What drives industrial energy prices? *

Maximo Camacho
Universidad de Murcia
mcamacho@um.es

Angela Caro
Universidad Carlos III de Madrid
angela.caro@uc3m.es

Daniel Peña
Universidad Carlos III de Madrid
daniel.pena@uc3m.es

October 11, 2022

Abstract

Understanding whether the drivers of industrial energy prices are worldwide, group-specific or country-specific is a key issue in economics. This requires flexible econometric models to examine large data sets containing a significant variety of industrial sectors in different countries. To this end, we propose an extension of a dynamic factor model with group structure to account for observable country-specific explanatory variables and develop Monte Carlo simulations to show its good finite sample performance. Using data from 12 industrial sectors in 30 countries during the period from 1995 to 2015, we find three drivers of energy prices: (i) a common factor, the main driving force, captures the worldwide dynamics; (ii) country-specific variables, mainly related to inflation and the use of renewable and waste resources; and (iii) group-specific factors, which are more related to country affiliation than to sector classification.

Keywords: Energy prices, Dynamic Factor Model, Clustering, Penalized regression.
JEL Classification: C32, C33, C55, C82, E32.

*We are grateful to Editor Prof. Sushanta Mallick, the Associate Editor, and two anonymous referees for very insightful and constructive comments that have improved our work. We are also very grateful to professors Andres Alonso, Tomohiro Ando and Pedro Galeano for making available their codes. M. Camacho is grateful for the support provided by grant PID2019-107192GB-I00 funded by MCIN/AEI/10.13039/501100011033, A. Caro was in receipt of an FPU grant from the Spanish Government and D. Peña for grant AEI/PID2019-109196GB-I00. All remaining errors are our responsibility. Data and codes that replicate our results are available on https://www.um.es/econometria/Maximo.

1 Introduction

Understanding the worldwide drivers of industrial energy prices is becoming a central issue for energy economists, politicians and energy traders. Energy prices represent a significant portion of our domestic expenditures and are a key factor in the competitiveness of many energy-intensive sectors, which tend to push industrial innovation efforts towards energy-efficient technologies as energy prices rise. The Paris Agreement has brought all countries together with the common objective of combating climate change. Aware of the importance of energy prices on environmental and economic performance, governments modulate the intensity of their climate and growth agendas to promote regulatory practices, taxes and subsidy policies with the aim of helping essential sectors and reallocating resources. Europe's Agency for the Cooperation of Energy Regulators in 2011 made market integration one of its top priorities and the European Green Deal pursues climate neutrality by 2050; the objectives of U.S. government policy are aligned with the concept of "energy dominance"; and one of the three pillars of the Australian Energy Policy Blueprint is putting consumers first by delivering reforms to put downward pressure on electricity and gas prices.

After three decades since the pioneering work of Griffin (1980) on energy economics and policy, the existing literature has focused on several issues related to energy prices. While not claiming to be exhaustive, there is a strand of literature that analyzes individual fuel types. Among others, Hailemariam and Smyth (2019) examine what drives natural gas prices in the US. Gao et al. (2021) propose flexible time-varying parameter models to forecast natural gas prices in the US, EU and Japan. Lyu et al. (2021) examine the role of economic uncertainty in oil prices fluctuations using US data and finds that they have countercyclical patterns. With respect to international oil price co-movements, Van Benthem and Romani (2009), investigate what drives energy demand in developing countries, and Aastveit et al. (2015), focus on the different responses to oil market shock across a set of geographical regions.

Another strand of the literature examines the links between energy prices, energy consumption and economic developments. For example, Mahadevan and Asafu-Adjaye (2007) find that a short- and long-run bidirectional causality exists between economic growth and energy consumption among energy exporters and importers in developed countries. In addition, Bertrand (2014) provides the first theoretical analysis of fuel switching within the European power sector, and Bretschger (2015) uses a panel of 37 developed economies over the period 1975-2009 to examine the impact of energy prices on long-run growth.

Binder and Makridis (2022) analyze the transmission of gas prices to consumer beliefs and expectations about the economy in the US. Correia-Fernandes et al. (2021) evaluate the dependence between electricity and natural gas prices using European data between 2015 and 2020. Ozcan et al. (2020) examine the interactions across energy consumption, economic growth and environmental degradation in OECD countries. Finally, Baumeister et al. (2022) evaluate the performance of global economic activity indicators for forecasting real oil prices and global petroleum consumption.

In terms of literature on the subject, to the best of our knowledge, very few have investigated the determinants of the most influential factors driving industrial energy prices for various countries and industrial sectors. With the aim of providing new insights into this issue, we focus on the panel data of the sector level Fixed Weights Energy Price Index (FEPI) for 12 industrial sectors in 30

countries for the period 1995 to 2015, which is a index of real energy prices recently developed by Sato et al. (2019).[1] This data set has a number of advantages over previous weighted energy price indices, such as covering four key types of fuel carriers, providing a large coverage of sectors and countries, addressing the endogeneity of fuel choice and being easily available for download.

With this data base in mind, we seek to answer three important empirical questions on the forces driving industrial energy prices. Firstly, we attempt to determine the optimal number of observable domestic driving forces of energy prices. In particular, we focus on the role of country-level development of energy imports, energy use, output per unit of energy use, renewable energy consumption, inflation rate, population growth, energy production from renewable sources and combustible renewables as potential country-specific variables driving industrial energy prices.

Secondly, the panel of industrial energy price indices potentially admits group structures with energy prices exhibiting strong cross-correlation among the series in the same group and separated from the cross-correlation observed in other groups. Thus, we focus on estimating the optimal number of groups among the large number of energy prices, determining how many group-specific factors explain the group-specific co-movements, providing group memberships and establishing the sources of group formation.

Thirdly, with the aim of computing worldwide industrial energy price indices, we need to have a grasp of the unobservable factors that behave as drivers of a common-to-all energy price index, as well as the national factors, the group-specific components and the idiosyncratic dynamics. Thus, we also ask if there are such common forces driving energy prices and how many factors explain the common co-movements.

In order to address these questions, our goal is to consider a framework which is general enough to let the data determine where the similarities and co-movements in industrial energy prices lie, while allowing for co-movements within and across countries and industrial sectors. To overcome the dimensionality problem of standard regression techniques, such as vector autoregressive models, when the cross-section dimension is even greater than the time-series dimension, we rely on factor models. In addition, we attempt to group the industrial price indices more flexibly than along the sector or country dimension only. Our proposal separates the drivers of energy prices into country-specific components, common co-movements and group-specific co-movements, where the matching of variables to groups, the number of groups, and the estimation of latent factors are data driven decisions.

For this purpose, we extend the grouped factor model proposed by Alonso et al. (2020) in several dimensions. Firstly, we follow Ando and Bai (2016) to allow for observable explanatory variables, in addition to common co-movements and group-specific co-movements. As the number of explanatory variables could potentially be large, we determine the number of non-zero coefficients through a regularization procedure. Secondly, we follow Bai (2009) and iterate the estimation algorithm by estimating the parameters of the country-specific components, given the factors and group estimates, and update factors/groups given the estimates of the country-specific components until convergence.

We evaluate the ability of our proposal to choose the appropriate number of groups and provide

[1] Due to data availability restrictions, we use only 30 out of the 48 countries included in the original data set.

adequate group membership by means of several Monte Carlo experiments, which are designed to capture some basic data problems that characterize economic data sets. We find a remarkable finite sample performance of our proposal. The mean of the estimated number of clusters barely deviates from the simulated number of clusters, with the proportion of times our proposal chooses the correct number of clusters being higher than 90% and tending to 100% when either the time series dimension or the cross-section dimension are large enough. In addition, we find a significant agreement between the label assignment obtained with the model and the label assignment of the different data generating processes. Noticeably, the documented good performance remains when the errors are not homoscedastic and/or do exhibit some cross-sectional correlation.

The empirical analysis indicates that the sources of the co-movement of cross-sectional and time-series variations in industrial energy prices are beyond the usual country- and sector-specific classifications. The main country-specific variables that work as drivers of energy prices are inflation rate, energy imports, and the use of renewable and waste resources. In addition, we find a common global factor that explains almost three quarters of the total variation for industrial energy prices. This factor loads positively with the energy prices of all the countries except Brazil and gives the worldwide evolution of energy prices. Finally, our results suggest that there are two separate groups of energy prices, whose dynamics are well captured by two distinct group-specific factors. Interestingly, we find that the estimated group membership is captured better by country affiliations than by sector classifications. These results illustrate that one of the key advantages of our proposal is that price indices can be grouped more flexibly than along the sector or country dimension only, and that the choice of clustering dimension is a data driven decision.

We find that our factor model explains three-quarters of the variance of energy prices, which indicates that the proposed model fits the data well. About two-quarters of the variance of energy prices is explained by the unobserved common factor and only one quarter is explained by the sum of the country-specific variables and the group-specific factors. This result may contain important implications and guidelines for policy authorities trying to mitigate the adverse effects of long periods of significant escalation in energy prices. In order to maximize the effectiveness of their measures, our results suggest the need to shift from country-specific strategies to worldwide measures common to all the industrial energy prices, regardless of sector, country or group.

The rest of the article is organized as follows. Section 2 specifies the model and describes an algorithm to select: the correct number of groups, the group membership, the number of factors to estimate model parameters and to compute inferences on unobserved factors. Section 3 illustrates the finite sample performance of the model through several Monte Carlo simulations. Section 4 describes the data set and examines the empirical results. Section 5 concludes.

2 Theoretical framework

2.1 Model specification

Let $t = 1, ..., T$ and $i = 1, ..., N$ represent the time and cross-section indices. To establish country memberships, the known number of countries is Q and $C = \{c_1, ..., c_N\}$ refers to the observed country membership, with $c_i \in \{1, ..., Q\}$. To determine group membership, the unknown number

of groups is S and $G = \{g_1, ..., g_N\}$ is the unobserved group membership, with $g_i \in \{1, ..., S\}$. There are N_j units within group j, such that $N = \sum_{j=1}^{S} N_j$. For example, $c_i = 1$ and $g_i = 2$ indicate that the unit i belongs to the first country and the second group.

