Local Principal Curves With Application to Econometric Data

Mohammad A. Zayed

Department of Mathematical Sciences, Durham University, UK

Durham-Newcastle Postgraduate Training

May 2010

Outline of presentation:

Introduction

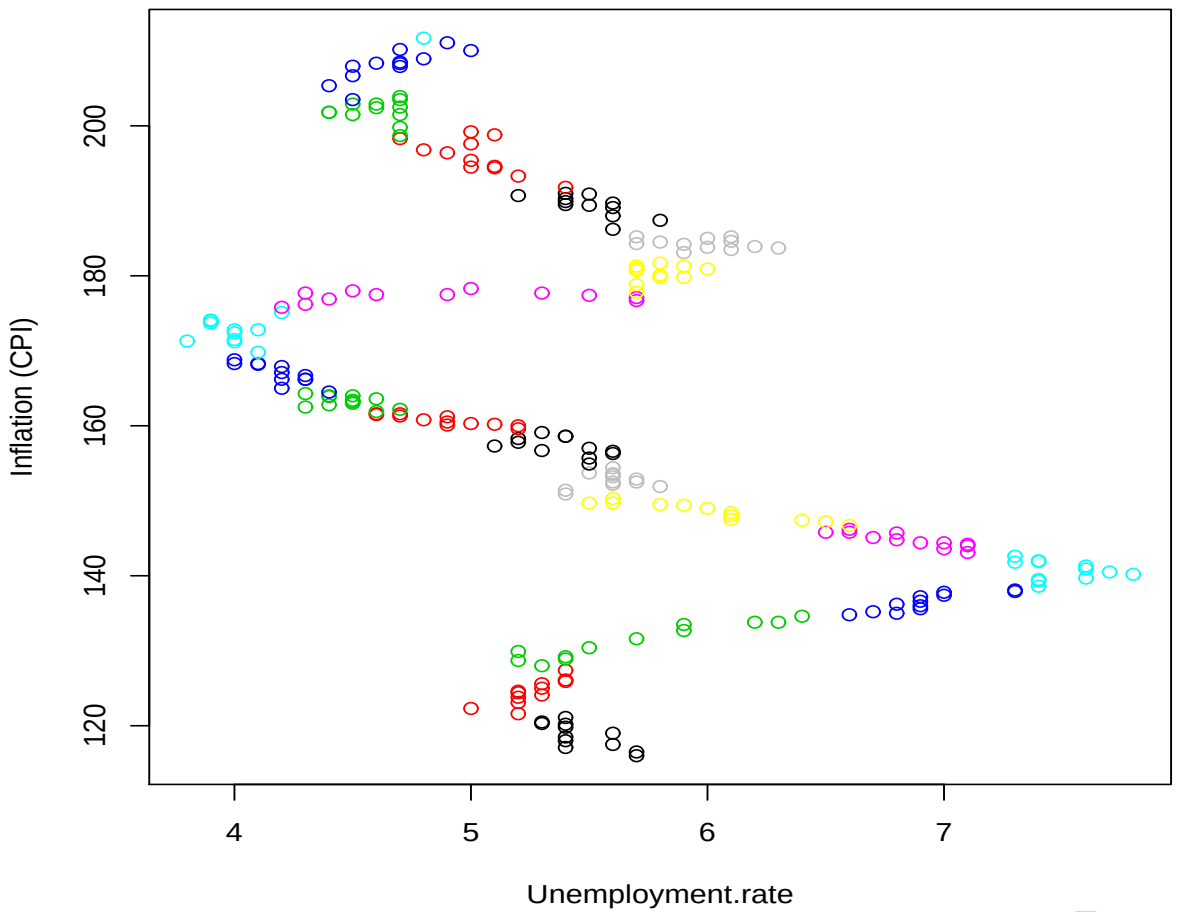
Local Principal Curves

An Econometric application

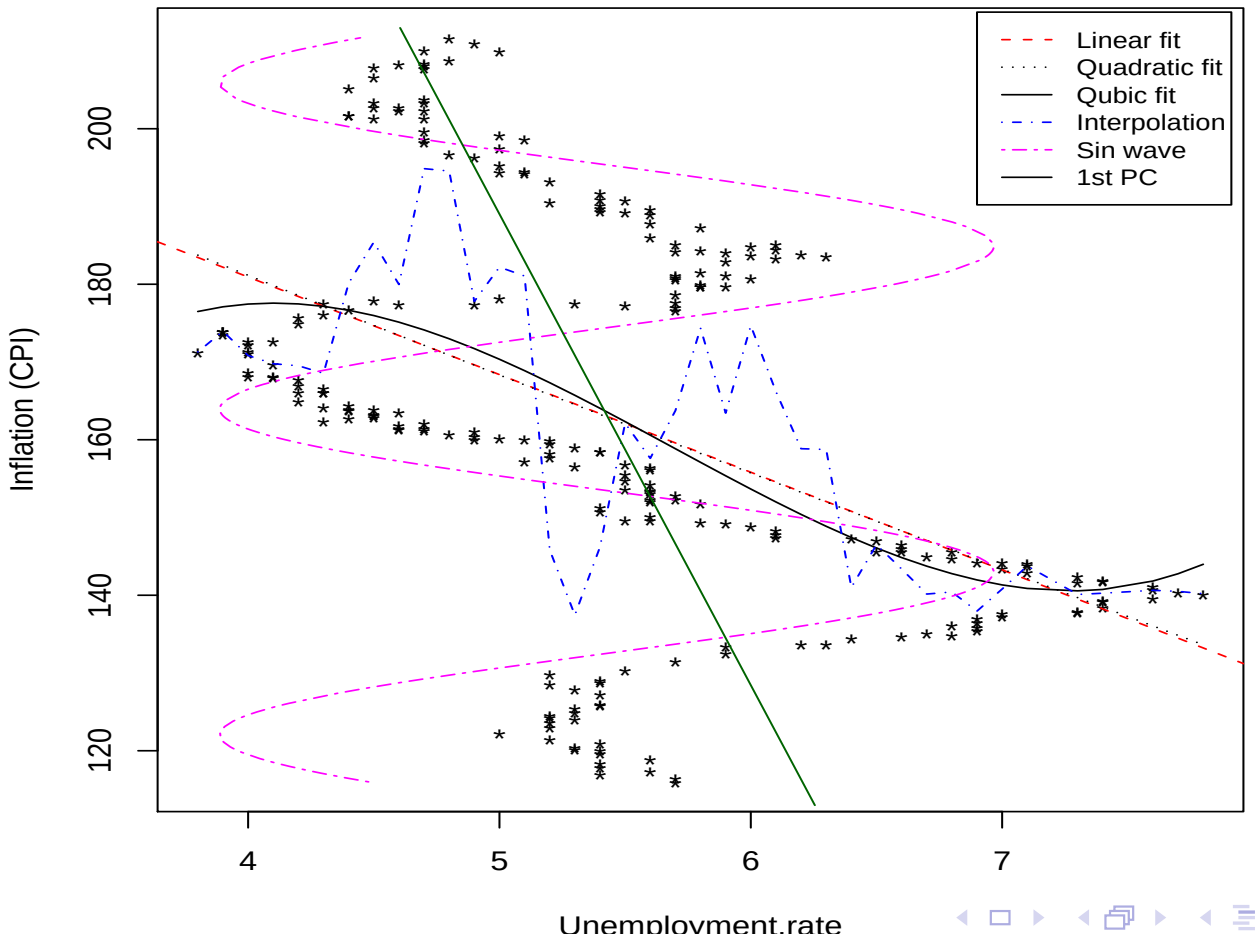
Conclusion

This brief presentation gives an insight into Local Principal Curves (LPCs) as a non-parametric tool for representing multidimensional data structures by a single smooth one-dimensional curve. We also explore some possible econometric applications in which LPCs might be a good choice as a method to represent data as well as (with some smart tools) predicting within data range as well.

