
# Department of Economics Discussion Paper Series

Modelling Non-stationary 'Big Data'

Jennifer L. Castle, Jurgen A. Doornik and David F. Hendry

Number 905
April 9, 2020

Department of Economics
Manor Road Building, Manor Road
Oxford, OX1 3UQ

ISSN 1471-0498

# Modelling Non-stationary 'Big Data'

Jennifer L. Castle[†], Jurgen A. Doornik[‡] and David F. Hendry[‡*]

[†] Magdalen College and Climate Econometrics, University of Oxford
[‡] Nuffield College and Climate Econometrics, University of Oxford

April 9, 2020

**Abstract**

Seeking substantive relationships among vast numbers of spurious connections when modelling Big Data requires an appropriate approach. Big Data are useful if they can increase the probability that the data generation process is nested in the postulated model, increase the power of specification and mis-specification tests, and yet do not raise the chances of adventitious significance. Simply choosing the best-fitting equation or trying hundreds of empirical fits and selecting a preferred one–perhaps contradicted by others that go unreported–is not going to lead to a useful outcome. Wide-sense non-stationarity (including both distributional shifts and integrated data) must be taken into account. The paper discusses the use of principal components analysis to identify cointegrating relations as a route to handling that aspect of non-stationary big data, along with saturation to handle distributional shifts, and models the monthly UK unemployment rate, using both macroeconomic and Google Trends data, searching over 3000 explanatory variables and yet identifying a parsimonious, well-specified and theoretically interpretable model specification.

*JEL classifications:* C51, Q54.
*Keywords:* Cointegration; Big Data; Model Selection; Outliers; Indicator Saturation; *Autometrics*.

## 1 Introduction

Interest in methods for analysing big data has proliferated in recent years, with economists and policy makers viewing big data as having the potential to improve the forecast accuracy of economic phenomena, see, e.g. Varian (2014) and Swanson and Xiong (2018). Doornik and Hendry (2015) address modelling time series in economics using automatic model selection for 'fat' data (defined by many variables, $N$, but fewer observations, $T$ where $N \gg T$). Much of the focus on methodology for big data

*Financial support from the Robertson Foundation (award 9907422), and the Institute for New Economic Thinking (grant 20029822) is gratefully acknowledged, as are helpful comments from Luke P. Jackson, Andrew B. Martinez, Felix Pretis, Giovanni Urga, Sébastien Laurent, Heather Anderson, and participants at the 2019 FUNCAS workshop on 40 years of cointegration, the Dynamic Econometrics Conference, 2019 and International Conference on Computational and Financial Econometrics, 2019. All calculations and graphs use *OxMetrics* (Doornik, 2018) and *PcGive* (Doornik and Hendry, 2018). Contact details: jennifer.castle@magd.ox.ac.uk; jurgen.doornik@nuffield.ox.ac.uk; david.hendry@nuffield.ox.ac.uk.

has centered on techniques for regularization to favour sparser models by either selection, shrinkage, or variable combination to fewer 'factors'.

Machine learning methods are seen as an efficient approach to dealing with big data sets, e.g. Athey (2018), but they often make assumptions about the properties of the data analysed, and in particular usually assume that the data generating mechanism (DGP) does not change over time. The absence of wide-sense non-stationarity requires assuming the data are integrated of order 0 (denoted as $\mathsf{I}(0)$, possibly after appropriate transformations) and are not subject to any distributional shifts. Such restrictive assumptions do not hold in most economic environments, and hence this paper explores how useful big data can be in non-stationary settings, allowing for stochastic trends and other distributional shifts. The approach we advocate is model selection using *Autometrics* (see Doornik, 2009), where the non-stationarities can be jointly modelled with all other aspects of specification, and big data can be handled efficiently by controlling the null retention rate. This is a form of machine learning over variables and non-stationarities for a given sample. The model selection algorithm enables learning what matters and what is changing in that sample, as it is imperative to establish what features are invariant to rely on the findings.

Economic data are approximate measurements subject to revisions on an evolving, high-dimensional, inter-correlated and probably non-linear system, that is prone to abrupt shifts; CovID-19 is the latest, horrible, example. To represent such a non-stationary process requires models that account for (a) all substantively important variables; (b) dynamics and unit roots; (c) outliers and location shifts; and (d) non-linear dependencies. Omitting key features from any model will inevitably result in erroneous conclusions as other aspects of that model will proxy the missing information. The modeller cannot know the correct specification in advance for observational data and therefore models must almost always be data-based on the available sample to discover what matters empirically. Given this framework, the natural question to ask is whether automatic methods applied to big data can help? Varian (2014) believes the answer is yes. There are many potential predictors, so some form of model selection is needed; but large datasets allow for more flexible models which could address all aspects (a)-(d).

The structure of the paper is as follows. Section 2 outlines the approach to modelling wide-sense non-stationary data that is large in dimension (large $N$ relative to $T$), discussing the importance of controlling the retention of irrelevant variables in selection (§2.1), how to deal with stochastic trends and the implementation of cointegration (§2.2), and summarising other possible approaches (§2.3). Section 3 questions whether the data reduction approach of principal components analysis can extract long-run information from a large data matrix in the form of the cointegrating relations. Theoretical analysis is applied to a bivariate model to identify the resulting factors, and simulated data is used to evaluate both single-equation cointegration analysis and factor approaches, before section 4 then illustrates the model selection algorithm *Autometrics* within a non-orthogonal cointegrated setting to evaluate the properties of selection applied to big data. Section 5 undertakes an empirical example, modelling the UK unemployment rate using macroeconomic and Google Trends data, searching over thousands of potential explanatory variables, and yet selection is able to isolate the cointegrating relation and result in a parsimonious, well-specified and theoretically interpretable model of monthly unemployment rates. Finally section 6 concludes.

## 2 Wide-sense non-stationary big data

The proposed approach deals with the case where $N$ (the number of regressors $z_{i,t}$) is large relative to $T$ and many of the $\mathbf{z}_i = (z_{1,i}, \ldots, z_{T,i})'$ are $\mathsf{I}(1)$ and/or subject to outliers or shifts in mean. We assume that the DGP of the variable of interest, $\mathbf{y} = (y_1, \ldots, y_T)'$ (which could be multivariate), has $n \ll T$ relevant regressors which are included in the initial set of $N$ candidate regressors, and $\mathbf{y}$ can also be subject to outliers and breaks. The problem is a model selection problem that must jointly deal with more regressors than observations, possible outliers and breaks, and an unknown DGP including possible cointegrating relations, dynamics, and non-linearities.

Doornik and Hendry (2015) outline the proposed approach to modelling big data, building on Doornik (2009) and Hendry and Doornik (2014), now implemented in Doornik and Hendry (2018). They highlight five problems that need to be addressed including ensuring all substantively relevant variables are included at the outset (formulation problem); using the appropriate model form for the underlying DGP (specification problem); eliminating irrelevant effects while retaining relevant influences (selection problem); checking the selected model is well-specified (evaluation problem); and handling huge numbers of candidate variables (computational problem). The issue for all statistical analyses of observational data, be it big or small data, is how to avoid concluding with a substantively mis-specified model, or a spurious relationship at the extreme.

We commence with very large general models (denoted the GUM: General Unrestricted Model), playing to the advantage of big data, but use rigorous model selection to impose sparcity. By commencing with very large models including all potentially relevant variables, dynamic reactions, non-linearities, outliers, shifts, unit roots and cointegration and carefully considering the endogeneity (or exogeneity) status of candidate variables, such initial models would satisfy the requirements for valid inference to ensure selection decisions are well based. The GUM is specified in $\mathsf{I}(1)$ space, with long lags to model dynamics and possibly non-linearities; differencing transformations are not applied initially. Outliers and shifts are detected in the first stage using saturation estimation (Hendry, Johansen, and Santos, 2008) including impulse indicator saturation (IIS: Johansen and Nielsen, 2009) and step indicator saturation (SIS: Castle, Doornik, Hendry, and Pretis, 2015), which tackle multiple outliers and shifts of unknown magnitudes at unknown locations at any point in the sample. Selection is undertaken using tight significance levels given $N \gg T$ (see §2.1). Cointegration is then checked, and the model is reformulated to a stationary representation.

Most selection tests remain valid in $\mathsf{I}(1)$ space, only tests for a unit root need non-standard critical values. Most diagnostic tests are also valid for integrated series, see Wooldridge (1999), although the Heteroscedasticity tests are an exception, see Caceres (2007). Outliers are likely after mapping to $\mathsf{I}(0)$ space (Hendry and Mizon, 2011), and the source of the outliers, be they measurement errors or shocks, can matter greatly. Location shifts in $\mathsf{I}(1)$ space are outliers in $\mathsf{I}(0)$ space and they must be modelled to ensure valid inference. The formulation of the deterministic terms is essential to ensure a congruent model, see e.g., Johansen (1992) and Johansen, Mosconi, and Nielsen (2000), and this equally applies to big data models where the data are often subject to sudden distributional shifts (e.g. Google Trends or Twitter data). Finally, big data often have big measurement errors. It is unclear if these will matter for large $N$ due to the law of large numbers. The wide-sense non-stationary analysis in Duffy and Hendry (2017) also suggests measurement errors are not crucial, considering the $T \gg N$ case. When shifts or trends in the data are large, cointegration analysis is not much affected by such measurement errors.

We next outline how automatic model selection is applied in §2.1, before considering how to deal

with stochastic trends and the implementation of cointegration in §2.2, and summarising other possible approaches in §2.3.

## 2.1 Automatic model selection for big data

Model selection is undertaken using *Autometrics*, which is a tree-search algorithm selecting congruent, parsimonious, encompassing representations. Implicitly, all $2^N$ possible models need checking but that is infeasible for large $N$ and $N > T$. When $N > T$, *Autometrics* uses a multiple-block search algorithm, see Hendry and Doornik (2014, ch.19). The essential aspect of the search algorithm is to control the gauge, defined as the empirical null retention frequency of the selection procedure. With large numbers of potential regressors the concern is that spurious relationships will be identified because of the vast number of multiple comparisons undertaken. With *Autometrics*, this risk is controlled by the selection significance which determines the gauge. Table 7.1 in Hendry and Doornik (2014) records the probabilities of all $2^N$ null rejection outcomes in t-testing at a critical value $c_\alpha$ (significance level $\alpha$) for $N$ irrelevant regressors in an I(0) setting. The average number of null variables retained ($k$) is:

$$k = \sum_{i=0}^{N} i \frac{N!}{i!\,(N-i)!} \alpha^i \,(1-\alpha)^{N-i} = N\alpha. \tag{1}$$

When $\alpha = 0.01\%$ with $N = 10,000$, the number of false rejections, namely $N \times \alpha$, results in $k = 1$ from (1). Despite more than $10^{3000}$ possible outcomes, only one irrelevant variable will be kept on average. However, for $N = 100,000$ at the same significance level, $k = 10$, so 10 irrelevant variables will be retained on average, albeit removing 99,990 irrelevant variables that have been checked for significance. Thus, spurious correlations can be controlled, but as $N$ gets larger the significance level must be tightened to control the gauge, which comes at the cost of a reduced probability of retaining relevant variables.

Under Normality, critical values increase slowly under the null as $\alpha$ decreases, as Table 7.2 in Hendry and Doornik (2014) shows for a range of significance levels. 50-fold reductions in $N\alpha$ can be attained for less than $50\%$ increases in $c_\alpha$. Even for $N = 5$ million, using a Normal approximation, under the null $\Pr(|\mathsf{t}| \geq 6) \simeq 2 \times 10^{-7}$, (see Selby, 1970, p. 933), then $\alpha N \simeq 1$. This relies heavily on Normality so it is essential to remove shifts, outliers, asymmetry and fat tails to ensure errors are 'essentially independent' approximately Normal martingale difference sequences.

| $\alpha$ | $c_\alpha$ | $\psi$ | $\mathsf{P}(\mathsf{t} \geq c_\alpha)$ | $[\mathsf{P}(\mathsf{t} \geq c_\alpha)]^4$ |
|---|---|---|---|---|
| 0.0001 | 4.00 | 3 | 0.16 | 0.001 |
| 0.0025 | 3.025 | 4 | 0.83 | 0.47 |
| 0.0001 | 4.00 | 4 | 0.50 | 0.063 |
| 0.0001 | 4.00 | 6 | 0.98 | 0.907 |

Table 1: t-test power at $T = 100$ using Normal critical values.

