İstatistik Araştırma Dergisi Cilt 02 No 01 Sayfa 79-87 Nisan 2003

Sürekli Değişken İçeren Grafiksel Modeller

Hülya BAYRAK* Fikri GÖKPINAR* Berrin ÖZKAYA**

ÖZET

Çok değişkenli normal dağılıma sahip değişkenler için elde edilen grafiksel modeller kovaryans seçimli modeller olarak adlandırılır. Kovaryans seçimli modeller değişken çiftlerinin koşullu bağımsızlığına dayanarak belirlenir. Bu çalışmada bir grafiksel modeldeki test işleminde kullanılan sapma istatistiğinin kovaryans seçimli modellerde irdelenmesi verildi. Uygulama olarak 30 farklı un örneği kullanıldı. Un örneklerinin protein miktarı, gluten miktarı ve sedimantasyon değerleri ile bu unlardan yapılan ekmeklerin, hacim verimi ve Dallman değerleri tesbit edildi. Elde edilen sonuçlar, kovaryans seçimli model ile irdelendi.

Anahtar kelimeler: *Kovaryans Seçimli Modeller, Sapma, Markov Özellikleri,Koşullu Bağımsızlık*

1. GİRİŞ

Çok değişkenli normal dağılıma sahip değişkenler için elde edilen grafiksel modeller kovaryans seçimli modeller olarak adlandırılır. Kovaryans seçimli modellerde, kovaryans yapısı, kovaryans matrisisinin tersinin elemanlarını sıfıra götürerek basitleştirilebilir. Kovaryans matrisinin tersinde sıfırların nasıl oluşturulacağını ortaya koyan bir kural Dempster(1972) tarafından verilmiştir.

Bir grafikte, V köşe kümesini, E köşeler arası kenar kümesini temsil eder. Eğer bir köşe çifti arası birden fazla kenar yoksa grafik basittir.

Bir kenar hem $(\alpha,\beta)\in E$ hem de $(\beta,\alpha)\in E$ şeklinde ise E yön verilmemiş kenar, $(\alpha,\beta)\in E$ şeklinde olup $(\beta,\alpha)\notin E$ ise yön verilmiş bir kenardır.

Bir 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.

* Gazi üniversitesi Fen Edebiyat Fakültesi İstatistik bölümü, Ankara, Türkiye, hbayrak@gazi.edu.tr-fikri@gazi.edu.tr

**Ankara üniversitesi Ziraat Fakültesi, Gıda Mühendisliği bölümü, Ankara, Türkiye. bozkaya@agri.ankara.edu.tr

α'dan β'ya bir ok varsa α'ya β'nın ailesi, β'ya da α'nın çocuğu denir. β'nın aile kümesini pa(β), α'nın çocuklar kümesini ch(α) şeklinde gösterilebilir.

α ile β arasında bir çizgi varsa α ve β'ya bitişik ya da komşu denir. α köşesinin komşu kümesi ne(α) şeklinde gösterilebilir.

Grafiksel modellerin temeli rasgele değişkenlerin koşullu bağımsızlığına dayanır. Bu modellerin grafikleri koşullu bağımsızlık ilişkilerini gösterir.

X, Y ve Z kesikli rasgele değişkenler olsun. X⊥Y/Z koşullu bağımsızlık ifadesi aşağıdaki gibi yazılır.

$$P(X=x,Y=y/Z=z)= P(X=x/Z=z)\ P(Y=y/Z=z)$$

Burada tüm z'ler için P(Z=z)>0'dır. X, Y, Z sürekli rasgele değişkenleri için X⊥Y/Z koşullu bağımsızlığı

$$X \perp Y/Z \Leftrightarrow f_{XY/Z}(x, y/z) = f_{X/Z}(x/z) f_{Y/Z}(y/z) \qquad (1)$$

biçiminde ifade edilir.

