Ortak Bilgi Miktarının Modelden-Bağımsız ve Hızlı Hesaplanması için Yeni Yöntemler

Novel Techniques for Model-Free and Fast Computation of Mutual Information

Serhat ÇAĞDAŞ ve Bilge KARAÇALI
Elektrik ve Elektronik Mühendisliği Bölümü
İzmir Yüksek Teknoloji Enstitüsü
İzmir, Türkiye
serhatcagdas@iyte.edu.tr, bilgekaracali@iyte.edu.tr

***Özetçe*—Bu çalışmada, iki rastlantısal değişken arasındaki ortak bilgi miktarının veri üzerinden hesaplanmasına yönelik yeni yaklaşımlar önerilmiştir. Bu yaklaşımlar, doğrusal dönüşüm altında diferansiyel entropinin gösterdiği özellikleri kullarak ve koşullu entropiyi modelden-bağımsız bir şekilde küçültmeye çalışarak kestirim yapacak şekilde kurgulanmıştır. Birim vektör parametrizasyonu ve veri oturtmaya dayanan tahmin edici olarak adlandırdığımız yöntemlerin, yaygın olarak kullanılan Kraskov yöntemiyle yapılan karşılaştırmalarda, örnek sayısı arttıkça işlem hızı açısından avantaj sağladığı görülmüştür.**

Anahtar Kelimeler — ortak bilgi miktarı; koşullu entropi, kestirici, model bağımsız, parametrik olmayan;

***Abstract*—In this study, two new approaches are proposed to calculate mutual information between two random variables from data. These approaches are constructed in a way to use the properties of the differential entropy under linear transformations and to try to minimize conditional entropy in a model-free manner. In comparisons with a widely used mutual information estimator, the Kraskov method, the methods that we termed as unit vector parametrization and data fitting based estimators, offered an advantage in terms of computation speed.**

Keywords — mutual information, conditional entropy, estimation, model free, non-parametric

I. GİRİŞ

Ortak bilgi miktarı, rastlantısal değişkenler arasındaki bağımlılığı ölçen araçlardan biridir. Ortak bilginin en temel kullanım alanlarından birisi, haberleşmede kanal giriş ve çıkış arasındaki belirsizliği ölçerek kanal kapasitesi hakkında fikir sahibi olunabilmesidir [1]. Bunun haricinde ses kodalama ve tanıma [2], EEG, MEG ve benzeri biyomedikal sinyallerin işlenmesi [3] gibi sinyal işlemenin bir çok alanında kullanılmakla birlikte, bağımsız bileşen analizi [4] gibi örünü tanıma ve makine öğrenmesi araçlarında da önemli bir yeri vardır [5]. Bu yöntemin tercih edilmesinin sebebi sadece doğrusal kovaryanstan kaynaklı bağımlılığı değil doğrusal olmayan bağımlılığı da ortaya çıkarmasıdır.

Ortak bilgi miktarı, kısaca bileşik olasılık yoğunluk fonksiyonunun marjinal yoğunluk fonksiyonlarının çarpımına olan uzaklığı olarak tanımlanabilir:

$$I(X,Y) = \iint f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \qquad (1)$$

Bu tanım, marjinal entropilerin toplamının bileşik entropiden farkı olarak da ifade edilebilir:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \qquad (2)$$

Birbirine denk bu iki ifadede olasılık yoğunluk fonksiyonlarının bilinmesi veya hesaplanması gerektiğinden dolayı ortak bilgi miktarını doğru bir şekilde bulmak bir zorluk olarak karşımıza çıkmaktadır. Bu konuda kernel, k-en yakın komşu (kNN) ya da histogram tabanlı yoğunluk fonksiyonu tahmini gibi yaklaşımların yanında maksimum olabilirlik kestirimi (ML) ya da Bayes kestirici gibi parametrik kestirim yöntemlerinin hız, işlem yükü ve doğruluk açısından karşılaştırıldığı bir inceleme sunulmuştur [6].

