Constructing Economic Summary Indexes via Principal Curves

Mohammad Zayed[1,2] and Jochen Einbeck[1]

[1] Department of Mathematical Sciences, Durham University, Science Laboratories, South Rd., Durham, DH1 3LE, UK, *jochen.einbeck@dur.ac.uk*
[2] Applied Statistics and Insurance Department, Mansoura University, Mansoura, 35516, Egypt, *m.a.zayed@dur.ac.uk*

Abstract. Index number construction is an important and traditional subject in both the statistical and the economical sciences. A novel technique based on *localized principal components* to compose a single summary index from a collection of indexes is proposed, which is implemented by fitting a (local) principal curve to the multivariate index data. We exploit the ability of principal curves to extract robust low-dimensional 'features' (corresponding to the summary index) from high-dimensional data structures, yielding further useful analytic tools to study the behaviour and composition of the summary index over time.

Keywords: summary indexes, feature extraction, principal component analysis, smoothing

1 Introduction

A standard problem in economics is the question of how to construct a single (summary) index from a series of individual (sub-)indexes. For instance, the main measure of inflation for national macro-economic purposes is the Consumer Price Index (CPI), which covers essentially the monetary expenditures on all goods and services by all households of a certain economy (for instance, the UK). This index, say X_0, is usually computed from sub-indexes $X = (X_1, \ldots, X_p)'$ by weighted averaging of type

$$X_0 = w_1 X_1 + \ldots + w_p X_p = w'X \tag{1}$$

where $w = (w_1, \ldots, w_p)'$ is a set of weights relating to the composition of expenditure, which is allowed to vary over time, i.e. $w = w(t)$. Economists have taken substantial efforts to derive formulas which give appropriate or 'representative' weights for a certain economy. The actual process of averaging in (1) is rather crude from a statistical perspective. It is highly dependent on outlying (potentially erroneous) data, it is not able to deal with missing data, it does not allow an analysis of the relative contribution of the sub-indexes over time, and does not take into account the differing variability

(information) contained in the indexes at different time points (other than through the weights, perhaps). A potential alternative addressing these issues was already suggested by Tintner (1946) and Moser (1984) in the context of production and price indexes, and labour market indicators, respectively. They proposed to construct a linear summary index by finding that linear combination $\gamma' X$ of $X_1, \ldots, X_p$ with maximal variance $\mathrm{Var}(\gamma' X)$ among all unit vectors γ. The solution to this problem is found via principal component analysis (PCA), and is given by the first eigenvector γ of the covariance matrix $\Sigma = \mathrm{Cov}(X)$ of X. Assuming the existence of a 'price line' $X = aX_0 + \epsilon$, with $a \in \mathbb{R}^p$, Theil (1960) developed a variant of PCA to estimate a and γ simultaneously. Neither of these authors used any additional weighting, though (external) weights w could be easily accommodated by considering $X_w = (w_1 X_1, \ldots, w_p X_p)'$ instead of X itself.

If we have a set of variables, each can be represented as a mix of a systematic component and an error, applying PCA to these variables results in constructing a number of independent factors, usually less in number than data dimension, which capture most of the total variance in the data set. This is done by finding some linear function of the variables in the data set, which is least subject to errors. Principal components are of interest mainly in cases where the variables under consideration, the values of which formulate the data cloud, are considered to be symmetric, rather than one or more variable being generated from the remaining ones.

PCA-based approaches have not yet found widespread application in the context of economic index data. One reason for that is that PCA will find that line through the multidimensional cloud of indexes which gives *globally* the best fit in terms of squared orthogonal distances; in other words 'one line has to fit it all'. The approximation done this way may be good in some parts of the data cloud but poor in others. As a consequence, the loadings $\gamma = (\gamma_1, \ldots, \gamma_p)'$ will reflect the contribution of the subindexes $1, \ldots, p$ towards the overall index not equally well over the full data range — actually, the amount of information that individual indexes contribute towards the overall index may vary greatly; an example for this is provided later in this article. Hence, what would be needed is a tool to maximize the variance locally, providing at each point the best local approximation to the data cloud. This implies that we need to fit a sequence of localized principal components, rather than one global principal component. The statistical concept corresponding to this viewpoint is a (local) principal curve.

