

GDP nowcasting with ragged-edge data: A Semi-Parametric Modelling

GUEGAN Dominique

`dguegan@univ-paris1.fr`

PSE - Centre Economique de la Sorbonne - University Paris 1 - Panthéon -
Sorbonne

*First Macro-Economic Forecasting Conference - 27th March 2009 - Roma
Joint Work with Laurent Ferrara and Patrick Rakotomaroahy*

ces.univ-paris1.fr

Plan

Introduction

Semi-Parametric Modelling (SPM) for GDP

Non-Parametric Modelling (NPM) for Monthly indicators

Carrying out of Non Parametric estimation

Semi-parametric modelling for GDP growth

Conclusion

Introduction: Monetary Policy Decisions in Real Time

- ▶ In the Euro-area zone, the first GDP estimation referred to as "flash estimate" is released by Eurostat around 43 days after the end of the reference quarter.
- ▶ The objectif: to provide an estimate of the current and future quarter GDP "flash estimate" release based on:
- ▶ Statistical models which deal with mixed frequency (using monthly data to nowcast quartely GDP).
 1. We use nowcast and forecast monthly indicators based on k -nearest neighbors (k -NN) method and radial basic functions (RBF) method.
 2. We plug those estimates, inside the bridge equations, permitting to get a semi-parametric forecast for the GDP growth (SPM).
 3. Indeed, we use predictive equations that bridge monthly information with quartely ones
- ▶ Validation of the SPM by carrying out a true real-time experience on the euro-area that uses vintage data.

Introduction: A quick review of the models

- ▶ Mixed Data sampling (MIDAS) regressions: Econometric approach. Regressions that directly accommodate variables sampled at different frequencies: Santa-Clara and Valkanov (2005).
- ▶ Amount of information necessary to obtain a robust estimate of the current GDP: no definitive answer. Marcellino, Stock and Watson (2001), Bernanke and Boivin (2003), Giannone, Lippi and Reichlin (2005), Boivin and Ng (2006), Giannone and Surico (2006).
- ▶ Modellings using large panels of time series. The use of Factor models for US GDP. Forni, Hallin, Lippi and Reichlin (2000, 2003), Bernanke and Boivin (2002), Stock and Watson (2002), Doz, Giannone and Reichlin (2006), Kapetanios and Marcellino (2006) and Doz, Giannone and Reichlin (2007).

Introduction: Why a new approach

Classical methods to estimate the GDP are regressions and factors models and are based on a lot of assumptions

1. Linearity (parametric models)
2. Stationarity (data sets)
3. Gaussianity (tools for estimation)

SPM: First step, the bridging equations

The bridging quarterly GDP with monthly data are based on linear regressions of quarterly GDP growth on a small set of key monthly indicators.

This simple modelling strategy has been popular among policy institutions which commonly pooled several GDP forecasts from bridge equation models so as to consider a large number of predictors.

SPM: A review on Bridging equations

1. Small set of key monthly indicators: Kitchen and Monaco (2003), Baffigi, Golinelli and Parigi (2004), Barhoumi, Darné, Ferrara and Pluyaud (2008).
2. Combination of small number of univariate equations: Runstler and Sedillot, (2003), Diron, (2008).
3. Factor models link with regressions: Angelini, Camba-Mendez, Giannone, Riechlin and Runstler (2008).
4. Factors through principal components : Stock and Watson (2002), Schumacher and Breitung (2006), Marcellino and Schumacher (2008).

SPM: Why?

- ▶ We want to use of small amount of information. Indeed, it is difficult to exploit all information available for nowcasting, as business cycle indicators are released in an asynchronous way.
- ▶ Due to these different publication lags, multivariate datasets typically exhibit complicated patterns of missing values at the end of the sample and imply unbalanced samples for estimation.
- ▶ This leads to the so-called 'ragged-edge' data problem in econometrics (Wallis, 1986), and nowcast methods are necessary that can tackle this issue.
- ▶ Thus, bridging equations appear as a possible deal to avoid all these problem.
- ▶ What is new here: non-parametric estimation (Silverman, 1952) of the monthly indicators.

