Çankaya Üniversitesi Fen-Edebiyat Fakültesi,
Journal of Arts and Sciences Sayı: 12 / Aralık 2009

Determination of Dependence Structure by Using Graphical Tools for Bivariate Continuous Data

Özlem Ege Oruç[*], *Zeynep F. Eren Doğu*[*]

Abstract

The dependence between a pair of continuous variates can be numerous with potentially surprising implications. Global dependence measures, such as Pearson correlation coefficient, can not reflect the complex dependence structure of two variables. For bivariate data set the dependence structure can not only be measured globally, but the dependence structure can also be analyzed locally. Thus, scalar dependence measures are extended to local dependence measures**.** In this study, determination of local dependence structure of bivariate data is discussed. For this, some graphical methods called chi-plot and local dependence map are used and examples with simulated and economic data set are illustrated.

Keywords: Local dependence, Correlation, Chi-plot, Local dependence map.

Özet

İki değişken arasındaki bağımlılık yapısı hakkında bilgi sahibi olmak oldukça önemlidir. Günümüzde yaygın olarak kullanılan global bağımlılık ölçüleri (Örneğin: Pearson korelasyon katsayısı gibi), iki değişken arasındaki karmaşık bağımlılık yapısını tam olarak yansıtamamaktadır. Bu nedenle iki değişken arasındaki bağımlılık yapısını sadece global bağımlılık ölçüleri ile değil lokal bağımlılık ölçüleri ile de incelemek gerekir. Çalışmada, iki değişken arasındaki bağımlılık yapısının belirlenmesinde kullanılan lokal bağımlılık ölçüleri incelenmiş ve bu yapının belirlenmesinde oldukça pratik araçlar olan ki-grafiği ve yerel bağımlılık haritaları kullanılarak uygulama yapılmış ve yorumlanmıştır.

Anahtar Kelimeler: Lokal bağımlılık, Korelasyon, Ki-grafiği, Lokal bağımlılık haritası.

* Dokuz Eylül University, Faculty of Arts and Sciences, Department of Statistics, Tinaztepe Campus 35160 Buca İzmir, TURKEY.
e-mails: ozlem.ege@deu.edu.tr, zeynepfiliz.eren@deu.edu.tr

1. Introduction

Dependence relations between random variables is one of the most widely studied topics in probability theory and statistics. For a bivariate data set the dependence structure can not only be measured globally, for example with the Pearson correlation coefficient, but the dependence structure can also be analyzed locally.

In this paper two graphical methods for analyzing dependence locally are discussed. These two methods indicate dependence quite differently. One graphical tool is the chi-plot introduced by (Fisher and Switzer, 1985). This utilizes the chi-measure to draw a plot which is approximately horizontal under independence. Another tool is local dependence map which is based on a new local dependence function which was introduced by (Bairamov et al., 2003). The new local dependence function is based on regression concepts and it can characterize the dependence structure of two random variables locally at the fixed points. It is possible to estimate the new local dependence function from data using, for example kernel methods. The local dependence map is constructed by applying this estimate. The local dependence map is a tool which makes local dependence easily interpretable. Via local permutation testing, local dependence maps simplify the estimated local dependence structure between variables by identifying regions of positive, zero and negative local dependence.

The organization of this paper is as follows: The next section introduces and discusses chi-plot. Section 3 introduces and discusses the local dependence map. Section 4 contains some examples of simulated and real data. Section 5 contains conclusion.

2. Chi-Plot

Fisher and Switzer (1985, 2001) introduce and discuss chi-plot that displays detailed and explicit information about the association between the two variables X and Y.

Chi-plot has some characteristic shapes depending on (a) whether the variables are independent or not, (b) have some degree of association, (c) have more complex dependence structure. The chi-plot does not depend on the value of pure data but it depends on data through the values of their ranks.

