İstatistik Araştırma Dergisi Cilt 02 No 02 Özel Sayı Sayfa 17-30 Ağustos 2003

Bant Genişliği Seçiminde Kullanılan Yöntemlerin Simetrik ve Simetrik Olmayan Dağılımlarda Karşılaştırılması

Derya GÖKMEN* Öniz TOKTAMIŞ**

ÖZET

Bu çalışmada olasılık yoğunluk fonksiyonunun çekirdek kestirimi için, bant genişliği seçiminde kullanılan "en küçük kareler çapraz geçerlilik", "yanlı çapraz geçerlilik", "düzleştirilmiş bootstrap" ve "plug-in" yöntemleri hakkında bilgi verilmiş, yapılan uygulama ile simetrik ve simetrik olmayan dağılımlardan alınan farklı büyüklüklerdeki örneklemlerde dört yöntem ile bant genişliği değerleri elde edilmiş, optimal bant genişliği değeri ile karşılaştırılmıştır.

Anahtar Kelimeler: *Çekirdek (kernel) kestirimi, bant genişliği, en küçük kareler çapraz geçerlilik, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap, plug-in.*

1. GİRİŞ

Uygulamalı istatistikte güncelliğini koruyan en önemli konulardan biri, olasılık yoğunluk fonksiyonunun kestirimidir. Olasılık yoğunluk fonksiyonunun kestirimi için kullanılan parametrik olmayan yöntemlerden biri olan ve ilk kez 1956 yılında Rosenblatt tarafından ortaya konulan "çekirdek kestirim yöntemi", üzerinde en çok çalışılan yöntemlerden biri olup, sahip olduğu özellikler bakımdan matematiksel olarak diğerlerinden daha iyi geliştirilmiştir.

$X_1, X_2, \ldots X_n$ bilinmeyen bir olasılık yoğunluk fonksiyonu f 'den alınan rasgele bir örneklem olmak üzere, herhangi bir x noktası için elde edilen çekirdek kestiricisi,

$$\hat{f}(x,h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) \qquad (1)$$

biçiminde verilmektedir. Burada K; $\int_{-\infty}^{\infty} K(u)du = 1$ koşulunu sağlayan çekirdek (kernel) fonksiyonu, h; pencere genişliği ve bant genişliği adlarını da alan düzleştirme parametresi, $\hat{f}(x,h)$; $X_1,\ldots,X_n$ gözlem değerlerine bağlı olduğu için bir raslantı değişkenidir.

* Ankara Üniversitesi, Tıp Fakültesi, Biyoistatistik Anabilim Dalı Araştırma Görevlisi, ANKARA
** Hacettepe Üniversitesi, Fen Fak., İstatistik Bölümü, Öğretim Üyesi, Beytepe-ANKARA

Olasılık yoğunluk fonksiyonunun çekirdek kestiriminde, çekirdek fonksiyonunun seçimi, bant genişliğinin seçimi kadar önemli olmayıp, hesaplama kolaylığı ve türevlenebilme özelliklerine göre yapılmaktadır. Çekirdek kestiriminde bant genişliğinin seçiminin önemli bir yeri vardır. Yapılan araştırmalar, bant genişliği değerindeki küçük bir değişikliğin, kestirimler üzerinde büyük değişiklikler meydana getirdiğini göstermiştir.

Çekirdek kestiricisinin performansını değerlendirmek için çeşitli ölçütler üzerinde durulmuştur. Bunlardan en yaygın olarak kullanılanları ilk kez Rosenblatt tarafından önerilen toplanmış hata kareler ortalaması (THKO) ve asimtotik toplanmış hata kareler ortalaması (ATHKO)'dır (Rosenblatt,1956).

Olasılık yoğunluk fonksiyonunun çekirdek kestiricisine ilişkin THKO,

$$THKO(h) = \int E\left[\hat{f}(x,h) - f(x)\right]^2 dx \qquad (2)$$

ve daha açık olarak,

$$THKO(h) = (nh)^{-1}\int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4}h^4\left\{\int_{-\infty}^{\infty} u^2 K(u) du\right\}^2 \int_{-\infty}^{\infty} f''(x)^2 dx + o\left\{(nh)^{-1} + h^4\right\} \qquad (3)$$

biçimindedir. Burada $u = \frac{x - X_i}{h}$'dir. Asimtotik toplanmış hata kareler ortalaması (ATHKO), Eşitlik 3'den yararlanılarak aşağıdaki biçimde yazılabilir:

$$ATHKO(h) = (nh)^{-1}\int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4}h^4\left[\int_{-\infty}^{\infty} u^2 K(u) du\right]^2 \int_{-\infty}^{\infty} f''(x)^2 dx \qquad (4)$$

THKO ve ATHKO ifadelerini minimum yapan h değeri, optimal bant genişliğidir. THKO ve ATHKO ifadelerini minimum yapan bant genişlikleri aşağıda verilmiştir:

$$h_{THKO} \cong \mu_2(K)^{-2/5}\left\{\int_{-\infty}^{\infty} K(u)^2 du\right\}^{1/5}\left\{\int_{-\infty}^{\infty} f''(x)^2 dx\right\}^{-1/5} n^{-1/5} \qquad (5)$$

$$h_{ATHKO} = \mu_2(K)^{-2/5}\left\{\int_{-\infty}^{\infty} K(u)^2 du\right\}^{1/5}\left\{\int_{-\infty}^{\infty} f''(x)^2 dx\right\}^{-1/5} n^{-1/5} \qquad (6)$$

Burada $\mu_2(K) = \int u^2 K(u)\, du$'dır. Eşitlik 5 ve 6'de verilen ifadelerde görüldüğü gibi optimal bant genişliği, bilinmeyen olasılık yoğunluk fonksiyonunun ikinci mertebeden türevinin karesinin integraline bağlı olup, bu eşitliklerden elde edilemez. Bu nedenle bant genişliğini elde etmek için çeşitli yöntemler önerilmiştir. Ancak genel

anlamda kabul görmüş bir yöntem bulunmamaktadır. Literatürde sık karşılaşılan bazı yöntemler aşağıda verilmektedir.

