha büyük özdeğerlerini değişmeden tutarak, daha az uygun ve daha küçük olan özdeğerlerini genişletmektir.

Bundan sonra bu kovaryans yapıları Hibrit Kovaryans Tahmin Edicisi-Hybrid Covariance Estimator (HCE)- $\hat{\Sigma}_{HCE}$ olarak adlandırılacaktır. Böylece, hibrit kovaryans tahmin edicileri $\hat{\Sigma}_{HCE}$ kullanılarak PPCA analizi yapıldığında, veri ne kadar yüksek boyutlu

olursa olsun, boyut sayısı kadar temel bileşen elde edilebilmekte ve klasik PCA uygulanması sonucunda ihmal edilebilecek yüksek boyutlardaki bazı gizli bilgilere de ulaşılabilmektedir. Artık bu kovaryans yapıları, temel bileşenler analizi ve bilgi kriterleri için gerekli olan tüm hesaplamalarda kullanılabilir.

2.4 Hibrit Boyut İndirgeme (Hybrid Dimension Reduction-HDR)

Buraya kadar yüksek boyutlu sınırlı örneklem büyüklüğüne sahip veri setlerinde, önemli bilgiyi ortaya çıkarmak ve sonuç olarak boyut indirebilmek amacıyla önerilen modelin teorik altyapısı üzerinde durulmuştur. Bundan sonra bu yöntem Hibrit Boyut İndirgeme (Hybrid Dimension Reduction-HDR) olarak adlandırılacaktır. Özetle HDR'nin hesaplama adımları şöyledir:

Aşama-1: Yüksek boyutlu kanser gen ifade verisi için maksimum entropi kovaryans matrisi hesaplanır

Aşama-2: Maksimum entropi kovaryans matrisi ile Bölüm 2.3.3'deki adımları kullanarak $\hat{\Sigma}_{HCE}$ kovaryans matrisi oluşturulur

Aşama-3: $\hat{\Sigma}_{HCE}$ ile PPCA analizi yapılır

Aşama-4: Bilgi Kriterleri ile uygun $\hat{\Sigma}_{HCE}$ yapısı ve boyut sayısına karar verilir.

HDR yönteminin başarısı ve uygulamadaki performansı Bölüm 3.1'de yapılacak simülasyon çalışmaları ile türetilmiş veri setleri üzerinde gösterildikten sonra, Bölüm 3.2'de altı tane kanser gen ifade verisi üzerinde gösterilecektir.

2.5 Aykırı Gözlem Tespiti (Detection of Outliers)

Genel olarak hata veya gürültü olarak düşünülen aykırı gözlemler (outliers) bazı durumlarda önemli bilgiye sahip olabilmektedirler. Aykırı gözlemlerin varlığı durumunda sonuçlar yanlış, parametre tahminleri yanlı ve uygulanan model hatalı belirlenmiş olabilir. Bu nedenle modelleme ve veri analizine başlamadan önce aykırı gözlemlerin tespitinin yapılması ve bu gözlemlerin etkilerinin neler olabileceğinin değerlendirilmesi oldukça önemlidir.

Bir aykırı gözlemin tanımı, verinin yapısı ve uygulanan aykırı gözlem tespit metoduyla ilgili saklı varsayımlara bağlıdır. Ancak bazı tanımlar çeşitli veri tipleri ve uygulanan metotlar için yeterli kabul edilmektedir. Hawkins (1980)'de bir aykırı gözlemi, diğer gözlemlerden çok fazla derecede ayrılan ve farklı bir mekanizmadan üretildiği düşünülen bir gözlem olarak tanımlar. Benzer şekilde Barnett ve Lewis (1994) ile Johnson (1992)'de aykırı gözlemi, verinin geri kalan üyeleri düşünüldüğünde onlara göre tutarsızmış gibi görünen gözlem olarak tanımlamışlardır. Yazarların bahsettiği "tutarsızmış gibi görünme" terimi aykırı gözlemler ile ilgilenildiği zaman temel problemdir ve aykırı gözlem tespit etme metotları bu problemi çözmeyi amaçlar.

Günümüze kadar, istatistiksel ve olasılıksal bilgi, uzaklık ve benzerlik-farklılık fonksiyonları, metrikler ve kerneller, verinin etiketlenmesi ile ilgilenildiği zaman doğruluk, ilişki kuralları ve örüntü özellikleri gibi farklı yöntemler kullanılarak bu probleme yaklaşılmıştır. Ayrıca klinik denemeler, veri temizleme, hava tahminleri, coğrafik bilgi sistemleri, bilgisayar ağ girişleri, kredi kartı dolandırıcılığı, medikal ve biyoinformatik deneyler olmak üzere çok çeşitli uygulama sahaları için de aykırı gözlem tespit metotları önerilmiştir. Dolayısıyla bu probleme yaklaşmak için kullanılan bilgi ve uygulanan verinin yapısı, aykırı gözlem tespit etme metotlarının sınıflanabilmesine imkan sağlamaktadır.

Tarihsel gelişimi içinde bu sahada yapılan ilk çalışmaların tek değişkenli veri setleri için önerilen yöntemler olduğu ve şuan günümüzde en çok çalışılan sahanın çok değişkenli metotlar olduğu görülmektedir. Burada da kullanılan yöntemin parametrik veya parametrik olmayan bir yöntem olmasına göre ayrım yapılabilir. Bilindiği üzere, istatistiksel parametrik metotlar, gözlemlerin bilinen bir dağılımdan geldiğini farzeder ve dağılımın bilinmeyen parametrelerinin istatistiksel tahminine dayanır. Aynı zamanda yoğunluk tabanlı yaklaşımlar (density based) olan parametrik metotlar, farzedilen dağılımdan ayrı davranış gösteren gözlemleri aykırı gözlem olarak tespit eder. Ancak bu metotlar, veri dağılımı hakkında önsel bilginin olmadığı durumlar ve yüksek boyutlu veri setleri için genellikle uygun değildir. Verinin dağılımı ile ilgili bilgi kullanmayan parametrik olmayan metotlar için uzaklık tabanlı yöntemler (distance based), derinlik tabanlı yaklaşımlar (depth based) ve kümeleme tabanlı (cluster based) teknikler örnek gösterilebilir. Uzaklık tabanlı yöntemlerde, bir uzaklık ölçüsüne göre (ki sıklıkla Mahalanobis ölçüsü kullanılır) verinin tüm gözlem noktalarının bir uzaklık ölçüsü hesaplanır ve en yüksek ölçüye sahip olan noktalar aykırı gözlem olarak tespit edilir. Derinlik tabanlı yaklaşımlarda ise veri noktalarına bir derinlik atanır ve aykırı

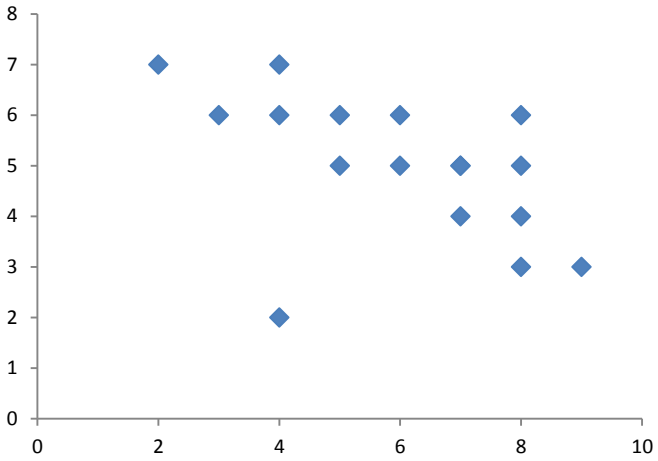
gözlemlerin en az derinliğe sahip oldukları düşünülür. Kümeleme tabanlı tekniklerde ise herhangi bir kümeye atanamayan gözlemin aykırı gözlem olabileceği veya küçük çaplı bir kümenin, kümelenmiş aykırı gözlemlerden oluşmuş olabileceği düşünülür. Bunların haricinde bir gözlemin komşuluğunda bulunan diğer gözlemlerin davranışına göre aykırı gözlem olarak tespit edilmesi gibi farklı bir yaklaşım da söz konusudur. Bu durumlarda, bir gözlemin komşuluğundaki diğer gözlemlere göre lokal durağansızlığı dikkate alınır ve bu gözlemin, tüm anakütle dikkate alındığında önemli bir şekilde aykırı bir davranış sergilemese bile, komşuluğundaki gözlemler açısından durağansızlık veya aşırılık sergilediği için aykırı gözlem olabileceği düşünülür [Ben Gal I., 2005; Escalante H.J., 2005].

Elbetteki burada bahsi geçen temel yaklaşımlardan farklı olarak çok çeşitli aykırı gözlem tespit metotları da önerilmiştir. Bu anlamda istatistik literatürü incelendiğinde, çok değişkenli metotlara nazaran tek değişkenli durumlar için geniş bir literatüre ulaşmak mümkündür. Özellikle Hawkins (1980) , Davies ve Gather (1993), Barnett ve Lewis (1994) konuyla alakalı belli başlı temel kaynaklar arasındadır. Son zamanlarda ortaya çıkan farklı veri setlerinin yapısı dikkate alındığında, yapay sinir ağları veya destek vektör makineleri gibi istatistiksel olmayan çeşitli makine öğrenme algoritmalarının kullanıldığı farklı yaklaşımlara da rastlamak mümkündür. Daha güncel ve geniş kapsamlı bir kaynak olarak, Chandola V. ve ark (2009) önerilebilir.

Bölümün ilerleyen başlıklarında, çok değişkenli durumlar için genel olarak kullanılan uzaklık tabanlı aykırı gözlem tespit metotları anlatıldıktan sonra yüksek boyutlu veri setleri için karşılaşılan zorluklardan bahsedilecektir. Bu aşamada bilgi kriterleri ile önerilen yaklaşımın detayları verilecektir.

2.5.1 Çok Değişkenli Veri Setlerinde Aykırı Gözlem Tespiti (Multivariate Outlier Detection)

Çok değişkenli gözlemlerin olduğu veri setlerinde aykırı gözlemler, her bir değişken bağımsız bir şekilde düşünülerek tespit edilemez. Aykırı gözlemler sadece çok değişkenli analiz uygulandığı zaman ve farklı değişkenler arasındaki ilişkiler veri sınıfı içinde karşılaştırıldığı zaman tespit edilebilir. Basit bir örnek, iki boyutlu uzayda iki ölçüye sahip olan veri noktaları için Şekil 2.4'de görülmektedir.



Şekil 2. 4: İki boyutlu uzayda iki ölçüye sahip olan veri noktaları.

Sol alt köşede olan noktanın, tek değişkenli durum için aykırı gözlem olmadığı ancak, çok değişkenli durum için bir aykırı gözlem olduğu açıktır. Eksenler boyunca, değerlerin yayılımına göre bir ölçü düşünüldüğü zaman, gözlemin tek değişkenli dağılımın merkezine yakın düştüğü görülebilmektedir. Bu yüzden çok değişkenli durumlar için, değişkenler arasındaki ilişkileri dikkate alan bir ölçünün kullanılması zorunludur.

İstatistiksel metotlar genellikle, veri dağılımın merkezine göre uzak noktaya yerleşmiş olan gözlemleri aykırı gözlem olarak tespit eder. Bu amaç için kullanılan bazı uzaklık ölçüleri vardır. Genel olarak kullanılan uzaklık kriteri, $X_i = X_1, X_2, \dots, X_p$ p-boyutlu uzayda n-gözleme sahip veri seti ele alındığı zaman,

$$MD_i \text{ veya } RD_i = \sqrt{(X_i - T(X))^T C(X)^{-1} (X_i - T(X))} \qquad (2.101)$$

ile tanımlı uzaklık ölçüsüdür. $T(X)$; X veri kümesinin ortalama-konum vektörü ve $C(X)$; kovaryans matrisidir. (2.101) ile verilen eşitlik, $T(X)$ için değişkenlerin aritmetik ortalamalarından oluşan ortalama vektörü ve $C(X)$ için klasik örnek varyans-kovaryans matrisi kullanıldığı zaman Mahalanobis Uzaklığı (Mahalanobis Distance-MD) olarak, bunlar için sağlam tahmin ediciler kullanıldığı zaman ise Sağlam Uzaklık (Robust Distance-RD) olarak ifade edilir.

Ayrıca çoklu aykırı gözlemlere veya aykırı gözlemlerden oluşan kümeye sahip olan veri setleri, maskeleme (masking) ve baskılama (swamping) etkilerine tabidir. Matematiksel olarak bir kesinliği olmamasına rağmen Acuna ve Rodrigez (2004)'de verdikleri tanımlar,

bu etkileri sezgisel olarak anlamaya yardımcı olabilmektedir. Diğer tanımlar için, Hawkins (1980), Davies ve Gather (1993), Barnet ve Lewis (1994) kaynaklarına bakılabilir.

Maskeleme: Bir aykırı gözlemin, ikinci bir aykırı gözlemin varlığından dolayı maskelenmesidir. Yani bir aykırı gözlemin, ancak diğer aykırı gözlem silindikten sonra ortaya çıkmasıdır.

Baskılama: Bir aykırı gözlemin, ikinci bir aykırı gözlemin varlığından dolayı baskılanmasıdır. Diğer bir ifade ile, bir aykırı gözlemin diğer aykırı gözlem silindikten sonra, artık aykırı bir gözlem olmaması durumudur.

Bu etkiler, bir grup aykırı gözlemin ya da aykırı gözlemlerden oluşan bir kümenin, ortalama ve kovaryans tahminlerini kendilerine doğru çekmesinden dolayı tahmini saptırmasından kaynaklanabilir. Bu saptırma neticesinde, maskeleme için, aykırı olan gözlemlerin ortalamadan uzaklıklarının küçükmüş gibi görünmesi ve onların aykırı gözlem olarak algılanmaması sonucunu ortaya çıkarır. Bu nedenle aykırı gözlem maskelenir. Baskılamada ise, bu saptırma, aykırı olmayan gözlemlerin ortalamadan uzaklıklarının büyükmüş gibi görünmesi ve onların aykırı gözlem olarak algılanması sonucunu ortaya çıkarır. Yani aykırı gözlem baskılanır. Dolayısıyla, sadece ortalama ve kovaryans matrisi hesabına dayanan uzaklık ölçüleri ile aykırı gözlem tespit metotları, baskılama ve maskeleme etkilerine tabidir [Ben Gal I., 2005].

Mahalanobis uzaklığı, çok değişkenli veri setlerinde aykırı gözlem tespiti için kullanılan ve çok iyi bilinen bir kriterdir. Ancak bu kriter, veri setinde bulunan muhtemel aykırı gözlemlerin varlığı altında hesaplandığı için ve bu gözlemler örnek kovaryans matrisi ile ortalama vektörünü etkiledikleri için sınırlamalara sahiptir. Ayrıca bu kriter yukarıdaki paragrafda tanımlanan maskeleme ve baskılama etkilerine tabidir. Hadi (1992)'de bu problemlerin etkisini azaltmak için iteratif olarak hesaplanan bir Mahalanobis ölçüsü önermiştir. Bu çalışmasında ortalama vektörü yerine değişken medyanlarının bir vektörünü koymayı, kovaryans matrisi için ise en küçük Mahalanobis uzaklığına sahip gözlemlerden oluşan kümeye ait olan kovaryans matrisinin kullanılmasını önermiştir. Hadi prosedürünün modifiye edilmiş bir hali ise Penny ve Joliffee tarafından (2001)'de önerilmiştir. Penny ve Joliffee kendi önerdikleri bu yaklaşımı ve Hadi prosedürü ile beraber bu konuda önerilen başka bazı yöntemlerin başarısını da karşılaştırmalı olarak analiz etmişlerdir.

Aykırı gözlem varlığında veri analizinin performansını artırmak için çok değişkenli dağılım parametreleri olan ortalama vektörü ve kovaryans matrisi için sağlam tahminlerin kullanımı da önerilmiştir. Bu amaçla literatürde $T(X)$ ve $C(X)$'in farklı tanımlamaları kullanılarak farklı RD ölçüleri elde edilmiştir. Minimum Hacimli Elipsoid (Minumum Volume Elipsoid -MVE), Rousseeuw tarafından (1983)'de önerilen ilk yüksek kırılma noktasına sahip olan sağlam tahmin edicidir. Hesaplama ve algoritmasının kullanışlı olması ve ortalama vektörü ve kovaryans matrisi için MVE'in yüksek dayanıklılığı nedeniyle aykırı gözlem tespit etmek için kullanışlı bir araçtır. MVE algoritmasında n gözlem içinden h birimli bir alt küme seçilir. Bu küme için ortalama vektörü ve kovaryans matrisi için klasik μ , Σ parametreleri hesaplanır. Burada Σ'nın pozitif tanımlı olması beklenir. Daha sonra μ , Σ kullanılarak bir elipsoid tanımlanır. Algoritma minimum hacimli elipsoidi elde edene kadar farklı alt kümelerin denenmesi şeklinde devam eder.

Minimum Kovaryans Determinant yöntemi (Minumum Covariance Determinant-MCD), her ne kadar Rousseeuw tarafından (1984)'de önerilen bir yöntem olsa da, yöntemin yaygın bir şekilde kullanımı Rousseeuw ve van Driessen'in (1999) FAST-MCD yöntemini önermesinden sonra gerçekleşmiştir. MVE yöntemine benzer bir yöntemle h (en az (n+p-1)/2 olmalı) boyutlu bir kümenin seçilmesi ile algoritma başlar. Algoritma yeni kümeyi belirlerken Mahalanobis ölçüsünü kullanır. En düşük uzaklığa sahip olan noktalar kümeye seçilir ve buradaki amaç en küçük kovaryans determinantına sahip olan kümeyi belirlemektir. Bu yöntemlerin h boyutlu küme için uygun örnekleme yöntemine ve h'ın uygun seçimine bağlı olduğu aşikardır. Caussinus ve Roiz'in (1990)'da gözlemlerin veri merkezine olan uzaklıklarına göre ağırlıklandırılmasına dayanan sağlam kovaryans tahmini ise önerilen bir başka yaklaşımdır. Billor, Hadi ve Velleman'ın (2000) BACON yaklaşımı ise bu gruptaki çalışmalara verilebilecek diğer bir örnektir. BACON yönteminde, gözlemlerin çok değişkenli eliptik dağılımdan geldiği varsayılarak Mahalanobis uzaklığından yararlanılmakta, kritik değer olarak da düzeltilmiş ki-kare değeri kullanılmaktadır. Gözlemlerin bloklanması nedeniyle hesapsal açıdan etkin bir yöntemdir. Diğer yöntemlere göre bu yöntemdeki iterasyon sayısı daha azdır.

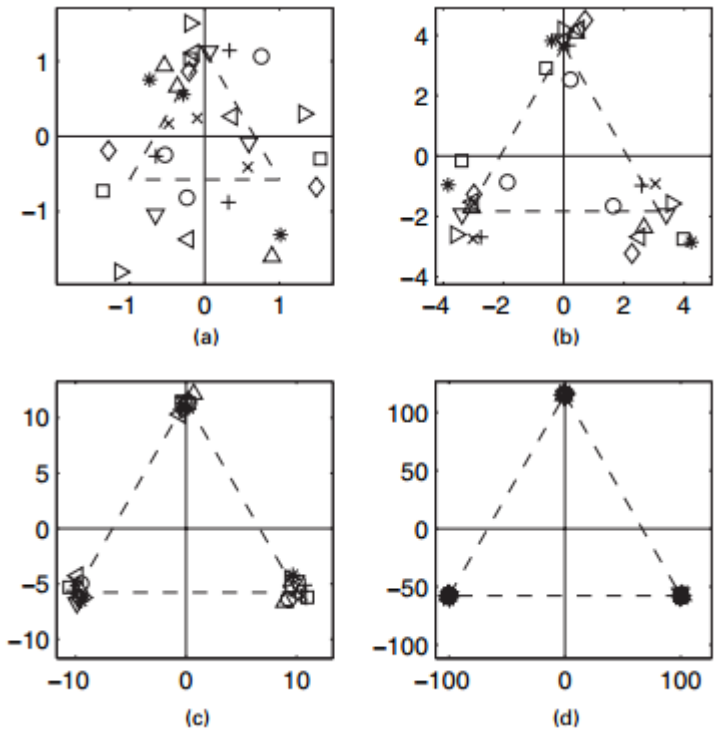
Buraya kadar anlatılanlardan anlaşılacağı üzere uzaklık tabanlı yaklaşımlar için ilk amaç $T(X)$ ve $C(X)$'in uygun bir tahminini bulabilmektir. Daha sonra denklem (2.101) ile hesaplanan uzaklıkların dağılım bilgisi gerekmektedir. Eğer X_i normal dağılıma sahip ise örnek ortalaması ve kovaryans matrisine dayanan klasik Mahalanobis uzaklığının karesi,

asimptotik olarak p serbestlik dereceli Ki-kare dağılımına sahiptir ve $MD_i^2 > \chi_{p,0.975}$ olan değerler aykırı gözlem olarak belirlenir. [Johnson ve Wichern, 1998]. Eğer aykırı gözlemlerin olmadığı normal dağılan büyük bir veri setine $T(X)$ ve $C(X)$ için sağlam tahmin ediciler uygulanmış ise Sağlam uzaklığın karesi F-dağılımına sahiptir. [Hardin ve Rocke,2005]. Bununla birlikte normal dağılmayan veri setleri için aykırı gözlemleri tespit edecek bir sınırın nasıl belirleneceği açık değildir. [Ben Gal I, 2005; Filzmoser ve ark., 2008; Aelst S.V. ve Rousseeuw P.J., 2009; Hubert M. ve Debruyne M., 2009].

