Celal Bayar University Journal of Science
Volume 13, Issue 3, p 803-813



# Aşırı Derecede Küçük Örneklem Problemi için Hibrit Regresyon Modeli

Esra Pamukçu*

Fırat Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 23119, Elazığ
epamukcu@firat.edu.tr
*İletişimden sorumlu yazar/Corresponding author

Geliş (Received): 12 Aralık (December) 2016
Kabul (Accepted): 29 Mayıs (May) 2017
DOI: 10.18466/cbayarfbe.339536

**Özet**

Geleneksel istatistik metodolojisinde, iyi seçilmiş değişkenlerin birkaç tane, örneklerin ise daha fazla olduğu farz edilir. Günümüzde ise birçok sahada, 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. Temel bileşenler analizi, faktör analizi, sınıflama ve kümeleme analizleri, regresyon katsayılarının çıkarımı ve tahmini gibi klasik çok değişkenli istatistiksel tekniklerin birçoğu, verinin kovaryans matrisinin ve/veya onun tersinin tahminini gerektirir. p değişken sayısı n örnek sayısından fazla olduğu durumlarda ise örnek varyans-kovaryans matrisi dejenere olur ve tersi hesaplanamaz. Bu, klasik istatistiksel metotlar açıcından karşılaşılabilecek en önemli zorluklardan biridir. Pamukçu ve ark tarafından (2015) yüksek boyutlu veri setlerindeki kovaryans probleminin üstesinden gelebilmek için, Hibrit Kovaryans Tahmin Edicisi (Hybrid Covariance Estimator-HCE) yöntemi geliştirilmiştir. HCE ile kovaryans yapısındaki bu bozulmanın önüne geçilmiş ve n<<p probleminin olduğu yüksek boyutlu veri setlerinin istatistiksel analizleri mümkün hale gelmiştir. HCE, aslında birçok farklı kovaryans yapısı ile elde edilebildiği için HCE ile yapılacak analizlerde önemli aşamalardan biri, veri setine uygun kovaryans yapısının belirlenmesidir. Bu aşamada ise model seçim kriterleri olarak da bilinen AIC, CAIC ve ICOMP gibi bilgi kriterleri ile uygun kovaryans yapısı seçilebilmektedir. Bu çalışmada, n<<p olan yüksek boyutlu veri setlerinde HCE ve bilgi kriterleri ile önerilen Hibrit Regresyon Modeli-HRM tanıtılmış ve hesaplama adımları verilmiştir. Simülasyon çalışması ile farklı senaryolarda farklı p/n oranlarına sahip veri setleri HRM ile analiz edilmiş, uygun kovaryans yapısının seçimi AIC, CAIC ve ICOMP bilgi kriterleri ile yapılmış ve sonuçlar klasik regresyon analizi yöntemi ile karşılaştırılmıştır.

**Anahtar Kelimeler** — Bilgi karmaşıklığı kriteri (ICOMP), Boyutsallık problemi, Hibrit kovaryans tahmin edicisi (HCE), Hibrit regresyon modeli (HRR), Küçük örnekem problemi.

# A New Hybrid Regression Model for Undersized Sample Problem

**Abstract**

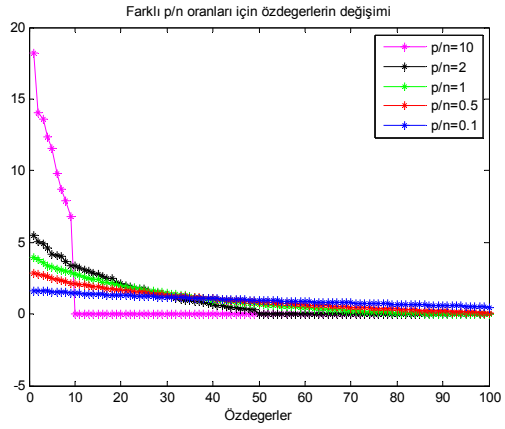
In traditional statistics, it is assumed that the number of samples which are available for study is more than number of well selected variables. Nowadays, in many fields, while the number of samples expressed in tens or hundreds, the single observation may have thousands even millions dimensions. The classical statistical techniques are not designed to be able to cope with this kind of data sets. Many of multivariate statistical techniques such as principal component analysis, factor analysis, classifiation and cluster analysis and the prediction of regression coefficients need estimation of the sample variance-covariance matrix or its inverse. When the number of observations is much smaller than the number of features (or variables), the usual sample covariance matrix degenerates and it can not be inverted. This is one of the biggest encountered obstacle to the classical statistical methods. To remedy the manifestation of the singular covariance matrices in high dimensional data, Hybrid Covariance Estimators (HCE) has been developed by Pamukcu et al.(2015). HCE has overcome the singularity problem of the covariance matrix and, thus, the multivariate statistical analysis for high dimensional data sets has been made possible. One of the most important process in statistical analysis using HCE is to select the appropriate covariance structure for the data set since HCE can in fact be obtained with many different covariance structures. It can be selected by using the information criteria such as Akaike Information Criteria, Information Complexity Criteria which are well known as model selection criteria. In this study, we introduce a new regression model with HCE and information criteria for n<<p undersized high dimensional data. We demonstrate our approach on simulation studies with different scenarious for p/n ratios. We use AIC,CAIC and ICOMP criteria to select appropriate HCE structure and compare the results with classical regression analysis.

**Keywords** — Curse of dimensionality, Hybrid covariance estimator (HCE), Hybrid regression model (HRM), Information complexity criterion (ICOMP), Undersized sample problem.

## 1 Giriş

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 [1]. 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 [2,3].