Given a vast number of variables, the need to use very tight significance levels to control 'false positives' inevitably entails that relevant variables with relatively small non-zero non-centralities, denoted $\psi$, will be harder to detect. Table 1 shows the approximate t-test power when a coefficient null hypothesis is tested once for the likely small values of $\alpha$ required when $N \geq 1000$. Although there is roughly a 50–50 chance of retaining a variable with $\mathsf{E}[\mathsf{t}] = \psi = c_\alpha$, there is little chance of keeping 4 such independent

variables until $\psi \gg c_\alpha$, as seen in the final column. These considerations apply to all search and learning algorithms.

## 2.2 Handling stochastic trends

There are two ways to deal with stochastic trends. Either the variables can be differenced a sufficient number of times to obtain $\mathsf{I}(0)$ data, or linear combinations of variables that reduce the order of integration could be obtained. The first approach is standard in factor analysis, see, e.g., Forni, Hallin, Lippi, and Reichlin (2000) and Stock and Watson (2002). If the target of selection is the DGP which has an underlying economic theory based on long-run equilibria, such relationships will not be modelled by a differencing approach.[1] Furthermore, the object of interest is the data in levels, so any transformation of the data to differences will need to be unravelled to interpret outcomes, and this transformation is particularly important for forecasting multi-steps ahead, where forecast errors in differences will cumulate and can result in systematic forecast failure in levels.

The origins of cointegration lie in the debates about viable links between levels and differences in time series that are non-stationary because of unit roots in their dynamics, see Hendry (2004). Imposing valid cointegration will generally help to obtain a clearer interpretation of the model as a result of a more orthogonal model formulation. Sims, Stock, and Watson (1990) show that transforming models to stationary form by differencing or cointegration is often unnecessary when testing autoregressive models with unit roots. Imposing cointegration will have little impact on the case of big $T$ and small $N$, as long as the appropriate distribution for the relevant test statistics is used. They show that the asymptotic distribution for the coefficients of the model without imposing cointegration is exactly the same as if the cointegrating relations were known *a priori*. Therefore, if the cointegrating relations automatically hold for big $T$, it is not necessary to impose them. However, this result is asymptotic in the direction of $T$ and it is not clear that the result would hold for big $N$ and small $T$. Hence, there may be some advantages to applying cointegrating transforms for 'fat' data. A lack of cointegration could be the more important aspect in big data, as nonsense relations need to be tightly controlled when there are no long-run links.

Our approach tests for cointegration so neither ignores it by differencing, nor imposes it at the outset. As $N$ is large and cointegration typically relates small numbers of variables, cointegration is tested after selection is initially applied, but we also investigate whether cointegrating relations can be detected within large $N$ in §3.

## 2.3 Other approaches for modelling non-stationary big data

Systems for huge data (large $N$, large $T$) will be very demanding and only possible by automatic modelling. One route forward is to use partial systems for subsets of endogenous variables $m < N$ and then combine. As an example of this approach, Hendry (2001) develops a model of inflation in which equilibrium-correction terms are developed for all different sectors of the economy, corresponding to excess demands for goods and services, factors of production, money, financial assets, foreign exchange, and government deficits. These are then included in a large model of inflation and model selection can be applied, allowing for additional exogenous variables (such as commodity prices), dynamics, breaks and outliers, and possible non-linearities. The methodology enables cointegration to be embedded within a

[1] As an example, in the energy-balance models of the Earth's climate, cointegrating vectors correspond to laws of conservation of energy (see Pretis, 2020).

big data approach to modelling phenomena of interest. This could be generalized from single-equation equilibrium-correction mechanisms to a multivariate approach as in Harbo, Johansen, Nielsen, and Rahbek (1998). However, shifts need to be handled in each sub-system, see Kurita and Nielsen (2019). Furthermore, all non-modelled variables will need to be forecast 'off-line', or outside of the partial system, in order to produce unconditional forecasts. This is a challenge for big data when $N - n$ is huge. Hendry and Mizon (2012) derive a forecast error taxomony for open systems and show that it can pay to omit the unmodelled regressors when forecasting, so big data may not help in this context.

A second approach is to use data reduction methods to obtain linear combinations of data forming latent 'factors'. The literature on factor approaches is huge, ranging from principal components (PCA: Stock and Watson, 1998), dynamic factor models (DFM: Forni, Hallin, Lippi, and Reichlin, 2000), partial least squares (PLS: Dijkstra, 1983) and factor augmented VARs (FAVAR: Bernanke, Boivin, and Eliasz, 2005). These approaches tend to ignore long-run information when computing the factors, but Castle, Clements, and Hendry (2013) demonstrate how factors can be included in initial general models jointly with large $N$ variables.

The factor-augmented error correction model (FECM) introduced by Banerjee and Marcellino (2009) combines equilibrium-correction, cointegration and dynamic factor models, see Banerjee, Marcellino, and Masten (2016) for a discussion. In the FECM approach cointegration is captured between the variables and the factors. The error correction term is given by $(X_{t-1} - \Lambda F_{t-1})$ where $X_t$ is an $N$ dimensional set of regressors and $F_t$ are the $\mathsf{I}(1)$ factors computed using the method in Bai (2004). This approach requires pre-testing to group the data into $\mathsf{I}(0)$ and $\mathsf{I}(1)$ categories, which is a potential drawback relative to our approach.

# 3 Using principal components to detect cointegrating relations

Principal components are used as a reduction device when forecasting with small $T$ relative to big $N$. By differencing $\mathsf{I}(1)$ data to remove unit roots, any cointegration will be lost, and if the data are originally $\mathsf{I}(2)$, then second differencing will do the same. When $N$ is large, the standard multivariate cointegration analysis of Johansen (1995) becomes intractable due to vanishing degrees of freedom. We next explore the use of principal components to identify the cointegrating relations, first in a small analytical exercise to see if the principal components do map to the cointegrating relations, and then using simulations to see if there are any regularities in which principal components (ordered by explained variance) isolate the cointegrating relations.

As $\mathsf{I}(1)$ time series have explosive variances but cointegrating vectors give linear combinations of these time series with finite variances, we might expect that choosing $r$ linear combinations with the minimum variance, i.e. the smallest $r$ principal components, should correspond to the estimated cointegrating vectors with no need for identifying restrictions. Harris (1997) gives a principal components estimator of cointegrating vectors that is consistent but asymptotically inefficient, and a modified estimator that is asymptotically efficient in a wide range of cases. Snell (1999) provides a test for cointegrating vectors from the principal components, where the $r$th principal component is $\mathsf{I}(0)$ under the null but $\mathsf{I}(1)$ under the alternative. Lansangan and Barrios (2008) show that principal components analysis can lead to one or very few components explaining the variability within the data if the data are non-stationary in the form of drifting mean, assigning similar loadings to all variables, and Zhao and Shang (2016) propose a method of detrended cross-correlation analysis to minimize the effects of the non-stationarity. The

literature using principal components to identify cointegrating relations relies on defining the 'correct' system of $n$ I (1) variables, i.e. all variables are non-stationary and enter the DGP. Our application of big data implies that irrelevant regressors may contaminate the principal components, so we investigate that situation in §3.2.

### 3.1 Principal components in a bivariate system in levels

Consider the simplest bivariate cointegrated system:

$$x_{1,t} = x_{2,t} + \epsilon_{1,t} \text{ where } \epsilon_{1,t} \sim \mathsf{IN}\left[0, \sigma_1^2\right], \tag{2}$$

$$x_{2,t} = x_{2,t-1} + \epsilon_{2,t} \text{ where } \epsilon_{2,t} \sim \mathsf{IN}\left[0, \sigma_2^2\right]. \tag{3}$$

$\mathsf{E}\left[\epsilon_{1,t}\epsilon_{2,s}\right] = 0 \; \forall t, s$, and $x_{1,0} = x_{2,0} = 0$. Although far from 'big data', unless $T$ is very large, the key ingredients for a principal components (PC) analysis of cointegration are present. First:

$$(x_{1,t} - x_{2,t}) = \epsilon_{1,t}, \tag{4}$$

so the cointegrating vector is $(1 \; : \; -1)$, with an equilibrium-correction representation which can be written as:

$$\Delta x_{1,t} = \Delta x_{2,t} - \alpha(x_{1,t-1} - x_{2,t-1}) + \epsilon_{1,t},$$

where $\alpha = 1$, and:

$$\Delta x_{2,t} = \epsilon_{2,t}, \tag{5}$$

so in expectation its second moment is:

$$\mathsf{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{t}\epsilon_{2,j}^2\right] = \sigma_2^2\sum_{t=1}^{T} t = \sigma_2^2 T(T+1)(2T+1)/6 \approx \sigma_2^2 T^3/3. \tag{6}$$

Next from the integral of (5):

$$\mathsf{E}\left[\sum_{t=1}^{T} x_{1,t}^2\right] \approx \sigma_2^2 T^3/3 + \sigma_1^2 T$$

and:

$$\mathsf{E}\left[\sum_{t=1}^{T} x_{1,t}x_{2,t}\right] \approx \sigma_2^2 T^3/3.$$

Combining:

$$\mathsf{E}\left[T^{-3}\begin{pmatrix} \sum_{t=1}^{T} x_{1,t}^2 & \sum_{t=1}^{T} x_{1,t}x_{2,t} \\ \sum_{t=1}^{T} x_{1,t}x_{2,t} & \sum_{t=1}^{T} x_{2,t}^2 \end{pmatrix}\right] \approx \frac{1}{3}\sigma_2^2\begin{pmatrix} 1+3\sigma_1^2\sigma_2^{-2}T^{-2} & 1 \\ 1 & 1 \end{pmatrix} = \frac{1}{3}\sigma_2^2\mathbf{M},$$

so that letting $3\sigma_1^2\sigma_2^{-2}T^{-2} = m$:

$$\begin{pmatrix} 1 & -1 \end{pmatrix}\mathbf{M} = \begin{pmatrix} 1 & -1 \end{pmatrix}\begin{pmatrix} 1+m & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} m & 0 \end{pmatrix},$$

confirming the second column delivers the cointegrating vector.