En Küçük Kareler Çapraz Geçerlilik (ÇG) Yöntemi

Bant genişliği seçiminde kullanılan yöntemler arasında, üzerinde en çok çalışılanı, en küçük kareler çapraz-geçerlilik yöntemidir. Bu yöntem, 1982 yılında Rudemo ve 1984 yılında Bowman tarafından önerilmiştir. En küçük kareler çapraz geçerlilik fonksiyonu ÇG(h),

$$ÇG(h) = \int \hat{f}(x,h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i,h) \qquad (7)$$

ile verilmektedir (Bowman,1984). Burada $\hat{f}_{-i}(X_i,h)$, x_i gözlemi dışarıda kalmak üzere diğer gözlem değerlerinden elde edilen çekirdek kestirimidir. Eşitlik 7 ile verilen ÇG(h) fonksiyonu, $E\int\left[\hat{f}(x,h)-f(x)\right]^2 dx + \int[f(x)]^2 dx = THKO(h) - \int[f(x)]^2 dx$ ifadesinin yansız bir kestirimidir (Cao et.al., 1994). ÇG(h) fonksiyonunu minimum yapan bant genişliği aynı zamanda THKO'sını da minimum yapan bant genişliğidir. En küçük kareler çapraz geçerlilik fonksiyonunu minimum yapan bant genişliğinin seçimi, simülasyon çalışması ile yapılmaktadır.

En küçük kareler çapraz geçerlilik fonksiyonu, birden fazla minimuma sahip olup, yapılan çalışmalar, ÇG(h) fonksiyonunu minimum yapan en büyük bant genişliği değerinin, optimal bant genişliği değerine daha yakın olduğunu göstermiştir. Ayrıca ÇG(h) fonksiyonundan elde edilen bant genişliğinin değişkenliği büyüktür (Wand and Jones, 1995).

Yanlı Çapraz Geçerlilik (YÇG) Yöntemi

Scott ve Terrell (1987) tarafından verilen yanlı çapraz geçerlilik yönteminde, performans kriteri olarak ATHKO kullanılmaktadır. Eşitlik 4'te verilen ATHKO değeri de THKO değeri gibi bilinmeyen bir nicelik olan $\int f''(x)^2\,dx = R(f'')$ ifadesine bağlıdır. Bunun için doğal bir kestirici $R(\hat{f}'')$'dır. Burada $\hat{f}$, bir çekirdek kestiricisidir. Scott ve Terrell bu kestiricinin asimtotik olarak yetersiz olduğunu, ATHKO ifadesinde f 'nin ikinci mertebeden türevinin karesinin integrali yerine, onun düzleştirilmiş bir biçimi olan,

$$\int_{-\infty}^{\infty} \tilde{f}''(x,h)^2 dx = \int_{-\infty}^{\infty} \hat{f}''(x,h)^2 dx - \frac{1}{nh^5}\int_{-\infty}^{\infty} K''(u)^2 du$$

ifadesinin kullanılması ile elde edilen kestiricinin daha iyi olduğunu belirtmişler ve bu eşitliği ATHKO ifadesinde yerine koyarak, yanlı çapraz geçerlilik fonksiyonunu,

$$YÇG(h) = \frac{1}{nh}\int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4} h^4 \mu_2(K)^2 \int_{-\infty}^{\infty} \tilde{f}''(x,h)^2 dx \qquad (8)$$

olarak elde etmişlerdir. Bu fonksiyonu minimum yapan bant genişliği değeri simülasyon yoluyla seçilir. Bu yöntem, çapraz geçerlilik ve ileride bahsedilecek olan plug-in yönteminin karışımıdır (Park and Marron, 1990).

ÇG ve YÇG yöntemlerinden hangisinin kullanılması gerektiğine karar verme problemi, verilen duruma bağlı olabilir. Olasılık yoğunluk fonksiyonu, oldukça çarpık bir dağılıma sahip ise, ÇG yöntemini kullanabiliriz. Diğer yandan simetrik bir olasılık yoğunluk fonksiyonu kullanıldığında,YÇG yöntemi uygun görülebilir (Härdle, 1991).

Düzleştirilmiş Bootstrap (B) Yöntemi

Bootstrap yöntemi, ana örneklemi kitle gibi kabul edip, ana örneklemden seçileni yerine koyarak çok sayıda örneklemin çekilmesi ve bu yeniden çekilen örneklemlerin her birinden ilgili tahmin edicilerin hesaplanması esasına dayanmaktadır. Eşitlik 2 ile verilen toplanmış hata kareler ortalaması, "yanın karesi" ve "varyans" bileşenlerine ayrılabilmektedir. Bu ifadedeki varyans terimi, alışılmış bootstrap yöntemi ile kestirilebilirken, yan terimi kestirilemez. Bu sorunu gidermek amacıyla Taylor(1989) ile Faraway ve Jhun (1990), "düzleştirilmiş bootstrap yöntemini" kullanmışlardır.

