CS229 Final project - Finding economic groupings over time

Stephen Reid

1 Introduction

People want to divide countries into economic groupings, whether it be along historic lines (First World, Third World) or because of strong trade links (NAFTA, MERCOSUR) or because of some whim of a momentarily famous investment banker trying to invent a catchphrase (BRICS). Groupings are created, adopted and some even persist in the popular consciousness. Many of these groupings are artificial and become obsolete as time passes and economies change.

The question arises whether one could find economic groupings based on how economies actually behave. Furthermore, how do these groupings change over time? Does the number of groups change over time? Or is there some fixed number of groups with countries moving in and out of them? My project has two components. The first is to fit a Gaussian mixture model (with tweaks) to data vectors of economic variables (like savings rates, consumption spending, export share) pertaining to countries at different times. The aim is to find a set of (possibly) time-varying groups to which the countries belong and then to see how the countries move between them. Tweaks are detailed in the next section.

The second component relates to visualisation of the groupings. A 2-dimensional plot (with animation over time) is desired. The canonical coordinate directions encountered in LDA, obtained by maximising the Rayleigh coefficients encountered in Hastie et al. (2001) will be employed for this purpose.

2 Model I: Gaussian mixtures with prior on cluster centres

Suppose we have data of the form x_{it}, $i = 1, 2, ..., n$ and $t = 1, 2, ..., T$, where x_{it} is a p-vector containing the values of p economic variables relating to country i at time t. The aim is, given only these vectors, to find sensible and objective clusters in the data over time. While the use of K-means (or K-medoids) clustering seems reasonable, I modelled the data using a Gaussian mixture model. Once the means of the mixture components are estimated, we have our K cluster centres. Furthermore, the flexibility allowed by this modelling paradigm allows for intuitive generalisations of a simple clustering model.

Specifically then, I assume that x_{it} comes from a mixture of multivariate Gaussian distributions. Let $z_{itk} = 1$ if country i belongs to cluster k at time t. Then I assume

$$x_{it}|z_{itk} = 1 \sim N_p(\mu_{tk}, \Sigma)$$

with $P(z_{itk} = 1) = \alpha_k$.

The z_{itk} are latent variables and we use the EM algorithm to maximise the loglikelihood. The EM criterion to maximise in the M-step becomes:

$$-\frac{nT}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{z}_{itk}(x_{it}-\mu_{tk})^T\Sigma^{-1}(x_{it}-\mu_{tk})$$

where $\hat{z}_{itk} = \frac{\alpha_k\phi_p\left(\Sigma^{-\frac{1}{2}}(x_{it}-\mu_{tk})\right)}{\sum_{j=1}^{K}\alpha_j\phi_p\left(\Sigma^{-\frac{1}{2}}(x_{it}-\mu_{tj})\right)}$ is computed in the E-step. $\phi_p(.)$ is the p-variate standard Gaussian density function.

There are issues with this setup. In particular, there is little to gain by having different, independently estimated cluster centres μ_{tk} for every time period. Not only do we estimate many parameters (leading to high variance of the estimates), but successive cluster centres will not necessarily relate to each other. So the decision is made to smooth the cluster centres. In particular, I maximise the criterion:

$$-\frac{nT}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{z}_{itk}(x_{it}-\mu_{tk})^T\Sigma^{-1}(x_{it}-\mu_{tk}) - \lambda\sum_{t=1}^{T}\sum_{k=1}^{K}||\mu_{t,k}-\mu_{t-1,k}||^2$$

Notice that when $\lambda = 0$, we recover the original problem. When $\lambda \to \infty$, we force $\mu_{t,k} = \mu_{t-1,k}$ and our cluster centres become μ_k and are constant over time. Any intermediate λ allows the cluster centres to change over time, but encourages centres at successive time steps to be close together.

There are at least two intuitive interpretations of this problem:

1. We are maximising a penalised loglikelihood, with a penalty directly geared toward ensuring that successive centres are close together.

2. We are computing a MAP estimate where our prior on the cluster centres is defined by the conditional distributions $\mu_{t,k}|\mu_{t-1,k} \sim N_p(\mu_{t-1,k}, \frac{2}{\lambda} I_p)$.

Considerations around the specification of this model include:

- The covariance matrix Σ. Note that this is constant over clusters. Having one for each cluster requires the estimation of too many parameters and this is prohibitive. Estimating a full Σ allows for elliptically shaped clusters in p-space. One could even make the further assumption that $\Sigma = I_p$ (after standardising the data), reducing the parameter burden even further. I tried both.

- The regularisation parameter λ. Currently this parameter is selected, rather unscientifically, to ensure a smooth progression of the cluster centres over time.