Obtaining the prinicipal components using a standard eigenvalue-eigenvector decomposition $\mathsf{E}[\mathbf{M}] = \mathbf{H}\mathbf{\Lambda}\mathbf{H}'$ with eigenvalues $\frac{1}{2}m$ and $\left(\frac{1}{2}m+2\right)$ and orthogonal cross product once normalized by $(m+2)^{-1/2}$:

$$\mathbf{h}_1 = \begin{pmatrix} \left(\frac{1}{2}m - \frac{1}{2}\sqrt{m^2+4}\right) \\ 1 \end{pmatrix}; \quad \mathbf{h}_2 = \begin{pmatrix} \left(\frac{1}{2}m + \frac{1}{2}\sqrt{m^2+4}\right) \\ 1 \end{pmatrix}$$

so:

$$\mathbf{h}_1'\mathbf{h}_2 = \left(\left(\frac{1}{2}m - \frac{1}{2}\sqrt{m^2+4}\right)\left(\frac{1}{2}m + \frac{1}{2}\sqrt{m^2+4}\right) + 1\right) = 0.$$

Below, $m^2$ is $\mathsf{O}(T^{-4})$ so is treated as negligible, in which case, as a check, $(m+2)^{-1}\mathbf{H}\mathbf{\Lambda}\mathbf{H}'$ becomes:

$$\begin{pmatrix} \left(\frac{1}{2}m-1\right) & \left(\frac{1}{2}m+1\right) \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \left(\frac{1}{2}m\right) & 0 \\ 0 & \left(\frac{1}{2}m+2\right) \end{pmatrix}\begin{pmatrix} \left(\frac{1}{2}m-1\right) & 1 \\ \left(\frac{1}{2}m+1\right) & 1 \end{pmatrix} \approx (m+2)\begin{pmatrix} 1+m & 1 \\ 1 & 1 \end{pmatrix},$$

using the approximation:

$$(m+2)^{-1}(3m+2) = \frac{2\left(1+\frac{3}{2}m\right)}{2\left(1+\frac{1}{2}m\right)} = \frac{\left(1+\frac{3}{2}m\right)}{\left(1+\frac{1}{2}m\right)} = \left(1+\frac{3}{2}m\right)\left(1-\frac{1}{2}m\right) \approx (1+m).$$

Thus approximate eigenvalues are $\frac{1}{2}m$ and $(\frac{1}{2}m+2)$ where the latter is bound to be the larger. The first row of $\mathbf{H}$ is close to $(1:-1)$ whereas the second row is $(1:1)$, and:

$$\mathbf{H}\mathbf{x}_t \approx (m+2)^{-1/2}\begin{pmatrix} mx_{2,t} - \left(1-\frac{1}{2}m\right)\epsilon_{1,t} \\ 2x_{2,t} + \epsilon_{1,t} \end{pmatrix},$$

so for $m \approx 0$, the first component estimates the cointegrating vector, but the second principal component has the largest variance, which does not involve the cointegrating vector. These results suggest that the largest principal component may not detect the long-run equilibrium relation, so current approaches to determining the optimal number of factors to include, such as a scree plot or total percentage variance explained, need to be reconsidered when the aim is to include long-run information in the model.

### 3.2 Simulations

We explore the analytic results further using Monte Carlo simulations for a more general DGP. The principal components approach is compared to computing the long-run cointegrating relation in a single-equation autoregressive distributed lag (ADL) model. The DGP is given by:

$$x_{1,t} = \beta_0 + \beta_1 x_{2,t} + \beta_2 x_{1,t-1} + \beta_3 x_{2,t-1} + v_{1,t}, \tag{7}$$

$$x_{2,t} = x_{2,0} + t\mu_2 + \sum_{j=1}^{t} v_{2,j}, \tag{8}$$

with:

$$(v_{1,t}, v_{2,t})' \sim \mathsf{IN}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{pmatrix}\right]. \tag{9}$$

We set $x_{2,0} = 0$ but discard 20 initial observations. The baseline parameters are $\beta_0 = 0.1$, $\beta_1 = 0.5$, $\beta_2 = 0.8$, $\beta_3 = 0.5$, $\mu_2 = 0.5$, $\sigma_{11}^2 = \sigma_{22}^2 = 1$, $\sigma_{12} = \sigma_{21} = 0$, and $T = 100$.

Additional irrelevant regressors are generated to 'contaminate' the variance-covariance matrix. Three cases are examined including:
(1) no additional regressors;
(2) one additional regressor, either $\mathsf{I}(0)$ or $\mathsf{I}(1)$; and
(3) ten additional regressors, either $\mathsf{I}(0)$ or $\mathsf{I}(1)$, with varying degrees of correlation between the additional regressors.
The additional regressors are uncorrelated with $x_1$ and $x_2$ in the population. We generate the additional regressors, $\mathbf{w}_t = (w_{1,t}, \ldots, w_{10,t})'$:

$$\Delta \mathbf{w}_t \sim \mathsf{IN}_{10}\left[\boldsymbol{\mu}_w, \boldsymbol{\Upsilon}\right] \tag{10}$$

where the elements of $\boldsymbol{\Upsilon}$ are $\upsilon_{ij} = 1$ for $i = j$ and $\upsilon_{ij} = \rho$ for $i \neq j$, and the $\mathsf{I}(1)$ additional regressors are obtained as $\mathbf{w}_t = \sum_{j=1}^{t} \Delta \mathbf{w}_j$, to give random walks with drift. We consider $\rho = 0, 0.5, 0.9$. All simulations are conducted with $M = 10,000$ replications.

### 3.2.1 Computing Principal Components

For the three cases outlined above let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2)'$ [no additional regressors, $K = 0$]; $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{w}_1)'$ [one additional regressor, $K = 1$]; or $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{w}_1, \ldots, \mathbf{w}_{10})'$ [ten additional regressors, $K = 10$], where $\mathbf{X}$ is of dimension $(T \times (K + 2))$.

Let $\widehat{\boldsymbol{\Omega}}$ denote the sample correlation matrix of $\mathbf{X}$. The eigenvalue decomposition is:

$$\widehat{\boldsymbol{\Omega}} = \widehat{\mathbf{H}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{H}}', \tag{11}$$

where $\widehat{\boldsymbol{\Lambda}}$ is the diagonal matrix of ordered eigenvalues $(\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_n \geq 0)$ and $\widehat{\mathbf{H}} = (\widehat{\mathbf{h}}_1, \ldots, \widehat{\mathbf{h}}_n)$ is the corresponding matrix of eigenvectors, with $\widehat{\mathbf{H}}'\widehat{\mathbf{H}} = \mathbf{I}_n$. The sample principal components are computed as:

$$\widehat{\mathbf{Z}} = \widehat{\mathbf{H}}'\widetilde{\mathbf{X}}, \tag{12}$$

where $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_T)'$ is the standardized data, $\widetilde{x}_{j,t} = (x_{j,t} - \overline{x}_j) / \widetilde{\sigma}_{x_j}$ $\forall j = 1, \ldots, K + 2$, where $\overline{x}_j = \frac{1}{T}\sum_{t=1}^{T} x_{j,t}$ and $\widetilde{\sigma}_{x_j} = \left[\frac{1}{T}\sum_{t=1}^{T}(x_{j,t} - \overline{x}_j)^2\right]^{1/2}$.

### 3.2.2 Evaluation

To evaluate whether the principal components measure the cointegrating relation, we compute the known cointegrating vector given the DGP for each draw of the simulation:

$$c_t = x_{1,t} - \kappa_0 - \kappa_1 x_{2,t} \tag{13}$$

where $\kappa_0 = \frac{\beta_0}{1-\beta_2} = 0.5$ and $\kappa_1 = \frac{\beta_1+\beta_3}{1-\beta_2} = 5$.

The absolute value of the correlation between the prinicipal components and the cointegrating relation is given by:

$$\widehat{\boldsymbol{\rho}}_j^{pc} = \left|\mathsf{cov}\left[\widehat{\mathbf{z}}_j, \mathbf{c}\right]\left(\mathsf{V}\left[\widehat{\mathbf{z}}_j\right]\mathsf{V}\left[\mathbf{c}\right]\right)^{-\frac{1}{2}}\right| \tag{14}$$

for $j = 1, \ldots, K + 2$, where bold denotes stacking over $t = 1, \ldots, T$.

### 3.2.3 Single-equation cointegration analysis

The principal components approach is compared to the estimated equilibrium correction mechanism ($ecm$) obtained by solving for the long-run solution. The general unrestricted model (GUM) is:

$$x_{1,t} = \beta_0 + \beta_1 x_{2,t} + \beta_2 x_{1,t-1} + \beta_3 x_{2,t-1} + \sum_{j=1}^{K} \sum_{l=0}^{1} \beta_{w,jl} w_{j,t-l} + u_t, \tag{15}$$

for $K = 0, 1$ or 10, and compute the $ecm$:

$$\widehat{c}_t = x_{1,t} - \widehat{\kappa}_0 - \widehat{\kappa}_1 x_{2,t} - \sum_{j=1}^{K} \widehat{\kappa}_{w,j} w_{j,t}, \tag{16}$$

where $\widehat{\kappa}_0 = \frac{\widehat{\beta}_0}{1-\widehat{\beta}_2}$, $\widehat{\kappa}_1 = \frac{\widehat{\beta}_1+\widehat{\beta}_3}{1-\widehat{\beta}_2}$, and $\widehat{\kappa}_{w,j} = \frac{\widehat{\beta}_{w,j0}+\widehat{\beta}_{w,j1}}{1-\widehat{\beta}_2}$. Note that $\mathsf{E}\left[x_{1,t} - \widehat{\kappa}_0 - \widehat{\kappa}_1 x_{2,t}\right] = 0$ under cointegration. The DGP weights are $\kappa_0 = 0.5$, $\kappa_1 = 5$ and $\kappa_{w,j} = 0$. We compute the absolute value of the correlation between the estimated $ecm$ from the ADL(1,1) model (16) and the known cointegrating relation (13):

$$\widehat{\boldsymbol{\rho}}^{adl} = \left| \mathsf{cov}\left[\widehat{\mathbf{c}}, \mathbf{c}\right] \left(\mathsf{V}\left[\widehat{\mathbf{c}}\right] \mathsf{V}\left[\mathbf{c}\right]\right)^{-\frac{1}{2}} \right|. \tag{17}$$

### 3.2.4 Results

If the vector of regressors contains the correct cointegrating variables and no others, the last principal component will measure the $ecm$. This result can be seen in figure 1, which records the histograms for the correlations between the principal components or long-run solutions and the DGP $ecm$. Panel a is the first principal component, which detects the common trend and not the cointegrating relation. Panel b shows that the second PC detects the cointegrating relation with reasonable accuracy. However, the ADL estimated $ecm$ is extremely precise, much more so than the second principal component. Increasing $\sigma_{11}^2$ and $\sigma_{22}^2$ makes the distinction between the first principal component estimating the long-run trend and the second estimating the cointegrating vector much less clear, but has less of an effect on the ADL estimated $ecm$.

For the case with one additional irrelevant $\mathsf{I}(1)$ regressor, the precision of the ADL model $ecm$ is unaffected by the additional regressor but the precision of the smallest principal component in estimating the $ecm$ worsens considerably, recorded in figure 2. Panel c records the correlation between the smallest principal component and the $ecm$ which has a more dispersed distribution away from unity compared to figure 1b when no irrelevant variables were included. Adding an additional stationary regressor did not affect the principal components approach to estimating the cointegrating relation to this extent as the smallest principal component remained reasonably accurate in estimating the $ecm$.

Finally, figure 3 records the correlation distributions with ten additional irrelevant $\mathsf{I}(1)$ regressors, correlated $\rho = 0.9$. None of the principal components manage to estimate a measure of the $ecm$, but the ADL model does a remarkably accurate job despite so many irrelevant correlated non-stationary regressors. The factor approach is unable to detect the $ecm$ in a given PC when additional variables are included regardless of whether the additional regressors are uncorrelated in the population or are highly correlated. The convenient result for the bivariate case that the smallest PC measures the $ecm$ is disrupted when the set of regressors is not exactly the DGP set, and varying the correlations and drift

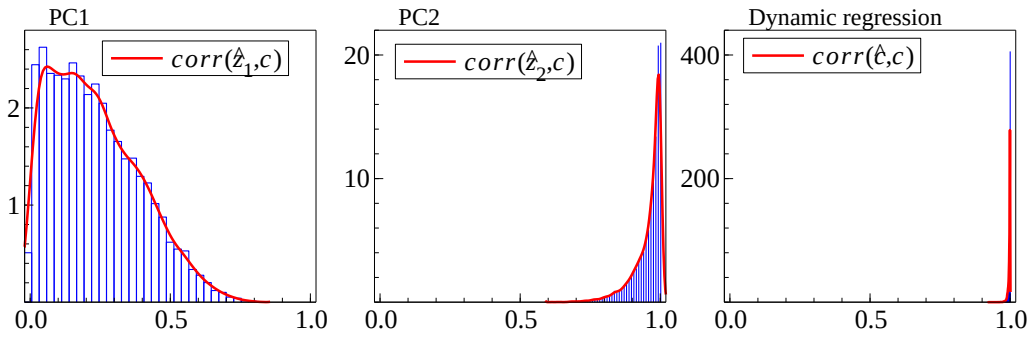


Figure 1: Dynamic DGP with no additional regressors. Panel a: absolute correlation between DGP $ecm$ and first principal component; Panel b: absolute correlation between DGP $ecm$ and second principal component; Panel c: absolute correlation between DGP $ecm$ and long-run solution from an ADL(1,1) model.