Faraway ve Jhun, bu yöntem için önce en küçük kareler çapraz geçerlilik yöntemi ile önsel bir bant genişliği "g" seçip, bu bant genişliği kullanılarak elde edilen çekirdek kestirimlerinden, bir algoritma ile bootstrap örneklemlerini oluşturmuşlardır (Silverman,1986). $X_1^*, X_2^*, \ldots X_n^*$, bir bootstrap örneklemi ve B, çekilen bootstrap örneklemlerinin sayısı olmak üzere, toplanmış hata kareler ortalamasının bootstrap kestirimi B(h),

$$B(h) = \frac{1}{B}\sum_{j=1}^{B}\int\left[\hat{f}_i^*(x,h) - \hat{f}(x,g)\right]^2 dx$$

biçimindedir. Burada $\hat{f}_i^*(x,h)$, i. bootstrap örnekleminden elde edilen çekirdek kestirimi, $\hat{f}(x,g)$, orijinal örneklem için çekirdek kestirimidir. B(h)'yı minimum yapan h bant genişliği değeri simülasyon ile elde edilmektedir (Faraway and Jhun,1990).

Taylor (1989), h=g alındığında ve Gaussian çekirdek fonksiyonu kullanıldığında yeniden örneklem seçimine gerek kalmadığını belirtmiş ve THKO'nın bootstrap kestirimini aşağıdaki gibi elde etmiştir:

$$B(h) = \frac{1}{2n^2 h\sqrt{2\pi}}\left[\sum_{i,j}\exp\left\{-\frac{(X_j - X_i)^2}{8h^2}\right\} - \frac{4}{\sqrt{3}}\sum_{i,j}\exp\left\{-\frac{(X_j - X_i)^2}{6h^2}\right\} + \sqrt{2}\sum_{i,j}\exp\left\{-\frac{(X_j - X_i)^2}{4h^2}\right\} + n\sqrt{2}\right] \qquad (9)$$

Eşitlik 9'dan görüldüğü gibi bu ifade yalnız gözlem değerlerine bağlıdır. Bu fonksiyonu minimum yapan h bant genişliği simülasyon yoluyla bulunur (Taylor, 1989). Çoğu zaman, küçük örneklemlerde bootstrap fonksiyonu birden fazla minimuma sahip

olabilir. Bu nedenle, düzleştirilmiş bootstrap yönteminin, örneklem hacmi 100'den büyük olduğu durumlarda kullanılması önerilmektedir.

Plug-in Yöntemi

Plug-in bant genişliği seçicileri, asimtotik olarak optimal bant genişliği için verilen formülde ortaya çıkan bilinmeyen nicelikler yerine, türev fonksiyonellerinin ($\psi_r = E\{f^{(r)}(x)\}$) kestirimlerinin devreye sokulması fikrine dayanır. ψ_r fonksiyonelinin çekirdek kestiricisi $\hat{\psi}_r$ için genel ifade,

$$\hat{\psi}_r(g) = n^{-1}\sum_{i=1}^{n}\hat{f}^{(r)}(X_i, g) = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}L_g^{(r)}(X_i - X_j) \qquad (10)$$

olarak verilmiştir (Wand and Jones, 1995). Burada $L_g = \frac{1}{g}L\left(\frac{X_i - X_j}{g}\right)$, L; genellikle K'dan farklı bir çekirdek fonksiyonu ve g; h'dan farklı bir bant genişliğidir. L, r türeve sahip k mertebeli (k=2,4,...) simetrik bir çekirdek fonksiyonu olmak üzere, asimtotik hata kareler ortalamasından g bant genişliği aşağıdaki biçimde elde edilir:

$$g_{AHKO} = \left[\frac{k!\,L^{(r)}(0)}{-\mu_k(L)\psi_{r+k}n}\right]^{1/(r+k+1)} \qquad (11)$$

Eşitlik 5'te, ikinci mertebeden çekirdek fonksiyonu kullanıldığında (k=2 ve $\mu_0(L) = 1$ $\mu_1(L) = 0, \mu_2(L) \neq 0$) $\int_{-\infty}^{\infty} f''(x)^2 dx = R(f'')$ ifadesinin ψ_4'e eşit olduğu görülür (Cula,1998). Bu durumda k= r= 2 olmak üzere ATHKO'ndan elde edilen bant genişliği,

$$h_{ATHKO} = \left\{\frac{R(L)}{\mu_2(L)^2\psi_4 n}\right\}^{1/5} \qquad (12)$$

biçiminde yazılabilir. ψ_4'ün, çekirdek kestiricisi $\hat{\psi}_4(g)$ ile yer değiştirmesi plug-in yönteminin temelini oluşturur (Wand and Jones, 1995). $\hat{\psi}_4(g)$, Eşitlik 12'de yerine konularak, bant genişliğinin plug-in kestiricisi $\hat{h}_{PI}$,

$$\hat{h}_{PI} = \left\{\frac{R(L)}{\mu_2(L)^2\hat{\psi}_4(g)n}\right\}^{1/5} \qquad (13)$$

olarak elde edilir. Park ve Marron (1990), $\hat{\psi}_4(g)$ kestiricisi yerine Eşitlik 10'daki ifadede $i \neq j$ için toplam alarak aşağıdaki kestiriciyi elde etmişlerdir:

$$\hat{\psi}_4(g) = \frac{1}{n^2 g^5}\sum\sum_{i\neq j} L\left(\frac{X_i - X_j}{g}\right)$$

Sheather ve Jones (1991), $\hat{\psi}_4(g)$ kestiricisi yerine Eşitlik 10'daki ifadede $i = j$ durumunu da toplama ekleyerek aşağıdaki kestiriciyi elde etmişlerdir:

$$\hat{\psi}_4(g) = \frac{1}{n^2 g^5} \sum_{i=1} \sum_{j=1} L\left(\frac{X_i - X_j}{g}\right)$$

Sheather ve Jones'un önerdiği kestiriciden elde edilen bant genişliği, Park ve Marron'un önerdiği kestiriciden elde edilen bant genişliğinden kuramsal olarak daha iyi performansa sahip olup, değişkenliği daha azdır (Sheather and Jones, 1991).

