Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Dynamic factor models with clustered loadings: Forecasting education flows using unemployment data

Francisco Blasques [a,b,1], Meindert Heres Hoogerkamp [a,c,d], Siem Jan Koopman [a,b,*], Ilka van de Werve [a,b,e]

[a] School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
[b] Tinbergen Institute Amsterdam, Gustav Mahlerplein 117, 1082 MS Amsterdam, The Netherlands
[c] Dutch Ministry of Education, Culture and Science, Rijnstraat 50, 2515 XP The Hague, The Netherlands
[d] Dienst Uitvoering Onderwijs (DUO), Kempkensberg 12, 9722 TB Groningen, The Netherlands
[e] Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), De Boelelaan 1077, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Dynamic factor models
Cluster analysis
Forecasting
Education
Unemployment

ABSTRACT

We propose a dynamic factor model which we use to analyze the relationship between education participation and national unemployment, as well as to forecast the number of students across the many different types of education. By clustering the factor loadings associated with the dynamic macroeconomic factor, we can measure to what extent the different types of education exhibit similarities in their relationship with macroeconomic cycles. To utilize the feature that unemployment data is available for a longer time period than our detailed education panel data, we propose a two-step procedure. First, we consider a score-driven model which filters the conditional expectation of the unemployment rate. Second, we consider a multivariate model in which we regress the number of students on the dynamic macroeconomic factor, and we further apply the k-means method to estimate the clustered loading matrix. In a Monte Carlo study, we analyze the performance of the proposed procedure in its ability to accurately capture clusters and preserve or enhance forecasting accuracy. For a high-dimensional, nation-wide data set from the Netherlands, we empirically investigate the impact of the rate of unemployment on choices in education over time. Our analysis confirms that the number of students in part-time education covaries more strongly with unemployment than those in full-time education.

[*] Corresponding author at: School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.

E-mail addresses: f.blasques@vu.nl (F. Blasques), m.hereshoogerkamp@vu.nl (M.H. Hoogerkamp), s.j.koopman@vu.nl (S.J. Koopman), i.vande.werve@vu.nl (I. van de Werve).

[1] F. Blasques is thankful to the Netherlands Organization for Scientific Research (NWO) for financial support (VI.Vidi.195.099). We would like to thank the handling Editors, the Associate Editor and the two referees for their careful reading of earlier versions of our paper and their many insightful comments.

https://doi.org/10.1016/j.ijforecast.2021.01.026

1. Introduction

Quality education is one of the Sustainable Development Goals of the United Nations. For example, the Dutch government allocated roughly 11% of its total expenditure in the national budget of 2019 towards education.[2,3] This illustrates the importance of education to our society. To secure a reliable budgetary policy, the Dutch government forecasts the numbers of students in each type of education at a nation-wide level. Although education systems are complex, dynamic, and evolving, accurate forecasts are of key importance for overall operational and financial planning, but also for providing insights into what drives participation in education. For the purpose of fiscal policy, it is valuable to understand the interaction between education participation and macroeconomic circumstances.

Although Spijkerman (2006) did not find a strong relation between macroeconomic indicators and the total number of students, macroeconomic circumstances do seem to affect the demand for certain types of education. In particular, the share of part-time education appears to be inversely related to unemployment rates, especially in vocational education. This analysis is relevant because educational institutions receive less funding for part-time students than for fulltime students. However, whether or not distinct groups react differently to macroeconomic circumstances has not been studied in full. The more recent availability of low-level data allows us to revisit this research question. Our data set for vocational and higher education is high-dimensional on the cross-section but low-dimensional on the time series. The models for such panel data sets typically strike a balance between interpretability and performance. However, to policymakers in government and educational institutions, both interpretability and performance are of interest.

In this article we develop a dynamic factor model where unemployment rates and education participation are modeled simultaneously. The model consists of two components. First, we focus on the time series dimension and model the historical unemployment rates through a score-driven local-level model as proposed by Creal, Koopman, and Lucas (2013). By using all available observations of the unemployment rate, we get more reliable results when filtering the factor than when using the few observations that correspond to the same years as those for which the education data are also available. Second, in the cross-section dimension of our methodology, we anticipate that many education flows respond in a similar way to changes in the unemployment rate. We take the extracted dynamic economic factor as given and model the education data set effectively as a multivariate regression model. We also propose clustering their dependence on the dynamic economic factor through the parameters of the loading matrix. This imposes a structure on the model that benefits interpretation. Moreover, since we represent many education series by a couple of cluster centroids, it is also more efficient with respect to forecasting education participation.

In accordance with these two components of our dynamic factor model, we propose a two-step estimation procedure. First, in the score-driven local-level model for the unemployment rates, we estimate the static parameters by maximum likelihood and extract the dynamic economic factor. This step is important to filter out the noise and preserve the signal in economic data such as the unemployment rate. Second, given the extracted dynamic economic factor, we estimate the multivariate regression model of the education data using the method of least squares. To gain insights into what types of education respond similarly to changes in unemployment rates, we perform a cluster analysis. By using the k-means method, we are able to represent all loading matrix elements by a few cluster centroids. This ability implies that cluster analysis can support the testing of joint significance for a group of variables without pre-imposition of group compositions. At the same time, we avoid the possible insignificance of individual variables, given that the education data for each variable are limited as the time series dimension is small. Once the clusters are identified, we can provide accurate forecasts for all series in the panel.

Dynamic factor models are well suited to extract common factors from large data sets; see Bai and Ng (2002), Jungbacker and Koopman (2015) and Stock and Watson (2002) amongst others. It has become more prevalent to estimate the parameters in dynamic factor models using a step-wise approach. For example, Doz, Giannone, and Reichlin (2011) first proxy the factors by principal components to estimate the static parameters, and then use Kalman filter techniques for signal extraction and forecasting. Bräuning and Koopman (2014) take a slightly different approach: they also first use a principal component analysis to obtain a dimension reduction, but then model all relevant variables jointly in a state-space framework such that parameter estimation, signal extraction, and forecasting are done by Kalman filter methods. The approaches in both papers are parameter-driven in which the stochastic processes of the factors have their own sources of error.

Our procedure differs in adopting an observation-driven approach: we allow the factors to evolve as dynamic processes that are formulated as functions of past data. In particular, we adopt the approach taken by Creal et al. (2013) and Harvey (2013), where the dynamic specification is based on autoregressive processes with the innovations defined as score functions with respect to the predictive likelihood function. We first model the unemployment rate data as a score-driven model to filter the dynamic factor, which we then consider as given in the second step of our proposed estimation procedure to estimate the model for the education data. Unrestricted parameter estimates of the loading matrix are then clustered to cluster types of education according to their dependence on the unemployment rate. Just as Stock and Watson (2008) do, we use the k-means algorithm in the cluster analysis, although their approach differs in that

[2] https://www.rijksfinancien.nl/visuele-begroting/2018/owb/u (visual in Dutch), last accessed 2020-04-30.

[3] https://www.rijksoverheid.nl/onderwerpen/financiering-onderwijs/overheidsfinanciering-onderwijs (in Dutch), last accessed 2020-04-30.

they base it on the residuals of the dynamic factor model. In our case, we represent the large vector of unrestricted loading estimates by a much smaller vector of cluster centroids. The introduction of clustering within a dynamic factor model has more recently been explored by Alonso, Galeano, and Peña (2020), Barnichon and Mesters (2018), and Hallin and Liška (2011). Similarly, Ando and Bai (2016) incorporate the k-means procedure within the dynamic factor model, and they also adopt an estimation procedure that consists of several steps.

In an extensive simulation study we assess the overall performance of our proposed estimation procedure and the accuracy of the forecasts. We discuss the most important insights from this study. First, the estimated parameters in our dynamic factor model show very small biases, while the clustered pattern in the loading matrix is correctly identified. Second, we can report high levels of forecast accuracy. We also find that the forecasts based on the clusters perform at least as well as the unrestricted forecasts. Indeed, we present improvements in forecast accuracy when a clear clustered structure is present in the data. Even for cases where clusters are less present in the data, we do not lose much on forecasting accuracy, yet we gain much on interpretation.

The remainder of this paper is organized as follows. We describe the education data and unemployment rate data in Section 2. In Section 3 we introduce the dynamic factor model and we discuss the features of the model: the two-step estimation procedure and the clustering method for forecasting. Section 4 describes the design and results of our Monte Carlo study. We apply the methodology to our education participation data in Section 5 and show that the proposed methodology can capture important clusters in the flows across the Dutch education system. Section 6 concludes and reviews the most important findings of our study.

2. Data

In 2019, 3.7 million students were registered at Dutch educational institutions: around 1.5 million (40.2%) were in primary education, 960,000 (25.6%) in secondary education, 500,000 (13.4%) in vocational education, 460,000 (12.4%) in higher vocational education (university of applied sciences), and 300,000 (8.2%) in university. For this study, we used data from DUO, the executive agency for the Dutch Ministry of Education, Culture and Science. Student registrations are observed on October 1st each year (reference date). The primary use of this data set is the Dutch student forecasts (Ministry of Education, Culture and Science, 2020) that feed the governmental education budget. Those in the Dutch population who are not students are labeled "outside education". Using information from two consecutive reference dates, flows through the educational system are constructed. The educational data set lists the number of people in each flow from one state to another, by age and sex, in a given year. Between these dates, students might obtain a diploma. They are said to transition from an origin state to a diploma state, and then from this diploma state to a destination state in or outside of education. Each state (one of 820 education types, 173 diploma types, or no education) has up to five descriptive labels: a sector (such as vocational education, university, etc.), type (such as on-the-job training, fulltime, etc.), level (such as bachelor's, master's, etc.), direction (such as health care, economics, etc.), and grade. Not all flows between the states are viable (for example, one cannot move from university to primary education) and some have not been observed. Since 2005, 326,000 unique transitions from origin to destination have been observed. Data from 2006 to 2019 are used.

