A New Method for Local Dependence Map and Its Applications

Lokal Bağlımlılık Haritaları İçin Yeni Bir Yöntem ve Uygulamaları

ABSTRACT Objective: This work introduces a new method to construct local dependence map based on the estimate for the linear local dependence function $H(x,y)$, which is generalization of Pearson correlation coefficient. The new local dependence map demonstrates a practical tool for local dependence structure between two random variables. The analysis of theoretical concepts is verified by an application based on real datasets in endocrinology. Material and Methods: The method, local dependence map, requires the estimation new local dependence function which is based on regression concepts. After this local dependence function must be converted with local permutation tests in local dependence map which make the local dependence function more interpretable by identifying the regions of positive, negative and zero local dependence. Results: Based on the proposed method and we give two examples based on the real data C-peptide, insulin and TSH, FT3, FT4 from endocrinology in order to show the advantageous of the current dependence maps. They show interesting local dependence features on the other hand overall correlation coefficient is not much informative. Conclusion: Scalar dependence measures such as correlation coefficient are often used as a measure of dependence for data in medical and biological science. However, they cannot reflect the complex dependence structure of two variables. Hence we are concerned exclusively with the statistical aspects of the dependence structure in dependence maps that will be constructed for the dataset. In this work a new method to construct local dependence map based on the regression concept for the linear local dependence function $H(x,y)$, which is generalization of Pearson correlation coefficient, is established. The proposed new local dependence map is devoted to two examples based on the real data C-peptide, insulin and TSH, FT3, FT4 from endocrinology in order to illustrate the usefulness of the current dependence maps. They show interesting local dependence features on the other hand overall correlation coefficient is not much informative.

Key Words: Correlation, correlation study, statistical data interpretation, local dependence map


Anahtar Kelimeler: Korelasyon, korelasyon çalışması, istatistiksel veri yorumu, yerel bağımlılık haritası

In recent years local dependence functions and their applications have arisen the interest of many statisticians. It is because scalar dependence measures, such as Pearson correlation coefficient and many others cannot always be adequate to explain the dependence structure between two random variables. In fact, there are generally some restrictions about not to use these scalar measures in many situations. Thus scalar dependence measures are extended to local dependence function.

Bjerve and Doksum, Doksum et al. and Blyth, determined a correlation curve

\[ \rho(x) = \frac{\sigma_1 \beta(x)}{\left(\sigma_1^2 \beta^2(x) + \sigma^2(x)\right)^{1/2}} \]  

(1)

as a generalization of the Pearson correlation coefficient where \( \beta(x) = \mu'(x) \) is the slope of the nonparametric regression \( \mu(x) = E(Y \mid X = x) \), \( \sigma^2(x) = Var(Y \mid X = x) \) is the nonparametric residual variance and \( \sigma_1^2 = Var(X) \). The correlation curve (1) is constructed from regression concept where Y is a response and X is a predictor variable.\(^1\)\(^3\) One may see that \( \rho(x) \) measures the strength of the association between X and Y locally at \( X = x \).

Holland and Wang introduced a local dependence function.\(^4\) Following their work Jones provided a motivation for the local dependence function by using mixed partial derivative of the log density

\[ \gamma(x, y) = \frac{\partial^2 \log f(x, y)}{\partial x \partial y} \]  

(2)

where \( \gamma(x, y) \) is a function of the conditional distribution of Y given X, or vice versa and is a \( f \) bivariate density function.\(^5\) So, \( \gamma \) may be applied to all bivariate distributions with specified conditionals. But in general, conditionally specified joint distributions are extremely constrained; see.\(^6\)

Bairamov and Kotz introduced a new local dependence function \( H(x, y) \), a generalization of Pearson correlation coefficient,\(^7\)\(^8\)

\[ H(x, y) = \frac{\rho + \varphi_x(y) \varphi_y(x)}{\sqrt{(1 + \rho^2)(1 + \varphi_x^2)(1 + \varphi_y^2)}} \]  

(3)

where, \( \varphi_x(y) = \frac{E(X) - E(X \mid Y = y)}{\sigma_x} \),

\[ \varphi_y(x) = \frac{E(Y) - E(Y \mid X = x)}{\sigma_y} \]  

and

\[ \rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} . \]