Let (X_i,Y_i), $(i=1,2,\ldots,n)$ be a random sample with bivariate distribution function, H. And let $I(Z)$ be an indicator function of the event Z. For each n bivariate sample points (X_i,Y_i), the (X,Y) plane is divided into quadrants by intersecting the regions $X \leq X_i$ and $Y \leq Y_i$. The cut point (X_i,Y_i) is just the i^{th} element of the random sample. Thus, we have $n-1$ remaining bivariate sample points. And these points are distributed among the four quadrants. By using these quadrant frequencies sample bivariate distribution function H_n and sample marginal distribution functions F_n and G_n are generated. For each data point (X_i,Y_i), $i=1,2,\ldots,n$;

$$H_n(X_i,Y_i) = H_{n_i} = \sum_{i=j} I(x_j \le x_i, y_j \le y_i)/(n-1)$$
$$F_n(X_i) = F_{n_i} = \sum_{i \ne j} I\left(x_j \le x_i\right)/(n-1) \quad (1)$$
$$G_n(Y_i) = G_{n_i} = \sum_{i \ne j} I(y_j \le y_i)/(n-1)$$

If the variables X and Y are statistically independent, we expect the joint bivariate distribution function H_n to be equal to the multiplication of the marginal distribution functions F_n and G_n at each of the sample cut point, (X_i, Y_i). With appropriate scaling, the differences $H_n - F_n(X_i) \times G_n(Y_i)$ will behave asymptotically like a standard normal variable under random sampling from independent marginals. The appropriate scaling factor that will transform the difference $H_n - F_n(X_i) \times G_n(Y_i)$ to a standard normal variable is:

$$\left(n^{\frac{1}{2}} \Big/ S_{n_i} \right)^{-1}, \text{ where } S_{n_i}^2 = F_{n_i}(1-F_{n_i}) \times G_{n_i}(1-G_{n_i}) \quad (2)$$

Summing all this information, we can find the signed and scaled measure: "standardized departure from bivariate statistical independence" by: $\chi_{n_i} = \left(H_{n_i} - F_{n_i} G_{n_i}\right) / S_{n_i}$, $i=1,2,\ldots,n$. At each sample point, χ_{n_i} actually acts like a correlation coefficient between dichotomized X values and dichotomized Y values. Hence, χ_{n_i} can take only the values in the interval $[-1,1]$. $\tilde{\chi}_{n_i}$ asymptotically approaches to ρ as $\lambda_{ni} \to 0$; i.e, $\lim_{\substack{n \to \infty \\ \lambda_{n_i} \to 0}} \left(\tilde{\chi}_{n_i}\right) = \rho$. If Y is a strictly increasing function of X, then $\chi_{n_i} = 1$ for all sample cut points, and vice versa. Under the independence of X and Y; if the marginal frequencies are not too close to zero or one, the approximate normal distribution of each χ_{n_i} $(i=1,2,\ldots n)$ will not be affected. Therefore, λ should be a real valued function of the marginal frequencies. If χ_{n_i} values are plotted against λ_{n_i} values, for $i=1,2,\ldots,n$, the plot will asymptotically show normal vertical scatter with variance $\frac{1}{n}$ around the horizontal $\chi = 0$, under the independence of X and Y. As a result, the association between X and Y can be determined by the departures from that zero-centered vertical scatter. (Fisher and Switzer, 1985) had chosen λ as:

$$\lambda_{n_i} = 4\,\mathrm{sgn}_{n_i} \max\left\{\left(F_{n_i} - \frac{1}{2}\right)^2, \left(G_{n_i} - \frac{1}{2}\right)^2\right\} \tag{3}$$

where $\mathrm{sgn}_{n_i} = \mathrm{sgn}\left\{\left(F_{n_i} - \frac{1}{2}\right), \left(G_{n_i} - \frac{1}{2}\right)\right\}$

Similar to χ_{n_i} values, all values of λ_{n_i} must also lie in the interval *[-1,1]*. When the bivariate data (X_i,Y_i), $(i=1,2,\ldots,n)$ are random bivariate sample from independent continuous marginals, then the values of λ_{n_i} are individually uniformly distributed. However when *X* and *Y* are somehow associated, then the values of λ_{n_i} may show clustering.