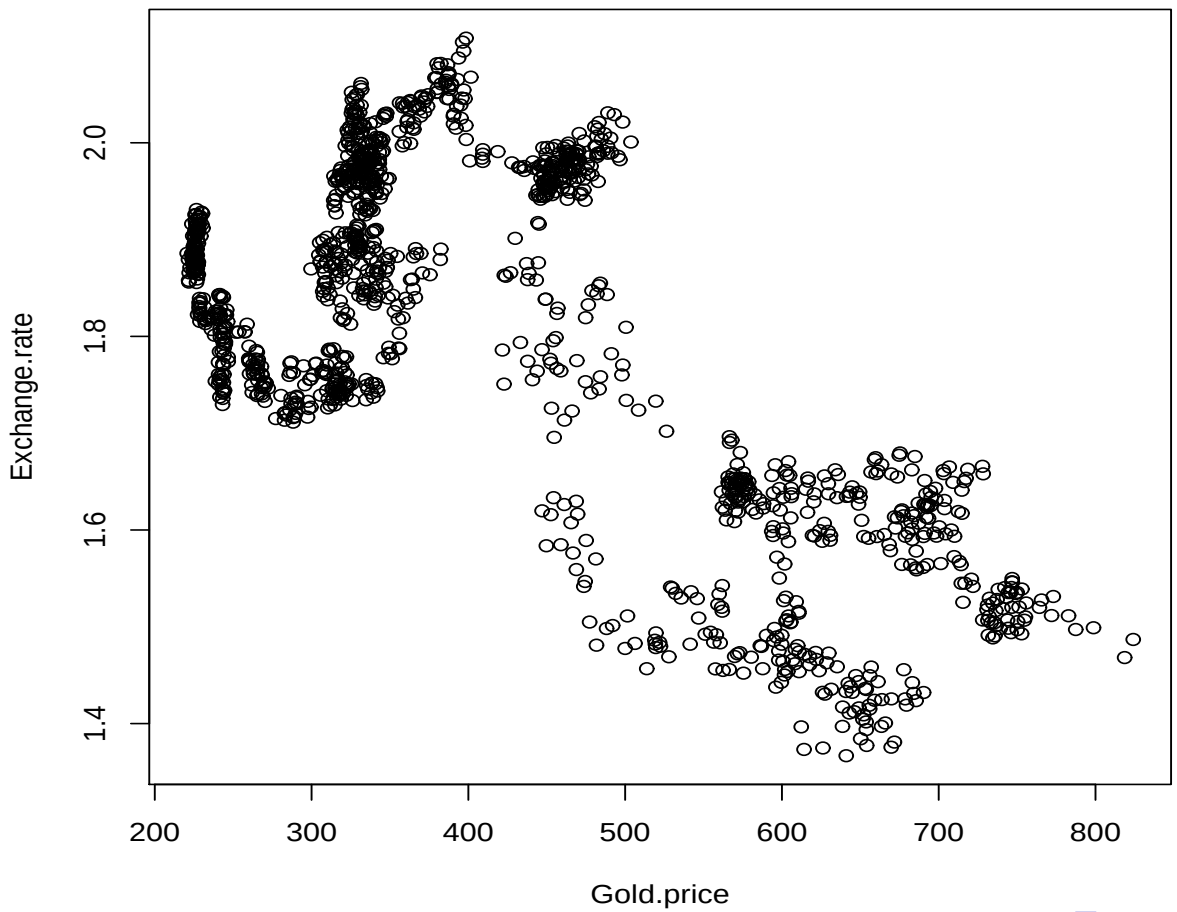
Unemployment rate vs. inflation - US (1981-2008)