Table 1
Descriptive statistics of the average level of 1155 time series.

count	1155
mean	239.6
std	487.3
min	14.9
25%	38.3
50%	82.8
75%	206.8
max	5658.0

Flows at the lowest level are filtered, aggregated, and transformed to form the data set for this study. We are interested in the short-term relation between unemployment rates and inflow into the first year of vocational and higher education. We include all inflow from the no-education and diploma categories. The origin states are aggregated by sector, category, and type. Furthermore, to reduce noise, some age groups (those that contain fewer first-year students) are combined as follows: <17, 23–25, 26–30, 31–40, and > 40 years old. In 2019, each age bin contained a total of between 6,900 and 60,600 students. This leaves $N = 1,155$ time series. This is a short, wide panel (large N, and small T), implying that many series have to be forecasted while limited data on the time dimension are available.

The average level of the 1,155 time series is 240 persons per year, with a minimum of 14.9 (22-year-old males moving from no education to fulltime school-based level 2 vocational education in the direction of technics/science) and a maximum of 5,658 ($<$17-year-old females moving from a preparatory vocational education diploma to fulltime school-based level 4 vocational education in the direction of healthcare). Of the time series, 50% contain on average between 38.3 and 206.8 persons per year (see Table 1). To illustrate, a random sample of 20 time series is plotted in Fig. 1. The left panel presents the actual time series (on a log scale), which may not all be stationary, as some may be subject to trend behavior. Transitions with a higher number of students on average have a higher variance. To normalize, each cross-sectional unit is divided by its average level: $\tilde{y}_{it} = \Delta y_{it}/\bar{y}_i$, where y_{it} denotes the number of people in transition i at time t. We assume that the transformed variables are stationary. Using the KPSS test, we find that only the null hypothesis of mean stationarity can be rejected ($p < 0.05$) for only 15 (1.2%) time series. These are still included, since one expects some rejections under the null when many tests are conducted. Most series show positive autocorrelation at the first lags (Fig. 2). This fits the autoregressive model developed in Section 3.

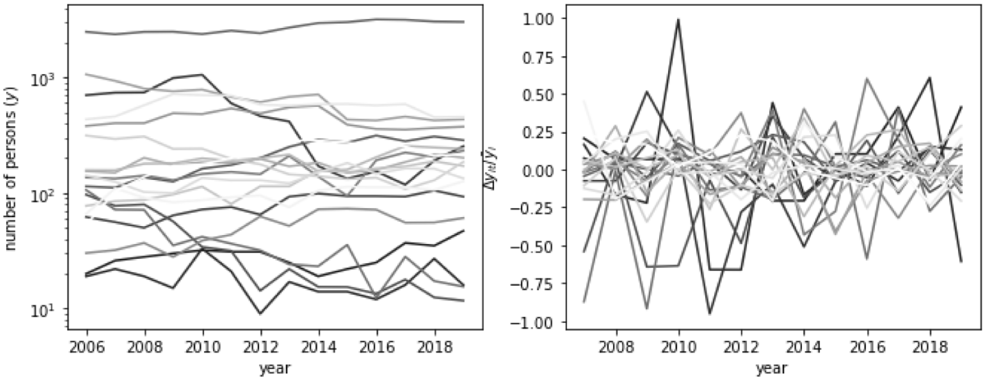


Fig. 1. Sample of 20 time series. Left: levels (log-scaled vertical axis). Right: transformed variable.

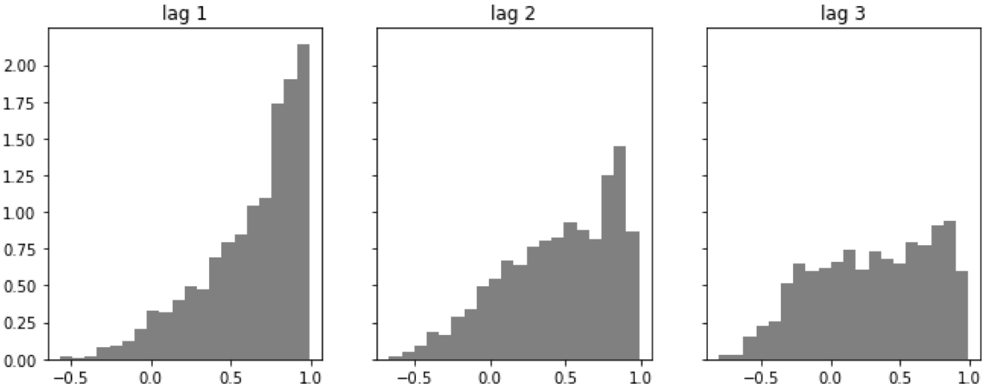
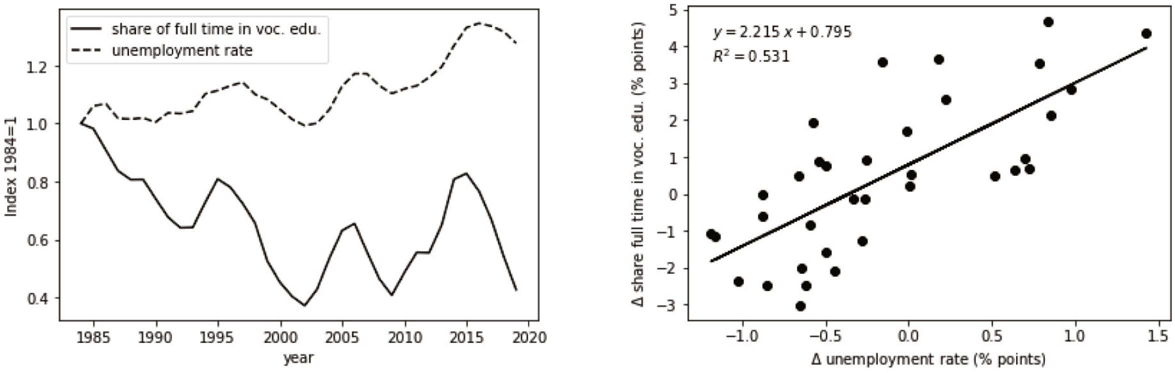


Fig. 2. Histogram of autocorrelations for all time series at the first three lags. At lag 1, most time series show positive autocorrelation. At higher lags, the estimates spread out more evenly and the mean decreases.



(a) Share of full-time in vocational education and unemployment rate (indexed at 1984=1)

(b) Increase in unemployment (x-axis) vs. share of full time in vocational education (y-axis)

Fig. 3. Share of full-time in vocational education and unemployment rate: (a) time series plot, (b) scatterplot.

We enrich the data by including historical macro-level data on the Dutch economy. Yearly unemployment rates (operationalized as the unemployed labor force divided by total labor force, aged 15–74 years) are obtained from 1970 to 2019 (Bureau for Economic Policy Analysis, 2019). In our modeling, we make use of the fact that the macroeconomic data are available for a much longer period than the education data. Fig. 3 provides the instigation for this particular research. Spijkerman (2006) did not find a strong correlation between unemployment and nationwide educational enrollment, but he did show that the share of fulltime education covaries positively with unemployment, especially in vocational studies. The left panel shows that peaks and valleys in the share of fulltime students in vocational studies indeed correspond with the labor market cycle. The right-side panel suggests a linear relationship in first differences.

In accordance with the scatterplot in Fig. 3 the first differences of both data sets are calculated. This also prevents regressing possibly (co)integrated time series. Our macro-economic time series has dimension $T_x = 49$, and the education panel has dimension $T_y = 13$, where the last T_y observations of both series refer to the same years.

3. Methodology for modeling and forecasting

Our education panel data set consists of many possible education flows, which can be selected on the basis of gender and age groups. We want to understand to what extent a macroeconomic variable (the unemployment rate) can help us to forecast the number of students for each category, for a number of years ahead, even though we only have data available for the last 13 years. The forecasting method is often regarded as more convincing when the model preserves a level of interpretability for policy purposes. We therefore develop a dynamic factor model where types of education are clustered based on their dependence on changes in the unemployment rate.

3.1. Dynamic factor modeling framework

Let y_t be the N-dimensional series of education flows in year t. The basic dynamic factor model is given by

$$y_t = \Lambda f_t + \varepsilon_t, \qquad t = 1, \ldots, T_y, \tag{1}$$

where Λ is an $N \times \ell$ loading matrix, $\ell \times 1$ vector f_t an unobserved factor, and the sequence $\varepsilon_1, \ldots, \varepsilon_{T_y}$ is independent and identically Gaussian distributed with mean zero and variance matrix $\sigma_\varepsilon^2 I_N$. Since the unobserved factor in our model is a proxy for the macroeconomic circumstances (measured by the unemployment rates), we have $\ell = 1$, implying that the loading matrix Λ is effectively a vector and f_t a scalar. A vector of unit-specific intercepts $\mu = (\mu_1, \ldots, \mu_N)'$ can be added to the model, such that we have $y_t = \mu + \Lambda f_t + \varepsilon_t$, but to facilitate the clustered forecasting method in our analysis, we assume that the data are demeaned and we have $\mu = 0$ without loss of generality.

To emphasize that the education and macroeconomic data have different time series dimensions, we introduce the following notation. The index t and time series dimension T_y correspond to the education data denoted by y_t, while index $\bar{t}$ and time series dimension T_x correspond to the unemployment rate data denoted by $x_{\bar{t}}$. Specifically, the macroeconomic data $x_{\bar{t}}$ is available at $\bar{t} \in \{1970, \ldots, 2019\}$ (so $T_x = 49$ years) while the education data y_t is available at $t \in \{2006, \ldots, 2019\}$ (so $T_y = 13$ years). The macroeconomic data are available over a longer time-span so that only the last T_y time periods coincide with the availability of the education data.

We let $x_{\bar{t}}$ be the growth in the unemployment rate in year $\bar{t}$; this time series of yearly observations has dimension T_x. The location model is given by

$$x_{\bar{t}} = f_{\bar{t}} + \xi_{\bar{t}}, \qquad \bar{t} = 1, \ldots, T_x, \tag{2}$$

where the signal $f_{\bar{t}}$ can be regarded as the time-varying mean of the observed time series $x_{\bar{t}}$, and where $\xi_1, \ldots, \xi_{T_x}$ is assumed to be an independent and identically distributed Gaussian sequence with mean zero and variance σ_ξ^2.