2.5.2 Yüksek Boyutlu Veri Setlerinde Karşılaşılan Zorluklar

Çoklu aykırı gözlemleri tespit etmek için önerilen tüm yöntemler, geçmişten beri veri boyutunun artışıyla daha da tırmanan bir hesaplama yükünden muzdariptirler. Ayrıca, genel aykırı gözlem metotları boyutsallık probleminden (curse of dimensionality) dolayı yüksek boyutlu verilerde etkili bir şekilde çalışamazlar. Bu temel olarak yüksek boyutlu veri setlerinin doğasında var olan aşağıdaki probemlerden kaynaklanır.

Uzaklık ve yoğunluk tabanlı olan genel teknikler, tüm boyutları dikkate alarak çalıştıkları için gürültü ve ilgisiz değişkenlerin varlığından etkilenirler. Yüksek boyutlarda en yakın ve en uzak noktalar benzer olduğu için, uzaklık veya en yakın komşuluk kavramları anlamsız olur. Bu durum Şekil 2.5'de gösterilmektedir [Kaynak: Hall P., Marron J.S., Neeman A., 2005].



Şekil 2. 5: n=3 için p=2, 20, 200, 20000 boyutlarında veri noktalarının saçılım grafiği.

Yukarıdaki şekilde (a)-(b)-(c)-(d) grafikleri n=3 gözlemin sırasıyla p=2, p=20, p=200 ve p=20000 boyutlar için saçılımını temsil etmektedirler. Boyutlar arttıkça gözlem noktalarının ayrımı imkansız hale gelmektedir. Bu durumun matematiksel ve geometrik açıklamaları için Hall ve ark. (2005) kaynağı iyi bir referans özelliği taşımaktadır. Yüksek boyutlu veri setlerinde gözlem noktaları seyrek (sparse) oldukları için tüm gözlem çiftleri neredeyse eşit uzaklıklara sahiptirler. Bu nedenle seyrek yüksek boyutlu veriler için anlamlı bir uzaklık ve yoğunluk kavramına ihtiyaç vardır. Eğer bu yoksa, uzaklık veya en yakın komşuluk bağlamında aykırı gözlem bulmak olası değildir.

Yüksek boyutlu veri setlerinde klasik kovaryans matrisi tam ranklı olmadığı için tersi alınabilir değildir. Bu özellikle yakınsama için birçok iterasyona gerek duyan iteratif metotlarda, her iterasyonda kovaryans matrisinin tersi hesaplandığı için yöntemlerin hesaplama yükünü artıran bir durumdur. Bu nedenle algoritmaların hesaplama zamanları n gözlem sayısının artışıyla lineer olarak, p değişken sayısının artışıyla kübik olarak artmaktadır. [Filzmoser ve ark., 2008]

2.5.3 Bilgi Kriterleri ile Aykırı Gözlem Tespiti

Veri setlerinin homojen olduklarını varsaymak alışılagelmiş bir durum olmasına rağmen, aslında veri setleri sıklıkla aykırı gözlemler veya alt gruplar ihtiva eden homojen olmayan bir yapıya sahiplerdir. Bu nedenle çoklu aykırı gözlemleri ve alt grupları tanımlamak için kullanılan metotlar, homojensizlik yüzünden bozulmayacak bir metrik kurmanın zorluğu ile ilişkilidir. Sağlam metotlar, homojenlik varsayımında esneme yapabilir. Fakat bu yöntemler de yaygın bir şekilde kabul görmemiştir. Bu kısmen, bu metotların aykırı gözlem tespitini, tahmin metodunun kara kutusu içinde saklı tutmalarından kaynaklanabilir. Ancak temel olarak bu, yöntemlerin yüksek boyutlardaki veri setleri için hesaplama açısından uygun olmamalarından kaynaklanır [Billor ve ark., 2000]. Oysaki bilgi kriterlerinin kullanımı ile böyle bir metrik kurmanın zorluğu ortadan kalkabilir. Mantıksal olarak özellikle verideki kovaryans yapısının kompleksliğini cezalandıran ICOMP tipi kriterler, zaten homojensizlik ölçümü yaptıkları için bu problemin üstesinden gelebilir.

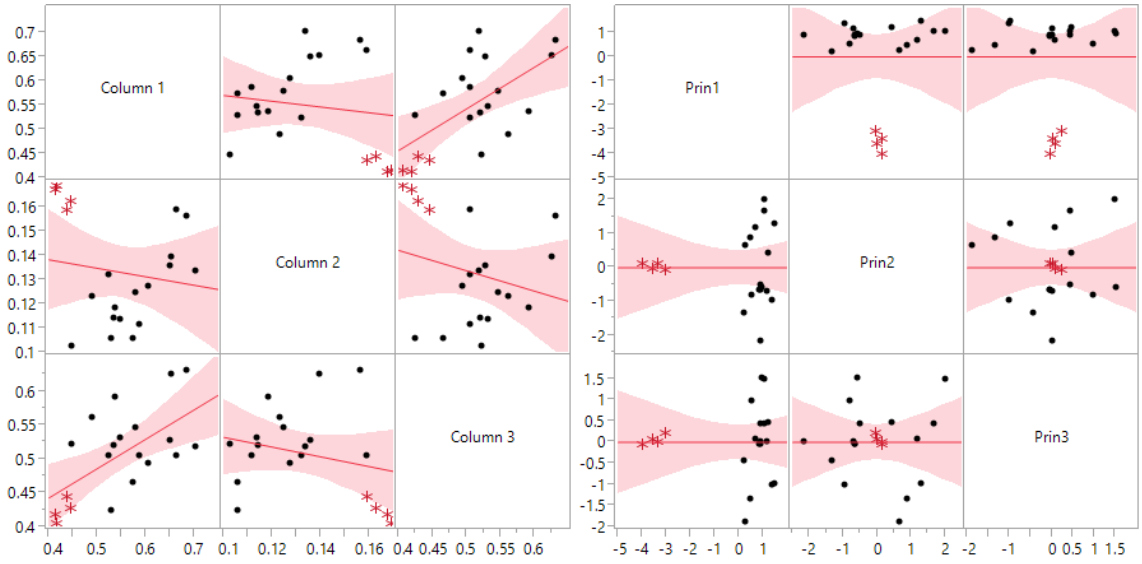
Bu amaçla bilgi kriterleinin aykırı gözlem tespitinde kullanılabilmesi için, veri setinde her bir gözlemin tek tek silinmesini ve geride kalan noktalar için bilgi kriterlerinin hesabını gerektiren bir yaklaşım öneriyoruz. Daha sonra bu hesapların kullanımıyla örnek olarak ICOMP kriteri için

$$ICOMP_r = \frac{ICOMP_i}{ICOMP_{tümdata}} \tag{2.102}$$

oranı hesaplanır. Burada $ICOMP_i$; i. Gözlemin veriden silinmesiyle geri kalan noktalar için hesaplanan değeri göstermektedir. $ICOMP_{tümdata}$; tüm gözlem noktalarının varlığında hesaplanan değer olup, $ICOMP_r$; ratio-oran kelimesine ithafen elde edilen oranı temsil emektedir. $ICOMP_r$'nin dağılımı bilinmediği için, bir üst ve alt limit tanımlanarak bu sınırlar dışında kalan gözlemler aykırı gözlem olarak tanımlanmaktadır. Gözlemler silinerek hesaplama yapmanın arkasında yatan düşünce, aykırı gözlemin veriden çıkarılmasıyla $ICOMP_i$ değerinin bundan çok etkilenecek olmasıdır. Bu değer, $ICOMP_{tümdata}$ ile oranlanarak bu değişim tüm veri setinin yapısıyla mukayese edilmekte, sonuç olarak aşırı değişimin olduğu gözlem değerleri tespit edilmektedir.

2.6 Hibrit Aykırı Gözlem Tespiti (Hybrid Outlier Detection-HOD)

Buraya kadar, literatürde aykırı gözlem tespiti için kullanılan en iyi bilinen yöntemler tanıtılmış ve özellikle yüksek boyutlu veri setlerinde karşılacak zorluklardan bahsedilmiştir. Tez kapsamında ilgilenilen bu problem için alternatif bir yöntem önerilmiştir. Temel bileşenler analizi PCA boyut indirgemede kullanılan bir yöntemdir ancak, yüksek boyutlu aykırı gözlemleri tanımlamak için alternatif bir yaklaşım olarak önerilebilir. PCA analizi sonucunda elde edilen temel bileşenler uzayında, aykırı gözlemler özellikle önemli bileşenler üzerinde çok belirgin bir şekilde görünürdür. Aykırı gözlemler varyansı kendi yönlerinde artırdıkları için, sezgisel olarak temel bileşenler uzayında, orijinal veri uzayına göre daha görünür olacaklardır. En azından maksimum varyansa sahip olan yönlerde muhtemel olarak aykırı gözlemler daha çok ortaya çıkacaklardır. Filzmoser ve ark. (2008)'de dikkat çektikleri bu durum modifiye edilmiş Wood Gravity veri seti üzerinde incelenmiştir. Sonuçlar Şekil 2.6'daki gibidir.



(a) (b)

Şekil 2. 6: Wood Gravity veri setinin aykırı gözlemlerinin orijinal boyutlar ve temel bileşenler üzerindeki görünümü

6 değişkenli 20 gözleme sahip bu özel veri seti Draper ve Smith (1966)'da önerilmiştir. Bu veri setinde 4, 6, 8 ve 19 nolu gözlemlerin aykırı gözlem oldukları bilinmektedir. Veri setine Rousseeuw P.J ve Leray A.M (2003)'ın "Robust Regression and Outlier Detection" kitabının 243. Sayfasından ulaşılabilir. Şekil 2.6'dan da anlaşılacağı üzere aykırı gözlemleri

temel bileşenler uzayında aramak, orijinal veri uzayında aramaktan daha kötü sonuçlar vermeyecektir.

PCA analizinin aykırı gözlem tespiti için sağladığı bu avantaj, açıkça bellidir ki yüksek boyutlu veri setlerinde PCA'nin doğru uygulanabilir olmasına bağlıdır. Bildiğimiz kadarıyla tez kapsamında önerilen HDR yaklaşımına kadar literatürde bu problem çözülebilmiş değildir. Bölüm 3.2'de gösterileceği gibi boyut indirgeme doğru bir şekilde uygulanabildikten sonra, yüksek boyutlu verinin boyutu HDR ile indirgenecek ve temel bileşenler üzerinde aykırı gözlem tespiti yapılabilecektir. Bundan sonra bu yöntem Hibrit Aykırı Gözlem Tespiti-Hybrid Outlier Detection (HOD) ile adlandırılacaktır. Özetle HOD şu adımlardan oluşmaktadır.

Aşama-1: HDR ile yüksek boyutlu verinin boyutu indirgenir

Aşama-2: HDR sonucu elde edilen yeni bileşenler üzerinde her bir gözlem tek tek dışarıda bırakılarak $ICOMP_r$ oranları hesaplanır.

Aşama-3: $ICOMP_r$'nin değerleri 0 ve 1 arasında olduğu için Beta dağılımından geldiği varsayılarak, bir üst ve alt limit tanımlanır ve bu sınırlar dışında kalan gözlemler aykırı gözlem olarak tanımlanır.

HOD yönteminin başarısı ve uygulamadaki performansı Bölüm 3.3'de yapılacak simülasyon çalışmaları ile türetilmiş veri setleri üzerinde gösterildikten sonra, Bölüm 3.4'de, HDR ile boyutları indirgenmiş altı tane kanser gen ifade verisi üzerinde gösterilecektir.

2.7 Sınıflandırma Prosedürleri

Uygulama olarak kullanılan gen ifade verileri, sağlam/hasta dokulardan veya farklı tümör dokularından alınan örnekler ile oluşturulduğu için farklı alt gruplara sahip olan veri setleridir. Dolayısıyla önerilen yöntem ile veri setinde var olan bilginin ortaya çıkarılıp çıkarılmadığı, ancak boyut indirgeme yapıldıktan sonra yeni transform edilmiş veri setini sınıflandırma prosedürlerinde giriş olarak düşünmek ve hatalı sınıflama oranlarını hesaplamakla mümkün olacaktır. Bu nedenle tezin bu bölümü, kullanılan sınıflandırma prosedürleri olan Lineer Diskriminant Analizi (Linear Discriminant Analysis-LDA) ve Kuadratik Diskriminant Analizi (Quadratic Diskriminant Analysis-QDA) yöntemlerinin tanıtılmasına ayırılmıştır.

Farklı sınıflardan veya gruplardan örnekler alınarak oluşturulan veri setlerinin analizinde en önemli amaçlardan biri, veri setinde var olan bilgi ile grup yapısının anlaşılması neticesinde gözlemlerin doğru gruplara veya sınıflara atanmasını sağlamaktır. Bu noktada, bir gözlemin hangi sınıfa ait olduğunun belirlemek için bir istatistiksel karar verme sınırının kullanılması gerekir.

Hatalı sınıflama riski, karar verme sürecinin veya sınıflayıcının performasının genel bir ölçüsüdür. Bu ölçü, sınıflanan örnek üzerinde beklenen hatalı sınıflama kaybı olarak tanımlanabilir. Herhangi bir hatayı kayıp, doğru kararı kayıp değil şeklinde atayan simetrik fonksiyon veya 0/1 kayıp fonksiyonu olarak düşünmek alışılagelmiştir. [Duda ve ark, 2001]. Bu nedenle karar verme sürecini maksimum doğru sınıflama veya buna denk olarak minumum hatalı sınıflama oranları ile değerlendirmek yaygındır.

Uygun istatistiksel karar verme sınırlarını tanımlamak için kullanılan birçok sınıflama kuralı vardır. Bir gözlemi en yüksek sonsal olasılıklı (posterior probability) sınıfa atayan Bayes seçim kuralı, tüm muhtemel kurallar arasında minimal hatalı sınıflama riskini başaranlardan ve en çok bilinenlerden biridir. [Anderson, 1984].

2.7.1 Bayes Seçim Kuralı

Bayes karar kuralının arkasındaki düşünce şudur: Grup üyeliği hakkındaki ulaşılabilen bilginin tamamı, şartlı (veya sonsal) olasılıkların bir dizisini içerir. Simetrik ve 0/1 şeklindeki kayıp fonksiyonunun olduğu durumlarda riski minimize eden Bayes kuralı aşağıdaki gibi tanımlanır.

$$P(\pi_i \mid x) > P(\pi_j \mid x) \tag{2.103}$$

Yani, $j \neq i$ ve $i, j = 1, 2, ..., g$ gruplar içinde bir girdi vektörü olan x'in π_i sınıfında olması için şart budur. x'in π_i grubunda bulunma olasılığı, π_j grubunda bulunma olasılığından fazla ise x, π_i grubuna atanır. Eğer en yüksek olasılığa sahip olan birden fazla grup varsa, o zaman bağlı gruplardan birine rasgele nesne tahsis edilmesiyle bu bağ kırılabilir [James, 1985].

Standart tahmin metotlarını kullanarak, $P(\pi_i \mid x)$ gibi olasılıkları hesaplamak zordur. $P(x \mid \pi_i)$ gibi olasılıklar için ise bu böyle değildir. π_i nesne sınıfından geldiği bilinen x girdi vektörü için bir dizi ölçü elde etme olasılığı, yani şartlı sınıf olasılığı $P(x \mid \pi_i)$ veya olabilirlik bilgisi, π_i sınıfından bir örüntü örneği alarak basitçe tahmin edilebilir. Genel olarak $P(\pi_i \mid x)$ ve $P(x \mid \pi_i)$ arasındaki ilişki Bayes teoremi olarak bilinir.

$$P(\pi_i \mid x) = \frac{P(x \mid \pi_i)P(\pi_i)}{\sum_{tümk} P(x \mid \pi_k)P(\pi_k)} \tag{2.104}$$

Burada, $k = 1, 2, ..., g$ gruplardır. Eşitlik (2.104)'ün sağ tarafındaki tüm terimler sayısal olarak hesaplanabilirdir ve bu yüzden örnekleme yapılarak bulunabilir. Ön olasılık olarak tanımlanan $P(\pi_i)$ terimi, herhangi bir bilgi varlığında π_i sınıfından gelen örüntü olasılığıdır. Yani tüm popülasyon içinde π_i sınıfının oranıdır. (2.103)'de verilen Bayes kuralı, (2.104)'de verilen Bayes teoremine uygulanırsa aşağıdaki kural verilebilir.

- $j \neq i$ ve $k = 1, 2, ..., g$ için x'in π_i sınıfına atanması

$$\frac{P(x \mid \pi_i)P(\pi_i)}{\sum_{tümk} P(x \mid \pi_k)P(\pi_k)} > \frac{P(x \mid \pi_j)P(\pi_j)}{\sum_{tümk} P(x \mid \pi_k)P(\pi_k)} \tag{2.105}$$

kuralına göre olur. Eşitsizliğin her iki tarafında paydalar bir birine eşittir. Bu nedenle Bayes kuralı aşağıdaki gibi rahatça yazılabilir.

- x'in π_i sınıfına atanması için

$$P(x \mid \pi_i)P(\pi_i) = \max_{1 \le j \le g} P(x \mid \pi_j)P(\pi_j) \tag{2.106}$$

şartı sağlanmalıdır. (2.106) ile tanımlanan sınıflama kuralı Optimal Bayes Karar Kuralı'nın son pratik formudur.

Bir sınıflayıcı dizayn etmek için bazı metotlar kullanılabilir. En uygun metodu seçmek için kullanılan stratejiler temel olarak şartlı sınıf olasılık yoğunluğu hakkındaki ulaşılabilen bilginin miktarına ve çeşidine bağlıdır. Yukarıdaki paragraflarda tanıtılan optimal bayes kuralı, şartlı sınıf yoğunluklarının tamamı özelleştirildiği zaman bir sınıflayıcı dizayn etmek

için kullanılabilir. Bununla birlikte, gerçek şartlı sınıf yoğunluğu bilinmez ve ulaşılabilen eğitim veya örneklem setinden tahmin edilmek zorunda olunabilir. Eğer en azından şartlı sınıf yoğunluklarının yapısı bilinirse (çok değişkenli normal dağılım gibi) fakat parametreler bilinmezse (ortalama vektör ve kovaryans matrisi gibi) bu problem bir parametrik karar problemi olarak adlandırılır. Bu problem ile mücadele etmek için yaygın bir strateji, eğitim veya örneklem setinden hesaplanan değerleri, onların tahmini olarak yoğunluk fonksiyonlarındaki bilinmeyen parametreler yerine koymaktır.

Şartlı sınıf yoğunluklarının formu bilinmez veya varsayımda bulunulamazsa, parametrik olmayan yöntemler düşünülmek zorundadır. Parametrik olmayan problemlerde, hem yoğunluk fonksiyonu kernel fonksiyonları kullanılarak tahmin edilmek zorundadır, hem de sınıf karar sınırı, ulaşılabilen eğitim örmeklerine dayalı olarak oluşturulmak zorundadır.