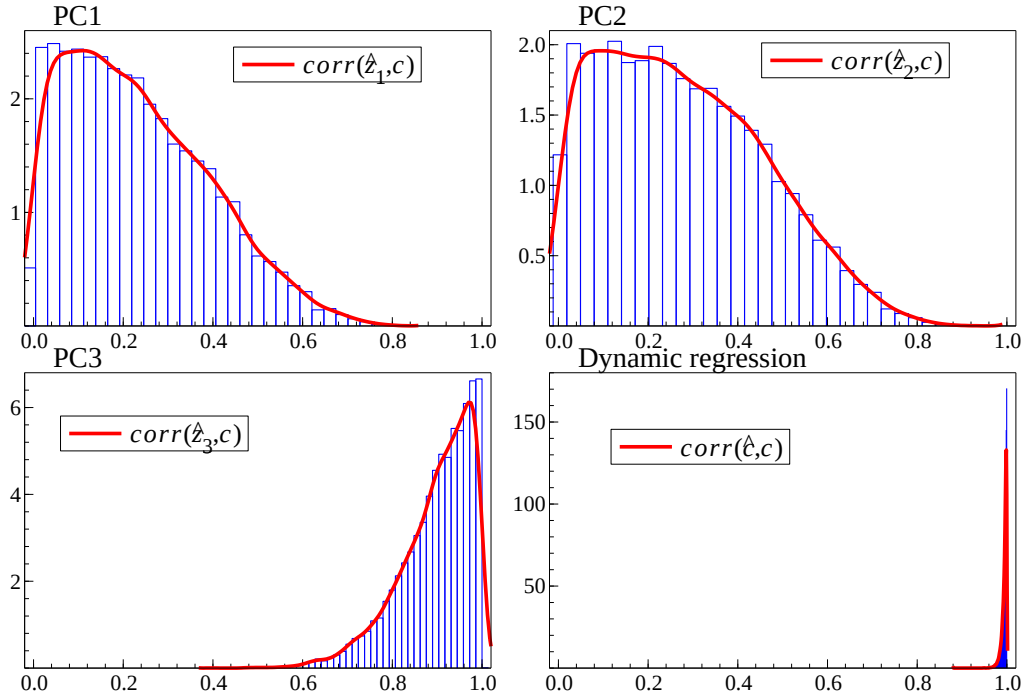


Figure 2: Dynamic DGP with one additional $\mathsf{I}(1)$ with drift regressor. Panel a: absolute correlation between DGP $ecm$ and first principal component; Panel b: absolute correlation between DGP $ecm$ and second principal component; Panel c: absolute correlation between DGP $ecm$ and third principal component; Panel d: absolute correlation between DGP $ecm$ and long-run solution from an ADL(1,1) model.

between the irrelevant variables can deliver almost any result for which factor measures the $ecm$, or none at all, whereas additional irrelevant variables do not effect the regression approach to estimating the long-run solution accurately, even if they are $\mathsf{I}(1)$ and highly correlated.

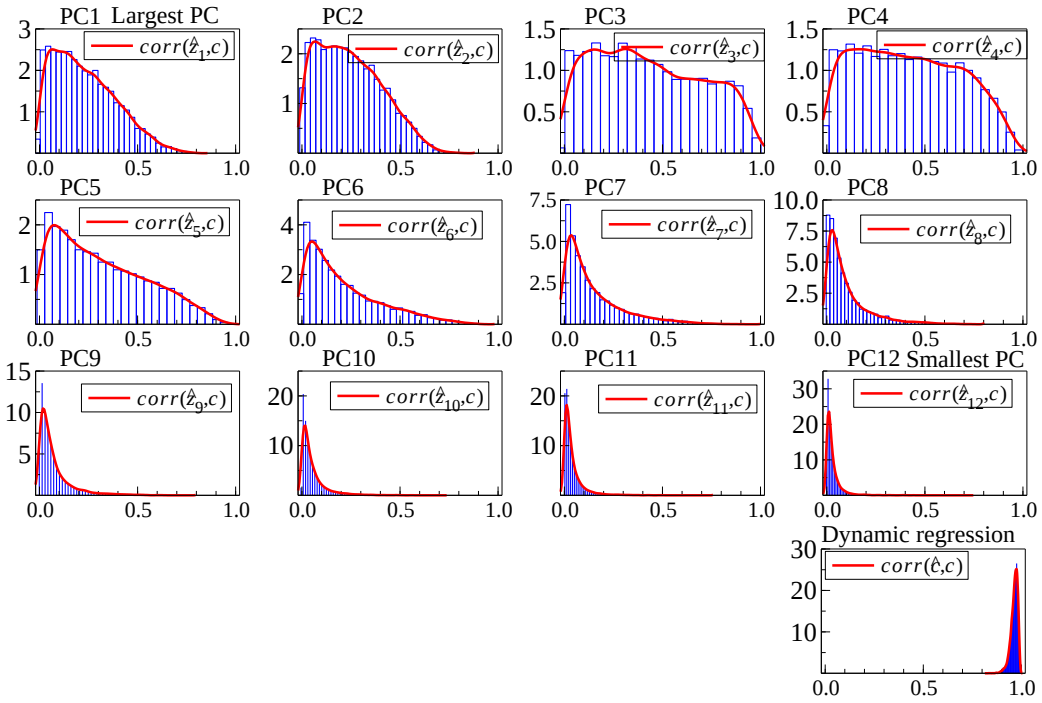


Figure 3: Dynamic DGP with ten additional $\mathsf{I}(1)$ with drift regressors correlated $\rho = 0.9$. Top three rows of panels: absolute correlation between DGP $ecm$ and principal components, ordered in rows from first to last PC; Bottom right panel: absolute correlation between DGP $ecm$ and long-run solution from an ADL(1,1) model.

### 3.3 Summary of using principal components to obtain cointegrating relations

If it is known which variables enter the cointegrating relation and no others are included, a factor approach is able to detect the $ecm$. It is often the last PC, which is frequently ignored in factor approaches where only the first few factors are used. However, in practice the factor approach conditions on numerous variables (large $N$) which are highly correlated. In this case, principal components is not able to accurately obtain the $ecm$ and is therefore not a valid approach for undertaking cointegration analysis. For the regression approach, if the DGP is dynamic, the long-run solution is accurately estimated despite 'contamination' with irrelevant, correlated $\mathsf{I}(1)$ regressors. These results point towards favouring our general approach to handling big data, perhaps augmented with latent common factors to detect many small effects (e.g., Castle, Clements, and Hendry, 2013).

# 4 Cointegration illustrations on I(1) big data using *Autometrics*

To investigate the ability of *Autometrics* to select the DGP facing I(1) 'big data', we conduct a simulation experiment with varying numbers of potential variables within a cointegrated single-equation setting. The cointegrated DGP is:

$$y_t = \sum_{i=1}^{20} \beta_i z_{i,t} + \beta_{21} y_{t-1} + \epsilon_t, \tag{18}$$

where:

$$z_{i,t} = z_{i,t-1} + \nu_t \;\; i = 1, \ldots, 20. \tag{19}$$

We set $\beta_{21} = 0.75$, $\boldsymbol{\nu}_t \sim \mathsf{IN}_{20}[\mathbf{0}, \mathbf{I}]$ and $\epsilon_t \sim \mathsf{IN}[0, 1]$. The relevant variables' coefficients are given in table 2, where the first 10 variables have non-zero non-centralities in pairs over $\pm 0.25$ to $\pm 0.35$. The additional 10 regressors $z_{11,t} - z_{20,t}$ have zero coefficients. $M = 1,000$ replications are conducted on a sample size of $T = 5000$. We evaluate selection using gauge (the empirical null retention frequency) and potency (the average non-null retention frequency).

| $z_{1,t}$ | $z_{2,t}$ | $z_{3,t}$ | $z_{4,t}$ | $z_{5,t}$ | $z_{6,t}$ | $z_{7,t}$ | $z_{8,t}$ | $z_{9,t}$ | $z_{10,t}$ |
|---|---|---|---|---|---|---|---|---|---|
| 0.25 | -0.25 | 0.30 | -0.30 | 0.275 | -0.275 | 0.325 | -0.325 | 0.35 | -0.35 |

Table 2: DGP coefficients for $\beta_i$.

The general unrestricted model (GUM) as the starting point for selection is given by:

$$y_t = \lambda_0 + \lambda_y y_{t-1} + \sum_{i=1}^{20} \sum_{s=0}^{S} \lambda_{is} z_{i,t-s} + \sum_{j=1}^{T} \delta_j I_{j=t} + \nu_t, \tag{20}$$

For the first case, we set $S = 1$ and do not apply IIS (so $\delta_j$ is set to $0 \; \forall j$) resulting in 42 regressors, of which 11 are relevant. Selection at $\alpha = 0.001$ delivered a gauge of $= 0.0027$ and a potency of 1.00. For one replication, selection took 0.13 seconds. The integrated DGP poses no problems for selection, with high potency and a gauge that is close to, although larger than, the selection significance level. However, $N = 42$ would not constitute 'big data'.

Next, we apply IIS, so keep $S = 1$ but include an indicator variable for every observation. Therefore, $N = 5042$, which is a lot of irrelevant variables to search over. Selection at $\alpha = 0.001$ delivered a gauge of 0.0029 and a potency of 0.998, with one replication taking 4 minutes and for that draw the DGP was exactly located. A tighter significance level of $\alpha = 0.0001$ would retain just one irrelevant variable on every other draw, despite searching over a mass of irrelevant variables.

To address the concern that the 'big data' aspect is the inclusion of orthogonal indicator variables which may behave differently to regressors, we next include many more lags of the relevant and irrelevant variables, setting $S = 50$ with IIS. This results in $N = 6072$ regressors. Selection at $\alpha = 0.001$ delivered a gauge of 0.0031 and a potency of 1, with one replication taking 5 minutes.

Finally, the GUM was then estimated in first differences with the ECM included at lag $k = 1$ or $k = 50$:

$$\Delta y_t = \gamma_0 + \gamma_1 ecm_{t-k} + \sum_{i=1}^{20} \sum_{s=0}^{50} \lambda_{is} \Delta z_{i,t-s} + \sum_{j=1}^{T} \delta_j I_{j=t} + \nu_t, \tag{21}$$

where $ecm$ is computed as the static long-run for the GUM (excluding impulse indicators). $N = 6022$ as the lagged dependent variables enter the long-run solution. For a single draw of the simulation the retained model selected 9 of the differenced DGP variables taking just 17 seconds. The $ecm$ located at $k = 1$ or $k = 50$ did not alter the selection results. The results show that *Autometrics* can handle large $N$ with unit roots, delivering properties of model selection close to those for small $N$ and stationary regressors. We next apply the method to an empirical example, modelling the UK unemployment rate.

# 5 Modelling the UK Unemployment Rate

Our approach to empirical modelling wide-sense non-stationary big data is illustrated in this section where we model the monthly UK unemployment rate. Clements and Hendry (2006) motivate our theory model, in which they show that the unemployment rate and the real interest rate minus the real growth rate, which we denote $Rr_t$, are cointegrated, or co-break. This 'structural' model is based on steady-state growth theory, such that the unemployment rate rises when the real interest rate exceeds the real growth rate, and vice versa. Hendry (2001) finds this relationship has held relatively constantly in the UK for 150 years over 1860–2010, with a long-run equilibrium unemployment rate of 5% having a 1-1 response to $Rr_t$.