Gerçek kovaryansın doğru bir tahminine dayanan çok değişkenli teknikler için büyük p-küçük n probleminin olduğu yerlerde, klasik örnek kovaryans matrisi sistematik olarak bozulan bir öz-yapıya sahip olur. 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 varyans kovaryans matrisinin maksimum olabilirlik tahmini, p/n büyük olduğu zaman yansız ve pozitif tanımlı olmasına rağmen, kovaryans matrisinin doğru bir tahmin edicisi değildir [4]. Bu durumda kovaryans matrisinin yapısı, en büyük özdeğerlerin yukarı yönde yanlı, en küçük özdeğerlerin aşağı yönde yanlı olması şeklinde bir bozulmaya uğrar [5]. Bu durumu gösterebilmek amacıyla farklı p/n oranlarına sahip veri setleri üretilmiş ve örnek kovaryans matrislerinin özdeğerleri hesaplanmıştır. Sonuçlar Şekil 1'deki gibidir.




**Şekil 1.** 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 [6].

Yüksek boyutlu veri setlerindeki kovaryans probleminin üstesinden gelebilmek için, bir dizi çalışmanın sonucu olarak Pamukcu ve ark., 2015 [7] tarafından literatürde ilk defa, Maksimum Entropi Kovaryans Tahmin Edicisi [3] ve onun bazı düzgünleştirilmiş kovaryans yapıları ile birlikte kullanımından elde edilen Hibrit Kovaryans Tahmin Edicisi (Hybrid Covariance Estimator-HCE) tanıtılmış, HCE ile kovaryans yapısındaki bu bozulmanın önüne geçilmiş ve n<<p probleminin olduğu yüksek boyutlu veri setlerinin çok değişkenli istatistiksel yöntemler ile analizleri mümkün hale gelmiştir.

HCE'nin kullanıldığı farklı çok değişkenli istatistiksel yöntemler geliştirilmeye ve tanıtılmaya devam etmektedir. Bu çalışmanın amacı, n<<p olan yüksek boyutlu veri setlerinde HCE ve bilgi kriterleri ile regresyon analizini tanıtmaktır. Çalışmanın geri kalan kısmı aşağıdaki başlıklardan oluşmaktadır. Bölüm 2.1'de literatürde var olan çeşitli düzgünleştirilmiş/sağlam kovaryans yapıları, Bölüm 2.2'de Maksimum Entropi Kovaryans matrisi ve Bölüm 2.3'de Hibrit Kovaryans Tahmin Edicisi (Hybrid Covariance Estimator-HCE) tanıtılacaktır. Bölüm 2.4'de model seçim kriterlerine genel bir bakış verildikten sonra çalışmada model seçim kriterlerinden biri olarak kullanılan Bozdogan'ın Bilgi Karmaşıklığı Kriteri (Information Complexity Criteria-ICOMP) tanıtılacaktır. Bölüm 2.5'de Bilgi kriterleri ve HCE ile önerilen yeni regresyon modeli olan Hibrit Regresyon Modeli'nin hesaplama adımları verilecektir. Bölüm 3'de simülasyon çalışması ile farklı senaryolarda farklı p/n oranlarına sahip veri setleri HRM ile analiz edilecek, uygun kovaryans yapısının seçimi AIC, CAIC ve ICOMP bilgi kriterleri ile yapılacaktır. Bölüm 4'de elde edilen sonuçlar tartışılacaktır.

## 2 Materyal ve Metot

### 2.1 Bazı Düzgünleştirilmiş Kovaryans Yapıları

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

$$\hat{\Sigma}_R = \hat{\Sigma}_{MLE} + \alpha I_p \quad (2.1)$$

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

$$CN = \lambda_{maks} / \lambda_{\min} \quad (2.2)$$

şeklindedir. Koşul sayısının tersi olarak kullanılan

$$\kappa(\Sigma) = \frac{1}{CN} \quad (2.3)$$

sayısı singülerliğin tanımı için kullanılabilir [9]. κ(Σ), 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. κ(Σ)<1e-10 ise
ii. Pozitif tanımlı değilse

kovaryans matrisine ridge düzenlileştirmesi uygulanabilir [10].

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

$$\hat{\Sigma}_S = (1-\hat{\rho})\hat{\Sigma}_{MLE} + \hat{\rho}\hat{D} \quad (2.4)$$

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,

$$\hat{D} = \frac{tr(\hat{\Sigma}_{MLE})}{p} I_p = (\frac{1}{p}\sum_{j=1}^{p}\lambda_j) I_p = \bar{\lambda} I_p \quad (2.5)$$

şeklindedir. Burada *tr(.)* matrisin izini, $\lambda_j$ , $j = 1,...,p$ için tahmin edilen örnek kovaryans matrisinin özdeğerlerini ve $\bar{\lambda}$ özdeğerlerin aritmetik ortalamasını gösterir. Denklem (2.4)'deki düzgünleştirilmiş kovaryans matrisinin tahmininde genel formdaki ağırlıklı ortalamanın kullanılması ile 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.