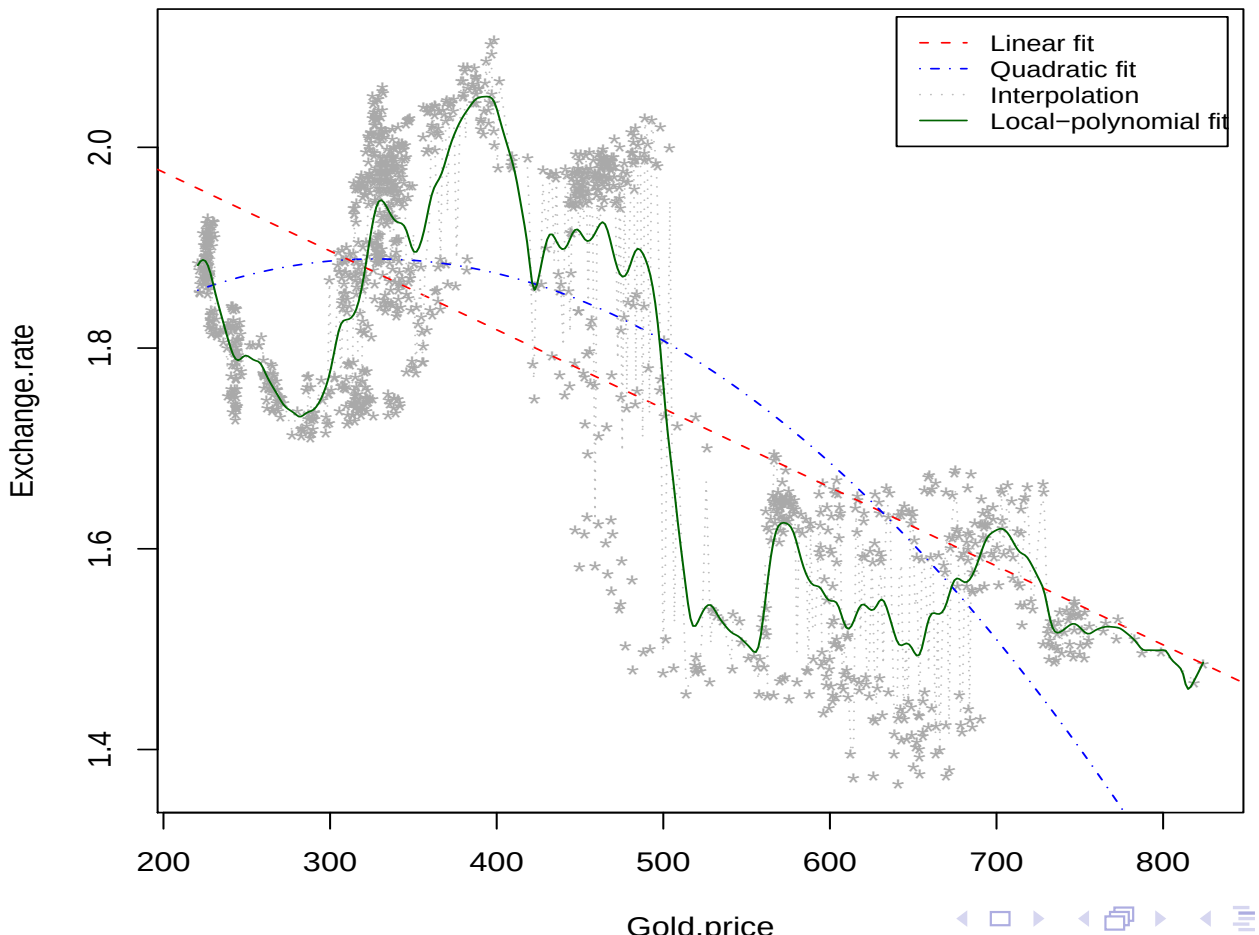
Fits for unemployment-inflation data



Gold prices and £/$ exchange rates Jan2005-May2010



Fits for gold-exchange.rates data



It is clear that none of the methods used to represent above data gives an adequate fit for the data, and this problem is expected to be more complicated in more complex data sets.

The basic problem of all fitting methods illustrated in the previous figures is that they use an asymmetric view on the variables, implying that each x (unemployment/gold-price) will be associated with exactly one estimated y (inflation/exchange-rate), which is clearly inadequate here.

Local Principal Curves

A group of non-traditional approaches that prove to be more efficient in the majority of these complex situations are methods based on what is called "Principal Curves".
Based upon the foundations of 'Principal Curves' and the idea of local modelling[**?**], and considering the situations of multi-dimensional data with symmetric components, Einbeck et al.[**?**] presented the idea of "Local Principal Curves", a flexible technique to model the complex data patterns arising in such situations.

If we have a d-dimensional data cloud $X_i \in \mathbb{R}^d, i = 1, ..., n$, where $X_i = (X_{i1}, ..., X_{id})$, we apply the LPC algorithm, to find a smooth curve which passes through the middle of the data cloud, as follows:

1. Choosing a suitable starting point $x_{(o)}$. Set $x = x_{(o)}$.
2. Calculating μ^x, the local center of mass around x.
3. Performing a principal component analysis locally at x.
4. Finding a new value for x by following the first local principal component starting at μ^x.
5. Repeating steps 2 to 4 until μ^x remains (approximately) constant.

There are several issues that are related to applying the LPC algorithm.

Starting point selection

There are two basic approaches to choose a starting point x_o from the data set, the first is to be chosen at random from the set of observations, and the second is to choose the point with the highest density as the initial starting point. Alternatively, a starting point out of the set of observations can be chosen, taking into consideration that it should lie within the multidimensional data range.