Özellikle makine öğrenmesi ya da veri madenciliği gibi büyük verilerin işlendiği durumlarda, kestirim yönteminin işlem yükü ve hesap zamanı önem kazanmaktadır. Bu çalışmada entropi ve ortak bilginin özellikleri kullanılarak modelden bağımsız ve hızlı bir biçimde ortak bilgi miktarının yeni kestirim yolları incelenmiş ve bu yöntemlerin hız ve doğruluk açısından karşılaştırılması yapılmıştır.

Bir sonraki bölümde yaygın kullanılan k-en yakın komşu algoritması tabanlı bir metodun yanında doğrusal dönüşüm altında entropinin gösterdiği özellikleri kullanan bir hesaplama yöntemi ve koşullu entropiyi küçültmeye çalışarak kestirim yapan bir yöntem tanıtılmıştır. Sonuçlar bölümünde Gauss, üstel ve tekdüze dağılıma ait değişkenlerin doğrusal dönüşümleri sonucu elde edilen simülasyon verileri altında doğruluk başarımları değerlendirilmiş, artan örnek sayısına bağlı olarak işlem hızları karşılaştırılmıştır. Tartışma kısmı ise genel bir son bakış ve gelecek araştırma doğrultularıyla bitirilmiştir.

II. MATERYAL VE METOD

A. Kraskov kNN Yöntemi

Kraskov'un metodu [7], k-en yakın komşu yöntemi ile Shannon entropisinin hesaplamasını yapan Kozachenko-Leonenko metodu üzerine kurulmuştur [8]. Bu yöntemin temel fikri k indisli yakın komşunun ortalama uzaklıkları üzerinde bir entropi hesabı yapmaktır:

978-1-5386-1501-0/18/$31.00 ©2018 IEEE

$$\hat{H}(X) = -\varphi(k) + \varphi(N) + \log c_d + \frac{d}{N}\sum_{i=1}^{N} \log \epsilon(i) \qquad (3)$$

Yukarıdaki ifadede, N toplam örnek sayısı, $\varphi(x)$ digama fonksiyonu, d rasgele değişken boyutu, c_d d boyutlu birim kürenin hacmi (2 boyut için dairenin alanı) ve $\epsilon(i)$, x_i örneğine ait k indisli en yakın komşu mesafesinin iki katı olarak tanımlanmaktadır.

Kraskov ve diğerleri, bu entropi kestirim formülü üzerinden birbirine benzer iki ortak bilgi hesaplama yöntemi belirlemişlerdir [7]. Biz bu çalışmada yöntem karşılaştırmaları için kendilerinin de daha sonradan Ortak Bilgiye Dayalı En Az Bağımlı Bileşen Analizi (MILCA) yönteminde kullandıkları aşağıdaki ifadeyi kullandık [4]:

$$\hat{I}(X,Y) = \varphi(k) - \frac{1}{k} - < \varphi(n_x) + \varphi(n_y) > + \varphi(N) \qquad (4)$$

Yukarıdaki formülde $\epsilon_x(i)$ ve $\epsilon_y(i)$, k indisli komşuyu barındıran en küçük dikdörtgene ait kenar uzunluları, $n_x(i)$ ve $n_y(i)$ de,

$$\|x_i - x_j\| \leq \epsilon_x(i)/2$$

ve

$$\|y_i - y_j\| \leq \epsilon_y(i)/2$$

şartlarını sağlayan eleman sayılarıdır. Ayrıca $<.>$ ortalama operatorü ve $i \in [1,2,\ldots,N]$ olarak tanımlanmıştır. Bu çalışma boyunca Kraskov ortak bilgi kestirimi için k parametresi 1 olarak alınmıştır.

