SÜREKLİ DEĞİŞKENLER İÇEREN GRAFİKSEL MODELLERDE KULLANILAN SAPMA VE F-İSTATİSTİKLERİNİN KARŞILAŞTIRILMASI

Fikri GÖKPINAR*

ÖZET

Çok değişkenli normal dağılıma sahip değişkenler için kullanılan grafiksel modeller kovaryans seçimli modeller olarak adlandırılır. Böyle modeller değişken çiftlerinin koşullu bağımsızlığına dayanarak belirlenir. Bu çalışmada kovaryans seçimli modeller için test işleminde kullanılan sapma ve F istatistikleri farklı örnek çapları, değişken sayıları ve koşullu bağımsızlık yapıları altında simülasyon yoluyla karşılaştırılmıştır.

Anahtar Kelimeler: *Koşullu bağımsızlık, Kovaryans seçimli modeller, Sapma, F istatistiği*

1. GİRİŞ

Çok değişkenli normal dağılıma sahip değişkenler için elde edilen grafiksel modeller kovaryans seçimli modeller olarak adlandırılır. Bu modellerde, ters kovaryans matrisinin elemanları sıfır olanlar dikkate alınarak model oluşturulabilir. Grafikte, V köşe kümesini, E köşeler arası kenar kümesini temsil eder. Eğer bir köşe çifti arası birden fazla kenar yoksa grafik basittir. Kenar $(\alpha,\beta)\in E$ ve $(\beta,\alpha)\in E$ şeklinde ise yön verilmemiş kenar, $(\alpha,\beta)\in E$ ve$(\beta,\alpha)\notin E$ ise yön verilmiş bir kenardır. Grafiğin tüm köşeleri arasında çizgi ya da ok varsa bu grafiğe tam grafik denir. V'nin bir alt kümesine, eğer buna ilişkin grafik tam ise, tamdır denir. V'nin tam alt kümelerine klik denir. α'dan β'ya bir ok varsa α'ya β'nın ailesi, β'ya da α'nın çocuğu denir. β'nın aile kümesini $pa(\beta)$, α'nın çocuklar kümesi $ch(\alpha)$ şeklinde gösterilebilir. α ile β arasında bir çizgi varsa α ve β'ya bitişik ya da komşu denir. α köşesinin komşu kümesi $ne(\alpha)$ şeklinde gösterilebilir. Grafiksel modellerin temeli rasgele değişkenlerin koşullu bağımsızlığına dayanır. Bu modellerin grafikleri koşullu bağımsızlık ilişkilerini gösterir.

X, Y, Z sürekli rasgele değişkenleri için $X \perp Y/Z$ koşullu bağımsızlığı

$$X \perp Y/Z \Leftrightarrow f_{XY/Z}(x, y/z) = f_{X/Z}(x/z) f_{Y/Z}(y/z) \qquad (1.1)$$

biçiminde ifade edilir.

$X \perp Y/Z$ ilişkisi aşağıdaki özelliklere sahiptir.

$$\left.\begin{array}{l} X \perp Y/Z \Rightarrow Y \perp X/Z\text{'dir.} \\ X \perp Y/Z \text{ ve } U=h(x) \Rightarrow U \perp Y/Z\text{'dir.} \\ X \perp Y/Z \text{ ve } U=h(x) \Rightarrow X \perp Y/(Z,U)\text{'dir.} \\ X \perp Y/Z \text{ ve } X \perp W/(Y,Z) \Rightarrow X \perp (W,Y)/Z\text{'dir.} \end{array}\right\} \qquad (1.2)$$

