Working paper

2020-07

Statistics and Econometrics
ISSN 2387-0303

Discovering general and sectorial trends in a large set of time series

Guillermo Carlomagno, Antoni Espasa

uc3m | Universidad Carlos III de Madrid

Discovering general and sectorial trends in a large set of time series

Guillermo Carlomagno[†] and Antoni Espasa[‡]

[†] *Central Bank of Chile: Agustinas 1180, Santiago, Chile, 8340454 (e-mail:gcarlomagno@bcentral.cl)*

[‡] *Department of Statistics and Instituto Flores de Lemus, University Carlos III of Madrid (e-mail:antoni.espasa@uc3m.es)*

Abstract

The objective of this research note is to extend the pairwise procedure studied by Carlomagno and Espasa (ming) to the case of general and sectorial trends. The extension allows to discover subsets of series that share general and/or sectorial stochastic trends between a (possible large) set of time series. This could be useful to model and forecast all of the series under analysis. Our approach does not need to assume pervasiveness of the trends, nor impose special restrictions on the serial or cross-sectional idiosyncratic correlation of the series. Additionally, the asymptotic theory works both, with finite N and $T \to \infty$, and with $[T; N] \to \infty$. In a Monte Carlo experiment we show that the extended procedure can produce reliable results in finite samples.

Keywords: Cointegration, Factor Models, Disaggregation, Pairwise tests, Heteroskedasticity.

JEL: C01, C22, C32, C53.

I Introduction

In this research note we extend the pairwise procedure of Carlomagno and Espasa (ming) —CE, hereafter— to the case of general and sectroial trends. With the objective of model and forecast all of the components an economic aggregate, CE study a procedure to discover subsets of components that share single common trends while neither assuming pervasiveness of the trends nor imposing special restrictions on the serial or cross-sectional idiosyncratic correlation. An important feature of their asymptotic theory is that it works both with fixed N and $T \to \infty$ and with $[T, N] \to \infty$.

CE focus on the specific case that the data set at hand contains several trends, among which some are common to reduced groups of series, such that each of those groups have only one common trend. To discover those subsets the authors adopt the pairwise procedure initially proposed by Espasa and Mayo-Burgos (2013). It consist of determining the cointegration rank in every possible pair of series, and then forming subsets in which all of the pairs are contegrated. They denote those subsets as *fully cintegrated*. Then, using the outcomes of the cointegration tests and the resulting fully cointegrated subsets, the final phase of the proposal is to estimate a single-equation model for each component, including as potential regressors all of the possibly relevant cointegration relationships found in the previous step. CE show that the pairwise procedure leads to more accurate forecasts of the US CPI than do other alternative methodologies, including dynamic factor models.

When dealing with a large data set of macro variables (not necessarily the components of a single one), the situation could be different. There seems to be agreement in the literature that a *general* factor that affects more or less all variables, plus *sectorial* factors that affect specific subsets is a sensible assumption (see, e.g., Karadimitropoulou and León-Ledesma (2013), Moench et al. (2013), and Breitung and Eickmeier (2015)).

If this is the situation, the pairwise procedure proposed by CE will not be useful. Since the only cointegrated pairs are those formed by series with a single common trend (e.g., series that have only the general factor and no sectorial one), the procedure will be unable to discover the 'true' data structure. The objective of this note is to extend the pairwise approach for this situation.

II Description of the strategy

II.1 General framework and assumptions

The general framework for the models we consider is given by a VAR model for all of the N series under analysis:

$$X_t = \mu_t + \Pi_1 X_{t-1} + ... + \Pi_k X_{t-k} + \epsilon_t \ \Rightarrow \Pi(L)X_t = \mu_t + \epsilon_t, \tag{1}$$

where X_t is a $N \times 1$ vector; Π_i are $(N \times N)$ coefficient matrices; ϵ_t is a vector of innovations with covariance matrix Σ_t; μ_t contains the deterministic components (constants, trends, seasonal dummies, and outlier and break indicators); $\Pi(z)$ is the characteristic polynomial; and L is the lag operator. If the system is cointegrated, it can be formulated as a vector equilibrium correction model (VEqCM):

$$\Delta X_t = \mu_t + \alpha\beta' X_{t-1} + \Phi_1 \Delta X_{t-1} + ... + \Phi_{k-1} \Delta X_{t-k+1} + \epsilon_t, \tag{2}$$

where α and β are $N \times r$ matrices, with $0 < r < N$; r is the number of cointegration relationships; $\alpha\beta' = -I_n + \Pi_1 + ... + \Pi_k$; and $\Phi_i = -\sum\limits_{j=i+1}^{k} \Pi_j$. The data structure for which our procedure is designed can be summarized in five assumptions:

Assumption 1 The N components are generated by the VEqCM in equation (2).

