Factor nowcasting of German GDP with ragged-edge data: A model comparison using MIDAS projections

Massimiliano Marcellino
Università Bocconi, IGIER and CEPR

Christian Schumacher
Deutsche Bundesbank

25 October 2007

1 Introduction

Bundesbank involved in disaggregated Euro area projection exercises (BMPE) together with European Central Bank (ECB) and other National Central Banks of the Eurosystem

In BMPE, Bundesbank has to provide forecasts for German GDP and inflation

Forecasts are monitored by other Eurosystem central banks and the public

$\longrightarrow$ There is always the need to be up-to-date with forecasting techniques

Presentation today part of bigger Bundesbank project on development of a range of nowcast and forecast models

Techniques are still under evaluation, not implementend yet in regular forecast exercises

Outline of presentation:

1. What is nowcasting?
2. Difficulties of now- and forecasting GDP
3. Discussion of three factor models that can tackle unbalanced data
4. MIDAS projection with factors as regressors: FACTOR-MIDAS
5. Empirical now- and forecast comparison
6. Conclusions

2 What is nowcasting?

Decision makers regularly request information on the current state of the economy

GDP is an important business cycle indicator, but sampled at quarterly frequency only and published with considerable delay

Economist's task: **Estimate current quarter GDP** using all information which is currently available

Example: in April, German GDP is available only for the fourth quarter of the previous year. To obtain a 2^{nd} quarter GDP nowcast, we have to make a projection with forecast horizon of two quarters from the end of the GDP sample.

3 Difficulties of now- and forecasting GDP

There are many difficulties in real-world applications, we discuss two:

1. GDP is quarterly data, many important indicators are sampled at monthly or higher frequency - **mixed-frequency problem**

2. Indicators for nowcasting are available with different publication lags $\longrightarrow$ leads to the so-called **ragged edge** of multivariate datasets, Wallis (1986)

Question here: How to nowcast quarterly GDP with factors estimated from a large set of monthly indicators?

4 Factor now- and forecasting: Two-step procedure

Boivin and Ng (2005) IJCB: In the single-frequency case, factor forecasting is often a two-step procedure

1. Estimation Step: Estimate factors

2. Forecast step: Factor-augmented (V)AR model

We follow the same strategy in the mixed-frequency case with ragged-edge data

5 Factor estimation with ragged-edge data

Monthly observations have a factor structure

$$\mathbf{X}_{t_m} = \mathbf{\Lambda}\mathbf{F}_{t_m} + \boldsymbol{\xi}_{t_m} \tag{1}$$

with factors $\mathbf{F}_{t_m} = (f'_{1,t_m}, \ldots, f'_{r,t_m})'$, loadings $\mathbf{\Lambda}$, idiosyncratic components $\boldsymbol{\xi}_{t_m}$

If data $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{T_m})'$ is balanced, there are different ways to estimate $\mathbf{F}$:

1. PCA as in Stock and Watson (2002) JBES

2. Dynamic PCA according to Forni et al. (2005) JASA

3. Subspace algorithms, Marcellino and Kapetanios (2006) CEPR WP

We consider three factors models for ragged-edge data:

1. Vertical realignment of the data and dynamic PCA, Altissimo et al. (2006) CEPR WP (New Eurocoin)

2. EM algorithm plus PCA, Stock and Watson (2002) JBES appendix, Bernanke and Boivin (2003) JME, Schumacher and Breitung (2006) BBK DP

3. Kalman smoother estimation using a large state-space factor model with PCA providing initial values, see Doz, Giannone and Reichlin (2006) ECB WP

5.1 Vertical realignment of the data and dynamic PCA

Altissimo et al. (2006) CEPR WP

Variable i is released with k_i months of publication lag $\rightarrow$ in period T_m, the final observation available is in period $T_m - k_i$