Kovaryans seçimli modellerin oluşturulmasına ilişkin kural Dempster(1972) tarafından verilmiştir. Ayrıca Wermuth(1980) ayrıştırılabilir modeller kovaryans seçimli modeller ile Wright(1923) tarafından önerilen path Analizi arasındaki ilişkiyi ortaya koymuştur. Daha sonra Cox ve Wermuth(1993) yaptıkları çalışmada kovaryans matrisi ile ilişkilendirilen çeşitli özel doğrusal yapıların modellenmesi için grafiksel modelleri kullanmışlardır. Whittaker(1990) bu modeller için sapma test istatistiği vermiştir. Ayrıca Eriksen(1996) sapma(d) test istatistiğine alternatif olarak geliştirdiği F test istatistiğini önermiştir. Fakat bu test istatistikler birbirleriyle testin gücü bakımından karşılaştırılmamış F ve d'nin hangi şartlar altında birbirlerine üstünlük sağladıkları ortaya konmamıştır. Bu çalışmada sürekli değişken içeren grafiksel modeller için kullanılan bu iki test istatistiği farklı örnek çapları ve farklı değişken sayıları için değişik kısmi korelasyon yapıları dikkate alınarak testin gücü bakımından karşılaştırılmış ve test istatistiklerinin daha iyi olan durumlar ortaya konmuştur.

2. SÜREKLİ DEĞİŞKEN İÇEREN GRAFİKSEL MODELLER

Grafiksel modeller sürekli değişkenler için kovaryans seçimli modeleri temel alır. Kovaryans seçimli modeller ilk önce Dempster(1972) tarafında kullanılmıştır. Bu modeller Whittaker(1990) tarafından geliştirilmiştir.

$Y=(Y_1,...,Y_q)$ q boyutlu rassal değişkenler vektörü olmak üzere

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ \mu_q \end{pmatrix} \tag{2.1}$$

ortalama vektörü ve

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & . & . & \sigma_{1q} \\ \sigma_{21} & \sigma_{22} & . & . & \sigma_{2q} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sigma_{q1} & \sigma_{q2} & . & . & \sigma_{qq} \end{pmatrix} \tag{2.2}$$

kovaryans matrisi ile çok değişkenli normal dağılıma sahiptir. Burada dikkate alınan kovaryans matrisinin tersidir.

$$K = \Sigma^{-1} = \begin{pmatrix} w^{11} & w^{12} & . & . & w^{1q} \\ w^{21} & w^{22} & . & . & w^{2q} \\ . & . & . & . & . \\ . & . & . & . & . \\ w^{q1} & w^{q2} & . & . & w^{qq} \end{pmatrix} \tag{2.3}$$

$$K = \Sigma^{-1} = \begin{pmatrix} w^{11} & w^{12} & . & . & w^{1q} \\ w^{21} & w^{22} & . & . & w^{2q} \\ . & . & . & . & . \\ . & . & . & . & . \\ w^{q1} & w^{q2} & . & . & w^{qq} \end{pmatrix} \tag{2.3}$$

Bu matris kesinlik matrisi ya da konsantrasyon matrisi olarak adlandırılır(Lauritzen and Wermuth,1989).

$(Y_3,....,Y_q)$ verilmişken (Y_1,Y_2)'in koşullu dağılımı (2.4)deki kovaryans matrisine sahip iki değişkenli normal dağılımdır.

$$\begin{pmatrix} w^{11} & w^{12} \\ w^{21} & w^{22} \end{pmatrix}^{-1} = \frac{1}{w^{11}w^{22} - (w^{12})^2} \begin{pmatrix} w^{22} & -w^{21} \\ -w^{12} & w^{11} \end{pmatrix} \tag{2.4}$$

(2.4)deki kovaryans matrisli dağılıma ilişkin korelasyon katsayısı (2.5)te verildiği gibidir.

$$p^{12.34...q} = \frac{-w^{12}}{(w^{11}w^{22})^{\frac{1}{2}}} \tag{2.5}$$

Ayrıca

$$p^{12.34...q} = 0 \Leftrightarrow w^{12} = 0 \tag{2.6}$$

denkliği (2.5)ten açıkça görülür.

Geri kalan değişkenler verilmişken iki değişkenin bağımsız olması için gerek ve yeter koşul konsantrasyon matrisinde karşılık gelen elemanların sıfır olmasıdır(Cox and Wermuth 1993). Bu şekilde konsantrasyon matrisinin elemanları, log lineer modeldeki iki faktörlü etkileşimlerle aynı rolü üstlenir.