Bu çalışmada ilgilenilen problem kapsamında, aşağıdaki düzenlileştirilmiş (regularized) veya düzleştirilmiş (smoothed) kovaryans tahmin edicileri kullanılmıştır. Bunlar literatürdeki orijinal isimleri ile: Empirical Bayes tahmin edicisi [11]

$$\hat{\Sigma}_{EB} = \hat{\Sigma} + \frac{p-1}{n.tr(\hat{\Sigma})} I_p \quad (2.6)$$

Stipulated Ridge tahmin edicisi [12]

$$\hat{\Sigma}_{SRE} = \hat{\Sigma} + p(p-1)\left[2n.tr(\hat{\Sigma})\right]^{-1} I_p \quad (2.7)$$

Stipulated Diagonal tahmin edicisi [12]

$$\hat{\Sigma}_{SDE} = (1-\pi)\hat{\Sigma} + \pi diag(\hat{\Sigma}) \quad (2.8a)$$

$$\pi = p(p-1)\left[2n.tr(\hat{\Sigma}^{-1}) - p\right]^{-1} \quad (2.8b)$$

Convex Sum tahmin edicisi [13; 14]

$$\hat{\Sigma}_{CSE} = \frac{n}{n+m}\hat{\Sigma} + \left(1 - \frac{n}{n+m}\right)\left[\frac{tr(\hat{\Sigma})}{p}\right] I_p \quad (2.9a)$$

$$0 < m < \frac{2\left[p(1+\beta)-2\right]}{p-\beta} \quad (2.9b)$$

$$\beta = \frac{tr(\hat{\Sigma})^2}{tr(\hat{\Sigma}^2)} \quad (2.9c)$$

Bozdoğan'ın Convex Sum tahmin edicisi [15]

$$\hat{\Sigma}_{BCSE} = \hat{\rho}\hat{\Sigma} + (1-\hat{\rho})\hat{D} \quad (2.10a)$$

$$\hat{\rho} = \frac{1}{\alpha} \quad (2.10b)$$

$$\alpha = \frac{1}{n-1}\sum_{j=1}^{p} Var(x_j) \quad (2.10c)$$

Ayrıca denklem (2.4) formunda olmayıp sadece kovaryans matrisinin özdeğerlerini daha stabil hale getiren bir başka yapı da Thomaz Stabilizasyonu'dur [16]. HCE'nin hesap-

lama adımlarından biri Thomaz Stabilizasyonu'na dayanmaktadır. Bu yapı aşağıdaki hesaplama adımlarından oluşmaktadır.

$\lambda_i$ , örnek kovaryans matrisinin i. özdeğeri, $\bar{\lambda}$ ; özdeğerlerin ortalaması, V; özdeğerlere karşılık gelen özvektörler olmak üzere,

$$\Lambda^* = \begin{pmatrix} \max(\lambda_1, \bar{\lambda}) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \max(\lambda_p, \bar{\lambda}) \end{pmatrix} \quad (2.11a)$$

2.11a denklemi ile, kovaryans matrisinin ortalamadan büyük özdeğerleri değişmez bırakılırken ortalamadan küçük özdeğerlerinin yerine özdeğerlerin ortalaması alınarak $\Lambda^*$ kösegen matrisi oluşturulur.

$$\hat{\Sigma}_{STA} = V\Lambda^* V \quad (2.11b)$$

Özdeğerleri stabil hale getirilmiş yeni kovaryans yapısı 2.11b denkleminde tanımladığı gibi hesaplanır.

**2.2 Maksimum Entropi Kovaryans Tahmin Edicisi**

Özdeğerlerin büyük çoğunluğunun negatif veya sıfır olarak elde edildiği durumlar için muhtemel bir çözüm Maksimum Entropi Kovaryans Matrisi'dir [3]. Bu kovaryans yapısı, örnek kovaryans matrisine göre daha fazla farklılığı ortaya koyar ve 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.

$$Z = \begin{bmatrix} x_1 & y_1 \\ . & . \\ . & . \\ . & . \\ x_n & y_n \end{bmatrix} \quad (2.12)$$

Adım-6: D ridge matrisi hesaplanır:

$$D = \begin{bmatrix} \frac{1}{12n}\sum\left(\xi_j - \xi_{j-1}\right)^2 + \frac{1}{6n}\left\{\left(\xi_1 - \xi_0\right)^2 + \left(\xi_n - \xi_{n-1}\right)^2\right\} & 0 \\ 0 & \frac{1}{12n}\sum\left(\eta_j - \eta_{j-1}\right)^2 + \frac{1}{6n}\left\{\left(\eta_1 - \eta_0\right)^2 + \left(\eta_n - \eta_{n-1}\right)^2\right\} \end{bmatrix} \quad (2.17)$$

Adım-1: Sıra istatistikleri oluşturulur:

$$x^0 \equiv x^1 < x^2 < \dots < x^n \equiv x^{n+1} \quad (2.13a)$$
$$y^0 \equiv y^1 < y^2 < \dots < y^n \equiv y^{n+1} \quad (2.13b)$$

Adım-2: Birincil orta noktalar hesaplanır:

$$\xi_j = \frac{1}{2}(x^j + x^{j+1}) \quad j = 0,1,\dots,n \quad (2.14a)$$
$$\eta_j = \frac{1}{2}(y^j + y^{j+1}) \quad j = 0,1,\dots,n \quad (2.14b)$$

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

$$\bar{x}^j = \frac{1}{2}\left(\xi_{j-1} + \xi_j\right) \quad j = 0,1,\dots,n \quad (2.15a)$$
$$\bar{y}^j = \frac{1}{2}\left(\eta_{j-1} + \eta_j\right) \quad j = 0,1,\dots,n \quad .15b)$$