SPM: The principle

We introduce a new statistical tool to complete the information set used to nowcast and forecast the quarterly GDP growth.

It exploits monthly data used to nowcast quarterly GDP and ragged edges.

1. First we provide robust forecasts for the monthly variables via non-parametric methods. (13 indicators)
2. Second we use the predictive bridge equations to bridge monthly information with quarterly ones. (8 equations) (Diron, 2008).

In fine this approach will be named "a semi-parametric modelling for the quarterly GDP" in the sense that we use non-parametric method to forecast the monthly variables and parametric equations to compute the estimation of the GDP, with least squares method.

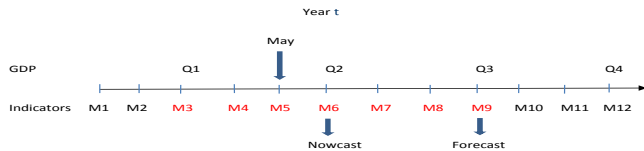
SPM: The model

If we divide the year in 4 quarters $Q_i, i = 1, \dots, 4$, and in 12 months $M_j, j = 1, \dots, 12$.

- ▶ At the year t , assuming that we are in month M_4 : we want to backcast the quarter Q1 and nowcast the quarter Q2 for GDP growth.
- ▶ At the year t , if we are in month M_5 , we want to nowcast the quarter Q2 and forecast the quarter Q3 for GDP growth.
- ▶ In that latter case, we will nowcast the 13 monthly variables until M_6 and forecast them until M_9 , and plug them in the bridge equations.

Objectives: the rag-edged monthly variables.

Nowcasting



NPM: Why the non-parametric methods

1. Working with non parametric approach permits to be free of a lot of assumptions concerning the datasets on which we work.
2. The methods are close to a non-linear modelling and appear more flexible than the classical linear regressions.
3. It is non necessary to make the data sets stationary: this means that we can work with data presenting some trend for instance.
4. We are free of the distribution law of the data set.

NPM: The basic method

We consider the simple problem which consists to estimate the relationship between two random variables, say X and Y .

Regression analysis is concerned with the question of how Y (the dependent variable) can be explained by X (the explanatory or regressor variable).

Thus this means a relationship of the form:

$$Y = m(X), \quad (1)$$

where $m(\cdot)$ is a function in the mathematical sense. We do not assume here any specific hypotheses on $m(\cdot)$.

The random variables X , in the expression (1), can be any variables and, in particular if we work at time t , they can be past values of the random variables Y_t .

NPM: The basic method

The underlying principle that theoretic laws usually do not hold in every individual case but merely on average is considered here and can be formalized as:

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \dots, T, \quad (2)$$

$$E[Y|X = x] = m(x). \quad (3)$$

Equation (2) says that the relationship $Y = m(X)$ does not need to hold exactly for the i th observation but is disturbed by the random variable ε .

NPM: The basic method

We assume that at time t , we observe X_t and its past and for estimation purpose we use the data set X_1, \dots, X_T . Thus, we estimate the function $m(\cdot)$ such that

$$X_{t+1} = m(\underline{X}_t) + \varepsilon_t, \quad (4)$$

where \underline{X}_t is a set of variables taken in the past of X_{t+1} .

We are going to estimate $m(\cdot)$ in the following way:

$$\hat{m}(x) = \frac{1}{T} \sum_{i=1}^T \omega_{i,T}(x) Y_i, \quad (5)$$

where $\omega_{i,T}$ are weights to specify and estimate.

NPM: The basic method

No transformation is necessary on the data sets to get the predictions and X_t will represent in the following the monthly indicator index known at time t and that we want to forecast at time $t + 1, , t + 2, \dots$, etc.

NPM: The k -Nearest neighbors method

The k -nearest neighbor (k -NN) estimator can be viewed as a weighted average of the response variables in a neighborhood around x , with the important fact that the neighborhood width is not fixed but variable (main difference with kernel approach, not considered here).

The values of X_{t+1} used in computing the average, are those which belong to the k observed values of X_t that are nearest the point x , at which we would like to estimate $m(x)$.