As in Ando and Bai (2017), we assume that the stationary energy price of the ith unit, observed at time t, y_{it}, admits a factor-representation with country-specific explanatory variables $(x_{c_i,t})$, group-specific unobserved factors $(f_{g_i,t})$ and a global unobserved factor across sectors and countries $(f_{0,t})$ which is expressed as follows

$$y_{it} = x'_{c_i,t}\beta_{c_i} + f'_{0,t}\lambda_{0,i} + f'_{g_i,t}\lambda_{g_i,i} + \varepsilon_{i,t}. \tag{1}$$

In this expression, $x_{c_i,t}$ is a $p \times 1$ vector of observed explanatory variables for country c_i; $f_{0,t}$ is a $r_0 \times 1$ vector of unobserved global factors affecting all energy prices; and, for example, if $g_i = j$ with $j \in \{1, ..., S\}$, $f_{j,t}$ is a $r_j \times 1$ vector of unobserved group-specific factors that affect the energy prices of only group j.[2] To facilitate economic interpretation, we assume that the $p \times 1$ vector of unknown regression coefficients β_{c_i} are fixed for each country c_i.[3] In addition, $\lambda_{0,i}$ are the common factor loadings and $\lambda_{g_i,i}$ measures the unknown sensitivity to unobservable group-specific factors. Finally, $\varepsilon_{i,t}$ is the unit specific error with zero mean, variance σ_{ε_i} and weak dependency as stated in Bai and Ng (2002) and Ahn and Horenstein (2013). It is assumed that $\varepsilon_{i,t}$ is independent of x_{it}, $f_{0,t}$ and $f_{g_i,t}$.[4]

Let us collect the common and group-specific factors and loadings in the $N \times r_0$ and $N \times r_j$ matrices Λ_0 and Λ_j, with $j = 1, \ldots, S$. Let Σ_{f_0} and Σ_{f_j} be the covariance matrices of the factors. As it is standard in factor models, we can only identify the factor space and not the individual factors themselves given that $\Lambda F_t = (\Lambda B^{-1})(BF_t)$, where B is any invertible $r \times r$ matrix. To overcome this drawback, we can always choose an orthogonal base in this space to represent the factors. Following Bai and Ng (2013), we identify the model as follows: (1) $\Lambda'_0\Lambda_0 = \mathbf{I}_{r_0}$, where $\mathbf{I}_{r_0}$ is the identity matrix of order r_0; (2) $\Lambda'_j\Lambda_j = \mathbf{I}_{r_j}$, where $\mathbf{I}_{r_j}$ is the identity matrix of order r_j for $j = 1, ..., S$; (3) $\Lambda'_0\Lambda_j = \mathbf{0}_{r_0 \times r_j}$; (4) $\Lambda'_j\Lambda_k = \mathbf{0}_{r_j \times r_k}$ for $j \neq k$; and (5) the r covariance matrices of the factors, Σ_{f_0}, Σ_{f_j} for $j = 1, ..., S$ are diagonal.

The factors, collected in $f_t = (f'_{0,t}, f'_{1,t}, ..., f'_{S,t})'$, are assumed to be stationary, independent and follow a diagonal Vector Autoregressive Moving Average model, $\Phi(B)f_t = \Theta(B)u_t$, where u_t is a white noise $N \times 1$ vector with mean equals to zero and diagonal covariance matrix Σ_u. If we collect the errors in (1) in the $N \times 1$ vector ε_t, it is also assumed that $E[\varepsilon_t u'_{t-h}] = \mathbf{0}_{N \times r}$ for all $h = 0, \pm 1, \pm 2, ...$

2.2 Model estimation

Alonso et al. (2020) developed an algorithm that estimates the number of groups, the number of factors and the model parameters and determines the group membership from large sets of time

[2] The total number of factors is $r = \sum_{j=0}^{S} r_j$.

[3] In the empirical application, we show that the data support this assumption.

[4] In the simulations, we examine the potential effects on model's performance of some forms of cross-sectional dependence and serial correlation in the errors.

series.[5] However, the implementation of their algorithm to our data set requires two important extensions. Firstly, we extend the algorithm to account for observed explanatory variables, others than unobserved common and group-specific factors. Secondly, for consistency purposes, the algorithm requires the country-specific explanatory variables to be independent of the global and group-specific factors, which is unlikely to hold in empirical implementations. To overcome this drawback, we follow Bai (2009) and iterate the algorithm, in the sense that the parameters of the explanatory variables are estimated given the factor and group estimates, and the factors/groups are obtained given these estimates. This procedure is iterated until convergence.

Our extension addresses the issue of estimating the country-specific regression coefficients for the control variables $\{\beta_1, \ldots, \beta_Q\}$, the unobservable factor structure, f_t for $t = 1, \ldots, T$, and the factor loadings $\{\Lambda_0, \Lambda_1, \ldots, \Lambda_S\}$. In addition, we jointly address the issue of selecting the optimal number of groups, S, the group memberships, G, and the dimension of the unobservable factors, $r_0, r_1, \ldots, r_S$.

To facilitate understanding, we list the steps involved in the estimation algorithm as follows:

Step 1. Clean the data from additive outliers and level shifts.

Step 2. Estimate the country-specific regression coefficients, $\beta_1, \ldots, \beta_Q$.

Step 3. Given the values of $\hat{\beta}_1, \ldots, \hat{\beta}_Q$, define the variable $y^*_{it} = y_{it} - x'_{c_i,t}\hat{\beta}_{c_i}$ and estimate an initial set of global factors $f_{0,t}$ and their corresponding loadings Λ_0.

Step 4. Given the values $\hat{\Lambda}_0$, $\hat{f}_{0,t}$, estimate the number of groups S and group membership G.

Step 5. Given the values S and G, estimate the group-specific factors $f_{g_i,t}$ and their corresponding loadings Λ_{g_i}.

Step 6. Given the values $\hat{\Lambda}_{g_i}$, $\hat{f}_{g_i,t}$, classify the estimated factors and re-estimate the group-specific factors using the series of the residuals $v_t = y^*_t - \hat{\Lambda}^*_0\hat{f}^*_{0,t}$.

Step 7. Given the values $\hat{\Lambda}^*_0$, $\hat{f}^*_{0,t}$ and $\hat{\Lambda}^*_{g_i}$,$\hat{f}^*_{g_i,t}$, estimate the country-specific regression coefficients, $\beta_1, \ldots, \beta_Q$.

Step 8. Repeat Steps $2 \sim 7$ until convergence.

Remark 1. Outliers and level shifts are quite common in energy prices, especially for developing countries. These extraordinary observations, usually associated with internal conflicts and changes in policies pose challenges in estimating highly structured models. As a result, prior to model estimation, we propose identifying and correcting for outliers with the Multivariate Additive Outlier (MAO) and the Multivariate Level Shift (MLS) approaches stated in Alonso et al. (2020).

Skipping over the details, which appear in the authors' contribution, this method searches for exogenous changes that affect the mean of the time series at a particular date or starting at a date and continuing until the end of the observed period. If MAO or MLS are detected, they are cleaned using an iterative algorithm that replaces the outliers with their interpolated values using Exponentially Weighted Moving Average (EWMA) smoothing.

Remark 2. To estimate the unknown regression coefficients, we define β as the $Q \times p$ matrix

[5] Using simulations, Alonso et al. (2020) showed that their algorithm exhibited better finite sample performance than the one proposed by Ando and Bai (2017).

that stacks the parameters β_{c_i} for each country $c_i = 1, \ldots, Q$. The estimates of these parameters are obtained by minimizing the penalized least-squares objective function

$$L(\beta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x'_{c_i,t}\beta_{c_i})^2 + NT \cdot P(\beta). \tag{2}$$

As the number of controls can be potentially large, the penalty function, $P(\beta)$, is used to identify the significant components of the regression coefficients. For this purpose, we propose a sparse penalized approach based on the SCAD (Smoothly Clipped Absolute Deviation) penalty of Fan and Li (2001)

$$P(\beta) = \sum_{c_i=1}^{Q} \sum_{j=1}^{p} P(|\beta_{c_i,j}|), \tag{3}$$

where $\beta_{c_i,j}$ refers to the j-th regression coefficient of country c_i and $P(|\beta_{c_i,j}|)$ is the quadratic spline function with knots at κ_{c_i} and $\gamma\kappa_{c_i}$

$$P(|\beta_{c_i,j}|) = \begin{cases} \kappa_{c_i}|\beta_{c_i,j}| & \text{if } |\beta_{c_i,j}| \leq \kappa_{c_i} \\ \frac{\gamma\kappa_{c_i}|\beta_{c_i,j}| - 0.5(\beta^2_{c_i,j} + \kappa^2_{c_i})}{\gamma - 1} & \text{if } \kappa_{c_i} < |\beta_{c_i,j}| \leq \gamma\kappa_{c_i} \\ \frac{\kappa^2_{c_i}(\gamma^2 - 1)}{2(\gamma - 1)} & \text{otherwise,} \end{cases}$$

for $\kappa_{c_i} > 0$ and $\gamma > 2$.[6]

In practice, this results in small coefficients being set to zero, a few other coefficients being shrunk towards zero while retaining the large coefficients as they are. Theoretically, the best pair (κ_{c_i}, γ) could be obtained using two-dimensional grid search. However, such an implementation could be computationally expensive. We set $\gamma = 3.7$ as Fan and Li (2001) suggested, and we select κ_{c_i} using cross validation methods. In the section devoted to simulations and in the Appendix, we show the good finite sample performance of SCAD, which outperforms other penalized approaches.

Given the estimates of the regression coefficients, $\hat{\beta}_{c_i}$, we define the variable $y^*_{it} = y_{it} - x'_{c_i,t}\hat{\beta}_{c_i}$. Then, the original factor model (1) reduces to

$$y^*_{it} = f'_{0,t}\lambda_{0,i} + f'_{g_i,t}\lambda_{g_i,i} + \varepsilon^*_{i,t}, \tag{4}$$

with $\varepsilon^*_{it} = \varepsilon_{it} + x'_{c_i,t}(\beta_{c_i} - \hat{\beta}_{c_i})$.

Remark 3. We propose using the method introduced in Caro and Peña (2021) to estimate the global factors, their factor loadings and the optimal number of common factors. Let $\mathbf{R}_{y^*}(k)$ be the lag k correlation matrix of the series y^*_t. For a pre-specified positive integer k_0, the estimates of the columns of the common loading matrix, $\hat{\Lambda}_0$, are the eigenvectors associated to the largest eigenvalues of the combined dynamic correlation matrix

$$\mathbf{R}_{k_0} = \sum_{k=0}^{k_0} w_k \mathbf{R}_y(k)\mathbf{R}_y(k)', \tag{5}$$

[6]The SCAD estimator has all of the desirable properties, including unbiasedness, sparsity, and continuity.

where the coefficients $w_k > 0$ are weights which verify $\sum_{k=0}^{k_0} w_k = 1$. Specifically, we consider the (standardized) asymptotic cross correlation coefficients for white noise stationary processes. Then, the number of global factors, r_0, is determined by using a test based on the ratios of consecutive eigenvalues of the combined dynamic correlation matrix. Finally, the factors are estimated by $\hat{f}_{0,t} = \hat{\Lambda}_0' y_t^*$.

It is worth emphasizing that the estimation does not disentangle group-specific factors from global factors. Thus, the initial estimate of the number of global factors is expected to include some of the group-specific factors and, for this reason, the number of global factors, r_0, at this step will, in general, be larger than the true number of global factors.

Remark 4. To provide group membership, we use the nonparametric algorithm proposed in Alonso and Peña (2019), which clusters the time series by their linear dependency. In short, the method consists of building an $N \times N$ dissimilarity matrix whose entries are calculated as one minus the generalized pairwise cross correlation of each pair of time series in the common component $\hat{\Lambda}_0 \hat{f}_{0,t}$. Using this matrix, hierarchical clustering with single linkage is used to find some dependence structures and to perform group membership, where the number of clusters are selected by the modification of the Silhouette statistics advocated by Rousseeuw (1987).