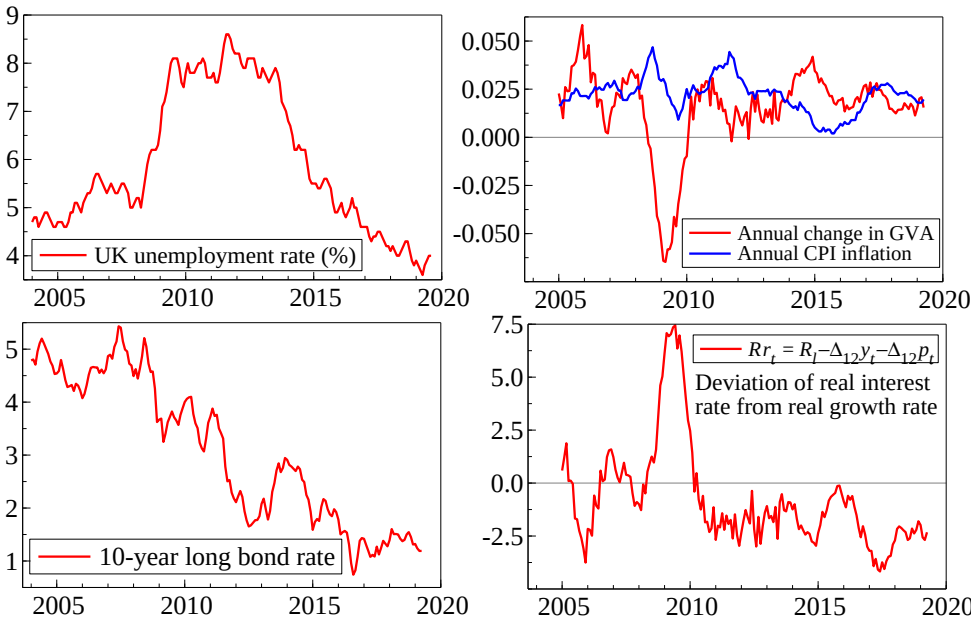


Figure 4: UK non-seasonally adjusted unemployment rate (%), annual change in GVA and annual CPI inflation, 10-year long bond rate, and $Rr_t$ in percent.

The data consist of monthly observations from $2004(1) - 2019(4)$ (chosen as the maximum available sample for the Google Trends data, which does not overlap the earlier study and is at a much higher frequency) and are plotted in figure 4, with table 3 in the Appendix recording data definitions. Let $Rr_t = R_{l,t} - \Delta_{12} y_t - \Delta_{12} p_t$, where $R_{l,t}$ is the long-term interest rate, $\Delta_{12} y_t$ is the annual change in

log Gross Value Added, and $\Delta_{12}p_t$ is the annual Consumer Price Index inflation rate. Ideally we would use consistent measures of output and inflation, i.e. GDP and its deflator, but monthly data precludes this so we use GVA and CPI inflation instead. As Duffy and Hendry (2017) show, measurement errors in $\mathsf{I}(1)$ time series need not substantively affect subsequent cointegration analysis. The unemployment rate exhibits substantial movements over the sample period, with a more than 3 percentage point increase in unemployment following the financial crisis, which persisted for several years before gradually falling back to the lowest levels over the sample period. We shall model this data using the theoretically motivated $Rr_t$ recorded in panel d, which captures the large decline in real growth relative to the real interest rate following the financial crisis.

## 5.1 Model selection on theory-motivated variables

We start by considering the relation between $Ur_t$ and $Rr_t$ as the benchmark 'theory' model. Commencing with a GUM given by an ADL specification with 13 lags of $Ur_t$ and $Rr_t$ including the contemporaneous $Rr_t$, an intercept and 11 seasonal dummies, we apply IIS and SIS, jointly selecting dynamics and indicators. Selection is undertaken using *Autometrics* at $\alpha = 0.001$ with a fixed intercept and seasonal dummies (i.e. the intercept and seasonals are not selected over). There are $N = 355$ variables with $T = 159$ observations (159 impulse indicators, 157 step indicators, 13 lags of the dependent variable, 14 coefficients for the conditioning variable, an intercept and 11 seasonal dummies). Two step indicators for 2011(5) and 2011(7) and an impulse indicator for 2011(8) were retained but were subsequently combined to a dummy variable taking the value 1 for 2011(5)-2011(8) and 0 otherwise ($D_{2011} = I_{2011(5)} + I_{2011(6)} + I_{2011(7)} + I_{2011(8)}$). The resulting selected model is:[2]

$$\begin{aligned}\widehat{Ur}_t \;=\;& \underset{(0.065)}{0.256} + \underset{(0.011)}{0.950}Ur_{t-1} + \underset{(0.006)}{0.050}Rr_t - \underset{(0.006)}{0.029}Rr_{t-5} + \underset{(0.004)}{0.017}Rr_{t-12} \\ & - \underset{(0.036)}{0.125}S_{2009(12)} + \underset{(0.033)}{0.136}S_{2013(11)} + \underset{(0.043)}{0.202}D_{2011} + \text{seasonals} \\ & \widehat{\sigma} = 0.080;\;\; \mathsf{F}_{\mathsf{AR}}(7,133) = 1.63;\;\; \mathsf{F}_{\mathsf{ARCH}}(7,145) = 1.04;\;\; \chi^2_{\mathsf{nd}}(2) = 0.49; \\ & \mathsf{F}_{\mathsf{Het}}(22,136) = 0.74;\;\; \mathsf{F}_{\mathsf{Reset}}(2,138) = 0.96;\;\; T = 2006(2) - 2019(4).\end{aligned} \tag{22}$$

The selected model is well-specified passing all diagnostic tests, and the solved long-run solution is:

$$\widehat{ecm} = Ur - \underset{(0.52)}{5.15} - \underset{(0.14)}{0.76}Rr + \underset{(0.43)}{2.52}S_{2010(1)} - \underset{(0.35)}{2.73}S_{2013(12)} \tag{23}$$

recorded in figure 5. The seasonals and impulse dummies do not enter the long-run solution, but the step dummies are included with a $1-$period lead as $\widehat{ecm}$ enters the dynamic model with a $1-$period lag. The long-run solution is remarkably similar to the annual model of excess demand for unemployment reported in Hendry (2001) over 1865–1991, with a mean of close to $5\%$ p.a. and a coefficient of $0.76$ on $Rr$ compared to the $0.77$ that Hendry (2001) found on annual data for the earlier sample.

[2]Estimated coefficient standard errors are shown in parentheses below estimated coefficients, $\widehat{\sigma}$ is the residual standard deviation, $\mathsf{R}^2$ is the coefficient of multiple correlation, $\mathsf{F}_{\mathsf{AR}}$ is a test for residual autocorrelation (see Godfrey, 1978), $\mathsf{F}_{\mathsf{ARCH}}$ tests for autoregressive conditional heteroscedasticity (see Engle, 1982), $\mathsf{F}_{\mathsf{Het}}$ is a test for residual heteroskedasticity (see White, 1980), $\chi^2_{\mathsf{nd}}(2)$ is a test for non-Normality (see Doornik and Hansen, 2008), and $\mathsf{F}_{\mathsf{Reset}}$ is the RESET test (see Ramsey, 1969).

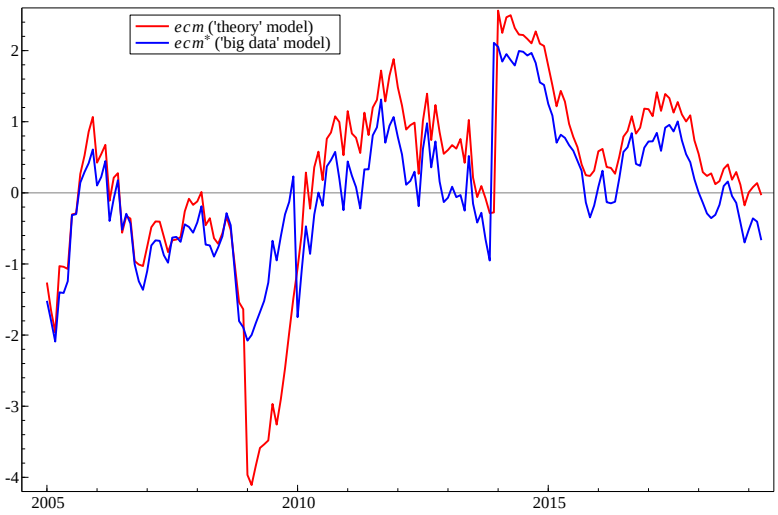


Figure 5: $ecm$s from (23) and (26).

## 5.2 Model selection on non-stationary big data

We next augment the general model to include two additional datasets that could contribute to the explanation of the unemployment rate including a large monthly macroeconomic database (§5.2.1) and Google Trends data (§5.2.2).

### 5.2.1 Macroeconomic time series

We have compiled a database of 72 monthly non-seasonally adjusted macroeconomic time series over the period $1995(1) - 2018(10)$, with data in broad categories including (a) output; (b) the labour market; (c) consumption, orders and inventories; (d) prices; (e) money and credit; (f) interest rates and exchange rates; and (g) stock prices. The list of variables is reported in table 4 but we exclude unemployment and the number of unemployed people per vacancy due to collinearity with the dependent variable, resulting in 70 variables.[3] Data transformations to logs are applied in many cases, but we do not apply differencing to I(0) representation initially, and we denote the variables as $z_{j,t}$, $j = 1, \ldots, 70$ where $z_{j,t}$ may be I(0) or I(1) and/or subject to outliers and breaks.

### 5.2.2 Google Trends time series

We have monthly time series from $2004(1) - 2019(4)$ on Google Trends data for 100 search terms relating to UK unemployment. To determine the list of top 100 search queries we obtained the most related search queries corresponding to the search term "unemployment". This included the top 25 searches that users did when they also searched "Unemployment" and the rising 25 searches when also searched "unemployment". The top searches refer to the most popular search queries. Scoring is on a

[3]The database is not fully comprehensive as some variables such as housing starts and inventories are only available at a quarterly frequency.

relative scale where a value of 100 is the most commonly searched query, 50 is a query searched half as often as the most popular query, and so on. The rising catagory includes queries with the biggest increase in search frequency since the last time period. We also apply this to the 25 top and 25 rising related topics to "unemployment". We deleted any overlapping terms and series with zero variation over the time period, replacing with additional correlated search terms. The resulting 100 time series are given in table 5, and are denoted $w_{j,t}$ for $j = 1, \ldots, 100$ below. The time series data are recorded on a scale from 0 to 100, with 100 representing the highest proportion for the terms queried within the selected region and time frame and zero the lowest. Figure 8 records a small selection of the time series as an example. Although the data are bound between 0 and 100, the data have differing properties including $\mathsf{I}(0)$ and $\mathsf{I}(1)$ for the sample given and with many structural breaks and seasonal patterns.

### 5.2.3 Model specification

The GUM is specified as:

$$\begin{aligned} Ur_t &= \underset{\{1\}}{[\beta_0]} + \sum_{i=1}^{13} \underset{\{13\}}{\beta_{u,i} Ur_{t-i}} + \underset{\{1\}}{\left[\beta_{r,0} Rr_t\right]} + \sum_{i=1}^{13} \underset{\{13\}}{\beta_{r,i} Rr_{t-i}} + \sum_{j=1}^{100}\sum_{i=0}^{13} \underset{\{100\times 14\}}{\beta_{wj,i} w_{j,t-i}} \\ &\quad + \sum_{j=1}^{70}\sum_{i=0}^{13} \underset{\{70\times 14\}}{\beta_{zj,i} z_{j,t-i}} + \sum_{k=1}^{T} \underset{\{159\}}{\delta_{1,k} 1_{\{k=t\}}} + \sum_{k=2}^{T-1} \underset{\{157\}}{\delta_{s,k} S_{\{k\geq t\}}} + \underset{\{11\}}{[\text{seasonals}]} + \epsilon_t \end{aligned} \tag{24}$$

for $T = 2006(2) - 2019(4)$, with subscript $\{\cdot\}$ denoting the number of regressors included and $[\cdot]$ denoting variables that are fixed in the selection algorithm. The intercept, seasonals and $Rr_t$ are fixed as deterministic terms or theory variables (see Hendry and Johansen, 2015), and we also fix the retained steps in (22), $S_{2009(12)}$ and $S_{2013(11)}$, to maintain comparability with (23) while not guaranteeing their significance, resulting in $N = 2735$ regressors for 159 observations.

