Advanced Revision Analysis for Economic Time Series and Their Role for Improving Forecast Accuracy

Alain Hecq
Maastricht University
Gian Luigi Mazzi
Eurostat

September 2008
Aim of the paper: we empirically investigate two questions?

1. What is/are the best "vintage" series to be considered:
   - The very last "cuvée" (saw it on Cirano and I like it (or "crus? ")),
   - The one from last year,
   - The last one without the last 3 years or data.

2. And for what purpose?
   - Long-run analyses,
   - Study of short-run co-movements,
   - Forecasting.
To answer these questions we analyze and illustrate what we think to be the most important time series features of «vintages» variables of the data revision process.

The starting point will be "the real-time data matrix" that we split into verticals and diagonals (my names).

result 1 (minor point): Different goals: we use verticals to have a good historical series, diagonals for forecasting, policy making...

RESULT 2: Criteria for selecting good "vintages" (diagonals or verticals):

- The presence of long-run commonalities: cointegration
- The presence of short-run commonalities (but only the WF works)

RESULT 2’: Dimensionality problems: which one to take out of 78 vintages for instance.

- Bivariate analysis first!!
- Univariate analysis even before: The final equation representation of multivariate systems to select series.
The starting point is the collection of the statistical information (the data) in a «real-time data set». (Yes we include the forecasts in this representation).

<table>
<thead>
<tr>
<th>$X_{t,v}$</th>
<th>1999Q2</th>
<th>1999Q3</th>
<th>1999Q4</th>
<th>2000Q1</th>
<th>2000Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999Q2</td>
<td>$x_{99Q,2,99Q}$</td>
<td>$x_{99Q,2,99Q}$</td>
<td>$x_{99Q,2,99Q}$</td>
<td>$x_{99Q,2,00Q}$</td>
<td>$x_{99Q,2,00Q}$</td>
</tr>
<tr>
<td>1999Q3</td>
<td>$\hat{x}_{99Q,3,99Q}$</td>
<td>$x_{99Q,3,99Q}$</td>
<td>$x_{99Q,3,99Q}$</td>
<td>$x_{99Q,3,00Q}$</td>
<td>$x_{99Q,3,00Q}$</td>
</tr>
<tr>
<td>1999Q4</td>
<td>$\hat{x}_{99Q,4,99Q}$</td>
<td>$\hat{x}_{99Q,4,99Q}$</td>
<td>$x_{99Q,4,99Q}$</td>
<td>$x_{99Q,4,00Q}$</td>
<td>$x_{99Q,4,00Q}$</td>
</tr>
<tr>
<td>2000Q1</td>
<td>$\hat{x}_{00Q,1,99Q}$</td>
<td>$\hat{x}_{00Q,1,99Q}$</td>
<td>$\hat{x}_{00Q,1,99Q}$</td>
<td>$x_{00Q,1,00Q}$</td>
<td>$x_{00Q,1,00Q}$</td>
</tr>
<tr>
<td>2000Q2</td>
<td>-</td>
<td>$\hat{x}_{00Q,2,99Q}$</td>
<td>$\hat{x}_{00Q,2,99Q}$</td>
<td>$\hat{x}_{00Q,2,00Q}$</td>
<td>$\hat{x}_{00Q,2,00Q}$</td>
</tr>
<tr>
<td>2000Q3</td>
<td>-</td>
<td>-</td>
<td>$\hat{x}_{00Q,3,99Q}$</td>
<td>$\hat{x}_{00Q,3,00Q}$</td>
<td>$\hat{x}_{00Q,3,00Q}$</td>
</tr>
<tr>
<td>2000Q4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>$\hat{x}_{00Q,4,00Q}$</td>
<td>$\hat{x}_{00Q,4,00Q}$</td>
</tr>
</tbody>
</table>
A real-time data set can be cut up in different ways with the goal to extracting some important information on the revision process.

The two most interesting frameworks are what we have called the vertical approach and the diagonal approach.

These simple names have been chosen to avoid the plethora of different confusing terminologies (for me) found in the literature.

In this paper concern (i) the quarterly real gross domestic product for the EU 12 area, (ii) the monthly industrial production index for Belgium and for (iii) the EU12.