Adım-4: Her bir gözlem çifti $P_k = (x_k, y_k)$ için, bazı i ve j değerleri vardır ki $(x_k, y_k) = (x^i, y^j)$ dir. Buradan ikincil orta noktalarla ilişkili olan $(\bar{x}^i, \bar{y}^j)$ 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:

$$C = \begin{bmatrix} \frac{1}{n}\sum(\bar{x}_k - \bar{x})^2 & \frac{1}{n}\sum(\bar{y}_k - \bar{y})(\bar{x}_k - \bar{x}) \\ \frac{1}{n}\sum(\bar{y}_k - \bar{y})(\bar{x}_k - \bar{x}) & \frac{1}{n}\sum(\bar{y}_k - \bar{y})^2 \end{bmatrix} \quad (2.16)$$

Adım-7: Maksimum entropi kovaryans matrisi C+D'dir. 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, aşırı derecede küçük örneklem problemine sahip veri setlerinde singüler ve kötü şartlandırılmış kovaryans matrisini pozitif tanımlı yapmasıdır. Ayrıca maksimum entropi kovaryans matrisi hakkında ilginç olan bir başka şey, lineer ve lineer olmayan sıra istatistiklerini kullanarak daha fazla bilgi kullanmasıdır. 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 Hibrit Kovaryans Tahmin Edicisi (Hybrid Covariance Estimator-HCE)

Bu çalışma kapsamında ilgilenilen problem olan yüksek boyutlu küçük örnekleme sahip veri setleri için, bu verilerin kovaryans matrislerinin maksimum olabilirlik tahminlerinin singüler yapıya sahip olup, pozitif tanımlı olmayacakları açıktır. Bu nedenle, maksimum olabilirlik tahminleri kullanılarak yapılacak klasik çok değişkenli istatistiksel analizlerin doğruluğu ve geçerliliği şüphelidir. Dolayısı ile veri girişinin kovaryans matrisi şeklinde olduğu çok değişkenli analizler için singülerlikten uzak iyi şartlandırılmış bir kovaryans yapısının elde edilmesi oldukça önemlidir. Bu noktadan hareketle Pamukçu ve ark. (2015) tarafından Hibrit Kovaryans Tahmin Edicisi-HCE geliştirilmiştir. HCE'nin hesaplama adımları aşağıdaki gibidir:

Adım-1. Maksimum olabilirlik veya maksimum entropi kovaryans tahmin edicisi hesaplanır
Adım-2: Thomaz stabilizasyon işlemi uygulanır
Adım-3: Stabilize edilmiş kovaryans yapısı Bölüm 2.1'de bahsi geçen diğer kovaryans yapıları için başlangıç kovaryansı olarak kullanılır

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

$$\Lambda^* = \begin{pmatrix} \max(\lambda_1, \bar{\lambda}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \max(\lambda_p, \bar{\lambda}) \end{pmatrix} \quad (2.18a)$$

$$\hat{\Sigma}_{ME\_STA} = V \Lambda^* V \quad (2.18b)$$

ile maksimum entropinin stabilizasyonu hesaplanır. Bu kovaryansın Convex Sum yapısında kullanılması ile HCE=ME_STA_CSE kovaryans matrisi elde edilmiş olur. Bu yapı denklem 2.19 (a-b)'da verildiği gibidir.

$$\hat{\Sigma}_{ME\_STA\_CSE} = \frac{n}{n+m}\hat{\Sigma}_{ME\_STA} + \left(1 - \frac{n}{n+m}\right)\left[\frac{tr(\hat{\Sigma}_{ME\_STA})}{p}\right] I_p \quad (2.19a)$$

$$0 < m < \frac{2\left[p(1+\beta)-2\right]}{p-\beta}, \quad \beta = \frac{tr(\hat{\Sigma})^2}{tr(\hat{\Sigma}^2)} \quad (2.19b)$$

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.

### 2.4 Model Seçim Kriterleri

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.Bilgi kriterlerine dayanan model seçim metotları, 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.

Akaike, 1973 ve 1974 yılları arasında ard arda yayınladığı makaleler [17,18] ile istatistiksel modelleme ve istatistiksel model tespiti veya değerlendirmesi alanlarındaki gelişmelere ön ayak olmuştur. Bu nedenle bu alandaki ilk araştırmacılardan biri olarak kabul edilmektedir. 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. $L(\hat{\theta})$ , maksimize edilmiş olabilirlik fonksiyonu, $\hat{\theta}$ , model altında $\theta$ parametre vektörünün maksimum olabilirlik tahmini ve k modeldeki bağımsız parametre sayısını

göstermek üzere,

$$AIC = -2\log L(\hat{\theta}) + 2k \qquad (2.20)$$

şeklinde tanımlanır. Akaike tipi bilgi kriterleri ise, Akaike'nin AIC kriterini temel alan kriterlere verilen genel bir addır. Bayes Bilgi Kriteri (Bayesian Information Criterion-BIC) olarak da bilinen ve Schwartz 1978 [19] tarafından geliştirilen Schwartz Bayesci Bilgi Kriteri ( Schwart's Bayesian Information Criterion-SBC),

$$SBC = -2\log L(\hat{\theta}) + k\log(n) \qquad (2.21)$$

ve denklem 2.22 ile tanımlanan Bozdogan, 1987 [20] tarafından geliştirilen Tutarlı Akaike Bilgi Kriteri (Consistent Akaike Information Criterion-CAIC),

$$CAIC = -2\log L(\hat{\theta}) + k\left[\log(n) + 1\right] \qquad (2.22)$$

Akaike tipi bilgi kriterlerinden bazılarıdır.

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 [21]. Rissanen ise (1978)'de bu sayının rasgele seçildiğini bildirmiştir [22]. Bu noktadan hareketle Bozdogan (1987)'de AIC kriterini geliştirmek, genişletmek ve onu asimptotik durumda tutarlı yapmak için önce Tutarlı Akaike Bilgi Kriteri CAIC'ni önermiştir [20].