3 Model II: Hidden Markov Model

Although the prior on the cluster centres imposed in the previous section speaks to the serial correlation in our constituent time series, a more direct way of modelling time series data is via a Hidden Markov Model (HMM). I do not have space here to describe the likelihood, optimisation procedures and other details, but there is a rather accessible treatment in Zucchini & MacDonald (2009).

Suffice it to say that the HMM is a mixure of Gaussians where $x_{it}|z_{itk} = 1 \sim N(\mu_{tk}, \Sigma)$ as before, but now we model the latent variables as a first order Markov Chain: $P(z_{itk} = 1|z_{i,,t-1,j} = 1) = \gamma_{jk}$. This induces the serial correlation in our modelled time series. There are now $K(K-1)$ additional parameters to be estimated (a $K \times K$ matrix Γ of transition probabilities with row sums equal to 1) and the number of parameters increases quite rapidly in the number of clusters.

Once its parameters are estimated, one can use the Viterbi algorithm to deduce the most likely sequence of latent states (cluster membership) for each of the data items. Again, I shall spare the details, which can be found in Zucchini & MacDonald (2009).

4 Visualisation

Once we have estimated the cluster centres $\hat{\mu}_{tk}$, we can turn to visualising the data. The problem here is that $p > 2$, making effective visualisation difficult, especially if we want some animated representation of the clusters over time. We require a reasonable method for projecting our data into lower dimensional subspaces. Principal component analysis is one possibility, but that only considers overall variation in the data. We have no guarantee that the cluster centres would spaced far apart in the lower dimensional subspace and visualisation would suffer.

A different method for selecting projection directions successively maximises (with unit length, mutually orthogonal vectors) the Rayleigh coefficient:

$$R(u) = \frac{u^T B u}{u^T W u}$$

where u is a unit norm vector. The matrices B and W are the between cluster and within cluster covariance matrices obtained from the decomposition (ignoring multiplicative constants):

$$\begin{aligned} T &= \sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x})(x_{it} - \bar{x})^T \\ &= \sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{z}_{itk}(x_{it} - \hat{\mu}_{tk})(x_{it} - \hat{\mu}_{tk})^T + \sum_{t=1}^{T}\sum_{k=1}^{K}\left(\sum_{i=1}^{n}\hat{z}_{itk}\right)(\hat{\mu}_{tk} - \bar{x})(\hat{\mu}_{tk} - \bar{x})^T \\ &= W + B \end{aligned}$$

A direction u chosen in this way simultaneously maximises the distance between cluster centres, while minimising the distance within each cluster between the points in that cluster and its centre. Clearly ideal for visualisation in our problem. The solution to this generalised eigenvalue problem is that the successive maxima are given by the eigenvalues of the matrix $W^{-1}B$ and the corresponding directions, the associated (unit vector) eigenvectors. As with PCA we can project our data onto (say) the first two of these directions to obtain a 2-dimensional plot. Plotting each year in isolation, then redrawing the next year after a delay creates an animated representation of the clustering.

5 Data and model fits

Versions of the models above were fit the to "Countries" dataset. This is a dataset comprising of economic data from $n = 23$ countries (see tables below for list). For each country, annual data is gleaned for the years 1985 - 2008 (so $T = 24$ years) on $p = 10$ economic time series, which include: *Real consumption share of GDP, Real fixed investment share of GDP, Real government consumption spending share of GDP, Openness (Exports + Imports to GDP), Annual percentage change in foreign exchange reserves, Unemployment rate, Inflation (measured by the percentage change in the GDP deflator), Financial account as percentage of GDP (both in US$), Current account as percentage of GDP (both in US$), Growth in real GDP.*

These variables were chosen because they are relatively easily obtained (from the IMF IFS database) and together provide a (very) general description of the economy of interest. A clustering based on these variables is not entirely devoid of content. Note that for policy formulation purposes, more variables should be obtained to reflect deeper attributes of the economies in question. However, data acquisition here is time intensive and it is difficult to collate data across countries for the majority of available time series.

The output from 3 models is presented below. Note that for each, I assumed that the covariance matrix $\Sigma = I_p$, because the dataset is not very large and I wanted to focus attention on the estimation of the cluster centres and transition probabilities. Also, the data was mean centered and standardised, so a unit variance assumption is probably not too stringent. These 3 models are: **Gaussian mixtures with constant cluster centres ($\lambda = \infty$), Gaussian mixtures with time varying clusters ($\lambda = 100$) and a first order HMM with constant cluster centres.**

Tables 1 - 3 show the estimated cluster membership (numbering 8, 8, and 6) for these 3 models respectively. Note that the cluster numbers themselves have no meaning other than being convenient labels. The same country may have different cluster labels in different model fits, but this does not mean that it belongs to qualitatively different cluster.