In our analysis, we take $x_{\bar{t}}$ as the yearly unemployment rate, which we consider to be a proxy of general macroeconomic circumstances. However, we note that we could generalize the model by allowing for multiple economic indicators (multivariate $x_{\bar{t}}$) and then still easily filter $\ell \geq 1$ factors f_t from it. This imposes extra structure on the model that requires extra assumptions to identify f_t. We leave this extension for future research.

Eq. (2) enables us to estimate the signal $f_{\bar{t}}$ in our modeling framework. We use the extracted signal for the estimation of loading parameters in Λ of Eq. (1) and for the forecasting of time series variables in y_t of Eq. (1). For the filtering of the factor, we adopt the score-driven model as introduced by Creal et al. (2013) and Harvey (2013). In an observation-driven approach, we introduce

$$\mathcal{X}_{\bar{t}-1} = \{x_1, \ldots, x_{\bar{t}-1}\} = \{\{x_1, \ldots, x_{\bar{t}-1}\}, \{f_1, \ldots, f_{\bar{t}-1}\}\},$$

so that the information set at time $\bar{t}$ is generated by $\{f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}\}$. Since the location model for $x_{\bar{t}}$ in Eq. (2) is linear Gaussian, we have $x_{\bar{t}} \sim p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)$, where $p(\cdot\,|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)$ is the univariate Gaussian density with mean $f_{\bar{t}}$, variance σ_ξ^2, and parameter vector θ. The unknown variance σ_ξ^2 is placed in the parameter vector θ, together with the unknown coefficients that we introduce below.

We follow Creal et al. (2013) in their formulation of the filtering or updating equations for $f_{\bar{t}}$; these are referred to as the generalized autoregressive score (GAS) model and are given by

$$f_{\bar{t}+1} = \omega + \sum_{i=1}^{p} \phi_i f_{\bar{t}+1-i} + \sum_{j=1}^{q} \alpha_j s_{\bar{t}+1-j}, \tag{3}$$

$$s_{\bar{t}} = S_{\bar{t}} \cdot \nabla_{\bar{t}}, \qquad S_{\bar{t}} = S(\bar{t}, f_{\bar{t}}, \mathcal{X}_{\bar{t}}),$$

$$\nabla_{\bar{t}} = \frac{\partial \log p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)}{\partial f_{\bar{t}}},$$

where ω is the intercept, $\phi_1, \ldots, \phi_p$ and $\alpha_1, \ldots, \alpha_q$ are the weight coefficients for the updating mechanism of $f_{\bar{t}+1}$, and $s_{\bar{t}}$ is the scaled score with the local score function $\nabla_{\bar{t}}$ and scaling term $S_{\bar{t}}$. Typically, we base the scaling on a variance measure of the score function. We refer to this score-driven model by GAS(p, q), where the integers $p \geq 0$ and $q \geq 0$ can be chosen on the basis of fit, residual diagnostics, and forecast performance considerations. In many cases, it is sufficient to take $p = q = 1$ and we have the GAS(1,1) model.

Given that $p(\cdot|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta)$ is the univariate Gaussian density with the time-varying mean (or location) $f_{\bar{t}}$, we have that

$$\begin{aligned} &\log p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1}; \theta) \\ &\quad = -\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log \sigma_\xi^2 - \tfrac{1}{2}(x_{\bar{t}} - f_{\bar{t}})^2/\sigma_\xi^2, \\ &\nabla_{\bar{t}} = (x_{\bar{t}} - f_{\bar{t}})/\sigma_\xi^2, \end{aligned}$$

$$S_{\bar{t}} = \mathcal{I}_{\bar{t}}^{-1} = \left(-\frac{\partial^2 \log p(x_{\bar{t}}|f_{\bar{t}}, \mathcal{X}_{\bar{t}-1};\theta)}{\partial f_{\bar{t}}^2} \right)^{-1} = \sigma_\xi^2,$$

$$s_{\bar{t}} = \sigma_\xi^2 \cdot (x_{\bar{t}} - f_{\bar{t}})/\sigma_\xi^2 = x_{\bar{t}} - f_{\bar{t}},$$

with parameter vector $\theta = \left(\omega\,,\,\phi_1\,,\,\ldots\,,\,\phi_p\,,\,\alpha_1\,,\,\ldots\,,\,\alpha_q\,,\,\sigma_\xi^2\right)'$. By placing the above elements in the filtering equation (3), it becomes apparent that the factor f_t depends on previous years of unemployment. It is further implied by the observation equation (1) that education participation is allowed to depend on past unemployment rates through the factor f_t. An alternative approach is to model these dynamics explicitly by allowing lags of the factor in the observation equation, but we leave this extension for future research.

Our dynamic factor modeling framework is represented by Eqs. (1), (2), and (3). This facilitates the linkage of the two available data sets (the education panel data y_t and the unemployment rate time series $x_{\bar{t}}$) and it provides feasible methods for parameter estimation and forecasting.

3.2. Two-step estimation procedure

We propose a two-step estimation procedure for our dynamic factor modeling framework given by Eqs. (1), (2), and (3). In the first step, we focus on the time series component and use the unemployment rate time series to estimate the score-driven model of Eqs. (2) and (3) using the method of maximum likelihood. The static parameters in $\theta = (\omega\,,\,\phi_1,\ldots,\phi_p\,,\,\alpha_1\,,\ldots,\,\alpha_q\,,\,\sigma_\xi^2)'$ are estimated via the maximization of the log-likelihood function as given by

$$\hat{\theta} = \arg\max_{\theta} T_x^{-1} \sum_{\bar{t}=1}^{T_x} \Big(-\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log\sigma_\xi^2$$
$$- \tfrac{1}{2}(x_{\bar{t}} - f_{\bar{t}})^2/\sigma_\xi^2\Big)\,,$$

where $f_{\bar{1}}$ is initialized by setting it as equal to the unconditional mean $\omega/(1-\sum_{i=1}^{p}\phi_i)$, and where $f_{\bar{t}}$ is obtained from the GAS filter (3), for a given θ and $\bar{t} = 1,\ldots,T_x$. When the maximum likelihood estimate $\widehat{\theta}$ is obtained, we denote the factors obtained from the GAS filter (3) with $\theta = \widehat{\theta}$ by $\hat{f}_{\bar{t}}$, with $\bar{t}$ as before. We consider $\hat{f}_{\bar{t}}$ as a proxy of macroeconomic circumstances.

In the second step we consider the education data, focus on the cross-section component, and view Eq. (1) as a multivariate regression model. We recall that the time series dimension of the unemployment data $x_{\bar{t}}$ is longer than of the education data y_t, so we replace the factors f_t by those estimated in the first step and only keep the last T_y periods such that $\hat{f}_t \equiv \hat{f}_{\bar{t}}$. The loadings in Eq. (1) are estimated by the method of least squares. In this way we obtain an estimate of the variance σ_ε^2 and the unrestricted estimate of the loading matrix as denoted by $\tilde{\Lambda} = (\tilde{\lambda}_1,\ldots,\tilde{\lambda}_N)'$. Next we carry out a cluster analysis on the estimated column representing the loading matrix.

Cluster analysis is often used in statistics and machine learning to partition many individuals, cities, or products into groups. As part of the second step in our estimation procedure, we propose clustering the elements of the loading matrix Λ such that it does not consist of N different elements, but of $K \ll N$ cluster centroids instead. This implies that we can analyze the similarities and differences between many types of education in their dependence on the macroeconomic data as measured by the dynamic factor. In addition, this implies that the number of unique forecasts decreases considerably.

There are several possibilities to proceed with a cluster analysis. We opted for the k-means algorithm because it provides a clear interpretation and it is computationally simple. The clustering method limits the number of forecasts that need to be produced, because we let a cluster of similar education flows be represented by one centroid. In the case of the k-means algorithm, we cluster the N different and unrestricted values of column $\tilde{\Lambda}$ into K cluster centroids for $\hat{\Lambda}$ by conducting the following steps:

1. Initialize the cluster centroids randomly from the data range: draw centroids $\delta_1,\ldots,\delta_K$ from $U(\tilde{\Lambda}_{min}, \tilde{\Lambda}_{max})$, where $U(\cdot)$ is the uniform distribution.
2. Obtain distances between data points and cluster centroids and add cluster labels to the data points: label $c^{(i)} = \arg\min_k \|\tilde{\lambda}_i - \delta_k\|^2, \quad \forall i$.
3. For each cluster, assign new centroids $\delta_k = \sum_{i=1}^{N} \mathbb{1}\{c^{(i)} = k\}\tilde{\lambda}_i \;/\; \sum_{i=1}^{N} \mathbb{1}\{c^{(i)} = k\}, \quad \forall k$.
4. Verify whether the cluster centroids changed.
5. Repeat steps 2–4 until convergence.
6. Replicate steps 1–5 for $R = 15$ different random seeds.
7. Keep the clustered loading matrix $\hat{\Lambda}$ with the shortest total distance to the cluster centroids.

In our empirical study, the number of clusters is not known beforehand. We therefore run this algorithm for several values of K and proceed heuristically by using the simple elbow method. The elbow method is a graphical representation of plotting the performance against several cluster sizes, where performance is usually measured as a distance metric. The typical pattern is a sharp decreasing distance for cluster sizes $1,\ldots,K$, with small K. For cluster sizes $K+1, K+2,\ldots$, the returns diminish. One then selects the kink or elbow in the graph as the optimal cluster size, because decreasing it would lead to a big performance loss while increasing it would lead to a negligible improvement.

The resulting small vector of cluster centroids facilitates interpretation for policymakers. It gives insights on which education flows covary similarly with changes in the unemployment rate. We want to emphasize that we do not attach any causal interpretation to it because the regressor might be endogenous owing to reverse causality. If we assume a simple labor supply model, then people spend time in either education or employment. So, it might well be that more education participation leads to higher unemployment rates. This is the reverse relationship of that in our model where we try to explain education by unemployment, so we emphasize that we measure associations in our current study.

