QUANTITATIVE METHODS IN ECONOMICS
Vol. XIV, No. 1, 2013, pp. 180 – 189

FUNCTIONAL EXPLORATORY DATA ANALYSIS OF UNEMPLOYMENT RATE FOR VARIOUS COUNTRIES

Stanisław Jaworski
Department of Econometrics and Statistics
Warsaw University of Life Sciences – SGGW
e-mail: stanislaw_jaworski@sggw.pl

Abstract:. Functional exploratory techniques are applied in the analysis of an unemployment rate. The rate is smoothed into differentiable function in order to facilitate the analysis. The main aim of the analysis is assigned to find out the unemployment curves which do not follow the same pattern as that of the other ones.

Keywords: B-splines basis system, smoothing with roughness penalty, functional principal component analysis, cluster analysis, depth measures, unemployment rate

INTRODUCTION

Unemployment rate represents unemployed persons as a percentage of the labour force based on International Labour Office definition. The labour force is the total number of people employed and unemployed. Unemployed persons comprise persons aged 15 to 74 who: a) are without work during the reference week; b) are available to start work within the next two weeks; c) have been actively seeking work in the past four weeks or had already found a job to start within the next three months.

Unemployment rate is an important economic indicator with wide range of social dimensions and is one of the primary goals of macroeconomic policy. It is also a key indicator of overall economic performance. A rising rate is seen as a sign of weakening economy that may call for cut in interest rate. A falling rate, similarly, indicates a growing economy which is usually accompanied by higher inflation rate and may call for increase in interest rates. Rapid change of the rate is a strong signal for a feasible grown or drop in a country's economy [Burgen et al. 2012].

The rate is not a perfect indicator of economic activity or inactivity. There are three main areas of criticisms regarding the measurement of the rate: survey accuracy, discouraged workers and underemployed workers. The criticism reflect the definitional and technical pitfalls involved in the preparation of several unemployment data emanating from different sources of various countries. Thus the interpretability of the rate should come not only from inspecting its level but also from its pace.

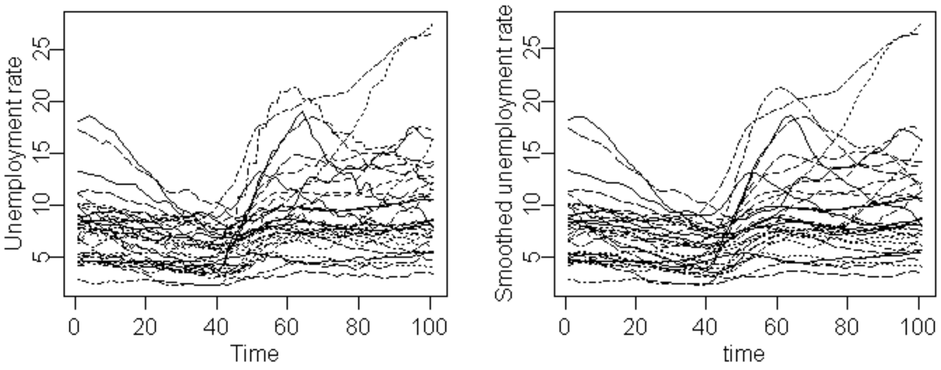
DATA ANALYSIS

The aim of the paper is to summarize main characteristics of the unemployment rate in various countries, mainly in Europe. The investigated data are presented in a seasonally adjusted form. They relate to Belgium, Bulgaria, Czech Republic, Denmark, Germany, Estonia, Ireland, Greece, Spain, France, Italy, Cyprus, Latvia, Lithuania, Luxembourg, Hungary, Malta, Netherlands, Austria, Poland, Portugal, Romania, Slovenia, Slovakia, Finland, Sweden, United Kingdom, Norway, Croatia, Turkey, United States and Japan. The series are monthly collected since 2005.01 to 2013.05.

Each unemployment curve is presented in a raw and smoothed form in Figure 1. Smoothness, in the sense of possessing a certain number of derivatives, is a property of a latent function. We assume the existence of the function and that it is giving rise to the observational unemployment rate. The approach reflects the idea that an unemployment rate is driven by an overall economic performance and should be filtered out of the noise coming from several shortcomings involved in the preparation of data. Moreover data smoothing allows us to inquire the dynamics of unemployment rates of various countries in a unified way by means of derivative of smoothed unemployment rate with respect to time.

