

The Multivariate k-Nearest Neighbor Model for Dependent Variables : One-Sided Estimation and Forecasting

Dominique Guegan, Patrick Rakotomarolahy

▶ **To cite this version:**

Dominique Guegan, Patrick Rakotomarolahy. The Multivariate k-Nearest Neighbor Model for Dependent Variables : One-Sided Estimation and Forecasting. 2009. halshs-00423871v2

HAL Id: halshs-00423871
https://shs.hal.science/halshs-00423871v2

Submitted on 5 Jan 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Documents de Travail du
Centre d'Economie de la Sorbonne

CES Working Papers

The Multivariate k-Nearest Neighbor Model for Dependent Variables : One-Sided Estimation and Forecasting

Dominique GUEGAN, Patrick RAKOTOMAROLAHY

2009.50

Version révisée

Maison des Sciences Économiques, 106-112 boulevard de L'Hôpital, 75647 Paris Cedex 13
http://ces.univ-paris1.fr/cesdp/CES-docs.htm
ISSN : 1955-611X

The Multivariate k-Nearest Neighbor Model for Dependent Variables: One-Sided Estimation and Forecasting

Dominique Guégan*, Patrick Rakotomarolahy†

Abstract

This article gives the asymptotic properties of multivariate k-nearest neighbor regression estimators for dependent variables belonging to $R^d, d > 1$. The results derived here permit to provide consistent forecasts, and confidence intervals for time series. An illustration of the method is given through the estimation of economic indicators used to compute the GDP with the bridge equations. An empirical forecast accuracy comparison is provided by comparing this non-parametric method with a parametric one based on ARIMA modelling that we consider as a benchmark because it is still often used in Central Banks to nowcast and forecast the GDP.

Keywords: Multivariate k-nearest neighbor - Asymptotic normality of the regression - Mixing time series - Confidence intervals - Forecasts - Economic indicators - GDP - Euro area.

JEL: C22 - C53 - E32.

1 Introduction

In econometrics, estimation and forecasting problems are common topics. There exists many methods to estimate and predict time series. The more popular being based on parametric models. With linear models, a fundamental reference is the book of Box and Jenkins (1970). Recent

*Paris School of Economics, CES-MSE, Université Paris 1 Panthéon-Sorbonne, 106 boulevard de l'Hopital 75647 Paris Cedex 13, France, e-mail: dguegan@univ-paris1.fr

†CES-MSE, Université Paris 1 Panthéon-Sorbonne, 106 boulevard de l'Hopital 75647 Paris Cedex 13, France, e-mail: rakotopapa@yahoo.fr

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

developments including non-linear models like the related GARCH models, SETAR-STAR models or Markov switching models have been also developed, Tong (1990), Krolzig (1998), Pena, Tiao and Tsay (2003). We can also make estimations and predictions from non-parametric techniques. Indeed, the use of non-parametric techniques has a long tradition in time series analysis. The advantage of non-parametric methods, unlike the methods mentioned previously is based on the fact that they let the data speak for themselves. Hence it avoids the subjectivity of choosing a specific parametric model before looking at the data. However there is the cost of more complicated mathematical arguments such as the selection of smoothing parameters. Nevertheless recent studies help to avoid these problems and also the speed of computers that can develop search algorithms from appropriate selection criteria, Silverman (1986), Devroye and Gyorfi (1985), Becker, Chambers and Wilks (1988). We favor here these nonparametric tools for time series study.

When the final objective of time series analyses is prediction, it is of interest to study the conditional means and conditional variances in some period, given the past of the process. When a point prediction is the final objective, an estimate of some conditional mean is desired, while the conditional variances are needed if interval forecasts are desired.

In this paper, we focus on the analysis of the conditional mean non-parametrically in order to build consistent forecasts. There are numerous non-parametric techniques used in time series analysis to estimate the conditional mean: the kernel methods, the wavelet techniques, the neural networks, the spline functions and the k-nearest neigbors method among others, Prakasa Rao (1983), Donoho and Johnstone (1992), Kuan and White (1994), Friedman (1988) and Mack (1981) for instance.