Daha önce de belirtildiği gibi, uygun bir istatistiksel sınıflama metodu seçmek için diğer bir ince nokta, ulaşılabilen bilginin miktarı ile ilgilidir. Bir problemin boyutsallığında veya özellik sayısında artış olduğu zaman (daha fazla detay) sezgisel olarak sınıflama performansının artması beklenirken, uygulamada gösterilmiştir ki bir noktadan sonra eğer ulaşılabilen ilave bilgi yoksa (daha fazla örnek) bunun tam tersi gerçekleşir. Bu yaygın bir paradokstur ve iyi bilinen bu davranış, boyutsallık problemi (curse of dimensionality) olarak adlandırılır. [Fukunaga, 1997; Jain ve ark., 2000; Bellman, 1961] Örneklem boyutu sınırlandırılıp sabitlendiğinde özelliklerin sayısı artmaya devam ederse, parametre tahminlerinin güvenilirliği azalır ve buna karşılık sınıflandırıcının performansı da düşer.

2.7.2 Bayes Eklentili Sınıflayıcılar (Bayes Plug-in Classifiers)

İstatistiksel örüntü tanıma sistemleri için kullanılan en yaygın parametrik yöntemlerden biridir. Bu sınıflayıcı, her bir grup veya sınıfın gerçek kovaryans matrisinin tersini içeren ölçünün benzerliğine dayanır. Uygulamada bu matrisler bilinmediği için, tahminler bir eğitim veya örneklem setinde ulaşılabilir bilgiye dayalı olarak hesaplanmak zorundadır. Gerçek kovaryans matrisini tahmin etmek için genel seçim, karşılık gelen örnek grup kovaryans matrisi tarafından tanımlanan maksimum olabilirlik tahmin edicisidir. Bununla birlikte örneklem büyüklüğünün sınırlı olduğu uygulamalarda, örneklem grup kovaryans tahmini yüksek derecede değişken olur ve hatta tersi bile alınamaz. Bu yüzden, hatırı sayılır bir çaba, özellikle sınırlı örneklem büyüklüğüne sahip yüksek boyutlu problemlerde diğer

konveksiyonel olmayan Bayes eklentili sınıflayıcıları geliştirmek için harcanmıştır. Bu bölümde konvensiyonel Bayes eklentili sınıflayıcılar Kuadratik ve Lineer Diskriminant Analizleri tanıtılacaktır.

2.7.2.1 Kuadratik Diskriminant Sınıflayıcısı

Bayes eklentili sınıflayıcılar aynı zamanda, Gauss maksimum olabilirlik sınıflayıcısı olarak da bilinir, çok değişkenli Normal veya Gauss sınıf yoğunluklarına dayanır.

$$P(x \mid \pi_i) = f_i(x \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \qquad (2.107)$$

Burada, μ_i ve Σ_i gerçek π_i popülasyon sınıfının ortalama vektörü ve kovaryans matrisidir. Denklem (2.107)'de tanımlanan $f_i(x \mid \mu_i, \Sigma_i)$ şartlı sınıf olasılık yoğunluğu aynı zamanda olabilirlik yoğunluk fonksiyonları olarak da bilinir. Daha önceki bölümde tanımlanan optimum Bayes kuralı içine denklem (2.107) yerleştirilirse aşağıdaki Bayes sınıflama kuralı elde edilir.

- x'in π_i sınıfına atanması için

$$f_i(x \mid \mu_i, \Sigma_i)P(\pi_i) = \max_{1 \le j \le g} f_j(x \mid \mu_j, \Sigma_j)P(\pi_j) \qquad (2.108)$$

şartı sağlanmalıdır. Tekrar hatırlatmak gerekirse, eşitlik (2.108)'de, $P(\pi_i)$; grup ile ilişkili önsel olasılık, g grup veya sınıf sayısıdır. (2.108) denklemini özelleştirmenin bir diğer yolu onun doğal logaritmasını almaktır.

$$\begin{aligned} d_i(x) &= \ln\left[f_i(x \mid \mu_i, \Sigma_i)P(\pi_i)\right] \\ &= \ln\left[\frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} P(\pi_i)\right] \\ &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln \mid \Sigma_i \mid -\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln P(\pi_i) \end{aligned} \qquad (2.109)$$

Denklem (2.109) ile elde edilen d_i sıklıkla *i.* sınıf için kuadratik diskriminant ölçüsü olarak adlandırılır. $\frac{n}{2}\ln(2\pi)$ tüm sınıflarda aynı olduğu için ihmal edilebilir. Buradan denklem

(2.106) ile tanımlanan optimal Bayes sınıflama kuralı daha basitleştirilirse, aşağıdaki kural elde edilir.

- x'in π_i sınıfına atanması için

$$\begin{aligned} d_i(x) &= \max_{1\le j\le g}\left[-\frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j) + \ln P(\pi_j)\right] \\ &= \min_{1\le j\le g}\left[\ln|\Sigma_j| + (x-\mu_j)'\Sigma_j^{-1}(x-\mu_j) - 2\ln P(\pi_j)\right] \\ &= \min_{1\le j\le g} d_j^*(x) \end{aligned} \tag{2.110}$$

sağlanmalıdır. Denklem (2.110)'da özelleştirilen Bayes sınıf kuralı, Kuadratik Diskriminant Kuralı olarak bilinir. $P(\pi_j)$ önsel olasılık bilgisi yoksa, $d_j^*(x)$ ölçüsü; x ve μ_j arasındaki genelleştirilmiş uzaklık ölçüsü olarak tercih edilir. İlk terim, j. grubun genelleştirilmiş varyansı ile ilgilidir ve ikinci terim, j. grubun ortalama vektörü ile x arasındaki Mahalanobis uzaklığıdır.

Bununla birlikte uygulamada, kovaryans matrisi ve ortalamanın yani μ_i ve Σ_i'nin değeri nadiren bilinir ve ulaşılabilen eğitim örneklerinden hesaplanarak yerine konulması gerekir. Bayes eklentili sınıflayıcıların eklenti (plug-in) terimiyle isimlendirilmesinin nedeni budur. Ortalama, μ_i'nin maksimum olabilirlik tahmin edicisi olan $\bar{x}_i$ örnek ortalaması kullanılarak tahmin edilir [Anderson,1984]. Yani

$$\mu_i \equiv \bar{x}_i = \frac{1}{n_i}\sum_{q=1}^{n_i} x_{i,q} \tag{2.111}$$

Burada, $x_{i,q}$, π_i sınıfından gelen q. gözlemdir ve n_i, π_i sınıfındaki gözlem sayısıdır. Kovaryans matrisi de, Σ_i'nin yansız maksimum olabilirlik tahmin edicisi olan S_i ile yani örnek grup kovaryans matrisi ile tahmin edilir.

$$\Sigma_i \equiv S_i = \frac{1}{n_i - 1}\sum_{q=1}^{n_i}(x_{i,q} - \bar{x}_i)(x_{i,q} - \bar{x}_i)' \tag{2.112}$$

Bunlara ilave olarak, verinin sadece normal dağılımlı olup olmadığı aynı zamanda istatistiksel olarak da bağımsız olduğu ve örnek ortalaması ve örnek grup kovaryans matrisi tahminlerinin, eğitim gözlemlerinin ortak olabilirliğini maksimize eden özelliğe sahip

olduğu farz edilir [Johnson ve Richern, 1998]. Yani, maksimum olabilirlik tahminleri, marjinal yoğunluk fonksiyonunu maksimize eder.

$$(\overline{x}_i, S_i) = \arg\max_{\mu_i, \Sigma_i} \prod_{q=1}^{n_i} f_i(x_{i,q} \mid \mu_i, \Sigma_i) \tag{2.113}$$

Bayes kuralında, kovaryans matrisi ve ortalamanın gerçek değerleri yerine bu tahminler yazılırsa aşağıdaki sınıflama kuralı elde edilir.

- x'in π_i sınıfına atanması için

$$d_i(x) = \ln|S_i| + (x_i - \overline{x}_i)' S_i^{-1} (x_i - \overline{x}_i) - 2\ln P(\pi_i) \tag{2.114}$$

minimize olmalıdır. Son denklem aynı zamanda standart veya konvensiyonel kuadratik diskriminant fonksiyonu sınıflayıcısı olarak da bilinir.

2.7.2.2 Lineer Diskriminant Sınıflayıcısı

S_i örnek grup kovaryans matrisinin durağan olmaması ve singülerliği ile ilgili karşılaşılabilecek problemlerin üstesinden gelebilmek için rutin olarak uygulanan basit metotlardan biri Lineer Diskriminant Sınıflayıcısı'dır.

Lineer diskriminant sınıflayıcısı, (2.114) denkleminde S_i'nin yerine aşağıda tanımlanan S_p örnek kovaryans matrisinin yazılması ile elde edilir.

$$S_p = \frac{1}{n-k}\sum_{i=1}^{k}(g_i - 1)S_i = \frac{(g_1 - 1)S_1 + (g_2 - 1)S_2 + \ldots + (g_k - 1)S_k}{n-k} \tag{2.115}$$

Burada $n = n_1 + n_2 + \ldots + n_k$ gruplardaki toplam gözlem sayısı, k ; toplam grup sayısı ve $i = 1, 2, \ldots, k$ için g_i gruplardır. Bu kovaryans matrisine birleştirilmiş kovaryans matrisi denir. S_p aslında her grup için ayrı ayrı hesaplanan S_i'lerin ağırlıklı ortalamasından başka bir şey değildir. S_p, potansiyel olarak S_i'den daha yüksek ranka, normal olarak full ranka sahip olacaktır.

Teorik olarak S_p, sadece $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$ olduğu zaman gerçek kovaryans matrisinin tutarlı bir tahmin edicisi olur. Ancak yapılan çalışmalar göstermiştir ki, lineer diskriminant sınıflayıcısı sadece basit ve kullanımı kolay değil, aynı zamanda küçük örneklem probleminin olduğu durumlarda özel olarak iyidir. Bu gibi durumlarda, gerçek kovaryans matrisi her bir grupta farklı olsa bile, konvensiyonel kuadratik diskriminant fonksiyonunu daha iyi yapabilir.

Sınıflama bağlamında, kovaryans matrisleri çok farklı olmadığı ve küçük örnek büyüklüğüne ulaşılabildiği sürece, normal dağılımla ilişkili olan elipsoidal simetri, onun şeklinden ziyade düşünme yönüyle ilişkilidir. Aslında James (1985)'de göstermiştir ki, lineer diskriminant analizi grupların en çok temsil edildiği bölgelerde, her bir grubun örnek ortalamasının yakınında bulunan gözlemler için benzer sonuçlar veren kuadratiklerden birine daha yakın olabilir. Bu bölgede gözlemlerin çoğu güya sınıflandırılmıştır. Dahası, kuadratik diskriminant sınıflayıcı lineer diskriminant sınıflayıcısına göre daha fazla örnek gerektirir ve normal dağılım varsayımının bozulmasına karşın daha hassastır. Bu avantajlarından dolayı lineer diskriminant sınıflayıcısı, en popüler sınıflayıcılardan biridir [Thomaz, 2004].

3. UYGULAMA ve BULGULAR

Buraya kadar yüksek boyutlu sınırlı örneklem büyüklüğüne sahip veri setlerinde, önemli bilgiyi ortaya çıkarmak ve sonuç olarak boyut indirgeyebilmek ve aykırı gözlemleri tespit edebilmek amacıyla önerilen modellerin teorik altyapısı üzerinde durulmuştur. Bu bölümde ise yöntemlerin hesaplama adımları hatırlatıldıktan sonra, öncelikle boyut indirgeme için önerilen HDR yönteminin performansı daha sonra aykırı gözlemlerin tespiti için önerilen HOD yönteminin performansı, hem yapılan simülasyon çalışmaları ile elde edilen türetilmiş veri setleri üzerinde, hem de altı tane mikrodizilim veri setleri üzerinde gösterilecektir. Model seçim kriterleri olarak kullanılan AIC, CAIC, ve CICOMP'ı hem temel bileşen seçimi için hem de aykırı gözlem tespiti için hesaplayan ve ayrıca PPCA üzerinde hibritleştirilmiş kovaryans tahmin edicilerinin kullanımı ile yapılan değişiklikleri uygulayan hazır bir paket program bulunmadığı için hesaplamalar MatLab programı üzerinde kodlanmıştır.

3.1 ICOMP ile Boyut Seçme: Simulasyon çalışması

ICOMP bilgi kriterinin PPCA ile boyut indirgeme yapılması esnasında, boyut sayısına karar verme aracı olarak kullanılabilmesi, boyutun araştırmacı tarafından keyfi olarak seçilmesinin önüne geçecek ve sahip olduğu teorik dayanak noktasından dolayı seçilen boyut sayısına duyulan güven artıracaktır. Bu amaçla, teorik alt yapısı Bölüm 2.2.8'de gösterilmiş olan yapının geçerliliğini test edebilmek amacıyla bir takım simülasyon çalışmaları yapılmıştır. Bu noktada amaç, yapısal olarak gerçek boyut sayısının önceden bilindiği veri setlerine bilgi kriterleri ile PPCA analizi yapmak, bilgi kriterleri ile boyut sayısına karar vermek ve yinelemeli olarak bu işlemler tekrar edildiğinde, bilgi kriterlerinin hangilerinin gerçek boyutu daha isabetli bir şekilde tespit edebildiğini ortaya koyabilmektir.

Simülasyon çalışmalarında iki farklı protokol kullanılmıştır. Bu protokollerde gerek gözlem sayıları için ($n = 100, 200, 500, 1000, 10000....$) gerekse varyans bileşenleri için ($\sigma^2 = 0.01, 0.1, 0.5, 1, 10...$) farklı senaryolar denenmiştir. Bu senaryoların hepsine ait sonuçlar elde edilmiştir. Ancak örnek olarak, Tablo 3.1'de kullanılan protokol ile boyut sayısı 9, gerçek boyut sayısı 3 olan $n = 1000$ gözleme sahip veri seti 100 defa üretildiğinde bilgi kriterlerinin gerçek boyutu isabet ettirmedeki performansları gösterilecektir.

Tablo 3. 1: Simülasyon Protokolü

$$S: \quad \Lambda = \begin{bmatrix} 0.6 & 0 & 0 \\ 0.7 & 0 & 0 \\ 0.8 & 0 & 0 \\ 0 & 0.6 & 0 \\ 0 & 0.7 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.6 \\ 0 & 0 & 0.7 \\ 0 & 0 & 0.8 \end{bmatrix}, \quad H = \begin{bmatrix} 1.0 & 0.8 & 0.72 \\ 0.8 & 1.0 & 0.84 \\ 0.72 & 0.84 & 1.0 \end{bmatrix},$$

$$\Sigma = \Lambda * H * \Lambda' + \sigma^2 I\,, \quad \mu = [0]_{1x9}$$

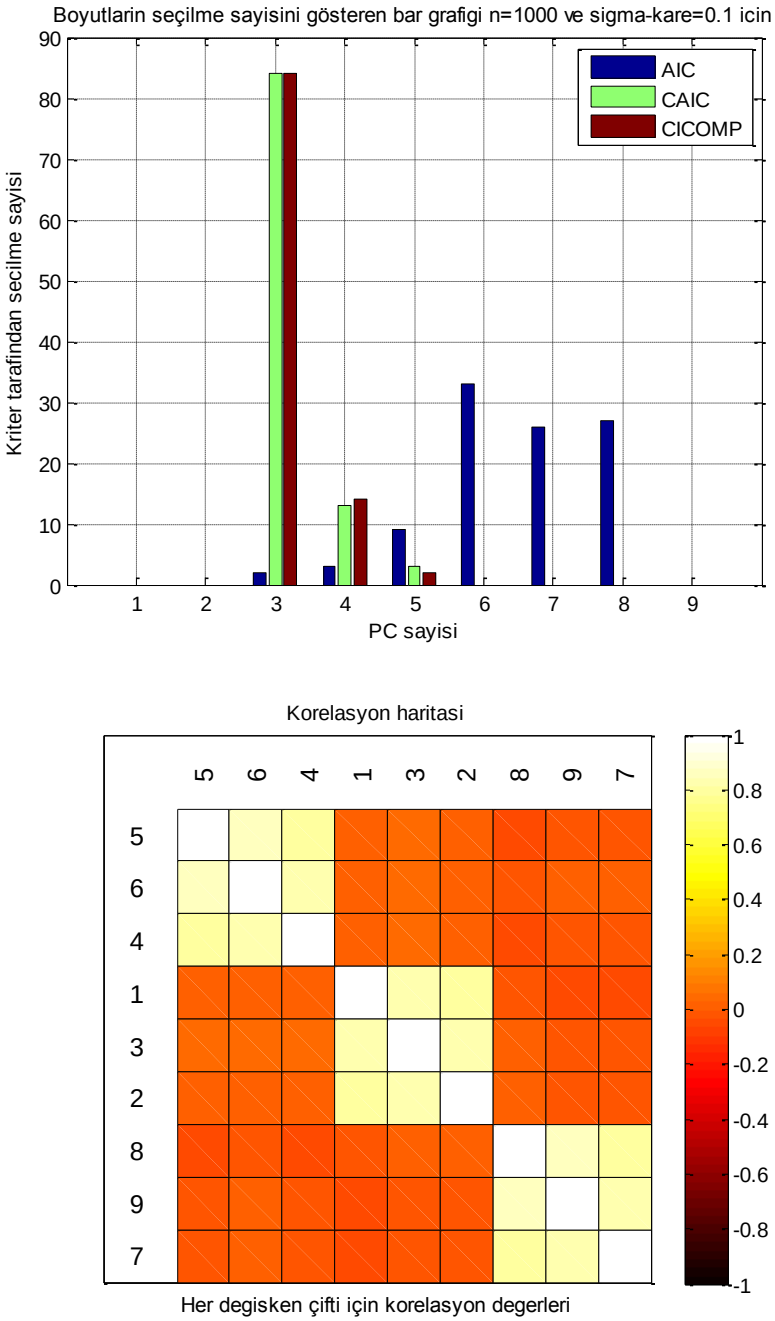
Tablo3.1'de H, gerçek boyutlar arasındaki korelasyonu gösteren ilişki matrisi, σ^2 kovaryans matrisinin ayrışımındaki varyans bileşeni olup örnek çalışmada 0.1 olarak alınmıştır. Sonuçlar aşağıdaki gibidir.

Tablo 3. 2: Türetilmiş veri setlerine bilgi kriterleri ile PPCA uygulanması sonucunda gerçek boyutu isabetli yakalama sayıları

n=1000	Boyut sayısı								
σ^2=0.1	1	2	3*	4	5	6	7	8	9
AIC	0	0	2	3	9	33	26	27	0
CAIC	0	0	84	13	3	0	0	0	0
CICOMP	0	0	84	14	2	0	0	0	0

CPU zamanı: 2,73 sn
*: Gerçek boyutu gösterir

n=1000 gözlem ve p=9 değişkene sahip gerçek boyutu 3 olan veri setinden 100 defa üretilip, 100 defa bilgi kriterleri ile PPCA'deki önemli bileşen sayısına, yani boyuta karar verilmiştir. Tablo 3.2'de verilen sonuçlara göre Bozdogan'ın kriterleri olan CAIC ve CICOMP'ın gerçek boyutu %84 oranında tespit ettikleri belirlenmiştir. Sonuç olarak Bozdogan'ın bilgi kriterleri gerçek boyut sayısını yakalama noktasında AIC bilgi kriterine göre belirgin bir üstünlük sağlamıştır. Özellikle AIC tipi kriterlerin fazla parametreye sahip olan modeli seçme yönündeki eğiliminin, burada da fazla boyut sayısına karar verme şeklinde ortaya çıktığı tespit edilmiştir. Sonuçlara ilişkin bar grafiği ve türetilen veri setlerinin korelasyon yapısı Şekil 3.1'de gösterilmiştir.



Şekil 3. 1: Bilgi kriterlerinin boyut tespit etme sayılarına ilişkin bar grafiği ve türetilen verinin korelasyon yapısı

Dolayısıyla, ICOMP tipi kriterlerin kullanılması ile verideki gerçek boyutun isabetli bir şekilde tespit edilebileceği söylenebilir.

3.2 HDR ile Mikrodizilim Veri Analizi

Tez kapsamında önerilen HDR yöntemi Bölüm 2.4'de de anlatıldığı gibi aşağıdaki hesaplama adımlarına sahiptir.