As a method of turning raw discrete data into smooth functions we chose smoothing by a B-spline basis with a roughness penalty (the penalty was based on the integral of the square of the derivative of order 2). A basis is a set of known functions $\{\phi_k\}_{k\in N}$ that any function $X(t)$ could be approximated by a linear combination of a sufficiently large K: $X(t) \approx \sum_{k=1}^{K} c_k \phi_k$ (see details in [Ramsay et al. 2005]. If smoothing penalization is required, apart of $\{c_k\}_{k=1,\ldots,K}$, an additional parameter λ is involved in estimation. The choice of the number of basis K and the smoothing penalty parameter was made according to the cross-validation criteria. The investigated data and method gave $K = 31$ and $\lambda = \mathbf{1.574804}$.

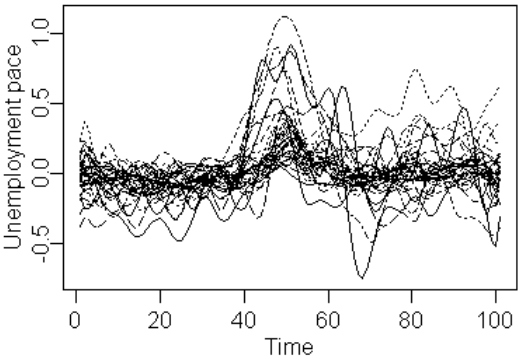
Figure 1. Raw and smoothed unemployment rate for various countries



Source: own preparation

The smoothed unemployment rate was used to represent an unemployment pace as a first derivative. The pace is presented in Figure 2.

Figure 2. The unemployment pace of various countries

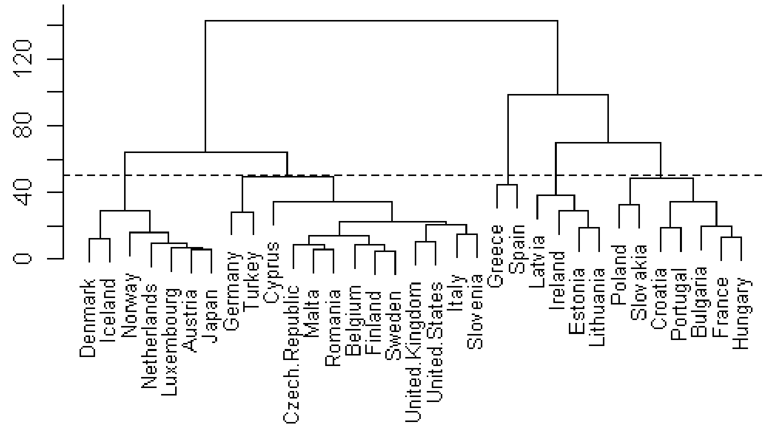


Source: own preparation

The plots in Figure. 1 and 2 are not particularly informative though some general conclusions can be drawn. First we can find unemployment convergence between countries covering the period 2005.01-2008.05 (Time 0-40 in Figure 1) and unemployment disparities intensified as of 2008.05. Second the corresponding unemployment pace in various countries was relatively low and stable until the first quarter of 2008 and then its level grew and became much more diversed. However, the curves layout in Figure 1. and Figure 2. suggests that there is a low-level unemployment dynamics in several countries.

By the use of metric and semi-metric functions, some additional information about the investigated unemployment rates can be extracted. Let us first consider the L2-metric $d(X_1, X_2) = \sqrt{\int_{Time} (X_1(t) - X_2(t))^2 dt}$ and the dissimilarity structure produced by the metric. The structure is presented in Figure 3 (a complete agglomeration method was used).