X⊥Y/Z ilişkisi aşağıdaki özelliklere sahiptir.

$$\left.\begin{array}{l} X \perp Y/Z \Rightarrow Y \perp X/Z\text{'dir.} \\ X \perp Y/Z \text{ ve } U=h(x) \Rightarrow U \perp Y/Z\text{'dir.} \\ X \perp Y/Z \text{ ve } U=h(x) \Rightarrow X \perp Y/(Z,U)\text{'dir.} \\ X \perp Y/Z \text{ ve } X \perp W/(Y,Z) \Rightarrow X \perp (W,Y)/Z\text{'dir.} \end{array}\right\} \qquad (2)$$

Rasgele değişkenler $(\chi_\alpha)_{\alpha \in V}$ koleksiyonu ve yön verilmemiş grafik için değişik markov özellikleri vardır. Bu özellikler aşağıdaki gibidir;

χ üzerindeki olasılık ölçümü *P* olmak üzere; (P) bitişik olmayan herhangi bir köşe çifti (α,β) için aşağıdaki koşul sağlanıyorsa ikili markov özelliğine sahiptir.

$$\alpha \perp \beta \ / \ V \backslash \{ \alpha, \beta \}$$

(L) herhangi bir köşe $\alpha \in V$ için aşağıdaki koşul sağlanıyorsa yerel markov özelliğine sahiptir.

$$\alpha \perp V \backslash cl(\alpha) \ / \ bd(\alpha)$$

(G) V'nin bağlantısız altkümelerin herhangi bir üçlüsü (A,B,S) aşağıdaki koşulu sağlıyorsa global markov özelliğine sahiptir.

$$A \perp B/S$$

$X=(X_1, X_2,...,X_p)$ vektörünün yoğunluk fonksiyonu $f_v(X_v)=f(X_1,...,X_p)$ mevcut ise, $f_v(x_v)$ yoğunluğu , G=(V,E) grafiğine göre aşağıdaki gibi faktorize edilebilir.

$$f_V(x_V)=\prod_{c\in C}\Psi_c(x_c) \tag{3}$$

Burada C, G'deki kliklerin kümesi ve $\Psi_c(x_c)$ x üzerinden x_c'ye bağlı negatif olmayan fonksiyonlardır. Bileşik yoğunluk $f_v(x_v)$ kesin pozitif ise faktorizasyon özelliği ve global markov özelliği denktir (Lauritzen,1996).

2 SÜREKLİ 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 olsun.

$$\mu=\begin{pmatrix}\mu_1\\ \mu_2\\ .\\ .\\ \mu_q\end{pmatrix} \tag{4}$$

ortalama vektörü ve

$$\Sigma=\begin{pmatrix}\sigma_{11} & \sigma_{12} & . & . & \sigma_{1q}\\ \sigma_{21} & \sigma_{22} & . & . & \sigma_{2q}\\ . & . & . & . & .\\ . & . & . & . & .\\ \sigma_{q1} & \sigma_{q2} & . & . & \sigma_{qq}\end{pmatrix} \tag{5}$$

kovaryans matrisine sahip çok değişkenli normal dağılıma sahiptir. Burada özellikle ilgilenilecek olan kovaryans matrisinin tersidir.

$$K=\Sigma^{-1}=\begin{pmatrix}w^{11} & w^{12} & . & . & w^{1q}\\ w^{21} & w^{22} & . & . & w^{2q}\\ . & . & . & . & .\\ . & . & . & . & .\\ w^{q1} & w^{q2} & . & . & w^{qq}\end{pmatrix} \tag{6}$$

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 'Eş.7'deki konsantrasyon matrisine sahip iki değişkenli normal dağılımdır.

$$\begin{pmatrix} w^{11} & w^{12} \\ w^{21} & w^{22} \end{pmatrix}^{-1} = \frac{1}{w^{11}w^{22} - (w^{12})^2} \begin{pmatrix} w^{22} & -w^{21} \\ -w^{12} & w^{11} \end{pmatrix} \tag{7}$$

bu iki değişkenli dağılıma ilişkin korelasyon katsayısı 'Eş 8'deki gibidir.

$$p^{12.34...q} = \frac{-w^{12}}{(w^{11}w^{22})^{\frac{1}{2}}} \tag{8}$$

Ayrıca

$$p^{12.34...q} = 0 \Leftrightarrow w^{12} = 0 \tag{9}$$

denkliği 'Eş 8''den açıkça görülür.