Plug-in yöntemi tamamen otomatik bir yöntem değildir, çünkü $\hat{h}_{PI}$, pilot bant genişliği g'nin seçimine bağlıdır. g'nin seçiminin bir yolu, $\hat{\psi}_4(g)$'nin kestirimi için AHKO-optimal bant genişliği formülüne başvurmaktır. Aynı ikinci mertebeden çekirdek fonksiyonu L, $\hat{\psi}_4(g)$'de kullanılırsa Eşitlik 11'den, AHKO-optimal bant genişliği,

$$g_{AHKO} = \left[\frac{2L^{(iv)}(0)}{-\mu_2(L)\psi_6 n}\right]^{1/7}$$

olarak elde edilir. Bununla birlikte, yukarıdaki ifade, $\hat{h}_{PI}$'daki gibi aynı kusura sahip olup, ψ_6 ile gösterilen bilinmeyen bir yoğunluk fonksiyoneline bağlıdır. ψ_6'yı diğer bir çekirdek kestirimini kullanarak tahmin edebiliriz, fakat onun optimal bant genişliği ψ_8'e bağlıdır. Bu problem, Eşitlik 11'den görüldüğü gibi ψ_r'nin kestirimi için verilen optimal bant genişliği, ψ_{r+2}'ye bağlı olduğu için ortadan kalkmayacaktır. Bu problemin üstesinden gelmek için alışılmış bir strateji, bir ψ_r fonksiyonelinin hızlı ve basit bir yöntem ile, örneğin standart bir dağılım kullanılarak, kestirilmesidir. Kuramsal çalışmalar fonksiyonel kestiriminin aşama sayısının "l", en az 2 olarak alınmasını desteklemekte olup, uygulamada bu değer yaygın olarak kullanılmaktadır. l'nin sayısı arttıkça bant genişliği seçicisi daha az yanlı olmakta, ancak seçicinin değişkenliği artmaktadır.

Plug-in yönteminin genelde, temel yoğunluk fonksiyonu yeterince düzgün olduğunda en etkili, fakat yeterli düzgünlük olmadığında daha az güçlü olduğu görülür. Loader (1999)'ın yaptığı çalışmada elde ettiği bulgular, plug-in yöntemlerini çeşitli açılardan değerlendirir. İlk olarak, plug-in yöntemleri, pilot bant genişliklerinin keyfi seçimine bağlıdır ve bu seçim yanlış olduğunda başarısız olur. İkinci olarak, üzerinde sıkça durulan çapraz geçerlilik yönteminin, kestirimleri az düzleştirilmiş olarak vermesi, bant genişliği seçiminin belirsizliğini yansıtır: plug-in yöntemleri bu belirsizliği zor problemler verildiğinde kestirimleri çok düzleştirerek ve önemli özellikleri göz ardı ederek yansıtır (Loader, 1999).

2. BANT GENİŞLİĞİ SEÇİM YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Bant genişliği seçim yöntemlerini karşılaştırmak için bir yaklaşım, yakınsama hızı kavramının kullanılmasıdır. Bu kavram, örneklem hacmi artarken kestiricinin hedef değerine ne kadar hızlı yaklaştığını göstermektedir ve çeşitli kestiricileri karşılaştırmak için çok yararlı olabilir. Yakınsama hızı asimtotik bir kavramdır ve asimtotik ifadeler büyük örneklem genişlikleri için geçerlidir. Bu nedenle, bu kavram küçük örneklem hacimleri ile yapılan uygulamalarda yeterli olmamaktadır. Küçük örneklemler için karşılaştırmalar benzetim çalışmalarına göre yapılmaktadır.

ÇG ve YÇG yöntemleri, oldukça zayıf koşullar altında optimuma yakınsama özelliğine sahiptir (Wand and Jones, 1995). Ayrıca ÇG yöntemi, büyük örneklem değişkenliğine sahip olduğu için, çoğu benzetim çalışmalarında ve gerçek veri örneklerinde bu yöntemin performansının hayal kırıklığına uğratıcı olduğu ortaya çıkmıştır. Yapılan çalışmalar YÇG yönteminden elde edilen bant genişliğinin, ÇG yönteminden elde edilen bant genişliğinden daha büyük değer aldığını göstermiştir. Ancak, Scott ve Terrell'e göre YÇG yönteminden elde edilen bant genişliğinin çekici tarafı, ÇG yöntemine göre asimtotik varyansının daha düşük olmasıdır ki, bu durumda YÇG yönteminden elde edilen bant genişliğinin daha kararlı olduğu söylenebilir. Bunun yanı sıra, ÇG ve YÇG fonksiyonları birden fazla yerel minimuma sahip olabilir. Yapılan çalışmalar bu durumda en büyük yerel minimuma sahip olan bant genişliğinin alınmasının uygun olduğunu ortaya koymuştur. Çünkü en büyük yerel minimumu veren h bant genişliği değeri, THKO'dan elde edilen optimal bant genişliğine en yakın olanıdır (Wand and Jones, 1995).