Remark 5. As in Step 3, given the group membership, G, and the number of groups, we can estimate the group-specific factor loadings $\hat{\Lambda}_{g_i}$ using the k lag combined dynamic correlation matrix of the time series y_t^* that belongs to group g_i, with $g_i \in \{1, \ldots, S\}$. The number of group-specific factors, r_{g_i}, is determined by testing the ratios of consecutive eigenvalues of the combined dynamic correlation matrix. Finally, the group-specific factors are estimated by $\hat{f}_{g_i,t} = \hat{\Lambda}_{g_i}' y_t^*$.

Remark 6. In Step 6, we need to decide whether a factor is global or group specific. Alonso et al. (2020) based this decision on the empirical canonical correlation between each of the r_0 factors calculated in the third step and the set of group specific factors $\hat{f}_{1,t}, \ldots, \hat{f}_{S,t}$. When the correlation of a factor with elements of a group is higher than 0.9, the factor belongs to this group. If the factor does not belong to any group or belongs to more than one group, then it is a global factor. After the application of these rules, the resulting set of global factors is denoted by $\hat{f}_{0,t}^*$ and their corresponding loadings by $\hat{\Lambda}_0^*$. Finally, the group-specific factors are re-estimated using the series of the residuals $v_t = y_t^* - \hat{\Lambda}_0^* \hat{f}_{0,t}^*$ that correspond to each group.

Remark 7. Consistency of the algorithm described in steps 2 to 7 would require independence between the country-specific explanatory variables, $x_{c_i,t}$, and the global and group-specific factors $(f_{0,t}', f_{g_i,t}')$. This is because, unless one has independence of factors and regressors, the $\hat{\beta}_{c_i}$ estimates obtained in step 2 are inconsistent, and since these inconsistent estimates are subsequently used to calculate the y_{it}^* variable from which factors and groups are estimated in following steps, the $\hat{\beta}_{c_i}$ inconsistency feeds into the later estimates, yielding overall inconsistency.

Since regressor and factor independence is rather unlikely in economic applications, we overcome this drawback by iterating the estimation procedure (see Bai, 2009). The idea is to update $\hat{\beta}_{c_i}$, given factor and group estimates, and update factors/groups given the new $\hat{\beta}_{c_i}$. This procedure is iterated until convergence. In the next section, which is devoted to simulations, we examine the finite sample performance of this iterative algorithm to provide accurate inferences on the number of clusters, group memberships and parameter estimates.

3 Monte Carlo experiment

In this section, we evaluate the finite sample performance of our contribution through several Monte Carlo simulations with data generated to mimic the structure of the data set used in the empirical application. In particular, we generate data from a dynamic factor model with grouped factor structure as in (1). We assume three countries ($Q = 3$) and each one includes N_i sectors, or observed time series. The $N{=}N_1 + N_2 + N_3$ series in the data set are affected by a common factor, $r_0 = 1$, and are clustered in three specific groups, ($S = 3$), each one with three specific factors $r_j = 3$, for $j = 1, 2, 3$.

The global factor $f_{0,t}$ is generated by an AR(1) model with an autoregressive parameter of 0.75, where the global factor noise is a vector of $U(0, 1)$ variables, and the elements of the loading matrix Λ_0 are $U(-2, 2)$. The group-specific factors are generated by an AR(1) model with an autoregressive parameter of 0.75, where the corresponding noises and each element of the group-specific factor loading matrices follow $N(0, 1)$ variables.

For each country, each of the $p = 10$ explanatory variables collected in the vector $x_{c_i,t}$ is generated as $x_{c_i,t} = f'_{0,t}\lambda_{0,i}R + v_{c_i,t}$, where R is a $(p \times 1)$ vector of ones and $v_{c_i,t}$ is a $(p \times 1)$ vector of $U(-4, 4)$ variables. The true number of predictors is only three and the non-zero parameter values of β_{c_i} are set to $(3, 3.5, 3)$ for country 1, to $(-2, -2, -2.5)$ for country 2, and to $(1, 0.5, 1.5)$ for country 3. We put the non-zero parameters into the first three elements of β_{c_i}. Thus, for example, the true parameter vector for country 1 is $\beta_1 = (3, 3.5, 3, 0, 0, ..., 0)'$.

We generate times series with grouped factor structure for various combinations of (T, N), including $T = (50, 100, 300)$ and $N = (300, 600)$. We set the same number of time series in each country ($N_1 = N_2 = N_3$) and, approximately, the same number of group-specific time series in each group. For example, when $N = 300$, we generate 100 series in each country with the first 33 series belonging to group 1, the next 33 series to group 2 and the last 34 series to group 3.

The noises are designed to see the effects of heteroscedasticity, cross-correlated and serially-correlated errors on the factor estimates. To facilitate comparisons, we design the Data-Generating Processes (DGP) as in Ando and Bai (2017). For this purpose, we set $\sigma_{\varepsilon_i} = 1$ and use three different data generating processes. In the first DGP, DGP_1, the noise term is homoscedastic and serially uncorrelated. Thus, ϵ_t is drawn from independent normal distributions with mean $\mathbf{0}$ and covariance matrix $\sigma_e^2\mathbf{I}_N$. In the second DGP, DGP_2, the noise is heteroscedastic and cross-sectionally correlated with $\varepsilon_{it} = 0.2e_{it}^1 + \delta_t e_{it}^2$, where $\delta_t = 1$ if t is odd and zero if t is even, and the N-dimensional vectors $e_t^1 = (e_{1t}^1, ..., e_{Nt}^1)'$ and $e_t^2 = (e_{1t}^2, ..., e_{Nt}^2)'$ follow multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $R = (R_{ij})$ with $R_{ij} = 0.3^{|i-j|}\sigma_e^2$ and e_t^1 and e_t^2 are independent. Finally, the third DGP, DGP_3, contains idiosyncratic errors that exhibit some serial and cross-sectional correlations. In this case, $\varepsilon_{it} = 0.2\varepsilon_{i,t-1} + e_{it}$, where $t = 1, ..., T$, the N-dimensional vector $e_t = (e_{1t}, ..., e_{Nt})'$ follows multivariate normal distributions with mean $\mathbf{0}$ and covariance matrix $R = (R_{ij})$ with $R_{ij} = 0.3^{|i-j|}\sigma_e^2$. In the three DGP the noise variance σ_e^2 takes the values 1 and 2.

We generate a total of 100 replications for each of the three data-generating processes. We apply our iterative algorithm to the simulated data in each one in order to simultaneously select the number of groups, the number of global common factors, the number of group-specific pervasive

factors and the size of the regularization parameter. We set the possible numbers of group-specific and global factors in a range from 0 to 20. The number of groups ranges from 2 to 20 and possible candidates for the regularization parameter are $\kappa_{c_i} = 10^{-3+0.25k}$ for $k = 0, ..., 12$.

The finite sample performance of our proposal is examined in Tables 1 to 3. In this paper, we concentrate our attention on correctly identifying the number of groups, on agreeing between real and estimated partitions, and on estimating the country-specific regression coefficients. Interested readers are referred to Alonso et al. (2020) for the analysis of other measures of the model's performance, such as the robustness for outlier detection, the accuracy of factor estimates and precision in providing the right number of common and group-specific factors.

The ability of the model to provide the correct identification of the number of groups is examined in columns 2 to 4 of Table 1. For each combination (T, N) and each data-generating process, the table reports the mean of the selected number of clusters (first row) and the number of iterations out of the 100 simulations where the true number of clusters was selected (second row). Regardless of the cross section and time series dimensions, the table shows that the mean estimated number of clusters barely deviates from the simulated number of clusters. In addition, the proportion of times the algorithm chooses the correct number of clusters is of almost 90% for the lowest combination $(T, N) = (100, 300)$ and tends to 100% across all simulations when either the time series dimension or the cross-section dimension increases.

We use the Adjusted Rand Index (ARI) proposed by Hubert and Arabie (1985) to compare agreement between real and estimated partitions. The index takes a value between 0 and 1, with 0 indicating that the generated and estimated partitions do not agree on any pair of points and 1 indicating that the two partitions are exactly the same. Let $K = (K_1, \ldots, K_S)$ be the partition suggested by the model, with K_j being the cluster of energy prices with group membership g_j for $j = 1, \ldots, S$, and let $K^0 = (K_1^0, \ldots, K_S^0)$ be the corresponding partition of the data generating process. Thus, ARI can be obtained as

$$ARI(K, K^0) = \frac{\sum_{i=1}^{S}\sum_{j=1}^{S^0}\binom{\#(K_i \cap K_j^0)}{2} - \sum_{i=1}^{S}\binom{\#(K_i)}{2}\sum_{j=1}^{S^0}\binom{\#(K_j^0)}{2}/\binom{N}{2}}{\left(\sum_{i=1}^{S}\binom{\#(K_i)}{2} + \sum_{j=1}^{S^0}\binom{\#(K_j^0)}{2}\right)/2 - \sum_{i=1}^{S}\binom{\#(K_i)}{2}\sum_{j=1}^{S^0}\binom{\#(K_j^0)}{2}/\binom{N}{2}},$$

where # denotes the cardinality.

Columns 5 to 7 in Table 1 report the means of the ARI measure from the 100 Monte Carlo simulations obtained for the estimated and generated partitions for each (T, N) combination and data generating process. The high values of ARI displayed in the table indicate significant agreement between the label assignment obtained with the model and the real label assignment when the errors are homoscedastic and serially uncorrelated. Remarkably, the agreement between the two label assignments diminishes only slightly when the errors are not homoscedastic and/or exhibit some cross-sectional correlation, indicating that the model performance is robust to these data features.

Simulation results for the parameter estimates of the regression coefficients are reported in Table 2. For this purpose, we compute the Mean Squared Error (MSE) between the true coefficients and

their estimates. We call p_{c_i} the number of nonzero parameters in country c_i, and the MSE is

$$MSE = \frac{1}{Q}\sum_{c_i=1}^{Q}\left[\frac{1}{p_{c_i}}\sum_{k=1}^{p_{c_i}}\left(\widehat{\beta}_{c_i,k} - \beta^0_{c_i,k}\right)\right]. \tag{6}$$

In addition, we also measure the identification performance using the True Negative Rate (TNR), which focuses on the proportion of times the proposed criterion is capable of setting the non-relevant observable regressors to zero. If $I(\cdot)$ is the indicator function, which takes a value of 1 if it is true and 0 otherwise, TNR can be obtained as follows

$$TNR = \frac{\sum_{c_i=1}^{Q}\sum_{k=1}^{p_{c_i}} I\left\{\widehat{\beta}_{c_i,k} = 0 \text{ and } \beta_{c_i,k} = 0\right\}}{\sum_{c_i=1}^{Q}\sum_{k=1}^{p_{c_i}} I\left\{\beta_{c_i,k} = 0\right\}}. \tag{7}$$

Entries in Table 2 report the average over the simulations of MSEs (columns 2 to 4) and TNRs (columns 5 to 7). The results reported in the table show that the parameters are well estimated in the simulation studies and that accuracy improves as the size of the panel increases both in the time-series and in the cross-section dimensions. In addition, the table shows that our iterative algorithm works well in determining the non-relevant regressors and that accuracy also increases with the N and T. Notably, the performance does not deteriorate significantly when the errors are not homoscedastic and/or exhibit some cross-sectional correlation. The results reported in the Appendix also show that that SCAD penalty achieves marginally better performance than MCP and LASSO.