Figure 3. Dendrogram of the smoothed unemployment rate



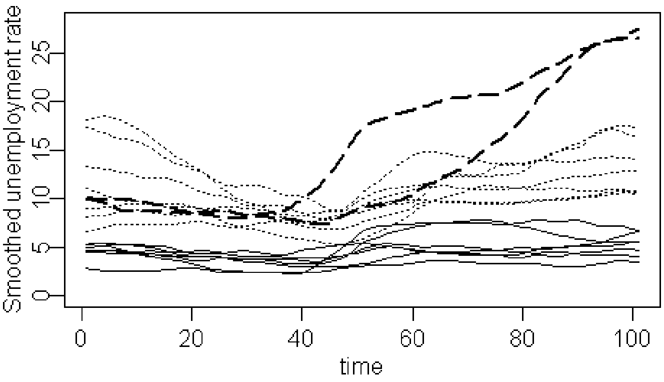
Source: own preparation

According to the proposed L2-metric, agglomerated countries reflect similar pattern of unemployment level in the investigated period 2005.01 - 2013.05. For example at the threshold 50 (dashed line in Figure 3) five groups are extracted. Let us consider the three of them: 1. Denmark, Iceland, Norway, Netherlands, Luxembourg, Austria, Japan; 2. Poland, Slovakia, Croatia, Portugal, Bulgaria, France, Hungary; 3. Greece, Spain. Each of the three groups are marked with a different style line in Figure 4. and represents curves which are close to each other over the whole time period. The lower-rate thresholds give finer and more homogenous divisions. For example the second group is divided into two parts at the threshold 40. The first group is not divided in the case because the group is less diverse.

The L2-metric takes into account all sources of variability between curves. In the case we want to investigate the variability in terms of its main sources, we can use the following semi-norms $d(X_1, X_2) = \sqrt{\sum_{k=1}^{q} \left(\int_{Time} (X_1(t) - X_2(t)) v_k(t) dt \right)^2}$, where $v_1, v_2, \ldots, v_q$ are orthonormal eigenfunctions of a covariance operator (see

[Horváth et al. 2012] and [Ferraty 2006]. Here, q is a tuning parameter indicating the resolution level at which the problem is considered. To understand its meaning a functional principal analysis should be applied (the details of the analysis are not presented in the paper; the interested reader is referred to [Furmańczyk et al. 2012]. According to the analysis the first two components (q=2) explain 94% of the total variability of the investigated curves: 78% - the first component and 16% -the second component. The first component reflects the differences between average levels of unemployment rates and the second one the effect of crossing unemployment rate curves in the period 2008-2009. These two sources of variability are depicted in [Furmańczyk et al. 2012], where the same set of unemployment rates is investigated but in the shorter time range, that is for 2005 till 2012.

Figure 4. Unemployment rates of Denmark, Iceland, Norway, Netherlands, Luxembourg, Austria, Japan (solid lines) and Poland, Slovakia, Croatia, Portugal, Bulgaria, France, Hungary (dotted lines), Greece and Spain (dashed lines).

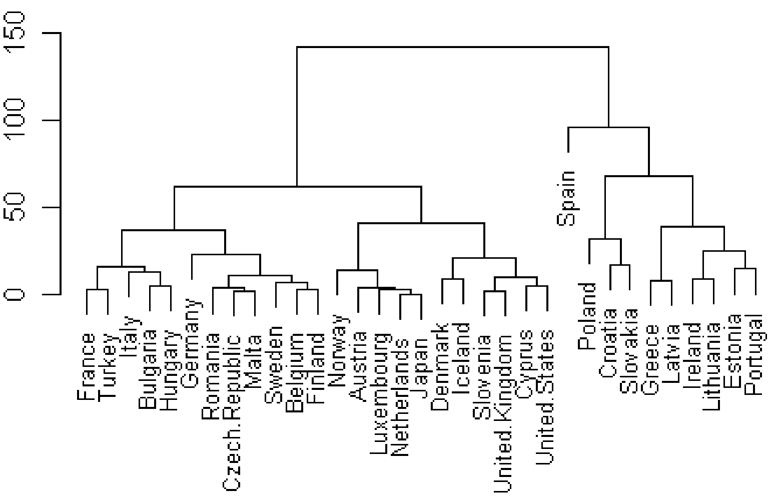


Source: own preparation

In [Furmańczyk et al. 2012] the weights of the first two sources of variability are accordingly 72% and 20%. It means that the changes of an unemployment rate in 2013 differentiated countries with respect to its first source of variability. The dissimilarity structure produced by the semi-metrics is exhibited in Figure 5. The threshold 50 gives five groups. Although the number of groups is the same as in the case of L_2-metric, the groups are different because the division does not take into account 6% of the remaining variability. For example the curves representing unemployment rates of Greece and Latvia are not close to each other though they are in the same cluster of the dendrogram in Figure 5. The reason is that the integrals $\int_{Time}(X_1(t)-X_2(t))v_k(t)dt$ for k=1,2 are close to zero in the case of the two curves. In fact the curves are crossing three times in the observed period (see Figure 6).