Bir başka deyişle, 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.

Bu sonucu bir başka yönden geliştirmek faydalı olur. Y'nin yoğunluğu

$$f(y) = \left|2\pi\Sigma_i\right|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(y-\mu)^1\Sigma^{-1}(y-\mu)\right\} \tag{10}$$

şeklinde yazılır. Bu yoğunluk düzenlenirse,

$$f(y) = \exp(g + h'y - \tfrac{1}{2}y'Ky) \tag{11}$$

ile ifade edilir. Burada $K=\Sigma^{-1}$, $h=\Sigma^{-1}\mu$ ve g normalleştirme sabiti olur ve

$$g = -\tfrac{1}{2}\ln\left|\Sigma\right| - \tfrac{1}{2}\mu'\Sigma\mu - \tfrac{q}{2}\ln(2\pi) \tag{12}$$

eşitliği ile verilir. Üstel aile terminolojisinde h ve K kanonik parametreler olarak adlandırılır(Edwards,2001). 'Eş. 11 tekrar dikkate alınarak,

$$f(y) = \exp\left(g + \sum_{j=1}^{q} h^j y_j - \tfrac{1}{2}\sum_{j=1}^{q}\sum_{k=1}^{q} w^{jk} y_j y_k\right) \tag{13}$$

şeklinde düzenlenebilir ve

$$Y_j \perp Y_k/(\text{geri kalan}) \Leftrightarrow w^{jk}=0 \tag{14}$$

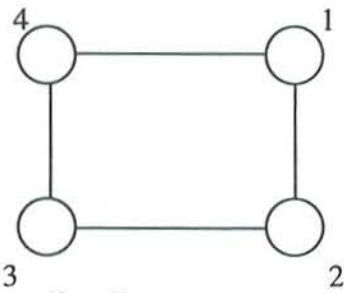
biçiminde ifade edilir.

Grafiksel Gauss modelleri, konsantrasyon matrisinin elemanları yani kısmi korelasyonların sıfır olmasıyla belirlenebilir. Örneğin q=4 iken modelde $w^{13}=w^{24}=0$ olduğu düşünülsün. Böylece konsantrasyon matrisi

$$K = \Sigma^{-1} = \begin{pmatrix} w^{11} & w^{12} & 0 & w^{14} \\ w^{21} & w^{22} & w^{23} & 0 \\ 0 & w^{23} & w^{33} & w^{34} \\ w^{41} & 0 & w^{43} & w^{44} \end{pmatrix} \tag{15}$$

şeklinde yazılır.

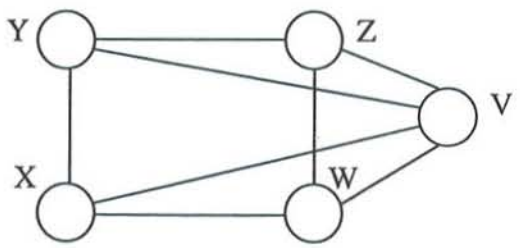
Bu modelin grafiği, sıfır olmayan kısmi korelasyonlara karşılık gelen çiftler arasında birer kenar konmasıyla edilir. Böyle bir modele ilişkin grafik aşağıdaki gibidir.



Şekil 1. $w^{13}=w^{24}=0$ Durumundaki Grafik

Grafiksel Gauss modellerin formüllendirilmesi için değişkenleri sayı yerine harf ile ve değişkenler kümesini Γ ile gösterilecektir.

Kesikli değişkenler içeren grafiksel modeller için olduğu gibi, model formülü, grafiğin klikleriyle verilen değişkenler kümesinin (üreteçler) listesini içerir. Örneğin Şekil 2'deki grafik göz önüne alınsın.