Given a time series, $X_1, \cdots, X_n$, in order to analyse the conditional mean non-parametrically one may consider the following representation for the regression function:

$$m(x) = E[X_{n+1}|X_n = x]. \tag{1.1}$$

Thus, model (1.1) has the format of a nonlinear regression problem for which many smoothing methods exist, Hart (1997).

In this paper we focus on dependent variables that may characterized by equation (1.1) and we want to reconstruct the function $m(.)$ using multivariate k-nearest neighbors. We limit ourselves to this method because it has many advantages in practice: among the non-parametric methods it is certainly the easiest to understand and implement. Working in a multivariate environment allows us to discover and take into account the structural behavior that can not always discern the path. Finally, recent results have also made significant contributions to select the number of neighbors within a given space, Ouyang, Li and Li (2006). In this paper, we focus on the multivariate k-NN method.

Re-exploiting the k-nearest neighbors (k-NN) method in order to reconstruct multivariate time series for forecasting, we focus on estimation problems for regression under weak assumptions, in the multivariate setting. We provide new results which concern the asymptotic normality of the function $m(.)$ for dependent variables, with respect to the bias-variance fit dilemma inherent in this kind of methodology. Our result permits a confidence interval, to be obtained which can be used to discriminate between classical methods.

Our proposal can be considered as a contribution of the general problem concerning the non-parametric estimate of a regression with k-NN method, extending well known results obtained for i.i.d. variables, Mack (1981) and Stute (1984) (they derived the asymptotic normality of the regression). Other classes of convergence for independent samples have also been developed by Stone (1977), and Devroye (1982). In case of dependent variables, Collomb (1984) provides piecewise convergence for univariate variables and Yakowitz (1987) gets the quadratic mean error for uniformly weighted k-NN estimates for univariate samples. Here, working with multivariate time series, we control the bias of a general multivariate k-NN estimate, using several weights, and we establish the asymptotic normality of this estimate which permits exact confidence intervals to be constructed.

Reconstruction of time series using multivariate k-NN method allows obtaining robust estimates under relatively weak assumptions whose use is interesting in practice. The method of nearest

neighbors has been used in finance and has shown the benefits of the method, Mizrach (1992), Meade (2002), Nowman and Saltoglu (2003), and Guégan and Huck (2005). In economics this method is still little known and therefore rarely used, an interesting review is Yatchew (1998). An area of very interesting application in economics is the forecast of GDP. Indeed, predict GDP is an important challenge for many institutions, particularly central banks. In these lattest institutions, many studies have been developed to solve the problem of nowcasting and forecasting the GDP. The studies focused on the number of economic indicators to be used, ranging from a large number or numbers rather parsimonious. But whatever the problem considered, parametric models have been generally accepted: we can cite the VAR models, Marcellino *et al.* (2006), the dynamic factor models, Bernanke and Boivin (2003), Forni *et al.* (2005), Kapetanios and Marcellino (2006), non-linear modelling, González, Hubrich and Teräsvirta (2009), or the bridge equations, Baffigi, Golinelli and Parigi (2004), Diron (2008). Few studies use a non-parametric approach, apart in our knowledge Blake (1999) or Tkacz and Hu (1999) with neural networks.

Here, we limit ourselves to the Diron approach to calculate GDP. Her method uses a limited number of economic indicators which are immersed in 8 equations from which an estimate of GDP is obtained. This approach has been used in several central banks to estimate the GDP using the economic indicators being estimated by simple linear models like ARIMA processes, Runstler and Sedillot (2003), and Darne (2008) for instance. To demonstrate the usefulness of the methodological approach developed in this paper, we estimate the economic indicators that occur in the Diron equations using the method of multivariate nearest neighbors, and we plug them into the eight equations for getting the GDP. We compare our result to that obtained when one considers the indicators estimated using an ARIMA modelling, considering this modelling as a benchmark. For the empirical study, each method is formulated and estimated on a sub-sample of the historical data, and its forecasts of the observation held back at the model specification stage are then evaluated.