To show the properties of the estimators, we need to take into account that we follow a two-step estimation approach. In the first step, we estimate the factor f_t by maximum likelihood, using a standard score-driven filter for the conditional expectation. Consequently, the asymptotic properties follow the usual results covered in the score-driven literature; see, for example, Blasques, Gorgi, Koopman, and Wintenberger (2018) and Blasques, Koopman, and Lucas (2014a). In the second step, we take the filtered f_t as given and attempt to estimate the unrestricted loadings in a simple linear regression model where the filtered f_t is already observed. The asymptotics of the second step estimator then follow standard regression conditions. The estimated unrestricted loadings are then clustered using the k-means algorithm. In Appendix A, we discuss the asymptotic properties in more detail. The finite-sample performance of the k-means clustering is analyzed in the Monte Carlo study below. We leave the theoretical characterization of the estimators in our two-step method for future research.

3.3. Clustered forecasting method

After the two-step estimation procedure, we use the estimated static parameters and filtered time-varying factor of the score-driven model to forecast future values of the factor in a recursive manner from Eq. (3). Next, together with the estimated clustered loading matrix, we forecast future education participation. Since the clustered loading matrix $\hat{\Lambda}$ consists of only K distinct values, we just need to forecast $K \ll N$ series. This saves computation time because we have thousands of education flows for each combination of age and gender.

4. Monte Carlo study

We carried out a Monte Carlo study to verify the performance of our estimation and measurement methodology. For this study, we considered the dynamic factor model given by Eqs. (1)–(3) with the GAS updating orders $p = 1$ and $q = 1$, that is, the GAS(1,1) specification. The dimensions in our simulation design were motivated by the empirical problem at hand. Hence, we set the cross-section dimension of y_t to be relatively large and the time series dimension to be relatively small. In particular, $N = 500$ and $T_y \in \{10, 20\}$. We set the time series dimension of $x_{\tilde{t}}$ to be moderate, with $T_x \in \{50, 100\}$. The last T_y time units of $x_{\tilde{t}}$ are equal to the T_y time units of y_t. To verify the forecasting performance, we set the forecast horizon to be $F = 3$ periods ahead. Moreover, we took the static parameters in the score-driven model as $\sigma_\xi = 1, \omega = 0.3, \phi = 0.95$ and $\alpha = 0.1$, making the unconditional mean of the process for the stationary factors equal to $\omega/(1-\phi) = 6$. The five equally sized clusters of the loading matrix have centroids 4, 11, 19, 23, and 35. This implies that the observations roughly vary between $4 \times 6 = 24$ and $35 \times 6 = 210$. Finally, we set the error variance to follow from $\sigma_\varepsilon = 20$. The reported Monte Carlo results in our study are based on $M = 1000$ simulations, for the different model specifications and data dimensions.

The specific choices of the static parameters and cluster centroids are for illustrative purposes. We do need to assume that there is some underlying form of clustering in the data present and we can visualize that by taking a non-zero unconditional mean of the factors and distinct choices of the centroids. However, the behavior over time can vary between the series within a cluster. To visualize that, we present a couple of example time series with varying variance σ_ε^2 in Fig. 4 for $T_x = 100, T_y = 10$, two clusters with loadings 23 and 35, and the remaining settings as above.

In these plots we see the trade-off when clustering is beneficial and when it is not. For a very small variance, such as the $\sigma_\varepsilon^2 = 10$ in the top-left plot, there is no real need for imposing the cluster analysis. By basically scanning over such data plots, one already gains the knowledge on patterns in the data. Practically, as there is basically no variation over time, the unrestricted estimate will be as good as the clustered one. At the other extreme, where the variance is very large, as in the bottom-right plot with $\sigma_\varepsilon^2 = 1000$, clustering is also not useful because there are no clusters to really distinguish. The large variation over time will already make it challenging to derive an unrestricted estimate and the cluster classification is therefore also difficult. More reasonable values in between, such as the plots with $\sigma_\varepsilon^2 \in \{60, 400\}$, show where extracting the clustered pattern can be of added value. With some variation over time, the unrestricted estimates may be a bit too far off in either direction for each of the series within a cluster. However, this is offset by using clustered estimates. In such cases, one might think that the unrestricted estimates are very different, but the clustered estimate clearly shows that their behavior is actually similar. In the discussion of the full simulation study below, we continue with $\sigma_\varepsilon^2 = 400$.

4.1. Parameter estimation results

The performance of the two-step estimation procedure can be visualized by densities of the estimated parameters and cluster centroids. We judge the clustering classification by confusion matrices.[4] Furthermore, we give in-sample statistics to compare the fit of the unrestricted and clustered models.

Figs. 5 and 6 give the density plots of the estimated static parameters and cluster centroids. For both figures we fix $T_y = 10$ and first consider $T_x = 50$ and then $T_x = 100$. We also obtained these results for $T_y = 20$ with $T_x \in \{50, 100\}$, but since the results were very similar, we provide them in Appendix. In each figure, the plotted densities of the estimated static parameters are given in the first set of results and the plotted densities of the cluster centroids in the loading matrix in the second set of results, all based on $M = 1000$ simulations.

If we focus on the first step of our proposed estimation procedure, then we consider the static parameter vector $(\sigma_\xi, \omega, \phi, \alpha)'$. Only the time series dimension T_x is of importance for these parameters of the score-driven model. Comparing the corresponding densities in Figs. 5 and 6 (or, similarly, comparing Figures B.1 and B.2 in the

[4] A confusion matrix gives insight on the correct assignment to the clusters, instead of the specific values of the centroids. It gives the counts of correct and incorrect assignments to the clusters with the smallest, second-to-smallest, …, and largest centroids. Perfect classification would give a diagonal matrix.

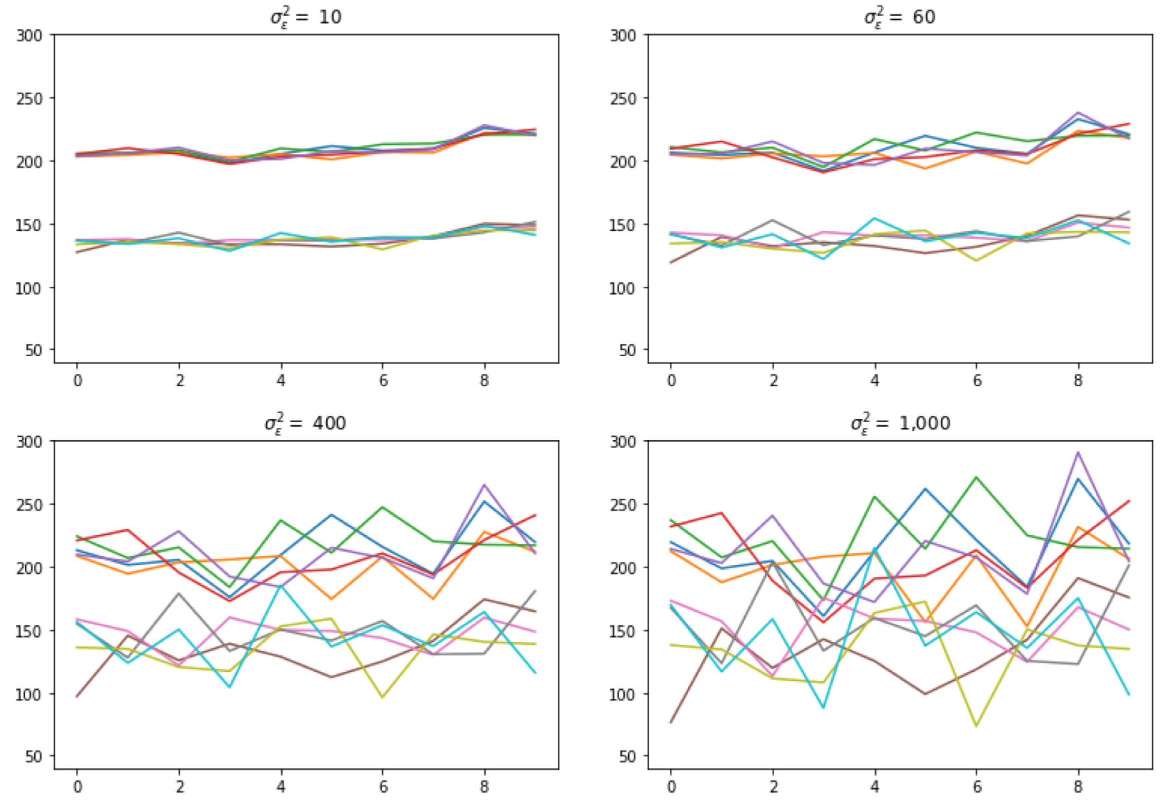
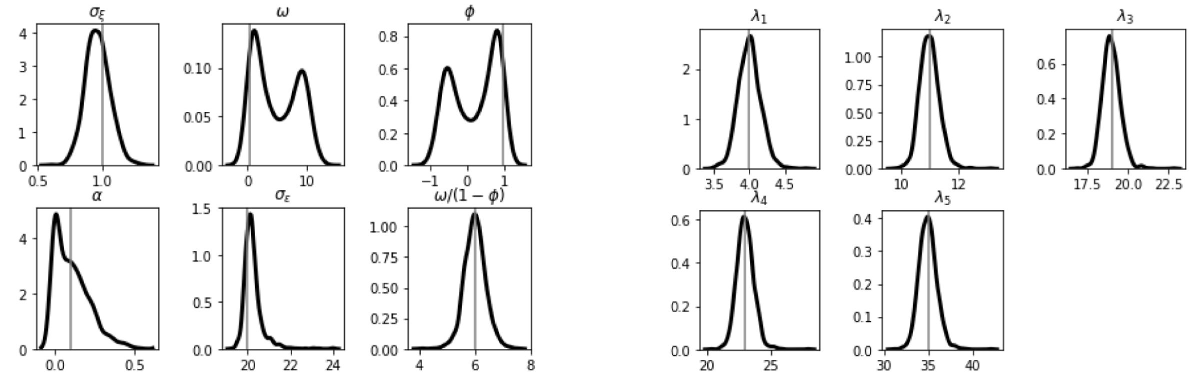


Fig. 4. Example time series y_t plotted against time. The same five time series from two clusters with loadings 23 and 35 are given in all plots. Only the variance σ_ε^2 differs over the plots. The other simulation settings are fixed ($T_x = 100, T_y = 10, N = 500, \sigma_\xi = 1, \omega = 0.3, \phi = 0.95, \alpha = 0.1$).