Assumption 2 The N components are $I(1)$.

Assumption 3 There is one subset of G series (with $0 \leq G < N$) in which the series share a unique common stochastic trend. We denote this trend a "global".

Assumption 4 There are J subsets (with $0 \leq J < N$), each one of size s_j (with $0 < s_j$, and $\sum_j s_j \leq N - G$) in which the series share two common trends, the global one, and one "sectorial".

Assumption 5 The innovations —say, e_t— of the bivariate systems that derive from equation (2) satisfy certain conditions that depend on the strategy used to test cointegration. The specific conditions are discussed below.

Assumption 6 The asymptotic behavior of the ratios $\dfrac{N}{G}$, $\dfrac{N}{s_j}$ and $\dfrac{N}{T}$ is limited by some rule that depends on the strategy used to test cointegration. The specific rules in each case are discussed below.

Remark 1 Apart from having all of its roots outside the unit circle, there is no restriction on the polynomial $(I - \Phi_1 L - ... - \Phi_k L)$, and the covariance matrix of ϵ_t (Σ) has no particular restrictions. Thus, we are not imposing any additional restrictions on the serial or cross-correlation of the series.

Remark 2 As stated in assumption 5, depending on the strategy used to test cointegration, we need the innovations of the bivariate models that derive from equation (2) to satisfy certain conditions. The large system in equation (2) is not estimable by conventional methods in real macroeconomic applications. As a consequence, the properties of its innovations, which are empirically untestable, are not our main interest. Instead, we specify conditions on the innovations of the bivariate systems, which are both empirically testable and sufficient for the asymptotic validity of our proposal.

II.2 The algorithm

We will use the notation G both as the name of the global-trend-subset that exist in the true DGP and as its cardinality. Analogously, we will use s_j as the name and cardinality of the subsets with the global and sectiorial trends. For the subsets constructed by the algorithm that we present below, we will use the notation $\hat{G}$ and $\hat{s}_j$. As we show in next subsection, under assumptions 1 to 6, the following algorithm can be applied to discover the subsets G and s_j:

(i) Apply the pairwise procedure of CE. This will lead us to discover the subset $\hat{G}$. (ii) Test for cointegration in all of the possible triplets formed by one series inside $\hat{G}$ and a pair of outsiders. For the triplets in which the outsiders have the same sectorial trend, we will find one cointegration relationship (two common trends). (iii) Construct a $(N - \hat{G}) \times (N - \hat{G})$ symmetric *adjacency matrix* for the series outside $\hat{G}$ such that each cell of this matrix represents a pair of the components outside $\hat{G}$. Each of those pairs belongs to $\hat{G}$ different triplets: one for each element of $\hat{G}$. Then, in each cell of the adjacency matrix, put a 1 if all of the corresponding $\hat{G}$ triplets have just one cointegration relationship; otherwise, put a 0. (iv) In previous adjacency matrix, look for the largest possible submatrix that is full of ones, using, for example, the Bron-Kerbosch algorithm (see Bron and Kerbosch, 1973 and Eppstein et al., 2010). This will lead us to discover the series in each sector.

Remark 3 By theorem 1 in CE, in point iii above, it would be asymptotically irrelevant if in testing cointegration in a given triplet formed by a pair outside $\hat{G}$ and an element inside $\hat{G}$ we do it (a) with all of the series in $\hat{G}$, (b) with some of them, or (c) with the estimated global trend in $\hat{G}$. When dealing with finite samples, requiring to find one cointegration relationship in all of the $\hat{G}$ triplets that contain the same pair of series outside $\hat{G}$ and one series of that subset (case a) may be too stringent. Instead, we could relax this requirement by allowing *a few* of those triplets to fail in showing the existence of one cointegration relationship (case b), or by testing testing cointegration in only one triplet (case c). We study cases a and b in next section.

This procedure contributes to the literature in one relevant aspect: while the usual practice is to assume the sectorial structure as given, we can estimate it. Ando and Bai (2015) estimate the sectorial structure but for stationary variables, with a size of sectors that goes to infinity (in their simulation experiments the smallest sector has 100 units) and restricted serial and cross-correlation of the error terms. The Global VAR models proposed by Pesaran et al. (2004) are also related to our proposal. Among other relevant differences, we determine the 'regions' (sectors) statistically and do not have restrictions on the number of variables per region, which can be large.