Düzleştirilmiş bootstrap yöntemi, büyük örneklem genişlikleri için, çoğu dağılımda çapraz geçerlilik yönteminden daha iyi işler. Düzleştirilmiş bootstrap yöntemi genelde, çapraz geçerlilikten daha büyük bir bant genişliği seçer, ancak bu bant genişliğinin değişkenliği daha azdır. Değişkenliğin daha az olması, bu yöntemin ÇG yönteminden daha üstün olduğunu göstermektedir. Ancak bootstrap yönteminin, ÇG yöntemine göre dezavantajı, hesaplama maliyetinin yüksek, işlemlerinin zaman alıcı olmasıdır (Taylor, 1989; Faraway and Jhun ,1990).

Düzleştirilmiş bootstrap ve plug-in yöntemleri, ÇG ve YÇG yöntemlerinden daha hızlı olarak optimuma yakınsama özelliğine sahiptirler. Ayrıca, plug-in yönteminden elde edilen bant genişliği seçicilerinin değişkenliği, ÇG ve YÇG yöntemlerinden elde edilen bant genişliklerininkinden daha azdır. Bu nedenle, bu bant genişliği, yeterince büyük örneklemler için her zaman üstündür.

Plug-in yöntemini önerenler, klasik yaklaşımlara eleştiride bulunmuşlardır. Örneğin, Park ve Marron (1990), çoğu benzetim çalışmasında ve gerçek veri örneklerinde, en küçük kareler çapraz geçerlilik yönteminin performansının hayal kırıklığına uğratıcı olduğunu, en küçük kareler çapraz geçerlilik yönteminden elde edilen bant genişliğinin değişkenliğinin büyük olmasından dolayı, bant genişliğinin seçimi için yeni arayışlara yönelmeler olduğunu ve bunların içinde en çok tutulanlarının plug-in ve yanlı çapraz geçerlilik yöntemleri olduğunu belirtmişlerdir. Loader (1999), plug-in yönteminin, pilot kestirimlerin belirlenmesi sırasında, gerekli bant genişliği hakkında etkili olarak önemli varsayımlar yaptığını ve bu bilgi yanlış olduğunda başarısızlığa uğradığını, bu yöntemin bilgilerinin çoğunu yüksek mertebeden pilot kestirimlerin kullanımı sayesinde veriden elde ettiğini, eğer klasik yaklaşımların,

yüksek mertebeden fonksiyonları incelemesine izin verilirse, daha iyi kestirimler bulunabileceğini ve plug-in yöntemlerinin daha iyi yakınsama hızları göstererek asimtotik analiz ile kurtarılamayacağını belirtmiştir.

3. UYGULAMA

Çekirdek kestirim yöntemi, olasılık yoğunluk fonksiyonunun kestirimi için verilen parametrik olmayan yöntemlerden biri olup, örneklemin alındığı kitlenin dağılımının bilinmediği durumda kullanılmaktadır.

Toktamış et.al. (1999) tarafından yapılan çalışmada, en küçük kareler çapraz geçerlilik, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap yöntemleri incelenmiş; simetrik dağılımlardan alınan örneklemler için en küçük kareler çapraz geçerlilik yönteminin diğerlerine göre daha iyi sonuç verdiği, simetrik olmayan dağılımlardan alınan örneklemler için de yanlı çapraz geçerlilik yönteminin diğerlerine göre daha iyi sonuç verdiği görülmüştür.

Bu çalışmada, en küçük kareler çapraz geçerlilik, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap yöntemlerine plug-in yöntemi de eklenmiş, bu yöntemlerden elde edilen bant genişliklerinin, optimal bant genişliği değerine ne kadar yakın olduğu araştırılmıştır.

Bu işlemler için öncelikle, SPSS paket programı kullanılarak, simetrik ve simetrik olmayan dağılımlardan 50, 100, 250 ve 500 birimden oluşan örneklemler seçilmiştir. 50, 100 ve 250 birimlik örneklemler için 100'er, 500 birimlik örneklemler için 50'şer tekrar yapılmıştır. Örneklemlerin çekildiği ***simetrik dağılımlar,*** "N(50,4) normal dağılım", "N(0,1) normal dağılım" ile parametreleri "$\alpha=3, \beta=1$ olan gamma dağılımı"; ***simetrik olmayan dağılımlar,*** parametreleri "$\alpha=2, \beta=1$ olan gamma dağılımı", "parametresi $\lambda=1$ olan üstel dağılım" ve parametreleri "k=3, $\alpha=2$ olan pareto dağılımı"dır. En küçük kareler çapraz geçerlilik, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap yöntemleri için bant genişliği değerleri, Toktamış et.al. (1999) tarafından yapılan çalışmadaki bilgisayar programları kullanılarak elde edilmiştir. Plug-in yöntemi için bant genişliği değeri elde edilirken, çekirdek fonksiyonu K=L, fonksiyonel kestiriminin aşama sayısı 2 alınmış, çekirdek kestiricisi olarak da Sheather ile Jones tarafından önerilen kestirici kullanılmıştır. Gökmen(2002) tarafından hazırlanan bilgisayar programları yardımıyla plug-in bant genişliği değerleri elde edilmiştir. Her bir dağılım için, dört yöntem kullanılarak elde edilen bant genişliklerinin ortalama ve varyansı bulunmuş, ortalama bant genişliği değeri, optimal bant genişliği değeri ile karşılaştırılmıştır. Elde edilen sonuçlar simetrik ve simetrik olmayan dağılımlar için ayrı ayrı değerlendirilmiştir.