The function \( H(x,y) \) in (3) is obtained from the expression of the linear correlation coefficient by replacing \( E(X) \) and \( E(Y) \) by the conditional expectations \( E(X \mid Y = y) \) and \( E(Y \mid X = x) \) respectively. In case that the local dependence function is specified with conditional moments, then it is expected that local dependence function is used for wider class of joint distributions. This local dependence measure has several useful properties. For instance, the new measure is symmetric in \( X \) and \( Y \) and its expected value is approximately equal the Pearson correlation coefficient. Another property is that, \( X^* = 0 \) and \( Y^* = 0 \) is a saddle point of \( H \) and \( H(X^*, Y^*) = \rho \) when \( (X, Y) \) has bivariate normal distribution with correlation coefficient \( \rho \). Further details can be seen in.\(^7\)

In general, it is difficult to interpret the whole dependence structure of the data from a local dependence function. However, dependence graphs have the potential to assess a far richer class of bivariate dependence structures. With this motivation, Jones and Koch introduced a new methodology about the interpretation of the dependence structure of bivariate data called “dependence map”.\(^9\) Dependence maps show us the estimated local dependence structure of the data by identifying regions positive, negative and zero local dependence with the help of local permutation test (Appendix 1).

In this paper, a new method to construct dependence maps of the given data is introduced by using the estimator of the local dependence function of Bairamov and Kotz with permutation test.
algorithm. The next section shows the details the proposed method and Section 3 is devoted to two examples based on the real data C-peptide, insulin and TSH, FT3, FT4 from endocrinology in order to illustrate the usefulness of the current dependence maps. They show interesting local dependence features on the other hand overall correlation coefficient is not much informative.

## CONSTRUCTING THE DEPENDENCE MAP

### ESTIMATION of $H(x,y)$

The local dependence function $H(x,y)$ is estimated via kernel methods. Bairamov and Kotz suggested an estimator for $H(x,y)$ by using Nadarya and Watson’s estimate for the regression functions $E(X|Y=x)$ and $E(Y|X=x)$. The estimators are

$$
\hat{A}_x(y) = \frac{1}{n} \sum_{i=1}^{n} X_i K \left( \frac{y - Y_i}{h_n} \right)
$$

and

$$
\hat{A}_y(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i K \left( \frac{x - X_i}{h_n} \right),
$$

where $(X_i, Y_i)$, $i=1,2,...,n$ are the dataset, $K$ is a kernel function, an integrable function with short tails, and $h_n$ is a bandwidth sequence tending to zero at appropriate rates, i.e. $h_n \to 0$ as $n \to \infty$. Using the equation (4) and taking the Gaussian kernel

$$
K_G(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)
$$
as $K$, we obtain an estimate $\hat{H}(x,y)$ for $H(x,y)$,

$$
\hat{H}(x,y) = \frac{\hat{\rho} + (\bar{x} - \hat{A}_x(y)) (\bar{y} - \hat{A}_y(x))}{S_x S_y}
$$

$$
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i, \bar{y} = \frac{1}{n} \sum_{i=1}^{n} Y_i, \text{ and }
$$

$$
S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2, S_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{y})^2.
$$

Then a usual for $h_n$ choice is the Gaussian bandwidth that gives the normal reference rule

$$
h_n = 1.06 \hat{\sigma} n^{-1/5},
$$

where $\hat{\sigma}$ is the standard deviation of the data. In what follows is the local permutation test for the dataset.

### LOCAL PERMUTATION TEST

Excel Visual Basic software (the code is given in Appendix) is used for this purpose. We first obtain the estimated local dependence measure for the given bivariate data. For every $(x_i, y_i), i=1,\ldots,n$, we compute $\hat{H}(x_i, y_i)$ using the bandwidth in equation (6). The samples satisfying $\hat{H}(x, y) = 0$ can be generated by permuting randomly $y_i$ since the local permutation test of the null hypothesis $H(x, y) = 0$ implies independence. This procedure is repeated $N$ times and in each step $\hat{H}_k(x_i, y_i)$ is computed for each permuted dataset for $k = 1, \ldots, N$. Then statistically significant $\hat{H}_k(x_i, y_i)$ is added to the list by comparing the observed $\hat{H}(x_i, y_i)$. When observed value of the estimated local dependence function is in the highest $(\alpha/2)\%$ of the simulated $\hat{H}_k(x_i, y_i)$, the map value is appointed to be $+1$ and when observed value of the estimated local dependence function is in the lowest $(\alpha/2)\%$ of the simulated $\hat{H}_k(x_i, y_i)$ then the map value is appointed to be $-1$ and otherwise it is $0$. From these values, we design the dependence map. With permutation test, local dependence maps simplify the estimated local dependence structure between two variables by identifying regions of positive (significance $+1$), zero (no significance) and negative (significance $-1$) local dependence.