Greece and Spain, as depicted by Figure 3, are candidates for outliers. By Figure 4, Spain is additionally depicted as an outlier in terms of the main sources of variability. In order to identify other outliers, depth measures can be used (see [Febrero-Bande et al. 2008] and [Cuevas et al. 2007] for theoretical details). A way to detect the presence of functional outliers is to look for curves with lower depth and is based on bootstrap. The result of the approach is presented in Figure 6. As a depth we used a random projection method.

Figure 5. Dendrogram of the smoothed unemployment rate based on first two principal components

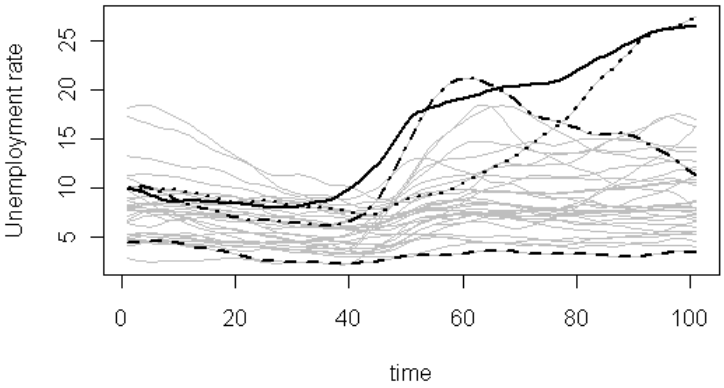


Source: own preparation

In the method a random direction a and projection of the data along the direction are taken: $\int_{Time} a(t)X_i(t)\,dt,\ i = 1,2,\dots,n$. Then the sample depth of a datum X_i is defined as the univariate depth of the corresponding one-dimensional projection (expressed in terms of order statistics so that the median is the deepest point). A single representative value is obtained by averaging on a. The direction of a is chosen according to Gaussian distribution. The bootstrap procedure is designed to select C such that, in the absence of outliers, the percentage of correct observations mislabeled as outliers is approximately equal to 1%: $\Pr(D(X_i) < C) = 0.01$, $i = 1,2,\dots,n$, where $D(X_i)$ denotes depth of the datum X_i. That is the cutoff C in the procedure is found by estimating this percentile, making use of the observed sample curves. In the case of our curves we used the smoothed bootstrap procedure based on trimming described in [Febrero–Bande et al. 2008].

The four exposed outlying curves in Figure 6 are 12% of total number of curves.

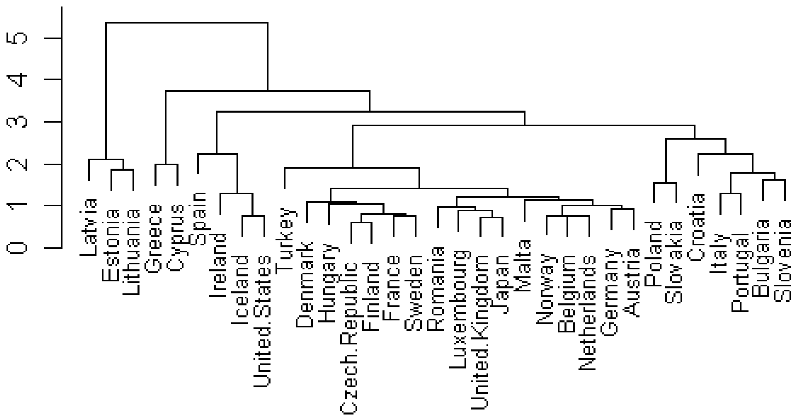
Figure 6. Outlying curves (bold lines). Spain (solid line), Norway (dashed line), Greece (dotted line), Latvia (dashed-dotted line)



Source: own preparation

We hinted in *Introduction* that interpretability of an unemployment rate should result not only from inspecting its level but also from its pace. This is the main reason why we represented unemployment rates as smooth differerentiable curves. The unemployment pace is thus well defined and we explored it by the same techniques as in the unemployment rate case (Figures 7, 8, 9).