The paper is organized as follows. In Section 2, we establish our main result: the asymptotic normality of the multivariate k-NN regression estimate for a mixing time series. In Section 3, an empirical forecast accuracy comparison of non-parametric and parametric approaches is provided. Section 4 concludes and Section 5 is devoted to the proofs.

2 Main result

We consider a time series observed in $\mathbb{R}$, and we transform the original data set by embedding it in a space of dimension d, building $\underline{X}_n = (X_{n-d+1}, \cdots, X_n) \in \mathbb{R}^d$. The embedding concept is important because it allows to take into account the characteristics of the series that are not observed on the trajectory in $\mathbb{R}$.

We are interested to get an estimate of $m(\underline{x})$, $\underline{x} \in \mathbb{R}^d$, using the k closest vectors to $\underline{X}_n = \underline{x}$ inside the training set $S = \{\underline{X}_t = (X_{t-d+1}, \cdots, X_t) \mid t = d, ..., n\} \subset \mathbb{R}^d$. We define a neighborhood around $\underline{x} \in \mathbb{R}^d$ such that $N(\underline{x}) = \{i \mid i = 1, \cdots, k(n)$ whose $\underline{X}_{(i)}$ represents the i^{th} nearest neighbor of $\underline{x}$ in the sense of a given distance measure$\}$. Then the k-NN regression estimate of $m(\underline{x})$, $\underline{x} \in \mathbb{R}^d$ is given by:

$$m_n(\underline{x}) = \sum_{\underline{X}_{(i)} \in S, i \in N(\underline{x})} w(\underline{x} - \underline{X}_{(i)}) X_{(i)+1}, \tag{2.1}$$

where $w(\cdot)$ is a weighting function associated with neighbors. We introduce the three most weights used in the literature for k-NN regression estimates, in which it is noteworthy that the parameter k needs to be estimated.

1. The first one corresponds to the so called uniformly k-NN weight :

$$m_n(x) = \frac{1}{k} \sum_{i \in N(x)} X_{(i)+1} \quad i.e \quad w(x - \underline{X}_{(i)}) = \frac{1}{k} \tag{2.2}$$

2. The second one is the non-uniformly k-NN weight not function of X_n:

$$m_n(x) = \sum_{i \in N(x)} w_i X_{(i)+1}, \quad w_i \in \mathbb{R}, \quad i.e \quad w(x - \underline{X}_{(i)}) = w_i. \tag{2.3}$$

3. The third one is the exponentially k-NN weight which takes into account the distance between the point observation and the neighbors, and depends on X_n:

$$m_n(x) = \sum_{i \in N(x)} \frac{exp(-||\underline{x} - \underline{X}_{(i)}||^2)}{\sum_{i \in N(x)} exp(-||\underline{x} - \underline{X}_{(i)}||^2)} X_{(i)+1} \tag{2.4}$$

A general form for the weights is:

$$w(\underline{x} - \underline{X}_{(i)}) = \frac{\frac{1}{nR_n^d} K(\frac{\underline{x} - \underline{X}_{(i)}}{R_n})}{\frac{1}{nR_n^d} \sum_{i=1}^n K(\frac{\underline{x} - \underline{X}_{(i)}}{R_n})}, \tag{2.5}$$

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

where R_n corresponds to the distance between $\underline{x}$ and the further k-nearest neighbors and, $K(\cdot)$ is a given weighting function. A link between this general form of weighting function and the previous three k-NN weights can be obtained by taking for the exponential weighting function, $K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n}) = exp(-||\underline{x} - \underline{X}_{(i)}||^2)$ and, for the uniform weighting function, $K(\frac{\underline{x}-\underline{X}_{(i)}}{R_n}) = \frac{1}{k}$.