Applying selection using *Autometrics* at $\alpha = 0.0001$ (implying less than 0.3 irrelevant variables would be retained under the null on average, and under the alternative, variables with a non-centrality of 4 would have a 50–50 chance of being retained) resulted in a terminal model being selected with $\widehat{\sigma} = 0.053$. Further reductions were applied given a marginal joint test of restrictions with $\chi^2(3) = 11.2^*$.[4] The resulting model is well-specified, passing all diagnostic tests, and is theoretically interpretable:

$$\begin{aligned} \widehat{Ur}_t &= \underset{(0.054)}{0.098} + \underset{(0.009)}{0.967} Ur_{t-1} + \underset{(0.004)}{0.014} Rr_t - \underset{(0.682)}{1.62}\, \Delta vacancies_t - \underset{(0.99)}{7.50}\, \Delta\Delta_{12} inactive_t \\ &\quad - \underset{(0.139)}{1.39}\, \Delta HoursUsual_{t-1} - \underset{(0.030)}{0.081} S_{2009(12)} + \underset{(0.028)}{0.102} S_{2013(11)} + \text{seasonals} \\ &\quad \widehat{\sigma} = 0.066;\ \ \mathsf{F}_{\mathsf{AR}}(7,133) = 0.81;\ \ \mathsf{F}_{\mathsf{ARCH}}(7,145) = 0.30;\ \ \chi^2_{\mathsf{nd}}(2) = 0.53; \\ &\quad \mathsf{F}_{\mathsf{Het}}(23,135) = 1.15;\ \ \mathsf{F}_{\mathsf{Reset}}(2,138) = 0.16;\ \ T = 2006(2) - 2019(4). \end{aligned} \tag{25}$$

Despite searching over 2735 regressors, the final model is parsimonious, with a reduction of $\widehat{\sigma}$ from 8% to 6.6% when commencing from the larger information set. The theory variable $Rr_t$ was fixed

[4] The additional restrictions include (a) replacing $vacancies_t$ and $vacancies_{t-1}$ with $\Delta vacancies_t$ [$\chi^2(1) = 3.42(p = 0.06)$]; (b) $HoursUsual_{t-1}$ and $HoursUsual_{t-2}$ with $\Delta HoursUsual_{t-1}$ [$\chi^2(1) = 2.11(p = 0.15)$]; and (c) $inactive_t$, $inactive_{t-1}$, $inactive_{t-12}$ and $inactive_{t-13}$ with $\Delta\Delta_{12} inactive_t$ [$\chi^2(1) = 0.31(p = 0.57)$], with $p-$values reported in parentheses.

in selection but is significant with $|\widehat{\mathrm{t}}| = 3.6$. Vacancies, usual weekly hours of work and the number of people who are economically inactive are all retained but enter in differences and therefore do not contribute to the long-run solution. Increasing numbers of vacancies reduces the unemployment rate contemporaneously. There is a seasonal effect of inactivity, with the change in the annual difference of number of economically inactive mattering for unemployment. Increasing usual hours worked in the previous month reduces the unemployment rate. The two step indicators retained in the theory model are significant although with smaller coefficients than in (22), but the dummy for 2011 is no longer needed. None of the Google Trends series are retained. The resulting long-run solution (excluding seasonals and difference terms which enter as short-run dynamics, adjusting the intercept to give a zero mean given the excluded differenced terms) is given by:

$$\widehat{ecm}^* = Ur - \underset{(0.95)}{5.29} - \underset{(0.10)}{0.44}\,Rr + \underset{(0.52)}{2.49}\,S_{2010(1)} - \underset{(0.41)}{3.15}\,S_{2013(12)}, \tag{26}$$

also reported in figure 5. The mean of unemployment is similar at $5.3\%$, but the coefficient on $Rr$ has fallen from $0.76$ to $0.44$. This has a noticable effect during 2009/10 when the theory model $\widehat{ecm}$ falls by more than double that of the $\widehat{ecm}^*$, reflecting the substantial rise in $Rr$ over this period due to the fall in GVA.

### 5.3 Transforming the model to a stationary representation

Next we use the $\widehat{ecm}^*$ obtained from the big data selection specification to re-parameterize to a stationary specification for the monthly change in the unemployment rate. We could apply selection to the $\mathsf{I}(0)$ equilibrium correction representation of (24) applying selection at a tight significance level again. However, given our well-specified model in levels, in which all macroeconomic and Google Trends data were searched over, we directly compute the equilibrium correction specification from the selected model (25), given by:

$$\begin{aligned}\widehat{\Delta Ur}_t \;=\;& -\underset{(0.019)}{0.116} - \underset{(0.007)}{0.027}\widehat{ecm}^*_{t-1} + \underset{(0.008)}{0.028}\Delta Rr_t - \underset{(0.997)}{7.34}\,\Delta\Delta_{12} inactive_t \\ & - \underset{(0.683)}{1.69}\,\Delta vacancies_t - \underset{(0.137)}{1.41}\,\Delta HoursUsual_{t-1} + \text{seasonals} \\ & \widehat{\sigma} = 0.066;\;\; \mathsf{F}_{\mathsf{AR}}(7,135) = 1.09;\;\; \mathsf{F}_{\mathsf{ARCH}}(7,145) = 0.14;\;\; \chi^2_{\mathsf{nd}}(2) = 0.14; \\ & \mathsf{F}_{\mathsf{Het}}(21,137) = 1.20;\;\; \mathsf{F}_{\mathsf{Reset}}(2,140) = 2.93;\;\; T = 2006(2) - 2019(4)\end{aligned} \tag{27}$$

The model fit, scaled residuals, residual density and residual autocorrelation for equation (27) are recorded in figure 6. The model captures much of the movement in the monthly change in the unemployment rate with an $\widehat{R}^2$ of $0.83$, and the resulting $\widehat{\sigma} = 0.066$. As a check, the variables in the model were fixed and IIS was applied to (27) at $\alpha = 0.001$ but no impulse indicators were retained. The $\widehat{ecm}^*$ is significant with a speed of adjustment of $-0.027$ on monthly data which cumulates to a $32\%$ adjustment to equilibrium per year. The short-run effects discussed in the ADL specification are highly significant and the profits proxy $Rr$ has a short-run impact as well as entering in the long-run equilibrium. Despite commencing from a 'big data' specification, and covering a period of turbulence over the financial crisis and great recession, selection has located a congruent, parsimonious and theoretically interpretable model.

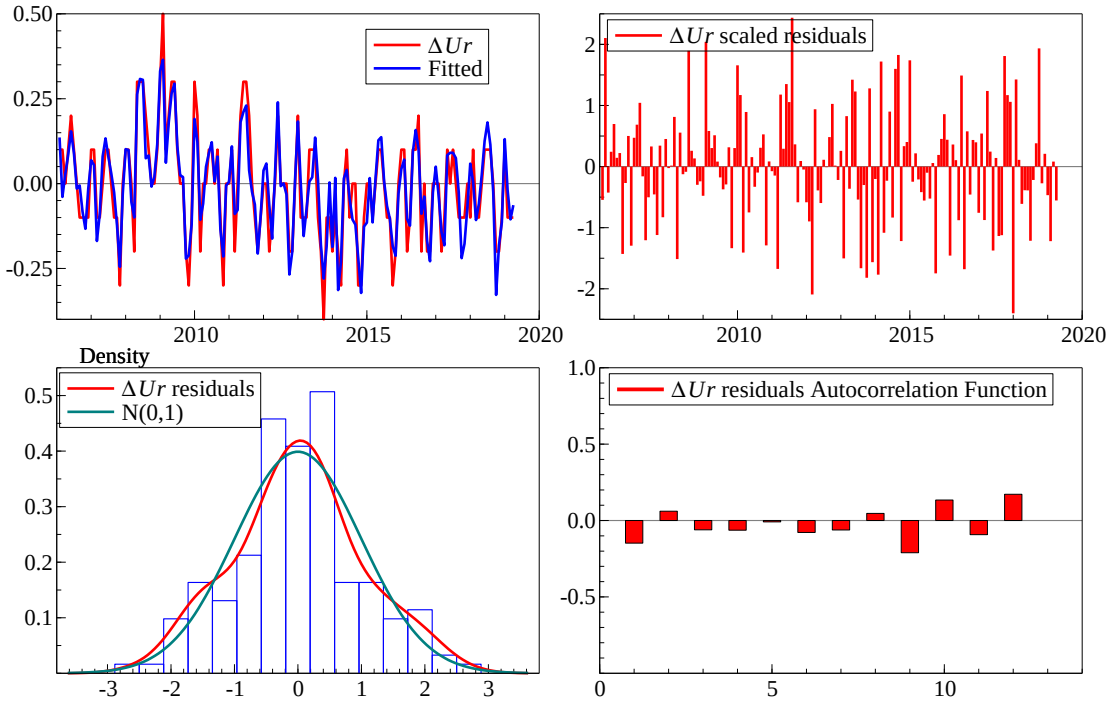


Figure 6: Model fit, residuals, residual density and residual autocorrelation for equation (27).

## 5.4 Using principal components to obtain cointegrating relations

We next consider whether the use of principle components can help to isolate the cointegrating relations, following §3. Define:

$$X_t = [Ur_t; Rr_t, z_{1,t} \ldots, z_{70,t}]$$

where the macroeconomic variables are retained in levels (with, in many cases, the log transformation applied). The principal components of $\mathbf{X}$ are extracted using the method outlined in §3.2.1 where the correlation matrix $\widehat{\mathbf{\Omega}}$ is estimated over $T = 2005(1) - 2019(4)$ to allow for lags of the principal components in the GUM.[5] The resulting 72 principal components are compared to the $\widehat{ecm}^*$ obtained in (26) to check whether any of the factors isolate the cointegrating relations. Figure 7 records the correlations between the principal components and $\widehat{ecm}^*$, with the ordered principal components along the horizontal axis; the largest correlation is $0.52$. It is clear that factor analysis is not able to capture long-run relations in the data when the data matrix includes lots of variables, many of which are non-stationary and irrelevant.

[5]Alternative methods for estimating the variance-covariance matrix to apply principal components could be used. For example, Croux, Filzmoser, and Fritz (2013) propose a method for estimating a robust sparse variance-covariance matrix which could be used to set some of the eigenvector weights to zero before computing the principal components.

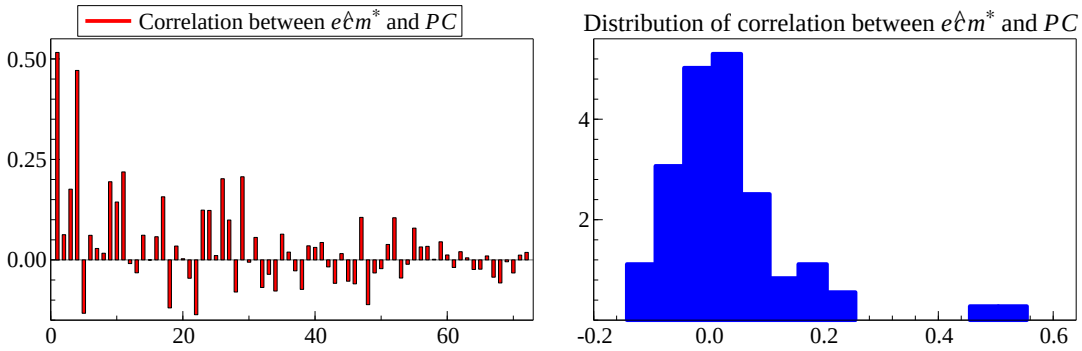


Figure 7: Panel (a): correlation between principal components and $\widehat{ecm}^*$ from (26); panel (b): histogram of correlation coefficients between principal components and $\widehat{ecm}^*$.