If *X* and *Y* are positively correlated, λ_{n_i} will tend to be positive and if negatively correlated, λ_{n_i} will tend to be negative. The contrary is an exception. Thus, we may think $|\lambda_{n_i}|$ as a "measure of the distance" from the sample point (X_i,Y_i) to the bivariate median of the distribution. As a result of this property of λ, sample points that cause departure from independence could be seen on the chi-plot. Moreover we can see the direction of those points with respect to the centre of the data set. Those sample points for which $|\lambda_{n_i}| \geq 4 \leq \left(\frac{1}{n-1} - \frac{1}{2}\right)^2$, have not been plotted in the illustrations in order to avoid any probable confusions. This censoring criterion will eliminate at most *8* points. If there is no degree of monotone dependence between *X* and *Y* but there is some dependence of a more complex nature, this will be manifested in the $\tilde{\chi}$ values in terms of the increased scatter, and possibly non-uniform increase in scatter, along the λ-axis.

3. Local Dependence Map

Local dependence map is another graphical tool for determination of local dependence structure of bivariate random variables. Unlike chi-plot and *K*-plot, local dependence map depends on the bivariate density function.

Local dependence function was introduced by (Holland and Wang, 1987) and is further developed in a series of subsequent papers by (Jones, 1996, 1998) and Jones and Koch (2002). It is based on the mixed partial derivatives of the logarithm of the bivariate density function. (Bairamov et al., 2003) introduced a new local dependence

function based on regression concepts. This measure is symmetric with respect to two random variables and its expected value is approximately equal to the Pearson correlation coefficient. It is possible to estimate the new local dependence function from data using the kernel methods.

(Bairamov and Kotz, 2000) suggested an estimator for H(x,y) by using Nadaraya and Watson's (1964) estimate for the regression functions $E(X|Y=y)$ and $E(Y|X=x)$.

$$\hat{A}_X(y) = \frac{\sum_{i=1}^{n} X_i K\left(\frac{y - Y_i}{h_y}\right)}{\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_y}\right)} \quad \text{and} \quad \hat{A}_Y(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h_x}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}\right)} \tag{4}$$

where (X_i,Y_i), $i=1,2,...,n$ are the data, K is a kernel function, an integrable function with short tails, and h_n is a width sequence tending to zero at appropriate rates. The estimate for $H(x,y)$ is:

$$\hat{H}(x,y) = \frac{\rho + \frac{(\overline{X} - \hat{A}_X(y))(\overline{Y} - \hat{A}_Y(x))}{S_X S_Y}}{\sqrt{1 + \left(\frac{\overline{X} - \hat{A}_X(y)}{S_X}\right)^2}\sqrt{1 + \left(\frac{\overline{Y} - \hat{A}_Y(x)}{S_Y}\right)^2}} \tag{5}$$

The first step to estimate $H(x,y)$ is to take the kernel to be the product of biweight univariate densities:

$$K_{Beta}(u) = (15/16)(1-u^2)^2, \ -1 \le u \le 1 \tag{6}$$

We prefer this kernel to estimate $H(x,y)$, because it has computational advantages over the normal kernel. After determining the kernel function, we determine that $h_n = \frac{1}{n}$.

Generally, it is hard to interpret the dependence structure of the bivariate random variables from their local dependence functions. It can be argued that the local dependence function convey information that is too detailed to be easily interpretable. This fact motivates (Jones and Koch, 2002) to make local dependence a more interpretable tool, by introducing so called dependence maps. In this study we construct a new dependence map by using local dependence function of (Bairamov et al.'s, 2003) which is based on regression concept. Via local permutation testing, dependence maps simplify the

estimated local dependence structure between variables by identifying regions of (significant) positive, (nonsignificant) zero, and (significant) negative local dependence (Ege Oruç and Üçer H., 2009).