(2.6)daki koşullu bağımsızlık Y'nin olasılık yoğunluk fonksiyonu dikkate alınarak kanonik parametreler yardımıyla görülebilir. Y'nin yoğunluğu

$$f(y) = (2\pi)^{-1}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^t \Sigma^{-1}(y-\mu)\right\} \tag{2.7}$$

şeklinde yazılır. Bu yoğunluk,

$$f(y) = \exp(g + h'y - \tfrac{1}{2}y'Ky) \tag{2.8}$$

ile ifade edilir. Burada $K=\Sigma^1$, $h=\Sigma^1\mu$ ve g normalleştirme sabiti olup

$$g = -\tfrac{1}{2}\ln|\Sigma| - \tfrac{1}{2}\mu'\Sigma\mu - \tfrac{q}{2}\ln(2\pi) \tag{2.9}$$

$Y_j \perp Y_k$/(geri kalan) $\Leftrightarrow w^{jk}=0$ (2.11)

biçiminde ifade edilir.

Grafiksel gauss modellerde log lineer modellerde olduğu gibi hiyerarşik ya da grafiksel olmayan model ayrımı yoktur. Tüm modeller grafikseldir ve grafiklerle modeller arasında birebir ilişki vardır. Grafiksel modeller koşullu bağımsızlık ilişkilerini ortaya koymak için d ve F istatistiklerinden faydalanır. d ve F istatistiklerinin hesaplanmasında en çok olabilirlik fonksiyonlarından faydalanılır. Bu fonksiyonu oluşturmak için N gözlemlik $y^{(1)},\ldots. y^{(N)}$ örneği alınsın. $\bar{y} = \sum_{k=1}^{N} y^{(k)} / N$ örnek ortalaması vektörüdür. Örnek kovaryans matrisi

$$S = \sum_{k=1}^{N} (y^{(k)} - \bar{y})(y^{(k)} - \bar{y})' / N$$

olsun. log-yoğunluk

$$\ln(f(y)) = -q\ln(2\pi)/2 - \ln|\Sigma|/2 - (y-\mu)'\Sigma^{-1}(y-\mu)/2$$

şeklinde yazılabilir. Ayrıca log-olabilirliği

$$l(\mu, K) = -Nq\ln(2\pi)/2 - N\ln|\Sigma|/2 - \sum_{k=1}^{N} (y^{(k)} - \mu)' K\ (y^{(k)} - \mu)/2 \qquad (2.12)$$

dir. Son terim

$$\sum_{k=1}^{N} (y^{(k)} - \mu)' K (y^{(k)} - \mu) = \sum_{k=1}^{N} (y^{(k)} - \bar{y})' K (y^{(k)} - \bar{y}) + N(\bar{y} - \mu)' K (\bar{y} - \mu)$$

şeklinde yazılabilir. Bu ifade iz fonksiyonu kullanarak aşağıdaki gibi basitleştirilebilir.

$$\sum_{k=1}^{N} (y^{(k)} - \bar{y})' K (y^{(k)} - \bar{y}) = Ntr(KS)$$

Böylece log-olabilirliği aşağıdaki eşitlik ile verilir.

$$l(\mu, K) = -Nq\ln(2\pi)/2 - N\ln|\Sigma|/2 - Ntr(KS)/2 - N(\bar{y} - \mu)' K (\bar{y} - \mu)/2 \qquad (2.13)$$

$a \subseteq \Gamma$ olan değişken altkümesi için, Σ^{aa}, S^{aa} a'ya karşılık gelen Σ ve S'in alt matrisleri olsun. $q_1, q_2, \ldots, q_t$ üreteçleriyle belirlenen model için, minimal yeterli istatistikler kümesinin, $\bar{y}$ örnek ortalaması ve üreteçlere karşılık gelen örnek kovaryans matrisinin marjinal alt kümelerinin kümesi (S^{aa}, $a=q_1, q_2, \ldots, q_t$ için) olduğu gösterilebilir. Olabilirlik eşitlikleri model altında beklenen değerleriyle birlikte bunları (minimal yeterli istatistik) eşitleyerek oluşturulur. Böylece $a=q_1, q_2, \ldots, q_t$ için $\hat{\mu} = \bar{y}$ ve $\hat{\Sigma}^{aa} = S^{aa}$ eşitlikleri elde edilir.