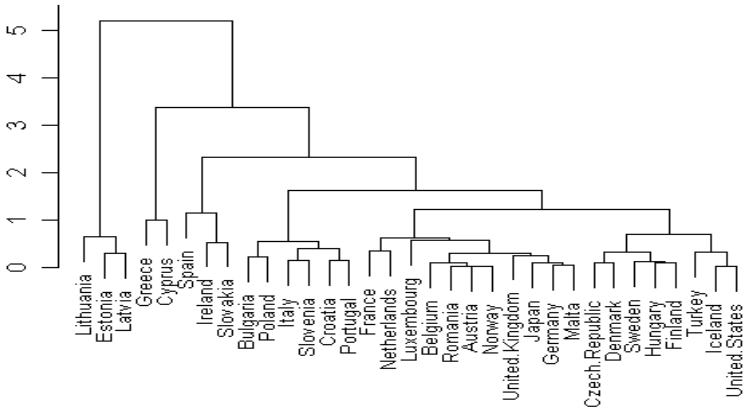
Figure 7. Dendrogram of the smoothed unemployment pace; being computed by means of the classical L_2-metric



Source: own preparation

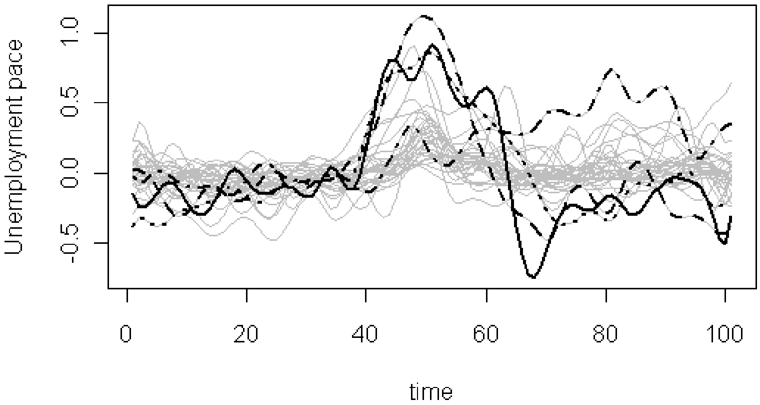
The placement of Latvia, Estonia and Lithuania in the dendrograms of Figure 7 and Figure 8 indicate that the unemployment pace of these countries is substantially different from the unemployment pace of the remaining countries. These countries are of potential outliers. Same Greece, Spain and Cyprus. Similar conclusions can be drawn from Figure 9. By cutting dendrogram in Figure 8 at level 1.5 we are receiving 5 groups, where the group with Norway is the biggest one and have a stable unemployment rate. The biplot in Figure 10 shows the placement of the extracted groups in terms of principal scores. Note that opposite to outliers the countries with stable unemployment pace are located close to point (0,0).

Figure 8. Dendrogram of the smoothed unemployment pace; being computed by means of the PCA-metric



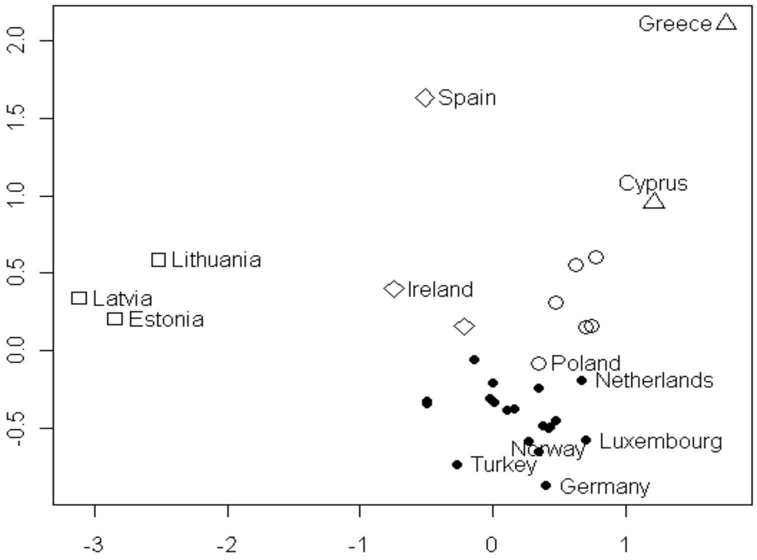
Source: own preparation

Figure 9. Outlying unemployment pace curves (bold lines). Estonia (solid line), Latvia (dashed line), Lithuania (dotted line), Greece (dashed-dotted line)