Local permutation test is applied to construct dependence maps. Local permutation test's null hypothesis $H(x,y)=0$ is equivalent to independence. Samples satisfying the hypothesis $H(x,y)=0$ can be generated. This procedure is repeated n times and it is computed for each permuted data set $H_P(x_i, y_i)$. When the estimated local dependence function is the highest $(\alpha/2)\%$ of the simulated $H_P(x_i, y_i)$ is designated to be $(+1)$ and also when the estimated local dependence function is the lowest $(\alpha/2)\%$ of the simulated $H_P(x_i, y_i)$ is designated to be (-1) and otherwise it is zero. Then by using these values, the dependence map could be designed easily. On the contour plot, the areas that have zero estimated local dependence are colored by light grey, the positive estimated local dependence are colored by white, and the negative estimated local dependence are colored by dark grey. In order to stabilize whether the estimated local dependence function $\hat{\gamma}$ is significant or insignificant (i.e.zero), multiple hypothesis are tested by Local Permutation Test (Üçer H. and Bayramoğlu, 2007). This makes the visual inspection much easier. If the estimated local dependence is near zero, local permutation test accepts it as zero dependence and avoids over interpretation of the sign of the local dependence when insignificant.

4. Examples

Example 1. The example 1 concerns the data set gasoline tax and price: We obtain the data from a report published by the U.S. Energy Information Administration (Petroleum Marketing Monthly). There are *100* observations on the following *2* variables: gasoline price and tax. Scatter plot shows nearly a perfect positive dependence between these two variables, and the correlation coefficient is found to be *0.876*, which is significantly different from zero. Chi-plot displays positive monotone association between the variables and moreover local dependence map also exhibits the positive dependence with dominant white area. But the dependence map is particularly informative here. Although there is a positive correlation between the variables, two different dependence structures are observed in Fig. 1. In the white region where the variables simultaneously take low, moderate and large values, positive dependence between gasoline "tax and price" appear. That is, as the tax increases (decreases) so does the price. The white region covering relatively larger than the others can be explained by the existence of highly positive correlation between the variables tax and price. In other words it is important to interpret in detail the correlation coefficient in this region. 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 tax and price. Further,

while tax takes moderate values, price takes moderate values at the same time. Thus, we may conclude that the amount of increase or decrease in tax is not affected by price. Both of the dependence graphs also show that the Pearson correlation coefficient is not suitable to explore the complex dependence structure between tax and price.