Simetrik Dağılımlar İçin Uygulama Sonuçları

Normal dağılım

Ortalaması 50 ve varyansı 4 olan normal dağılımdan örneklem büyüklüğü 50, 100, 250 ve 500 olan rasgele örneklemler alınmıştır. Çekirdek fonksiyonu standart normal dağılım alınarak, çapraz geçerlilik, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap ve plug-in yöntemleri ile bant genişliği değerleri bulunmuş ve sonuçlar Tablo 1'de sunulmuştur.

Tablo 1. Normal dağılımdan alınan örneklemler için dört yöntem ile elde edilen bant genişliği değerlerinin ortalama ve varyansları

Örneklem Büyüklüğü	Optimal Bant Genişliği	Bant Genişliklerinin Dağılımına Ait	Yöntemler			
			ÇG	YÇG	B	PLUG-IN
50	0,969240	Ortalama	1,207000	1,095900	1,427475	0,939700
		Varyans	0,037588	0,066970	0,020523	0,019352
100	0,843773	Ortalama	1,048750	0,980500	1,079900	0,822100
		Varyans	0,012256	0,028750	0,007662	0,010984
250	0,702486	Ortalama	0,896341	0,891600	0,828200	0,695000
		Varyans	0,004649	0,004951	0,002146	0,004007
500	0,611549	Ortalama	-	0,853600	0,690800	0,601600
		Varyans	-	0,001399	0,001261	0,001928

Tüm örneklem büyüklüklerinde plug-in yöntemi optimal h değerine daha yakın sonuç vermiş, örneklem büyüklüğü 500 durumu hariç, tüm örneklem büyüklüklerinde plug-in yönteminden elde edilen bant genişliği seçicilerinin varyansı diğerlerine göre daha küçük elde edilmiştir. Örneklem büyüklüğü 500 için düzleştirilmiş bootstrap yönteminden elde edilen bant genişliği seçicilerinin varyansının diğerlerine göre daha küçük olduğu görülmüştür.

Standart normal dağılım

Standart normal dağılımdan örneklem büyüklüğü 50, 100, 250 ve 500 olan örneklemler alınmıştır. Çekirdek fonksiyonu standart normal dağılım alınarak, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap ve plug-in yöntemleri ile bant genişliği değerleri bulunmuş, çapraz geçerlilik yöntemi ile bant genişliği değeri elde edilememiştir. Sonuçlar Tablo 2'de sunulmuştur.

Tablo 2. Standart normal dağılımdan alınan örneklemler için üç yöntem ile elde edilen bant genişliği değerlerinin ortalama ve varyansları

Örneklem Büyüklüğü	Optimal Bant Genişliği	Bant Genişliklerinin Dağılımına Ait	Yöntemler		
			YÇG	B	PLUG-IN
50	0,380503	Ortalama	0,914032	0,594400	0,465800
		Varyans	0,009411	0,002192	0,005647
100	0,331247	Ortalama	0,924138	0,582222	0,412300
		Varyans	0,006604	0,003526	0,002408
250	0,275781	Ortalama	0,935068	0,574464	0,344300
		Varyans	0,004373	0,001287	0,001009
500	0,240081	Ortalama	0,942000	0,582195	0,305000
		Varyans	0,002039	0,001143	0,000323

Tüm örneklem büyüklüklerinde plug-in yönteminin optimal h değerine daha yakın sonuç verdiği ve örneklem büyüklüğü 50 durumu hariç diğer durumlarda plug-in yönteminden elde edilen bant genişliği seçicilerinin varyansının diğerlerine göre daha küçük olduğu görülmüştür. n=50 durumunda düzleştirilmiş bootstrap yönteminden elde edilen bant genişliği seçicilerinin varyansı diğerlerine göre daha küçüktür. Plug-in

yönteminden sonra, optimal bant genişliğine en yakın değer, düzleştirilmiş bootstrap yönteminden elde edilmiştir. Tüm örneklem büyüklüklerinde, yanlı çapraz geçerlilik yönteminden elde edilen bant genişlikleri, hem optimal değerden uzak sonuçlar vermiş, hem de varyansı büyük bulunmuştur.

Gamma dağılımı

$\alpha = 3, \beta = 1$ parametreli gamma dağılımından (yaklaşık olarak simetrik) örneklem büyüklüğü 50, 100, 250 ve 500 olan örneklemler alınmıştır. Çekirdek fonksiyonu standart normal dağılım alınarak, çapraz geçerlilik, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap ve plug-in yöntemleri ile bant genişliği değerleri bulunmuş, örneklem büyüklüğü 250 ve 500 için çapraz geçerlilik yönteminden bant genişliği değeri elde edilememiştir. Sonuçlar Tablo 3'te sunulmuştur.