2 Principal curves

The concept of *principal curves* was introduced in the Statistics literature by Hastie and Stuetzle (1989) (hereafter: HS) as a nonparametric extension of PCA. A principal curve is descriptively defined as a smooth curve $f : \mathbb{R} \longrightarrow \mathbb{R}^p, \lambda \mapsto f(\lambda)$ that passes through the 'middle' of a $p-$variate data set, providing a nonlinear summary of the data. For HS curves, the notion of the

'middle' of the data cloud is implemented via the concept of self-consistency (Tarpey & Flury, 1996), meaning that each point on the curve is the average of all points that project there.

Principal curves have recently attracted interest particularly in the engineering literature (Ming-Ming et al., 2010) due to their ability to extract low-dimensional 'features' from high-dimensional data structures via the curve parametrization λ. In particular, for $X \in \mathbb{R}^p$, one defines the *projection index* as the parameter of the closest point on the curve to X, i.e.

$$\lambda_f(X) = \sup_{\lambda}\{\lambda : \|X - f(\lambda)\| = \inf_{\eta} \|X - f(\eta)\|\}. \tag{2}$$

In our context, the extracted feature $\lambda_f(X)$ would be corresponding to the summary index of X, as we will illustrate in the following section. However, we are not only interested in this overall index, but also in the local contributions of the individual sub-indexes, for which we need to determine loadings in terms of localized eigenvectors. The original algorithm by HS does not compute these, neither explicitly nor implicitly, so it is of limited use for our development. An alternative concept, which is explicitly based on localized PCA, is the local principal curve (LPC) algorithm (Einbeck et al., 2005):

Given n replicates of X, forming a $p-$variate data cloud $x_i, i = 1, ..., n$, where $x_i = (x_{i1}, ..., x_{ip})'$, a smooth curve which passes through the middle of the data cloud is found as follows:

1. Choose a suitable starting point $x_{(0)} \in \mathbb{R}^p$, either by hand or at random from the data cloud. Set $x = x_{(0)}$.
2. Calculate μ^x, a local mean around x.
3. Perform a principal component analysis locally at x, yielding a localized eigenvector γ^x.
4. Find a new value for x by following γ^x a predetermined step size, starting at μ^x.
5. Repeat steps 2 to 4 until μ^x remains (approximately) constant.

The local principal curve is determined by the series of the μ^x values. The actual localization in 2. and 3. is performed through multivariate kernel functions, see Einbeck et al. (2005) for details on these steps. After termination of the algorithm, the parametrization λ is calculated retrospectively through the Euclidean distances between neighboring μ^x, and interpolated between the μ^x through linear segments or cubic splines (Einbeck et al., 2009), yielding a fully parametrized one-dimensional curve $f(\lambda)$ through p-dimensional space, which passes precisely through all the local means μ^x. Note that the algorithm is robust to outlying data points due to the localized way of averaging.

There is one important adjustment that is useful to be made for index data: Normally, there is some reference date for which all sub-indexes take a baseline value, say 100, and also the overall index takes this value. Hence, also the parametrized principal curve has to reflect this property and this can be

realized through an *anchor:* This is a point of predetermined coordinates, say $x_{(0)} = (100, \ldots, 100)'$, and predetermined parameter value ('index') $\lambda = 100$, through which the curve *is forced to pass.* This is implemented by inverting steps 2 and 3 above, and recalculating λ by integrating over the arc length of the curve starting with the anchor point. Of course, this method is only feasible when the baseline time point is part of the time interval considered. We illustrate this algorithm, and its functionality as a 'feature extractor' for the summary index, in the subsequent section.