The following result extends the asymptotic convergence of the estimation of the regression when it is estimated by nearest neighbors belonging to $\mathbb{R}^d$, to the case of dependent variables. This result is crucial for applications since it can provide estimates and therefore forecasts for the conditional mean of time series whose use in economics and finance is common. It can therefore leave the independent framework without having to filter the observed data. The knowledge of the bias and speed rate of variance provide consistent estimates, and the asymptotic normality of the estimates permits to build confidence intervals. The building of confidence intervals is often used to compare the quality of forecasts obtained from different methods, and permit to enhance the comparison of several methods (parametric and non-parametric methods), beyond point forecast, in forecasting. Indeed, no test is available to discuss the choice between the parametric and the non-parametric approaches. To establish our main result, we assume that the data $(X_n)_n$ are strictly stationary time series. They are characterized by an invariant measure with density f, the random variable $X_{n+1} \mid (\underline{X}_n = \underline{x})$ has a conditional density $f(y \mid \underline{x})$, and the invariant measure associated to the embedded time series $\{\underline{X}_n = (X_{n-d+1}, \cdots, X_n)\}$ is h.

Theorem 2.1. *Assuming that $\{X_n\}$ is a stationary time series, and that the following assumptions are verified:*

(i) $(X_n)_n$ is ϕ-mixing.

(ii) $m(\underline{x}), f(y \mid \underline{x})$ and $h(\underline{x})$ are p continuously differentiable.

(iii) $f(y \mid \underline{x})$ is bounded,

(iv) the sequence $k(n) < n$ is such that $\sum_{i=1}^{k(n)} w_i \to 1$ as $\quad n \to \infty$,

then k-NN regression function $m_n(\underline{x})$ defined in (2.1) verifies:

$$\sqrt{n^Q}(m_n(\underline{x}) - Em_n(\underline{x})) \to_{\mathcal{D}} \mathcal{N}(0, \sigma^2), \tag{2.6}$$

with

$$E[(m_n(\underline{x}) - m(\underline{x}))^2] = O(n^{-Q}), \tag{2.7}$$

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

where $0 \leq Q < 1$, $Q = \frac{2p}{2p+d}$, *and*

$$\sigma^2 = \gamma^2(Var(X_{n+1} \mid \underline{X}_n = \underline{x}) + B^2),$$

with $B = O(n^{-\frac{(1-Q)p}{d}})$, *and* γ *a positive constant which is equal to 1 when we use uniform weights.*

The proof of the theorem is postponed to the end of the article.

The conditions under which this result is established are quite weak in time series analysis. We know that mixing conditions permit to consider data sets which are asymptotically independent. Parametric processes like the bilinear models including ARMA models, related GARCH processes and Markov switching processes are known to be mixing, Guégan (1983) and Carrasco and Chen (2002). The condition (iv) is in particular verified for the weights introduced in (2.2) - (2.4), and (2.5). Note that the parameter γ represents the correlations between the vectors $\underline{X}_n$. This theorem is interesting because it provides asymptotic normality for the estimate $m_n(x)$ under regular conditions and thus permits to build confidence intervals. In order to construct a confidence interval for the k-NN estimate regression, we establish the following result.

Corollary 2.1. *Under the assumptions of Theorem (2.1), a general form of the confidence interval for* $m(\underline{x})$, *for a given* α, *is :*

$$m(\underline{x}) \in [m_n(\underline{x}) - B - \frac{\hat{\sigma} z_{1-\frac{\alpha}{2}}}{\sqrt{k}}, m_n(\underline{x}) + B + \frac{\hat{\sigma} z_{1-\frac{\alpha}{2}}}{\sqrt{k}}] \tag{2.8}$$

where $z_{1-\frac{\alpha}{2}}$ *is the* $(1-\frac{\alpha}{2})$ *quantile of the Student law,* $\hat{\sigma}$ *is an estimate for* σ *and* B *is such that:*

1. B *is negligible, if* $\frac{k(n)}{n} \to 0$, $n \to \infty$,
2. *If not,* $B = O(r^p)$, *with* $r = \left(\frac{k(n)}{(n-d)\hat{h}(\underline{x})c}\right)^{\frac{1}{d}}$ *where* $c = \frac{\pi^{d/2}}{\Gamma((d+2)/2)}$, *and* $\hat{h}(\underline{x})$ *is an estimate for the density* $h(\underline{x})$.