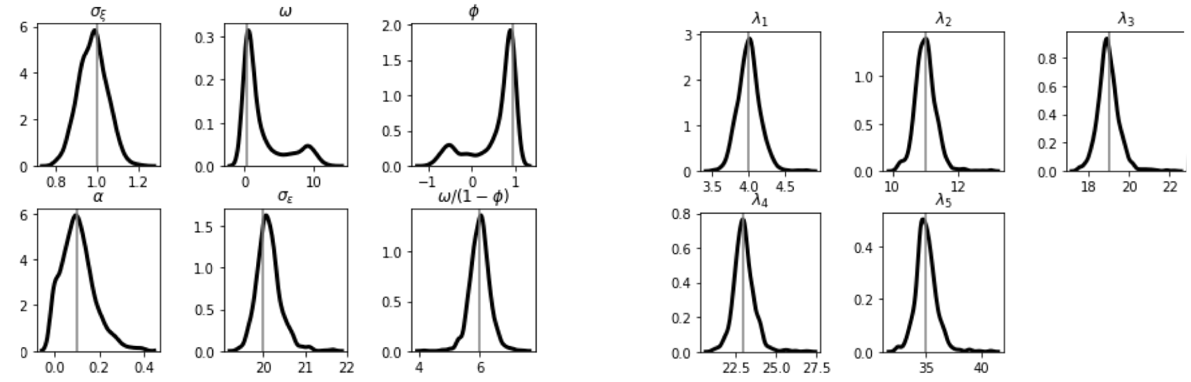
(a) Densities of estimated static parameters.

(b) Densities of estimated cluster centroids.

Fig. 5. Parameter estimation results of $M = 1000$ simulations for $N = 500, T_y = 10, T_x = 50$. Vertical lines represent true values.

Appendix, since the varying time series dimension T_y is not relevant in this step) clearly shows that the parameter estimates become much more precise as the time series dimension T_x increases. We obtain more improvements when we take T_x even larger, but such cases do not match our empirical study, so we do not consider this further.

It is apparent from the density plots for $T_x = 50$, but also for $T_x = 100$ to some extent, that the densities of ω and ϕ appear to be bimodal. Since the density plot of the unconditional mean $\omega/(1-\phi)$ is unimodal around the true value, this suggests that in small samples it is challenging to empirically separate the two parameters ω and ϕ. When the time series dimension increases, our results show that the bimodality vanishes and we get the more unimodal results, as expected. In both cases, the filtered factors are well estimated. Hence, there are no further consequences for our parameters of interest, such as the cluster centroids. All estimated cluster centroids are centered around their true values with merely small deviations.



(a) Densities of estimated static parameters.

(b) Densities of estimated cluster centroids.

Fig. 6. Parameter estimation results of $M = 1000$ simulations for $N = 500, T_y = 10, T_x = 100$. Vertical lines represent true values.

Table 2
Confusion matrices of estimated cluster centroids. All results are based on $M = 1000$ simulations with $N = 500$. The top panels have $T_x = 50$ while $T_y \in \{10, 20\}$ varies, and the bottom panels have $T_x = 100$ fixed while $T_y \in \{10, 20\}$ varies. In each panel, the row labels indicate the true clusters and the columns labels to the assigned clusters in estimation. Frequencies are given (here also equal to percentages), and perfect classification would be $100I_5$. From smallest to largest, the true values are 4, 11, 19, 23, and 35.

	$T_y = 10, T_x = 50$					$T_y = 20, T_x = 50$				
	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
C1	99.960	0.040	0	0	0	100	0	0	0	0
C2	0.057	99.931	0.012	0	0	0.001	99.999	0	0	0
C3	0	0.009	96.938	3.053	0	0	0	99.574	0.426	0
C4	0	0	2.954	97.046	0	0	0	0.378	99.622	0
C5	0	0	0	0	100	0	0	0	0	100
	$T_y = 10, T_x = 100$					$T_y = 20, T_x = 100$				
	E1	E2	E3	E4	E5	E1	E2	E3	E4	E5
C1	99.952	0.048	0	0	0	100	0	0	0	0
C2	0.043	99.945	0.012	0	0	0	100	0	0	0
C3	0	0.010	96.925	3.065	0	0	0	99.602	0.398	0
C4	0	0	2.996	97.004	0	0	0	0.383	99.617	0
C5	0	0	0	0	100	0	0	0	0	100

Besides the estimated values of the cluster centroids, the assignment to the correct cluster is also of importance. For this purpose we present confusion matrices in Table 2. In the columns we vary the time series dimension $T_y \in \{10, 20\}$, while in the rows, $T_x \in \{50, 100\}$ varies. First, we recall at this point that the results of the first step of the estimation procedure are taken as given and the time series dimension of the second step, denoted by T_y, is now of interest. For the case of $T_y = 10$, the confusion matrices are already satisfactory, with lower boundaries of 96.9% being correctly assigned; for the case of $T_y = 20$, the lower boundaries are even as high as 99.6%. This confirms our earlier finding that even though it might be empirically challenging to separate some of the score-driven model parameters, it does not lead to any problem in identifying the clusters and cluster centroids in the second step. Hence we can correctly identify the structure in the data.

The estimated cluster centroids and confusion matrices are rather precise. However, our key interest is the comparison between the unrestricted model and the clustered model. For that matter, we present in-sample statistics in Table 3 for time series dimension $T_y = 10$. We again found very similar results for $T_y = 20$, so Table B.1 in the Appendix shows the corresponding tables. In the simulation study, we used five equally sized clusters with centroids 4, 11, 19, 23, and 35 as the loading matrix. We report for each of the clusters, and the full sample, the mean squared error, and mean absolute error of the unrestricted estimates and estimated cluster centroids compared to the true values and the difference between the two estimates. Overall, the estimated cluster centroids always outperform the unrestricted estimates. Furthermore, we report the contribution of each cluster to the log likelihood. We can then compare the models by the AIC and, since the clustered model is a restricted version of the unrestricted model, the LR-test. In all cases, the LR-test provides evidence that there is no significant difference between the log-likelihood values. However, taking into account that the clustered model has much fewer parameters than the unrestricted model, the AIC shows that the clustered model should be preferred. For each combination of time series dimensions, the differences between the clusters are small. This is because the clusters are here chosen to be the same in size and without deviation around the cluster centroid, but in empirical studies such statistics

Table 3
Model fit for unrestricted and clustered model. All results are based on $M = 1000$ simulations with $N = 500$ and $T_y = 10$, with $T_x = 50$ in the top panel and $T_x = 100$ in the bottom panel. Each row in a panel represents a cluster and the last row is the full sample. The columns show loss functions given of the unrestricted loadings compared to the true ones ("Unr."), the clustered centroids compared to the true ones ("Cl."), and the unrestricted loadings minus the clustered centroids ("Diff."). The first three columns have the MSE as the loss function and the last three columns show the MAE. For both models, the log likelihood and AIC are given and they are compared via the LR-statistic in the last column. For the latter, the critical values are 123 for each cluster (100-1 degrees of freedom) and 548 overall (500-5 degrees of freedom), at the 5% significance level.

$T_y = 10, T_x = 50$											
	MSE			MAE			Unrestricted		Clustered		
	Unr.	Cl.	Diff.	Unr.	Cl.	Diff.	LL	AIC	LL	AIC	LR
C1	1.134	0.047	1.106	0.849	0.127	0.839	−4,324	8,850	−4,375	8,754	102
C2	1.234	0.150	1.117	0.884	0.267	0.842	−4,325	8,852	−4,376	8,756	102
C3	1.432	0.827	1.015	0.948	0.538	0.817	−4,332	8,866	−4,378	8,760	92
C4	1.575	0.929	1.008	0.992	0.623	0.813	−4,335	8,872	−4,381	8,766	92
C5	2.168	1.053	1.115	1.157	0.786	0.839	−4,350	8,902	−4,399	8,802	98
Full	1.508	0.601	1.072	0.966	0.468	0.830	−21,884	44,770	−22,130	44,272	492
$T_y = 10, T_x = 100$											
	MSE			MAE			Unrestricted		Clustered		
	Unr.	Cl.	Diff.	Unr.	Cl.	Diff.	LL	AIC	LL	AIC	LR
C1	1.137	0.047	1.111	0.849	0.121	0.840	−4,323	8,848	−4,374	8,752	102
C2	1.216	0.134	1.109	0.876	0.243	0.839	−4,323	8,848	−4,374	8,752	102
C3	1.408	0.802	1.010	0.937	0.508	0.814	−4,326	8,854	−4,373	8,750	94
C4	1.519	0.888	1.004	0.971	0.575	0.811	−4,328	8,858	−4,375	8,754	94
C5	2.054	0.950	1.104	1.114	0.705	0.837	−4,338	8,878	−4,388	8,780	100
Full	1.467	0.564	1.067	0.949	0.430	0.828	−21,856	44,714	−22,105	44,222	498

will give insight on the homogeneity of the different clusters.

4.2. Forecasting results

By enforcing the clustered structure in our modeling framework, we do not want to compromise forecasting performance. To judge the accuracy of clustered forecasting, we compute loss functions of forecasting with and without clustering. We then consider the fraction of the two loss functions and prefer clustered forecasting if

$$\text{Relative Accuracy} = M^{-1}\sum_{m=1}^{M}\left(\frac{LF_m^{clustered}}{LF_m^{unrestricted}}\right) < 1,$$

where the numerator reflects clustered forecasting (with $\hat{\Lambda}$ after running k-means), and the denominator reflects unrestricted forecasting (with $\tilde{\Lambda}$ directly after least squares) for loss function $LF \in \{MSE, MAE\}$. We use $M = 1000$ simulations to obtain the relative accuracy.