Aşama-1: Yüksek boyutlu kanser gen ifade verisi için maksimum entropi kovaryans matrisi hesaplanır

Aşama-2: Maksimum entropi kovaryans matrisi ile Bölüm 2.3.3'deki adımları kullanarak $\hat{\Sigma}_{HCE}$ kovaryans matrisi oluşturulur

Aşama-3: $\hat{\Sigma}_{HCE}$ ile PPCA analizi yapılır

Aşama-4: Bilgi Kriterleri ile uygun $\hat{\Sigma}_{HCE}$ yapısı ve boyut sayısına karar verilir.

Bu bölümde ise HDR'nin geçerliliğini ve etkinliğini gösterebilmek amacıyla altı tane mikrodizilim gen ifade verileri üzerinde uygulama yapılmıştır. Analizlerde kullanılan veri setleri Tablo 3.3'de verilmiştir.

Tablo 3. 3: Çalışmada kullanılan veri setleri

Veriseti	Kaynak	n	p	Grup	Grup tanımları
Leukemia(Lösemi)	Golub et.al(1999)	72	3571	2	Farklı lösemi tipleri
Colon(Kolon)	Alon et al.(1999)	62	2000	2	Sağlam ve kanserli doku
Prostate(Prostat)	Singh et. al. (2002)	102	6033	2	Sağlam ve kanserli doku
Lymphoma(Lenfoma)	Alizadeh et al. (2000)	62	4026	3	Farklı lenfoma türleri
SRBCT	Khan et al(2001)	83	2308	4	Farklı tümör dokuları
Brain(Beyin)	Pomeroy et al (2002)	42	5597	5	Farklı tümör dokuları

Lösemi veri seti, iki farklı tümör dokularından alınan örneklerle oluşturulmuştur. Bunlar akut lenfoblastik lösemi (ALL) için 47 hasta ve akut miyeloid lösemi (AML) 25 hasta şeklindedir. Toplam 72 hastanın herbirinden elde edilen kemik iliği örnekleri Affymetrix Hgu6800 çipleri ile analiz edilerek bu dokular üzerinde 3571 gen ifadesi incelenmiştir.

Kolon veri seti, 22 gözlem sağlam ve 40 gözlem tümörlü dokulardan olmak üzere iki alt gruptan oluşmaktadır. Toplam 62 gözlemin herbirinden elde edilen dokular Affymetrix oligonükleotid dizisi ile analiz edilerek bu dokular üzerinde 2000 gen ifadesi incelenmiştir.

Prostat veri seti için Singh ve ark. (2002), 1995 ve 1997 yılları arasında 235 cerrahi hastadan radikal prostatektomi örneklerini inceleyerek bunlardan 102 tanesinin yüksek kaliteye sahip olduklarını belirtmişlerdir. 52 tanesi prostat tümör örnekleri 50 tanesi ise sağlam doku örnekleri olmak üzere toplam 102 gözlem üzerinde oligonükleotid mikrodiziler ile 6033 gen ifadesini incelemişlerdir.

Lenfoma veri seti, 42 örnek yaygın büyük B-hücreli lenfoma (diffuse large B-cell lymphoma-DLBCL), 9 örnek foliküler lenfoma (follicular lymphoma -FL),11 örnek kronik lenfositik lösemi (chronic lymphocytic leukemia-CLL) olmak üzere 63 örnek ihtiva etmektedir. Bu tümör çeşitlerine ait dokular üzerinde 4026 gen ifadesi incelenmiştir.

SRBCT (Small Round Blue Cell Tumor) veri seti, dört farklı tümör dokularından alınan örneklerle oluşturulmuştur. Bunlar ewing family of tumors (EWS) için 29 hasta, neuroblastoma (NB) için 18 hasta, rhabdomyosarcoma (RMS) için 25 hasta ve non-hodgkin lymphoma (NHL) için 11 hasta şeklindedir. Toplam 83 hastanın herbirinden elde edilen dokular üzerinde 2308 tane gen ifadesi incelenmiştir.

Beyin veri setinde, 5 farklı tümör grubunun incelenmiştir. Alt gruplar; 10 örnek medülloblastomlar, 5 örnek CNS AT/RTs, 5 örnek böbrek ve böbrek dışı rabdoid tümörler, 8 örnek supratentoriyal PNET, 10 örnek embriyonel olmayan beyin tümörleri (malignant glioma) and 4 örnek normal insan serebellası şeklindedir. Affymetrix oligonükleotit çipler kullanılarak bu dokular üzerinde 5597 gen ifadeleri incelenmiştir.

Tablo 3.3'den de anlaşıldığı üzere tüm veri setleri ciddi şekilde küçük örneklem problemine sahiptir. Daha da önemlisi çeşitli alt gruplardan oluşan bu veri setleri için, 5 örnek içeren bir alt grubun varlığı gözönüne alındığında, alt grupların daha da ciddi bir şekilde küçük örneklem problemine sahip olduğu aşikardır.

Veri setlerinde ilk genlerdeki bilgi miktarı ile sınıflandırma yapmanın zorluğunu ortaya koyabilmek için ilk 5, 10 ve 15 gen kullanılarak LDA ve QDA ile sınıflandırma yapılmıştır. Sonuçlar Tablo 3.4'deki gibidir.

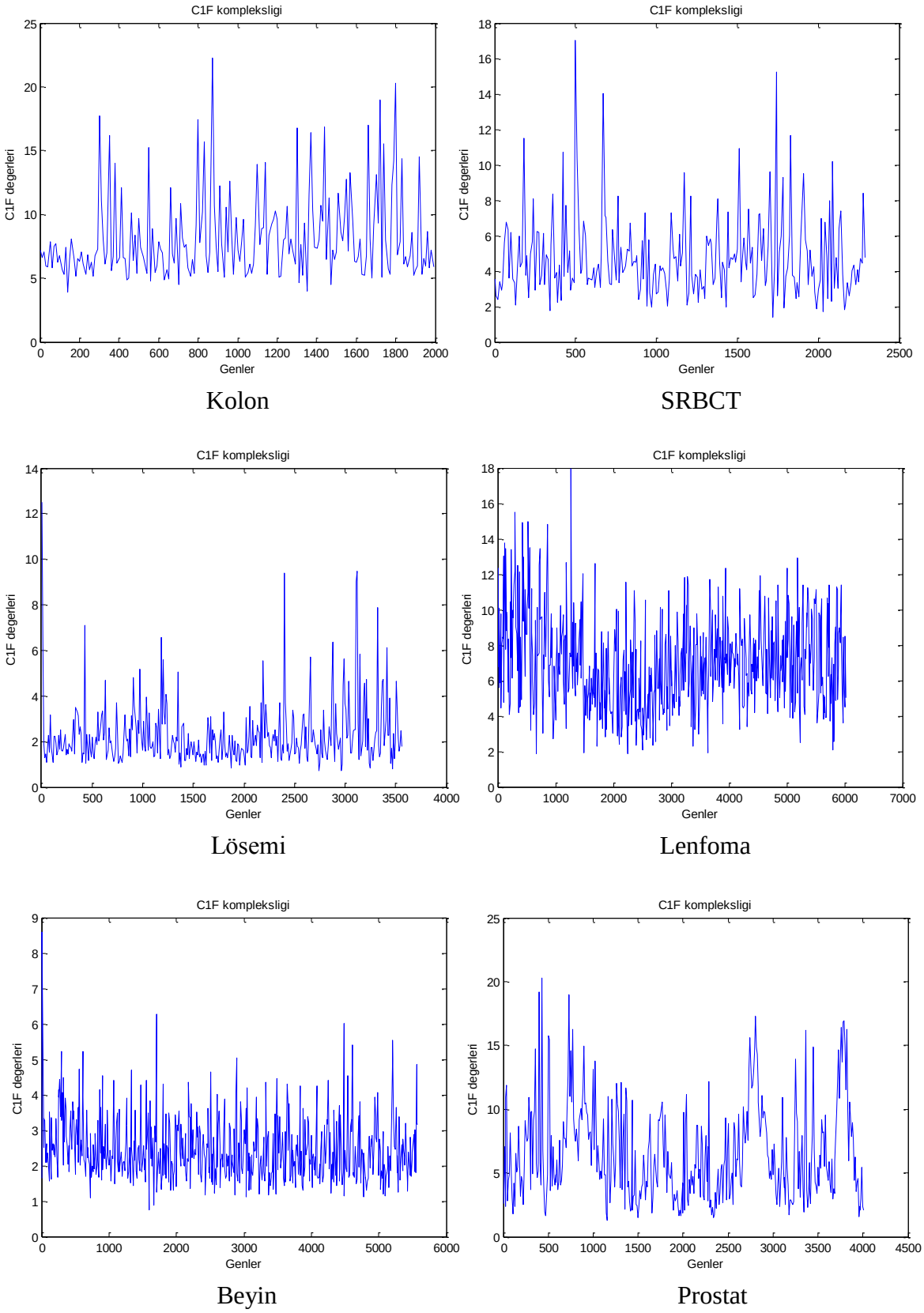
Tablo 3. 4: Tüm veri setlerinde ilk genler için sınıflandırma sonuçları

Veri setleri	Orijinal Boyut	LDA hatalı sınıflama yüzdesi			QDA hatalı sınıflama yüzdesi		
		5 gen	10 gen	15 gen	5 gen	10 gen	15 gen
Lösemi	3571	43.1%	33.3%	20.1%	31.9%	22.2%	6.9%
Kolon	2000	35.4%	32.2%	25.8%	32.2%	16.1%	16.1%
Prostat	6033	34.3%	36.27%	18.6%	30.4%	25.5%	18.6%
Lenfoma	4026	30.6%	29.0%	16.1%	19.4%	1.61%*	3.22%*
SRBCT	2308	31.8%	12.7%	17.5%	15.9%	3.18%*	0.0%*
Beyin	5597	45.2%	30.9%	14.2%	42.8%*	9.5%*	16.6%*

*: Singülerliği gösterir

Elde edilen sonuçlar incelendiğinde, hatalı sınıflama oranının yüksekliği dikkat çekmektedir. Bu veri setlerinde var olan kompleksliği görsel olarak ortaya koyabilmek için Bozdogan (1988) tarafından geliştirilen C_{1F} maksimal kovaryans kompleksliği kullanılmıştır.

Şekil 3.2, tüm veri setleri için 10'arlı gruplar halinde kümelenerek hesaplanan C_{1F} komplekslik değerlerine ait grafikleri göstermektedir. Bu grafikler veri setlerinin hangi kısımlarında kompleksliğinin azaldığını ve arttığını göstermektedir. Yer yer dalgalanmaların söz konusu olduğu grafiklerden de anlaşılmaktadır. Bu nedenle; bu grafikler, veri setinde gizli olan bilginin verinin değişik bölgelerinde olabileceği hususunda bize görsel bir imkan sağlamaktadır.



Şekil 3. 2: Çalışmada kullanılan tüm veri setleri için C_{1F} komplekslik grafikleri

Verilerin yüksek boyutlarındaki bilgilere de ulaşabilmek için Maksimum Entropi Kovaryans Matrisi ve mevcut kovaryans yapılarının muhtemel tüm hibritleşmeleri için farklı $\hat{\Sigma}_{HCE}$ hesaplanarak tek tek analiz yapılmış ve PPCA analizi için hangi bilgi kriteri ile hangi kovaryans yapısının kullanılabileceği incelenmiştir. En iyi sonuçları veren 2 farklı yapı ME/STA/CSE ve ME/STA/BCSE olarak elde edilmiştir. Tablo 3.5'den Tablo 3.10'a kadar her veri setinde kullanılan $\hat{\Sigma}_{HCE}$ kovaryans matrisinin özdeğerler üzerinde yaptığı iyileştirmeler ve buna bağlı olarak $\hat{\Sigma}_{HCE}$ için singülerlik durumunun incelemeleri gösterilmektedir.

Tablo 3. 5: Lösemi veri seti için $\hat{\Sigma}_{HCE}$ ile özdeğer iyileştirmesi

$\hat{\Sigma}$ için özdeğerler	$\hat{\Sigma}_{ME}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_CSE}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_BCSE}$ için özdeğerler
1.o*e-13	1.0*e-5	1.0*e+0	1.0*e+1	1.0*e+0
-5.329900	2.411268	8.489431	1.104235	8.559948
-4.548523	2.775117	8.489431	1.104235	8.562411
-3.911744	2.926139	8.489431	1.104235	8.570907
-3.844771	3.283898	8.489431	1.104235	8.584887
-3.644393	3.284871	8.489431	1.104235	8.589654
-3.574917	3.411935	8.489431	1.104235	8.590812
-3.369492	3.997277	8.489431	1.104235	8.599542
-3.126794	4.069785	8.489431	1.104235	8.602612
-2.981659	4.108430	8.489431	1.104235	8.604998
-2.941157	4.226116	8.489431	1.104235	8.610829
$\kappa(\Sigma) =$ -7.32E-16	$\kappa(\Sigma) =$ 2.29E-09	$\kappa(\Sigma) =$ 0.000807842	$\kappa(\Sigma) =$ 0.001515024	$\kappa(\Sigma) =$ 0.022982951

$\hat{\Sigma}_{ME}$: *Maksimum Entropi Kovaryans matrisi*
$\hat{\Sigma}_{ME\_STA}$: *Maksimum Entropi Thomaz Stabilizasyonu*
$\hat{\Sigma}_{ME\_STA\_CSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Convex Sum Hibrit yapısı*
$\hat{\Sigma}_{ME\_STA\_BCSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Bozdogan Convex Sum Hibrit Yapısı*

Tablo 3. 6: Kolon veri seti için $\hat{\Sigma}_{HCE}$ ile özdeğer iyileştirmesi

$\hat{\Sigma}$ için özdeğerler	$\hat{\Sigma}_{ME}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_CSE}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_BCSE}$ için özdeğerler
1.0*e+1	1.0*e+1	1.0*e+13	1.0*e+13	1.0*e+13
-4.17449	-2.209985	5.631931	6.734519	5.631931
-0.446566	-0.459016	5.631931	6.734519	5.631932
-0.237418	0.0231485	5.631931	6.734519	5.631932
-0.161746	0.0373740	5.631931	6.734519	5.631932
-0.139540	0.0412454	5.631931	6.734519	5.631933
-0.118291	0.0548244	5.631931	6.734519	5.631933
-0.107965	0.0612780	5.631931	6.734519	5.631933
-0.101894	0.0616895	5.631931	6.734519	5.631934
-0.080946	0.0618432	5.631931	6.734519	5.631934
-0.071084	0.0679490	5.631931	6.734519	5.631934
$\kappa(\Sigma) =$	$\kappa(\Sigma) =$	$\kappa(\Sigma) =$	$\kappa(\Sigma) =$	$\kappa(\Sigma) =$
-8.85E-16	-3.99E-16	0.00101724	0.001523292	0.001590843

$\hat{\Sigma}_{ME}$: *Maksimum Entropi Kovaryans matrisi*
$\hat{\Sigma}_{ME\_STA}$: *Maksimum Entropi Thomaz Stabilizasyonu*
$\hat{\Sigma}_{ME\_STA\_CSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Convex Sum Hibrit yapısı*
$\hat{\Sigma}_{ME\_STA\_BCSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Bozdogan Convex Sum Hibrit Yapısı*

Tablo 3. 7: Prostat veri seti için $\hat{\Sigma}_{HCE}$ ile özdeğer iyileştirmesi

$\hat{\Sigma}$ için özdeğerler	$\hat{\Sigma}_{ME}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_CSE}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_BCSE}$ için özdeğerler
1.0*e-12	1.0*e-5	1.0*e+1	1.0*e+1	1.0*e+1
-3.746241	1.166101	1.122460	1.384135	1.125114
-1.531800	1.271342	1.122460	1.384135	1.126098
-1.229232	1.441267	1.122460	1.384135	1.128947
-1.105239	1.468678	1.122460	1.384135	1.130653
-1.102578	1.651203	1.122460	1.384135	1.130916
-1.034967	1.745122	1.122460	1.384135	1.131110
-1.007104	1.865418	1.122460	1.384135	1.131887
-0.941359	1.938657	1.122460	1.384135	1.131948
-0.909864	1.974393	1.122460	1.384135	1.132049
-0.831655	2.062129	1.122460	1.384135	1.132166
$\kappa(\Sigma) =$	$\kappa(\Sigma) =$	$\kappa(\Sigma) =$	$\kappa(\Sigma) =$	$\kappa(\Sigma) =$
-1.11E-15	4.92E-10	0.000473552	0.000765251	0.018089278

$\hat{\Sigma}_{ME}$: *Maksimum Entropi Kovaryans matrisi*
$\hat{\Sigma}_{ME\_STA}$: *Maksimum Entropi Thomaz Stabilizasyonu*
$\hat{\Sigma}_{ME\_STA\_CSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Convex Sum Hibrit yapısı*
$\hat{\Sigma}_{ME\_STA\_BCSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Bozdogan Convex Sum Hibrit Yapısı*

Tablo 3. 8: Lenfoma veri seti için $\hat{\Sigma}_{HCE}$ ile özdeğer iyileştirmesi

$\hat{\Sigma}$ için özdeğerler	$\hat{\Sigma}_{ME}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_CSE}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_BCSE}$ için özdeğerler
1.0*e-11	1.0*e-4	1.0*e+1	1.0*e+2	1.0*e+1
-1.194355	7.939347	8.175630	1.032737	8.230170
-0.5869232	8.593514	8.175630	1.032737	8.232448
-0.5734302	9.212570	8.175630	1.032737	8.237471
-0.5101604	9.572241	8.175630	1.032737	8.240535
-0.4849419	9.577893	8.175630	1.032737	8.243370
-0.4658723	9.666184	8.175630	1.032737	8.246258
-0.4262829	10.00798	8.175630	1.032737	8.246487
-0.4005819	10.01572	8.175630	1.032737	8.252508
-0.3870031	10.07410	8.175630	1.032737	8.252535
-0.3815725	10.08569	8.175630	1.032737	8.253432
$\kappa(\Sigma) =$ -7.66E-16	$\kappa(\Sigma) =$ 5.77E-10	$\kappa(\Sigma) =$ 0.000593795	$\kappa(\Sigma) =$ 0.001023289	$\kappa(\Sigma) =$ 0.013422542

$\hat{\Sigma}_{ME}$: *Maksimum Entropi Kovaryans matrisi*
$\hat{\Sigma}_{ME\_STA}$: *Maksimum Entropi Thomaz Stabilizasyonu*
$\hat{\Sigma}_{ME\_STA\_CSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Convex Sum Hibrit yapısı*
$\hat{\Sigma}_{ME\_STA\_BCSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Bozdogan Convex Sum Hibrit Yapısı*

Tablo 3. 9: SRBCT veri seti için $\hat{\Sigma}_{HCE}$ ile özdeğer iyileştirmesi

$\hat{\Sigma}$ için özdeğerler	$\hat{\Sigma}_{ME}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_CSE}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_BCSE}$ için özdeğerler
1.0*e-12	1.0*e-6	1.0*e+1	1.0*e+1	1.0*e+1
-4.318355	6.614255	4.851013	6.602953	4.851423
-3.404719	10.073861	4.851013	6.602953	4.851547
-2.267305	10.089544	4.851013	6.602953	4.851626
-1.621497	10.166824	4.851013	6.602953	4.851690
-1.594498	10.205549	4.851013	6.602953	4.851712
-1.393361	10.219934	4.851013	6.602953	4.851724
-1.266208	10.229870	4.851013	6.602953	4.851726
-1.158224	10.231940	4.851013	6.602953	4.851730
-1.111982	10.232613	4.851013	6.602953	4.851731
-1.028455	10.330809	4.851013	6.602953	4.851837
$\kappa(\Sigma) =$ -2.43E-16	$\kappa(\Sigma) =$ 2.01E-10	$\kappa(\Sigma) =$ 0.001476385	$\kappa(\Sigma) =$ 0.003190884	$\kappa(\Sigma) =$ 0.00280856

$\hat{\Sigma}_{ME}$: *Maksimum Entropi Kovaryans matrisi*
$\hat{\Sigma}_{ME\_STA}$: *Maksimum Entropi Thomaz Stabilizasyonu*
$\hat{\Sigma}_{ME\_STA\_CSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Convex Sum Hibrit yapısı*
$\hat{\Sigma}_{ME\_STA\_BCSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Bozdogan Convex Sum Hibrit Yapısı*