The proof of this lemma is moved to the end of the article.

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

3 An application: the forecast of the Euro-area GDP

Information on the current state of economic activity is a crucial ingredient for policy making. Economic policy makers, international organisations and private sector forecasters commonly use short term forecasts of real gross domestic product (GDP) growth based on monthly indicators. For users, an assessment of the reliability of these tools, and of the source of potential forecast errors is essential. There exists many studies proposing real-time modelling in order to take into account some complexity inherent to the computation of the GDP which are: the number of economic indicators, the modelling for GDP and the impact of data revisions. In the exercise that we present below, we wish to show that beyond the model chosen to calculate the GDP in the end, the forecasts of monthly economic indicators used in the final model are fundamental and may be misleading not negligible if they are not properly estimated.

We therefore consider the approach of bridge equations to calculate the GDP in the final stage, following the original work of Runstler and Sedillot (2003). We limit ourselves to the 8 equations introduced in the paper of Diron (2008), each equation providing a model of GDP, denoted $Y_t^i, i = 1, \cdots 8$. They are finally aggregated consistently to provide a final value of GDP, denoted Y_t (These equations are repeated at the end of paper). Each equation is calculated from monthly economic indicators, 13 in total denoted $X_t^i, i = 1, \cdots, 13$. We are interested here in forecasting these indicators. We propose to estimate and provide forecasts of these indicators from three models: the ARIMA modelling, still widely used in central banks, and the approach of nearest neighbors. For the latter method we distinguish forecasts obtained without embedding (d = 1) from forecasts obtained when d$>$ 1. It appears clearly that using the method of nearest neighbors with embedding, we called multivariate nearest neighbors above, can significantly improve the quality of the forecast of GDP in fine. The thirteen economic indicators that we consider are listed in a table at the end of the paper.

For this exercise,, we use the real-time data base provided by EABCN through their web site [1]. The real-time information set starts in January 1990 when possible (exceptions are the confidence indicator in services, that starts in 1995, and EuroCoin, that starts in 1999) and ends

[1] www.eabcn.org

in November 2007. The vintage series for the OECD composite leading indicator are available through the OECD real-time data base [2]. 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. To solve this issue, we apply the three methodologies to forecast the monthly variables in order to complete the values until the end of the current quarter for GDP nowcasts and until the end of the next quarter for GDP forecasts, then we aggregate the monthly data to quarterly frequencies.

The parametric modelling is based on ARIMA(p,d,0) processes. For each economic indicator, we selected the best ARIMA model under the criterion of Akaike, Akaike (1974). That means we do not necessarily use the same ARIMA model for all 13 indicators. The parameters of the models are estimated by least square estimation method. Regarding the method of nearest neighbors, when d = 1, we determine the number k of neighbors by minimizing the following criterion (mean square error criterion, RMSE): $\sqrt{\frac{1}{n-k}\sum_{t=k}^{n}||\hat{X}_{t+1}^{i} - X_{t+1}^{i}||^2}$, $i = 1, \cdots 13$, where n is the sample size, $\hat{X}_{t+1}^{i}$ is the estimate of the i-th economic indicator X_{t+1}^{i} calculated from the expression (2.1), when $d = 1$. The number k of nearest neighbors determined by this criterion at the horizon h = 1 is used to calculate the forecasting capabilities for h $>$ 1. Of course this work is done for each economic indicator, and therefore the number of neighbors k may not be the same for all indicators. In the case of the multivariate approach ($d > 1$), we now describe the algorithm used to determine the embedding dimension and the number of neighbors used to obtain the best predictor for X_{n+h}^{i} in the sense of RMSE. We present the method for all indicators and thus, for simplicity, we drop the index i in the algorithm. We assume that we have a data set $X_1, ..., X_n$ in $\mathbb{R}$.