Table 4 reports these average fractions; in the columns we vary time series dimension $T_y \in \{10, 20\}$, while in the rows we vary $T_x \in \{50, 100\}$. Furthermore, row f in any cell of Table 4 reports the performance for forecasting $f \in \{1, 2, 3\}$ steps ahead.

Above, we saw that the time series dimension T_x of the first step in our proposed estimation procedure does not have a big impact on the estimation results of the second step. This is also confirmed by the forecasting results. An improvement of more than 4% in the mean squared error is obtained if $T_y = 20$. This becomes 7% if T_y is only half of it. This reveals the strength of our proposed procedure: for data sets where forecasting is of interest but the time series dimension is small although the cross-section dimension is large, it is beneficial to exploit the clustered nature of the data. As the time series dimension T_y increases, clustered forecasting goes to unrestricted forecasting because the unrestricted estimates of the loadings become less biased in the second step. For a small time series dimension T_y, the clustering averages out these biases such that the forecasting performance improves.

Table 4
Forecasting performance of clustered forecasting versus unrestricted forecasting. All results are based on $M = 1000$ simulations with $N = 500$. The top panel has $T_x = 50$ while $T_y \in \{10, 20\}$ varies, and the bottom panel has $T_x = 100$ fixed while $T_y \in \{10, 20\}$ varies. Clustered forecasting is preferred if $M^{-1}\sum_{m=1}^{M}\left(\frac{LF^{clustered}}{LF^{unrestricted}}\right)_m < 1$, where the numerator reflects clustered forecasting (with $\hat{\Lambda}$ after running k-means), and the denominator reflects unrestricted forecasting (with $\tilde{\Lambda}$ directly after least squares) for loss function $LF \in \{MSE, MAE\}$. The first row in each cell represents one-step-ahead forecasting, the second row two steps ahead, and the last row three steps ahead.

		$T_y = 10$		$T_y = 20$	
		MSE-ratio	*MAE-ratio*	*MSE-ratio*	*MAE-ratio*
$T_x = 50$	$f = 1$	0.930	0.963	0.957	0.978
	$f = 2$	0.929	0.963	0.957	0.978
	$f = 3$	0.929	0.963	0.958	0.978
$T_x = 100$	$f = 1$	0.928	0.962	0.956	0.978
	$f = 2$	0.928	0.962	0.956	0.977
	$f = 3$	0.930	0.963	0.958	0.978

4.3. Sensitivity analysis

In the Monte Carlo study, we assumed that the number of clusters is known. This is not the case in an empirical study. To learn about the sensitivity of assuming the number of clusters as known, we present results for cases where we deviate from the true number of clusters. We keep simulating data from our model with five clusters, but now we also let the k-means algorithm use another

Table 5
Estimation and forecasting results of using 3, . . . , 7 clusters in the *k*-means algorithm, while the data is simulated using five equally sized clusters with loadings 4, 11, 19, 23, and 35. All results are based on $M = 1000$ simulations with $N = 500, T_x = 50, T_y = 10$. The first column gives the number of clusters used in the algorithm, where the row of 5 clusters in italics indicates the true number of clusters. The average estimated cluster centroids are given in the second column and the average MSE in the third column. The log likelihood (LL) is given in the fourth column and the relative accuracy (the MSE-ratio) for forecasting one step ahead is given in the last column.

K	Average clustered centroids							MSE	LL	$f = 1$
3	7.495	21.001	35.000					6.905	−23,273	1.437
4	4.002	11.010	21.012	35.000				2.032	−22,470	1.045
5	*4.002*	*10.998*	*18.974*	*23.032*	*35.000*			*0.601*	*-22,130*	*0.930*
6	3.860	9.948	16.479	20.642	24.612	35.111		0.797	−22,091	0.946
7	3.657	8.315	13.110	18.928	21.760	27.276	35.323	0.960	−22,055	0.958

number of clusters. We present summary statistics of the model performance in Table 5 for $N = 500, T_x = 50, T_y = 10$, and static parameters as before. The table reports the average cluster centroid values, its average MSE, the log likelihood of the clustered model, and the MSE-ratio of forecasting $f = 1$ step ahead, all based on $M = 1000$ simulations. The true number of clusters is five and we show results based on three to seven clusters.

The simulations are done with cluster centroids 4, 11, 19, 23, and 35, and with equally sized clusters, such that the average value of the loadings is 18.4. When using fewer than five clusters, we find that the average means are on the higher side, implying a high MSE and low log-likelihood values. When using three clusters instead of five, the forecasting performance decreases considerably (MSE-ratio of 1.437). However, when using four instead of five clusters, the performance of clustered forecasting is as good as for unrestricted forecasting (MSE-ratio ≈ 1). This implies that the accuracy loss is small when imposing more structure, while computational efficiency is higher and interpretation remains. The overall performance is higher when using more clusters than five, rather than using fewer clusters than five, both for in-sample (lower MSE of the clustered centroids) and out-of-sample (MSE-ratios <1) criteria. Given that more clusters can embed a structure with fewer clusters, we can expect that the latter case shows better results. Of course, using the true number of five clusters remains optimal, and using too many clusters leads to overfitting.

Finally, in our main Monte Carlo study, the time series dimensions T_x and T_y and the cross-section dimension N were chosen in line with the empirical study. However, the static parameters and loadings were chosen under the assumption that the data consist of clearly identifiable clusters. In practice, this might be less the case. For example, in our empirical study, we find $\omega \approx 0, \phi \approx 0.35, \alpha \approx 0.9$, and the cluster centroids are smaller and also zero or negative. To verify whether such other settings alter the performance of our proposed methodology, we repeated the analysis with parameters $(\sigma_\xi, \omega, \phi, \alpha, \sigma_\varepsilon)' = (1, 0, 0.35, 0.9, 1)'$ and with cluster centroids −4, −2, 0, 3, and 5. We present these results in the Appendix; they did not lead to very different conclusions regarding the performance of our proposed methodology. Figures B.3 to B.6 show the densities of the estimated parameters and Table B.2 summarizes the forecasting performance.

For this other set of parameter values and cluster centroids, the estimation results are even more precise than those for the original set, even for the smaller time series dimensions. The forecasting results are somewhat less strong (MSE-ratios $\rightarrow 1$), but with clustered forecasting we still do not lose on accuracy compared to unrestricted forecasting. This would mean that unrestricted and clustered forecasting should be equally preferred if one is only interested in forecasting. However, in the policy-relevant context of our empirical study, we are especially interested in the interpretation. By the structure that we put on the model, we gain a lot on interpretation while we do not pay for forecasting performance. This, in combination with the decent improvements in forecasting performance that we found in the main simulation study above, gives enough evidence that our methodology can be generalized and applied in different contexts in future research. All Monte Carlo study results together give sufficient evidence that we can rely on the results from our empirical study.

5. Empirical study

In a study on education enrollment in the Netherlands, Spijkerman (2006) found that educational choices are related to unemployment rates, in particular in part-time and on-the-job education. In the literature, proposed causal relations usually concern demand for education vis-a-vis the supply of labor. Economists typically assume substitution: people allocate their time towards education and the labor market. In this line of thought, enrollment is understood as an investment decision (Clark, 2011, pp. 524–525). Labor market characteristics such as vacancies, unemployment rates, and wages influence the choice made at any given moment. For example, higher demand for labor increases the opportunity costs of education, decreasing its relative preference. Conversely, when confronted with a weak labor market, young students are more likely to remain in education; see Lamb, Walstab, Teese, Vickers, and Rumberger (2004), Clark (2011, p. 523). For post-initial education, a tight labor market might induce people to re-educate, look for better job prospects, or protect their position (Groenez, Desmedt, & Nicaise, 2007). There are effects on the supply side of education as well, especially in on-the-job learning. When demand for labor is high, employers are more likely to provide apprenticeship opportunities. This relationship is assumed to have multiple causes: apprenticeships can be a substitute for hard-to-find workers (especially for middle-skill vacancies), or they can be a way to attract talent. This is the reasoning behind the correction for

Table 6
Parameter estimates and metrics for several GAS specifications.

	GAS(1,1) w/o ω	GAS(1,1)	GAS(2,1) w/o ω	GAS(1,2) w/o ω
$\hat{\sigma}^2_\xi$	0.005	0.005	0.005	0.005
$\hat{\omega}$	–	−0.000	–	–
$\hat{\phi}_1$	0.365	0.365	0.355	0.329
$\hat{\phi}_2$	–	–	−0.117	–
$\hat{\alpha}_1$	0.933	0.937	0.956	0.936
$\hat{\alpha}_2$	–	–	–	0.039
logL	−181.90	−181.06	−181.35	−181.01
AIC	369.80	370.12	370.70	370.02

unemployment vocational education in the Dutch student forecasts (Ministry of Education, Culture and Science, 2020).

There is limited recognition of issues related to the non-stationarity in the regressed time series (participation in education and economic indicators) in this literature; see, for example, Lamb et al. (2004, pp. 126–132). As a result, regressions might be spurious. It is further suggested that causal relations are based on cross-country comparative analysis, in which the participation choices of individuals cannot be distinguished from the varying institutional landscapes; see, for example, Groenez et al. (2007, p. 2). We use data from the Netherlands only, and increase N by lowering the level of analysis. In particular, we study transitions from one type of education to another. The change in the unemployment rate is regressed on the changes in transitions into first-grade studies in Dutch vocational and higher education. Fig. 3 suggests a linear relation in differenced unemployment rates and a differenced share of fulltime students in vocational education. We will not test causal claims, for which a structural causal model should be developed.