III Asymptotic properties

In this section we argue that our procedure is asymptotically valid to discover subsets with general and sectorial trends.

We evaluate the asymptotic properties in three dimensions: (i) *Potency:* The proportion of correct series that are included in $\hat{G}$ and $\hat{s}_j$. (ii) *Gauge:* The proportion of wrong series that are included in $\hat{G}$ and $\hat{s}_j$[1]. (iii) *False discovery:* The discovery of subsets in which none of the pairs is cointegrated.

Studying false discoveries and gauge separately is relevant, in as much as it allows us to explicitly analyze how the procedure would work when there are no truly cointegrated pairs.

[1] The terms "gauge" and "potency" are borrowed from Castle et al. (2011).

The asymptotic theory developed in CE applies directly to our case of interest. This means that, under the conditions we comment below, cointegration ranks can be determined either by the Johansen test, the information criteria procedure proposed by Cavaliere et al. (2016), or the nonparamteric approach of Poskitt (2000) and Athanasopoulos et al. (2016).

In the case of the Johansen and the nonparametric approaches, the innovations of the bivariate and trivariate subsystems must be iid. The information criteria approach allows a wide class of conditional and unconditional heteroskedasticity that include multiple covariance shifts, variances with broken trends, smooth variance shifts, and GARCH and stochastic volatility processes (see assumption H in Cavaliere et al., 2016).

When N is finite and T goes to infinity, those conditions on the innovations, together with assumptions 1 to 4, are sufficient to show that the three alternatives have high potency, and low gauge and false discovery (see details in CE).

When $[N, T] \to \infty$, we need assumption 6 in order to control gauge and false discovery. When using the Johoansen approach we need three conditions: $\frac{G}{N^{1-1/\kappa}} \to\geq c$, $\frac{s_j}{N^{1-1/\kappa}} \to\geq c$, and $\frac{T}{N^{1/\kappa}} \to\geq c$, for some $c > 0$ and $\kappa > 0$. Under the information criteria approach, when using the BIC, we need the ratios N/G and N/s_j to be $Op(log(T))$. Finally, in the nonparametic approach, we need the ratios N/s_j and N/G to $Op(P_T/loglog(T))$, where P_T is the penalty function associated with the nonparametric strategy (Poskitt, 2000). See CE for the technical details of these conditions.

IV Monte Carlo experiments

As we mentioned above, the generalization for the case of general and sectorial trends requires testing cointegration not only in pairs, but also in some triplets of series. Thus, the computational cost rises with respect to the pure pairwise approach. Assume a case with $N = 100$, $G = 10$, and $\hat{G} = 10$. After testing cointegration in all of the $4,950$ pairs that exist between the 100 series, the procedure requires making other $10 \times 90(90-1)/2 = 40,005$ cointegration tests. However, as highlighted in Remark 3, this issue could be mitigated by not testing cointegration in *all* of the possible triplets. We do not explore this possibility here.

We consider two DGPs. The DGPs are DGP-1 and DGP-3 described in CE, modified to have general and sectorial trends. We denote these modified DGPs as DGP-a and DGP-b, respectively. DGP-a represents a process in which each variable in G, or in some sector, reacts only to one cointegration relationship, and idiosyncratic components are independent. In DGP-b, variables in G or in some sector react to more than one cointegration relationship, and there is idiosyncratic cross-correlation between all of the N variables. The rest of this subsection is devoted to describe the DGPs in more detail.

Let s_i be the number of variables that, in addition to the general trend, also have the trend of sector i. Using the same normalization for matrix β as in CE, without loss of generality, we normalize all cointegration relationships with respect to one of the variables in G. To have a simple example of $\beta' s$ structure, assume that $N = 10$, $G = 3$, $s_1 = 3$, $s_2 = 3$, and that the remaining series have its own trend. In this case, we can set β such that:

$$\beta' = \begin{pmatrix} \beta_{11} & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{21} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta_{31} & 0 & 0 & \beta_{34} & 1 & 0 & 0 & 0 & 0 & 0 \\ \beta_{41} & 0 & 0 & \beta_{44} & 0 & 1 & 0 & 0 & 0 & 0 \\ \beta_{51} & 0 & 0 & 0 & 0 & 0 & \beta_{57} & 1 & 0 & 0 \\ \beta_{61} & 0 & 0 & 0 & 0 & 0 & \beta_{67} & 0 & 1 & 0 \end{pmatrix} \tag{3}$$

An important difference with respect to CE is that we cannot set the coefficients β_{ij} equal to -1 because the series in G would be cointegrated with all of the other series in the system. To avoid this, we need some variation in the coefficients β_{ij}. Thus, we take those coefficients from the uniform distribution with parameters $[-5, -0.1]$. For DGP-a (DGP-b), matrix Φ_1 is generated in the same way as in DGP 1 (DGP 3) in CE.