Bandwidth selection

To apply the LPC algorithm, a bandwidth matrix, H, needs to be determined. This matrix contains a set of bandwidths (or square bandwidths), $h_1, h_2, ..., h_d$, that corresponds to the number of variables (data dimensions), d. Each bandwidth determines the size of local neighborhood around each point in a certain direction. The optimal choice of the bandwidth matrix depends to a great extent upon the nature of the data set under consideration.

Kernel function

To compute the local mean around a point, a multidimensional kernel function is needed to produce some weighting around the chosen point. There are several types of kernel functions that are normally used. The LPC algorithm uses a multidimensional Gaussian kernel, $K_H(.)$. The one-dimensional Gaussian kernel usually takes the form: $k(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$, and K_H is obtained from this as a product kernel.

$$k_h(u) = \frac{1}{h}k(u), \quad K_H(u) = k_{h_1}(u) \times k_{h_2}(u) \times ... \times k_{h_d}(u) \qquad (1)$$

After choosing a starting point, x_o, and a d-dimensional kernel, $K_H(.)$, the LPC algorithm computes the local center of mass around the starting point then around each other point chosen in successive computations. The local center of mass around a point, x, is given by:

$$\mu^x = \frac{\sum_{i=1}^{n} K_H(X_i - x)X_i}{\sum_{i=1}^{n} K_H(X_i - x)} \tag{2}$$

The next step after calculating μ^x is to perform a principal component analysis around x. Denote by Σ^x the local covariance matrix of x, and let γ^x be the first eigenvector of Σ^x, we then obtain an updated value of x, $\mu^x + t_o\gamma^x$, where a suitable value of t_o is to be chosen. The LPC algorithm stops when μ^x remains constant, and the local principal curve is determined by the set of μ^x values.

Local principal curves can provide a useful tool in a wide range of highdimensional data studies, specially when aiming to visualize and analyze this data with significantly reduced number of dimensions. On one hand, merely looking at the data in a simple one dimensional curve/plot will provide a good idea about the basic pattern/shape of the data set, which is useful for doing some simple statistical inference.

On the other hand, the fact that the curve is parametrized over some parameter λ is of great importance if a link is recognized between the curve parametrization and some real variable(s) which are thought to be related to the data set under study. This link can be used in predicting real multidimensional data points with information about the link variable(s) only.

Phillips Curves

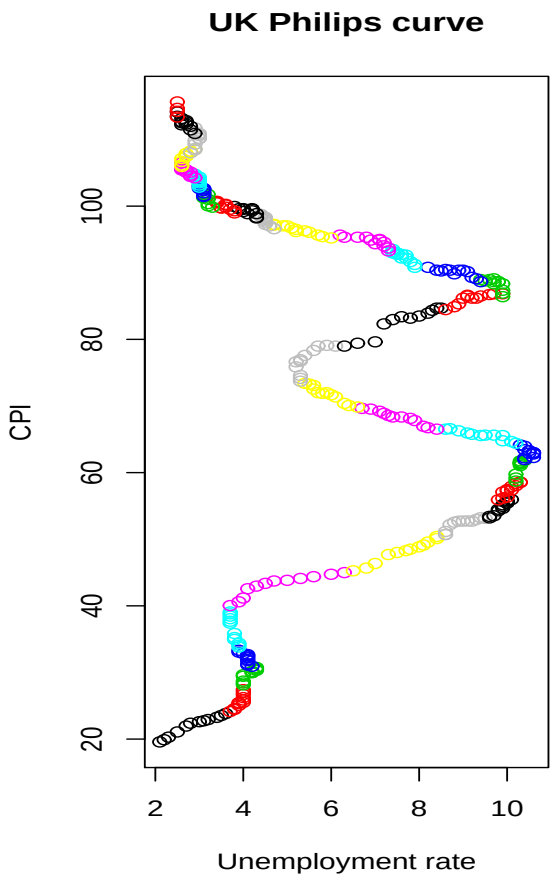
Philips curves is a famous term in economics. It refers to curves that study the relation between unemployment and the rate of inflation in an economy. It was named after the economist *Alban William Phillips* who was responsible for the first appearance of Phillips curves when he wrote a paper in 1958 in which he observed an inverse relationship between money wage changes and unemployment in the British economy over the period 1861-1957.