Tablo 3. 10: Beyin veri seti için $\hat{\Sigma}_{HCE}$ ile özdeğer iyileştirmesi

$\hat{\Sigma}$ için özdeğerler	$\hat{\Sigma}_{ME}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_CSE}$ için özdeğerler	$\hat{\Sigma}_{ME\_STA\_BCSE}$ için özdeğerler
1.0*e-12	1.0*e-5		1.0*e+1	
-1.773675	1.670242	8.642946	**1.196218**	8.649957
-1.089631	2.397314	8.642946	**1.196218**	8.652881
-1.037029	2.426444	8.642946	1.196218	8.653174
-1.023396	2.494041	8.642946	1.196218	8.656057
-0.932597	2.646266	8.642946	1.196218	8.660741
-0.807212	2.735081	8.642946	1.196218	8.661537
-0.795727	3.194235	8.642946	1.196218	8.661939
-0.776082	3.348018	8.642946	1.196218	8.662572
-0.760816	3.464996	8.642946	1.196218	8.664153
-0.739112	3.623167	8.642946	1.196218	8.664671
$\kappa(\Sigma) =$ -1.39E-15	$\kappa(\Sigma) =$ 9.04E-10	$\kappa(\Sigma) =$ 0.000467536	$\kappa(\Sigma) =$ 0.001054974	$\kappa(\Sigma) =$ 0.051677161

$\hat{\Sigma}_{ME}$: *Maksimum Entropi Kovaryans matrisi*
$\hat{\Sigma}_{ME\_STA}$: *Maksimum Entropi Thomaz Stabilizasyonu*
$\hat{\Sigma}_{ME\_STA\_CSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Convex Sum Hibrit yapısı*
$\hat{\Sigma}_{ME\_STA\_BCSE}$:*Maksimum Entropi Thomaz Stabilizasyonu ile Bozdogan Convex Sum Hibrit Yapısı*

Yapılan hesaplamalar, klasik örnek kovaryans matrisinin pozitif olmayan (non-positive) ve kötü şartlanmış (ill-conditioned) olduğunu ve özdeğerlerin büyük bölümünün negatif veya sıfır olduğunu göstermiştir. Yukarıda verilen tablolar, özellikle negatif özdeğer problemine vurgu yapabilmek için ve veri setlerinin boyutları çok yüksek olduğu için, örnek kovaryans matrisi ve diğer kovaryans matrisleri kullanıldığında elde edilen özdeğerlerden sadece en küçük 10 tanesinin değerlerini göstermektedir. Ancak örnekle ifade etmek gerekirse, SRBCT verisinde 2308 özdeğerden 1122 tanesinin negatif çıktığı belirlenmiştir. Ayrıca virgülden sonra 6 basamak yazıldığı için bazı özdeğerler eşit gibi görülmektedir. Ancak, Tablo 3.10'da koyu olarak vurgulanan özdeğerlerin gerçek değerleri 11.9621771439795 ve 11.9621771439805 şeklinde olup bu özdeğerler farklıdır. Kullanmış olduğumuz kovaryans matrislerinin klasik kovaryans matrisinde olduğu gibi dejenere olmadıkları, pozitif özdeğerlere sahip oldukları anlaşılmıştır. $\kappa(\Sigma)$ değerleri de, önerdiğimiz $\hat{\Sigma}_{HCE}$ kovaryans yapılarının klasik örnek kovaryans matrisine göre singülerlikten uzak olduklarını göstermektedir. Artık $\hat{\Sigma}_{HCE}$ ile PPCA analizi yapılarak boyut indirgenebilir.

Daha sonraki aşama, önemli olan temel bileşen sayısına karar vermektir. Literatürde var olan klasik yöntemlerden farklı olarak, önemli bileşen sayısı, bilgi kriterlerinin yardımı ile tespit edilmiştir. Bu amaç için Bölüm 2.2.8'de detayları anlatılan kriterlerin boyut sayısına karar vermede kullanılma şekli şöyledir: PPC bileşenleri tek tek ilave edilerek oluşturulan modellerin bilgi kriterleri için değerleri hesaplanır. Minumum değerin elde edildiği modelde bulunan bileşen sayısı, önemli bileşen, dolayısıyla boyut sayısı olarak tespit edilir. Bu sonuçlara ilişkin grafikler bu bölümün sonunda tek bir seti için örnek olarak verilecektir.

Böylelikle, verideki korelasyon yapısının azaltılması ve önemli bileşenlerin seçilmesiyle, yüksek boyutlu sınırlı örneklem büyüklüğüne sahip veri setlerinin boyut indirgemesi yapılmıştır. Tüm veri setleri için bilgi kriterlerine göre ayrı ayrı elde edilen sonuçlar Tablo 3.11 ve Tablo 3.12'de verildiği gibidir.

Tablo 3. 11: AIC bilgi kriteri ile seçilen boyut sayıları ve kullanılan kovaryans yapıları

Veri setleri	Orijinal Boyut	$\hat{\Sigma}_{HCE}$	İndirgenen Boyut	LDA hatalı sınıflama yüzdesi	QDA hatalı sınıflama yüzdesi
Lösemi	3571	ME/STA/BCSE	6	15.28%	26.3%
Kolon	2000	ME/STA/CSE	14	19.35%	12.9%
Prostat	6033	ME/STA/BCSE	6	36.27%	34.31%
Lenfoma	4026	ME/STA/BCSE	15	4.83%	1.613%
SRBCT	2308	ME/STA/BCSE	15	9.52%	0.0%
Beyin	5597	ME/STA/CSE	14	2.38%	4.76%

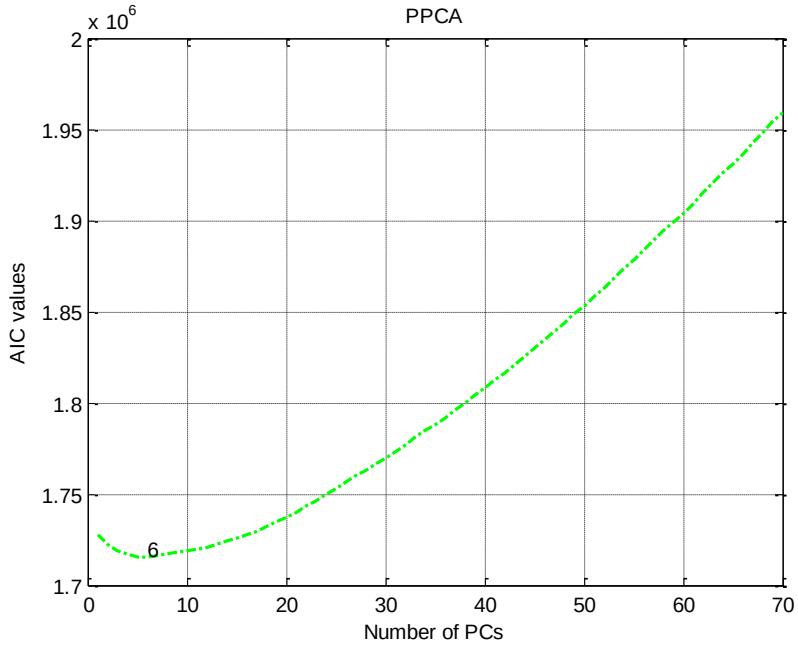
Tablo 3. 12: CAIC ve CICOMP bilgi kriteri ile seçilen boyut sayıları ve kullanılan kovaryans yapıları

Veri setleri	Orijinal Boyut	$\hat{\Sigma}_{HCE}$	İndirgenen Boyut	LDA hatalı sınıflama yüzdesi	QDA hatalı sınıflama yüzdesi
Lösemi	3571	ME/STA/CSE	8	1.38%	0.0%
Kolon	2000	ME/STA/CSE	10	22.58%	%14.52
Prostat	6033	ME/STA/CSE	12	3.92%	7.84%
Lenfoma	4026	ME/STA/CSE	8	0.0%	0.0%
SRBCT	2308	ME/STA/CSE	8	0.0%	0.0%
Beyin	5597	ME/STA/CSE	4	9.52%	4.76%

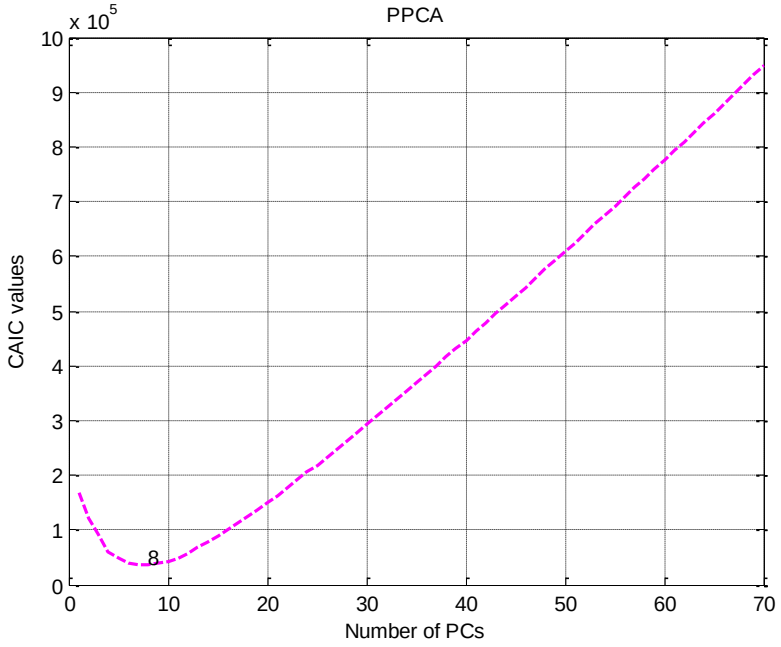
Tablo 3.11 ve 3.12'deki sonuçlar, Tablo 3.4'de ki sınıflama sonuçları ile kıyaslandığında, HDR ile seçilen boyutların sınıflandırma üzerinde elde ettikleri başarı oldukça belirgindir. Araştırmacı için karar kuralı boyut indirgemeden dolayı göze alınacak bilgi kaybı miktarı ile

sınıflama başarısı arasındaki alışveriş (trade-off) veya dengeyi dikkate almaktır. Kolon kanser verisindeki çok az bir farklılık dışında Bozdogan tarafından önerilen CAIC ve CICOMP kriterlerinin seçtiği kovaryans yapılarının ve boyutların sınıflama sonuçları üzerindeki üstün başarısı dikkat çekmektedir. Özellikle Lösemi veri seti, LDA için orijinal ilk 10 gen ile %33.3 ve AIC'nin seçtiği boyutlar ile %15.28 sınıflama hatası verirken bu oran CICOMP'ın seçtiği boyutlar ile %1.38'e gerilemiştir. QDA için çok daha çarpıcı bir şekilde %22.2 ve %26.3'den %0'a düşmüştür. Benzer ilgi çekici sonuçlar, Lenfoma ve SRBCT veri setleri için de elde edilmiştir. Tüm bu sonuçlar önerdiğimiz HDR yaklaşımının başarısını ortaya koymaktadır.

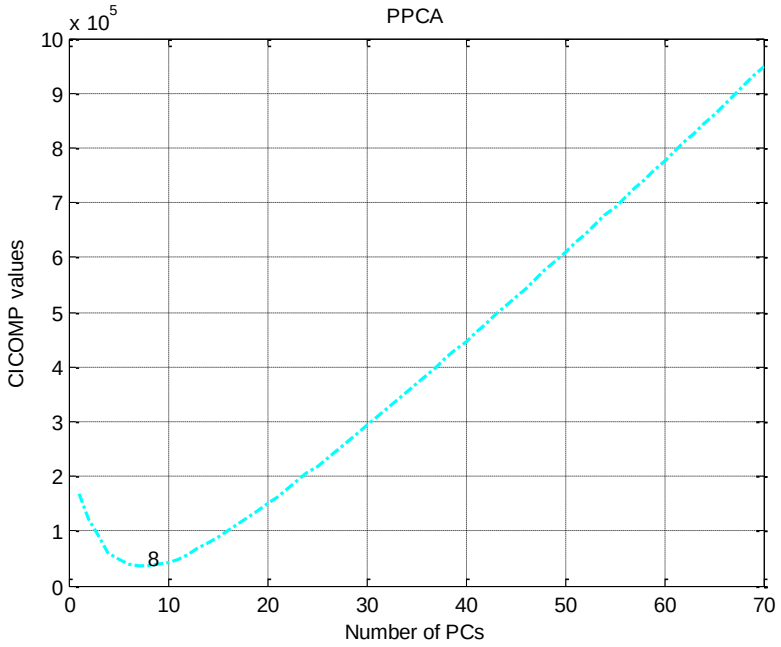
Son olarak sadece Lösemi veri setinin analizinde bilgi kriterleri için boyut seçmede kullanılan grafikler Şekil 3.3, 3.4 ve 3.5'de verilmiştir.



Şekil 3. 3: PPCA analizinde önemli boyut sayısına karar vermek için hesaplanan AIC değerleri grafiği.



Şekil 3. 4: PPCA analizinde önemli boyut sayısına karar vermek için hesaplanan CAIC değerleri grafiği.



Şekil 3. 5: PPCA analizinde önemli boyut sayısına karar vermek için hesaplanan CICOMP değerleri grafiği.

3.3 ICOMP ile Aykırı Gözlem Tespiti: Simülasyon Çalışması

Genel olarak $ICOMP_r$ şeklinde tanımlanan yapı, çalışma kapsamında kullanılan diğer kriterler için de benzer şekilde $AIC_r, CAIC_r$ ve $CICOMP_r$ şeklinde tanımlanabilir. Bu oranların aykırı gözlemleri tespit etmedeki başarısını ve yöntemin kontrol grafikleri ile nasıl çalıştığını gösterebilmek amacıyla aşağıdaki protokol çerçevesinde bir simülasyon çalışması yapılmıştır. Sonuçlar ayrıca geleneksel yöntemlerden biri olan Mahalanobis uzaklığı ile karşılaştırılmıştır.

Aykırı gözlemlerin olduğu bir veri seti türetmenin yolu, öncelikle normal yapıdaki gözlemleri türetmek ve daha sonra farklı mekanizmalar tanımlayarak buradan elde edilecek aykırı gözlemleri veriye eklemektir. Aykırı gözlemler verideki yerleri belli olacak şekilde rassal olarak dağıtıldıktan sonra, yöntemin bu gözlemlerin yerlerini tespit edip etmediğine bakılarak performans değerlendirmesi yapılabilir. Bu amaçla öncelikle, Bölüm 3.1'de belirlenen protokol çerçevesinde $N(\mu, \Sigma)$ dağılımından aykırı olmayan gözlemler türetilmiştir. Sonra aşağıda belirtilen mekanizmalar kullanarak aykırı gözlemlerin setleri oluşturulmuştur. Simülasyon çalışmasında veriye eklenecek q-tane aykırı gözlem, $N(\mu_q, \Sigma_q)$ dağılımından üretilmiştir. Bunun için Fritsch ve ark. (2012)'de tanımladıkları aşağıdaki protokol kullanılmıştır.

Senaryo-1: Aykırı gözlemler $\alpha > 1$ için $\Sigma_q = \alpha\Sigma$ ve $\mu_q = 0_q$ yapısından türetilir.

Senaryo-2: Aykırı gözlemler $\Sigma_q = \Sigma$ ve $\mu_q \neq 0_q$ yapısından türetilir.

Senaryo-3: Aykırı gözlemler $\mu_q \neq 0_q$ ve $\Sigma_q = \Sigma + \alpha a a^T$ yapısından türetilir. Burada, $a = \frac{a_{rand}}{\|a_{rand}\|}$ şeklinde olup a_{rand} ise $\mathcal{B}(\frac{1}{2})$ Bernoulli dağılımından çekilen p-boyutlu bir vektördür.

Farklı n-gözlem sayılarında, şekil ve lokasyon parametreleri için yukarıdaki senaryolar kullanılarak ve aykırı gözlemlerin bulaşma oranları da değiştirilerek çok farklı yapılarda veri setleri türetilmiştir. Her bir senaryo için sonuçlar Tablo 3.13, 3.14 ve 3.15'deki gibidir.

Tablo 3. 13: Senaryo-1 için elde edilen sonuçlar.

n	Cr	#of Out	AIC	CAIC	CICOMP	MD	CPU(sn)
40	5	2	86	86	86	53	2.46
	10	4	73	73	72	13	2.59
80	5	4	86	86	84	34	4.63
	10	8	64	64	61	3	4.92
100	5	5	74	74	76	26	6.24
	10	10	61	61	61	2	6.28
200	5	10	65	65	65	3	17.49
	10	20	36	36	37	0	17.57

$\Sigma_q = 8*\Sigma$ *, İterasyon sayısı=100, n: Gözlem sayısı*

C_r *:Bulaşma oranı (Contamination rate), #of Out: Aykırı gözlem sayısı*

Tablo 3. 14: Senaryo-2 için elde edilen sonuçlar.

n	Cr	#of Out	AIC	CAIC	CICOMP	MD	CPU(sn)
40	5	2	92	92	100	9	2.49
	10	4	3	3	100	0	2.52
80	5	4	74	74	100	0	4.89
	10	8	0	0	100	0	5.05
100	5	5	41	41	100	0	6.20
	10	10	0	0	100	0	6.37
200	5	10	11	11	100	0	17.19
	10	20	0	0	100	0	17.43

$\mu_q = 8*[1]$ *, İterasyon sayısı=100, n: Gözlem sayısı*

C_r *:Bulaşma oranı (Contamination rate), #of Out: Aykırı gözlem sayısı*

Tablo 3. 15: Senaryo-3 için elde edilen sonuçlar.

n	Cr	#of Out	AIC	CAIC	CICOMP	MD	CPU(sn)
40	5	2	98	98	98	93	2.40
	10	4	83	83	82	33	2.42
80	5	4	95	95	95	70	4.94
	10	8	81	81	79	12	5.00
100	5	5	91	91	90	55	6.36
	10	10	77	77	77	10	6.28
200	5	10	87	87	87	34	17.53
	10	20	34	34	34	0	17.54

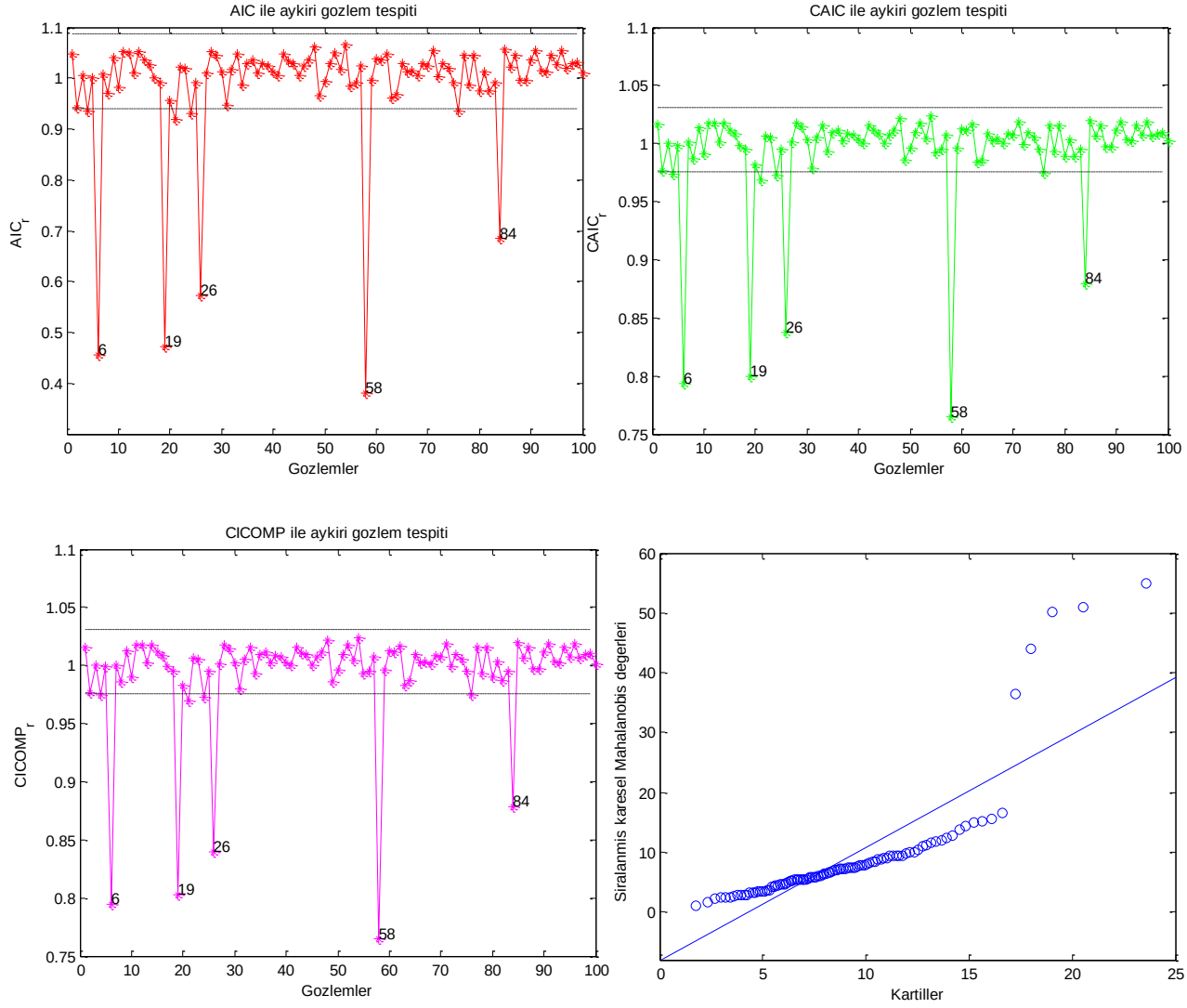
$\Sigma_q = \Sigma + 5aa^T$ *, İterasyon sayısı=100, n: Gözlem sayısı*

C_r *:Bulaşma oranı (Contamination rate) #of Out: Aykırı gözlem sayısı*

Tablo 3.13, 3.14 ve 3.15'de gösterilen sonuçlar, simülasyon protokolünün 100 defa tekrarlanması sonucunda, verideki aykırı gözlemlerin tamamını doğru şekilde yakalayabildikleri iterasyon sayılarıdır. Sonuçlar oldukça ilgi çekicidir. Farklı senaryolar ile aykırı gözlemlerin yapıları değiştirildiği için sonuçlar farklılık göstermektedir. Ancak tüm senaryolarda bilgi kriterlerinin, klasik yaklaşım olan Mahalanobis uzaklığına göre üstün başarı sergiledikleri açıktır. Özellikle senaryoların hepsinde, verideki aykırı gözlemlerin sayısı arttıkça, Mahalanobis uzaklığının performansı belirgin bir şekilde düşmektedir. Ancak bilgi kriterleri bu durumdan Mahalanobis uzaklığı kadar etkilenmemektedir.