Tablo 3. Gamma dağılımından (simetrik) alınan örneklemler için dört yöntem ile elde edilen bant genişliği değerlerinin ortalama ve varyansları

Örneklem Büyüklüğü	Optimal Bant Genişliği	Bant Genişliklerinin Dağılımına Ait	Yöntemler			
			ÇG	YÇG	B	PLUG-IN
50	0,496856	Ortalama	0,921875	0,886700	1,176750	0,690000
		Varyans	0,014415	0,018471	0,047088	0,009604
100	0,432538	Ortalama	0,804091	0,842400	0,864100	0,587300
		Varyans	0,002892	0,006170	0,020428	0,002755
250	0,360111	Ortalama	-	0,819200	0,607000	0,469300
		Varyans	-	0,000668	0,001906	0,001237
500	0,313495	Ortalama	-	0,794800	0,499200	0,319400
		Varyans	-	0,000087	0,000354	0,002087

Tüm örneklem büyüklüklerinde plug-in yönteminin optimale yakın sonuç verdiği, n= 50 ve 100 için plug-in yönteminden, n= 250 ve 500 için YÇG yönteminden elde edilen bant genişliği seçicilerinin varyansının daha küçük olduğu görülmüştür. ÇG ile YÇG yöntemi karşılaştırıldığında, ÇG yönteminden elde edilen bant genişliği seçicilerinin varyansı daha küçüktür. n=250 ve 500 için YÇG yönteminden elde edilen bant genişliği değeri, düzleştirilmiş bootstrap yönteminden elde edilenden daha büyük olup, optimalden uzak sonuçlar vermiştir.Ancak YÇG yönteminden elde edilen bant genişliği seçicilerinin varyansı daha küçüktür.

Simetrik Olmayan Dağılımlar İçin Uygulama Sonuçları

Gamma dağılımı

$\alpha = 2, \beta = 1$ parametreleri ile gamma dağılımından örneklem büyüklüğü 50, 100, 250 ve 500 olan örneklemler alınmıştır. Çekirdek fonksiyonu standart normal dağılım alınarak, çapraz geçerlilik, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap ve plug-in yöntemleri ile bant genişliği değerleri bulunmuş, örneklem büyüklüğü 100, 250 ve 500 için çapraz geçerlilik yönteminden bant genişliği değeri elde edilememiştir. Sonuçlar Tablo 4'te sunulmuştur.

Tablo 4. Gamma dağılımından (simetrik olmayan) alınan örneklemler için dört yöntem ile elde edilen bant genişliği değerlerinin ortalama ve varyansları

Örneklem Büyüklüğü	Optimal Bant Genişliği	Bant Genişliklerinin Dağılımına Ait	Yöntemler			
			ÇG	YÇG	B	PLUG-IN
50	0,339585	Ortalama	0,780000	0,857900	0,932346	0,522500
		Varyans	0,005155	0,011487	0,035236	0,005722
100	0,295626	Ortalama	-	0,822000	0,679700	0,431600
		Varyans	-	0,001875	0,011649	0,001583
250	0,246124	Ortalama	-	0,807400	0,533300	0,334700
		Varyans	-	0,000217	0,000641	0,000603
500	0,214264	Ortalama	-	0,804200	0,504800	0,281800
		Varyans	-	0,000106	0,000140	0,000293

Tüm örneklem büyüklüklerinde plug-in yönteminin optimal değere yakın sonuç verdiği, n= 50 için ÇG ve plug-in yönteminden, n= 100 için plug-in yönteminden, n= 250 ve 500 için YÇG yönteminden elde edilen bant genişliği seçicilerinin varyansının daha küçük olduğu sonucuna ulaşılmıştır. Örneklem büyüklüğünün 50'den büyük olduğu durumlarda YÇG yönteminden elde edilen bant genişliği değerlerinin optimal h değerinden çok uzak sonuçlar verdiği görülmüştür.

Üstel dağılım

$\lambda = 1$ parametreli üstel dağılımdan örneklem büyüklüğü 50, 100, 250 ve 500 olan örneklemler alınmıştır. Çekirdek fonksiyonu standart normal dağılım alınarak, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap ve plug-in yöntemleri ile bant genişliği değerleri bulunmuş, çapraz geçerlilik yönteminden bant genişliği elde edilememiştir. Sonuçlar Tablo 5'te sunulmuştur.

Tablo 5. Üstel dağılımdan alınan örneklemler için üç yöntem ile elde edilen bant genişliği değerlerinin ortalama ve varyansları

Örneklem Büyüklüğü	Optimal Bant Genişliği	Bant Genişliklerinin Dağılımına Ait	Yöntemler		
			YÇG	B	PLUG-IN
50	0,407868	Ortalama	1,046400	0,726364	0,290300
		Varyans	0,084117	0,029965	0,002716
100	0,355070	Ortalama	1,041892	0,686000	0,236500
		Varyans	0,054244	0,017760	0,000940
250	0,295615	Ortalama	1,000952	0,616000	0,176400
		Varyans	0,022759	0,003827	0,000187
500	0,257348	Ortalama	1,043636	0,624000	0,139600
		Varyans	0,015418	0,008930	0,000065

Tüm örneklem büyüklüklerinde plug-in yönteminin optimal değere yakın sonuç verdiği ve bu yöntemden elde edilen bant genişliği seçicilerinin varyansının diğerlerinden daha küçük olduğu, YÇG yönteminden elde edilen bant genişliklerinin optimalden çok uzak değerler verdiği, düzleştirilmiş bootstrap yönteminden elde edilen bant genişliklerinin YÇG yönteminden elde edilenlerden daha küçük olduğu görülmüştür.