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 [23-28,15,9].

ICOMP, (I; Information-COMP; Complexity) 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.

$$Kayıp = \underbrace{\underbrace{Uyum\ eksikliği}_{-2\log L(\hat{\theta})} + \underbrace{Parametrelerin\ ceza\ terimi}_{2k}}_{AIC} + \underbrace{Kompleksli\dot{k}\ ölçüsü}_{C_1(\hat{F})}$$

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

$$ICOMP = -2\log L(\hat{\theta}) + 2C_1(\Sigma) \qquad (2.23)$$

şeklinde tanımlanır. Eşitliğin ikinci terimi olan $C_1(\Sigma)$ Bozdoğan (1990) tarafından ortaya atılan çok değişkenli normal dağılımlı doğrusal veya doğrusal olmayan modelin karmaşıklığının en büyük bilgi teorik ölçümü (maximal information theoretic measure of complexity) olarak adlandırılır [24] ve

$$C_1(\Sigma) = \frac{p}{2}\log\left[\frac{tr(\Sigma)}{p}\right] - \frac{1}{2}\log|\Sigma| \qquad (2.24)$$

şeklinde hesaplanır. Denklem 2.24 için gerekli düzenlemeler yapılırsa,

$$C_1(\Sigma) = \frac{1}{2}\log\frac{\left[\frac{tr(\Sigma)}{p}\right]}{|\Sigma|} \qquad (2.25)$$

ifadesi de yazılabilir. $|\Sigma|$ , genelleştirilmiş varyans ve $\frac{tr(\Sigma)}{p}$ toplam varyans ortalaması olduğu için kompleksli, genelleştirilmiş varyans ile toplam varyans ortalamasının geometrik ortalaması arasındaki logaritmik oran olarak da yorumlanabilir. Dahası, eğer $\Sigma$ 'nın özdeğerleri $\lambda_1, \lambda_2, ..., \lambda_p$ olarak alınırsa, o zaman $\frac{tr(\Sigma)}{p} = \bar{\lambda}_a = \frac{1}{p}\sum\lambda_i$ özdeğerlerin aritmetik ortalaması,

$|\Sigma|^{1/p} = \bar{\lambda}_g = \left(\prod \lambda_j\right)^{1/p}$ özdeğerlerin geometrik ortalaması olmak üzere, $\Sigma$ 'nın kompleksliği

$$C_1(\Sigma) = \frac{p}{2}\log\frac{\bar{\lambda}_a}{\bar{\lambda}_g} \quad (2.26)$$

olarak da yorumlanabilir. $C_1(\Sigma)$, $\Sigma$ 'nın özdeğerlerinin nasıl eşit olmadığını ölçer ve tek bir fonksiyon içinde determinant ve iz olarak adlandırılan çok değişkenli saçılımın en basit iki ölçeğini içerir. $\Sigma$ 'nın tam ranklı olamadığı durumlarda p değeri genellikle s=rank(Σ) olacak şekilde değiştirilir. Bu konuyla alakalı daha detaylı bilgi için [23,24,28] kaynaklarına başvurulabilir. ICOMP'ın farklı uygulama sahalarında tanımlanan ve uygulanan başka yapıları da söz konusudur. Ters Fisher Bilgi Matrisi (Inverse Fısher Information Matrix-IFIM)'in kullanıldığı ICOMP_IFIM, sonsal beklenen faydayı (Posterior Expected Utility_PEU) maksimize etmek amacıyla ICOMP'ın Bayes uyarlaması ICOMP_PEU, hatalı belirlenen modellere karşı dayanıklı formu ICOMP_Miss bunlardan sadece bazılarıdır. Bu kriterler ICOMP tipi kriterler olarak adlandırılır. Daha detaylı bilgi için [23-28,15,9] kaynaklarından faydalanılabilir.

### 2.5 ICOMP ve HCE ile Önerilen Yaklaşım: Hibrit Regresyon Modeli (HRM)

Örnek varyans-kovaryans yapısında singülerlik probleminin olduğu yüksek boyutlu küçük örnekleme sahip n<<p olan veri setinde;

Adım-1: Veri seti için farklı $\hat{\Sigma}_{HCE}$ kovaryans matrisleri tahmin edilir.

Adım-2: $\hat{\Sigma}_{HCE}$, çoklu regresyon çözümlemesinde $X'X$ Gram matrix yerine kullanılır.

Adım-3: Farklı kovaryans yapıları ile oluşturulan modeller arasından ICOMP değerinin minimum olduğu model en iyi model olarak belirlenir.

Çalışmada, ICOMP'ın yanı sıra Akaike tipi bilgi kriterlerinden AIC ve CAIC kriterleri de alternatif olarak kullanılmıştır. Dolayısı ile HRM analizinde model seçimi için kullanılacak bilgi kriterlerinin türetilmesi gerekmektedir. Bu kriterler; k parametre sayısı, n gözlem sayısı olmak üzere aşağıdaki gibi tanımlanır.

AIC:

$$AIC(HRM) = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2k \quad (2.27)$$

CAIC:

$$CAIC(HRM) = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + k(\log(n)+1) \quad (2.28)$$

şeklinde hesaplanır. Burada $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ 'dır.