### 5.5 Big data: Over 3000 variables with 159 observations.

Finally, we tackle an extreme case of Big Data, including 12 lags of the macroeconomic principal components along with the macroeconomic variables and Google Trends data. We specify the GUM in I(0) space to reduce the degree of collinearity which increases the speed of the selection path search. The GUM is given by:

$$\begin{aligned}\Delta Ur_t \;=\; & \underset{\{1\}}{[\gamma_0]} + \sum_{i=1}^{12} \underset{\{12\}}{\gamma_{ui}} \Delta Ur_{t-i} + \underset{\{1\}}{\left[\gamma_{ecm} \widehat{ecm}^*_{t-1}\right]} + \sum_{i=0}^{12} \underset{\{13\}}{\gamma_{ri}} \Delta Rr_{t-i} + \sum_{j=1}^{100} \sum_{i=0}^{12} \underset{\{100\times 13\}}{\gamma_{wji}} \Delta w_{j,t-i} \\ & + \sum_{j=1}^{70} \sum_{i=0}^{12} \underset{\{70\times 13\}}{\gamma_{zji}} \Delta z_{j,t-i} + \sum_{j=1}^{72} \sum_{i=0}^{12} \underset{\{72\times 13\}}{\gamma_{pcji}} pc_{j,t-i} + \sum_{k=1}^{T} \underset{\{159\}}{\delta_k} 1_{\{k=t\}} + \underset{\{11\}}{[\text{seasonals}]} + \epsilon_t,\end{aligned} \tag{28}$$

The $z_{j,t}$ and $w_{j,t}$ are differenced where appropriate to remove unit roots, see table 4 in the Appendix for transformations. This leads to $N = 3343$ variables (of which 13 are fixed), but selection at $\alpha = 0.0001$ takes only 2.48 seconds. With an initial search space of $2^8$, search resulted in two terminal models differing only by lag length. The final model retained is similar to (27), with no Google Trends, impulse indicators, or principal components retained.

$$\begin{aligned}\widehat{\Delta Ur}_t \;=\; & \underset{(0.020)}{-0.002} - \underset{(0.006)}{0.028} \widehat{ecm}^*_{t-1} + \underset{(0.007)}{0.022} \Delta Rr_t - \underset{(0.311)}{0.384} \Delta awe_t - \underset{(1.53)}{17.7} \Delta inactive_t \\ & + \underset{(1.03)}{3.27} \Delta inactive_{t-12} - \underset{(0.14)}{2.14} \Delta HoursUsual_{t-1} + \text{seasonals} \\ & \widehat{\sigma} = 0.056; \;\; \mathsf{F}_{\mathsf{AR}}(7,134) = 2.57^*; \;\; \mathsf{F}_{\mathsf{ARCH}}(7,145) = 0.74; \;\; \chi^2_{\mathsf{nd}}(2) = 0.20; \\ & \mathsf{F}_{\mathsf{Het}}(23,135) = 1.31; \;\; \mathsf{F}_{\mathsf{Reset}}(2,139) = 2.21; \;\; T = 2006(2) - 2019(4)\end{aligned} \tag{29}$$

The change in the log of average weekly earnings is retained despite it being insignificant as its removal leads to some evidence of residual autocorrelation. $\Delta inactive_t$ and $\Delta inactive_{t-12}$ were both retained but tests of restrictions to combine to a second difference as applied in (25) were rejected. The model has

an equation standard error of $\widehat{\sigma} = 0.056$. $\widehat{ecm}^*_{t-1}$ is highly significant ($|\widehat{t}| = 5.0$) with a similar speed of adjustment to the previous model where over $30\%$ of disequilibria is corrected per annum, along with short-run effects including usual hours worked and the number of people who are inactive in the labour market. Despite commencing from over three thousand possible regressors we obtain a well-specified model with just $18$ parameters to estimate. Thus, cointegration and shifts can be handled and model selection is not hindered by commencing from large datasets.

# 6 Conclusions

There are many important requirements of any procedure searching for a data-based relationship using non-stationary big data. The paramount considerations include embedding all candidate variables in general initial models, which clearly favours big data; using high quality data on all variables, which could be problematic for some sources of big data; enforcing very tight significance levels to avoid an excess of 'false positives' when $N$ is large at a cost in not selecting 'low significance' effects; applying an effective and computationally feasible selection procedure; handling both stochastic trends by cointegration or differencing and location shifts by saturation estimation; and testing the specification of putative models. Some theory-based insights can also help if not merely imposed but retained as part of the formulation.

We examine the ability of principal components to detect cointegrating relations in a single-equation simulation exercise. The simulations highlighted the problems with using a factor approach to modelling cointegration. The cointegrating relations can be identified accurately by principal components in some specific settings but the structure of the data matrix (both its dimension and correlation structure) affects which principal component detects the cointegrating relations and the accuracy with which it does so. Contamination of the variance-covariance matrix with many irrelevant variables (likely in big data settings) results in the principal components approach being unable to identify the cointegrating relations.

The approach outlined in Doornik and Hendry (2015) suggests that big data can be handled using selection as the method of regularization. It is the only approach that focuses on congruent and encompassing model specifications for big data, and can be used for I(1) cointegrated levels, as demonstrated by the empirical example modelling UK unemployment. Therefore big data, if handled carefully, can be an asset and does not jeopardize the chances of locating a good model.

# References


Athey, S. (2018). *The Impact of Machine Learning on Economics*, pp. 507–547. University of Chicago Press.

Bai, J. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics 122*, 137–183.

Banerjee, A. and M. Marcellino (2009). Factor augmented error correction models. pp. 227–254. Oxford: Oxford University Press.

Banerjee, A., M. Marcellino, and I. Masten (2016). An overview of the factor-augmented error-correction model. In Hillebrand, E. and Koopman, S. J. (Ed.), *Dynamic Factor Models*, Chapter 1, pp. 3–41. Bingley: Emerald Publishing.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics 120*, 387–422.

Caceres, C. (2007). Asymptotic properties of tests for mis-specification. Unpublished doctoral thesis, Economics Department, Oxford University.

Castle, J. L., M. P. Clements, and D. F. Hendry (2013). Forecasting by factors, by variables, by both, or neither? *Journal of Econometrics 177(2)*, 305–319.

Castle, J. L., J. A. Doornik, D. F. Hendry, and F. Pretis (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics 3(2)*, 240–264.

Castle, J. L. and N. Shephard (Eds.) (2009). *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press.

Clements, M. P. and D. F. Hendry (2006). Forecasting with breaks. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Econometrics on Forecasting*, pp. 605–657. Amsterdam: Elsevier.

Croux, C., P. Filzmoser, and H. Fritz (2013). Robust sparse principal component analysis. *Technometrics 55(2)*, 202–214.

Dijkstra, T. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics 22*, 67–90.

Doornik, J. A. (2009). Autometrics. See Castle and Shephard (2009), pp. 88–121.

Doornik, J. A. and H. Hansen (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics 70*, 927–939.

Doornik, J. A. and D. F. Hendry (2015). Statistical model selection with big data. *Cogent Economics and Finance*, DOI:10.1080/23322039.2015.1045216.

Doornik, J. A. and D. F. Hendry (2018). *Empirical Econometric Modelling using PcGive: Volume I.* (8th ed.). London: Timberlake Consultants Press.

Duffy, J. A. and D. F. Hendry (2017). The impact of near-integrated measurement errors on modelling long-run macroeconomic time series. *Econometric Reviews 36*, 568–587.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity, with estimates of the variance of United Kingdom inflation. *Econometrica 50*, 987–1007.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized factor model: Identification and estimation. *The Review of Economics and Statistics 82*, 540–554.

Godfrey, L. G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica 46*, 1303–1313.

Harbo, I., S. Johansen, B. Nielsen, and A. Rahbek (1998). Asymptotic inference on cointegrating rank in partial systems. *Journal of Business and Economic Statistics 16(4)*, 388–399.

Harris, D. (1997). Principal components analysis of cointegrated time series. *Econometric Theory 13*, 529–557.

Hendry, D. F. (2001). Modelling UK inflation, 1875-1991. *Journal of Applied Econometrics 16*, 255–275.

Hendry, D. F. (2004). The Nobel Memorial Prize for Clive W.J. Granger. *Scandinavian Journal of Economics 106*, 187–213.

Hendry, D. F. and J. A. Doornik (2014). *Empirical Model Discovery and Theory Evaluation*. Cambridge, Mass.: MIT Press.

Hendry, D. F. and S. Johansen (2015). Model discovery and Trygve Haavelmo's legacy. *Econometric Theory 31*, 93–114.

Hendry, D. F., S. Johansen, and C. Santos (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics 33*, 317–335. Erratum, 337–339.

Hendry, D. F. and G. E. Mizon (2011). Econometric modelling of time series with outlying observations. *Journal of Time Series Econometrics 3 (1)*, DOI: 10.2202/1941–1928.1100.

Hendry, D. F. and G. E. Mizon (2012). Open-model forecast-error taxonomies. In X. Chen and N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 219–240. New York: Springer.

Johansen, S. (1992). Determination of cointegration rank in the presence of a linear trend. *Oxford Bulletin of Economics and Statistics 54*, 383–398.

Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Johansen, S., R. Mosconi, and B. Nielsen (2000). Cointegration analysis in the presence of structural breaks in the deterministic trend. *Econometrics Journal 3*, 216–249.

Johansen, S. and B. Nielsen (2009). An analysis of the indicator saturation estimator as a robust regression estimator. See Castle and Shephard (2009), pp. 1–36.

Kurita, T. and B. Nielsen (2019). Partial cointegrated vector autoregressive models with structural breaks in deterministic terms. *Econometrics 7*, 42: https://doi.org/10.3390/econometrics7040042.

Lansangan, J. R. G. and E. B. Barrios (2008). Principal components analysis of nonstationary time series data. *Statistics and Computing 19*(2), 173–187.

Pretis, F. (2020). Econometric models of climate systems: The equivalence of two-component energy balance models and cointegrated vector autoregressions. *Journal of Econometrics 214*, 256–273.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B, 31*, 350–371.

Selby, M. S. (Ed.) (1970). *Handbook of Tables for Mathematics*. Cleveland, Ohio: The Chemical Rubber Co.

Sims, C. A., J. H. Stock, and M. W. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica 58*, 113–144.

Snell, A. (1999). Testing for r versus r-1 cointegrating vectors. *Journal of Econometrics 88*, 151–191.

Stock, J. H. and M. W. Watson (1998). Diffusion indexes. Working paper No. 6702, NBER.

Stock, J. H. and M. W. Watson (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

Swanson, N. R. and W. Xiong (2018). Big data analytics in economics: What have we learned so far, and where should we go from here? *Canadian Journal of Economics 51*(3), 695–746.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*, 3–28.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica 48*, 817–838.

Wooldridge, J. M. (1999). Asymptotic properties of some specification tests in linear models with integrated processes. In R. F. Engle and H. White (Eds.), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive W.J. Granger*, pp. 366–384. Oxford: Oxford University Press.

Zhao, X. and P. Shang (2016). Principal component analysis for non-stationary time series based on detrended cross-correlation analysis. *Nonlinear Dynamics 84*(2), 1033–1044.

# 7 Appendix

| Label | Description | Source: Code |
|---|---|---|
| $Ur$ | ILO Unemployment rate for UK: All aged 16 & over (NSA) | ONS: MGUK |
| $Y$ | Chained volume index of gross value added | ONS: MGDP |
| $P$ | Consumer price index, all items (2015=100) | ONS: D7BT |
| $R_l$ | Long-Term Government Bond Yields, 10-year, %, (monthly) | FRED: IRLTLT01GBM156N |
| $z_i$ | Macroeconomic Variables | See table 4 |
| $w_i$ | Google Trends data, $i = 1, \ldots, 100$ | See table 5 |
| $pc_i$ | Principal components of $z_i$ excluding 16 and 17 in table 4 | |

Table 3: Lower cases represent logs and $\Delta x_t$ represents the monthly change in $x_t$. NSA = not seasonally adjusted.