Balancing by 'vertical' realignment

$$\widetilde{x}_{i,T_m} = x_{i,T_m-k_i} \tag{2}$$

Applying this procedure for each series and harmonising at the beginning of the sample, yields a balanced data set $\widetilde{\mathbf{X}}_{t_m}$

Factor estimation from $\widetilde{\mathbf{X}}_{t_m}$ by Dynamic PCA, Forni et al. (2005) JASA

5.2 EM algorithm and static PCA

Stock and Watson (2002) JBES appendix, Schumacher and Breitung (2006) BBK DP

We want a full data column vector $\mathbf{X}_i = (x_{i,1}, \ldots, x_{i,T_m})'$, due to the ragged-edge problem we observe only $\mathbf{X}_i^{\text{obs}}$

relationship between observed and unobserved

$$\mathbf{X}_i^{\text{obs}} = \mathbf{A}_i \mathbf{X}_i \tag{3}$$

where $\mathbf{A}_i$ is a matrix that can tackle missing values or mixed frequencies

EM algorithm:

1. Initial (naive) guess of observations $\widehat{\mathbf{X}}_i^{(0)} \forall i$ yields a balanced dataset $\widehat{\mathbf{X}}^{(0)}$, static PCA provides initial monthly factors $\widehat{\mathbf{F}}^{(0)}$ and loadings $\widehat{\mathbf{\Lambda}}^{(0)}$

2. **E-step:** Expectation of $\mathbf{X}_i$ conditional on observations $\mathbf{X}_i^{\text{obs}}$, factors $\widehat{\mathbf{F}}^{(j-1)}$ and loadings $\widehat{\mathbf{\Lambda}}_i^{(j-1)}$ from the previous iteration

$$\widehat{\mathbf{X}}_i^{(j)} = \widehat{\mathbf{F}}^{(j-1)} \widehat{\mathbf{\Lambda}}_i^{(j-1)} + \mathbf{A}_i' (\mathbf{A}_i' \mathbf{A}_i)^{-1} \left(\mathbf{X}_i^{\text{obs}} - \mathbf{A}_i \widehat{\mathbf{F}}^{(j-1)} \widehat{\mathbf{\Lambda}}_i^{(j-1)} \right) \tag{4}$$

3. **M-step:** Repeat the E-step for all i providing again a balanced dataset. Reestimate $\widehat{\mathbf{F}}^{(j)}$ and $\widehat{\mathbf{\Lambda}}^{(j)}$ by PCA, and go to step 2 until convergence

5.3 Kalman smoother estimates in a large state-space model

Doz et al. (2006) ECB WP

Model:

$$\mathbf{X}_{t_m} = \mathbf{\Lambda}\mathbf{F}_{t_m} + \boldsymbol{\xi}_{t_m} \quad (5)$$
$$\mathbf{\Psi}(L_m)\mathbf{F}_{t_m} = \mathbf{B}\boldsymbol{\eta}_{t_m} \quad (6)$$

VAR for factors with $\mathbf{\Psi}(L_m) = \sum_{i=1}^{p} \mathbf{\Psi}_i L_m^i$ and $L_m x_{t_m} = x_{t_m - 1}$, q-dimensional vector $\boldsymbol{\eta}_{t_m}$ contains dynamic shocks that drive factors, identification matrix $\mathbf{B}$ is $(r \times q)$-dimensional

Model has state-space representation with factors as states

Estimation 'trick': Coefficients estimated outside state-space model, no iterative ML

QML to estimate the factors:

1. Estimate $\widehat{\mathbf{F}}_{t_m}$ using PCA as an initial estimate

2. Estimate $\widehat{\mathbf{\Lambda}}$ by regressing $\mathbf{X}_{t_m}$ on the estimated factors $\widehat{\mathbf{F}}_{t_m}$. The covariance of the idiosyncratic components $\widehat{\boldsymbol{\xi}}_{t_m} = \mathbf{X}_{t_m} - \widehat{\mathbf{\Lambda}}\widehat{\mathbf{F}}_{t_m}$, denoted as $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\xi}}$, is also estimated