As a final remark of this section, we need to verify that the method does not impose a group structure and group factors when actually none are present. For this purpose, in Table 3, we report the simulations of a data generating process that does not contain a group structure, which implies that $S = 1$. To facilitate comparisons, the errors, country structure, elements of $x_{c_i,t}$ and parameters of β_{c_i} are generated as DGP_1 describes. In this case, the common factor $f_{0,t}$ is a vector of $U(0,1)$ variables and the elements of the loading matrix Λ_0 are $U(-2,2)$ and we focus on only two combinations of $(T,N) = (100,300)$ and $(T,N) = (200,300)$. On average, results in Table 3 show that the iterative algorithm we propose correctly identifies no group structure in the DGP because the average number of clusters is close to one. In addition, the algorithm exhibits satisfactory clustering performance. Finally, the reported values of MSEs and TNRs show the good performance of the SCAD penalty in estimating the country-specific regression coefficients. These results are robust regardless of the DGP used in the simulations.

4 Empirical application

4.1 Data description

We analyzed the determinants of the common drives in industrial energy prices by using the Fixed Energy Price Index ($FEPI$) constructed by Sato et al. (2019).[7] The data set provided widespread

[7]The authors also constructed a Variable weights Energy Price Level (VEPL) index, which is useful for cross-sectional analysis. The FEPI is more appropriate for our panel analysis.

coverage of sectors, countries and years. In particular, the index of industrial energy prices is a weighted average of four fuel-specific prices (electricity, natural gas, coal and oil), with weights given by the share of fuel consumption in the sector's energy mix.

The fixed-weight price index captures the variation in fuel prices alone, including policies and taxation, and switches off the source of price variation that is endogenously related to the technological choices of the firm. Let P_{it}^{j} be the real price of fuel type j per tonnes of oil equivalent (toe) for aggregate industry in country i at time t in constant 2010 USD. Let F_{iq}^{j} be the input quantity of fuel type j in toe for the industrial sector q in country i. For each country i, sector q and year t, the index is constructed according to the following equation:

$$FEPI_{iqt} = \sum_{j} w_{iq}^{j} \cdot log(P_{it}^{j}), \tag{8}$$

where weights $w_{iq}^{j} = \sum_{j}(F_{iq}^{j} / \sum_{j} F_{iq}^{j})$ are fixed over time. Thus, the $FEPI$ captures only energy price changes that come from changes in fuel prices, and not through changes in the mix fuel inputs.

Due to data availability restrictions, the effective data set used in this paper includes 30 OECD and non-OECD countries and 12 sectors for the period 1995-2015. Thus, our panel has a cross-sectional dimension of $N = 30 \cdot 12 = 360$ and a time-series dimension of $T = 21$ years. Countries and sectors included in the sample and their respective acronyms are listed in Table 4.

We focused on Dahl (2015) to construct the set of observable control variables that have potential influence on industrial energy prices. The long list that we extracted from his book could have included population growth, demographic shifts and elongation of human life, income growth, environmental concerns, technology (investment capital available), renewable energies, waste storage and proliferation, government intervention, transportation/travel (moving freight, commuting, recreation and tourism, socializing, shopping, other services, industry travel), household heating, cooling, transport, and the availability of nuclear energy.

However, given the scarcity of data on some of the countries in our data set, we limited the list to energy imports, energy intensity level of primary energy, GDP per unit of energy use, renewable energy consumption, inflation rate, population growth, electricity production from renewable sources, and combustible renewables and waste. All of these time series, with the exception of inflation rate and population growth were input into the model on growth rates to achieve stationarity.

4.2 Empirical results

We implemented the factor model with group structure described in Section 2 to the Sato et al. (2019) data set with three objectives. Firstly, we determined the main observable country drivers of industrial energy prices. Secondly, we investigated whether international energy prices would admit a grouped factor model structure, which involved examining whether there are only unobservable common international drivers or whether some group-specific drivers emerge from the model. Thirdly, if there are group-specific drivers of energy prices, we examined whether they depend on country patterns or on industrial sectors. To help in understanding the last issue, in Figure 1 we plotted the 12 energy price indices for Australia and, in Figure 2, the energy prices of the industrial sector Construction for the 30 countries in our sample. These figures showed that

the potential group-specific unobservable drivers could emerge across countries or across industrial sectors.

To begin with, it is worth evaluating the empirical reliability of assuming country-specific regression coefficients. For this purpose, we changed (1) and assumed individual specific coefficients β_i for each energy price indicator y_i, for $i = 1, ..., 360$, and displayed the estimated coefficients in the plots of Figure 3. Each plot shows the non-zero impacts of the explanatory variables on the energy price indeces, and in the x-axis, the 12 energy price indices are grouped by countries. Although the estimated coefficients were heterogeneous across countries, we detected that the energy prices belonging to the same country were affected by the same explanatory variable. For example, the 12 energy prices in the UK tended to be explained by the energy imports (blue bars) and the magnitude of their respective regression coefficients were similar. However, we discarded the idea of assuming the same regression coefficients for each explanatory variable because, as documented in the bottom plots of the figure, the regression coefficients varied substantially across countries.

The time series of energy prices could be affected by univariate and/or multivariate additive outliers and/or level shifts due to many reasons such as political strategies, jobs shifting, economic recessions and market issues. These outliers may affect the sample estimates of autocovariance matrices and model parameters. To avoid this risk, we first cleaned the data of additive outliers and level shifts, which represented 0.60% of the total number of observations in the sample. In addition, as the $FEPI$ is expressed in logarithmic terms, with all the outlier corrected series, we took first differences to have stationary time series.[8]

Figure 4 plots the time series evolution of the cross-sectional average of the 360 price indices. As observed in this figure, the estimated average tracks the energy price evolution well, with obvious and pronounced drops corresponding to the energy crisis at the beginning of the 21^{st} century and the 2008-2009 financial crisis. Since 2014, the average of the price indices has exhibited an uninterrupted drop in energy prices, which is mainly driven by the worldwide oil price collapse.

Next, we fitted model (2) by minimizing the objective function. We set $\lambda = 3.7$ and estimated the coefficients paths for $SCAD$-penalized regression models over a grid of values for the regularization parameter, which were automatically selected by the function *penalized*.[9] The $SCAD$ penalty function helped us to identify the set of observable country-specific variables that actually drives the energy prices because it shrinks small coefficients toward zero and leaves large coefficients not penalized.

The countries for which the eight observable county-specific explanatory variables exhibited non-zero coefficients are reported in Table 5. The main results are summarized as follows. First, Australia, Austria, Czech Republic, Finland, Mexico, Portugal and Romania retained non-zero coefficients for inflation rate, maybe due to the strong correlation between energy prices and inflation rate. Second, the method does not shrink the coefficients of energy imports towards zero in countries surrounded by water such as Canada, Greece and United Kingdom. Third, the energy intensity level of primary energy, which indicates how much energy is used to produce one unit of economic output, is significant in Denmark, Korea, the Netherlands and Norway. Fourth, the

[8]Note that the change in $FEPI$ reflects a ratio that is consistent with usual index calculations when used in regression analysis.

[9]We followed McIlhagga (2016) and optimized κ_{c_i} using the MATLAB toolbox *penalized*.

energy price indices for Croatia, Italy and the US are influenced by GDP per unit of energy use. Fifth, renewable energy consumption explains the energy prices in Belgium, which has reduced its use of fossil fuels; Brazil, which is well-known as a low carbon economy; Sweden, which is a global leader in decarbonization; and Switzerland, which has a carbon-free electricity sector dominated by nuclear and hydro generation. Sixth, population growth is significant in Germany, Japan, and Poland. Seventh, electricity production from renewable sources is significant in France and New Zealand, which relies on hydropower and geothermal energy. Finally, the increase in renewable and waste resources, which enable higher efficiency and lower environmental impact engines, drives energy prices in Cyprus, Hungary, Slovakia and Turkey.

Given the estimates of the regression coefficients, we fitted the group-specific factor model stated in (4) to estimate an initial set of global factors and their corresponding factor loadings. For this purpose, we considered the two-step estimation procedure proposed in Lam and Yao (2012) for the presence of factors with different degrees of strength. In Figure 5, we ploted the ratios of consecutive eigenvalues (in descending order) for the combined dynamic correlation matrix (5). It is clear that the ratio-based estimators provide evidence in favor of choosing an initial value for r_0 of two common factors, one common factor in each step. Thus, we used principal components to obtain an initial estimate of the two factors $\hat{f}_{0,t}$ and their respective loadings $\hat{\Lambda}_0$.

The first estimated common factor explains 77.5% of the total variance of $y^*_{it} = y_{it} - x'_{c_i,t}\hat{\beta}_{c_i}$. The loadings this factor on the 360 energy price indices are plotted in Figure 6 as a bar chart, where the price indices are grouped by country to make them easier to understand. This factor loaded positively on all the countries except Brazil, indicating that this factor will reproduce the worldwide evolution of industrial energy price variations.

The dynamics of the first two estimated common factors appear in the top panel of Figure 7. Visually, the evolution of the first factor, blue line, and the average price index displayed in Figure 4 is very similar across the entire sample and these two time series have a correlation exceeding 0.99. Thus, we interpreted the estimated world common component as an aggregate that captured the evolution of the worldwide industrial energy price dynamics.

The second estimated factor shows fairly different factor model characteristics. This factor explains only 6.6% of the total variability of y^*_{it}. Figure 6 reports that the factor loaded positively on some countries and negatively on others, with a quite heterogeneous magnitude. In addition, the time series evolution of the factor, displayed in Figure 7, did not allow us to interpret this factor as a worldwide price energy factor. According to the estimation steps stated in Section 2, this initial estimate of the second factor was a candidate to be recombined through the next steps.

Given the value of the country-specific regression coefficients and the common factor structures, we calculated the generalized pairwise cross correlation for the components of the common factors to estimate the number of groups S and the group membership G using the Silhouette statistic. We added the restriction that the clusters must have a minimum size to avoid spurious cluster formation. In particular, we implemented this restriction by omitting time series in the dendrogram with a relatively small dependency on the rest (90% percentile of the dendrogram's unions), which were reassigned to the closer clusters.[10] In addition, we imposed that each cluster should have a

[10]We checked that cluster formation was robust to a wide range of restrictions.

minimum size constraint of 5% of the total industrial energy price indices. The Silhouette statistic estimated that the optimal number of groups in the international industrial energy prices is 2 out of a maximum number of 20 groups.