For DGP-a, matrix α has exactly the same structure as in DGP 1 of CE, except that the number of columns (r) is now $G + s_1 + s_2 - 3$. With this structure, the series in first position of G, s_1 and s_2 are weakly exogenous. The other series, react to a single cointegration relationship that affects itself, the series in the first position of G, and, for the series that belong to some sector, the first series of the sector (see matrix β in equation (3)).

In DGP-b, we set matrix α such that each variable j belonging to G or to some sector reacts to $q_j + 1$ cointegration relationships and there are no weakly exogenous variables.

To get a visual example, assume that apart from the general common trend, there are two sectorial ones. In this case, matrix α can be partitioned as follows:

$$\alpha = \begin{bmatrix} A1_{G\times(G-1)} & 0 & 0 \\ 0 & A2_{s_1\times(s_1-1)} & 0 \\ 0 & 0 & A3_{s_2\times(s_2-1)} \\ 0 & 0 & 0_{(N-(G+s_1+s_2))\times(s_2-1)} \end{bmatrix}, \tag{4}$$

where $A1$, $A2$ and $A3$ have the same structure as matrix α of DGP 3 in CE.

We consider four scenarios and one sample size, $T = 400$. In the four scenarios, there is a single general trend, two sectors, and some series with their own trends. In *scenario 1* we set $N = 35$, $G = 10$, $s_1 = 10$, $s_2 = 10$, and the remaining five series have their own trends. In *scenario 2* we add more noise; instead of only five series with their own trends, we have 30, thus, $N = 60$. In *scenario 3*, $N = 80$, $G = 25$, $s_1 = 25$, $s_2 = 25$, and the remaining five series have their own trends. In *scenario 4*, $N = 105$, $G = 25$, $s_1 = 25$, $s_2 = 25$, and the remaining 30 series have their own trends.

We use scenarios 1 and 2 both for DGP-a and DGP-b. For saving computing time, we simulate scenarios 3 and 4 only for DGP-b.

Results

In this section we apply the algorithm described in section II.2 to the simulated data. For determining the cointegration rank of the pairs and the triplets, we use only the Johansen test. Table 1 includes the *gauge* and *potency*, false discovery is very small in all cases, so we do not report it. Figures under 'Sectors' columns are averages for the two sectors. As the table shows, in general, the procedure has high potency for discovering the true series in each sector, with little cost in terms of gauge.

In DGP-a ('simple' matrix α and diagonal Φ_1), potency for G is close to 99% in both scenarios, and gauge is 1%. For the sectors, when we require a cointegration relationship in *all* of the triplets formed by a pair of series outside $\hat{G}$ and each of the insiders, potency is somewhat lower, but still high (92%). When we allow some of those triplets to fail in showing a cointegration relationship (see remark 3), potency figures of the sectors get close to those of G. This improvement in potency is costless in terms of gauge.

In DGP-b (‘complex’ matrix α and non-diagonal Φ_1), potency for G is almost the same as in DGP-a. Gauge is somewhat larger, but we still have acceptable low figures. Potency results for the sectors show a relevant deterioration that is mitigated by allowing some failures in the cointegration tests of the triplets. This improvement in potency is costless in terms of gauge, which is somewhat larger than in DGP-a but is still acceptably low. Note that a gauge of 0.05 in scenarios 1 and 2 implies an average of 0.5 wrong series in the estimated subsets. For scenarios 3 and 4 the same gauge implies an average of 1.25 wrong series.

TABLE 1:
Gauge and potency of the pairwise procedure for discovering general and sectorial trends (nominal size $\varphi = 0.01$, $T = 400$)