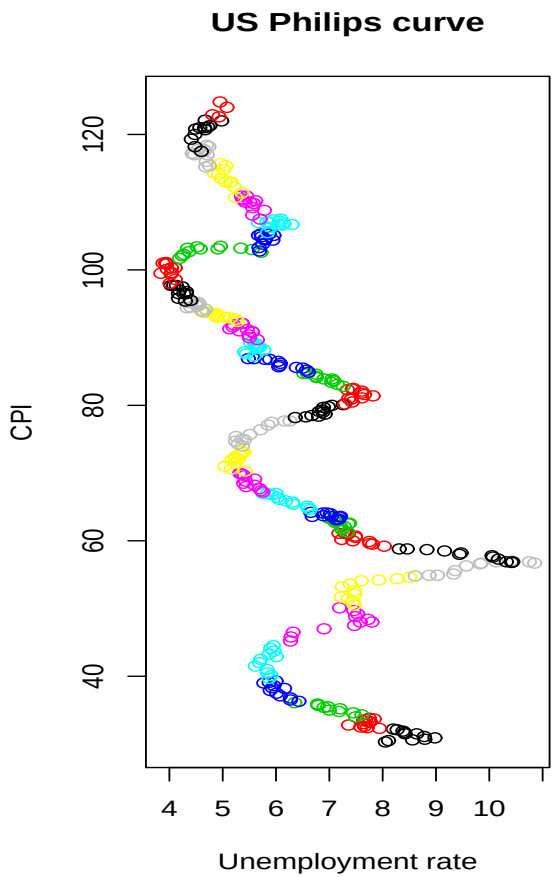
In our case, we will look at the Phillips curves of the relation between unemployment rate and consumer price index (CPI) as the most commonly used measure of inflation. The data sample used consists of:

- Consumer price indices (all goods and services) for both UK and US. (monthly data)
- Unemployment rates for both UK and US. (monthly data)

Our data sample covers the period from Jan. 1975 to April 2008.

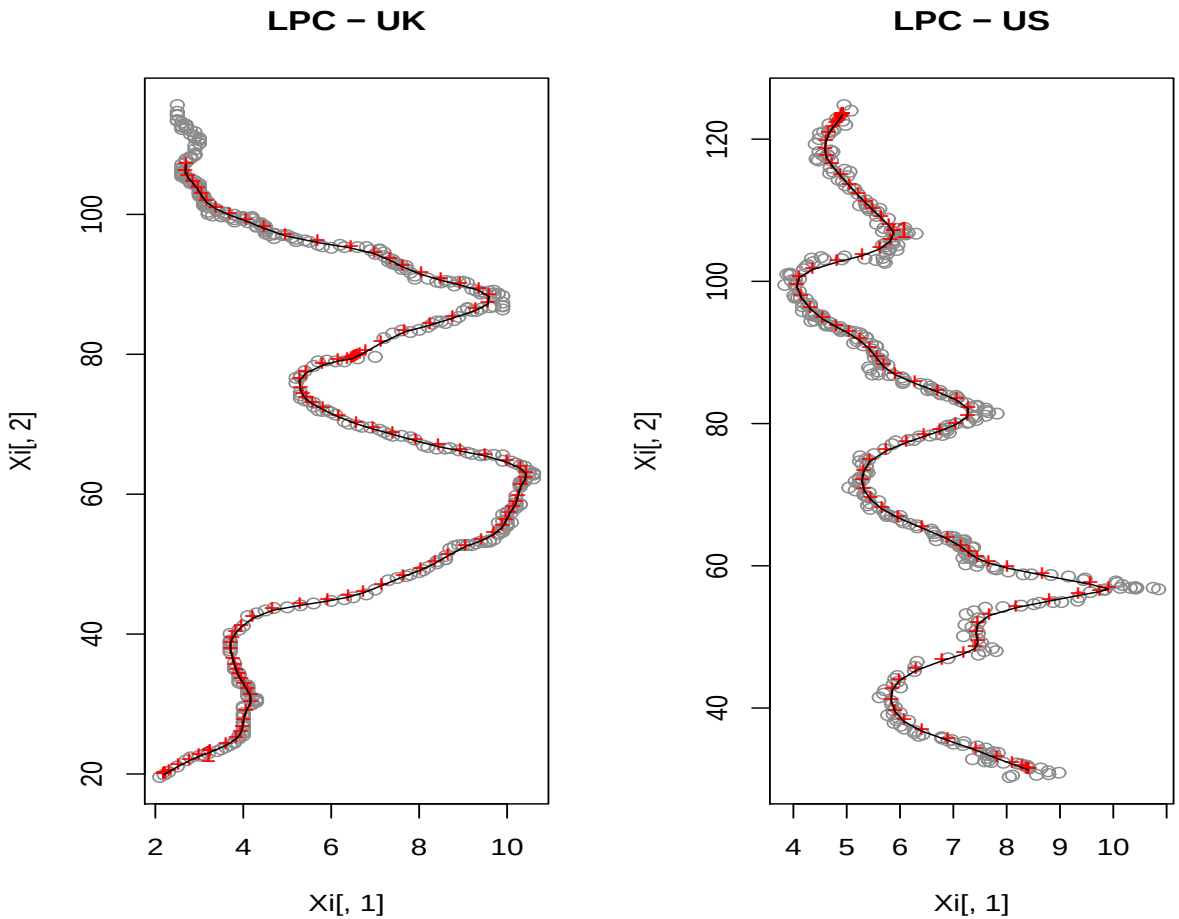
The first step before fitting a curve through our data is exploring the dat first by a simple scatter plot