Transitions with a higher number of students on average have a higher variance. To normalize, each cross-sectional unit is divided by its average level: $\tilde{y}_{it} = \Delta y_{it}/\bar{y}_i$, where y_{it} denotes the number of people in transition i at time $t = 1, \dots, T_y$. Several specifications of the GAS model have been tested; see Table 6. Based on the AIC, for our yearly observed macroeconomic time series, the GAS(1,1) without intercept seems to be the most suitable. The filter is initialized using the unconditional mean (0). The right panel in Fig. 7 presents the differenced unemployment rate (solid) and the estimated common factor (dashed) over time.

With $\hat{\phi}_1 = 0.365$, extrapolating from this filter results in rapid regression towards the zero mean. Comparing clustered and unclustered loadings will not be very meaningful when the forecasted factor is close to zero. Moreover, the model with one factor is sensitive to variance in the forecast of x_t. Since we have a short panel and little space to vary T_x when forecasting many steps ahead, the out-of-sample tests are somewhat less reliable.[5] Instead, we rely on in-sample metrics to compare the clustered and unrestricted models.

The unrestricted model reveals that the unemployment rate explains about 7.7% of the total variation in the education flows of y_t. This low number is to be expected, since unemployment is a relevant factor for only a part of the transitions into education. The R^2 of the restricted model is naturally lower, but it converges to this rate of 7.7% as the number of clusters increases. Similarly, the MSE- and MAE-ratios in the right panel converge to 1; see Fig. 8. We aim for a relatively small number of centroids, since the predictive performance worsens when increasing the number of clusters. To find the optimal number of clusters, the adjusted R^2 ($\bar{R}^2$) can be used, which has its maximum value at $K = 19$. The MSE-ratio is 1.001, meaning that the predictive performance of the clustered model is almost on par with the unrestricted one. Thus, clustering the loading matrix is an effective way to reduce the number of parameters in the model. We can refer to this procedure as the "elbow method"; see Section 3.2. Alternatively, other criteria to select the optimal number of clusters can be adopted, such as the information criterion AIC.

Fig. 9 presents the estimated loading coefficients. It shows a density of unrestricted loadings and a stepwise cumulative distribution of clustered loadings. The estimated unrestricted loadings have a sample average of −2.67 (std. dev. 13.95), and the clustered loadings have an average of −2.68 (std. dev. 13.89). The smallest and largest loadings are −60.58 and 47.32, respectively, meaning that for these transitions 1% of the change in the unemployment factor corresponds to a >40% change (relative to a transition's sample average) in student counts. The unrestricted loading coefficients do not reveal obvious clusters (multimodal distribution). This is possibly due to high variances of the estimates (small sample) and to model misspecification. We considered a simple linear model with constant loadings. In reality, the student's decision depends on more than only the unemployment rate. This model therefore is at present not suitable for structural or causal interpretation.

The main advantage of this model is that decreasing the number of estimated coefficients reduces model complexity, without losing notable explanatory performance. Clustering might be helpful when model simplicity is considered attractive. In the context of the governmental student forecasts, model interpretability is highly valued (Ministry of Education, Culture and Science, 2018, p. 67). Since the governmental student forecasts feed the education budget, and thus the allocation of scarce public resources, transparency and explainability are key. We

[5] In an exploration of forecasting performance on random subsets of the data, we found that the clustered model did not underperform for the unrestricted one. MSE-ratios centered closely around 1.0. More research is needed to draw conclusions.

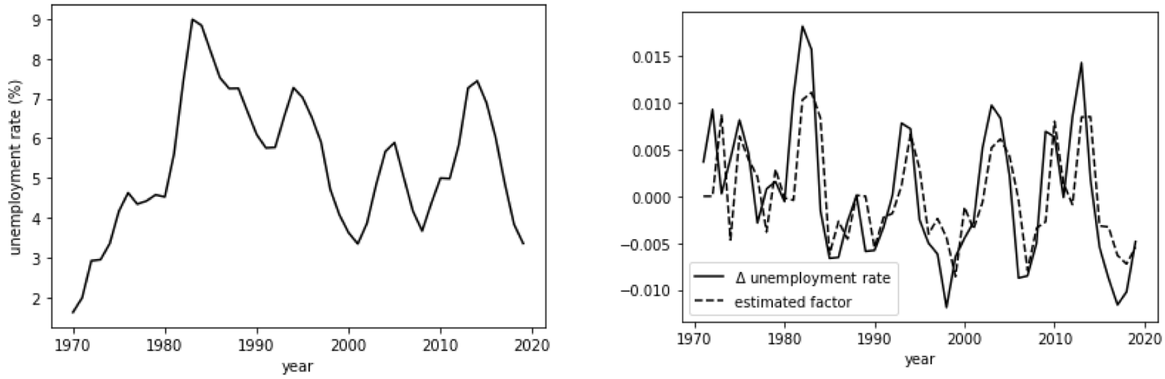


Fig. 7. Unemployment rate and estimated factor. Left: unemployment rate at from 1970 to 2019. Right: differenced unemployment rate and estimated factor $\hat{f}_t$ from 1971 to 2019 ($t = 1, 2, \dots, T_x$, $T_x = 49$).

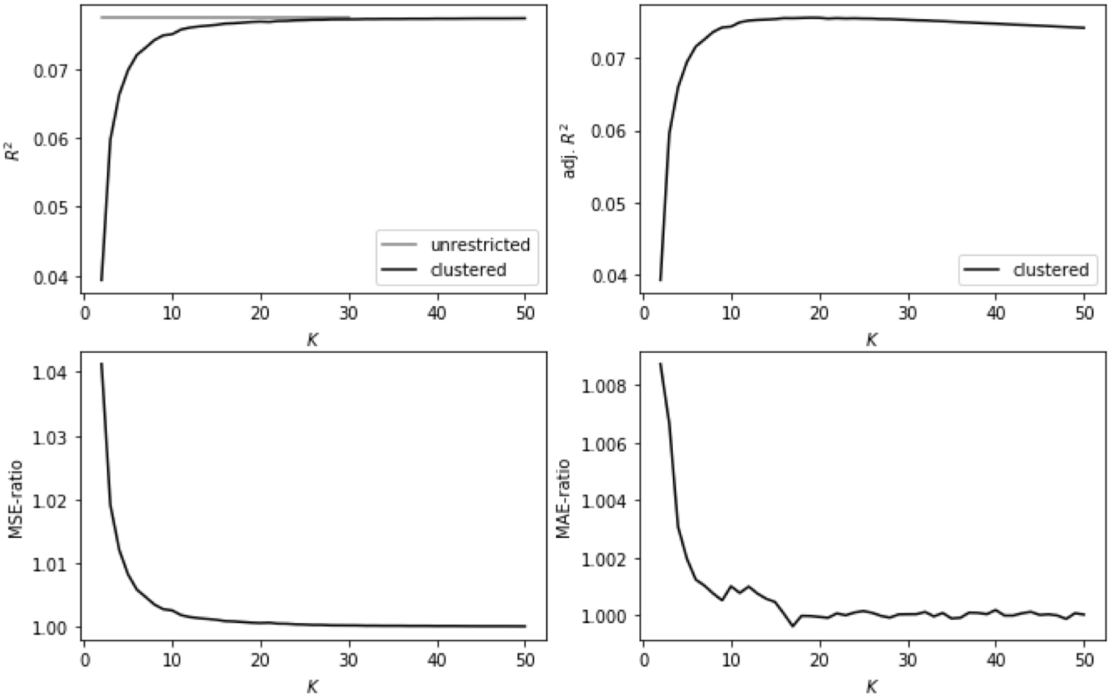


Fig. 8. Various metrics (R^2, $\bar{R}^2$, MSE-ratio, and MAE-ratio) for $K = 2, \dots, 50$ (with $R = 1{,}000$ repetitions using different centroid seeds).

argue that a clustered model is easier to convey, especially if the clusters themselves are meaningfully presented. Both the linear model and k-means clustering are transparent and widely known. Little explanatory performance is traded in for gains in model simplicity. Restricting the model does not lead to notable performance loss.

Fig. 10 presents the R^2 per cluster. As one would expect, the common factor is the most relevant for clusters with larger loadings. In one cluster, 32.1% of variance is explained by the filtered unemployment factor. In the left tail we find transitions into vocational education, which seem to be negatively related to unemployment rates. This supports the hypotheses that participation in on-the-job learning depends on the availability of apprenticeships and/or that low unemployment induces people to learn market-oriented skills. Similarly, the right tail of the distribution contains many transitions into school-based vocational education, moving pro-cyclical with unemployment.

Across the panel, we find that part-time/on-the-job education tends to covary negatively with unemployment; see Table C.1. Similarly, the results indicate that unemployment rates most affect people aged 31 years and above; see Fig. 11. Also, those not in education are less likely to study with increasing unemployment, whereas transitions from diploma origins do not show a strong covariance. Thus, the results do not suggest that a weak labor market induces people to reorient. Instead, the results favor the notion that people are more likely to participate in post-initial education when they have better job prospects. Alternatively, post-initial education depends on the availability of apprenticeship positions. Although

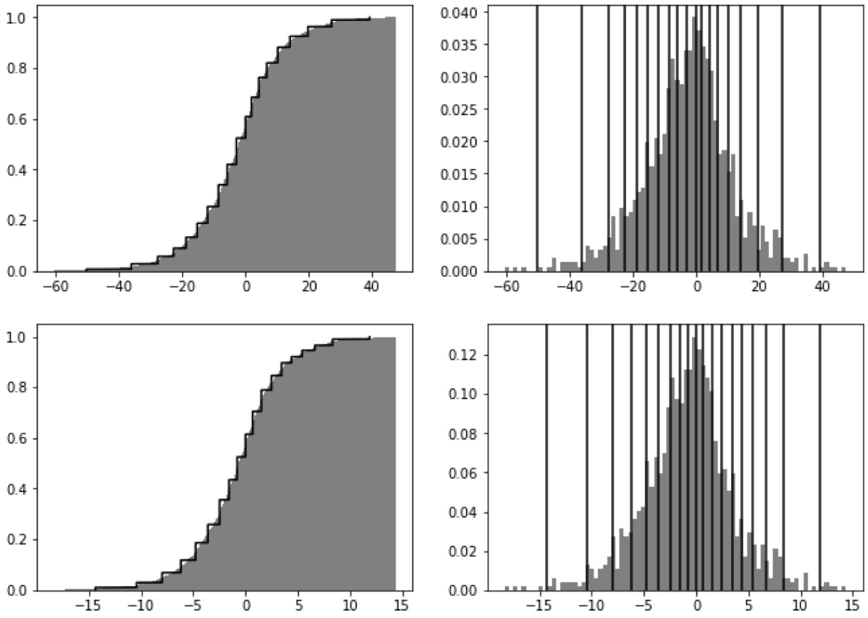


Fig. 9. Distribution of loadings. Up with OLS, down with ridge regression ($\alpha = 0.01$). Left: cumulative distribution. The stepwise line indicates the position of the clusters and the distribution of clustered loadings. Right: density. The vertical lines indicate estimated cluster positions.