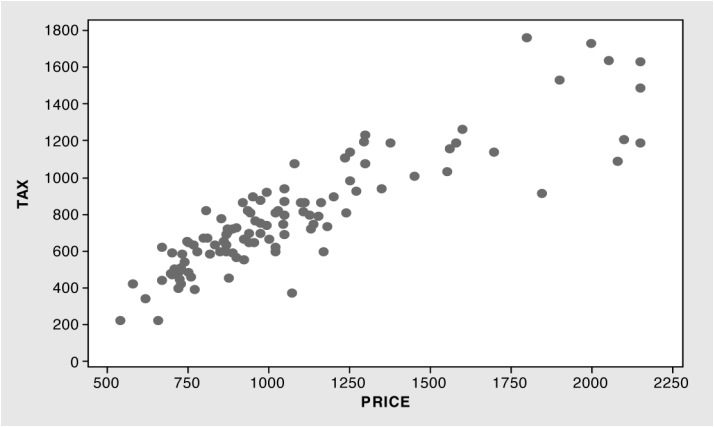

Fig.1-a. Scatter plot for Example 1

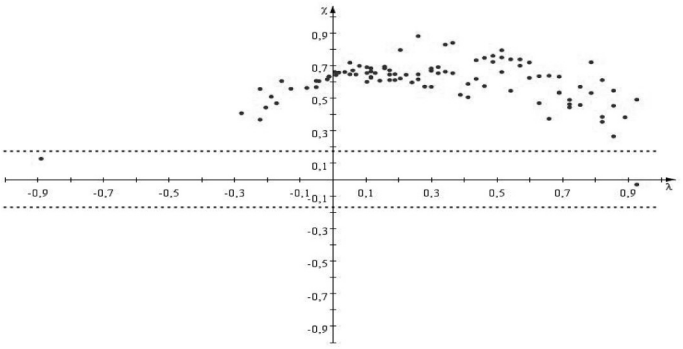

Fig.1-b. Chi-plot for Example 1

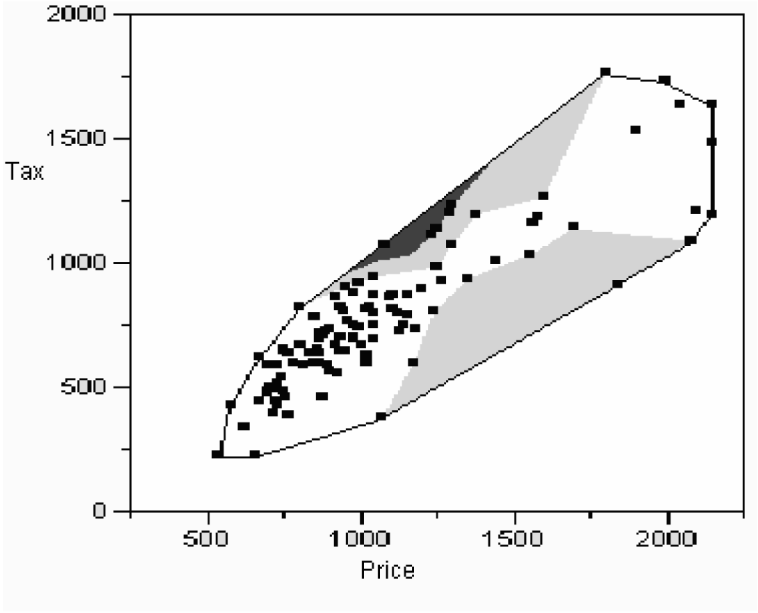

Fig.1-c. Local dependence map for Example 1

Example 2. In Fig.2 we investigate the dependence structure between X and Y which are generated by bivariate normal variables with correlation coefficient is *-0.10* that is too small. Scatter plots and chi-plots coincide with each other. Scatter plot does not have a pattern, this means that there is no association between the variables. Also the points on the chi-plot are appropriately distributed along the lines, which mean that there is no association between the variables too. The dependence map is particularly informative here. The overall correlation between *X* and *Y* is $\rho = -0.10$ and this implies weak negative linear dependence. Two main areas of data points are deemed to have zero and negative local dependence. Negative dependence region colored dark grey occur for moderato values of *X* and large values of *Y*. Although the correlation coefficient is greater in magnitude, the region negatively correlated is smaller. Moreover, for large and small values of *Y* one cannot expect local dependence in any values of *X*, implying zero local dependence.