1. We embed this data set in a space of dimension d, $2 < d \leq 10$. We obtain a sequence of vectors in $\mathbb{R}^d$: $\{\underline{X}_d, \underline{X}_{d+1}, ..., \underline{X}_n$, where $\underline{X}_i = (X_{i-d+1}, ..., X_i)\}$.

2. We determine the k nearest vectors of $\underline{X}_n$ among the above vectors. If we denote $r_i = \|\underline{X}_n - \underline{X}_i\|$, $i = d, d+1, ..., n-1$, the distance between these vectors, we build the sequence $r_d, r_{d+1}, ..., r_{n-1}$ ordered in an increasing way: $r_{(d)} < r_{(d+1)} < ... < r_{(n-1)}$.

3. We detect the k vectors, denoted $\underline{X}_{(j)}$, corresponding to $r_{(j)}$, $j = d, d+1, ..., d+k-1$ (the

[2] http://stats.oecd.org/mei/

k smallest distance).

4. To compute $m_n(\underline{X}_n) = \hat{X}_{n+1}$, we use the following expression:

$$\hat{X}_{n+1} = \sum_{j=d}^{k+d-1} w(\left\|\underline{X}_n - \underline{X}_{(j)}\right\|)X_{(j)+1} \tag{3.1}$$

   By the way, we obtain the one step ahead forecast.

5. Now, we consider as information set: $X_1, ..., X_n, \hat{X}_{n+1}$ and redo step 1 to step 4, we get the two step ahead forecast. We obtain the forecast of third step ahead in a similar way as the obtention of two step ahead forecast. And so on $\cdots$.

We restrict ourselves to exponential weights because they can give more weight to nearest neighbors: it is a property specific to the method of nearest neighbors that we want to favor here by choosing that type of weights rather than the uniform weights that give the same importance to all neighbors. For each indicator $X_t^i, i = 1, \cdots 13$, the best pair (d, k) is determined again by minimizing the criterion $\sqrt{\frac{1}{n-k-d}\sum_{t=k+d}^{n}||\hat{X}_{t+1}^i - X_{t+1}^i||^2}$. Once the pair (d, k) is found, it is used for all prediction horizons. Note that the pair (d, k) may be different for all indicators.

As soons as the three modellings are retained, we compute 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 previous forecasts of the monthly indicators. According to this scheme, the monthly series have to be forecast for an horizon h varying between 3 and 6 months in order to complete the data set at the end of the sample. Recall that the h-step-ahead predictor for $h > 1$ is estimated recursively starting from the one-step-ahead formula.

Using five years of vintage data, from the first quarter 2003 to the third quarter 2007, we provide RMSEs for the Euro area flash estimates of GDP growth in genuine real-time conditions, $\hat{Y}_t$. We have computed the RMSEs for the quarterly GDP flash estimates, obtained with three forecasting methods used to complete adequately in real-time the monthly indicators, that is ARIMA modelling and k-NN method ($d = 1$ and $d > 1$). More precisely, we provide the RMSEs of the combined forecasts based on the arithmetic mean of the eight equations of Diron (2008). Thus, for a given forecast horizon h, we compute $\hat{Y}_t^j(h)$ which is the predictor stemming from Diron's equation $j == 1, \cdots 8$, in which we have plugged the forecasts of the monthly economic

indicators, and we compute the final estimate GDP at horizon h: $\hat{Y}_t(h) = \frac{1}{8}\sum_{j=1}^{8} \hat{Y}_t^j(h)$. the RMSE criterion for the final GDP is

$$RMSE(h) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{Y}_t(h) - Y_t)^2}, \tag{3.2}$$

where T is the number of quarters between Q1 2003 and Q4 2007 (in our exercise, $T = 19$) and Y_t is the Euro area flash estimate for quarter t. The RMSE errors for final GDP are provided in table 1 and comments follow.

h	ARIMA	k-NN(1)	k-NN(d>1)
6	0.249	0.198	0.214
5	0.221	0.203	0.192
4	0.216	0.202	0.196
3	0.195	0.186	0.177
2	0.191	0.176	0.177
1	0.175	0.174	0.171

Table 1: RMSE for the estimated mean quarterly GDP Y_t computed from equation (3.2), using ARIMA(p,d,0) modelling (column 2) and k-NN predictions ($d = 1$ (column 3), and $d > 1$ (column 4)), for the monthly economic indicators $X_t^i, i = 1, \cdots 13$, h is the monthly forecast horizon.