Pareto dağılımı

$\alpha = 3, k = 2$ parametreli pareto dağılımından örneklem büyüklüğü 50, 100, 250 ve 500 olan örneklemler alınmıştır. Çekirdek fonksiyonu standart normal dağılım alınarak, yanlı çapraz geçerlilik, düzleştirilmiş bootstrap ve plug-in yöntemleri ile bant genişliği değerleri bulunmuş, örneklem büyüklüğü n=50, 100 ile 250 için düzleştirilmiş bootstrap yöntemi ve tüm örneklem büyüklükleri için çapraz geçerlilik yönteminden bant genişliği değeri elde edilememiştir. Sonuçlar Tablo 6'da sunulmuştur.

Tablo 6. Pareto dağılımından alınan örneklemler için üç yöntem ile elde edilen bant genişliği değerlerinin ortalama ve varyansları

Örneklem Büyüklüğü	Optimal Bant Genişliği	Bant Genişliklerinin Dağılımına Ait	Yöntemler		
			YÇG	B	PLUG-IN
50	0,222999	Ortalama	2,307222	-	1,052800
		Varyans	1,494344	-	0,463319
100	0,194132	Ortalama	2,504677	-	0,909100
		Varyans	0,570701	-	0,371261
250	0,161625	Ortalama	2,360000	-	0,666200
		Varyans	0,378373	-	0,039240
500	0,140703	Ortalama	2,264706	1,090909	0,487600
		Varyans	0,160535	0,062909	0,014092

YÇG yönteminden elde edilen bant genişliklerinin plug-in yöntemi ile elde edilenden daha büyük olduğu, n=500 için düzleştirilmiş bootstrap yönteminden elde edilen bant genişliğinin YÇG yöntemi ile elde edilenden daha küçük olduğu saptanmıştır. Tüm örneklem büyüklüklerinde plug-in yönteminin YÇG yöntemine göre optimale yakın sonuç verdiği, fakat optimalden yine de çok uzak olduğu görülmüştür. Plug-in yönteminden elde edilen bant genişliği seçicilerinin varyansının, YÇG yönteminden elde edilen bant genişliği seçicilerinin varyansına göre daha küçük olduğu da elde edilen sonuçlar arasındadır.

4. SONUÇ

Bant genişliğinin seçimi için verilen dört yöntem karşılaştırıldığında *tüm örneklem büyüklüklerinde, simetrik ve simetrik olmayan dağılımlardan alınan tüm örneklemlerde* plug-in yönteminin, optimal bant genişliğine en yakın sonuçları verdiği görülmüştür.

KAYNAKLAR

BOWMAN, A. W.,1984, *An Alternative Method of Cross-validation for the Smoothing of Density Estimates*, Biometrika, 71 ,2, 353-360.

CAO, R., CUEVAS A., MANTEIGA W.G., 1994, *A Comparative Study Of Several Smoothing Methods In Density Estimation*, Computational Statistics & Data Analysis, 17, 153-176.

CULA, S.G., 1998, *Çok Değişkenli Olasılık Yoğunluk Fonksiyonunun Çekirdek Fonksiyonlarıyla Kestirimi*, Doktora Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 125s.

FARAWAY, J.J., JHUN M., 1990, *Bootstrap Choice of Bandwidth for Density Estimation*, Journal of the American Statistical Association, 85, 1119-1122.

GÖKMEN, D, 2002, *Bant Genişliği Seçiminde Kullanılan Yöntemlerin Simetrik ve Simetrik Olmayan Dağılımlarda Karşılaştırılması*, Bilim Uzmanlığı Tezi, Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, Ankara, 100s.

HÄRDLE, W., 1991, *Smoothing Techniques with Implementation in S*, Springer-Verlag New York Inc.

LOADER, C.R., 1999, *Bandwidth Selection: Classical or Plug-in,* Ann.Statist., 27, 415-438.

PARK, B.U.and MARRON J.S., 1990, *Comparison of Data-Driven Bandwidth Selectors*, J.Am. Statist. Assoc., 85, 66-72.

ROSENBLATT, M., 1956. *Remarks on some nonparametric estimates of a density function.* Annals Math.Statist. , 27, 832-837.

SHEATHER, S.J., JONES M.C., 1991, *A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation*, J.R.Statist. Soc., B 53, 683-690.

SILVERMAN, B.W., 1986, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

TAYLOR, C.C., 1989, *Bootstrap Choice of the Smoothing Parameter in Kernel Density Estimation, Biometrika*, 76, 705-712.

TOKTAMIŞ, Ö., CULA, S., KURT, S., 1999, *Comparison of Bandwidth Selection Methods for Kernel Estimation of Probability Density Function*, İstatistik, Journal of the Turkish Statistical Association,Vol. 2, Number 2, 107-121.

SCOTT, D. W., TERRELL, G. R, 1987, *Biased and unbiased cross-validation in density estimation,* Journal of the American Statistical Association, 82, 400, 1131-1146.

WAND, M.P., JONES, M.C., 1995, *Kernel Smoothing,* Chapman and Hall, New York.

Comparison Of Bandwidth Selection Methods For Symmetric And Asymmetric Distributions

ABSTRACT

In this study, the information about "least squares cross validation", "biased cross validation", "smoothed bootstrap" and "plug-in" methods which are used in the bandwidth selection for kernel estimation of probability density function was given and doing an application, in samples with different sample sizes, obtained from symmetric and asymmetric distributions, the bandwidth values were calculated using these four methods and compared with optimal bandwidth value.

Key Words: *Kernel estimation, bandwidth, least squares cross validation, biased cross validation, smoothed bootstrap, plug-in.*