When fitting the LPC through the data, the algorithm usually starts with a random starting point and an arbitrary choice of the bandwidth vector. We then check if the bandwidth selection and the chosen starting point is suitable for the data. After some trials, one should reach a reasonable combination of bandwidths that gives a 'good-looking' curve. The LPC for our data is plotted in the following figure

LPC for Phillips data

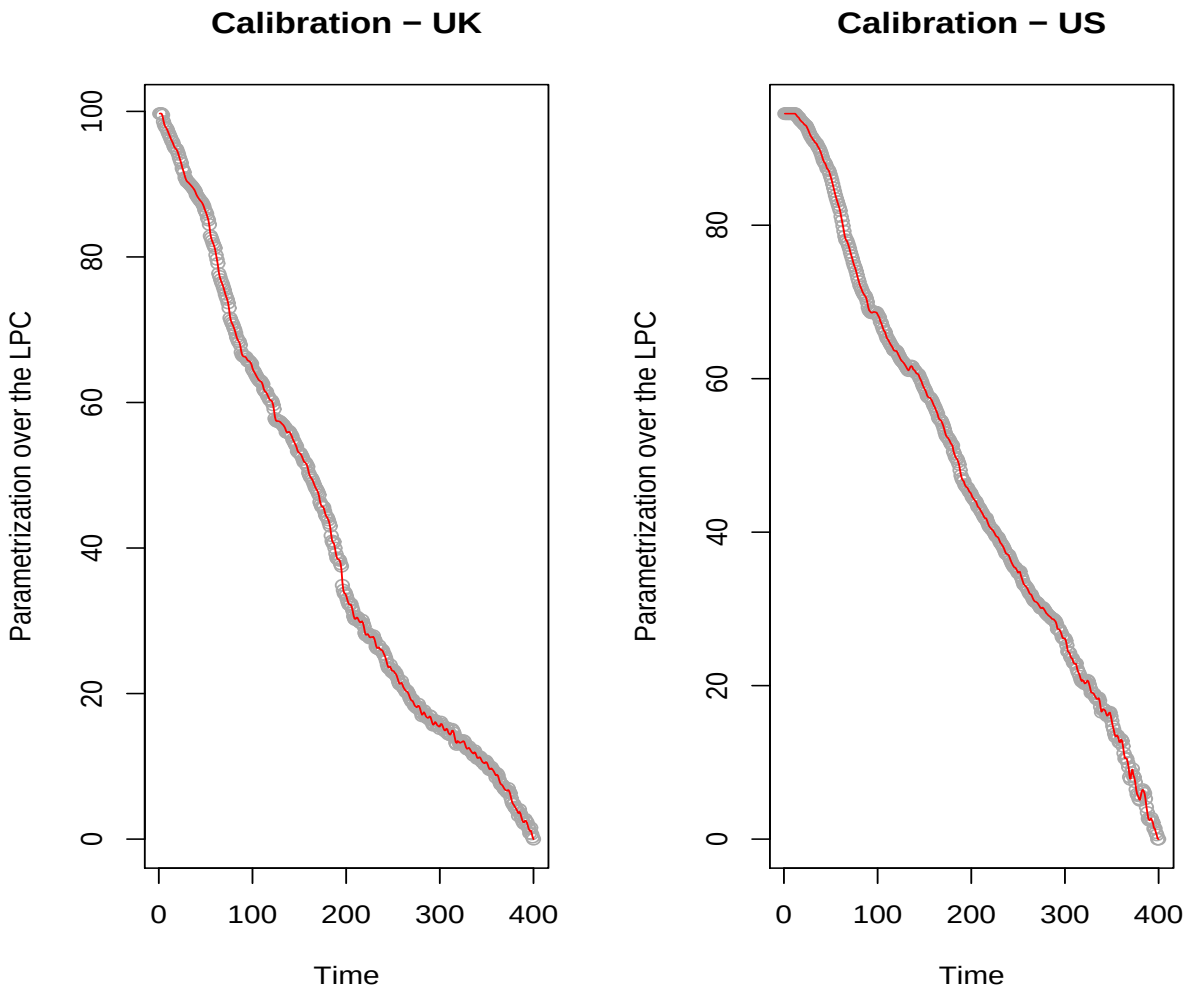


Note: The red "+" symbols correspond to the μ^x.

After fitting the LPC, one should go further and think about prediction. The key issue now is to look for some link between the curve (actually what is meant here is the parametrisation along the curve) and another real variable. In most econometric application, thinking about 'time' to play this link rule is a natural choice. Time variable has a major rule in economics as it is expected to have a role in analysing past and hence expected future behaviour of some economic indicators. It can also explain and clarify some facts about systematic cyclic behaviour.

The link between time and parametrization over the LPC is worth trying, since both of them increase in a monotone way. For our example, this link can be represented and referred to as a 'calibration curve'. If the link is looking reliable and seems to be unique for each value, we should then create some functional form or fit a spline to represent the relation between time and LPC parametrization. Prediction will be possible once a spline has been fit.

Calibration curves for Phillips data



The prediction process typically goes as follows:

(i) Use the calibration curve to predict curve parametrization given time.

(ii) Predict multidimensional data points simultaneously through the fitted LPC.

To check for this, the US data is used (as it looks more complicated). Assume that we want to do the prediction process given that time = 200. The following representation shows the prediction process from time=200 to the corresponding real 2-dimensional data point:

$$time = 200 \Rightarrow LPC\ parameter = 44.96$$

$$\Rightarrow unemployment\ rate = 6.8685, CPI = 79.2974$$