Data used throughout this paper are from the Euro Area Business Cycle Network (EABCN). Eurostat data are under investigation (under manipulation say).
The Euro area Real Time database (project co-ordinated by the EABCN), constructed jointly by the DG Research and DG Statistics of the ECB, consists of vintages, or snapshots, of time series of several macroeconomic variables, based on series reported in the ECB’s Monthly Bulletins.

The database now comprises more than 200 variables and all of them are downloadable by EABCN fellows.

There are 38 variables for the EU area.

The country real time database, assembled by the National Central Banks (NCB), comprises more than 200 variables (approximately 25-30 variables for each country) and all of them are downloadable by EABCN fellows. Data are currently available for The Netherlands, Ireland, Luxembourg, Spain, Belgium, Italy and UK.
Time series univariate features

- Whatever the framework used (i.e. the vertical or the diagonal approach) and the variable analyzed (IIP, GDP), the main univariate statistical features of the data are:
- The presence of a unit root in the log-levels of the series, namely that the series are non stationary.
- The presence of serial correlation in the first differences.
- Normality, ARCH, linearity are OK
Our proposed time series criteria for selecting vintages (min. requirement):

1. The vertical and/or diagonal vintages being time series of the same underlying process, long-run and short-run co-movements must be observed.

2. For both types of relationships we expect a unique common trend and a unique common cycle generating $n$ series, i.e. $r = n - 1$ cointegrating vectors and $s = n - 1$ common cyclical feature vectors.

3. Each relationship cannot be so different from the (1 -1) relationships.
BUT:

- 2. and 3. are not possible in a multivariate framework with both cointegration and serial correlation common features (Engle-Kozicki)
- Multivariate analysis can be confronted to the curse of dimensionality.

Example: Belgian monthly IPI, verticals.
Criteria 1: Looking/Testing for co-movements in Belgian IIP: The verticals

Figure: Log-levels of the Belgian IPI - 9 verticals
Figure: Growth rates of vertical vintages (Belgium)
Commonalities in the trend and the cycle components are obvious on figures but for instance the Johansen’s multivariate test would not formally find $r = 8$ cointegrating vectors for the nine vintages.

The graph of the cointegrating vectors I don’t see how we would favour 6 and not 8 for instance.

**Figure:** Eight cointegrating vectors
## Table: Unrestricted Cointegration Rank Test (Johansen’s trace test)

<table>
<thead>
<tr>
<th>Nb. of coint.</th>
<th>Eigenvalue</th>
<th>Statistics</th>
<th>Critical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>* 0.9785</td>
<td>545.245</td>
<td>197.370</td>
</tr>
<tr>
<td>At most 1</td>
<td>* 0.9394</td>
<td>368.573</td>
<td>59.529</td>
</tr>
<tr>
<td>At most 2</td>
<td>* 0.8634</td>
<td>239.602</td>
<td>125.61</td>
</tr>
<tr>
<td>At most 3</td>
<td>* 0.6654</td>
<td>148.029</td>
<td>95.753</td>
</tr>
<tr>
<td>At most 4</td>
<td>* 0.6291</td>
<td>97.657</td>
<td>69.818</td>
</tr>
<tr>
<td>At most 5</td>
<td>* 0.4209</td>
<td>52.025</td>
<td>47.856</td>
</tr>
<tr>
<td>At most 6</td>
<td>0.2794</td>
<td>26.895</td>
<td>29.797</td>
</tr>
<tr>
<td>At most 7</td>
<td>0.2213</td>
<td>11.822</td>
<td>15.494</td>
</tr>
<tr>
<td>At most 8</td>
<td>0.0067</td>
<td>0.313</td>
<td>3.841</td>
</tr>
</tbody>
</table>

*Note: * denotes rejection of the hypothesis at the 0.05 level.*
The second type of relationships is the existence of common cyclical features, namely the study of short-run co-movements.

- For the VECM representation of a CI(1,1)-VAR($p$)

\[ \Delta Z_t = \alpha \beta' Z_{t-1} + \Phi_1 \Delta Z_{t-1} + \ldots + \Phi_p \Delta Z_{t-p+1} + \varepsilon_t \]

SCCF for synchronous cycles (Vahid and Engle, 1993):

\[ \delta' \Delta Z_t = \delta' \varepsilon_t \]

- In the multivariate setting the number of serial correlation common feature relationships (the usual approach) is bounded by $n - r$, with $n$ and $r$ respectively the number of variables and cointegrating vectors.