Using the number of groups and the group membership, we obtained the principal component's estimate of the group-specific factors. For this purpose, we determined that the possible number of group-specific pervasive factors r_j would range from 0 to 20, with $j = 1, 2$. Thus, we chose the initial set of group-specific factors by using the ratios of two consecutive eigenvalues of the group-specific combined dynamic correlation matrices.

Finally, we re-examined the factors to decide whether each of the two common factors was actually global or specific. As expected, using the empirical canonical correlation between each common factor and the set of specific factors, we decided that the first factor was actually global and reallocated the second common factor to the first group. Finally, the specific factors were re-estimated and the final estimated numbers of groups-specific factors were $\hat{r}_1 = \hat{r}_2 = 1$, explaining 27.6% and 46.3% in the variance of the energy prices that belong to each group, respectively. The estimated group-specific factors were plotted in the bottom panel of Figure 7.

To provide the group formation with economic meaning, Figure 8 reports a classification matrix that displays the group membership of the 360 energy prize indices using colors. At first glance, the table paints a picture of the classification results. Each cell represents an energy price index and is shown in orange if it is assigned to Group 1 and in blue if it is assigned to Group 2. Because the country affiliations and sector classifications are known, the two-way table of the estimated group membership g_i against these classifications can be seen in the rows and columns in the table. It shows that the set of energy prices is partitioned into 30 disjoint classes when the focus is on the rows, and into 12 disjoint classes when the focus is on the columns.

To examine whether the group formation agreed with country affiliation or with sector classification, we focused on the concept of entropy. This is a general measure of impurity in classification statistics when the data are classified into different categories and grouped in disjoint classes. Let E_j be the entropy of class j and let p_i^j be the frequentist probability of group i of price indices lying in class j, with $i = 1, 2$, and $j = 1, \ldots, 12$ when the classifications are analyzed through sectors and $j = 1, \ldots, 30$ when they are analyzed through countries. The entropy, defined as

$$E_j = -p_1^j log_2(p_1^j) - p_2^j log_2(p_2^j), \tag{9}$$

appears in the last column and in the last row of Figure 8.

This quantity ranges from 0 to 1, with 0 indicating that the class is pure in the sense that it contains the energy price indices for only one group, and 1 indicating that the energy price indices are distributed equally between the two groups. For example, the entropy of the last row of Figure 8 was 0 because all the sectors in the United States belonged to Group 1, while the entropy of France was 1 because the factor model assigned half of the sectors to each group.

On average, the entropy of the participation of energy prices by countries was 0.46 while it reached 0.85 when the partition is by sectors. This implies that country affiliation was significantly more decisive than sector association when determining group classification. This result agreed with Sato et al. (2019), who found that country variation matters more than sector variation as a

source of cross-sectional variation in energy prices.

The size of Group 1 was 116 while Group 2 was about 50% larger and contained 244 energy prices indices. This implies that Group 2's degree of heterogeneity was much larger than that of Group 1. In particular, Group 1 included US energy prices, together with those of its largest trading partner, Mexico. It also included the two allied countries Cyprus and Greece, which together with Israel, form the natural gas extraction plan, the so-called Energy Triangle.

Group 1 also included most energy prices in Croatia, Romania and the Netherlands. The first two countries, Croatia and Romania, showed higher levels of energy poverty compared to the rest of the European Union members, as documented by Lenz and Grgurev (2017), and the latter country, is a major producer of natural gas in Europe.

Although most of the central European countries lay in Group 2, this group was more heterogeneous, and the membership results did not support a common energy policy in the European Union, given that some European Union countries did not belong to this group. Interestingly, the group included Finland, Norway and Sweden, environmentally sensitive countries that are legislating to become zero-carbon economies. Finally, the Southeast Asian countries considered, Japan, Korea, Australia and New Zealand, were grouped together as representatives of the trade agreement partners involving the Association of Southeast Asian Nations (ASEAN).

The time series evolution of the group-specific factors is displayed in the bottom panel of Figure 7. The factor of Group 1, the orange line, captures the North American natural gas crisis from 2000-2008, and shows the ups and downs in the price of natural gas. The co-movements across the energy prices included in this group reveals that this group experienced the consequences of the financial crisis with a delay, with their decline starting in 2010. They have still not recovered since then.

By contrast, the factor of Group 2, the blue line, shows a sharp decline from 1998 to 2000 and in 2008. Given that this cluster includes Southeast Asian countries, the drop at the beginning of the sample may represent the decline in crude oil prices due to the impact of the Asian financial crisis in 1997 (see OECD, 2007). The figure also shows a rise in energy prices, reflecting the increase in the price of crude oil in the UK in 2000.

We are now in the condition to answer the question of the manuscript's title: "What drives industrial energy prices?" To address this issue, we first compute the percentage of the variance in the energy prices that the independent variables collected in $x_{c_i,t}$, $\hat{f}_{g_i,t}$, and $\hat{f}_{0,t}$ explain collectively. The R-squared statistics is 76,4%, which indicates that the proposed model fits the data well. Now, we regress the energy prices on the country-specific explanatory variables, the set of estimated group-specific factors and the estimated common factor, achieving R-squared statistics of 14.6%, 15.3%, and 46.4%, respectively.

The prominent role of worldwide factors to explain energy prices is consistent with the findings provided in Mumtaz and Theodoridis (2017). The latter found that the cross-country co-movement in volatility of real and financial variables had increased over time with the common component becoming more important over the last decade. In a more specific setup, Bastianin et al. (2019) also found the existence of price-growth convergence in natural gas prices.

This result may contain important implications and guidelines for policy authorities trying to

mitigate the adverse effects of long periods of significant escalation in energy prices. In order to maximize the effectiveness of their measures, our results suggest the need to shift from country-specific strategies to worldwide measures common to all the industrial energy prices, regardless of sector, country or group.

5 Conclusion

In this paper we postulate that there are four potential driving forces to consider: (i) country-specific explanatory variables, which affect the energy prices of a given country; (ii) group-specific unobserved drivers, which impact on energy prices that exhibit a strong cross-correlation between the series in the same group and are separated from the cross-correlation observed in other groups; (iii) global or worldwide unobserved drivers, which are common to all the industrial energy prices, regardless of sector, country or group; and (iv) idiosyncratic unobserved components, which refer to the particular dynamics of the industrial energy price indices.

With the aim of shedding some lights on this issue, we proposed a novel extension of the dynamic factor model with group structure advocated by Alonso et al. (2020) to account for observable country-specific explanatory variables that potentially drive the industrial energy prices. We also proposed iterating the model-selection and model-estimation algorithm to overcome consistency problems. Using Monte Carlo simulations, we documented a remarkable finite sample performance of our proposal.

Our data source is the novel repository provided by Sato et al. (2019), which was constructed to enable international comparisons. The effective data set used in the empirical analysis contained sector level energy prices for 12 industrial sectors in 30 countries for the period 1995 to 2015, where the price indices were constructed as weighted averages of fuel-specific prices by fuel consumption.

We found a variety of results, of which we have only briefly mentioned three. Firstly, price inflation and the use of renewable and waste resources played a prominent role in the country-specific drivers of energy prices. Secondly, we estimated a common-to-all energy price indices factor that captured the major energy price developments over the period from 1995 to 2015. Thirdly, we found two distinct group of energy price indices. A two-way table of the grouping output against the energy prices classified by countries and sectors showed that our procedure is successful in recognizing the fundamental differences of energy price indices across countries. Further research may be able to address how to forecast fuel prices using country-specific, group-specific and global drivers.

In summary, this article has three main contributions: firstly, it adds, to the academic literature investigating worldwide energy prices, and the study of global and group-specific co-movements together with macroeconomic variables; secondly, the clustering helps to provide a better understanding of the underlying connections of energy prices, which is crucial for public and private investment; thirdly, the results help public policy decision makers to understand the hidden international dynamics in the energy market, given the importance of trading as a pillar of energy policy.

References

[1] Aastveit, K., Bjornland, H., and Thorsrud, L. 2015. What drives oil prices? Emerging versus developed economies. *Journal of Applied Econometrics*, 30(1): 1013-1028.

[2] Ahn, S., and Horenstein, A. 2013. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3): 1203-1227.

[3] Alonso, A., and Peña, D. 2019. Clustering time series by linear dependency. *Statistics and Computing*, 29(4): 655-676.

[4] Alonso, A., Galeano, P., and Peña, D. 2020. A robust procedure to build dynamic factor models with cluster structure. *Journal of Econometrics*, 216(1): 35-52.

[5] Ando, T., and Bai, J. 2016. Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1): 163-191.

[6] Ando, T., and Bai, J. 2017. Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *Journal of the American Statistical Association*, 112(519): 1182-1198.

[7] Bai, J. 2009. Panel data models with interactive fixed effects. *Econometrica*, 77(4): 1229–1279.

[8] Bai, J., and Ng, S. 2002. Determining the number of factors in approximate factor models. *Econometrica*, 70(1): 191-221.

[9] Bai, J., and Ng, S. 2013. Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1): 18-29.

[10] Bastianin, A., Galeotti, M., and Polo, M. 2019. Convergence of European natural gas prices. *Energy Economics*, 91: 793-811.

[11] Baumeister, Ch., Korobilis, D., and Lee, T. 2022. Energy Markets and Global Economic Conditions. *The Review of Economic and Statistics*, forthcoming.

[12] Bertrand, V. 2014. Carbon and energy prices under uncertainty: A theoretical analysis of fuel switching with heterogeneous power plants. *Resource and Energy Economics*, 38: 198-220.

[13] Binder, C., and Makridis, Ch. 2022. Stuck in the Seventies: Gas Prices and Consumer Sentiment. *The Review of Economic and Statistics*, forthcoming.

[14] Bretschger, L. 2015. Energy prices, growth, and the channels in between: Theory and evidence. *Resource and Energy Economics*, 39: 29-52.

[15] Caro, A. and Peña, D. 2021. A test for the number of factors in dynamic factor models. Working Paper of the Statistics Department at Carlos III University of Madrid.

[16] Correia Fernandes, M., Dias, J., and Vidal-Nunes, J. 2021. Modeling energy prices under energy transition: A novel stochastic-copula approach. *Economic Modelling*, 105: 105671.

[17] Dahl, C. 2015. *International energy markets: understanding pricing, policies, and profits*. PennWell Books.

[18] Fan, J., and Li, R. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348-1360.

[19] Gao, S., Hou, Ch., and Nguyen, B. 2021. Forecasting natural gas prices using highly flexible time-varying parameter models. *Economic Modelling*, 105: 105652.

[20] Ghoshray, A., and Johnson, B. 2010. Trends in world energy prices. *Energy Economics*, 32(5): 1147-1156.

[21] Griffin, J. 1980. *Energy Economics and Policy*, (New York: Academic Press, 1980).

[22] Hailemariam, A., and Smyth, R. 2019. What drives volatility in natural gas prices? *Energy Economics*, 80: 731-742.

[23] Hubert, L., and Arabie, P. 1985. Comparing partitions. *Journal of Classification*, 2(1): 193-218.

[24] Lam, C., and Yao, Q. 2012. Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2): 694-726.