Formally the k -NN estimator of $m(x)$ can be written as (5) where the weights $\omega_{i,T}(x)$ have to be specified.

NPM: The k -Nearest neighbors method

To estimate $m(\cdot)$ given in (4), assuming that we are at time t , among the observations X_1, \dots, X_{t-1} , we are looking at the k closest neighbors of X_t .

If we denote these k points by $X_{(i)}$ $i = 1, \dots, k$, then an estimate of $m(\cdot)$ by the k -NN method permits to compute \hat{X}_{t+1} such that

$$\hat{X}_{t+1} = \sum_{i=1}^k w(\|X_t - X_{(i)}\|) X_{(i)+1}$$

where $\|\cdot\|$ is the Euclidean norm and the weights $w(\cdot)$ can be chosen in different ways.

NPM: The k -Nearest neighbors method

The weights are usually either exponential or uniform functions :

- ▶ exponential weight : $w(\|x - X_{(i)}\|) = \frac{\exp(-\|x - X_{(i)}\|^2)}{\sum_{i=1}^k \exp(-\|x - X_{(i)}\|^2)}$
- ▶ uniform weight : $w(\|x - X_{(i)}\|) = \frac{1}{k}$.

If we estimate $m(\cdot)$ at a point x where the data are sparse then it might happen that the k -nearest neighbors are rather far away from x (and each other), thus consequently we end up with a wide neighborhood around x for which an average of the corresponding values of X_{t+1} is computed. The parameter k need to be estimated.

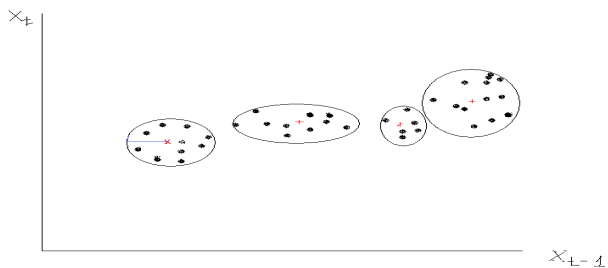
NPM: The k -Nearest neighbors method

Notice that the variance of this estimator does not depend on the distribution of the data set which makes it interesting. It is also a great difference with the kernel method.

NPM: The Radial basis function (RBF) method

- ▶ The radial basis function method is a particular case of what we call the spline smoothing.
- ▶ The motivation is based on the estimation of $m(x)$ interpolating the data, without exploiting any structure that might be present in the data.
- ▶ Working with this method, we embed the data in a space of dimension d , d is unknown.
- ▶ In this space, we cluster the data using a spline function, for instance, a radial basis function.

NPM: RBF Method



NPM: The Radial basis function (RBF) method

A radial basis function ϕ is an application defined from R^d to R^d and is characterized by its centroid c and its width r .

Using the same information set as before, the estimation of m by a set of k clusters through a radial basis function is :

$$\hat{X}_{t+1} = \sum_{i=1}^k w_i \phi(\|\underline{X}_t - c_i\|, r_i), \quad (6)$$

where $\underline{X}_t = (X_t, X_{t-1}, \dots, X_{t-(d-1)}) \in R^d$, ϕ is the radial basis functions.

The parameters of the model are k , c_i , r_i and w_i .

NPM: The Radial basis function (RBF) method

The radial basis function can be chosen among the following functions:

- ▶ Spline function : $\phi(x, r) = \frac{x^2}{r^2} \log(\frac{x}{r})$
- ▶ Gaussian function : $\phi(x, r) = \exp(-\frac{x^2}{2r^2})$
- ▶ Multiquadric function : $\phi(x, r) = \sqrt{x^2 + r^2}$
- ▶ Inverse multiquadric : $\phi(x, r) = \frac{1}{\sqrt{x^2 + r^2}}$

The procedure to estimate the parameters follows the *k*-means cluster method.

Application: NPM for six economic indexes

In a first exercise, we apply the two previous methods in order to estimate six economic indexes, for a given set of information and a given time horizon.

The six indices are Industrial Production Index, Unemployment Index, Industrial Confidence Index, Consumer Index, Retail Index and Building Index (1985 - 2008).