- This explains why we observe only a small number of short-run co-movements in empirical investigations (Patterson for instance).
Using SCCF, not only we have $s \leq n - r$ but SCCF approach is not able to obtain a unit elasticity both for the long-run and the short-run relationships (independence of the cointegrating and the common feature space).

For investigating the presence of the short-run co-movements, we prefer to use the weak form common cyclical feature model (Hecq, Palm and Urbain (2000, 2006))

$$
\delta' \Delta Z_t = \delta' \alpha \beta' Z_{t-1} + \delta' \varepsilon_t
$$
Table: Summary of common cyclical feature test statistics for 9 vintages

<table>
<thead>
<tr>
<th>Nb. of CF</th>
<th>LR</th>
<th>LR$^{cor}$</th>
<th>AIC</th>
<th>HQ</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCCF</td>
<td>$s = 0$</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>WF</td>
<td>$s = 1$</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Note: For LR tests, results when a p-value higher than 5%. We take $r=8$ cointegrating vectors and 2 lags on the first differences.

Result: although the number of short-run co-movements is higher using the WF than the SCCF, we are quite far from $n - 1$. 
Are we lost?

- **Our option:** Let’s choose and use the series such that the number of co-movements is maximized.

- **How:**
  1. Bivariate analysis before (sometimes instead) multivariate models (technical reasons such as dimensionality, WF vs. SCCF).
  2. Univariate models prior to a multivariate framework (FE, Cubadda, Hecq and Palm 2008).
Using a bivariate cointegration analysis, out of the 9 verticals, we keep 5 series (not necessarily the last ones, the first ones...).

**Table: Cointegration test statistics for 5 vintages**

<table>
<thead>
<tr>
<th>Nb. of coint.</th>
<th>Eigenvalue</th>
<th>Statistics</th>
<th>Critical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>* 0.807</td>
<td>163.27</td>
<td>69.81</td>
</tr>
<tr>
<td>At most 1</td>
<td>* 0.593</td>
<td>80.91</td>
<td>47.85</td>
</tr>
<tr>
<td>At most 2</td>
<td>* 0.326</td>
<td>35.90</td>
<td>29.79</td>
</tr>
<tr>
<td>At most 3</td>
<td>* 0.240</td>
<td>16.11</td>
<td>15.49</td>
</tr>
<tr>
<td>At most 4</td>
<td>0.046</td>
<td>2.36</td>
<td>3.84</td>
</tr>
</tbody>
</table>

* denotes rejection of the hypothesis at the 0.05 level.
Table: Summary of common cyclical feature test statistics for 5 vintages

<table>
<thead>
<tr>
<th>Nb. of CF</th>
<th>LR</th>
<th>LR$^{ss}$</th>
<th>AIC</th>
<th>HQ</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCCF</td>
<td>$s = 1$ (0.53)</td>
<td>1 (0.64)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>WF</td>
<td>$s = 4$ (0.11)</td>
<td>1 (0.18)</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Note: 5 vintages For LR tests, number for a p-value higher than 5%. We take $r=4$ cointegrating vectors and 1 lags on the first differences.

- Result: $n = 5$, $r = 4$ and $s = 4$ (WF)
- Note: only one SCCF (by definition)
To work with diagonals, a rebasing is a question of the utmost importance.
Figure: Cointegrating relationships on raw diagonals
On rebased diagonals:

- \( n = 8 \)
- \( r = 7 \)
- \( s = 7(WF) \)
Partial conclusion on co-movements:

- It is wise to first look at bivariate relationships before multivariate analyses.
- Indeed, a bivariate investigation of all possible vintages and diagonals helps to determine a set of variables we could next use in a multivariate setting.
- Work with the WF for short-run co-movements.
- We cannot do it with all possible vintages (I’ve got 78 verticals for instance) => Final equations: univariate models compatible with multivariate frameworks.
The Final equations (FE) of the VAR Presentation