Bu sonuçları kullanarak, model altında olabilirliği maksimize etmek için ifade basitleştirilebilir. $\hat{\mu} = \bar{y}$ olduğunda (2.13)deki son terim yok olur ve $\hat{\Sigma}$ ve S, $w^{ij}=0$ olan

$a \subseteq \Gamma$ olan değişken altkümesi için, Σ^{aa}, S^{aa} a'ya karşılık gelen Σ ve S'in alt matrisleri olsun. $q_1,q_2,....,q_t$ üreteçleriyle belirlenen model için, minimal yeterli istatistikler kümesinin, $\bar{y}$ örnek ortalaması ve üreteçlere karşılık gelen örnek kovaryans matrisinin marjinal alt kümelerinin kümesi (S^{aa}, $a=q_1,q_2,....,q_t$ için) olduğu gösterilebilir. Olabilirlik eşitlikleri model altında beklenen değerleriyle birlikte bunları (minimal yeterli istatistik) eşitleyerek oluşturulur. Böylece $a=q_1,q_2,...,q_t$ için $\hat{\mu} = \bar{y}$ ve $\hat{\Sigma}^{aa} = S^{aa}$ eşitlikleri elde edilir.

Bu sonuçları kullanarak, model altında olabilirliği maksimize etmek için ifade basitleştirilebilir. $\hat{\mu} = \bar{y}$ olduğunda (2.13)deki son terim yok olur ve $\hat{\Sigma}$ ve S, $w^{ij}=0$ olan elemanlar için kesin olarak farklılık gösterdiğinde $tr(\hat{K}S) = tr(\hat{K}\hat{\Sigma}) = q$ elde edilir Böylece model altında maksimize edilen log-olabilirlik

$$l_0 = -Nq\ln(2\pi)/2 - N\ln\left|\hat{\Sigma}\right|/2 - Nq/2$$

ifadesine basitleştirilebilir. Tam model M_1 altında $\hat{\Sigma} = S$ 'dir. Böylece bu modelin maksimize edilmiş log-olabilirliği

$$l_1 = -Nq\ln(2\pi)/2 - N\ln\left|\hat{S}\right|/2 - Nq/2$$

dir.

Grafiksel modellerde bir kenarın çıkarılıp çıkarılmayacağına ilişkin test için sapma istatistiği aşağıdaki gibi tanımlanır.

Bir M_0 modelinin sapması M_1 doymuş modeline karşı M_0'ın olabilirlik oranı testidir. Böylece modelin sapması

$$G^2 = 2(\hat{l}_1 - \hat{l}_0)$$
$$= N\ln\left(\left|\hat{\Sigma}\right|/\left|S\right|\right) \qquad (2.14)$$

olur. Burada S, Σ'nın doymuş model altındaki tahminidir $\hat{\Sigma}$ ise karşıt (doymamış) model altındaki tahminidir. S bilinen örnek kovaryans matrisidir. $\hat{\Sigma}$ ise modelde hangi değişkenler arasında koşullu bağımsızlık olduğu varsayılıyorsa S'in tersinde o elemanlara karşılık gelen elemanların 0 olarak belirlenmesiyle elde edilir.

$M_0 \subseteq M_1$ için sapma farkı ise

$$d = N\ln\left(\left|\hat{\Sigma}_0\right|/\left|\hat{\Sigma}_1\right|\right) \qquad (2.15)$$

3. UYGULAMA

Uygulama olarak, bölüm 2'de tanımlanan d ve F test istatistikleri 3, 4 ve 5 değişkenli modellerde farklı bağımlılık yapıları ve örnek çapları verilmişken testin gücü bakımından karşılaştırılacaktır.