We provide statistics and modelling of the series using ARMA regression as a benchmark, then k -NN and RBF methods. The RMSE permits to compare the estimates obtained for these six indexes through the three previous methods.

Application: Statistics for six economic indexes

	T	Mean	Std dev	Skewness	Kurtosis	Min	Max
IPI	217	93.63	9.91	0.20	1.85	77.97	114.50
UI	181	12 636	974	-0.40	1.81	10 700	13 990
ICI	279	-5.98	8.27	-0.70	3.50	-30.20	6.50
CI	279	-11.10	6.66	-0.39	3.06	-29	2
RI	279	-6.28	6.82	-0.32	2.61	-27.10	7.40
BI	279	-16.83	12.98	-0.37	2.10	-48.10	4.60

Table: descriptive statistics of the six indices

These series cannot be characterized by a Gaussian distribution. IPI is right skewed whereas the other series are left skewed. Their kurtosis imply that the tail of these series are more thinner than the Gaussian tail except for the ICI which presents an excess of kurtosis.

Application: Modelling for six economic indexes

- ▶ ARMA modelling. For IPI, we retain an ARIMA(6,1,0) model, for ICI an AR(4) model, for RI an AR(3) model and for UI, CI and BI an AR(1) model.
- ▶ k -nearest neighbor method. We use exponential and/or Uniform weights, k varies from 1 to 5 and the predictive horizon h from 1 to 18.
- ▶ Radial basis function method. We need to vary the embedding dimension d and the number of clusters k . For all the data sets, we make d varying from 1 to 5. The number of clusters k is equal to 3,4, 5 and 7 and, we consider the following predictive horizons $h = 1, 2, 3, 6, 9, 12, 18, 24, 30$. We use different radial basic functions.

Application: Nowcasting RMSE for the six economic indexes

We provide the value of the forecasting RMSE for each series, for $h = 1$.

Series	Best k -NN	best RBF	AR
IPI	1.37	1.05	5.16
UI	114	61	95
ICI	1.22	1.06	1.62
CI	1.28	0.60	1.13
RI	3.36	2.14	2.47
BI	2.70	1.18	1.65

Table: RMSE for the three modellings when $h = 1$ for the six indexes

Application: Nowcasting of six economic indexes

This study shows the interest to use these two non-parametric methods to forecast monthly indicators.

We have found that forecasting by RBF method gives better results than using ARIMA processes. We found also that forecasting by k -NN gives better results compared with ARIMA for large horizon h .

We are going to develop this approach to nowcast and forecast the monthly indicators in order to nowcast and forecast the quarterly GDP growth.

Nowcasting: The principle

In order to nowcast and forecast the quarterly GDP quarter-over-quarter growth rate, assuming that the indicators are chosen and estimated using non-parametric methods, we use the parametric modelling proposed in the paper of Diron (2008), marginally adapted in order to take new values into account.

We use eight equations and we are dealing with a panel of $n = 13$ series of monthly variables.

We consider linear combinations of 'hard' data which are variables on actual production and demand, such as industrial production, retail sales, business and consumer confidence surveys. These indicators are used in the individual equations. We specify them now:

Nowcasting: The data sets

1. Industrial Production Index (IP),
2. Industrial Production Index in Construction (CTRP),
3. Confidence Indicator in Services (SER-CONF),
4. Retail Sales (RS),
5. New Passenger Registrations (CARS),
6. Confidence Indicator in Industry (MAN-CONF),
7. Confidence Indicator in Building (BUI-CONF),
8. Consumers Confidence Indicator (CONS-CONF)
9. Confidence Indicator in Retail Trade (RET-CONF)
10. Effective Exchange rate (EER)
11. Eurostock Index deflated bby Harmonized Index of Consumption Prices (SPI),
12. The OECD Composite Leading Indicator, trend restored (OECD-CLI),
13. The EuroCoin Indicator (EUROCOIN).