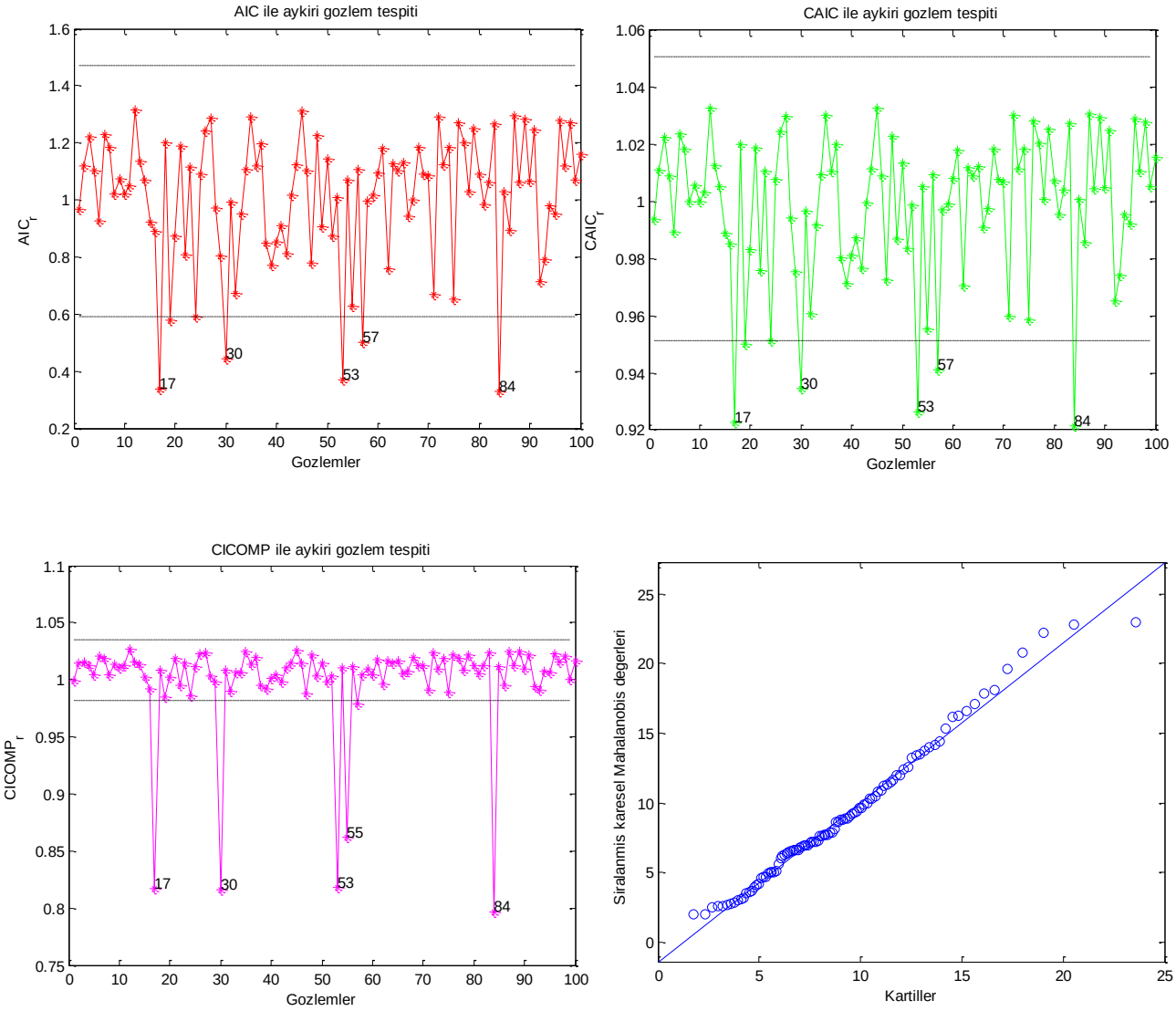
Kriterlerin performansları kendi aralarında karşılaştırılacak olursa, AIC tipi kriterler olan AIC ve CAIC tamamen aynı performansı göstermiştir. Senaryo-1 ve Senaryo-3 için, kriterler arası performans farklılıkları yok denecek kadar azdır. Ancak Senaryo-2'de CICOMP, AIC tipi kriterlere göre olağanüstü başarı sergilemiş ve tüm aykırı gözlemleri, sayısı ne olursa olsun doğru bir şekilde belirlemiştir. Bu nedenle uygulamada, performans olarak aykırı gözlemlerin sayısından ve yapısından etkilenmediği için CICOMP'ın kullanımı önerilebilir. Şekil 3.6, 3.7 ve 3.8'de her bir senaryo için tek bir iterasyondan elde edilen kontrol grafikleri verilmiştir.

Senaryo-1: $\Sigma_q = 8\Sigma$, aykırı gözlemler: 6- 19- 26- 58- 84



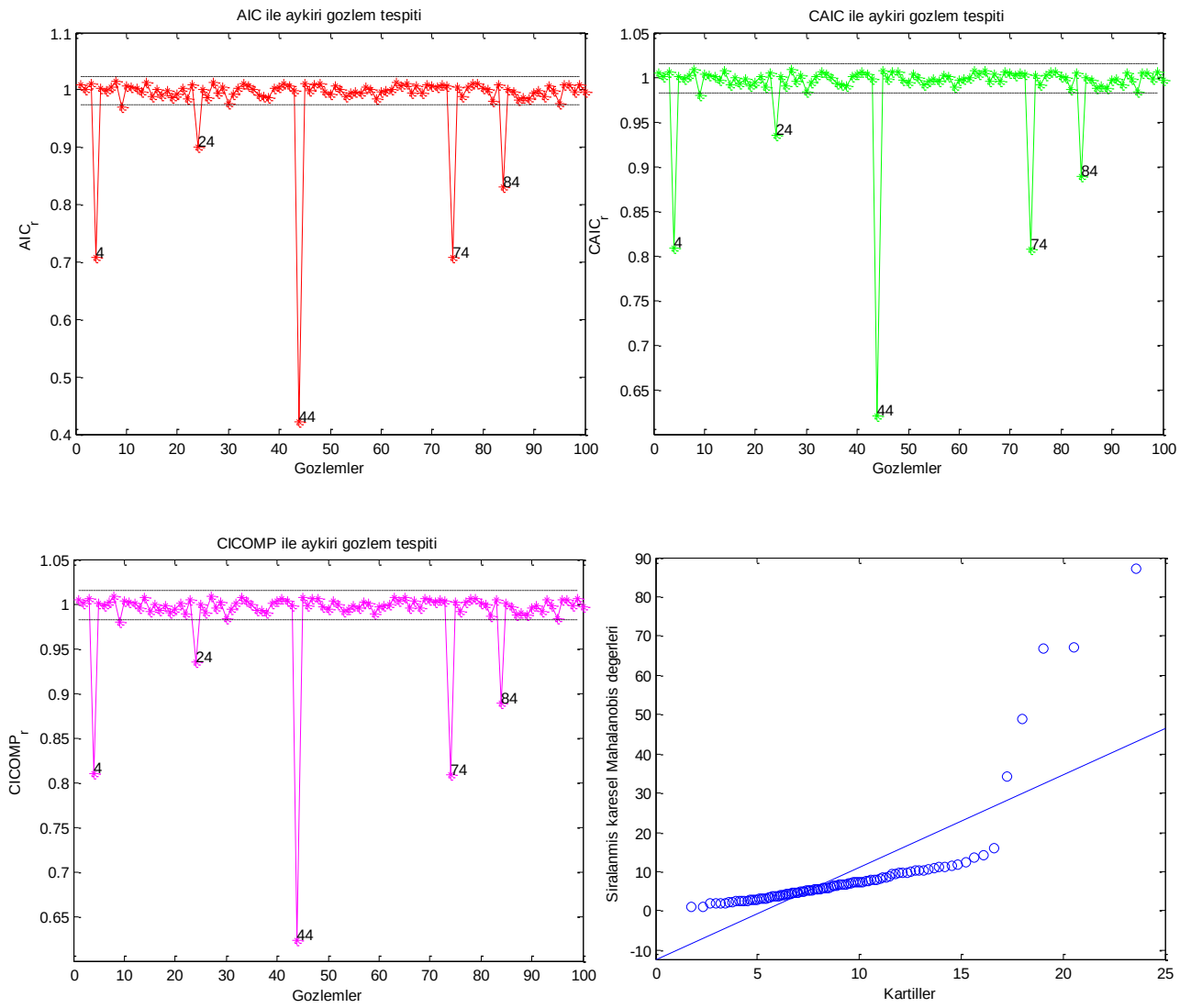
Şekil 3. 6: Senaryo-1'e ait aykırı gözlemlerin bilgi kriterleri ile tespiti

Senaryo-2: $\mu_q = 8\,[1]$, aykırı gözlemler: 17- 30- 53- 55- 84



Şekil 3. 7: Senaryo-2'ye ait aykırı gözlemlerin bilgi kriterleri ile tespiti

Senaryo-3: $\Sigma_q = \Sigma + 5aa^T$ ve $a = \frac{a_{rand}}{\|a_{rand}\|}$, $a_{rand} \sim \mathcal{B}(\frac{1}{2})$ için aykırı gözlemler: 4-24 –44-74-84



Şekil 3. 8: Senaryo-3'e ait aykırı gözlemlerin bilgi kriterleri ile tespiti

3.4 HOD ile Mikrodizilim Veri Analizi

Mikrodizilim verilerinde iki tür aykırı gözlem vardır. İlki, veride farklı bir sınıfa ait olan hatalı etiketlenmiş gözlemlerdir. Bu gözlemler normal doku olarak etiketlenen bir tümör dokusu gibi yanlış bir sınıfa atanan gözlemlerdir. Bu aykırı gözlemler kullanılan sınıflama prosedürü ile tespit edilebilirler. İkincisi ise veride herhangi bir sınıfa ait olmayan anomali gözlemlerdir. Bu aykırı gözlemlerin kaynağı daha çok belirsizdir, fakat keşfedilmemiş bir biyolojik sınıftan, zayıf sınıf tahmininden, deneysel hatalardan veya aşırı biyolojik çeşitlilikten kaynaklanabilir. Anomali gösteren bir gözlemin sınıfı belli olmadığı zaman, etiketlendiği sınıfın doğruluğuna tutarlı bir şekilde itiraz da edilemeyecektir. Örnek olarak, bir gözlem gerçekten tümör dokusu olabilir fakat onun ifade düzeyi diğer tümör dokularının ifade düzeylerinden oldukça farklı olabilir.

Mikrodizilim veri setlerinde aykırı gözlemlerin belirlenmesi ve ayıklanması, verinin geri kalanıyla karşılaştırıldığında tutarsız davrandıkları için önemlidir. Aykırı gözlem varlığında bir model uygulamak, parametre tahminlerini çarpık yapabileceği gibi yanlış çıkarsamalara da sebep verebilirler. Literatürde mikrodizilim gen ifade verilerinin analizleri ve yorumlanması ile ilgili olarak sınıflama ve gen seçilimi hakkında çok fazla çalışma bulunmaktadır. Ancak aynı şeyi gen ifade verilerinde aykırı gözlem tespiti için söylememiz mümkün değildir [Shieh ve Hung, 2009].

Tez kapsamında önerilen HOD yöntemi Bölüm 2.6' da anlatıldığı gibi aşağıdaki hesaplama adımlarına sahiptir.

Aşama-1: HDR ile yüksek boyutlu verinin boyutu indirgenir

Aşama-2: HDR sonucu elde edilen yeni bileşenler üzerinde her bir gözlem tek tek dışarıda bırakılarak $ICOMP_r$ oranları hesaplanır.

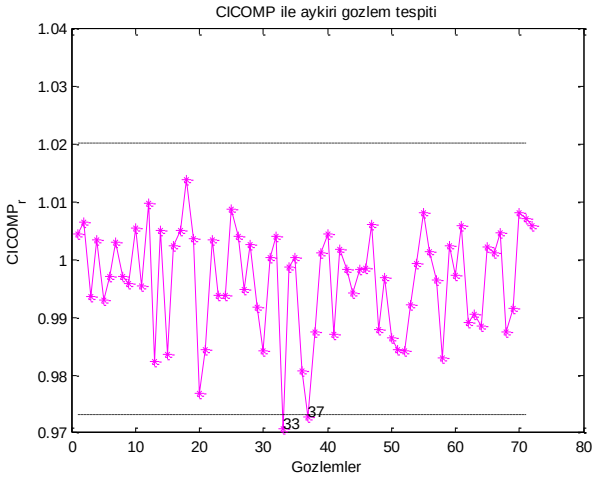
Aşama-3: $ICOMP_r$'nin değerleri 0 ve 1 arasında olduğu için Beta dağılımından geldiği varsayılarak, bir üst ve alt limit tanımlanır ve bu sınırlar dışında kalan gözlemler aykırı gözlem olarak tanımlanır.

Bu bölümde HOD yaklaşımı ile Bölüm 3.2'de boyutları indirgenen altı tane mikrodizilim veri seti analiz edilerek aykırı gözlemler tespit edilmiştir. Simulasyon sonuçları dikkate alındığında CICOMP ile farklı türlerdeki aykırı gözlemler başarılı bir şekilde tespit edilebildiği için, Tablo 3.16'daki sonuçlar CICOMP tarafından tespit edilen aykırı gözlemleri vermektedir. Ayrıca HDR ile boyut indirgendiği için, Mahalanobis uzaklığı ile de aykırı gözlem tespiti yapılmış ve sonuçlar birlikte verilmiştir.

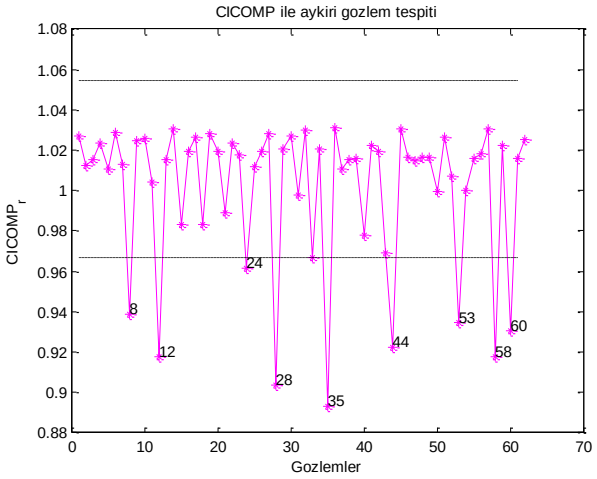
Tablo 3. 16: Mikrodizilim veri setlerinde HOD ile tespit edilen aykırı gözlemler

Veri seti	Gözlem sayısı	CICOMP	MD
Lösemi	72	33,37	21
Kolon	62	8,12,24,28,35,44,53,58,60	35,40
Prostat	102	3,7,9,10,14,16,17,19,22,29,42 64,68,92	42,92
Lenfoma	62	54,55,56,57,58,59	-
SRBCT	63	10,55,57,58,59,60,62,63	60,63
Beyin	42	32	30,37,39,41

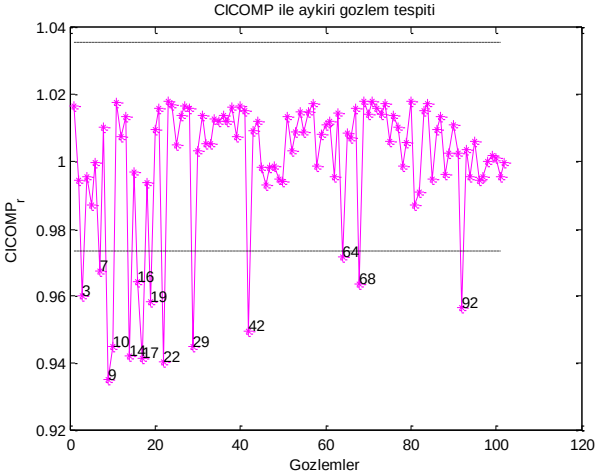
Bu sonuçlara ilişkin kontrol grafikleri Şekil 3.9'da verilmiştir.



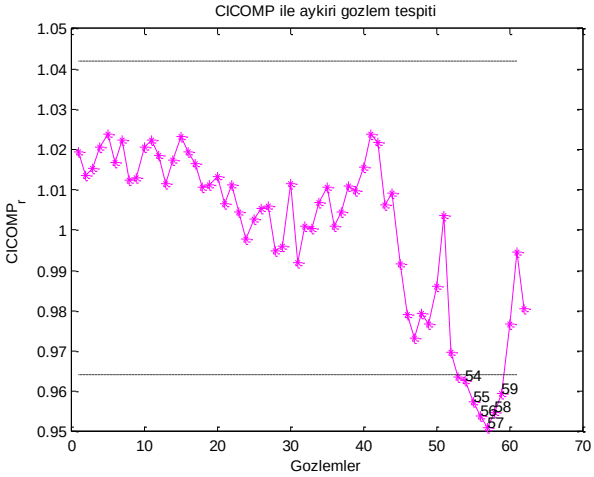
Lösemi



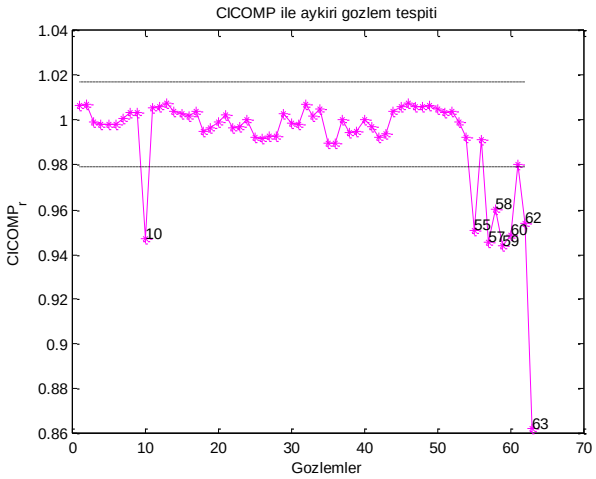
Kolon



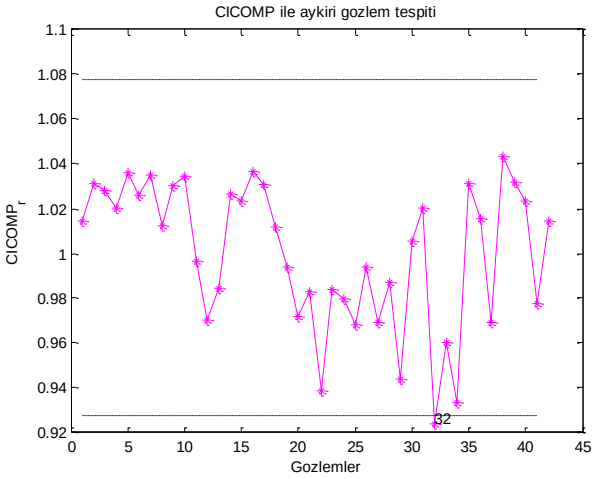
Prostat



Lenfoma



SRBCT



Beyin

Şekil 3. 9: HOD ile mikrodizilim veri analizi

Aykırı gözlemlerin tespiti istatistiksel analizlerde önemli bir rol oynamaktadır. Klasik istatistiksel modeller aykırı gözlemler ihtiva eden veri setlerine körü körüne uygulanırsa, sonuçlar en iyi ihtimalle yanıltıcı olacaktır. Buna ek olarak, aykırı gözlem tespiti için ortalama vektörü ve kovaryans matrisine dayalı klasik araçlar maskeleme ve baskılama etkilerinden dolayı tüm aykırı gözlemleri nadiren tespit edebilmektedirler.