### APPLICATIONS

Scalar dependence measures such as correlation coefficient are often used as a measure of dependence for data in medical and biological science. Ho-
However, they cannot reflect the complex dependence structure of two variables. Moreover, linear correlation coefficient does not always determine an accurate association between two random variables since the dependence may not exist in the entire plane or it may not linear or the variance $Y$ given $X=x$ may not be a constant. Hence we are now concerned exclusively with the statistical aspects of the dependence structure in dependence maps that will be constructed for the dataset.

Example 1: C-Peptide and Insulin Data

C-peptide is an indicator of insulin stock for insulin-treated diabetics. The level of c-peptide in the blood can indicate how much insulin is being produced by the pancreas. In cases when insulin level cannot be determined due to the existence of insulin antibodies in the blood or in the patients under the treatment of diabetes, it is important to measure the level of c-peptide.  

In this example we construct dependence map using c-peptide and insulin from the data supplied by DEU faculty of medicine endocrine laboratory. 108 patients are examined for this purpose (we take $N=500$ and $\alpha=0.05$). The dependence map is particularly informative here. The overall correlation coefficient between c-peptide and insulin is 0.657, significantly different from zero. Although there is a positive correlation between the variables, two different dependence structures are observed in Figure 1. In the white region where the variables simultaneously take low, moderate and large values, positive dependence between insulin and c-peptide appear. That is, as the level of insulin increases, so does the level of c-peptide. The white region covering relatively larger than the others can be explained by the existence of highly positive correlation between the variables c-peptide and insulin. In other words it is important to interpret in detail the correlation coefficient in this region. When insulin level cannot be determined appropriately it is meaningful to see the level of c-peptide in order to decide the level of insulin. However, an analogous interpretation will not be possible for the other regions in the dependence map. It is shown in color light grey where there is no dependence between insulin and c-peptide. Further, while c-peptide takes moderate and large values insulin takes small and moderate values or while c-peptide values are small and moderate insulin takes moderate values at the same time. Thus, we may conclude that the amount of increase or decrease in insulin is not affected by c-peptide. Note that there are also black regions in the map which shows insufficient data.

In Figure 2, the contour plot of the estimated local dependence values $H(x_i, y_j)$ for $(x_i, y_j)$ is illustrated. It helps us see the level of insulin and c-
peptide of a patient. For example, the variables having the value (2.1, 10.4) indicate positive local dependence approximately 0.75.

Example 2: TSH, FT3 and FT4

Thyroid-stimulating hormone (TSH) is released by the pituitary gland and circulates in the bloodstream to the thyroid where it controls the release of the thyroid hormones triiodothyronine (T3) and thyroxine T4. The level of free thyroxine (FT4), free triiodothyronine (FT3) and TSH identify how well the thyroid gland is working. Beyond the normal values of TSH it is necessary to consider the values of FT3 and FT4. Decreased level of TSH together with increased level of FT3 and/or FT4 shows hyperthyroidism. On the contrary, the increased level of TSH together with the decreased level of FT3 and/or FT4 shows hypothyroidism.

In this example we investigate the dependence structure in both TSH-FT3 and TSH-FT4. The data are supplied by DEU the faculty of medicine endocrine laboratory. 300 patients’ test results are examined and use 0.01-significance level is used for this aim. The Figure 3 displays the dependence map between TSH and FT3. Even though a weak negative correlation \( r = -0.1527 \) exists, two different dependence structures happen as in the previous example. Negative local dependence is seen in the region colored dark grey for (i) small values of TSH and moderate or large values of FT3 (ii) for moderate and large values of TSH and small values of FT3. In accordance with this observation, hyperthyroidism and hypothyroidism arise respectively. But one cannot make a similar consequence for both moderate values of the variables because it is not possible to mention a correlation between them. The Figure 4 displays the contour plot of estimated values \( \hat{H}(x, y) \). The values obtained from the local dependence function are accumulated around -0.15 which is the correlation coefficient.