[25] Lenz, N. V., and Grgurev, I. 2017. Assessment of energy poverty in new European union member states: The case of Bulgaria, Croatia and Romania. *International Journal of Energy Economics and Policy*, 7(2): 1-8.

[26] Lyu, Y., Yi, H., and Yang, M. 2021. Revisiting the role of economic uncertainty in oil price fluctuations: Evidence from a new time-varying oil market model. *Economic Modelling*, 103: 105616.

[27] Mahadevan, R., and Asafu-Adjaye, J. 2007. Energy consumption, economic growth and prices: A reassessment using panel vecm for developed and developing countries. *Energy Policy*, 35(4): 2481-2490.

[28] McIlhagga, W. H. 2016. Penalized: A Matlab toolbox for fitting generalized linear models with penalties. *Journal of Statistical Software*, 72.

[29] Mumtaz, H., and Theodoridis, K. 2017. Common and country specific economic uncertainty. *Journal of International Economics*, 105: 205-216.

[30] OECD, 2007. *OECD factbook 2007: Economic, environmental and social statistic*. OECD Publishing, Paris.

[31] Ozcan, B., Tzeremes, P. and Tzeremes, N. 2020. Energy consumption, economic growth and environmental degradation in OECD countries. *Economic Modelling*, 84: 203-213.

[32] Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53-65.

[33] Sato, M., Singer, G., Dussaux, D., and Lovo, S. 2019. International and sectoral variation in industrial energy prices 1995-2015. *Energy Economics*, 78: 235-258.

[34] Van Benthem, A., and Romani, M. 2009. Fuelling growth: what drives energy demand in developing countries? *The Energy Journal*, 30(3): 91-114.

Table 1: Clustering performance.

	Number of clusters			ARI		
(T, N)	DGP_1	DGP_2	DGP_3	DGP_1	DGP_2	DGP_3
$(50, 300)$	2.91	2.91	2.90	0.99	0.99	0.99
	93	95	92			
$(100, 300)$	2.93	2.99	2.84	0.97	0.98	0.95
	95	98	84			
$(200, 300)$	3.01	3.05	3.01	0.99	1.00	0.99
	99	96	99			
$(50, 600)$	2.89	2.87	2.88	0.98	0.99	0.98
	94	92	91			
$(100, 600)$	2.89	3.00	2.88	0.98	0.99	0.96
	94	100	91			
$(200, 600)$	3.00	3.00	3.00	0.99	1.00	0.99
	100	100	100			

Notes: Letters T and N are time-series and cross-section dimensions, respectively. For each cell in columns 2 to 4, the first row is the mean of the selected number of clusters and the second row is the number of the simulations out of 100 for which the model chooses the true number of clusters, $S = 3$. Columns 5 to 7 report the means of the Adjusted Rand Index measure. The three Data Generating Processes, DGP_1, DGP_2, DGP_3, are stated in the text.

Table 2: Country-specific beta coefficients estimation.

	Mean Squared Error			True Negative Rate		
(T, N)	DGP_1	DGP_2	DGP_3	DGP_1	DGP_2	DGP_3
$(50, 300)$	0.0926	0.0921	0.0931	0.7543	0.7781	0.7490
$(100, 300)$	0.0026	0.0026	0.0026	0.9162	0.9119	0.8762
$(200, 300)$	0.0024	0.0024	0.0024	0.8890	0.9019	0.8714
$(50, 600)$	0.0826	0.0822	0.0823	0.7195	0.7167	0.7162
$(100, 600)$	0.0011	0.0011	0.0012	0.9176	0.9162	0.8976
$(200, 600)$	0.0013	0.0013	0.0013	0.8957	0.8676	0.8971

Notes: Letters T and N are time-series and cross-section dimensions, respectively. The cells in columns 2 to 4 report the averages of the mean squared errors between the true coefficients and their estimates. The cells in columns 5 to 7 report the means of the true negative rates, which measure the proportion of times the proposed criterion is capable of setting to zero the false observable regressors. The three Data Generating Processes, DGP_1, DGP_2, DGP_3, are stated in the text.

Table 3: Clustering performance and country-specific beta coefficients with no group structure.

(T, N)	DGP_1	DGP_2	DGP_3	DGP_1	DGP_2	DGP_3
	Number of clusters			ARI		
$(100, 300)$	1.11	1.03	1.08	1.00	1.00	1.00
	89	97	92			
$(200, 300)$	1.07	1.07	1.09	1.00	1.00	1.00
	94	88	91			
	Mean Squared Error			True Negative Rate		
$(100, 300)$	0.0023	0.0023	0.0024	0.9160	0.9028	0.9058
$(200, 300)$	0.0023	0.0021	0.0022	0.8269	0.8116	0.8310

Notes: See notes of Tables 1 and 2.

Table 4: Countries and industrial sectors acronyms.

Countries		**Sectors**
Australia AUS	Japan JAP	Chemical & petrochemical CHE
Austria AUT	Korea, Republic of KOR	Construction CONS
Belgium BEL	Mexico MEX	Food & tobacco FOOD
Brazil BRA	The Netherlands NLD	Iron & steel IRN
Canada CAN	New Zealand NZL	Machinery MCH
Croatia HRV	Norway NOR	Mining & quarrying MIN
Cyprus CYP	Poland POL	Non-Ferrous metals NFER
Czech Republic CZE	Portugal PRT	Non-metallic minerals NMET
Denmark DNK	Romania ROU	Paper, pulp & print PAP
Finland FIN	Slovakia SVK	Textile & leather TEX
France FRA	Sweden SWE	Transport equipment TRNS
Germany DEU	Switzerland CHE	Wood & wood products WOOD
Greece GRC	Turkey TUR	
Hungary HUN	United Kingdom UK	
Italy ITA	United States US	

Note: The table shows the 30 countries and 12 industrial sectors with their respective acronyms.

Table 5: SCAD-penalized regression results.

Energy imports → CAN, GRC, UK
Energy intensity level of primary energy → DNK, KOR, NLD, NOR
GDP per unit of energy use → HRV, ITA, US
Renewable energy consumption → BEL, BRA, SWE, CHE
Inflation rate → AUS, AUT, CZE, FIN, MEX, PRT, ROU
Population growth → GER, JPN, POL
Electricity production from renewable sources → FRA, NZL
Combustible renewable and waste → CYP, HUN, SVK, TUR

Note: Each row refers to a country-specific explanatory variable and the countries for which this variable shows non-zero coefficients.

Appendix

The purpose of this appendix is to compare the performance of the smoothly clipped absolute deviation (SCAD) penalty approach (Fan and Li 2001), which we used in the iterative algorithm, against different regularization techniques, such as the least absolute shrinkage and selection operator (LASSO) and the minimax concave penalty (MCP).

In words, LASSO uses a constant rate of penalization, MCP initially applies the same rate of penalization as the LASSO technique, but continuously reduces the rate of penalization until the rate becomes 0 for a threshold value, and SCAD aims to eliminate unimportant predictors from the model while leaving the important predictors unpenalized.

As the regularization techniques rely on oracle properties for being selection consistent and the oracle properties are asymptotic properties, in this appendix we have carried out a simulation analysis that tries to mimic the simulations and data used in this paper. For this purpose, we relied on DGP_1, where the noise term was homoscedastic and serially uncorrelated, and focused on the combination $(T, N) = (100, 300)$ as Section 3 describes.

Table A1 shows the average across simulations of the Mean Squared Error (MSE) between the true coefficients and their estimates and the proportion of times the proposed criterion is capable of setting the non-relevant observable regressors to zero (True Negative Rate, TNR). Our results suggest that the SCAD penalty outperforms the other two alternative regularization techniques.

Table A1: Penalty regression performance.

	LASSO	MCP	SCAD
MSE	0.0027	0.0027	0.0026
TNR	0.8348	0.6048	0.9129

Notes: Averages across simulations of the Mean Square Error (MSE) and the True Negative Rate (TNR). The Data Generation Process is stated in the text.

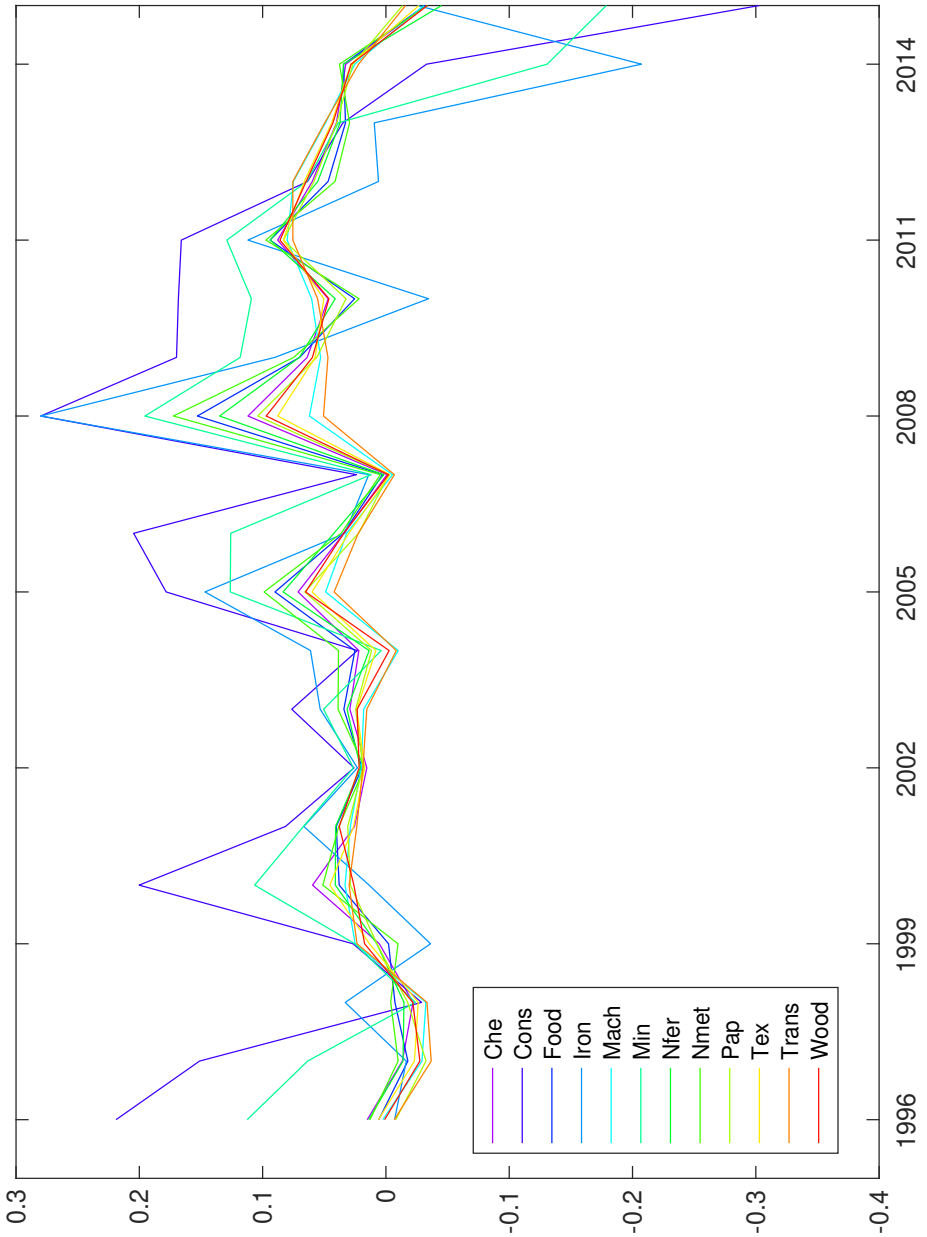


Figure 1: Australian price indeces for the 12 industrial sectors from 1995 to 2015.
Note: Acronyms are listed in Table 4.