The corresponding real values for unemployment rate and CPI was 6.87 and 79.30 respectively. This shows that using calibration and fitting splines do produce very good predictions.

One last thing that needs more investigation regarding this data is the cyclic behaviour of data in both countries, UK and US. However, one can already notice that:

- Cycles happen in both countries.
- Cycles happen rapidly in the US compared to the UK (US cycle-time length is less).
- A cycle starts first in the US, then, after some time, a corresponding cycle starts in the UK.

The matter of detecting and measuring cycles within the context of LPCs and comparing two or more curves will be part of the expected future research.

Conclusions

Local principal curves provide a useful and flexible tool to represent multidimensional and complex structures. It is important to apply the algorithm with the optimal options for the specific data set of interest. LPCs could be useful in modelling many econometric data, especially when we can create a link between the parametrization of the curve and another external variable that is expected to be related to our data set. This link can be represented through a 'calibration' curve, and a good link should give a precise and unique calibration as possible. In this context, We should keep in mind the smoothness trade-off between the LPC and the calibration curve.

When the prediction works fine with regard to predicting existing data points, it is then possible to think about other issues regarding prediction, such as:

- Given a certain data point, which may or may not be a part of the original data cloud, and after projecting this point onto the curve, we could try to reconstruct the time at which this observation could have occurred.
- Using a boundary correction, if necessary, which is done by reducing the bandwidth consecutively when the curve starts to converge.
- Estimating future observations, at least in the short run. (i.e. extrapolating)


Einbeck, Jochen, Tutz, Gerhard, and Evers, Ludger.
Local principal curves.
Statistics and Computing, 15(4):301–313, October 2005.


Fan, J. and Gijbels, I.
Local Polynomial Modelling and Its Applications.
Chapman and Hall, 1995.