3, 4 ve 5 değişkenli modeller için yığın konsantrasyon matrisleri sırasıyla

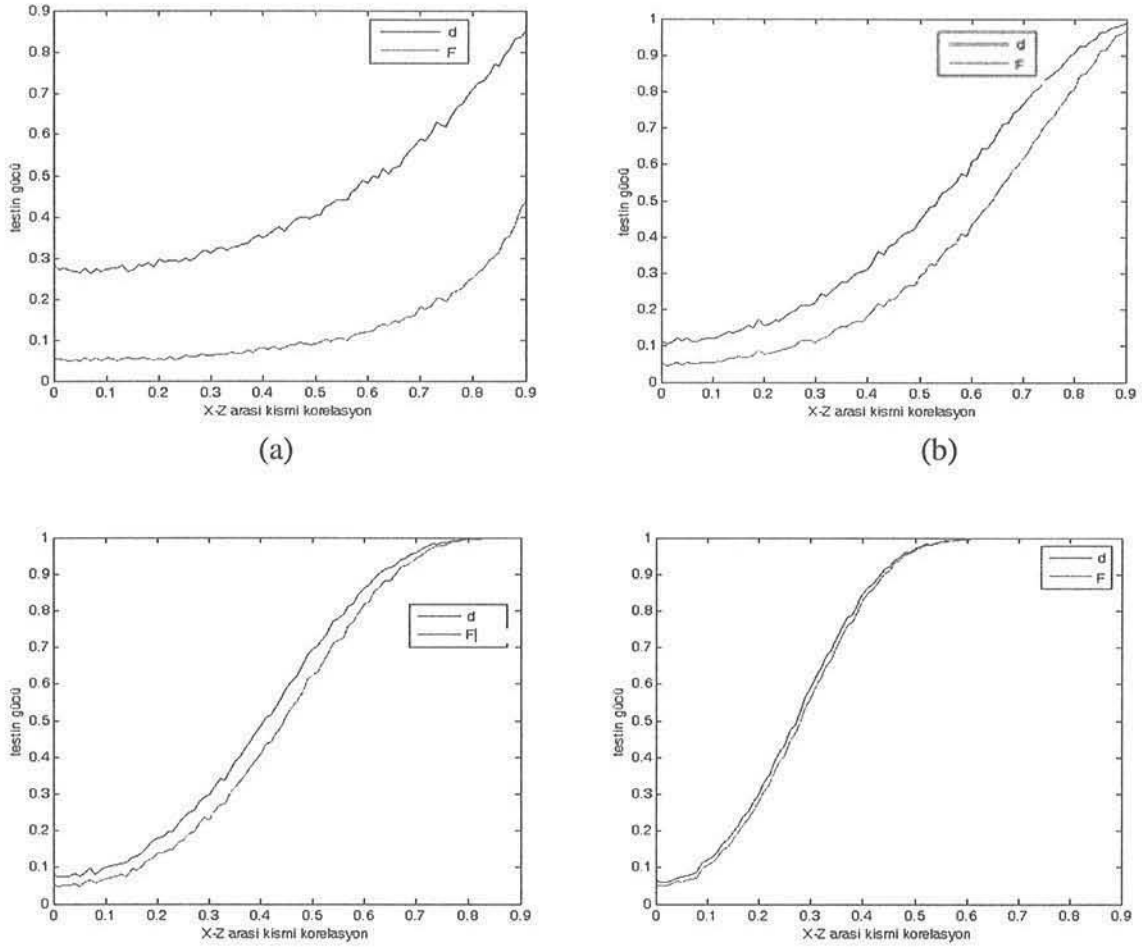
$$K = \begin{pmatrix} 0.5301 & 0.2471 & w_{xz} \\ 0.2471 & 1.0465 & 0.5673 \\ w_{zx} & 0.5673 & 1.5854 \end{pmatrix}$$

$$K = \begin{pmatrix} 0.029 & -0.016 & -0.008 & w_{xz} \\ -0.016 & 0.050 & -0.006 & -0.015 \\ -0.008 & -0.006 & 0.039 & -0.037 \\ w_{xz} & -0.015 & -0.037 & 0.089 \end{pmatrix}$$