B. Birim Vektör Parametrizasyonu

Denklem (2)'ye geri dönülecek olursa, ortak bilgi miktarının kestirimi için marjinal ve ortak entropilerin hesaplanması yeterlidir. Bu aşamada, sürekli rastlantısal değişkenlere ait marjinal entropinin kestirimi için literatürde birçok yöntem bulunmaktadır [9]. Bu yöntemler içerisinde Kozachenko-Leonenko entropi kestirimi [8] gibi aralıklar üzerinden kestirim yapan Vasicek entropi kestirici oldukça hızlı ve doğru sonuçlar vermektedir [10]. Herhangi bir X sürekli rasgele değişkeni için $x_1, x_2, \ldots, x_n$ örnekler ve $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ bu örnekler üzerinden oluşturulmuş sıralı örnekler olarak kabul edilirse n örnek sayılı bu değişken için Vasicek entropi kestirimi aşağıdaki denklemle gerçeklenir:

$$V_{m,n} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n}{2m}\left(x_{(i+m)} - x_{(i-m)}\right)\right) \qquad (5)$$

Örnek sayısı n ve m değeri sonsuza giderken aynı zamanda $m/n \to 0$ koşulu sağlandığında $V_{m,n}$ değeri gerçek entropi değerine yakınsamaktadır [10]. Kestiricinin tutarlılığını sağlamak için de formüldeki sapmayı düzelten aşağıdaki kestirim formulü önerilmiştir:

$$\begin{aligned} H_V(X) = V_{m,n} &- \log(n) + \log(2m) \\ &- \left(1 - \frac{2m}{n}\right)\varphi(2m) + \varphi(n+1) \\ &- \frac{2}{n}\sum_{i=1}^{m} \varphi(i+m-1) \end{aligned} \qquad (6)$$

Bu yöntemin bileşik entropi hesabı için kullanılması konusunda Voronoi hücreleri ve Delaunay bölgeleri ile bölütleme temelli bir yöntem önerilmiştir [11]. Fakat bu yöntem oldukça karmaşık ve zaman alıcı olduğu için uygulamada çok tercih edilen bir çözüm olamamıştır.

Birim vektör parametrizasyonu yönteminde, entropinin doğrusal transformasyon altında gösterdiği özellik kullanılıp bileşik entropinin minimize edilmesi ile ortak bilgi değerine hızlı bir yakınsama amaçlanmaktadır. X ve Y sürekli rasgele değişkenlerine uygulanan

$$\begin{bmatrix} U \\ V \end{bmatrix} = A \begin{bmatrix} X \\ Y \end{bmatrix}$$

doğrusal dönüşümü sonucunda $H(U,V)$ ve $H(X,Y)$ bileşik entropi değerleri arasında 2×2'lik A matrisinin determinantının logaritması kadar bir değişim gözlenmektedir:

$$H(U,V) = H(X,Y) + \log|\det(A)| \qquad (7)$$

Bileşik entropi ile ilgi bir diğer özellik ise bu değerin marjinal entropilerin toplamından her zaman daha küçük olduğudur:

$$H(U,V) \leq H(U) + H(V) \qquad (8)$$

Yukarıdaki ifadelerden, $H(X,Y)$ değeri ile doğrusal dönüşüm altındaki marjinal entropi değerleri arasında aşağıdaki bağıntı kolayca kurulabilir:

$$H(X,Y) \leq H(U) + H(V) - \log|\det(A)| \qquad (9)$$

Dolayısıyla denklemin sağ tarafını doğrusal dönüşüm altında

$$\hat{H}(X,Y) = \min_A\left(H_V(U) + H_V(V) - \log|\det(A)|\right) \qquad (10)$$

ile ne kadar küçültebilirsek bileşik entropinin gerçek değerine o kadar yaklaşmış oluruz ve bunu Vasicek entropi hesabı ile ortak bilgi kestirminde kullanabiliriz:

$$\hat{I}(X,Y) = \mathrm{H_V(X)} + H_V(Y) - \hat{H}(X,Y) \qquad (11)$$

Yöntem içerisinde minimum entropi değerini sağlayan A matrisinin aranması da önemli bir hususdur. Bu çalışmada A matrisi bir rotasyon matrisine benzer biçimde, $\theta_1, \theta_2 \in [-\pi, \pi)$ olacak şekilde kosinus ve sinus çiftleri üzerinden

$$A(\theta_1,\theta_2) = \begin{bmatrix} \cos(\theta_1) & \sin(\theta_1) \\ \cos(\theta_2) & \sin(\theta_2) \end{bmatrix} \qquad (12)$$

ile oluşturulmuş ve minimum bileşik entropiyi hesaplayan matris bu matris kümesi içerisinde, θ_1 ve θ_2 üzerinden aranmıştır.