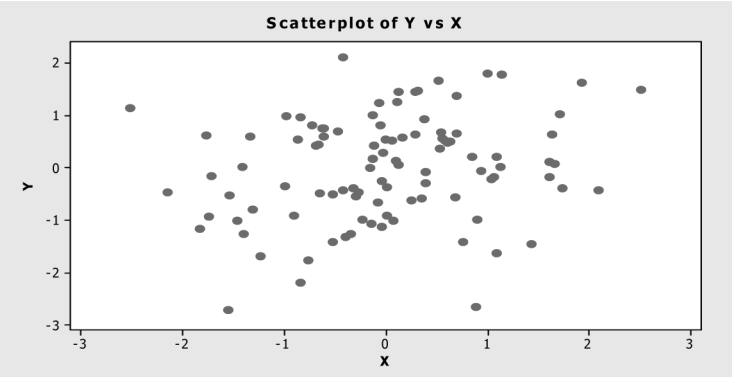

Fig.2-a. Scatter plot for Example 2

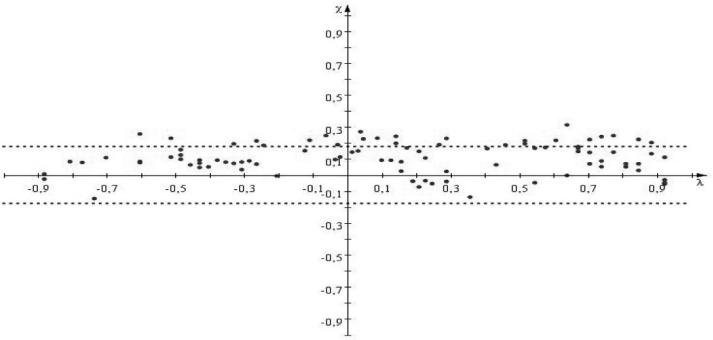

Fig.2-b. Chi-plot for Example 2

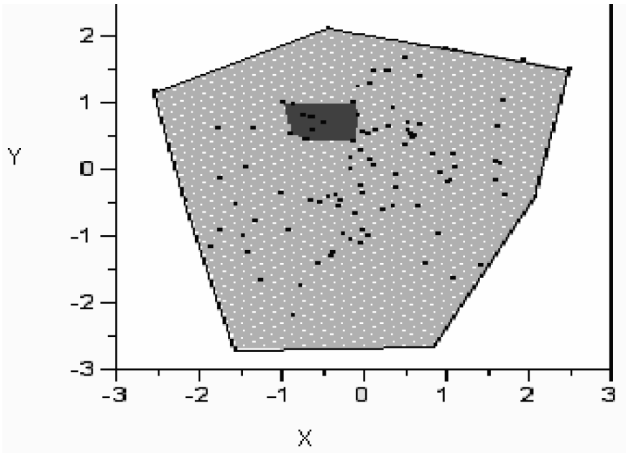

Fig.2-c. Local dependence map for Example 2

5. Conclusion

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

In previous sections two methods for local analysis of dependence were presented and illustrated with real and simulated data examples. The first method, the Chi-plot, is simple to calculate and is well suited to recognize dependence in the tails of the distribution. The second method, new local dependence map, although is quite difficult to calculate the regions on it can be indicated for the dependencies which are easy to interpret.

Note: To construct the chi-plots, the Java program written by Z. F. Eren and C. Kınacı is used. On the other hand, in order to plot local dependence map, we use a Visual Basic code in MS Excel for applying permutation test to our data which was written by (Oruç and Üçer, 2009).

References

1. D.Mari and S.Kotz, *Correlation and Dependence (London, Imperial College Press, 2001).*
2. B. Hüdaverdi Üçer and Bayramoğlu İ., 'A New Epsilon-Local Dependence Measure and Dependence Maps', *SJAM*, 8(2) (2007), pp.3-12.
3. I. Bairamov, S. Kotz and T. Kozubowski, 'A New Measure of Linear Local Dependence', *Statistics*, 37(3) (2003), pp. 243-258.
4. I. Bairamov, S. Kotz, 'On Local Dependence Function for Multivariate Distributions', *New Trends in Prob. and Stat.*, 5(2000), pp.27-44.
5. K. Abberger, 'Exploring Local Dependence', *University of Konstanz*, (2003) pp.1-14.
6. M.C.Jones , I. Koch, 'Dependence maps: Local Dependence in Practice', *Statistics and Computing*, 13(2003), 241-255.
7. N. I. Fisher, P. Switzer, 'Graphical Assessment of Dependence: Is a Picture Worth 100 Tests?, *The American Statistician*, 55(3) (2001), pp.233-239.
8. N. I. Fisher, P. Switzer, 'Chi-plots for Assessing Dependence', *Biometrika*, 72(2) (1985), pp. 253-265.
9. Ö. Ege Oruç, B. Hüdaverdi Üçer, 'A New Method For Local Dependence Map and Its Applications', *Türkiye Klinikleri Biyoistatistik Dergisi*, 1(1) (2009), pp. 1-8.