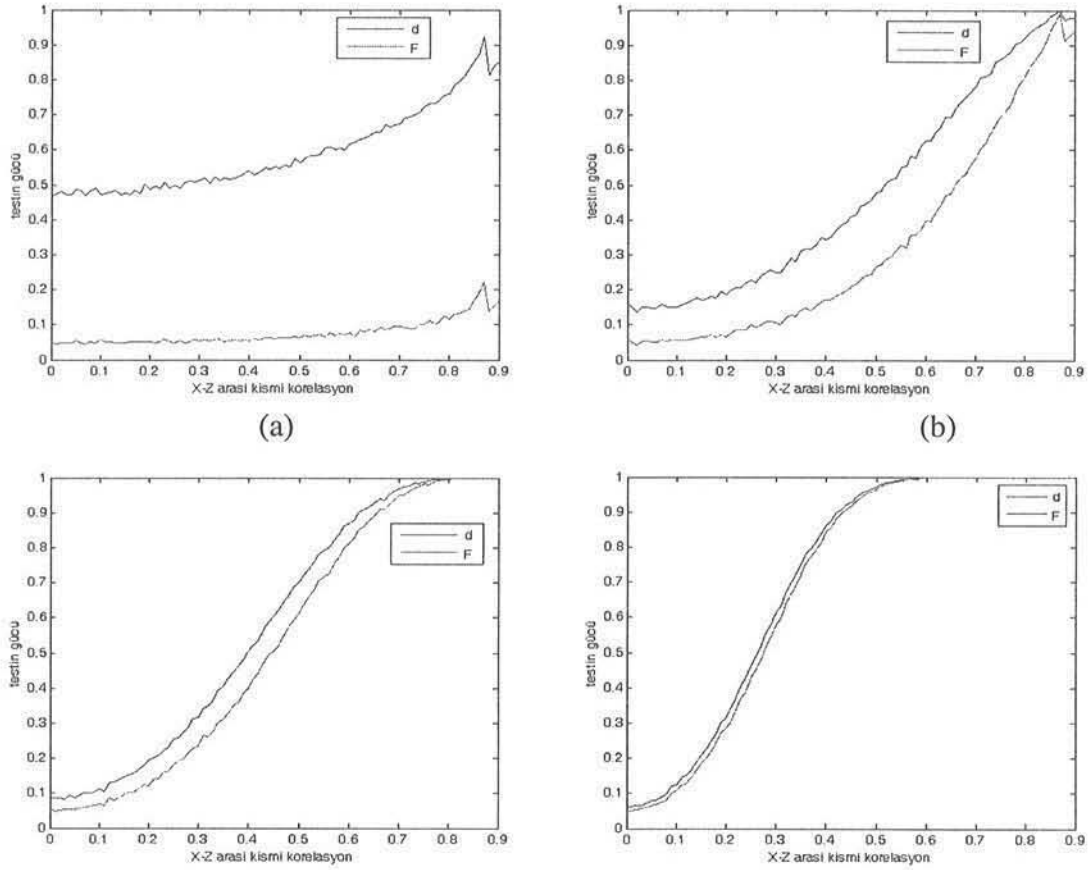
$$K = \begin{pmatrix} 0.011 & -0.005 & -0.001 & -0.001 & w_{xz} \\ -0.005 & 0.027 & -0.007 & -0.005 & -0.003 \\ -0.001 & -0.007 & 0.010 & -0.020 & 0.010 \\ -0.001 & -0.005 & -0.020 & 0.007 & 0.020 \\ w_{xz} & -0.003 & 0.010 & 0.020 & 0.005 \end{pmatrix}$$

olarak alınmıştır.