C. Veri Oturtmaya Dayanan Yöntem

Birim vektör parametrizasonu yöntemi, aralarındaki bağıntı doğrusal olan değişkenler için ortak bilgi kestiriminde iyi sonuçlar verse de ilişkileri doğrusal olmayan değişkenler için ortak bilgi kestirim başarımı düşebilir. Bu yüzden bu çalışmada, değişkenler arasında gözlenen şartlılığı ortadan kaldırmaya çalışarak koşullu entropi üzerinden bir ortak bilgi hesabı yapan, veri oturtmaya dayalı bir diğer ortak bilgi kestirim yöntemi daha değerlendirilmiştir. Bu yöntem için Denklem (2)'ye uygun bir biçimde ortak bilgi miktarının, koşullu entropi üzerinden bir başka ifadesi kullanılmıştır:

$$I(X,Y) = H(Y) - H(Y|X) \qquad (13)$$

Burada X ve Y'nin birbirinden bağımsız olduğu durumda $H(Y|X) = H(Y)$ eşitliği ortaya çıkmaktadır. Bu doğrultuda Y değişkenini X değişkeninden olabildiğince bağımsız hale getirebilmek için, değişkeni, X değişkenine göre koşullu beklenen değerinden ve koşullu varyasından kurtarmaya çalışılabilir. Bu işlemler, iki değişkeni tamamen bağımsız hale

getirmese bile koşullu entropiyi yeterince küçültmemizi sağlayabilir:

$$Y' = \frac{Y - m(x)}{\sigma(x)} \tag{14}$$

Yukarıdaki ifadede $m(x) = E[Y|X = x]$ ve $\sigma(x) = \sqrt{Var(Y|X = x)}$ olarak tanımlanmıştır. Bu durumda ortak bilgi miktarının kestiriminde $H(Y'|X) \approx H(Y')$ kabulünü yapabiliriz. Dolayısıyla koşullu entropinin aşağıdaki iki özelliğini kullanarak ortak bilgi miktarı için alternatif bir kestirim yöntemi daha ortaya koyulabilir:

1. X ve Y iki rastlantısal değişken iken, X'in herhangi bir değeri için Y'nin ötelenmesi koşullu entropi değerini değiştirmez:

$$H(Y + c|X = x) = H(Y|X = x) \tag{15}$$

2. X'in herhangi bir değeri için Y'nin sabit bir sayıyla çarpımı, koşullu entropi değerinde çarpımın logaritması kadar bir değişime sebep olur:

$$H(aY|X = x) = H(Y|X = x) + \log|a| \tag{16}$$

Doğrulukları kolaylıkla gösterilebilecek yukarıdaki özellikler göz önünde bulundurulduğuda Y' değişkenine ait koşullu entropi ile Y değişkeninin koşullu entropisi arasında koşullu standart sapma değerinin logaritmasının beklenen değeri kadar bir fark ortaya çıkacaktır:

$$H(Y|X) = H(Y'|X) + E[\log \sigma(x)] \tag{17}$$

Daha önceki $H(Y'|X) \approx H(Y')$ kabulümüzü de göz önünde bulunudurursak Vasicek entropi hesabı ile ortak bilgi değeri için aşağıdaki kestirimi yapmış oluruz:

$$\hat{I}(X,Y) = H_V(Y) - H_V(Y') - E[\log \sigma(x)] \tag{18}$$

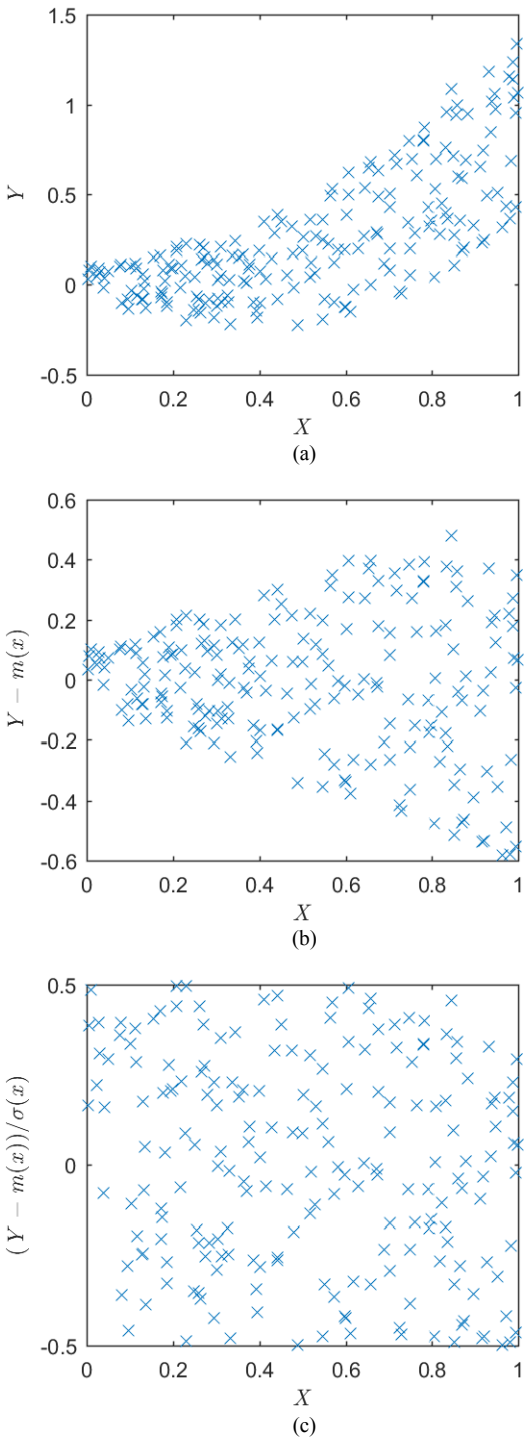
Şekil 1'de bir örnek data için değişkenlerin ortalama ve varyans olarak şartlılıktan arındırma işlemi gösterilmiştir. Bu örnek için Y üzerindeki değişiklikler sonucu elde edilen Y' ve X rastlantısal değişkenlerinin birbirinden bağımsız hale geldiği görülebilir. İşlem sonunda artık Vasicek entropi hesabı ile marjinal entropiler hesaplanıp ortak bilgi miktarına ulaşılabilir.

Denklem (18)'i incelendiğinde kestirim yönteminin artık bir koşullu varyans tahminine indirgendiği görülebilir. Çalışma içerisinde varyans tahmini için polinomsal bağlanım ve kayan ortalama düzleştirici yöntemleri kullanılmıştır [5],[12]. Polinomsal bağlanım yönteminde derece seçimi için veriler sıralanıp birer atlanarak iki kümeye ayrılmış, ilk kısmı polinom denkleminin kestiriminde ikinci kısım ise kestirimin hatasının hesaplanmasında kullanılmıştır. Kayan ortalama düzleşitirici ile ortalama ve varyans kestiriminde de tekdüze bir kernel kullanılmış, kernel genişliğinde ise yine minimum karesel hatayı veren kernel eleman sayısı seçilmiştir.

III. SONUÇLAR

Yöntemlerin karşılaştırılması için Gauss, üstel ve tekdüze olmak üzere üç farklı dağılımlı değişken ile, 10 derece aralıklarla birbirine ters açılı rotasyon vektörleri kullanılarak doğrusal dönüşüme uğratılmış ve aralarında ilinti bulunan başka iki değişken oluşturulmuştur. Bu yol izlenerek Gauss-Gauss, Gauss-tekdüze, ve Gauss-üstel ikililerinden her farklı örnek sayısı ve dönüşüm açısı için 100 tekrarlı test verisi üretilmiş, bu veriler üzerinden yöntemlerin doğruluk ve hız karşılaştırmaları gerçekleştirilmiştir.