ICOMP:

$$ICOMP_{Miss}(HRM) = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2C_1(Cov(\hat{\beta}_{HRM})_{Misspec}) \quad (2.29)$$

şeklinde hesaplanır. Burada,

$$S_k = Çarpıklık\ katsayısı = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^3\right)}{\hat{\sigma}^3} \text{ ve}$$

$$K_t = Basıklık\ katsayısı = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^4\right)}{\hat{\sigma}^4} \text{ olmak üzere}$$

$$Cov(\hat{\beta}_{HRM})_{Misspec} = \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} \begin{bmatrix} \frac{n}{\hat{\sigma}^2} & \frac{nS_k}{2\hat{\sigma}^3} \\ \frac{nS_k}{2\hat{\sigma}^3} & \frac{n(K_t-1)}{4\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} \quad (2.30)$$

şeklinde tanımlanır.

## 3 Bulgular

### 3.1 Simülasyon Çalışması

$\hat{\Sigma}_{HCE}$ için teorik olarak veri setine uygun kovaryans yapılarından herhangi biri seçilebilir. $\hat{\Sigma}_{HCE}$, bugüne kadar bir çok farklı veri seti üzerinde farklı yöntemleri denemek amacıyla bir çok kez test edilmiştir. Bu nedenle, simülasyon çalışmasında şimdiye kadar yapmış olduğumuz çalışmalardan edindiğimiz bilgiler ışığında aşağıdaki kovaryans yapıları kullanılmıştır. Hesaplamalarda Matlab programı kullanılmıştır.

**Tablo 1.** HRM için simülasyon çalışmasında kullanılan kovaryans yapıları

| | Hibrit Kovaryans | Model |
|---|---|---|
| 1 | ME* | HRM1 |
| 2 | MLE_STA_CSE | HRM2 |
| 3 | ME_STA_CSE | HRM3 |
| 4 | MLE_STA_BCSE | HRM4 |
| 5 | ME_STA_BCSE | HRM5 |

*ME: Maksimum Entropi Tahmini, MLE: Maksimum Olabilirlik Tahmini, STA: Thomaz Stabilizasyonu, CSE: Convex Sum Tahmini, BCSE: Bozdogan Convex Sum Tahmini.

Simülasyon Protokolu:

- $P_1$: Birbiriyle korelasyona sahip ilgili değişkenlerin sayısı
- $P_2$: İlgisiz ya da gereksiz değişkenlerin sayısı
- P: Veri setindeki toplam değişken sayısı= $P_1$+ $P_2$
- n: Gözlem sayısı
- $X_{1(nxp_1)} \sim MVN(0,\Sigma)$
- $X_{2(nxp_2)} \sim U(0,1)$
- r: Korelasyonlu değişkenler arasındaki korelasyonu belirleyen sayı=0.5
- $\sigma$ : Hata varyansı=0.25
- $\varepsilon$ : Hata
- $y = X\beta + \varepsilon\sigma$

Bu protokol için farklı p/n oranlarını elde edebilmek amacıyla çeşitli senaryolar denenmiştir.

Senaryo-1: $P_1$=10, $P_2$=20, P= $P_1$+$P_2$=30 için n=5, 10, 15, 20 alınarak (nxp)=(5x30), (10x30), (15x30), (20x30) veri setleri üretilmiştir.

Senaryo-2: $P_1$=20, $P_2$=30, P= $P_1$+$P_2$=50 için n=10, 20, 30, 40 alınarak (nxp)=(10x50), (20x50), (30x50), (40x50) veri setleri üretilmiştir.

Senaryo-3: $P_1$=80, $P_2$=20, P= $P_1$+$P_2$=100 için n=10, 20, 21, 50 alınarak (nxp)=(10x100), (20x100), (21x100), (50x100) veri setleri üretilmiştir.

Son senaryoda gözlem sayısındaki bir birimlik değişimin etkisi ilerleyen paragraflarda tartışılacaktır.

İlgisiz değişkenlerin oluşturduğu (nx$P_2$) boyutundaki veri seti U(0,1) tekdüze dağılımdan türetilirken, ilgili değişkenlerin oluşturduğu (nx$P_1$) boyutundaki veri seti, sıfır ortalamalı ve $\Sigma$ varyans-kovaryans matrisine sahip çok değişkenli normal dağılımdan türetilmiştir. Değişkenler arasında korelasyon olması açısından $\Sigma$ aşağıdaki gibi tanımlanmıştır. Örnek olarak r=0.5 ve $P_1$=5 alınmıştır.

$$\Sigma = \begin{bmatrix} 1 & r & \frac{r}{2} & \frac{r}{4} & \frac{r}{8} \\ r & 1 & r & \frac{r}{2} & \frac{r}{4} \\ \frac{r}{2} & r & 1 & r & \frac{r}{2} \\ \frac{r}{4} & \frac{r}{2} & r & 1 & r \\ \frac{r}{8} & \frac{r}{4} & \frac{r}{2} & r & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.500 & 0.250 & 0.125 & 0.062 \\ 0.500 & 1 & 0.500 & 0.250 & 0.125 \\ 0.250 & 0.500 & 1 & 0.500 & 0.250 \\ 0.125 & 0.250 & 0.500 & 1 & 0.500 \\ 0.062 & 0.125 & 0.250 & 0.500 & 1 \end{bmatrix} \quad (2.31)$$