Number of clusters is selected using the BIC. There are some details about how to count the number of parameters in the smoothed cluster centre model (see appendix). Conclusions are similar for the three model fits. Additional clusters in the Gaussian mixture models are "crisis" clusters - they are occupied temporarily by countries experiencing some extreme economic events. The HMM has only one such crisis cluster, whereas the Gaussian mixture models have three apiece. This allows one to characterise the type of crisis experienced by the country more fully.

There seems to be a fairly stable cluster containing "Western" or "First World" economies like France, Canada, Australia, Spain and the United Kingdom. It is somewhat surprising that some "Middle Income" economies, like Poland, South Africa and Brazil also spend much time in this cluster. Notice that Brazil bounces around quite dramatically before settling down. The economic instability in this country during the 80s and 90s is well documented. Germany and Austria are interesting exceptions in this cluster. They spend most of their time there, but end up in a "Open, Industrialised" cluster with the Netherlands and Belgium (who never leave this cluster). "Open" economies are those characterised by large proportions of exports and imports relative to the overall size of the economies. This observation also accords with the conventional wisdom these economies tend to be major export hubs.

Furthermore, one can recognise a "Developing, Third World" cluster with members Egypt, Turkey, Argentina (also tumultuous, like Brazil - a South American curse), India, Mexico and Indonesia. Finally, one notices an "Asian Tiger" cluster (characterised by rapid growth, high savings and investment rates) containing, at various times, China, South Korea, Japan, Thailand.

It is interesting to consider the Asian economies and their progression, especially after the Asian Crisis of 1997-1998. Korea moves briefly to a crisis cluster in 1998, but soon returns to the Tiger cluster. Indonesia initially spends time in the Tiger cluster, experiences the crisis and then seems to "drop down" to the developing economy cluster. It seems to have struggled to cope with the fallout from the crisis. In contrast, Thailand starts off as a developing economy, but then seems to graduate to the open, industrialised cluster with Belgium and Netherlands. This reflects the differing responses these economies had to the crisis and the extent of their subsequent success.

Visualisation techniques described in Section 4 were used to create animations of these tables, colour coding different clusters. These are posted on YouTube on my channel. Follow `http://www.youtube.com/watch?v=9_dmDFZLJAA` for Table 1 and navigate to my channel from there to view the other two. Cluster centres are marked by "X" and country abbreviations label the countries. Note that some distance information is lost when projecting onto lower dimensions, so countries coloured according to one cluster may appear closer to a cluster centre of a different colour.

6 Discussion

Unsupervised clustering algorithms were applied to an economic dataset with time series data. The challenge here is the modelling of cluster centres over time so that they properly reflect the serial correlation encountered in most time series data. Smoothing penalties and HMMs were used to model the serial correlation. Superficially interesting clusters were discovered in the data. Obvious extensions of this would be the automatic selection of the regularisation parameter in the Gaussian mixture model and an objective criterion to compare different model types, perhaps aiding the selection of a single "best" clustering model.

Country	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08
Australia	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Canada	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
France	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Spain	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
U. Kingdom	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Poland	5	5	6	5	5	5	5	7	5	5	5	5	5	5	5	5	5	5	5	5	5
S. Africa	5	5	5	5	5	5	5	5	5	7	5	5	5	5	5	5	5	5	5	5	5
U. States	5	5	5	5	5	5	5	5	5	5	8	8	8	5	5	5	5	5	5	5	5
Germany	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	1	1	1
Austria	5	5	5	5	5	5	5	5	5	5	5	5	5	5	1	5	1	1	1	1	1
Japan	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Brazil	3	3	2	5	3	2	2	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Mexico	8	8	8	8	8	8	8	6	8	8	8	8	8	8	8	8	8	8	8	8	8
Egypt	8	8	8	6	6	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Turkey	8	8	8	8	8	8	8	8	8	8	8	5	8	6	8	8	8	8	8	8	5
India	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	4	4	4	4
Indonesia	8	4	4	4	4	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
Argentina	3	2	2	8	8	8	8	5	8	8	8	5	5	5	6	8	8	8	8	8	8
China	4	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Korea	7	4	4	4	4	4	4	4	4	4	6	4	4	4	4	4	4	4	4	4	1
Thailand	4	4	4	4	4	4	4	4	4	1	6	1	1	1	1	1	1	4	1	1	1
Belgium	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Netherlands	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 1: *Estimated cluster membership of the country dataset over the* 8 *clusters selected by the Gaussian mixture model with constant cluster centres. Output shown only for* 1988 - 2008. *Table is arranged by cluster in which countries spent most of their time, ranked within each cluster in descending order of time spent there. Solid lines separate clusters, dashed lines special cases within each cluster.*