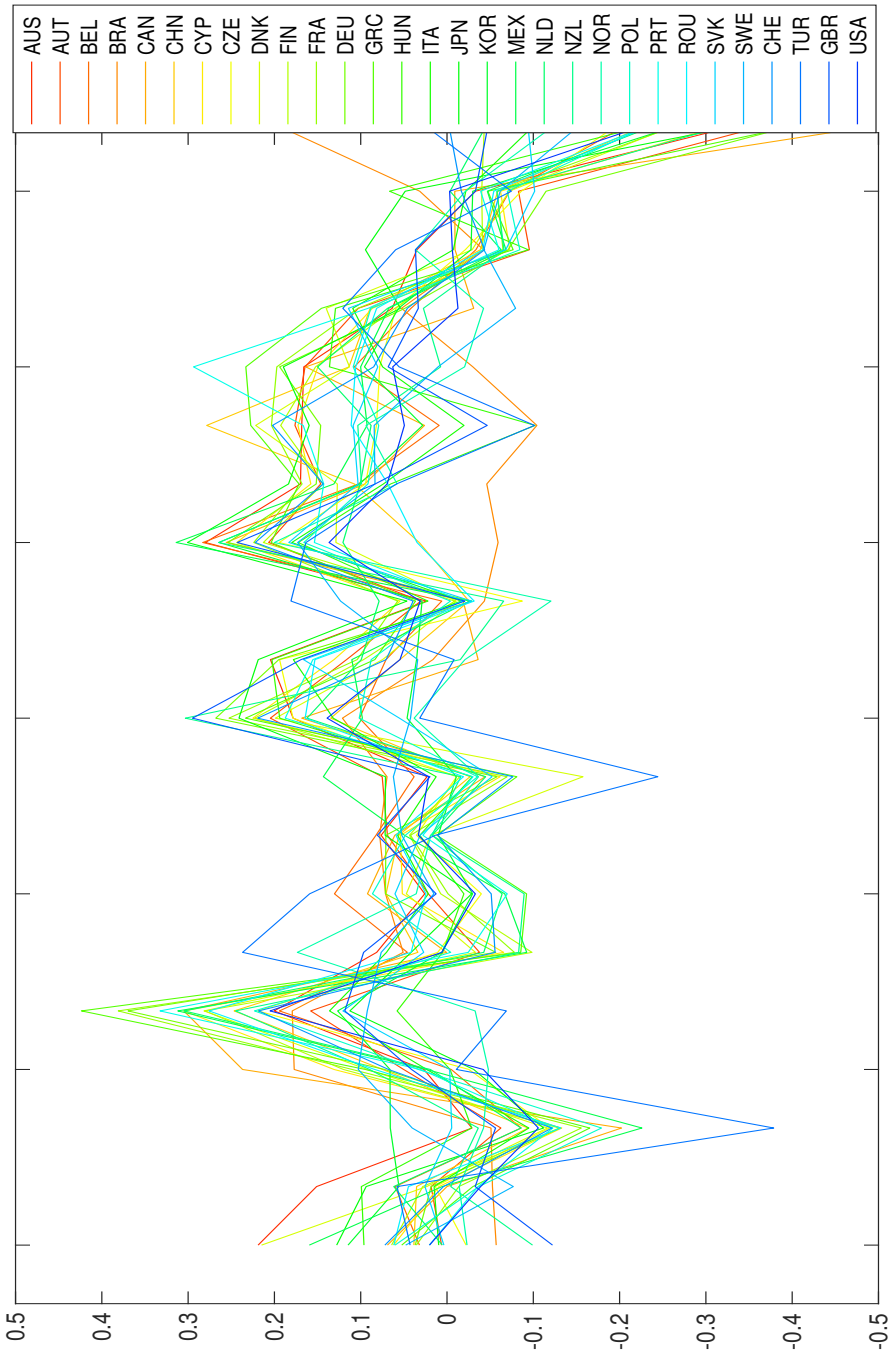


Figure 2: Construction price indeces for the 30 countries of the sample from 1995 to 2015.
Note: Acronyms are listed in Table 4.

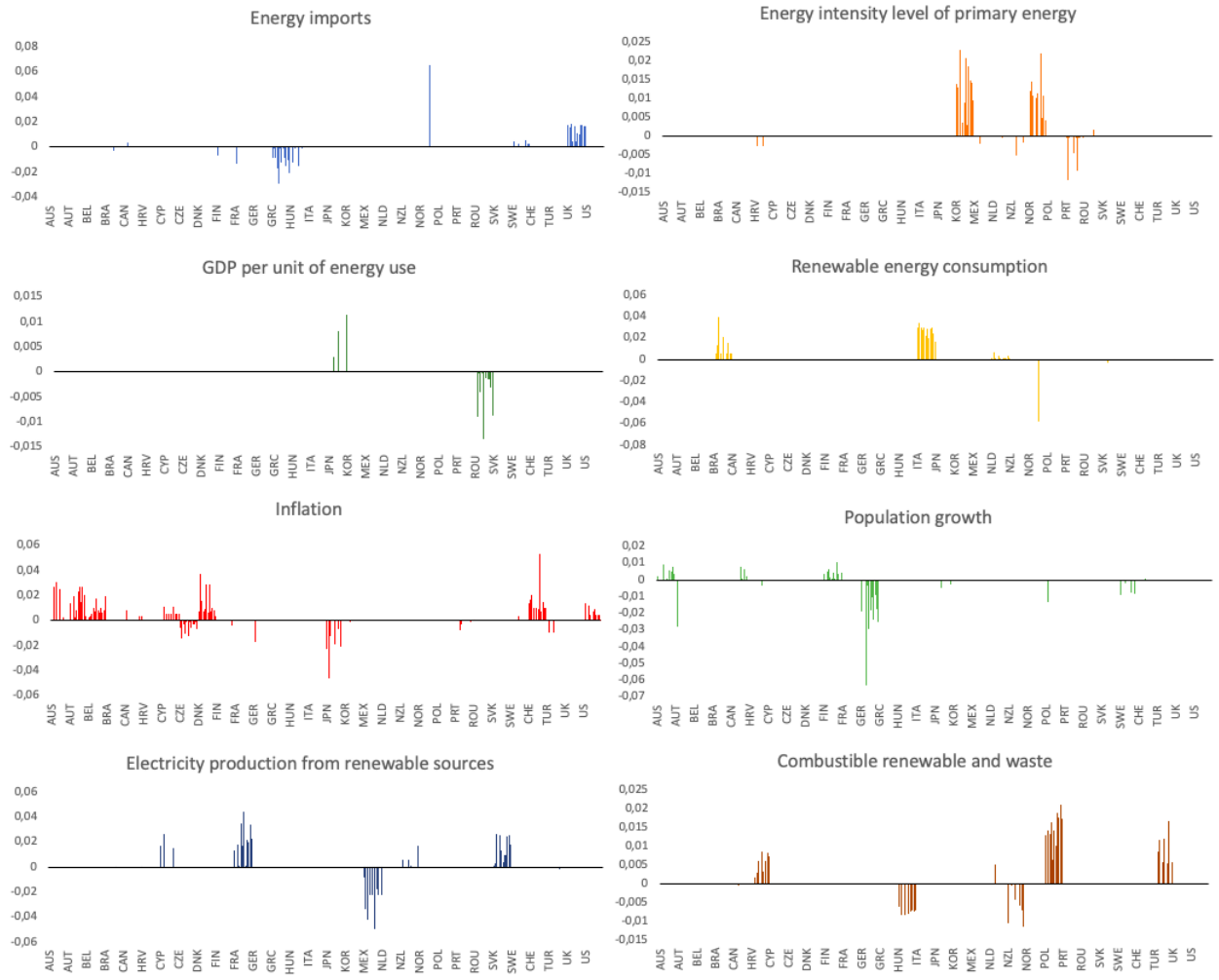


Figure 3: Estimated non-zero impacts of the observed explanatory variables.

Notes: Each plot shows the non-zero impacts of the explanatory variables on the energy price indeces for each country. In the x-axis, the 12 energy price indices are grouped by countries. Acronyms are listed in Table 4.

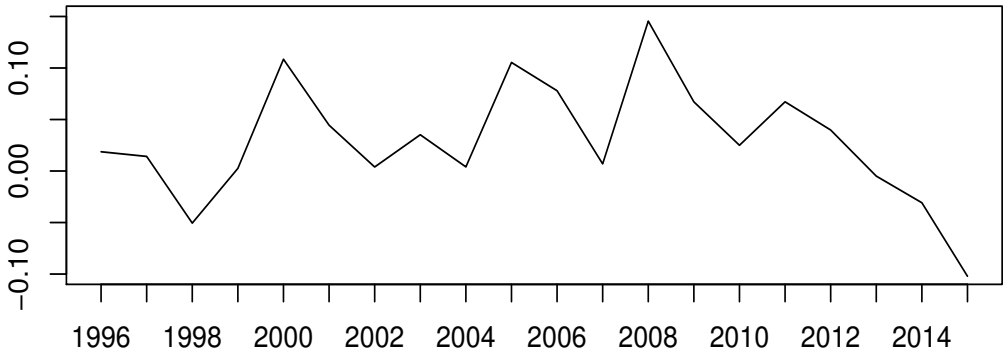


Figure 4: Time series evolution of the cross-section average of the energy price indeces from 1995 to 2015.

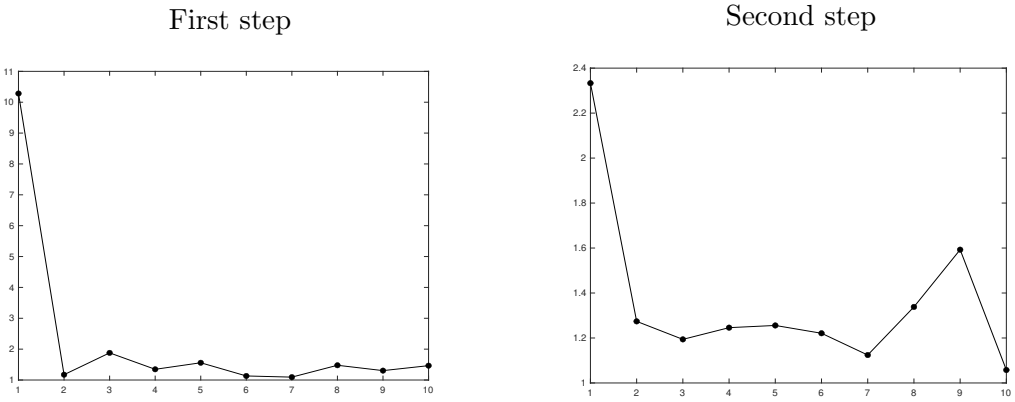


Figure 5: Ratios of two consecutive eigenvalues of the combined dynamic correlation matrix.