**Tablo 2:**. Senaryo 1 için HRM modellerinin bilgi kriteri değerleri ($P_1$=10, $P_2$=20, P= $P_1$+$P_2$=30)

| n=5 | | HRM1 | HRM2 | HRM3 | **HRM4** | HRM5 |
|---|---|---|---|---|---|---|
| | AIC | 137.616 | 128.653 | 146.122 | **126.713** | 146.451 |
| | CAIC | 155.899 | 146.936 | 164.405 | **144.997** | 164.734 |
| | ICOMP | 78.616 | 69.653 | 87.122 | **67.713** | 87.451 |
| n=10 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 234.728 | **178.531** | 235.809 | 181.830 | 234.782 |
| | CAIC | 273.806 | **217.608** | 274.887 | 220.907 | 273.860 |
| | ICOMP | 175.728 | **119.530** | 176.809 | 122.829 | 175.782 |
| n=15 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 326.611 | **201.015** | 314.254 | 206.445 | 311.583 |
| | CAIC | 377.853 | **252.256** | 365.496 | 257.686 | 362.825 |
| | ICOMP | 267.611 | **142.013** | 255.254 | 147.444 | 252.583 |
| n=20 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 403.324 | **286.657** | 412.244 | 302.203 | 412.197 |
| | CAIC | 463.196 | **346.529** | 472.116 | 362.075 | 472.069 |
| | ICOMP | 344.324 | **227.657** | 353.244 | 243.203 | 353.197 |

**Tablo 3.** Senaryo 2 için HRM modellerinin bilgi kriteri değerleri ($P_1$=20, $P_2$=30, P= $P_1$+$P_2$=50)

| n=10 | | HRM1 | HRM2 | HRM3 | **HRM4** | HRM5 |
|---|---|---|---|---|---|---|
| | AIC | 279.558 | 248.454 | 295.389 | **247.408** | 296.374 |
| | CAIC | 344.687 | 313.583 | 360.518 | **312.538** | 361.503 |
| | ICOMP1 | 180.558 | 149.454 | 196.389 | **148.408** | 197.374 |
| n=20 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 508.892 | **340.844** | 497.480 | 350.530 | 488.077 |
| | CAIC | 608.679 | **440.630** | 597.267 | 450.317 | 587.863 |
| | ICOMP1 | 409.892 | **241.843** | 398.480 | 251.530 | 389.077 |
| n=30 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 651.883 | **503.700** | 679.901 | 529.626 | 681.774 |
| | CAIC | 771.942 | **623.760** | 799.961 | 649.686 | 801.833 |
| | ICOMP1 | 552.883 | **404.700** | 580.901 | 430.626 | 582.774 |
| n=40 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 990.660 | **576.723** | 971.535 | 612.517 | 975.226 |
| | CAIC | 1125.10 | **711.167** | 1105.979 | 746.961 | 1109.671 |
| | ICOMP1 | 891.660 | **477.723** | 872.535 | 513.517 | 876.226 |

**Tablo 4.** Senaryo 3 için HRM modellerinin bilgi kriteri değerleri ($P_1$=80, $P_2$=20, P= $P_1$+$P_2$=100)

| n=10 | | HRM1 | HRM2 | HRM3 | **HRM4** | HRM5 |
|---|---|---|---|---|---|---|
| | AIC | 456.5085 | 393.2942 | 447.1549 | **383.6019** | 438.1423 |
| | CAIC | 586.767 | 523.5527 | 577.4134 | **513.8604** | 568.4008 |
| | ICOMP1 | 257.5085 | 194.2942 | 248.1549 | **184.6019** | 239.1423 |
| n=20 | | HRM1 | HRM2 | HRM3 | **HRM4** | HRM5 |
| | AIC | 738.1502 | 566.6225 | 719.7299 | **565.9338** | 696.6322 |
| | CAIC | 937.7234 | 766.1957 | 919.3032 | **765.5071** | 896.2054 |
| | ICOMP1 | 539.1502 | 367.6225 | 520.7299 | **366.9338** | 497.6322 |
| n=21 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 748.8445 | **565.6283** | 730.654065 | 565.7666 | 706.0466 |
| | CAIC | 953.2967 | **770.0806** | 935.1063088 | 770.2189 | 910.4989 |
| | ICOMP1 | 549.8445 | **366.6283** | 531.654065 | 366.7666 | 507.0466 |
| n=50 | | HRM1 | **HRM2** | HRM3 | HRM4 | HRM5 |
| | AIC | 1478.337 | **1007.262** | 1452.223 | 1043.292 | 1423.937 |
| | CAIC | 1769.54 | **1298.464** | 1743.425 | 1334.495 | 1715.139 |
| | ICOMP1 | 1279.337 | **808.2622** | 1253.223 | 844.2925 | 1224.937 |

Üç farklı senaryoda üretilen 12 farklı veri seti hem klasik regresyon ile hem de Tablo-1'de verilen 5 farklı kovaryans ryans kovaryans yapısındaki singülerlik probleminden dolayı, rank sayısı kadar değişkenin sıfırdan farklı olduğu, diğerlerinin sıfır olarak elde edildiği görülmüştür. HRM ile analiz yapıldığında ise, kovaryans yapısındaki singülerlik problemi ortadan kalktığı için değişken sayısı kadar sıfırdan farklı regresyon katsayısı elde edilebilmiştir.

yapısı kullanılarak analiz edilmiştir. Klasik regresyon yöntemi ile analiz edildiği zaman, örnek va

Hangi kovaryans yapısının en iyi sonucu verdiği ise bilgi kriterleri kullanılarak belirlenmiştir. Sonuçlar Tablo-2,3 ve 4'deki gibidir. Aynı veri seti için kullanılan farklı kovaryans yapıları arasından bilgi kriterlerinin minimum olarak elde edildiği model en iyi model olarak seçilmiştir ve koyu renk yazılarak belirtilmiştir. Tüm senaryolarda en iyi sonucları veren HCE yapıları MLE_STA_CSE ve MLE_STA_BCSE

olarak tespit edilmiştir. p/n oranına bağlı olarak iki kovaryans yapısından biri uygun olmaktadır. İlk iki senaryoda p/n oranının 5 olduğu sezgisel olarak anlaşılmıştır. Yani değişken sayısı gözlem sayısının 5 katı ve üstünde ise MLE_STA_BCSE kovaryans yapısı, bu oran düştükçe MLE_STA_CSE kovaryans yapısı iyi sonuç vermektedir. Son senaryoda, bu durum n=20 ve n=21 alınarak test edilmiştir. Gerçektende, n=21 olduğu yani p/n oranı 5'in altına düştüğü anda uygun kovaryans yapısı değişmiştir.