Appendix - Number of parameters in BIC

The BIC criterion is used to determine the number of clusters in each model. The criterion, for number of clusters K, has the form:

$$-2\text{loglik}(K) + \log(npT) * \text{number\_of\_parameters\_in\_model}$$

Clearly, we need to count the number of parameters in the fit. This is easy for the constant cluster centre Gaussian mixture and the HMM. However, the regularisation in the smoothed cluster centre Gaussian mixture complicates matters. I used the rule-of-thumb proposed by Hastie et al. (2001), Chapter 3, modified slightly for this problem. Since we use the identity covariance matrix, the problem separates into p regularisations where:

$$\hat{\mu}_{jk} = (D_k + \lambda M)^{-1} v_k$$

where $\hat{\mu}_{jk}$ is the T-vector of cluster centre estimates of the j^{th} variable in cluster k, D_k a $T \times T$ diagonal matrix with entries $n_{tk} = \sum_{i=1}^{n} z_{itk}$, v_k a T-vector with entries $\sum_{i=1}^{n} z_{itk} x_{itj}$ and M the tridiagonal matrix associated with the quadratic form in the penalty. Using the formula for the eigenvalues of tridiagonal matrices, a modified approximate parameter count is:

$$\begin{aligned} \text{df}(\lambda) &= \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{K} \text{trace}\left(\text{cov}(x_{ij}, \hat{\mu}_{jk})\right) \\ &= \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{n_{tk}}{n_{tk} + 2\lambda\left(1 + \cos\left(\frac{t\pi}{T+1}\right)\right)} \end{aligned}$$

References

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer Verlag, New York.

Zucchini, W. & MacDonald, I. (2009), *Hidden Markov Models for Time Series; An Introduction Using R*, Chapman and Hall, CRC, Boca Raton.

Country	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08
Australia	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
France	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Spain	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
U. Kingdom	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
S. Africa	3	3	3	3	3	3	3	3	3	7	3	3	3	3	3	3	3	3	3	3	3
U. States	3	3	3	3	3	3	3	3	3	2	2	2	3	3	3	3	3	3	3	3	3
Poland	5	2	5	3	3	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3
Canada	3	3	3	3	3	3	3	3	3	3	3	3	4	3	3	3	3	2	2	2	2
Germany	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4
Austria	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4
Brazil	2	5	8	3	7	8	8	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Thailand	6	1	1	1	1	1	1	1	1	5	5	4	4	4	4	4	2	1	2	4	4
India	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	6	2
Mexico	2	2	2	2	2	2	2	7	2	2	2	2	2	2	2	2	2	2	2	2	2
Egypt	2	2	5	5	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Indonesia	2	6	6	2	6	2	2	2	2	2	5	2	2	2	2	2	2	2	2	6	2
Turkey	2	2	2	2	2	2	2	2	2	2	2	2	2	5	2	2	2	3	3	3	3
Argentina	7	8	8	2	2	2	2	3	2	2	2	3	3	3	5	2	2	2	2	2	2
Japan	6	6	6	6	2	2	2	2	2	2	3	2	2	3	3	2	2	2	2	2	2
China	6	2	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
Korea	7	6	6	6	6	6	6	6	6	6	7	6	6	2	2	2	2	2	2	2	4
Belgium	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Netherlands	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

Table 2: *Estimated cluster membership of the country dataset over the* 8 *clusters selected by the Gaussian mixture model with smoothed time-varying cluster centres. Output shown only for* 1988 - 2008. *Table is arranged by cluster in which countries spent most of their time, ranked within each cluster in descending order of time spent there. Solid lines separate clusters, dashed lines special cases within each cluster.*

Country	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08
Australia	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Canada	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
France	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Spain	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
U. Kingdom	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
U. States	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Poland	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
S. Africa	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2
Germany	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	5	5	5
Austria	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	5	5	5	5	5
Brazil	3	6	6	2	1	6	6	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Turkey	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Egypt	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Mexico	3	3	3	3	3	3	3	1	3	3	3	3	3	3	3	3	3	3	3	3	3
Argentina	3	6	6	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
India	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4
Indonesia	3	4	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3	3	3	3
China	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
Korea	1	4	4	4	4	4	4	4	4	4	1	4	4	4	4	4	4	4	4	4	4
Japan	4	4	4	4	4	4	4	4	4	4	4	4	2	2	2	2	2	2	2	2	2
Thailand	4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5
Belgium	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
Netherlands	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

Table 3: *Estimated cluster membership of the country dataset over the* 6 *clusters selected by the first order HMM. Output shown only for* 1988 - 2008. *Table is arranged by cluster in which countries spent most of their time, ranked within each cluster in descending order of time spent there. Solid lines separate clusters, dashed lines special cases within each cluster.*