Notes: In each plot, the x-axis refers to the number of factors while the y-axis refers to the ratio of two consecutive eigenvalues. The two steps refer to the estimation procedure suggested in Lam and Yao (2012).

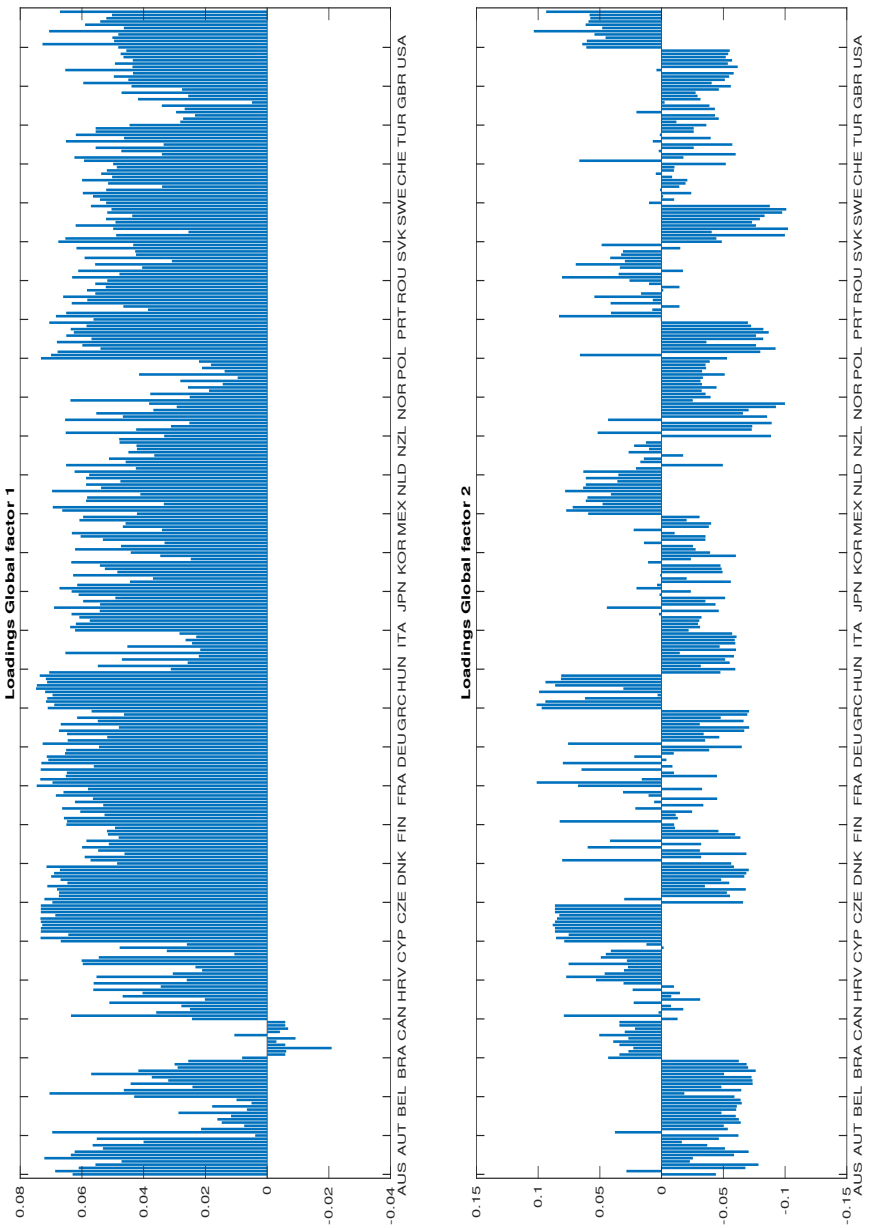
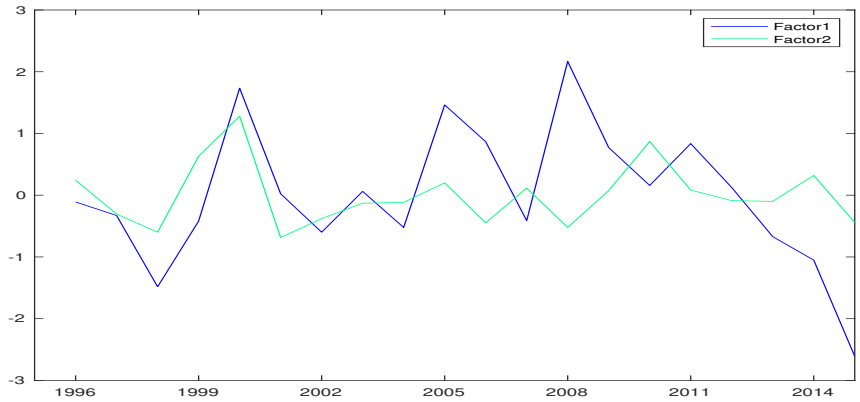


Figure 6: Estimated loadings of the two initial global factors over the 360 energy price indeces.

Notes: In the x-axis, the 12 energy price indices are grouped by countries. Acronyms are listed in Table 4.

Initial common factors



Final group-specific factors

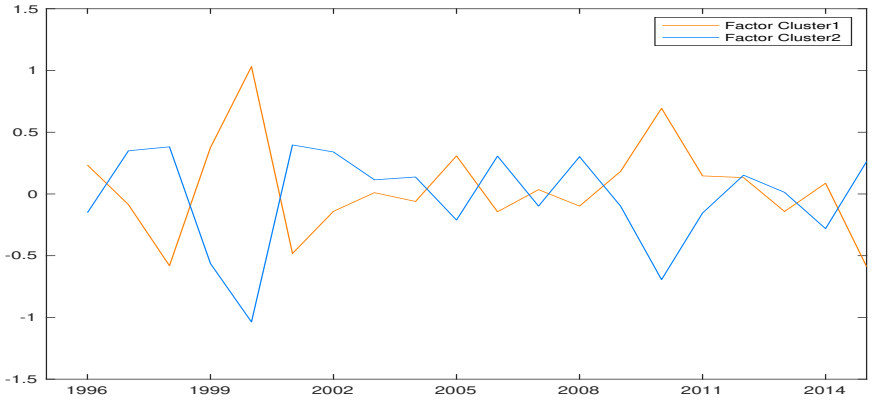


Figure 7: Estimated common and group-specific factors from 1995 to 2015.

Notes: The upper plot presents the estimation of initial common factors. The bottom plot presents the estimation of the group-specific factors.

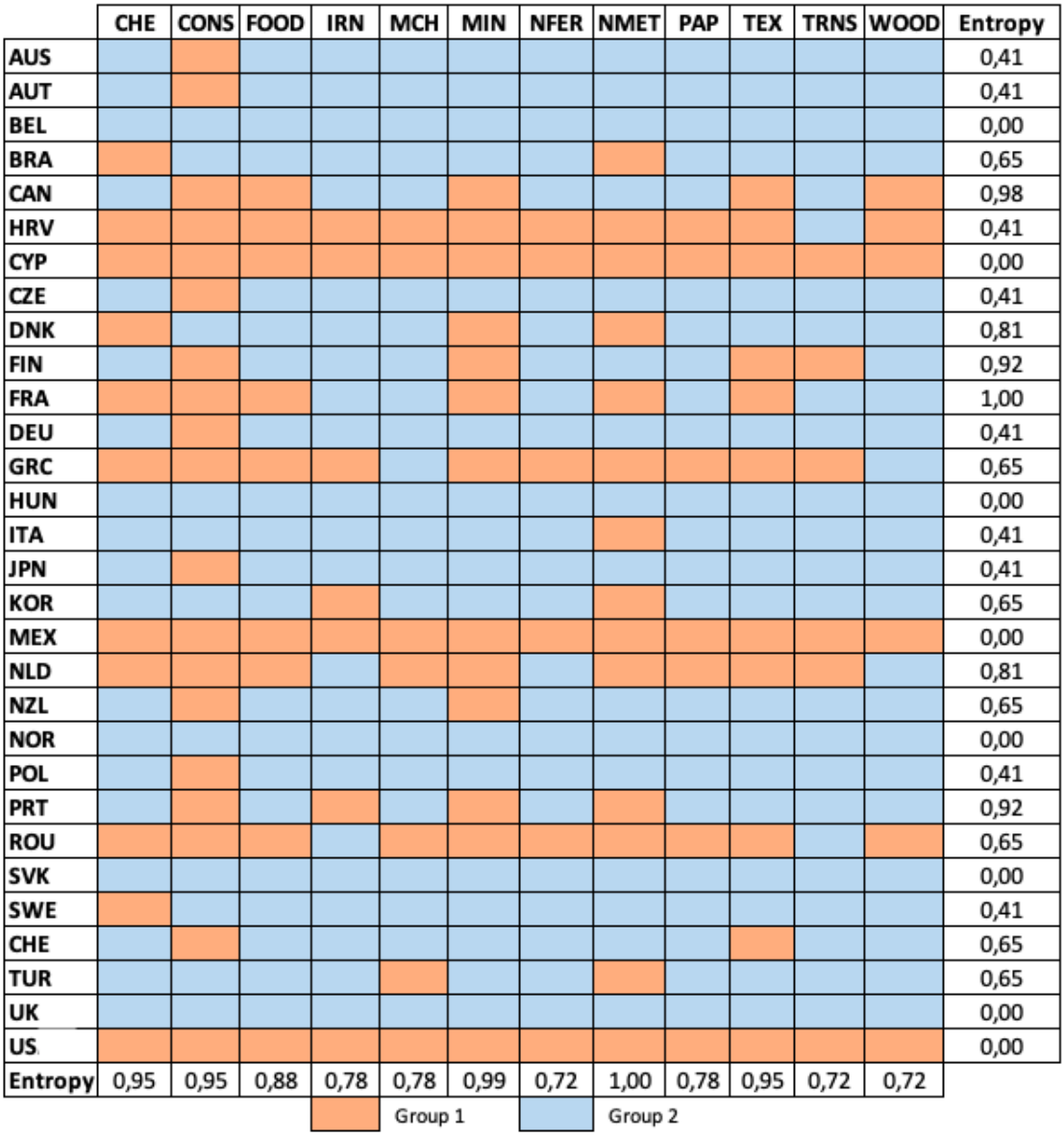


Figure 8: Matrix classification map.

Notes: The 360 price indeces are classified into the two groups determined by the group-specific factor model. In the table's rows, the energy price indeces are grouped by country and in the table's columns are grouped by industrial sectors. Entropy ranges from 0 to 1, with 0 indicating that the class is pure, and 1 indicating complete/total uncertainty/disorder/randomness. Acronyms are listed in Table 4.