3. Estimate factor VAR(p) on the factors $\widehat{\mathbf{F}}_{t_m}$ yielding $\widehat{\mathbf{\Psi}}(L_m)$ and the residual covariance of $\widehat{\boldsymbol{\varsigma}}_{t_m} = \widehat{\mathbf{\Psi}}(L_m)\widehat{\mathbf{F}}_{t_m}$, denoted as $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varsigma}}$

4. To obtain an estimate for $\mathbf{B}$, apply an eigenvalue decomposition of $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\varsigma}}$

5. State-space is fully specified numerically $\rightarrow$ The Kalman smoother provides estimates of factors

5.4 Discussion

VA-DPCA: easy to use, but

- availability of data determines dynamic cross-correlations between variables!
- statistical release dates for data are not the same over time, for example due to major revisions $\longrightarrow$ dynamic correlations within the data change and factors can change over time

EM-PCA

- interpolation of missing values in line with factor model

- no dynamics

- sometimes convergence problems (‘partial factors’, large r)

KFS-PCA

- optimality properties of Kalman filter/smoother

- explicit dynamics of the factors

- assumptions on idiosyncratic components often not fulfilled

- more (dynamic) structure, more auxiliary parameters to fix and perhaps subject to misspecification, see Boivin and Ng (2005) IJCB

6 Factor forecasting: MIDAS-basic

Ghysels, Sinko, Valkanov (2007) EctrRev

Our contribution: Clements and Galvão (2007) WP use single macro variables, we use factors $\rightarrow$ FACTOR-MIDAS

Three ways:

1. MIDAS-basic

2. MIDAS-smooth

3. Unrestricted MIDAS

6.1 Factor forecasting: MIDAS-basic

Ghysels, Sinko, Valkanov (2007) EctrRev, Clements and Galvão (2007) WP

MIDAS Model with one factor $\widehat{f}_{t_m}$ for forecast horizon h_q quarters ($h_q = h_m/3$):

$$y_{t_q+h_q} = \beta_0 + \beta_1 b(L_m, \boldsymbol{\theta})\widehat{f}^{(3)}_{t_m} + \varepsilon_{t_m+h_m} \tag{7}$$

$$b(L, \theta) = \sum_{k=0}^{K} c(k, \boldsymbol{\theta}) L_m^k \qquad c(k, \boldsymbol{\theta}) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{k=0}^{K} \exp(\theta_1 k + \theta_2 k^2)} \tag{8}$$

Quarterly GDP $y_{t_q+h_q}$ is directly related to the monthly factor $\widehat{f}^{(3)}_{t_m}$ and lags, where $\widehat{f}^{(3)}_{t_m}$ is a skip-sampled version of $\widehat{f}_{t_m}$

Estimation by nonlinear least squares (NLS)

Device for **direct** forecasting depending on h_q, see Marcellino, Stock, Watson (2006)

$$y_{T_q+h_q|T_q} = \widehat{\beta}_0 + \widehat{\beta}_1 b(L_m, \widehat{\boldsymbol{\theta}}) \widehat{f}_{T_m} \tag{9}$$

6.2 The projection in New Eurocoin: MIDAS-smooth

Altissimo et al. (2006) CEPR WP

Projection:

$$y_{T_q+h_q|T_q} = \widehat{\mu} + \left[\widetilde{\mathbf{\Sigma}}_{y\mathbf{F}}(h_m) \times \widehat{\mathbf{\Sigma}}_{\mathbf{F}}^{-1} \right] \times \widehat{\mathbf{F}}_{T_m} \qquad (10)$$

$\widehat{\mu}$ is GDP sample mean, $\widehat{\mathbf{\Sigma}}_{\mathbf{F}}$ is sample covariance of factors