Şekil 2. VWX, VWZ, VZY,VXY Kiklerine Sahip Olan Grafik

Kayıp kenarlar WY ve XY'dir. Yani w^{wy} ve w^{xz} sıfır olarak belirlenir. Grafiklerin klikleri {V,W,X},{ V,W,Z},{V,Z,Y},{V,X,Y}'dir. Böylece modelin formülü

VWX, VWZ, VZY,VXY

şeklindedir.

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.

2.1 Sürekli Modeller için Olabilirlik

N gözlemlik $y^{(1)},\ldots. y^{(N)}$ bir örnek alınsın. $\bar{y} = \sum_{k=1}^{N} y^{(k)} / N$ örnek ortalaması vektörüdür. $S = \sum_{k=1}^{N} (y^{(k)} - \bar{y})(y^{(k)} - \bar{y})' / N$

örnek kovaryans matrisi 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

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

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.

$$l(\mu, K) = -Nq\ln(2\pi)/2 - N\ln|\Sigma|/2 - Ntr(KS)/2 - N(\bar{y} - \mu)' K (\bar{y} - \mu)/2 \qquad (16)$$

$a \subseteq \Gamma$ 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 'Eş. 16'daki son terim yok olur ve $\hat{\Sigma}$ ve S, $w^{ij}=0$ olan elemanlar için kesin olarak farklılık gösteriyorsa $tr(\hat{K}S) = tr(\hat{K}\hat{\Sigma}) = q$ olur. Böylece model altında maksimize edilen log-olabilirlik

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

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

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

dir.

Grafiksel modellerde bir kenarın çıkarılıp çıkarılmayacağına ilişkin testte sapma istatistiğinden faydalanılır. Bir M_0 modelinin sapması M_f kısıtsız(doymuş) modeline karşı M_0'ın olabilirlik oranı testidir. Böylece modelin sapması

$$\begin{aligned} G^2 &= 2(\hat{l}_f - \hat{l}_m) \\ &= N\ln(|\hat{\Sigma}|/|S| \end{aligned} \qquad (17)$$

olur. $M_0 \subseteq M_1$ için sapma farkı

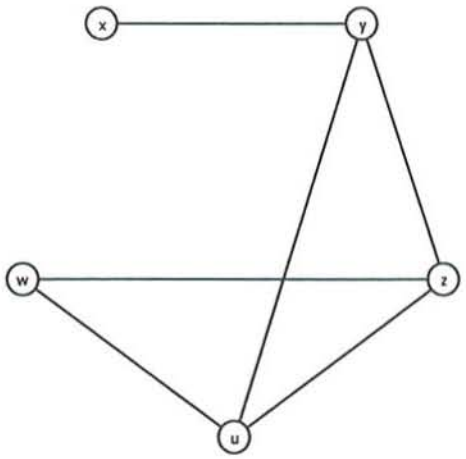
$$d = \ln\left(|\hat{\Sigma}_0|/|\hat{\Sigma}_1|\right)$$

dir (Edwards,2001). $\hat{\Sigma}_0$ ve $\hat{\Sigma}_1$, sırasıyla M_0 ve M_1 altında Σ'ın tahminleridir. M_0 altında d, M_0 ile M_1 arasında parametre sayılarındaki (köşe sayısı) fark kadar serbestlik dereceli asimptotik χ^2 dağılımıdır.

3 UYGULAMA

Ülkemizde günlük kalorinin sağlanmasında ve halkın beslenmesinde ekmeğin çok önemli bir yeri olmasına karşın ülkemizde üretilen ekmeklerin kalitesi istenen düzeyde değildir. Bunun en önemli nedenlerinden birisi kullanılan buğdayların teknolojik kalitesinin yeterli olmamasıdır. Buğday ve bunlardan elde edilen unların teknolojik kalitelerini belirleyen faktörlerin başında protein ve gluten miktarları ile bunların kalitesi gelir. Protein miktarları birbirine eşit olan buğdayların teknolojik özellikleri arasında da önemli farklar olabilir. Bu çoğu kez gluten miktar ve kalitesindeki farklılıktan kaynaklanır. Bu uygulamada 30 farklı un örneği kullanılmıştır. Un örneklerinin protein miktarı (x), gluten miktarı (y) ve sedimentasyon değerleri (z) ile bu unlardan yapılan ekmeklerin, hacim verimi (w) ve Dallman değerleri (u) tesbit edilmiştir(Özkaya,B. ve diğerleri).