		Scenario 1		**Scenario 2**		**Scenario 3**		**Scenario 4**	
		G	*Sectors*	*G*	*Sectors*	*G*	*Sectors*	*G*	*Sectors*
DGP-a: All	*Potency*	0.98	0.92	0.99	0.92	—	—	—	—
	Gauge	0.01	0.02	0.01	0.01	—	—	—	—
DGP-a: All but one	*Potency*	0.98	0.95	0.99	0.96	—	—	—	—
	Gauge	0.01	0.01	0.01	0.01	—	—	—	—
DGP-a: All but 2	*Potency*	0.98	0.95	0.99	0.96	—	—	—	—
	Gauge	0.01	0.01	0.01	0.01	—	—	—	—
DGP-b: All	*Potency*	0.97	0.82	0.96	0.73	0.96	0.68	0.96	0.60
	Gauge	0.05	0.03	0.03	0.03	0.04	0.01	0.04	0.01
DGP-b: All but one	*Potency*	0.97	0.88	0.96	0.84	0.96	0.82	0.96	0.76
	Gauge	0.05	0.03	0.03	0.02	0.04	0.01	0.04	0.01
DGP-b: All but two	*Potency*	0.97	0.89	0.96	0.86	0.96	0.87	0.96	0.83
	Gauge	0.05	0.03	0.03	0.02	0.04	0.01	0.04	0.01

- $Gauge = \frac{100}{(N-G)Nexp} \sum_{i=1}^{Nexp} Z_{2,i}$. - $Pot = \frac{100}{GNexp} \sum_{i=1}^{Nexp} Z_{1,i}$. - Z_2 = number of wrong series included in $\hat{n}_1$. - Z_1 = number of correct series included in $\hat{n}_1$. - $Nexp$ = number of experiments (500). - G is the group of series that have the general trend only. - *Scenario 1:* $N = 35$, $G = 10$, $s_1 = s_2 = 10$. - *Scenario 2:* $N = 60$, $G = 10$, $s_1 = s_2 = 10$. - *Scenario 3:* $N = 80$, $G = 25$, $s_1 = s_2 = 25$. - *Scenario 4:* $N = 105$, $G = 25$, $s_1 = s_2 = 25$. - Figures in ‘Sectors’ columns are averages for the two sectors. - *All but one* and *All but two* rows indicate that we are allowing one or two triplets to fail in showing a cointegration relationship (see remark 3).

V Concluding remarks

In this research note we extended the pairwise approach studied by Carlomagno and Espasa (ming) for the case of general and sectorial trends. Our extension allows to discover subsets of series that share general and/or sectorial common trends from a possible large set of time series. The asymptotic theory works both with a fixed corss-sectional dimension and when it goes to infinity, and it does not need to assume pervasiveness neither of the global nor of the sectorial trends. Additionally, the dynamic behavior of the idiosyncratic components is not specially restricted.

We studied the finite-sample properties of our proposal in a Monte Carlo experiment, which show that our approach can also produce reliable results in finite-samples.

References

Ando, T. and J. Bai (2015). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*.

Athanasopoulos, G., D. S. Poskitt, F. Vahid, and W. Yao (2016). Determination of long-run and short-run dynamics in ec-varma models via canonical correlations. *Journal of Applied Econometrics 31*(6), 1100–1119.

Breitung, J. and S. Eickmeier (2015). Analyzing business cycle asymmetries in a multi-level factor model. *Economics Letters 127*, 31–34.

Bron, C. and J. Kerbosch (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM 16*(9), 575–577.

Carlomagno, G. and A. Espasa (forthcoming). Discovering specific common trends in a large set of disaggregates: Statistical procedures, their properties, and an empirical application. *Oxford Bulletin of Economics and Statistics*.

Castle, J., J. Doornik, and D. Hendry (2011). Evaluating automatic model selection. *Journal of Time Series Econometrics 3*(169), 239–246.

Cavaliere, G., L. De Angelis, A. Rahbek, and A. R. Taylor (2016). Determining the cointegration rank in heteroskedastic var models of unknown order. *Econometric theory*, 1–34.

Eppstein, D., M. Löffler, and D. Strash (2010). Listing all maximal cliques in sparse graphs in near-optimal time. pp. 403–414.

Espasa, A. and I. Mayo-Burgos (2013). Forecasting aggregates and disaggregates with common features. *International Journal of Forecasting 29*(4), 718–732.

Karadimitropoulou, A. and M. León-Ledesma (2013). World, country, and sector factors in international business cycles. *Journal of economic dynamics and control 37*(12), 2913–2927.

Moench, E., S. Ng, and S. Potter (2013). Dynamic hierarchical factor models. *Review of Economics and Statistics 95*(5), 1811–1817.

Pesaran, M. H., T. Schuermann, and S. M. Weiner (2004). Modeling regional interdependencies using a global error-correcting macroeconometric model. *Journal of Business & Economic Statistics 22*(2), 129–162.

Poskitt, D. S. (2000). Strongly consistent determination of cointegrating rank via canonical correlations. *Journal of Business & Economic Statistics 18*(1), 77–90.