When comparing column 2 on one side and columns 3 and 4 of the other side, we find that forecast errors are always lower with the method of nearest neighbors. For ARIMA modelling and k-NN method with $d > 1$ the RMSE becomes lower when the forecast horizon reduces from $h = 6$ to $h = 1$, illustrating that the accuracy of the nowcasting and forecasting increases as soon as the information set grows thanks to the released monthly data. Indeed, few days before the publication of the flash estimate (around 13 days with $h = 1$), the lowest RMSE is obtained with the k-NN method (RMSE=0.171), with $d > 1$. If now we restrict to the nonparametric procedures, we obtain smaller errors if we work with the multivariate setting than with the univariate one. This result shows the superiority of the method developed in a space of dimension $d > 1$. Indeed, we believe that in terms of predictions, any method developped in an area that does not restrict the use of information given by only one path must lead to improved forecasts. This is confirmed when one compares, for the same method, the forecast errors obtained only in

$\mathbb{R}$ with the error calculated from a treatment in $\mathbb{R}^d$: in this case the errors are always smaller (e.g. compare the columns 3 and 4 of Table 1). This idea is also present inside the works of Kapetanios and Marcellino (2006), for example, when working with multivariate factor models, they improve predictions of GDP comparing it with those of univariate models. Interesting work is now to compare parametric multivariate models with our approach. The great difference still lies in the fact that factor models typically use a large number of factors to be efficient, which is not the case here. However, this work is done and will be a companion paper.

4 Conclusion

We have proved the consistency (bias, L^2 convergence) and the asymptotic normality of multivariate k-NN regression estimate for dependent time series. We also provided interval forecasts for such regression estimates. To illustrate our new methodology, we provide an exercise permitting to show that this method can consistently improve the nowcast and forecast of GDP.

In that last exercise, we have been particularly interested in detecting the best predictor for economic indicators using the RMSE criterion. During the period of estimation, we were not interested to get an economic interpretation for the couples (d, k) that achieve the result. For this part of our work applications, we are quite close to philosophy developped in the works of data mining, focusing on the relevant set of data permitting to solve a specific problem with respect to appropriate criteria, Han et al. (1996) and, for a deeper discussion on this last subject, Hoover and Perez (1999).

Concerning the use of non-parametric techniques, we can recall that the more data is large, the better the forecasts: this problem seems to be more sensitive when using non-parametric techniques than parametric modelling, although no detailed quantitative study have only been conducted on the subject. We beleive that the method of nearest neighbors on both its simplicity still requires less data to provide robust estimates. Nevertheless the problem with amount of data would needs to be studied more thoroughly, but this subject is outside the scope of this paper.

The application we developed here also shows the interest of the method to relatively weak as-

sumptions we have to check to get good predictions. Nevertheless it is important to note that we work with stationary data. Indeed, we make the data stationary (by differentiation) and then we calculate the predictions regardless of the method used. However we are liberated from the assumption of Gaussianity for both methods (parametric and non-parametric). Finally, we believe that the method of nearest neighbors in higher dimension can significantly improve the forecasts, for work in embedding allows to take into account non-linear structures that are not taken into account either by ARIMA model, nor by the nearest neighbors method working on the trajectory.

A possible extension of this work is to compare these results with other methods, parametric and nonparametric developed in higher dimension, for example with the method of radial functions for the latter approach. This will be the subject of a future paper. It would be interesting to have a robust estimation method for detecting the same time, the number of neighbors, the embedding dimension and the weight functions. An open problem also when working with the nearest neighbors method is to find the right approach in the presence of nonstationary data, especially when a trend is present on some parts of the observed trajectory. This problem is not addressed in this paper.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control,* AC.19.