$\widetilde{\mathbf{\Sigma}}_{y\mathbf{F}}(k)$ is cross-covariance between **smoothed** GDP and factors with k lags

$$\widetilde{\mathbf{\Sigma}}_{y\mathbf{F}}(k) = \frac{1}{2H+1} \sum_{j=-H}^{H} \alpha(\omega_j) \widehat{\mathbf{S}}_{y\mathbf{F}}(\omega_j) e^{i\omega_j k} \qquad (11)$$

with $\alpha(\omega_j) = 1 \, \forall \left| \omega_j \right| < \pi/6$ and zero otherwise, and cross-spectral matrix $\widehat{\mathbf{S}}_{y\mathbf{F}}(\omega_j)$

If we disregard smoothing and use

$$\widehat{\mathbf{\Sigma}}_{y\mathbf{F}}(k) = \frac{1}{T^* - 1} \sum_{t_m=M+1}^{T_m} y_{t_m} \widehat{\mathbf{F}}^{(3)\prime}_{t_m-k} \tag{12}$$

where $T^* = \text{floor}[(T_m - (M+1))/3]$ in

$$y_{T_q+h_q|T_q} = \widehat{\mu} + \left[\widehat{\mathbf{\Sigma}}_{y\mathbf{F}}(h_m) \times \widehat{\mathbf{\Sigma}}_{\mathbf{F}}^{-1}\right] \times \widehat{\mathbf{F}}_{T_m} \tag{13}$$

$\longrightarrow$ Both in basic MIDAS and New Eurocoin: $\widehat{\mathbf{\Sigma}}_{y\mathbf{F}}(k)$ can be estimated consistently although y_{t_q} and $\widehat{\mathbf{F}}_{t_m}$ have different sampling frequencies

$\longrightarrow$ Projection in New Eurocoin and basic MIDAS projection follow the same idea!

6.3 Unrestricted MIDAS

Unrestricted lag order model

$$y_{t_q+h_q} = \beta_0 + \mathbf{D}(L_m)\widehat{\mathbf{F}}_{t_m}^{(3)} + \varepsilon_{t_m+h_m}, \tag{14}$$

where $\mathbf{D}(L_m) = \sum_{k=0}^{K} \mathbf{D}_k L_m^k$ is an unrestricted lag polynomial of order K.

Estimation of $\mathbf{D}(L_m)$ and β_0 by OLS

We consider fixed lag orders with $k = 0$ and $k = 1$

Note that for $k = 0$, we consider only t_m-dated factors for forecasting $\longrightarrow$ case $k = 0$ is like MIDAS-smooth without smoothing

7 Empirical now- and forecast comparison

7.1 Data

German quarterly GDP from $1992Q1$ until $2006Q3$

111 monthly indicators

Data: No real-time, we generate ragged-edge vintages from final data

Solution: Missing values at the end of the full, final sample are used to identify missing values in pseudo real-time subsamples, see Banbura and Rünstler (2007) ECB WP

7.2 Forecast design

Recursive design with increasing sample size

Evaluation sample from $1998Q4$ until $2006Q3$

Each month, we compute new now- and forecasts with monthly horizon $h_m = 1, 2, \ldots, 9$

Models are estimated using fixed specification in terms of numbers of factors r and q, and using information criteria, see Bai and Ng (2002, 2007)

7.3 Empirical results

Results based on fixed specifications for $r = 1, 2$

relative MSE to variance of GDP, number of factors $r = 1$

		nowcast current quarter			forecast 1 quarter			forecast 2 quarters		
	horizon h_m	1	2	3	4	5	6	7	8	9
1.a. MIDAS-U0	VA-DPCA	0.71	0.86	0.89	0.90	1.05	0.98	1.05	1.09	1.12
	EM-PCA	0.58	0.65	0.72	0.92	0.93	0.79	1.10	1.10	1.05
	KFS-PCA	0.68	0.85	0.80	0.95	1.01	0.93	1.08	1.09	1.06
1.b. Ranking	VA-DPCA	3	3	3	1	3	3	1	2	3
	EM-PCA	1	1	1	2	1	1	3	3	1
	KFS-PCA	2	2	2	3	2	2	2	1	2
2. Benchmarks	AR	1.02	1.17	1.17	1.17	1.08	1.08	1.08	1.08	1.08
	in-sample mean	1.03	1.04	1.04	1.04	1.05	1.05	1.05	1.06	1.06