In Figure 5 we investigate the dependence structure between TSH and FT4. The overall correlation between TSH and FT4 is \( r = -0.286 \) and this implies weak linear negative dependence. Two main areas of data points are deemed to have zero and negative local dependence. Negative dependence region colored dark grey occur for small values of FT4 and almost all values of TSH and also for small values of TSH in large values of FT4. This again gives us hyperthyroidism and hypothyroidism respectively. The difference here from TSH–FT3 is that although the correlation coefficient is greater in magnitude, the region negatively correlated is smaller. Moreover, for moderate values of FT4 one cannot expect local dependence in any values of TSH, implying zero local dependence.
In the last graph (Figure 6), the estimated local dependence values are given. For the smallest value of FT4 and highest value of TSH, the greatest in magnitude negative correlation (approximately -0.75) is observed. The minimum local dependence is attained at moderate values of FT4.

**SUMMARY**

Scalar dependence measures such as correlation coefficient can be an important part of medical and biological science studies. But these measures cannot be adequate to summarize complex dependence structure. The dependence between a pair of variables can be various with potentially surprising aspects. For bivariate data set the dependence structure can not only be measured globally, but the dependence structure can also be analyzed locally.

The aim of this study is to introduce a method to local analysis of dependence were presented and illustrated at hormone data examples. The method, local dependence map, requires the estimation new local dependence function which is based on regression concepts. After this local dependence function must be converted with local permutation tests in local dependence map which make the local dependence function more interpretable by identifying the regions of positive, negative and zero local dependence. Another advantage of this map is to decide the dependence structure for every observed pair of the data. This dependence map may play a considerable role in determining the dependence structure especially in medical and biological sciences.

**Acknowledgement**

The authors would like to thank Dr. Dilek Çimrin for supplying data from DEU Faculty of Medicine Endocrine Laboratory.

**APPENDIX 1: Local Permutation Test Code.**

(The code written in Excel Visual Basic software)

```vba
Option Explicit
Const MyError = -9
Dim StDevX, StDevY, XAve, HY, YAve As Double
Dim N As Integer
Sub TakeSpareColumn (What, Where)
Dim WhatS = What & "2:" & What & N + 1
WhatS = What & "2:" & What & N + 1
Range (WhereS).Value = Range(WhatS).Value
End Sub
Sub DoSample (SourceColIndex, TargetColIndex)
Dim SourceList As New Collection
Dim TargetList As New Collection
Dim I As Integer
Dim WhereS
WhereS = Where & "2:" & Where & N + 1
WhatS = What & "2:" & What & N + 1
Range (WhereS).Value = Range(WhatS).Value
End Sub
```