Nowcasting: The data sets

- ▶ The real-time information set starts in January 1990 when possible and ends in November 2007. Confidence Indicator in Services starts in 1995 and EuroCoin in 1999.
- ▶ We use data base provided by EABCN through their web site (www.eacbn.org).
- ▶ The vintage series for the OECD composite leading indicator are available through the OECD real-time data base (<http://stats.oecd.org/mei/>).
- ▶ The EuroCoin index is taken as released by the Bank of Italy.
- ▶ The vintage data base for a given month takes the form of an unbalanced data set at the end of the sample.

Nowcasting: The data sets

Our aim is to estimate the GDP flash estimates that were released in real-time by Eurostat from the first quarter of 2003 to the third quarter of 2007 using the non-parametric forecasts of the monthly indicators that we have previously introduced.

We use three different ways to forecast the monthly variables :

1. ARIMA(6,1,0) model,
2. k -NN method : $k = 2$
3. Radial basis function method: ϕ is a Gaussian radial basis function, $k = 6$ and $d = 3$.

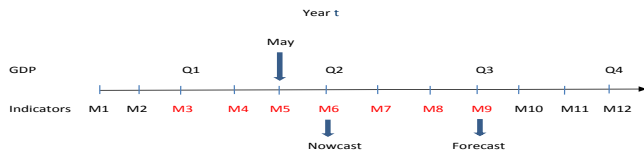
Nowcasting: The Principle

- ▶ We provide a true real-time analysis.
- ▶ We assume that GDP nowcasts and forecasts are done at each end of the month, as soon as the opinion surveys are released by the European Commission, that is around the last working day of the month.
- ▶ Thus, for a given month, we assume that we have only access to the information available at this time.
- ▶ Moreover, GDP values are those that were available at this exact date.

Nowcasting: The Principle

- ▶ For a given quarter, six GDP estimates are provided, namely 3 nowcasts and 3 forecasts.
- ▶ The first nowcast is estimated at the end of the second month of a this quarter,
- ▶ the second at the end of the third month of this quarter and
- ▶ the last at the the end of first month of the next quarter.
- ▶ It turns out that the last nowcast is done around 13 days before the flash estimate release.

Nowcasting: Comparison of the forecasting RMSE



Nowcasting: The Principle

Using five years of vintage data, from 2003 to 2007, we provide quadratic forecast errors for the euro area flash estimates of GDP growth in genuine real-time conditions.

More precisely, we provide the RMSEs of the estimates stemming from the eight equations in which we have plugged the corresponding forecast monthly indicators, as well as the RMSEs of the combined forecasts based on the arithmetic mean of these eight equations ($h = 6$ corresponds to six months before the release quarter GDP).

Nowcasting: Forecasting RMSE for GDP using AR predictions

	Fore6	Fore5	Fore4	Fore3	Fore2	Fore1
EQ1	0.21	0.21	0.23	0.21	0.22	0.22
EQ2	0.21	0.22	0.25	0.21	0.21	0.21
EQ3	0.33	0.23	0.23	0.28	0.22	0.23
EQ4	0.21	0.21	0.20	0.17	0.18	0.18
EQ5	0.24	0.24	0.23	0.20	0.21	0.22
EQ6	0.23	0.22	0.22	0.22	0.22	0.22
EQ7	0.24	0.24	0.24	0.26	0.24	0.24
EQ8	0.22	0.22	0.19	0.19	0.19	0.19
Mean	0.190	0.191	0.190	0.186	0.177	0.171

Table: RMSE for GDP using AR monthly predictions

Nowcasting: Forecasting RMSE for GDP using k -NN predictions

	Fore6	Fore5	Fore4	Fore3	Fore2	Fore1
EQ1	0.34	0.38	0.37	0.44	0.33	0.24
EQ2	0.39	0.44	0.41	0.52	0.36	0.24
EQ3	0.23	0.28	0.25	0.26	0.22	0.23
EQ4	0.21	0.23	0.18	0.17	0.18	0.18
EQ5	0.21	0.23	0.21	0.20	0.21	0.22
EQ6	0.23	0.22	0.22	0.22	0.22	0.22
EQ7	0.21	0.21	0.21	0.22	0.22	0.23
EQ8	0.22	0.21	0.20	0.19	0.19	0.19
Mean	0.211	0.217	0.189	0.207	0.179	0.161