## 4 Tartışma ve Sonuç

Klasik çok değişkenli istatistiksel yöntemlerin birçoğu, $X'X$ matrisinin ve onun tersinin hesaplanmasını gerektirir. Ancak n<<p olduğu durumlarda $X'X$ matrisinin tersinin (precision matrix) hesaplanamaması, bu veri setlerinde klasik yöntemlerin kullanılması açısından karşılaşılabilen en büyük zorluklardan biridir. Literatürde, n<<p için çeşitli kovaryans matrisleri tanıtılmaya devam edilmektedir. Bunlardan bir tanesi de Pamukçu ve ark. tarafından 2015'de önerilen Hibrit Kovaryans Tahmin Edicisi (Hybrid Covariance Estimator-HCE)'dir. Bu çalışmada, ilk olarak n<<p problemine sahip veri setleri için, HCE kovaryans matrisinin yapısı tanıtılmıştır. n<<p olan veri setleri için, HCE özdeğerlerinin daha durağan bir yapıya sahip olduğu ve singüler olmayan bir kovaryans matrisidir. HCE ile çok değişkenli istatistiksel yöntemler geliştirilmeye ve tanıtılmaya devam edilmektedir. Bunlardan bir tanesi olarak bu çalışmada Hibrit Regresyon Modeli (HybridRegression Model-HRM) tanıtılmıştır. n<<p için, farklı n ve p oranlarında türetilen veri setleri HRM ile analiz edilmiştir. Klasik regresyon modeli, bu tarz verilerle başa çıkabilecek şekilde tasarlanabilmiş değildir. Nitekim, klasik regresyon modeli ile katsayıların büyük çoğunluğu sıfır elde edilmiştir. HRM için farklı 5 model denenmiş, hepsinde de değişken sayısı kadar katsayı hesaplanabilmiştir. Yapılan simülasyon çalışması neticesinde, en iyi kovaryans yapısının p/n oranına bağlı olarak değişkenlik gösterdiği ve p/n ≤ 5 için MLE/STA/CSE, p/n > 5 için MLE/STA/BCSE en iyi kovaryans yapıları olarak tespit edilmiştir.

Bundan sonraki aşama katsayıları hesaplanabilen p adet değişken arasından cevap değişkeni üzerinde en çok etkiye sahip olan değişkenlerin tespit edilmesi aşamasıdır. Model seçimi, temelde açıklayıcı değişkenlerden hangi bir veya birkaçının cevap değişkeni üzerinde etkili olduğunu ortaya koyarak mevcut modeller altkümesi içinden en iyi modelin bulunması ile ilgili bir süreçtir. Özellikle açıklayıcı değişkenlere ilişkin alternatif alt küme sayısı, yüksek boyutlu veri setleri için milyonları bulabildiği ( p=20 için $2^{20}$ =1048576 adet değişken kombinasyonunun olması gibi) bir gerçektir. Bu ölçüde geniş bir çözüm uzayı içinden en iyi modeli seçebilmek için klasik ileri doğru, geri doğru ya da adımsal tekniklerden ziyade ICOMP tipi bilgi kriterlerinin amaç fonksiyonu olarak kullanıldığı genetik algoritma gibi bazı nümerik optimizasyon teknikleri ve stratejilerinin kullanılması önerilmektedir. Tüm değişkenler içinden en iyilerinin tespit edilmesi probleminin çözülmesinin ardından yöntemin etkinliği gerçek veriler üzerinde gösterilebilecektir.

## Teşekkür

Araştırma probleminin belirlenmesi, kaynakların temini ve simülasyon protokolü için University of Tennessee Department of Business Analytics & Statistics öğretim üyesi Prof. Dr. Hamparsum Bozdoğan'a teşekürlerimi sunarım. Ayrıca makaleyi okuyup son şeklini almasında desteklerini esirgemeyen Fırat Üniversitesi Fen Fakültesi İstatistik Bölümü öğretim üyesi Doç.Dr. Mehmet Gürcan'a teşekkür ederim.

## Referanslar

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

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

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

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

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

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

**7.** Pamukçu, E.; Bozdogan, H., Çalık, S. A Novel Hybrid Dimension Reduction Technique for Undersized High Dimensional Gene Expression Data Sets Using Information Complexity Criterion for Cancer Classification. *Computational and Mathematical Methods in Medicine*. Volume 2015 (2015), Article ID 370640, 14 pages

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

**9.** Bozdogan, H.; Information Complexity and Multivariate Learning in High Dimensions with Applications in Data Mining. Forthcoming book. 2017

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

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

**12.** Shurygin, A.; The linear combination of the simplest discriminator and Fisher's one. Nauka (ed). Applied Statistics. Moscow. Rusia. 1983; 144-158

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

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

**15.** Bozdogan, H.; 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*. 2010; 39(2), 370-398

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

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

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

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

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

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

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

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

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

**25.** Bozdogan, H.; 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, 1994.

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

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

**28.** Bozdogan, H.; 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,2004.