3 Analysis of CPI data

In the applied part of this work, two sets of consumer price indexes have been used, the first, as an introductory example, is a two dimensional set, and the second is a twelve dimensional set. All data are monthly UK data published through 'National Statistics Online' covering the period from January 1988 until December 2008. Both sets of indexes are complemented subsets of the same total summary index, which is the total CPI for 'All Items'. The indexes used for analysis are: (2005=100 for all indexes)

```
D7BT: CPI INDEX 00 : ALL ITEMS
D7BU: CPI INDEX 01 : FOOD AND NON-ALCOHOLIC BEVERAGES
D7BV: CPI INDEX 02 : ALCOHOLIC BEVERAGES,TOBACCO & NARCOTICS
D7BW: CPI INDEX 03 : CLOTHING AND FOOTWEAR
D7BX: CPI INDEX 04 : HOUSING, WATER AND FUELS
D7BY: CPI INDEX 05 : FURN, HH EQUIP & ROUTINE REPAIR OF HOUSE
D7BZ: CPI INDEX 06 : HEALTH
D7C2: CPI INDEX 07 : TRANSPORT
D7C3: CPI INDEX 08 : COMMUNICATION
D7C4: CPI INDEX 09 : RECREATION & CULTURE
D7C5: CPI INDEX 10 : EDUCATION
D7C6: CPI INDEX 11 : HOTELS, CAFES AND RESTAURANTS
D7C7: CPI INDEX 12 : MISCELLANEOUS GOODS AND SERVICES
D7F4: CPI INDEX: ALL GOODS
D7F5: CPI INDEX: ALL SERVICES
```

3.1 Index construction from two sub-indexes

We aim to reconstruct the overall index (`CPI INDEX 00: ALL ITEMS`) using two sub-indexes: the `CPI INDEX: ALL GOODS` and the `CPI INDEX: ALL SERVICES`. We use the modified LPC algorithm using an anchor at $x_{(0)} = (100, 100)'$ and $\lambda = 100$ (corresponding to the reference point January 2005), as outlined in Section 2. For simplicity, a constant weight $w = (547, 453)'$ for all years is used. Now, applying this adjusted LPC algorithm to fit a summary curve through the two weighted indexes, one obtains the fit produced in Figure 1. It seems to give a reasonable summary for the two-dimensional data set in the form of a one-dimensional curve.

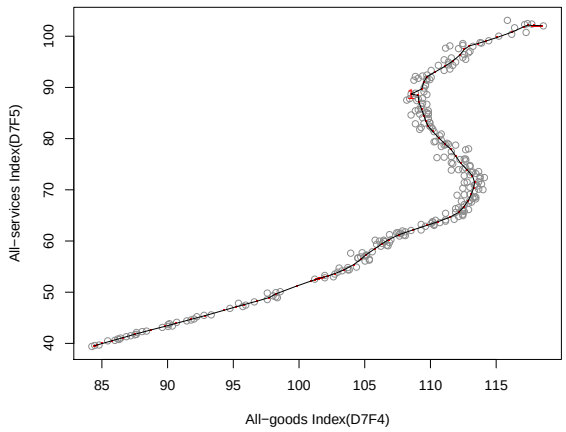

Fig. 1. LPC fit for 2D CPI data.

A first property of interest when using this statistical approach in CPI context could be: compared to the original overall index, how well is the resulting fit capturing the overall index behavior? Figure 2 compares how the projection indexes $\lambda_f(X)$ and the original `CPI INDEX OO` change over time. Figure 2 suggests that the statistically fitted overall index captures most movements in the true index, which is a desirable situation. Also, it can be seen that the fitted index looks smoother than the original index, due to the underlying smoothing properties implied by using the LPC algorithm.

The other useful informative tool accompanying the use of LPCs is related to the total variance explained by the curve and how each variable (sub-index) contributes to the fitted overall index. This is statistically measured through 'loadings', i.e. the entries of the (local) eigenvectors. At every point on the curve, the sum of squared loadings of the first eigenvector should be equal to one. This 'unity' property of eigenvectors provides a good tool to indicate how the sub-indexes influence the fitted overall index at each point (time). Figure 3 shows the cumulative squared loadings of first eigenvectors for our example. Useful interpretations could be derived from such a figure, for instance, around the fitted curve's parameter values of 80 and 100, the second sub-index has a dominating effect on the fitted overall index.

3.2 Index construction from twelve sub-indexes

Adopting the same techniques used in the previous example, the LPC algorithm was applied to fit the overall consumer price index from the twelve sub-indexes (`INDEX 01, INDEX 02, ..., INDEX 12`). Main indicators from the resulting fit are shown in Figure 4. We can study the index behaviour and the dominating underlying factors affecting it over time.