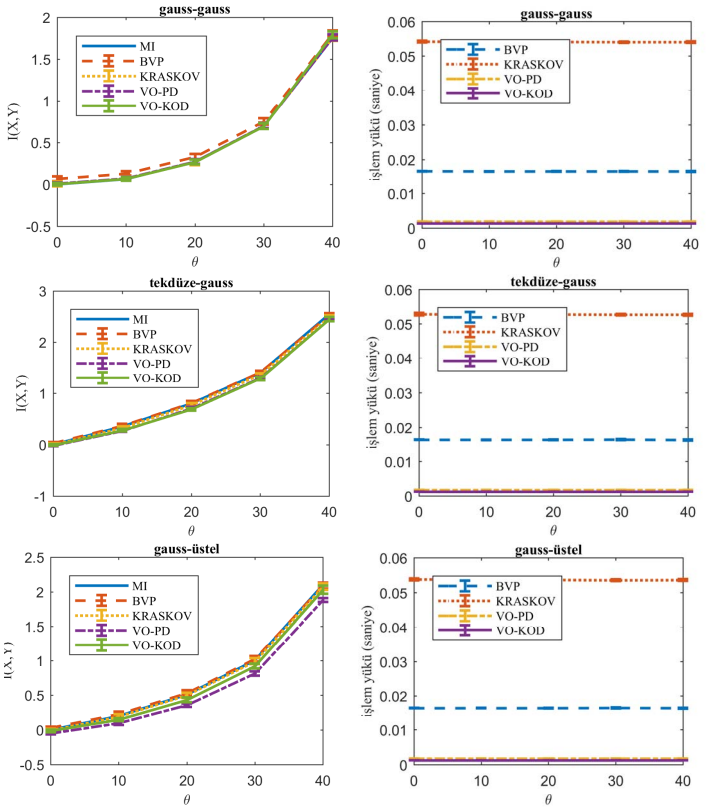
Şekil 2'de 1000 örnekli veriler için Kraskov ve birim vektör parametrizasyon (BVP) yöntemine ek olarak veri oturtmaya dayalı yöntemin polinom düzleştirici (VO-PD) ve kayan ortalama düzleştirici (VO-KOD) yöntemleri ile koşullu varyans kestirimi yoluyla elde edilen sonuçların ve işlem sürelerinin



Şekil 1. Rasgele Değişkenlerin Koşullu Beklenen Değer ve Varyanstan Kurtarılması

zaman grafikleri görünmektedir. Birim vektör parametrizasyon yönteminin oluşturduğu ortak bilgi miktarı kestirimleri, yaklaşımın kurgusundan beklendiği gibi genel olarak gerçek değerin üzerinde sonuçlar vermiştir.Polinom oturtma ile gerçeklenen veri oturtmaya dayalı kestirim yöntemi ise Gauss-Gauss dağılımlı veriler için başarılı sonuçlar verse de genel

olarak kayan ortalama düzleştirici yöntemi ile gerçeklenen yönteme göre hata oranı daha yüksek olmuştur.

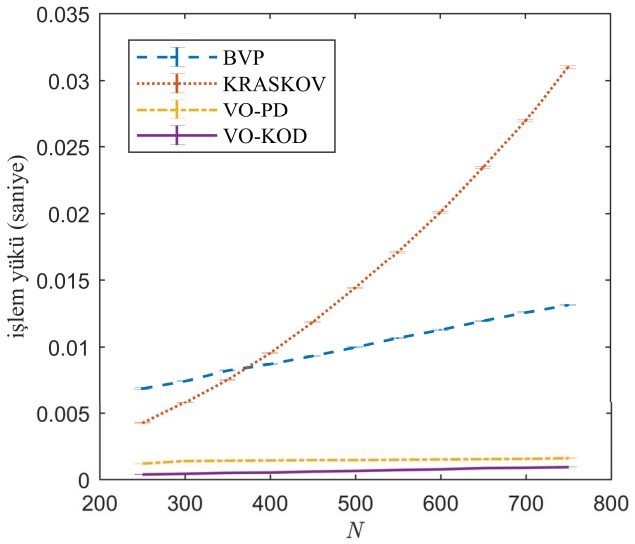


Şekil 2. Örnek sayısı $N = 1000$ için farklı dağılım ve projeksiyon açısına ait ortak bilgi kestirimi değeri ve zamanı grafikleri (MI çizgisi gerçek değeri temsil etmektedir.)