**Figure 6:** Contour plot of linear local dependence estimate and TSH-FT4 data.
Dim Index As Integer
' Getting data into the list
For I = 2 To N + 1
    SourceList .Add (Cells (I, SourceColIndex))
Next
' Getting data to the sample list (without replacement)
Randomize
For I = 1 To N
    Index = Int(((N - I + 1) * Rnd) + 1)
    TargetList. Add (SourceList (Index))
    SourceList. Remove (Index)
Next
For I = 1 To N
    Cells (I + 1, TargetColIndex) = TargetList(I)
Next I
End Sub
Function Sqrt (ByVal Numb As Double) As Double
    Sqrt = Numb ^ (0.5)
End Function
Function g() As Double
    g = Cells (6, 2)
End Function
Function Xi(I As Integer) As Double
    Xi = Cells (1 + I, 4)
End Function
Function Yi(I As Integer) As Double
    Yi = Cells (1 + I, 5)
End Function
Function UY(Y As Double, I As Integer) As Double
    UY = (Y - Yi(I)) / HY
End Function
Function UX(X As Double, I As Integer) As Double
    UX = (X - Xi(I)) / HX
End Function
Function AXY(Y As Double) As Double
    Dim Numerator As Double
    Dim Denominator As Double
    Dim Val As Double
    Dim I As Integer
    Numerator = 0
    Denominator = 0
    For I = 1 To N
        Val = (1 / Sqrt(44# / 7#)) * Exp((-1 * (UX(X, I)) ^ 2) / 2)
        Numerator = Numerator + (Xi(I) * Val)
        Denominator = Denominator + Val
    Next
    AXY = (Numerator/Denominator)
End Function
Function AYX(X As Double) As Double
    Dim Numerator As Double
    Dim Val, Denominator As Double
    Dim I As Integer
    Numerator = 0
    Denominator = 0
    For I = 1 To N
        Val = (1 / Sqrt(44# / 7#)) * Exp((-1 * (UY(Y, I)) ^ 2) / 2)
        Numerator = Numerator + (Yi(I) * Val)
        Denominator = Denominator + Val
    Next
    AYX = (Numerator/Denominator)
End Function
Function HXY(X As Double, Y As Double) As Double
    Dim AXY_Y, AYX_X As Double
    Dim SQ1, SQ2 As Double
    Dim Result As Double
    AXY_Y = AXY(Y)
    AYX_X = AYX(X)
    SQ1 = 1 + ((Xlave - AXY_Y) / StDevX) ^ 2
    SQ2 = 1 + ((Ylave - AYX_X) / StDevY) ^ 2
    If (SQ1 < 0) Or (SQ2 < 0) Then
        Result = MyError
    Else
        Result = (Sqrt(SQ1) * Sqrt(SQ2))
    End If
    HXY = Result
End Function
Sub DoTest()
    Dim TestNumber
    Dim I As Integer
    Dim J As Integer
    Dim Index As Long
    Dim HXY Number As Long
    Dim Sum HXY As New Collection
    Dim Min HXY As New Collection
    Dim Original HXY As New Collection
    Dim Maximum HXY As Double
    Dim Minimum HXY As Double
    Dim Temp HXY As Double
    Dim Sure As Date
    YAve = Cells (3, 2)
    XAve = Cells (2, 2)
    N = Cells (1, 2)
    HX = Cells (7, 2)
    HY = Cells (8, 2)
    StDevX = Cells (4, 2)
    StDevY = Cells (5, 2)
    Sure = Now
    Application. StatusBar = "Original H(x,y) is computing to compare"
    For I = 1 To N
        Original HXY. Add (HXY(Xi(I), Yi(I)))
    Next I
    Test Number = Cells (9, 2)
    Take Spare Column "E", "F"
    For I = 1 To Test Number
        Application. StatusBar = "  Getting Sample " & I
        DoSample 5, 7
        TakeSpareColumn "G", "E"
    Next I
End Sub
Application. Status Bar = "HXY is computing for sample * & I"
DoEvents
For J = 1 To N
    TempHXY = HXY(X(J), Y(J))
    If (TempHXY <> MyError) and (Original HXY. Item(J) <> MyError) Then
        If Abs(TempHXY) >= Abs(OriginalHXY. Item(J)) Then
            SumHXY. Add (TempHXY)
        End If
    End If
Next J
Next I
Columns(11). Clear
For I = 1 To SumHXY.Count
    Cells(I, 11) = SumHXY.Item(I)
Next
Columns (11). Select
Selection. Sort Columns (11)
HXYNumber = Int(SumHXY.Count * 0.025)
If HXYNumber = 0 Then
    HXYNumber = 1
End If
MinimumHXY = Cells(HXYNumber, 11)
MaximumHXY = Cells(ToplamHXY.Count - HXYNumber + 1, 11)
Cells (2, 9) = "Minimum HXY"
Cells (2, 10) = MinimumHXY
Cells (3, 9) = "Maximum HXY"
Cells (3, 10) = MaximumHXY
Cells (4, 9) = "Sampling Number"
Cells (4, 10) = HXYNumber
Application. StatusBar = "Observed H(x,y) is computing"
TakeSpareColumn "F", "E"
For I = 1 To N
    TempHXY = OriginalHXY.Item(I)
    If TempHXY <> MyError Then
        Cells(I + 1, 6) = TempHXY
        Cells(I + 1, 7) = 0
        If Cells(I + 1, 6) > MaximumHXY Then
            Cells(I + 1, 7) = 1
        End If
        If Cells(I + 1, 6) < MinimumHXY Then
            Cells(I + 1, 7) = -1
        End If
    Else
        Cells(I + 1, 6) = "*
        Cells(I + 1, 7) = "*
    End If
Next I
Application. StatusBar = ""
MsgBox "Done...", vbOKOnly
End Sub

REFERENCES