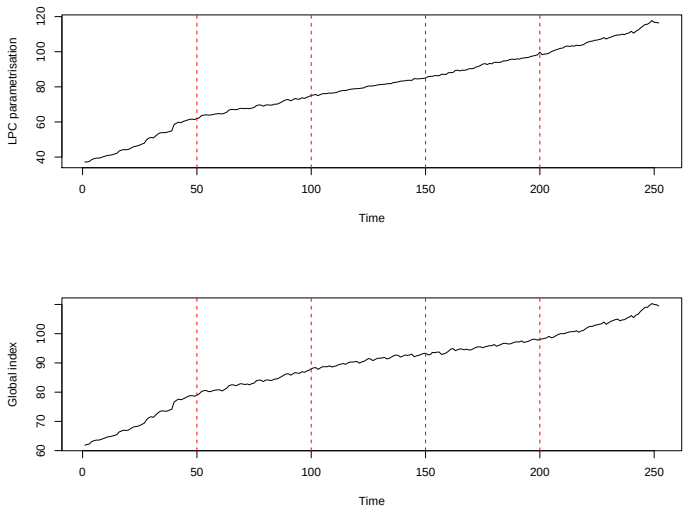

Fig. 2. LPC-based (top) and average-based (bottom) CPI behavior over time.

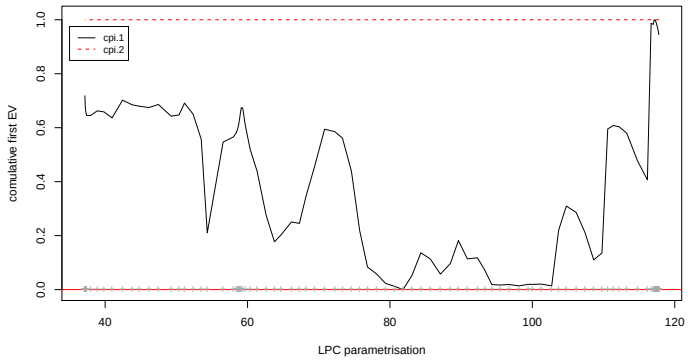

Fig. 3. Cumulative squared loadings of first eigenvectors - 2D fit.

The bottom part of Figure 4 allows to assess the contributions of the 12 sub-indexes over time. For example, it can be seen that the third index has the largest effect on the fitted overall index around the LPC parameter value 50 (which corresponds to some time point near 150), and the same can be said about index four around parameter values of 120 and 178 (times: 19 and 249) and that the first index alone contributes by 30% in the fitted index around parameter value of 169 (time near 246), and so on. All such interpretations can have useful meanings in the econometrics context.