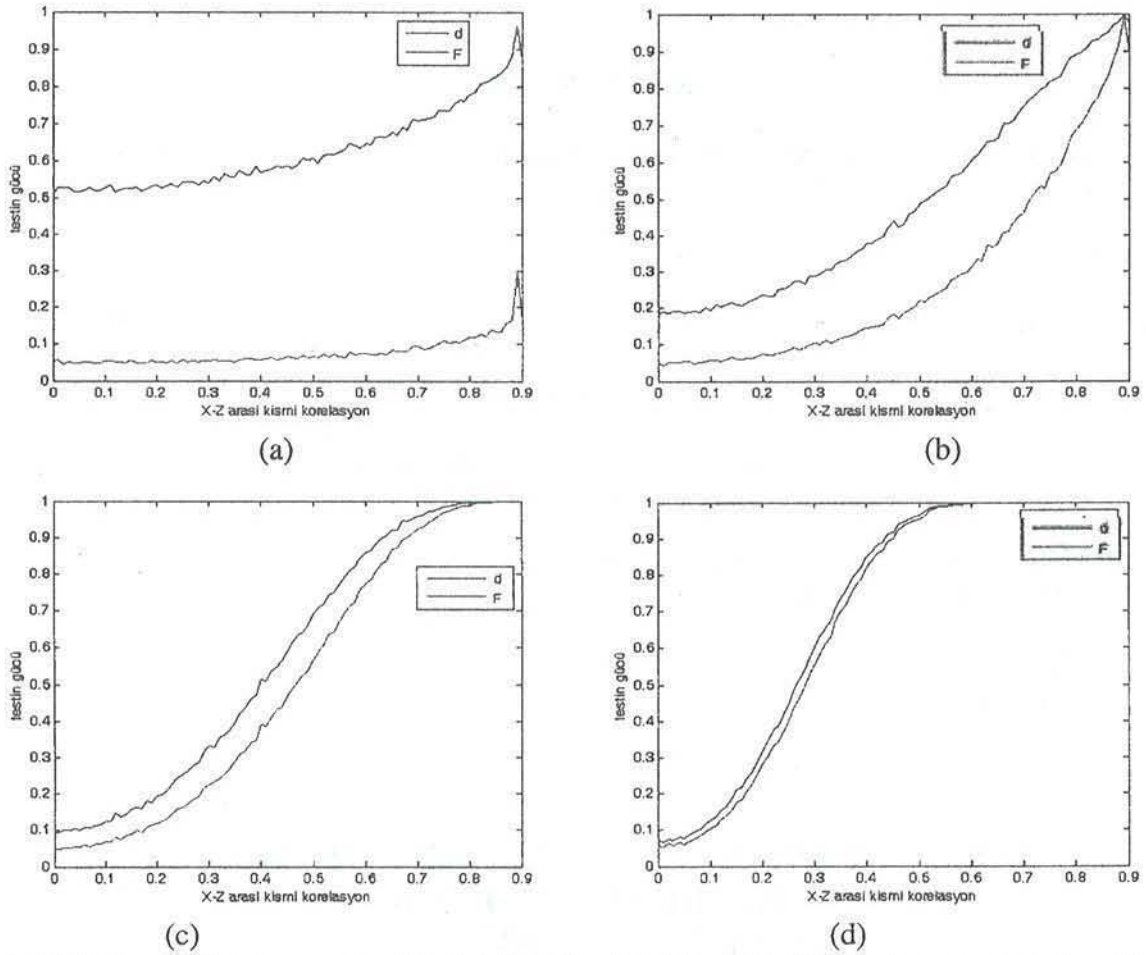
İki test istatistiği için kullanılan formüllerde yğın ortalaması olmadığından yığın ortalamaları tüm değişkenler için 50 alınmıştır. Örnek çapları 5, 10, 20 ve 50 göz önüne alınmış ve X ile Z arasındaki kısmi korelasyona farklı değerler verilerek Şekil 3.1-Şekil 3.3 elde edilmiştir.