Un ve ekmeğin kalitesini etkileyen kriterler arasında nasıl bir ilişki olduğu ve bu ilişkinin özelliği Kovaryans seçimli model ile incelenmiş ve bu kriterler arasındaki ilişkileri gösteren grafiksel model Şekil 3'de verilmiştir.



Şekil 3. Unun Ekmeklik Kalitesini Etkileyen Faktörler Arası İlişkileri Gösteren Model

Şekil 3 incelendiğinde unun gluten miktarı ve sedimentasyon değeri ile bu undan yapılan ekmeğin Dallman değeri arasında tam bir ilişki bulunmaktadır.

Un örneklerinin protein miktarı (x), gluten miktarı (y) ve sedimentasyon değerleri (z) ile bu unlardan yapılan ekmeklerin, hacim verimi (w) ve Dallman değerleri (u) değişkenlerine ilişkin koşullu bağımsızlıkları gösteren Markov özellikleri Tablo 1'deki gibi yazılabilir.

Tablo 1. Grafiğe İlişkin Koşullu Bağımsızlıklar

İkili Markov Özelliği	Yerel Markov Özelliği	Global Markov Özelliği
$x \perp w/(u,y,z)$	$x \perp (w,u,z)/y$	$(x,y) \perp w)/(u,z)$
$x \perp z/(u,y,w)$	$y \perp w/(x,z,u)$	
$x \perp u/(w,y,z)$		
$y \perp w/(x,z,u)$		

Tablo 1'e göre ikili, yerel ve global markov özellikleri dikkate alındığında şu yorumlar yapılabilir: Unun protein miktarının, gluten miktarı verildiğinde, sedimantasyon değerinden ve bu undan yapılan ekmeğin hacim veriminden koşullu olarak bağımsız olduğu, aynı zamanda unun protein miktarının, gluten miktarı verildiğinde, bu undan yapılan ekmeğin Dallman değerinden de koşullu olarak bağımsız olduğu görülmektedir. Yani unun protein miktarı verildiğinde sedimantasyon değeri hakkında bir değerlendirme yapabilmek için unun gluten miktarının bilinmesine gereksinim bulunmaktadır. Global markov özelliğinden dalman değeri ve sedimantasyon değeri verildiğinde unun protein miktarının ve gluten miktarının bu

unlardan yapılan ekmeklerin hacim veriminden koşullu olarak bağımsız olduğu söylenebilir.

Bu çalışma sonuçlarının kovaryans seçimli model ile irdelenmesinin, un ve ekmek kalitesini gösteren kriterler arasında ilişki olup olmadığını belirlemek bakımından olduğu kadar unun basit yöntemlerle tespit edilen bazı kimyasal ve fizikokimyasal özelliklerine bakılarak ekmeğin kalitesini tahmin edebilme açısından da yararlı olacağı düşünülmektedir.

KAYNAKLAR

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

LAURITZEN S.L., (1996), *Graphical models*, Clarendon press, London

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

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

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

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

ÖZKAYA, B., BAYRAK, H., GÖKPINAR F., (2002), *Unun Ekmeklik Kalitesini Etkileyen Bazı Faktörler Arasındaki İlişkilerin Kovaryans Seçimli Model İle Belirlenmesi*, Hububat 2002 Hububat işleme teknolojileri kongre ve sergisi, bildiriler kitabı, 549-555 Gaziantep

Graphical Models for 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, which is used to test conditional independences, was given. Thirty different flour samples were used in this study. Protein, wet gluten and sedimentation values of flour samples along with the loaf volume and Dallman values of bread made of these flours were investigated. Covariance selection models evaluated the obtained results

Keywords: *Covariance Selection Model, Deviance, Markov Properties, Conditional Independence*