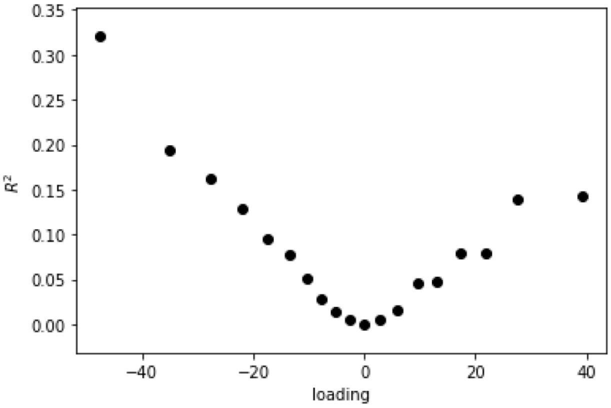


Fig. 10. R^2 per cluster.

clustering strongly reduces the number of parameters, Fig. 11 indicates that the structure of the loading matrix remains intact. Loadings also vary to some extent across directions in education, but these results do not provide enough basis to make conclusions.

A drawback of the linear regression model on the short panel is that the estimated parameter coefficients are largely affected by random noise. One way to reduce model variance is to include a penalty in the objective criterion. For an illustration, see the right panel of Fig. 9 for the distribution of estimated loadings using ridge regression with $\alpha = 0.001$. Model variance might also be reduced by using the descriptive labels of the educational time series. The categories could be included in a k-means-type algorithm; see Huang (1998) for an extension to clustering in a setting with mixed categorical and numerical data.

To complete the empirical analysis, we present impulse response functions (IRFs) as produced by a unit shock in the differenced unemployment rate x_t at time $t = 0$ ($\xi_0 = 1$). Fig. 12 plots these IRFs, which show the dynamic impact of the unemployment shock in education flows. Through the updating equation, the common factor f_t responds with a one-step delay. Additionally, the education flows predicted by the linear regression model covary synchronously with f_t. Some of the clusters covary positively and some negatively. The right panel shows the effects on the differenced transitions into education for the 19 clusters.

6. Conclusion

In this paper we introduced a novel dynamic factor model that is capable of forecasting the number of students across the many different types of education.

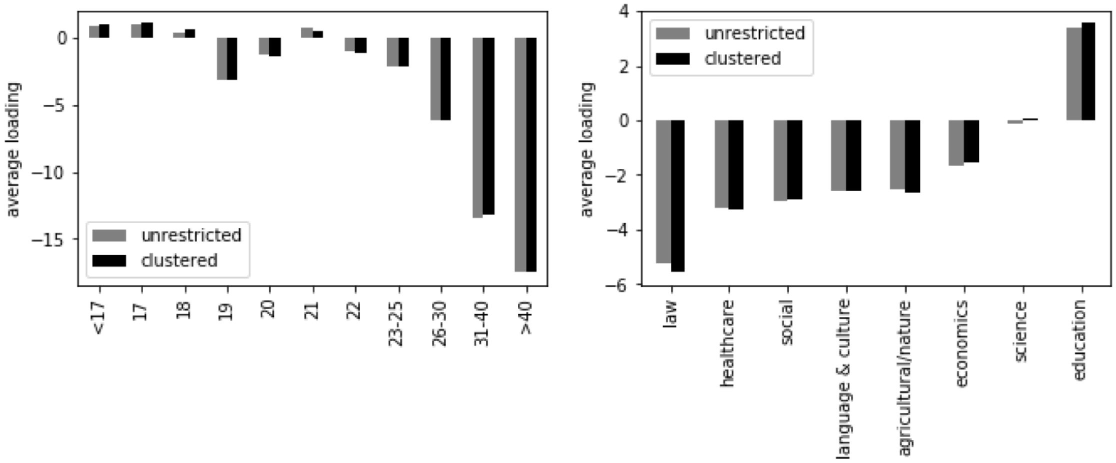


Fig. 11. Average loadings in unrestricted and clustered model (weighted by average level of the transition $\bar{y}$) for age (groups) and directions.

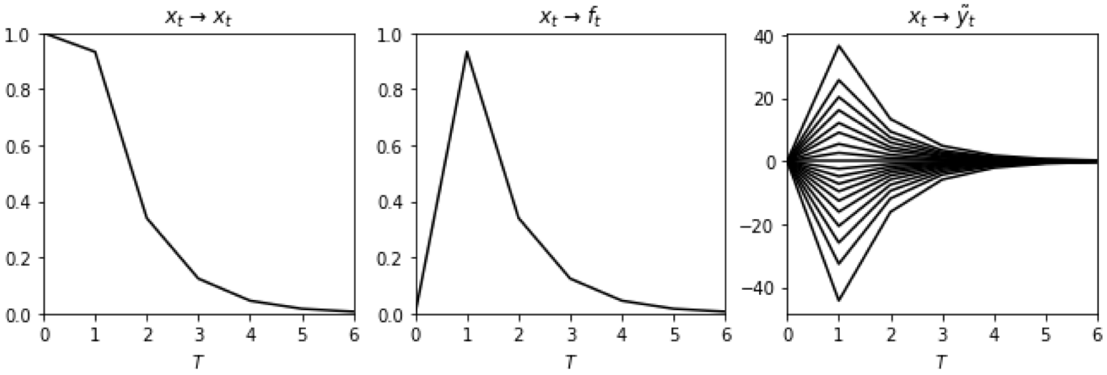


Fig. 12. Impulse response functions for unit shock in differenced unemployment rate x_t on (from left to right) x_t, common factor f_t, and differenced transitions into education $\tilde{y}_t$.

The model can also be used to analyze the relationship between education participation and relevant macroeconomic variables such as the unemployment rate. We further proposed an econometric treatment for this flexible modeling framework. An empirical analysis was carried out for a large data set for the educational system in the Netherlands. We found that, overall, changes in the unemployment rate accounted for approximately 7.7% of the changes in the flows across the educational system. Given that the panel data dimension is huge, we allowed for clustering in the factor loadings that are associated with the dynamic macroeconomic factor. As a result we could measure the extent to which the different types of education exhibit similarities in their relationship with macroeconomic cycles. In the empirical study we highlighted the practical feasibility and good forecasting performance of our modeling framework. In future research, we plan to generalize the methodology further and verify its theoretical properties.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2021.01.026.

References

Alonso, A. M., Galeano, P., & Peña, D. (2020). A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics*, *216*(1), 35–52.

Ando, T., & Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, *31*(1), 163–191.

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.

Barnichon, R., & Mesters, G. (2018). On the demographic adjustment of unemployment. *Review of Economic Statistics*, *100*(2) 219–231.

Blasques, F., Gorgi, P., Koopman, S. J., & Wintenberger, O. (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, *12*(1), 1019–1052.

Blasques, F., Koopman, S. J., & Lucas, A. (2014a). *Maximum likelihood estimation for generalized autoregressive score models*: *Tinbergen institute discussion paper (TI 2014-029/III)*.

Bräuning, F., & Koopman, S. J. (2014). Forecasting macroeconomic variables using collapsed dynamic factor analysis. *International Journal of Forecasting*, *30*(3), 572–584.

Bureau for Economic Policy Analysis (2019). MLT-raming November 2019, cijfers, November. Retrieved 2020-05-21, from https://www.cpb.nl/middellangetermijnverkenning-2022-2025#docid-160027.

Clark, D. (2011). Do recessions keep students in school? The impact of youth unemployment on enrolment in post-compulsory education in England. *Economica*, *78*(311), 523–545, Publisher: Wiley Online Library.

Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*(5), 777–795.

Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.

Groenez, S., Desmedt, E., & Nicaise, I. (2007). Participation in lifelong learning in the EU-15: The role of macro-level determinants. In *Paper for the ECER conference*.

Hallin, M., & Liška, R. (2011). Dynamic factors in the presence of blocks. *Journal of Econometrics*, *163*, 29–41.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: With applications to financial and economic time series*: *Vol. 52*, Cambridge University Press.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304.

Jungbacker, B., & Koopman, S. J. (2015). Likelihood-based dynamic factor analysis for measurement and forecasting. *The Econometrics Journal*, *18*(2), C1–C21.

Lamb, S., Walstab, A., Teese, R., Vickers, M., & Rumberger, R. (2004). *Staying on at school: Improving student retention in Australia*. Brisbane: Queensland Department of Education and the Arts.

Ministry of Education, Culture and Science (2018). *Referentieraming 2018*. September. Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2018/09/18/referentieraming-ocw-2018.

Ministry of Education, Culture and Science (2020). *Referentieramingen 2020*. June.

Spijkerman, M. (2006). *De invloed van conjunctuureffecten op onderwijsdeelname*. SEOR. Erasmus Universiteit Rotterdam, October.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–1179.

Stock, J. H., & Watson, M. W. (2008). In T. Bollerslev, J. Russell, & M. Watson (Eds.), *Volatility and time series econometrics: Essays in honor of Robert F. Engle*, *The evolution of national and regional factors in U.S. housing construction*. Oxford University Press.