- Let us consider the $n$ dimensional VAR($p$)

$$Z_t = \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \ldots + \Phi_p Z_{t-p} + \varepsilon_t,$$

$$\Phi(L)Z_t = \varepsilon_t$$

- Premultiplying both sides by $\Phi(L)^{adj}$, the adjoint of $\Phi(L)$ such that $\Phi(L)^{adj} = \text{det}[\Phi(L)]\Phi(L)^{-1}$ one obtains the FE

$$\text{det}[\Phi(L)]Z_t = \Phi(L)^{adj}\varepsilon_t,$$

In which:
The Final equations (FE) of the VAR

Implications

1. Every FE is an univariate ARMA\((p^*, q^*)\) although it is derived from the VAR\((p)\).

2. They have identical autoregressive parameters because \(\text{det}[\Phi(L)]\) is the same of the \(n\) equations.

The paradox

1. **Theoretical** ARMA\((np, (n - 1)p)\) which could be large.

2. **Empirical** parsimonious ARIMA.
VAR of order 1 for \( n = 3 \) variables

\[
\begin{bmatrix}
Z_{1t} \\
Z_{2t} \\
Z_{3t}
\end{bmatrix}
= \begin{bmatrix}
0.5 & -0.5 & 0.5 \\
0.25 & -0.25 + \omega & 0.25 \\
0.5 & -0.5 & 0.5 + \omega
\end{bmatrix}
\begin{bmatrix}
Z_{1t-1} \\
Z_{2t-1} \\
Z_{3t-1}
\end{bmatrix}
+ \varepsilon_t,
\]

where \( Z_t = (Z_{1t}, Z_{2t}, Z_{3t})' \). For that VAR

\[
\det[I - \Phi_1 L] = -0.5 L^3 \omega^2 + (1.25 \omega + \omega^2) L^2 - (0.75 + 2 \omega) L + 1
\]

such that we have three ARMA(3,2).
But if $\omega = 0$.

The VAR can be written in its reduced rank form

$$Z_t = \begin{bmatrix} 0.5 \\ 0.25 \\ 0.5 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} Z_{t-1} + \varepsilon_t$$

$$Z_t = \begin{bmatrix} 0.5 \\ 0.25 \\ 0.5 \end{bmatrix} Common\_Cycle + \varepsilon_t$$
General results

**Table:** Maximum ARMA orders of univariate series generated by a stationary VAR(p) and CI(1,1) VAR(p) with cofeature restrictions

<table>
<thead>
<tr>
<th>Models</th>
<th>AR order</th>
<th>I(d)</th>
<th>MA order</th>
</tr>
</thead>
<tbody>
<tr>
<td>I(0)-VAR(p)</td>
<td>$np$</td>
<td>0</td>
<td>$(n - 1)p$</td>
</tr>
<tr>
<td>SCCF</td>
<td>$(n - s)p$</td>
<td>0</td>
<td>$(n - s)p$</td>
</tr>
<tr>
<td>PSCCF(1)</td>
<td>$(n - s)p + s$</td>
<td>0</td>
<td>$(n - s)p + (s - 1)$</td>
</tr>
<tr>
<td>C(1,1)-VAR(p)</td>
<td>$n(p - 1) + r$</td>
<td>1</td>
<td>$(n - 1)(p - 1) + r$</td>
</tr>
<tr>
<td>SCCF</td>
<td>$(n - s)(p - 1) + r$</td>
<td>1</td>
<td>$(n - s)(p - 1) + r$</td>
</tr>
<tr>
<td>PSCCF(1)</td>
<td>$(n - s)(p - 1) + r + s$</td>
<td>1</td>
<td>$(n - s)(p - 1) + r + s - 1$</td>
</tr>
<tr>
<td>WF</td>
<td>$(n - s)(p - 1) + r$</td>
<td>1</td>
<td>$(n - s)(p - 1) + r$</td>
</tr>
</tbody>
</table>
Empirical example
Growth rates of GDP in Canada and the US (no coint in the levels)

Figure: Quarterly growth rates of industrial production indexes (industrial sector)
Empirical example