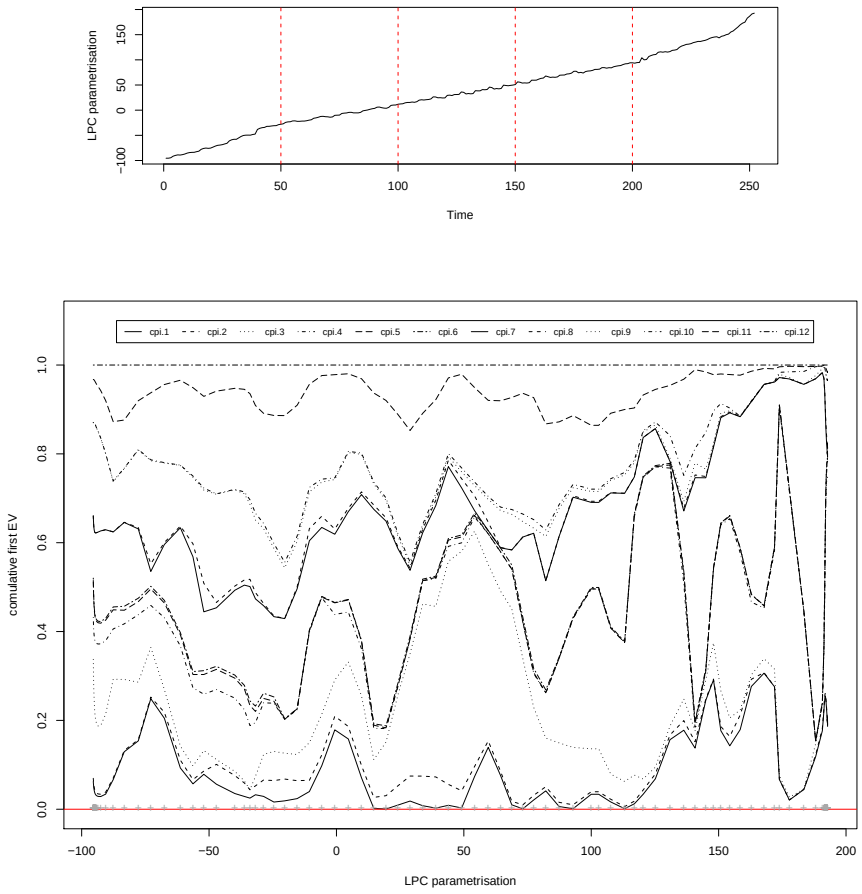

Fig. 4. A 12-D example. Top: reconstructed summary index (LPC parametrisation over time); bottom: cumulative squared loadings (first eigenvector) over time.

One remaining important feature of the proposed technique is the ability to predict missing data points at any given time (discrete or continuous) within the data range. This is achieved, technically, through 'calibration' of time and the LPC parametrisation (by plotting them against each other and using a nonparametric smoother to find the functional relationship). Having done this, if we assume that we want to predict the data point that corresponds to, say, time = 220.5, we plug this value in the calibrated object which gives a parameter value of 126.3425, then we get the corresponding estimated 12-dimensional weighted point on the fitted curve, and applying a simple adverse-weight formula to each index ($cpi = weighted.cpi * average.weight/current.index.weight$), we get the real time

estimated sub-indexes' values (101.78, 102.53, 96.37, 108.04, 99.48, 102.38, 102.93, 100.1, 98.87, 104.82, 102.84, 103.38). This could be useful in handling missing data as well as predicting any assumed in-between data points (for instance, holidays).

4 Conclusion

The work presented in this paper is merely a statistically-based approach to fit and analyse main economic indexes. The computed index using the LPC algorithm has the ability to capture the basic trend of the original corresponding index over time. Being based upon principal component analysis, it allows to detect the influence of all variables (sub-indexes) on the fitted index at all points (time), and would furthermore allow to assess the degree of 'local linearity' of the index, in terms of total local variance explained, at each point in time by looking at the localized first eigenvalues. The main novel feature of the proposed technique is that it is nonlinear and even non-parametric, while the traditional PCA-based methods are linear, which may be of limited accuracy in particular if the time range considered is quite large.

It should be noted that the proposed technique, just as PCA itself and the modified version by Theil (1960), is an 'ex-post' algorithm, i.e. one needs to have the full data available in order to reconstruct the indexes retrospectively. However, unlike other principal curve algorithms, the LPC methodology would in principle allow for an updating algorithm, which would enable to extend the previously fitted curve and the associated statistics once new data have come in. This is a matter of future research.

References

EINBECK, J., TUTZ, G. and EVERS, L. (2005): Local principal curves. *Statistics and Computing 15 (4), 301-313.*

EINBECK, J., EVERS, L., and HINCHLIFF, K. (2009): Data compression and regression based on local principal curves. In: Fink et al. (Eds): Advances in Data Analysis, Data Handling and Business Intelligence, Springer, Heidelberg, 701–712.

HASTIE, T. and STUETZLE, W. (1989): Principal curves. *Journal of the American Statistical Association 84 (406), 502-516.*

MING-MING, Y., JIAN, L., CHUAN-CAI, L. and JING-YU, Y. (2010): Similarity preserving principal curve: an optimal one-dimensional feature extractor for data representation. *IEEE Transactions on Neural Networks*, to appear.

MOSER, J. W. (1984): A principal component analysis of labor market indicators. *Eastern Economic Journal X (3), 243-257.*

TARPEY, T. and FLURY, B. (1996): Self-consistency: a fundamental concept in statistics. *Statistical Science 11 (3), 229-243.*

THEIL, H. (1960): Best linear index numbers of prices and quantities. *Econometrica 28 (2), 464-480.*

TINTNER, G. (1946): Some applications of multivariate analysis to economic data. *Journal of the American Statistical Association 41 (236), 472-500.*