Bu çalışmada, PPCA ve bilgi kriterlerinin birlikte uygulanması sonucu önerilen HOD yöntemi ile etkin ve hızlı bir aykırı gözlem tespit etme tekniği geliştirilmiştir. Önerilen yöntemin performansı çeşitli simüle edilmiş veri setleri üzerinde ve yüksek boyutlu gen ifade verileri üzerinde gösterilmiştir. Elde edilen sonuçlara göre, HOD yönteminin yüksek boyutlu gen veri setlerindeki aykırı gözlemleri tespit etmek için kullanımı önerilebilir.

4. SONUÇ VE TARTIŞMA

Bu çalışmada, yapısal olarak aşırı derecede küçük örneklem problemine sahip veri setleri için en temel problemler olan boyut indirgeme ve aykırı gözlemlerin tespiti için literatürde ilk defa, bilgi karmaşıklığı kriteri ICOMP'ın kullanımı ile Hibrit Boyut İndirgeme (Hybrid Dimention Reduction-HDR) ve Hibrit Aykırı Gözlem Tespiti (Hybrid Outlier Detection-HOD) yöntemleri önerilmiştir. Yöntemlerin performans değerlendirmesi hem simülasyon çalışmaları hem de mikrodizilim veri setleri üzerinde gösterilmiştir.

Literatürde bugüne kadar, yüksek boyutlu mikrodizilim gen verilerinde var olan bilgiyi ortaya çıkarabilmek, önemsiz ve ilgisiz olan değişkenleri ayıklayabilmek ve sonuç olarak minumum hata oranıyla sınıflama yapabilmek için bazı yaklaşımlar önerilmiştir. Bu yaklaşımlar incelendiğinde, bazılarının sınıflamadan önce ilk aşama olarak gen seçilimi yapmaya çalışarak boyut indirgedikleri bazılarının ise verinin ortogonal yeni eksenler üzerine dönüşümünü yapan PCA gibi yöntemleri kullanarak boyut indirgedikleri tespit edilmiştir. Gen seçilim yöntemleri, genlerin tamamının bir ölçek dahilinde skorlanmasına ve daha sonra en iyi skora sahip olan genlerin seçilimine dayanmaktadır. Bu amaçla t-testi yaklaşımı [Hedenfalk ve ark., 2001] veya Wilcoxon's rank toplam istatikleri [Dettling ve Bühlmann, 2003] gibi farklı yaklaşımlar önerilmiştir. Bu yaklaşımlar genellikle sınıflama üzerinde etkisi olabilecek tek değişkenli gen seçilimi yapabilmek için kullanılır. Ancak bir veri niteliği tek başına performansa çok az bir etki yaratırken, diğer niteliklerle birleştiğinde tüm veri kümesini temsil edecek bir alt grup oluşturabilir [Lu ve Han, 2003]. Dolayısıyla genleri tek tek ele almak bu açıdan bakıldığında problemli olabilir. Genlerin sayıları binler hatta onbinlerle ifade edildiği için, çoklu karşılaştırma yapabilmenin imkansızlığı da açıkça bellidir. Sonuç olarak gen seçilimi için, genler arası etkileşimleri göz ardı etmeyen bir ölçü kullanmak daha avantajlı olabilecektir.

Boyut indirgeme bağlamında, bu konuda yapılan bazı çalışmalarda örnek kovaryans matrisinin singülerlik probleminin ve bunun neticesinde ortaya çıkan negatif özdeğer probleminin çözümünün yapılmadığı hatta bu problemle karşılaşmamak için n gözlem sayısına bağlı olarak sadece pozitif özdeğerlerin dikkate alındığı gözlenmiştir. Bu özdeğerlerin en büyük varyansa sahip oldukları, sıfır ve negatif çıkanların ise önemsiz bilgi taşıdıkları için ihmal edilebileceği belirtilmiştir [Filzmoser ve ark., 2008]. Halbuki bu

durum, Bölüm 2.1.3'de de belirtildiği gibi, pozitif yarı tanımlı matris olan kovaryans matrislerinin doğasına aykırı bir durumdur. Ghosh (2002)'de mikrodizilim veri setlerinde SVD ayrışımını kullanmayı ve buradan elde edilecek pozitif tekil değerler ile PCA yapılmasını tavsiye etmiştir. Kovaryans matrisinin rankı kadar pozitif tekil değer elde edileceği için, rank probleminin üstesinden gelinmediği ve diğer tekil değerlerin ihmal edilerek çözüm yapıldığı aşikardır. Nguyen ve Rocke (2002)'de kısmi en küçük kareler (Partial Least Square-PLS) yöntemini önerirken aynı şekilde matris rankını dikkate almışlardır. Chiaromon ve Martinelli (2002)'de dilimlenmiş ters regresyon (Sliced Inverse Regression-SIR) yöntemini önerirken veri matrisinin rankına bağlı olarak yapısal uzay diye adlandırdıkları bir alt kümeyi dikkate almışlardır. Bu yöntemlerin performans karşılaştırması Dai ve ark. tarafından (2006)'da yapılmıştır. Bu çalışmada ise, random örnekleme yöntemi ile gen seçilimi yapıldıktan sonra yöntemler karşılaştırılmıştır. Elde edilen alt kümenin optimal alt küme olmasının önemli olmadığı, sadece yöntemlerin performans değerlendirmesinin yapıldığı belirtilmiştir. PPCA yöntemi, Oba ve ark. tarafından (2003)'de kayıp gözlemlerin olduğu gen ifade verilerinde bu gözlemlerin tahmini için önerilmiştir. Ancak bu çalışmada da tüm gen verisi değil, alt kümeler kullanılarak analiz yapılmıştır. Tahmin edilen kayıp gözlemlerin analiz sonuçlarına etkisinin olup olmadığı incelenmiştir. Scholz ve ark. (2005)'de yapay sinir ağları tabanlı lineer olmayan PCA kullanımını önermişlerdir. Yapılan çalışmada en yüksek varyans oranına sahip olan genler dikkate alınmıştır. Liu ve ark. (2008)'de negatif değerler içermeyen veri setlerinin analizi için önerilen Non-negative matrix factorization-NMF yöntemini mikrodizilim verilerinin boyutunu indirgemek için önermişler ve bu yaklaşımı PCA ile karşılaştırmışlardır. PCA için boyutsallık probleminden kaynaklı problemlerin olduğunu vurgulamalarına rağmen uygulamada klasik PCA kullanmışlardır. Hatta elde edilen grafiklerde bile gözlem sayısı kadar PC'nin olduğu dikkat çekmektedir. 11 tane gen ifade verisinden elde edilen sonuçlar incelendiğinde, benzer kullanılan gen verilerinden Prostat için PCA+k-means ile %15, NMF+k-means ile %13, SRBCT verisi için PCA+k-means ile %45, NMF+k-means ile %28 hatalı sınıflama oranlarını elde etmişlerdir. Bu oranlar Tablo 3.12'deki sonuçlar ile karşılaştırıldığında HDR'nin başarısı bir kez daha ortaya çıkmaktadır. Shi ve Luo (2010)'da gen verilerini görsel olarak sunabilmek için en fazla 2 ve 3 boyut üzerinde çalıştıkları lineer olmayan bir boyut indirgeme yöntemini önermişlerdir. PCA için tek eleştiri genlerin lineer kombinasyonunu içermesi şeklinde yapılmış ve uygulamada yine klasik PCA kullanılmıştır. Beş kanser verisine yapılan analizlerin sonuçları incelendiğinde, benzer kullanılan gen

verilerinden Lösemi ve SRBCT verileri için en iyi %52.2'lik hatalı sınıflama oranı elde etmişlerdir. Bu sonuçlar tekrar HDR'nin başarısını vurgulamak için dikkat çekicidir. Rezghi ve Obulkasim (2014)'de gürültüsüz PCA yapısını önermişlerdir. PCA analizi gürültü içeren veri setlerine uygulandığı zaman, bunların lineer birleşimleri olan PC skorlarına da bu gürültünün bulaşacağı ve bunun da PC skorları bir sınıflandırma prosedürü için giriş vektörleri olarak düşünüldüğü zaman performans kaybına sebep olacağını vurgulamışlardır. Bu nedenle PCA için SVD çözümünde, gürültünün cezalandırılması için bir ceza terimi ekleyerek noisy-free PCA önermişlerdir. Ancak çözüm SVD'ye dayanmakta ve tekrar pozitif tekil değerler dikkate alınmaktadır.

Buraya kadar örneklendirilen çalışmalardan, genel olarak gen seçiliminin yapıldığı ya da matris rankı kadar genle çalışıldığı anlaşılmaktadır. Dolayısıyla bildiğimiz kadarıyla, aşırı derecede küçük örneklem probleminin olduğu gen verilerinde negatif özdeğer probleminin üstesinden gelerek boyut indirgemenin yapılabildiği bir çalışmaya rastlanmamıştır. Bu anlamda tez kapsamında önerilen Hibrit Boyut İndirgeme-Hybrid Dimension Reduction-HDR literatüre önemli bir yenilik katmıştır.

Aykırı gözlem tespiti bağlamında, veri setinin satırlarında (gözlemler) veya sütunlarında (genler) aykırılığın aranması şeklinde farklı yaklaşımlar önerilmiştir. Esasen aykırı gözlemler terimiyle vurgulanmak istenen farklı olan gözlemlerin tespit edilmesi olarak yorumlanabileceği için, tez kapsamında da bu problem dikkate alınmıştır. İkinci durum literatürde farklı gen ifadelerinin tespiti olarak adlandırılmıştır. İlk olarak Tomlins ve ark. (2005)'de cancer outlier profile analysis-COPA'yı ardından Tibshirani ve Hastie (2006)'da outlier sum-OS istatistiğini ve sonra Wu (2007)'de outlier robust t-statistic (ORT)'yi önermişlerdir. Her ne kadar outlier kelimesi bu çalışmalarda kullanılmış olsa da bunların sütunlarda bulunan genler arasından farklı yapıya sahip olanları tespit etmek için önerildikleri hatırlatılmalıdır.

Bu hatırlatmanın ardından gözlemler arasından aykırı olanların tespiti yani ilk durum için, yüksek boyutlu sınırlı örneklem problemine sahip veri setlerinde karşılaşılan zorluklar, uzaklık, yoğunluk, en yakın komşuluk tabanlı yaklaşımların yüksek boyutlarda neden anlamsızlaştığı Bölüm 2.5.2'de tartışılmıştı. Benzer eleştiriler, Aggarwal ve Yu (2002)'de de yapılmıştır. Aynı çalışmada yüksek boyutlu veriler için evrimsel bir aykırı gözlem tespit etme metodu önerilmiştir. Genetik algoritma ile önerilen yaklaşımın, gerçek aykırı

gözlemleri %98 oranında tespit ettiğini, %2 oranında aykırı olan gözlemleri normal gözlem olarak belirlediğini, %9.5 oranında ise, aykırı olmayan noktalarını aykırı gözlem olarak belirlediğini bildirmişlerdir. Yüksek boyutlardaki problemlerden dolayı bazı çalışmalarda ön hazırlık aşaması olarak boyut indirgemenin yapılması ve daha sonra indirgenmiş uzay üzerinde aykırı gözlem tespiti yapılması önerilmiştir. Her ne kadar, tez kapsamında önerilen HOD ile fikir anlamında benzerlik taşısa da, açıkça bellidir ki bu düşüncesinin başarısı hem uygulanan boyut indirgeme prosedürünün doğru bir şekilde uygulanabilmiş olmasına hem de kullanılacak aykırı gözlem tespiti metodunun yapısına bağlıdır. Bu bağlamda, Filzmoser ve ark. (2008)'de indirgenmiş uzay üzerinde çalışmanın faydalarından bahsetmişler ve PCOut yöntemini önermişlerdir. Saçılmış aykırı gözlemlerin farklı bir kovaryans matrisi yapısından geldiklerini, lokal aykırı gözlemlerin ise, farklı bir lokasyon parametresine sahip olduklarını vurgulayan yazarlar, önerdikleri yaklaşım ile iki aşamada, her iki tür aykırı gözlemi tespit edebildiklerini öne sürmüşlerdir. İlginç olan ise bu yöntemin düşük boyutlarda zayıf performans göstermesine karşılık yüksek boyutlarda iyi performans gösterdiğini ifade etmeleridir. Çalışmaya göre p=50 boyut için aykırı gözlemler hatalı negatif ve hatalı poziif olmak üzere toplamda %56 yanlış belirlenirken, p=2000 boyutlarda bu oran %3 olmaktadır. Lösemi veri seti için PCOut yöntemini genler için uygulamışlar ve 7129 genden 2609 tanesini aykırı olarak tespit etmişlerdir. Daha sonra bu sayıyı aşırı aykırı olanlar olarak 296 taneye filtrelemişlerdir. Bu çalışma da ikinci grup problem için örnek olarak verilebilir. Oh ve Gao (2009)'da kernel tabanlı bir aykırı gözlem tespit metodu KLOD'u önermişlerdir. Mahalonobis uzaklığı için karşılaştırmanın da yapıldığı bu çalışmada, hem Lösemi hem de Kolon kanser verisi üzerinde uygulama yapılmıştır. KLOD ile Lösemi veri seti için 9, Kolon için 6 aykırı gözlem tespit edilirken Mahalonobis ile Lösemi için 14, Kolon için 3 aykırı gözlem tespit edildiği bildirilmiştir. Fakat bunların hangi gözlemler oldukları rapor edilmemiştir. Debruyne (2009)'da Destek Vektör Makinesi – Support Vector Machine (SVM) ile sınıflama tabanlı bir aykırı gözlem haritalama tekniği önermişdir. Kolon ve Lösemi verileri üzerine uygulama yapılan çalışmada, Lösemi verisi için aykırı gözlem tespit edilmediği ifade edilirken, Kolon verisi için Alon ve ark. (1999)'da aykırı gözlem olarak etiketledikleri gözlemlerin aynılarının tespit edildiği belirtilmiştir.

Buraya kadar yapılan literatür taramasının sonucu olarak, Shieh ve Hung (2009)'da belirttikleri gibi, mikrodizilim gen ifade verilerinin analizleri ve yorumlanması ile ilgili olarak sınıflama ve gen seçilimi hakkında çok fazla çalışma bulunmasına rağmen aynı şeyi

gen ifade verilerinde aykırı gözlem tespiti için söylememiz mümkün değildir. Bu özellikle yüksek boyutlu veri setlerinin yapısında var olan boyutsallık probleminden kaynaklanmaktadır. Esasen aykırı gözlemlerin ne literatürde genel geçer olarak kabul edilen bir tanımı ne de aynı şekilde tespit etme yöntemi olarak kabul görmüş bir yöntemi vardır. Düşük boyutlu veri setlerinde en temel klasik yaklaşım olan Mahalonobis uzaklığının bile aykırı gözlemleri tespit etmede ki performansının sorgulandığı ve bunun neticesinde sağlam yöntemlerin önerildiği düşünülecek olursa, bu problemlerin üstüne yüksek boyutlarda bir de boyutsallık probleminin eklenmesi, bu konuda rapor edilen çalışmalarının nispeten daha az olmasının bir sebebi olarak gösterilebilir.

Çoklu aykırı gözlemleri ve alt grupları tanımlamak için kullanılan metotlar, homojensizlik yüzünden bozulmayacak bir metrik kurmanın zorluğu ile ilişkilidir. Sağlam metotlar, homojenlik varsayımında esneme yapabilir. Fakat bu yöntemler de yaygın bir şekilde kabul görmemiştir. Bu durum onların yüksek boyutlardaki veri setleri için hesaplama açısından uygun olmamalarından kaynaklanır. Oysaki bilgi kriterlerinin kullanımı ile böyle bir metrik kurmanın zorluğu ortadan kalkabilir. Mantıksal olarak özellikle verideki kovaryans yapısının kompleksliğini cezalandıran ICOMP tipi kriterler, zaten homojensizlik ölçümü yaptıkları için bu problemin üstesinden gelebilir. Bu noktada, tez kapsamında önerilen HOD yaklaşımı, aykırı gözlemlerin tespiti için literatüre farklı bir bakış açısı kazandırmıştır. Özellikle HDR ile indirgenmiş uzay üzerinde çalışmasından dolayı, çok yüksek boyutlu veri setleri için uygulanabilir ve hesaplama açısından hızlı ve etkindir.

Sonuç olarak hem yapılan simülasyon çalışmalarından elde edilen benzetim verileri üzerinden hem de mikrodizilim veri setleri üzerinden elde edilen sonuçlar, yöntemlerin başarılarını açık bir şekilde ortaya koymaktadır. Uygulama için altı tane mikrodizilim veri seti kullanılmış olsa da, HDR ve HOD teknikleri bu veri setleri ile benzer yapıya sahip, aşırı derecede küçük örneklem problemi olan veri setlerinin hepsine uygulanabilir.

5. ÖNERİLER

Bu tez çalışması, ileride yapılacak çalışmalar için bir mihenk taşı vazifesi görmektedir. Sonuçlar her nekadar tatmin edici olsa da, önerilen yöntemlerin performanslarını artırabilmek veya genelleştirebilmek için aşağıdaki öneriler uygulanabilir.

- İlk aşama olarak kullanılan PCA ileri ki çalışamalarda non-Gaussian PCA, kernel yoğunluklu PPCA, olasılıklı bağımsız bileşenler analizi- Probabilistic Independent Component Analysis (PICA), denetimsiz karma model kümeleme analizi- Unsupervised mixture cluster analysis, yöntemlerine uyarlanması öngörülmektedir.
- ICOMP'ın uygunluk fonksiyonu olarak kullanılacağı bir genetik algoritma ve HDR'nin birlikte kullanımı ile, genler arasından en mükemmel sonucu ortaya çıkaracak alt kümenin belirlenmesi sağlanabilir. Böylelikle dönüştürülmüş verinin, gerçek uzayda hangi genlere tekabül ettiği belirlenebilir.
- Sınıflandırma noktasında HDR'nin etkinliği, farklı sınıflandırma prosedürlerinin kullanımı ile daha çok ortaya çıkabilir.

6. KAYNAKLAR

Abdi, H. 2007. Singular value decomposition and generalized singular value decomposition. *Encyclopedia of Measurements and Statistics.* Neil Salking(ed). Sage. USA

Aelst, S.V. ve Rousseeuw, P.J. 2009. Minumum volüme ellipsoid. Advanced Review. *WIREs Comp. Stat.* **1:** 171-182

Aggarwal, C., C. ve Yu,P.,S. 2001. Outlier detection for high dimensional data. *ACM SIGMOD record.* **30(2):**37-46

Akaike, H. 1973. Information theory and extension of the maximum likelihood principle. 2nd International Symposium on Information Theory. Budapest: Academiai Kiado. 267-281

Akaike,H. 1974. A new look at the statistical model identification. *IEEE Transaction and Automatic Control.* AC-19:719-723

Akaike, H. 1981. Likelihood of a model and information criteria. *Journal of Econometrics.***16**:3-14

Akbilgiç, O. 2011. Hibrit Radyal Tabanlı Fonksiyon Ağları ile Değişken Seçimi ve Tahminleme: Menkul Kıymet Yatırım Kararlarına İlişkin Bir Uygulama. Doktora tezi. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü. İstanbul

Alizadeh, A.,A. et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.***403:**503-511.