Growth rates of GDP in Canada and the US

- The estimation by OLS of the VAR(1) in first differences delivers (standard errors in brackets)

\[
\begin{bmatrix}
\Delta \ln USA_t \\
\Delta \ln CA_t
\end{bmatrix} = \begin{bmatrix}
0.003 \\
0.003
\end{bmatrix} + \begin{bmatrix}
0.333 & 0.273 \\
0.265 & 0.360
\end{bmatrix} \begin{bmatrix}
\Delta \ln USA_{t-1} \\
\Delta \ln CA_{t-1}
\end{bmatrix}.
\]

- Theoretical FE orders: ARMA(2,1) processes?
- SCCF test statistics: we cannot reject the null that there exists a single SCCF vector (p-value = 0.31)
It emerges that this gives us a theoretical framework for aggregating vintages (diagonals or verticals)

Doing so we only work with individual series, look if they are alike and average them.

We do not lose long and short-run co-movements, they are already included in the individual ARMA (they imply their orders actually)

Using these three elements (i.e. cointegration, common cycles and the final equations) we have determined that the vertical and the diagonal approaches should be used for different purposes.
The vertical approach is best for determining the vintage series that must be finally taken for an historical analysis of the economy (or for an econometric model).

We have proposed a method to choose in an optimal manner the vintage or the combination of different vintages (sub-period that increases).

- Preselect with FE (stability of ARMA orders).
- Cointegration.
- Common cycles.

It is very important to notice that the conclusion will depend on whether long-run properties or short-run business cycle properties are considered. For instance to analyze the business cycle properties of some series, the vintage published in \( t + 11 \) months is the best vintage for the historical series while it is the vintage \( t + 23 \) months for long-run properties.
- We think however that it could be time consuming to systematically apply this approach to every series.
- We would recommend to redo this analysis for a set of real variables, financial series, monetary variables, sectorial data, prices, etc and to evaluate the accuracy of the historical series for each of these categories.
- Then researches will have some indication about the series to work with. For instance if someone work with prices we could recommend to skip the last three months in the last vintage.
• The diagonal approach gives an alternative interesting viewpoint.
• With the diagonals we obtain an indication about the moment at which we can trust the data in the revision process.
• This is very important when working with short-run indicators, for early warning systems and for forecasting.
• This is the information that will be used by policy makers and economic agents in general.
• For our series it was not possible to prefer a particular diagonal from a cointegration analysis because such relationships were present for any combination.
• However a weak form common feature analysis and the identification of individual ARMA models reveal that releases in $t + 6$ or $t + 7$ months are trustable revisions for the number at time $t$.
Using such an analysis on diagonals we can obtain an indicator of the accuracy of every series.

This would tell us the data that is worth improving with priority for the trade off accuracy/timeliness and would give us an index of the accuracy of different variables.
Example: EU IIP diagonals

**Table: 2001 vintages IIP EU12l selection**

<table>
<thead>
<tr>
<th>Diags</th>
<th>SARMA</th>
<th>dummies</th>
<th>SARMA</th>
<th>dummies</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t, t + 3$</td>
<td>(0,0,1)(0,0,0)</td>
<td>-</td>
<td>$t, t + 11$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
<tr>
<td>$t, t + 4$</td>
<td>(0,0,1)(0,0,0)</td>
<td>2001:10</td>
<td>$t, t + 12$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
<tr>
<td>$t, t + 5$</td>
<td>(1,0,0)(0,0,0)</td>
<td>2001:10</td>
<td>$t, t + 13$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
<tr>
<td>$t, t + 6$</td>
<td>(3,0,0)(0,0,1)</td>
<td>-</td>
<td>$t, t + 14$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
<tr>
<td>$t, t + 7$</td>
<td>(3,0,0)(0,0,1)</td>
<td>-</td>
<td>$t, t + 15$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
<tr>
<td>$t, t + 8$</td>
<td>(0,0,1)(0,0,1)</td>
<td>2001:10</td>
<td>$t, t + 16$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
<tr>
<td>$t, t + 9$</td>
<td>(0,0,1)(0,0,1)</td>
<td>2001:10</td>
<td>$t, t + 17$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
<tr>
<td>$t, t + 10$</td>
<td>(0,0,1)(0,0,1)</td>
<td>2001:10</td>
<td>$t, t + 18$</td>
<td>(0,0,1)(0,0,1)</td>
</tr>
</tbody>
</table>