relative MSE to variance of GDP, number of factors $r = 2$, $q = 1$

		nowcast current quarter			forecast 1 quarter			forecast 2 quarters		
	horizon h_m	1	2	3	4	5	6	7	8	9
4.a. MIDAS-U0	VA-DPCA	0.75	0.82	0.87	0.78	1.01	0.94	1.28	1.13	1.15
	EM-PCA	0.66	1.07	0.85	0.98	0.96	0.73	1.26	1.00	2.30
	KFS-PCA	0.71	1.06	0.87	0.94	0.96	0.69	1.17	1.11	1.52
4.b. Ranking	VA-DPCA	3	1	2	1	3	3	3	3	1
	EM-PCA	1	3	1	3	1	2	2	1	3
	KFS-PCA	2	2	3	2	2	1	1	2	2
6. Benchmarks	AR	1.02	1.17	1.17	1.17	1.08	1.08	1.08	1.08	1.08
	in-sample mean	1.03	1.04	1.04	1.04	1.05	1.05	1.05	1.06	1.06

relative MSE to variance of GDP, number of factors $r = 1$

		nowcast current quarter			forecast 1 quarter			forecast 2 quarters		
	horizon h_m	1	2	3	4	5	6	7	8	9
1.a. VA-DPCA	MIDAS-basic	0.71	1.01	1.06	0.94	1.18	1.05	1.16	1.24	1.30
	MIDAS-smooth	0.69	0.92	0.87	0.95	1.10	1.20	1.18	1.12	1.19
	MIDAS-U0	0.71	0.86	0.89	0.90	1.05	0.98	1.05	1.09	1.12
1.b. Ranking	MIDAS-basic	3	3	3	2	3	2	2	3	3
	MIDAS-smooth	1	2	1	3	2	3	3	2	2
	MIDAS-U0	2	1	2	1	1	1	1	1	1
2.a. EM-PCA	MIDAS-basic	0.62	0.69	0.78	1.07	0.99	1.01	1.30	1.09	1.05
	MIDAS-smooth	0.70	0.73	0.84	0.94	0.95	1.00	1.05	1.09	1.13
	MIDAS-U0	0.58	0.65	0.72	0.92	0.93	0.79	1.10	1.10	1.05
2.b. Ranking	MIDAS-basic	2	2	2	3	3	3	3	1	1
	MIDAS-smooth	3	3	3	2	2	2	1	2	3
	MIDAS-U0	1	1	1	1	1	1	2	3	2

7.4 More results

Nowcasts based on information-criteria model selection for r and q have no information content

We also checked PCA vs DPCA together with vertical realignment of the data $\longrightarrow$ no big differences

We checked the information content of ragged-edge data vs. balanced data $\longrightarrow$ ragged-edge contains in general useful information for nowcasting

We compared an intgrated state-space model which also interpolates GDP from Banbura and Rünstler (2007) with the two-step factor nowcast here $\longrightarrow$ no big differences

8 Conclusions

Factor models considered here can address nowcasting questions with ragged-edge and mixed-frequency data

Models with only one or two factors ($r = 1, 2$) perform best

Differences between factor estimation methods (vertical realignment, EM, state space) are minor

Simplest MIDAS projections with few lags do better than exponential lag versions

Projections are informative for the nowcast and forecast one quarter ahead $\rightarrow$ factor models can be regarded as **short-term forecast models** only