Şekil 3.1: Üç değişkenli yapıda a)N=5, b)N=10, c)N=20, d) N=50 örnek çaplarında d ve F için testin güçleri



Şekil 3.2: Dört değişkeli yapıda a)N=5, b)N=10, c)N=20, d) N=50 örnek çaplarında d ve F için testin güçleri



Şekil 3.3: Beş değişkeli yapıda a)N=5, b)N=10, c)N=20, d) N=50 örnek çaplarında d ve F için testin güçleri

Şekil 3.1-Şekil 3.3 incelendiğinde örnek çapı küçük iken d ile F arasındaki fark çok büyüktür. Fakat iki tahmin edici için de testin gücü yüksek değerler almamaktadır. Burada dikkat edilecek nokta küçük örnek çaplarında iki değişken arasındaki kısmi korelasyon 0'a yakın iken bile d'nin F'ye göre çok yüksek değerler aldığıdır. Bu fark değişken sayısı arttıkça artmaktadır. Şekil 3.1a, 3.2a, 3.3a incelendiğinde kısmi korelasyon 0 iken F için elde edilen testin gücü değeri 0.05 yakın değerler alırken, d için testin gücü değeri değişken sayısı arttıkça yükselmektedir. Bu durum istenilen bir sonuç değildir. Kısmi korelasyon 0'a yakınken testin gücünün 0'a yakın olması beklenir. Ancak kısmi korelasyon değeri yüksek iken testin gücünün de yüksek olması düşünülür. Fakat N=5 için F, kısmi korelasyon değeri 0.90 iken testin gücünün 0.2 civarı değer aldığı görülüyor. Bu durumda çekilen örnek çapı küçükken, iki test istatistiği farklı durumlarda değişik sonuçlar vermektedir. Küçük yığın kısmi korelasyonları içim F daha iyi sonuçlar verirken yüksek kısmi korelasyon değerlerinde d daha uygun sonuçlar vermektedir. Örnek çapı arttıkça d, kısmi korelasyonun düşük olduğu durumlarda F'e yaklaşırken, kısmi korelasyonun yüksek olduğu durumlarda F, d'ye yakın

kısmi korelasyonları içim F daha iyi sonuçlar verirken yüksek kısmi korelasyon değerlerinde d daha uygun sonuçlar vermektedir. Örnek çapı arttıkça d, kısmi korelasyonun düşük olduğu durumlarda F'e yaklaşırken, kısmi korelasyonun yüksek olduğu durumlarda F, d'ye yakın sonuçlar vermektedir. Özellikle örnek çapı 50 iken iki test istatistiği arasında farkın çok az olduğu görülmüştür.

KAYNAKLAR

COX, D.R. and WERMUTH N., 1993, *Linear dependencies represented by chain graph (with discussion).* Statistical Science, 8, 204-218

DEMPSTER, A.P., 1972, *Covariance selection, Biometrics*, 28, 157-75

EDWARDS, D., 2001, *Introduction to graphical modelling*. 2nd Edition, Springer-Verlag, New York

ERIKSEN P.S.,1996, *Tests in covariance selection models.* Scand. Journal of Statistics. 23, 275-284

LAURITZEN, S.L. and WERMUTH, N., 1989, *Graphical models for associations between variables*, some of which are qualitative and quantative, Annals of Statistics, 17, 31-57

WERMUTH, N., 1980, *Linear recursive equations, covariance selections and path analysis*, Journal of American Statistical Asociation, 75, 963,72

WHİTTAKER, J., 1990, *Graphical models in applied multivariate statistics*, John Wiley and Sons, Chichester

WRIGHT, S., 1923, *The Theory of Path coefficients: a reply to Niles' criticism.* Genetics, 8, 239-55

COMPARISON OF DEVIANCE AND F STATISTICS USED IN GRAPHICAL MODELS CONTAINS CONTINUOUS VARIABLES

ABSTRACT

Graphical models for variables, which was distributed multinormal is called covariance selection models. Covariance selection models were determined by conditional independences. In this study, deviance and F, which is used to test conditional independences, is compared under different sample size, different number of variables and different conditional independence structures by simulation.

Key Words: *Conditional independence, Covariance selection model, Deviance, F-statistics*