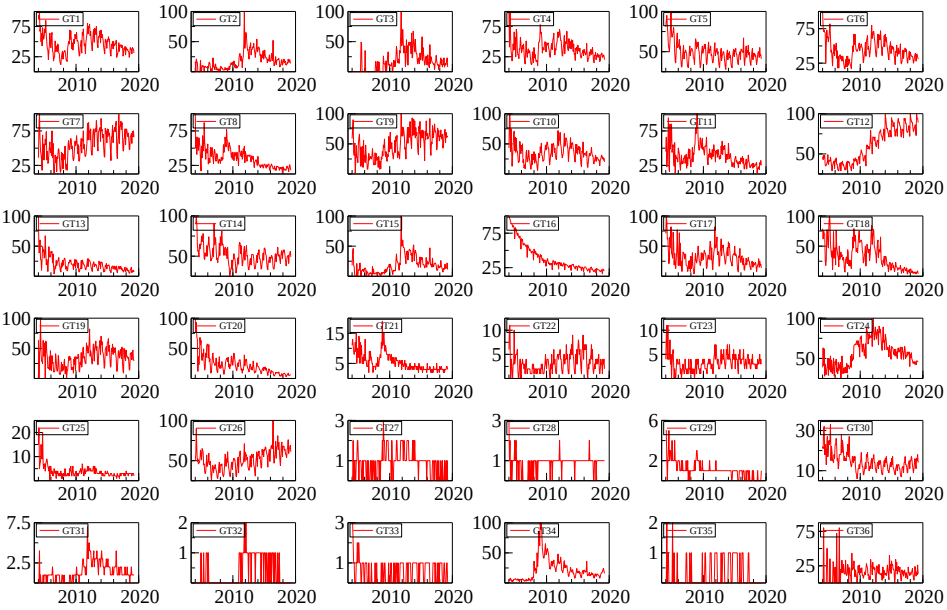


Figure 8: A sample of 36 of the Google Trends time series obtained from the queries defined in table 5.

| | Name | Description | Source | Transformation |
|---|---|---|---|---|
| 1 | IPIntermediate | Index of Production: Intermediate Goods | ONS | 1 |
| 2 | IPCapital | Index of Production: Capital Goods | ONS | 1 |
| 3 | IPNondurables | Index of Production: Consumer durables | ONS | 1 |
| 4 | IPDurables | Index of Production: Consumer non-durables | ONS | 1 |
| 5 | IPEnergy | Index of Production: Energy | ONS | 1 |
| 6 | IPEAI | Index of Production: Engineering and Allied Industries | ONS | 1 |
| 7 | IPOther | Index of Production: Other Manufacturing | ONS | 1 |
| 8 | IPMining | Index of Production: Mining | ONS | 1 |
| 9 | IPManufacture | Index of Production: Manufacturing | ONS | 1 |
| 10 | IPProduction | Index of Production: Production | ONS | 1 |
| 11 | IOS | Index of Services | ONS | 1 |
| 12 | BCI | Business Confidence Index | OECD | 2 |
| 13 | CCI | Consumer Confidence Index | OECD | 2 |
| 14 | CLI | Composite Leading Indicator | OECD | 2 |
| 15 | vacancies | All vacancies (in thousands) Total for UK aged 16 and over | ONS | 2 |
| 16 | UnempPerVac | Number of Unemployed people per vacancy | ONS | 2 |
| 17 | U | ILO Unemployed | ONS | 1 |
| 18 | U_up6m | Unemployment up to 6 months duration | ONS | 1 |
| 19 | U_6to12m | Unemployment 6 to 12 months duration | ONS | 1 |
| 20 | U_over12m | Unemployment over 12 months duration | ONS | 1 |
| 21 | U_over24m | Unemployment over 24 months duration | ONS | 1 |
| 22 | Inactive | Economically inactive | ONS | 1 |
| 23 | Hours | Actual Weekly Hours of Work | ONS | 2 |
| 24 | HoursFull | Actual Weekly Hours of Work for Full time workers | ONS | 2 |
| 25 | HoursPart | Actual Weekly Hours of Work for Part time workers | ONS | 2 |
| 26 | HoursUsual | Total Usual weekly hours of work | ONS | 2 |
| 27 | AWE | Average Weekly Earnings | ONS | 1 |
| 28 | AWEPrivate | Average Weekly Earnings for private sector | ONS | 1 |
| 29 | AWEPublic | Average Weekly Earnings for public sector | ONS | 1 |
| 30 | CPI | Consumer Price Index | ONS | 1 |
| 31 | RPI | Retail Price Index | ONS | 1 |
| 32 | PPIManu | Producer Price Index: Manufacturing | ONS | 1 |
| 33 | PPIOutput | Producer Price Index: Output | ONS | 1 |
| 34 | CPIGoods | Consumer Price Index: Goods | ONS | 1 |
| 35 | CPIServices | Consumer Price Index: Services | ONS | 1 |
| 36 | CPIEnergy | Consumer Price Index: Energy | ONS | 1 |
| 37 | CPIDurables | Consumer Price Index: Durables | ONS | 1 |
| 38 | CPOINondurables | Consumer Price Index: Non-durables | ONS | 1 |
| 39 | CPIFood | Consumer Price Index: Food | ONS | 1 |
| 40 | CPIHousing | Consumer Price Index: Housing | ONS | 1 |
| 41 | CPITransport | Consumer Price Index: Transport | ONS | 1 |
| 42 | CPIClothing | Consumer Price Index: Clothing | ONS | 1 |
| 43 | CPIHealth | Consumer Price Index: Health | ONS | 1 |
| 44 | RPIPCE | Retail Price Index: Personal Expenditure | ONS | 1 |
| 45 | FTSE | FTSE 100 Index | YF | 2 |
| 46 | S&P500 | S&P500 Index | YF | 2 |
| 47 | DJIA | Dow Jones Industrial Average in USD | YF | 2 |
| 48 | Yield20 | 20 year Nominal Yield | BOE | 2 |
| 49 | Yield10 | 10 year Nominal Yield | BOE | 2 |
| 50 | Yield5 | 5 year Nominal Yield | BOE | 2 |
| 51 | IRhousehold | Interest rate for overdrafts to households | BOE | 2 |
| 52 | BankRate | Official Bank Rate | BOE | 2 |
| 53 | 3MEuroDollar | 3 month Euro-Dollar deposit interest rate | BOE | 2 |
| 54 | TB3M | 3 month Treasury Bill rate | BOE | 2 |
| 55 | SONIA | Sterling overnight index average lending rate | BOE | 2 |
| 56 | EER | Effective exchange rate index | BOE | 2 |
| 57 | EERBroad | Broad Effective exchange rate index | BOE | 2 |
| 58 | XRCanada | Spot exchange rate, Canadian Dollar into Sterling | BOE | 2 |
| 59 | XR_Euro | Spot exchange rate, Euro into Sterling | BOE | 2 |
| 60 | XR_Japan | Spot exchange rate, Japanese Yen into Sterling | BOE | 2 |
| 61 | XR_Swiss | Spot exchange rate, Swiss Franc into Sterling | BOE | 2 |
| 62 | XR_US | Spot exchange rate, US$ into Sterling | BOE | 2 |
| 63 | Approvals | Total sterling approvals for house purchase | BOE | 1 |
| 64 | M4 | Amounts outstanding of M4 | BOE | 3 |
| 65 | M0 | Amounts outstanding of total sterling notes and coin in circulation | BOE | 3 |
| 66 | SecuredLending | Total sterling approvals for secured lending to individuals | BOE | 1 |
| 67 | M4(Breakadj) | Break adjusted level of M4 liabilities | BOE | 3 |
| 68 | M3 | Amounts outstanding of M3 | BOE | 3 |
| 69 | RSGoods | Retail sales: household goods Index | ONS | 1 |
| 70 | RSNonfood | Retail sales: other non-foods Index | ONS | 1 |
| 71 | RSClothing | Retail Sales: Clothing and footwear index | ONS | 1 |
| 72 | RSFood | Retail sales: food drink & tobacco Index | ONS | 1 |

Table 4: Brief description of non-seasonally adjusted monthly macroeconomic data (details including source codes in the online appendix). ONS: Office for National Statistics; BOE: Bank of England; YF: Yahoo Finance; OECD: OECD Database (UK data search). Transformations for the EqCM respresentation include $1 = \Delta \log z_t$; $2 = \Delta z_t$; and $3 = \Delta^2 \log z_t$. Lower cases represent logs.

| | | | |
|---|---|---|---|
| 1 | UK unemployment | 51 | france unemployment rate |
| 2 | Youth unemployment | 52 | fiscal policy |
| 3 | Youth unemployment in the United Kingdom | 53 | uk unemployment rate 2012 |
| 4 | Unemployment in the United Kingdom | 54 | china unemployment rate |
| 5 | Economy | 55 | can i claim unemployment benefit |
| 6 | unemployment uk | 56 | Unemployment |
| 7 | unemployment rate | 57 | Rate |
| 8 | unemployment benefit | 58 | Employment |
| 9 | uk unemployment rate | 59 | inflation |
| 10 | unemployment in uk | 60 | Statistics |
| 11 | unemployment benefits | 61 | Youth unemployment |
| 12 | benefits | 62 | Youth |
| 13 | unemployment rates | 63 | Economics |
| 14 | inflation | 64 | Economy |
| 15 | youth unemployment | 65 | Employee benefits |
| 16 | employment | 66 | Gross domestic product |
| 17 | unemployment in the uk | 67 | Government |
| 18 | unemployment figures | 68 | Jobseekers Allowance |
| 19 | what is unemployment | 69 | Wage |
| 20 | unemployment statistics | 70 | Policy |
| 21 | unemployment insurance | 71 | Office for National Statistics |
| 22 | unemployment definition | 72 | Interest |
| 23 | unemployment rate in uk | 73 | Structural unemployment |
| 24 | unemployed | 74 | Immigration |
| 25 | unemployment benefit uk | 75 | Fiscal policy |
| 26 | gdp | 76 | Sole proprietorship |
| 27 | bbc unemployment | 77 | Macroeconomics |
| 28 | us unemployment | 78 | unemployment economics |
| 29 | unemployment office | 79 | frictional unemployment |
| 30 | inflation rate | 80 | uk economy |
| 31 | youth unemployment | 81 | how much is unemployment benefit |
| 32 | uk youth unemployment | 82 | claim unemployment benefit |
| 33 | definition of unemployment | 83 | reasons for unemployment |
| 34 | recession | 84 | unemployment benefit calculator |
| 35 | define unemployment | 85 | voluntary unemployment |
| 36 | current unemployment rate uk | 86 | uk youth unemployment rate |
| 37 | how much is unemployment benefit | 87 | costs of unemployment |
| 38 | spain unemployment rate | 88 | minimum wage |
| 39 | greece unemployment | 89 | minimum |
| 40 | long term unemployment | 90 | wages |
| 41 | tutor2u | 91 | minimum wage uk |
| 42 | graduate unemployment | 92 | average wage |
| 43 | england unemployment rate | 93 | living wage |
| 44 | unemployment tutor2u | 94 | what is minimum wage |
| 45 | german unemployment | 95 | national minimum wage |
| 46 | tutor2u | 96 | minimum wage uk |
| 47 | graduate unemployment | 97 | jobs |
| 48 | uk unemployment graph | 98 | vacancies |
| 49 | claiming unemployment benefit | 99 | regional unemployment |
| 50 | london unemployment rate | 100 | number unemployed |

Table 5: 100 search queries to generate the Google Trends time series data. All searches for UK over $2004(1) - 2019(4)$.