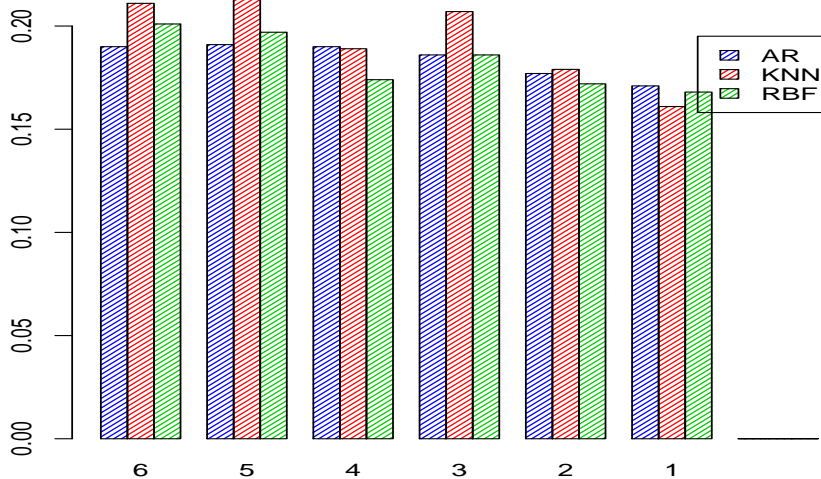
Table: RMSE for GDP using k -NN monthly predictions

Nowcasting: Forecasting RMSE for GDP using RBF predictions

	Fore6	Fore5	Fore4	Fore3	Fore2	Fore1
EQ1	0.25	0.25	0.24	0.28	0.25	0.23
EQ2	0.29	0.27	0.26	0.31	0.27	0.23
EQ3	0.25	0.25	0.23	0.24	0.22	0.23
EQ4	0.21	0.21	0.19	0.18	0.18	0.18
EQ5	0.21	0.20	0.20	0.20	0.21	0.22
EQ6	0.28	0.27	0.27	0.27	0.27	0.27
EQ7	0.25	0.26	0.26	0.30	0.24	0.24
EQ8	0.22	0.21	0.20	0.19	0.19	0.19
Mean	0.201	0.197	0.174	0.186	0.172	0.168

Table: RMSE for GDP using RBF monthly predictions

Nowcasting: Comparison of the forecasting RMSE



Summary of the results

h	ARIMA(6,1,0)	RBF ($k=6, d=3$)	2-NN
6	0.190	0.201	0.211
5	0.191	0.197	0.217
4	0.190	0.174	0.189
3	0.186	0.186	0.207
2	0.177	0.172	0.179
1	0.171	0.168	0.161

Table: RMSE with AR, NN and RBF prediction for the estimated mean quarterly GDP

Nowcasting: Comments on the results

1. We beat one-step ahead ARIMA forecasts and more: $h = 1$ to $h = 4$.
2. GDP forecasts for longer horizon ($h = 5$ and $h = 6$) are less accurate with k -NN and RBF methods : we can improve these results using the best forecasts for each method.
3. We observe that for some of the eight equations, the results with non-parametric methods are not optimal. Especially, equations 1 and 2 with the k -NN method: this result can be improved with other values of k .

Nowcasting: Some remarks

1. Exercise to improve using different sets of parameters for the both methods.
2. No transformation of the data sets even when there is evidence of non-stationarity. Exercise to perform with k -NN and RBF.
3. Extensions of the work with other NP methods: Wavelets and neural networks.

Conclusion: Main contribution

- ▶ The contribution of this paper is a new way to rag edge monthly data which appear in the computation of the quartely GDP growth, using non-parametric methods based on nearest neighbors and radial basis function approaches.
- ▶ We show that, with this new approach, we beat the predictions done with ARIMA models for short horizon, this suggests that we would be able also to beat multi step.
- ▶ The interest of the method is that we are free of specific assumptions to obtain the relevant forecasts.

Conclusion: Other key features

A number of studies have used univariate forecasting ("bridge") equations to obtain short-term predictions of GDP from monthly indicators: we are in this context.

1. Number of equations?
2. Choice of the indicators?
3. Are regressions relevant in fine?