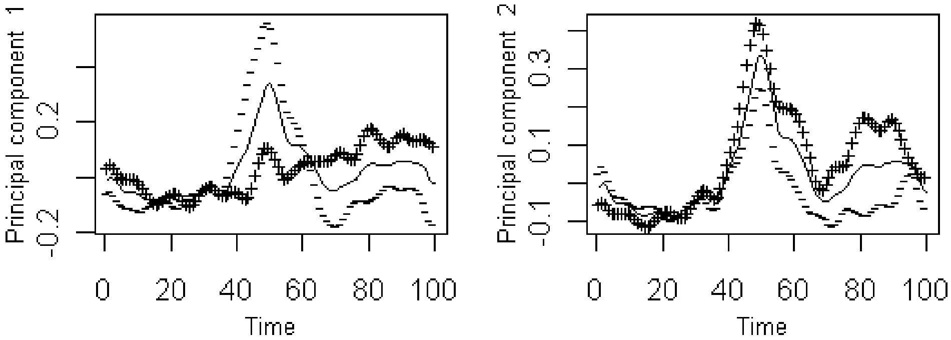
Source: own preparation

Figure 10. Biplot of the first two principal scores for unemployment pace. Different symbols are used for the five clusters separated by 1.5 threshold in the dendrogram of Figure 8.



Source: own preparation

Figure 11. First two principal components as perturbations of the mean unemployment pace.



Source: own preparation

According to Figure 10 and Figure 11 we can reflect relative dissimilarity of outliers. Latvia, Lithuania and Estonia have relatively high positive unemployment pace for 2009 till 2010 and negative after the period. Greece, Cyprus and Spain have relatively positive unemployment pace as of 2008.

SUMMARY

Functional exploratory techniques are applied in the analysis of an unemployment rate and pace. The rate is smoothed into differentiable function. The first derivative of the smoothed rate represents the unemployment pace.

An insight into the differences between countries is made by cluster analysis, functional principal components analysis and depth measures analysis. The hierarchical cluster analysis is used on a set of dissimilarities, where a dissimilarity structure is produced by the classical L_2 norm and principal components semi-norm. The depth notion is applied to outlier detection procedure.

The techniques detect Estonia, Latvia, Lithuania, Greece, Norway and Spain as outliers and show countries with comparable unemployment rate and pace. In the case of outliers the analysis reflects the nature of the depicted dissimilarities. It is shown that the dissimilarities begun as of 2008-2009 and still hold, which is a strong signal for a feasible changes in the countries' economy.

REFERENCES

Burgen E., Meyer B., and Tasci M. (2012) An Elusive Relation between Unemployment and GDP Growth: Okun’s Law. Cleveland Federal Reserve Economic Trends.

Cuevas A., Febrero-Bande M., Fraiman R. (2007) Robust Estimation and Clasification for Functional Data via Projction-Based Depth Notions, Computational Statistics, 22(3), 481-496.

Febrero-Bande M., Galeano P., Gonzales-Manteiga W. (2008) Outlier Detection in Functional Data by Depth Measures, with Application to Identity Abnormal NOx Levels, Environmetrics, 19(4), 331-345

Ferraty F., Vieu P. (2006) Nonparametric Functional Data Analysis, Springer Series in Statistics. Springer-Verlag, New York. Theory and practice.

Furmańczyk K., Jaworski S. (2012) Unemployment rate for various countries since 2005 to 2012: comparison of its level and pace using functional principal analysis. Quntitative Methods in Economics, Vol. XIII, No 2, pp. 40-47.

Horváth L, Kokoszka P. (2012) Inference for Functional Data with Applications, Springer Series in Statistics

Ramsay J.O., Silverman B. W. (2005) Functional Data Analysis. Second Edition, Springer, NY.