Alon et al. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA.* **96:**6745-6750

Alpar, R.,C. 2011. Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. Detay Yayıncılık. 3. Baskı. Ankara

Anderson, T.,W. ve Rubin H. 1956. Statistical inference in factor analysis. J Neyman (ed). Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability. Volume 5. U. Cal, Berkeley. 111-150

Anderson, T.W. 1984. Introduction to Multivariate Statistical Analysis. 2. Baskı. John Wiley and Sons. New York.

Barnett, V. ve Lewis, T. 1994. Outliers in Statistical Data. 2. Baskı. John Wiley and Sons. New York.

Bellman, R. 1961. Adaptive Control Processes: A Guided Tour. Princeton University Press.

Beltrami, E. 1873. Sulle Funzioni Bilineari. Giornale di Matematiche ad USO Degli Studenti Delle Universita. **11**: 98-106

Ben Gal I. 2005. Outlier Detection. In Data Mining and Knowledge Discovery Handbook. Oded Maiman and Liar Rakoch (eds): 131-146

Bhansali, R.,J. ve Downham, D.,Y. 1977. Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika.***64:**547-551

Bickel, P.,J. ve Li, B. 2006. Regularization in statistics. *Sociedad de Estadistica e Investigacion operativa. Test.* **15(2)**: 271-344

Billor et.al. 2000. BACON: blocked adaptive computationally efficient outlier nominators. *Computational Statistics and Data Analysis.* **34:**279-298

Boyce, D., E., Farhi, A. ve Weischedel, R. 1974. Optimal Subset Selection: Multiple Regression, Interdepence and optimal network algortihms. Springer Verlag.New York

Bozdogan, H. 1987. Model selection and Akaike's İnformation Criterion (AIC): the general theory and its analytical extensions. *Psychometrika.* **52(3):**345-370

Bozdogan, H. 1988. ICOMP: A new model selection criterion. *Classification and Related Methods of Data Analysis.* 599-608

Bozdogan, H. 1990. On the information based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics: Theory and Methods.***1:**221-278

Bozdogan, H. 1994. Choosing the number of clusters, subset selecyion of variables and outlier detection in the standard mixture model cluster analysis. Invited paper in *New Approaches in Classification and Data Analysis.* Springer Verlag. New York

Bozdogan, H. ve Haughton, D.,M.,A. 1998. Information complexity criteria for regression models. *Computational Statistics and Data Analysis.* **28**: 51-76

Bozdogan,H. 2000. Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology.***44**: 62-91

Bozdogan, H. 2004. Intelligent Statistical Data Mining with Information Complexity and Genetic Algortihm. In Statsitical Data Mining and Knowledge Discovery. H. Bozdogan (ed). Chapman and Hall/CRC. Florida

Bozdogan, H. 2010. A new class of information complexity (ICOMP) criteria with an application to customer profiling and segmentation. Invited paper. In *Istanbul University Journal of the School Business Administration.* **39(2)**:370-398

Bozdogan, H. ve Howe, J.,A. 2012. Misspecified multivariate regression models using the genetic algorithm and information complexity as the fitness function. *European Journal of Pure and Applied Mathematics.* **5(2)**: 211-249

Bozdogan, H., 2013. İkili görüşme. University of Tennesse. Knoxville/USA

Bozdogan, H. 2014. Information Complexity and Multivariate Learning in High Dimensions with Applications in Data Mining. Forthcoming book.

Bryant, E., H. ve Atchley, W.,R. 1975. Multivariate Statistical Methods within Groups Covariation. Dowden,Hutchins and Roos Publishing. New York

Burnham, K., P. ve Anderson, D.,R. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research.* **28:**111-119

Caussinus and Roiz. 1990. Interesting projections of multidimensional data by means of generalized principal component analyses. *Comp. Stat.* 121-126

Chandola, V. et. al. 2009. Anomaly detection: A survey. *Journal ACM Computing Surveys-CSUR.* **41(3)**

Chen, M. 1976. Estimation of covariance matrices under a quadratic loss function. Research Report S-46. Department os Mathematics. SUNY at Albany.

Chen, Y. 2010. Robust shrinkage estimation of high dimensional covariance matrices. *IEEE Workshop on Sensor Array and Multichannel Signal Processing (SAM).*

Chiaromonte, F. ve Martinelli, J. 2002. Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences.***176:**123-144

Chu, F. ve Wang, L. 2005. Application of support vector machines to cancer classification with microarray data. *International Journal of Neural Systems.* **15(6):**475-484

Cooley, W.,W. ve Lohnes, P.,R. 1971. Multivariate Data Analysis. John Wiley and Sons. New York

Cunningham, P. 2007. Dimension Reduction. Technical Report.UCD-CSI-2007-7. University College Dublin.

Dai, J.J., Lieu, L. ve Rocke, D. 2006. Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetic and Molecular Biology.* **5(1)**

Davies, L. ve Gather, U. 1993. The identification of multiple outliers. *Journal of American Statistical Association*. **88(423)**:782-92

Debruyne, M. 2009. An outlier map for support vector machine classification. *The Annals of Applied Statistics.* **3(4):** 1566-1580

Deniz, A., E. 2007. Yapısal Eşitlik Modellerinde Bilgi Kriterleri. Doktora tezi. Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü. İstanbul

Dettling and Bühlman. 2003. Boosting for tumor classification with gene expression data. *Bioinformatics.* **19(9):** 1061-1069

Donoho, D.,L. 2000. High dimensional data analysis: The curses and blessings of dimensionality. statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf. 23 Ocak 2013.

Draper, N.R. ve Smith, H. 1966. Applied Regression Analysis. Wiley. New York.

Duda, R.O., Hart, P.E. ve Stark, D.G. 2001. Pattern Classification. 2. Baskı. John Wiley and Sons. New York

Erbaş, Ü. 2010. Entropi İlkelerinin Boyut İndirgeme Uygulamaları. Doktora tezi. Marmara Üniversitesi Sosyal Bilimler Enstitüsü. İstanbul.

Escalante, H.J. 2005. A comparison of outlier detection algortihms for machine learning. Citeseer.

Fiebig, D., G. 1984. On the maximum entropy approach to undersized samples. *Applied Mathematics and Computation.***14:**301-312

Filzmoser et. al. 2008. Outlier identifaciton in high dimensions. *Computational Statistics and Data Analysis.* **52(3):**1694-1711

Fisher, R. 1936. IRIS data set. archive.ics.uci.edu/ml/datasets/Iris. 10 Ocak 2014.

Frisch, R. 1929. Correlation and scatter in statistical variables. *Nordic Stat J.***8**:36-102

Fritsch et.al. 2012. Detecting outliers in high dimensional neuroimaging datasets with robust covariance estimators. *Medical Image Analysis.* **16:**1359-1370

Fukunaga, K. 1990. Introduction to Statistical Pattern Recognition. 2. Baskı. Academic Press. Boston.

Girshick, M., A. 1936. Principal components. *Journal of American Statistical Association.* **31(195)**:519-528

Girshick, M., A. 1939. On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics.* **10(3):**203-224

Ghosh,D. 2002. Singular value decomposition regression models for classification of tumors from microarray experiments. *Pacific Symposium on Biocomputing*. Kauai,Hawai

Golub, T.,R. et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.***286:**531-537

Gower, J.,C. 1966. Some distance of latent root and vector methods used in multivariate analysis. *Biometrika.* **53 (3/4):**325-338

Hadamard, J. 1902. Sur les problemes aux derivees partielles et leur signification physique. *Princeton University Bulletin.* 49-52

Hadi, A.S. 1992. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series-B.* **54**: 761-771

Hadi, A.S. 1994. A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society. Series-B.* **56(2)**

Haff, L.,R. 1980. Emprical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics.***8(3)**:586-597

Hall,P. ,Marron, J.S. ve Neeman, A. 2005. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society. Series-B.* **67(3):** 427-444

Hannan, E.,J. ve Quinn, B., G. 1979. The determination of the order of an autoregression. *Journal of the Royal Society Series-B.***41:** 190-195

Hardin and Rocke. 2005. The distribution of robust distances. *Journal of Computational and Graphical Statistics.* **14(4)**: 928-946

Harris, C., J. 1978. An Information Theoretic Approach to Estimation. M.,J. Gregson (ed.) In Recent Theoretical Developments in Control. Academic Press. London.

Hawkins, D.M. 1980. Identification of Outliers. Chapman and Hall.

Hedenfalk et. al. 2001. Gene expression profiles in hereditary breast cancer. *New Eng. Jour. Medicine.* **344:** 539-548

Hocking, R.,R. 1983. Developments in linear regression methodology. *Technometrics.* **25**:219-230

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology.* **24(6):**417-441

Hubert, M. ve Debruyne, M. 2009. Minumum covariance determinant. Advanced Review. *WIREs Comp. Stat.* **2:** 36-43

İdil, N.,B. 2009. Gen İfade Verileri ve İşlemsel Kanser Sınıflandırılması. Yüksek Lisans Tezi. Başkent Üniveristesi Fen Bilimleri Enstitüsü.

Jain, A.K., Duin, R.P.W. ve Mao, J. 2000. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **22(1):**4-37

James, M. 1985. Classificaiton Algortihms. William Collins Sons. London.

Jeffer, J.,N.,R. 1967. Two case studies in the application of principal component analysis. *Journal of Royal Statistical Society Series-C.* **16(3)**: 225-236

Johnson, R.A. ve Wichern, D.W. 1998. Applied Multivariate Statistical Analysis. 4. Baskı. Prentice Hall. New Jersey.

Joliffee, I.,T. 2002. Principal Component Analysis. Springer Series in Statistics. 2. Baskı. New York.

Jordan, C. 1874. Memoirre sur les formes bilinearies. *Journal de Mathematiques Pures et Appliques.*35-54

Khan, J. et al. 2001. Classification and diagnostic prediction of cancer using gene expression profiling and artifical neural networks. *Nat. Med.* **6:**673-679

Kullback, S. ve Leibler, R.,A. 1951. On information and sufficiency. *The Annals of Mathematical Sciences.***22(1)**:79-86

Kullback, S. 1968. Information Theory and Statistics. Dover. New York.

Kumar, D., Rai, C., S. ve Kumar, S. 2008. Principal component analysis for data compression and face recognition. *Infocomp Journal of Computer Sciences.* **7(4):** 48-59

Lawley, D.,N. 1953. A modified method of estimation in factor analysis and some large sample results. *In Uppsala Symposium on Psychological Factor Analysis.* Number-3 in Nordisk Psykologi Monograph Series. 35-42. Uppsala Almqvist and Wiksell

Ledoit, O. ve Wolf, M. 2004. A well conditioned estimator for large dimensional covariance matrices. *Journal of Multivariate Analysis.* **88**:365-411

Linhart, H. ve Zucchini, W. 1986. Finite sample selection criteria for multinomial models. *Statistiche Hefte***. 27**:173-178

Liu, Y. 2007. Robust and Misspecification Resistant Model Selection in Regression Models with Information Complexity and Genetic Algorithm. Doktora tezi. University of Tennessee. Knoxville/USA.

Liu,W., Yuan, K. ve Ye, D. 2008. Reducing microarray data via non-negative matrix factorization for visualization and clustering analysis. *Journal of Biomedical Informatics.* **41:**602-606

Lu, Y. ve Han, J. 2003. Cancer classification using gene expression data. *Information Systems.***28:**243-268

Miguel, A. ve Carreira, P. 1997. A review of dimension reduction techniques. Technical Reports. CS-96-09. University of Sheffiled.

Montel, N. 1970. Why stepdown procedures in variable selection. *Technometrics.* **12**:591-612

Nguyen, D.V. ve Rocke, D.M. 2002. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics.* **18(1):**39-50

Nyamundada, G., Brennon, L. ve Gormley, I., C. 2010. Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics.* **11:** 571-582

Oba, S. et al. 2003. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics.***19(16):**2088-2096

Oh, J.H. ve Gao, J. 2008. A kernel based approach for detecting outliers of high dimensional biological data. *BMC Bioinformatics.* **10**

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine.***2:**559-572

Penny and Joliffee. 2001. A comparison of multivariate outlier detection methods for clinical laboratory safety data. *The Statistician.* **50(3):**295-308

Press, S. 1975. Estimation of a normal covariance matrix. Technical Report. University of British Columbia.

Pomeroy, S.,L. et al. 2002. Prediction of central nervous system embryonal tumor outcome based on gene expression. *Nature.* **415:** 436-442

Pourahmadi, M. 2012. Covariance estimation: The GLM and regularization perspectives. Statistical Science.

Rao,R. 1964. The use and interpretation of principal component analysis in applied research. *The Indian Journal of Statistics Series-A.***26(4):** 1961-2002

Rezghi, M. ve Obulkasim A. 2014. Noise free principal component analysis: A efficient dimension reduction technique for high dimensional molecular data. *Expert Systems with Applications.* **41:**7797-7804

Rissanen, J. 1978. Modeling by shortest data description. *Autometica.* **14**: 465-471.

Rodrigez, L.A. ve Acuna, E. 2004. Meta analysis study of outlier detection methods in classification. Technical Paper. Department of Mathematics. University of Puerto Rico at Mayagues.

Rousseeuw, P.J. 1984. Least median of squares regression. *Journal of the American Statistical Association.* **79(388):**871-880

Rousseeuw, P.J. 1985. Multivariate estimation with high breakdown point. *Mathematical Stattistics and Applications-(B)*: 283-297

Rousseeuw, P.J. ve Driessen, K.V. 1990. A fast algortihm for minumum covariance determinant estimator. *Technometrics.* **41:** 212-223

Rousseeuw, P.J. ve Leray, A.M. 2003. Robust Regression and Outlier Detection. Wiley. New York.

Scholz,M. et al. 2005. Non-linear PCA: a missing data approach. *Bioinformatics.***21(20):** 3887-3895

Schwarz, G. 1978. Estimating the dimension of model. *Annals of Statistics.* **6:**461-464

Shannon, C.,E. 1948. A mathematical theory of communication. *Bell System Technical Journal.***27**: 379-423

Shi, J. ve Luo, Z. 2010. Non-linear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. *Computers in Biology and Medicine.***40:** 723-732

Shibata, R. 1989. Statistical Aspects of Model Selection. In Data to Modeling. J.C. Willems (ed). Springer Verlag. Berlin. 216-240

Shieh, A.D. ve Hung, Y. S. 2009. Detecting outlier samples in microarray data. *Statistical Application in Genetics and Molecular Biology.***8(1)** Art.13

Shlens, J. 2009. A tutorial on principal component analysis. www.snl.salk.edu/~shlens/pca.pdf. 11 Şubat 2013.

Shurygin, A. 1983. The linear combination of the simplest discriminator and Fisher's one. Nauka (ed). *Applied Statistics.* Moscow. Rusia.

Singh, D. et al. 2002. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell.* **1:**203-209

Sokal,R.,R. ve Rohlf, F.,J. 1981. Biometry. (2. Baskı). W.H. Freeman Compony. New York

Stein, C. 1956. Some problems in multivariate analysis. Part-1. Technical Report-6: Department of Statistics. Stanford University

Stein, C. 1975. Estimation of covariance matrix. *Rietz Lecture.* 39th Annual Meeting IMS. Atlanta, Georgia.

Stewart, G.,W. 1993. On the early history of the singular value decomposition.*SIAM Review.***35(4):**551-566

Strong, G. 2003. Introduction to Linear Algebra. Wellesley Cambridge Press.

Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and finite corrections. *Communications in Statistics: Theory and Methods.***7(1)**:13-26

Takeuchi, K. 1976. Distribution of the information statistics and criteria for adequacy of models. *Mathematical Science.***153:**12-13. Japanese

Thomaz., C.,E. 2004. Maximum Entropy Covariance Estimate for Statistical Pattern Recognization. Doktora tezi. Department of Computing Imperial College. University of London. UK.

Tibshirani, R. Ve Hastie, T. 2007. Outlier sums for differential gene expression analysis. *Biostatistics.* **8(1):**2-8

Tipping, M., E. ve Bishop, M., C. 1997. Probabilistic principal component analysis, Technical Report NCRG/97/10

Tipping, M., E. ve Bishop, M., C. 1999. Probabilistic principal component analysis, *Journal of the Royal Statistical Society, Series-B.* **61(3):**611-622

Thurstone, L.,L. 1931. Multiple factor analysis. *Psychol Rev.* **38:**406-427

Tomlins, S. A. et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science.***310:** 644-648

van Emden, M.,H. 1971. An Analysis of Complexity. Mathematical Centre Tracts. Amsterdam.

Wall, M., E., Rechtsteiner, A. ve Rocha, M.,L. 2003. Singular value decomposition and principal component analysis. A practical approach to microarray data analysis. Berrar, D., P, Dubitzky, M. ve Gronzow, M (eds). Kluwer: Norwel. 91-109

Warton, D., I. 2008. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the Ameican Statistical Association.* **103(481)**:340-349

Watanabe, S. 1985. Pattern Recognition: Human and Mechanical. John Wiley and Sons. New York.

Williams, J.,L., Bozdogan, H. ve Aiman-Smith, L. 1995. Inference problems with equivalent models. Lawrence Erlboum Associates. New Jersey.

Wu, B. 2007. Cancer outlier differential gene expression detection. *Biostatistics.* **8(3):** 566-575

URL-1: Web kaynağı, www.turkkanser.org.tr/newsfiles/61_dunya_kanser_istatistikleri.pdf 12 Mart 2014.

URL-3: Web kaynağı, Zhau Z. Probabilistic PCA. enbub.fulton.asu.edu/cseml/08. spring_group_meeting/slides/ppca.pdf. 19 Haziran 2013.

Yu, S. 2006. Advanced Probabilistic Models for Clustering and Projections. Doktora tezi. Dissertation im Fach Informatik an der Fakultat für Mathematik, İnformatik und Statistik der Ludwig Maximilians Universitat München.

Zhou, S., 2003. Probabilistic analysis of kernel principal components: mixture modeling, and classiffication

http://scholar.google.com.tr/scholar?hl=tr&q=Probabilistic+analysis+of+kernel+principal+components%3A+mixture+modeling%2C+and+classiffication&btnG=&lr 13 Temmuz 2012.

Zhou, S.K., Chellappa R., Moghaddam B., Probabilistic analysis of kernel principal components, http://scholar.google.com.tr/scholar?q=%22Probabilistic+analysis+of+kernel+principal+components%22&btnG=&hl=tr&as_sdt=0 13 Temmuz 2012.

ÖZGEÇMİŞ

1982 yılında Elazığ'da doğan Esra Pamukçu, ilk orta ve lise öğrenimini Elazığ'da tamamladıktan sonra 2000 yılında Fırat Üniversitesi Matematik bölümünü kazanmıştır. 2004 yılında bölüm derecesi ile mezun olan Pamukçu, aynı yıl özel öğretim kurumlarında Matematik-Geometri öğretmenliğine başlamıştır. 4 yıl bu görevi sürdürdükten sonra 2008 yılında Fırat Üniversitesi İstatistik Bölümü Uygulamalı İstatistik anabilim dalında yüksek lisansa başlamıştır. Kasım, 2009'da aynı anabilim dalına araştırma görevlisi olarak atanan Pamukçu, 2010 yılı Ocak ayında yüksek lisansı bitirdikten sonra aynı yılın Şubat ayında doktoraya başlamıştır. 2012 yılının Şubat döneminde Prof. Dr. Hamparsum Bozdoğan ile doktora çalışmalarına başlayan yazar, daha sonra 2013 Nisan ve Temmuz arasındaki dönemde University of Tennessee'de doktora araştırma asistanı olarak görev almıştır. Prof. Bozdogan'ın danışmanlığında tez çalışmalarını tamamlayan Pamukçu, aynı zamanda uluslararası ICOMP Research Group'un koordinatörüdür. Halen Fırat Üniversitesi İstatistik Bölümü'nde araştırma görevlisi olarak görev yapan yazar, evli ve bir çocuk annesidir.

Esra Pamukçu