Hız olarak değerlendirildiğinde ise yine kayan ortalama düzleştirici ile gerçeklenen veri oturtma yöntemi polinom uydurma yöntemine göre daha hızlı sonuç verirken 1000 örnekli veriler için Kraskov yöntemine göre yaklaşık 13 kat daha hızlı biçimde işlemini tamamlamıştır. Şekil 3'te değişen örnek sayısına göre çizilen işlem süresi grafiğinde görüldüğü gibi yine kayan ortalama düzleştirici yöntemi ile koşullu varyans kestirimi yapılan veri oturtmaya dayalı ortak bilgi tahmin yöntemi, diğer yöntemlere göre daha hızlı sonuç vermiştir. Kraskov yöntemi, örnek sayısı azken birim vektör parametrizasyonu yöntemi ile karşılaştırıldığında daha çabuk hesaplama yapıyorken, yöntemin işlem yükü, örnek sayısıyla üstel orantılı olarak arttığı örnek sayısı arttıkça kestirim süresi de görece diğer yöntemlere göre uzun sürmüştür.

IV. TARTIŞMA

Bu çalışmada, literatürdeki çalışmalarda ortak bilgi miktarının eldeki veri üzerinden hesaplanması amacıyla sıklıkla kullanılan Kraskov yöntemine alternatif olarak iki yeni yöntem önerilmiş ve bu yöntemler, hem doğruluk hem de işlem yükü açılarından karşılaştırılmıştır. Bu karşılaştırmalar sonucunda birim vektör parametrizasyonu yöntemi ve veri oturtmaya dayalı kestirim yöntemi, örnek sayısı arttıkça gerçek değere asimptotik yakınsamayı garanti etmemekle birlikte ortak bilgi hesabında hızlı ve kullanılabilir yöntemler olarak karşımıza çıkmaktadır. Özellikle yüksek boyutta veri için bu hesaplamalar yapılırken bu yöntemler tercih edilebilir. Veri oturtmaya dayalı kestirim yönteminde kullanılacak farklı düzleştirici ya da indirgeme metodları, yöntemin doğruluğunu ve işlem hızını olumlu yönde etkileyebilir. Veri oturtma modelinin hangi değişken üzerinden uygulanacağının seçimi için uygun kriterler türerilebilirse bu da doğruluğu artırmada faydalı olabilir.



Şekil 3. Farklı Örnek Sayıları İçin Ortalama İşlem Zamanları

KAYNAKLAR

[1] Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[2] Deller Jr, J. R., Proakis, J. G., & Hansen, J. H. (1993). *Discrete time processing of speech signals*. Prentice Hall PTR.

[3] Sakkalis, V. (2011). Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in biology and medicine*, *41*(12), 1110-1117.

[4] Stögbauer, H., Kraskov, A., Astakhov, S. A., & Grassberger, P. (2004). Least-dependent-component analysis based on mutual information. *Physical Review E*, *70*(6), 066123.

[5] Christopher, M. B. (2016). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

[6] Walters-Williams, J., & Li, Y. (2009, July). Estimation of mutual information: A survey. In *International Conference on Rough Sets and Knowledge Technology* (pp. 389-396). Springer, Berlin, Heidelberg.

[7] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, *69*(6), 066138.

[8] Kozachenko, L. F., & Leonenko, N. N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, *23*(2), 9-16.

[9] Wieczorkowski, R., & Grzegorzewski, P. (1999). Entropy estimators□ improvements and comparisons. *Communications in Statistics-Simulation and Computation*, *28*(2), 541-567.

[10] Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54-59.

[11] Miller, E. G. (2003, April). A new class of entropy estimators for multi-dimensional densities. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on* (Vol. 3, pp. III-297). IEEE.

[12] Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. John Wiley & Sons, Inc..