Baffigi, A.R., R. Golinelli and G. Parigi (2004). Bridge models to forecast the euro area GDP. *International Journal of Forecasting 20,* 447-60.

Becker, R.A., J.M. Chambers and A.R. Wilks (1988). *The new S language.* Chapman and Hall: New York.

Blake, A.P. An Artificial Neural Network System of Leading Indicators. *National Institute of Economic and Social Research, unpublished paper.*

Bernanke, B.S. and J. Boivin (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics 50,* 525-46.

Box, G.E.P. and G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control.* Holden Day: New York.

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various GARCH and Stochastic Volatility models. *Econometric Theory 18,* 17-39.

Collomb, G. (1984). Nonparametric time series analysis and prediction: Uniform almost sure convergence of the window and k-NN autoregression estimates. *Math Oper. Stat., Ser. Statistics 1984.*

Darne, O. (2008). Using business survey in industrial and services sector to nowcast GDP growth: The French case. *Economics Bulletin 3 (32),* 1-8.

Devroye, L.P. (1982). Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates. *Z. Wahrscheinlichkeneitstheorie Verw Gebiete 61*, 467-81.

Devroye L.P. and L. Györfi (1985). *Nonparametric Density Estimation: the L1 View.* Wiley: New York.

Diron, M. (2008). Short-term forecasts of Euro area real GDP growth: an assessment of real-time performance based on vintage data. *Journal of Forecasting 27,* 371-90.

Donoho, D.L. and I.M. Johnstone (1992). Minimax estimation via wavelet shrinkage. Technical report 402, Dept. Stat. Stantford university.

Forni, M., D. Giannone, M. Lippi and L. Reichlin (2005). Opening the black box: structural factor models with large cross-sections. *European Central Bank WP, No 571.*

Friedman, J.H. (1988). Multivariate adaptative regression splines (with discussion). *Annals of Statistics 19,* 1-141.

González, A., K. Hubrich and T. Teräsvirta (2009). *Forecasting inflation with gradual regime shifts and exogenous information.* CREATES Research Papers 2009-03, School of Economics and Management, University of Aarhus.

Guégan, D. (1983). Une condition d'ergodicité pour des modèles ilinéaires à temps discret. *CRAS Série 1, 297-301.*

Guégan, D. and N. Huck (2005). On the use of nearest neighbors in finance. *Revue de Finance 26,* 67-86.

Han, J., Y. Fu, W. Wang, K. Koperski and O. Zaine (1997). DMQL: a data mining query language for relational databases. WP Simon Fraser University, B.C. Canada.

Hart, J.D. (1997). *Nonparametric smoothing and lack-of-fit tests.* Springer verlag, New York.

Hoover, K.D. and S.J. Perez (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal 2,* 167–91.

Kapetanios, G. and M. Marcellino (2006). A parametric estimation method for dynamic factors models of large dimensions. *IGIER WP, No 305.*

Krolzig, H.M. (1998). Predicting Markov-switching vector autoregressive processes. *Mimeo, Institute of Economics and Statistics, University of Oxford.*

Kuan, C.M. and H. White (1994). Artificial neural networks : an econometric perspective. *Econometric Reviews 13,* 1-91.

Mack, Y.P. (1981). Local properties of k-NN regression estimates. *SIAM Journal on Algebraic and Discrete Methods 2,* 311-23.

Marcellino, M., J.H. Stock and M.W. Watson (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics 135,* 499-526.

Meade, M. (2002). A comparison of the accuracy of short term foreign exchange forecasting methods. *International Journal of forecasting 18,* 67-83.

Mizrach, B. (1992). Multivariate Nearest-Neighbour Forecasts of EMS Exchange Rates. *Journal of Applied Econometrics 7,* Supplement: Special Issue on Nonlinear Dynamics and Econometrics (Dec., 1992), S151-S163.

Ouyang, D., D. Li and Q. Li (2006). Cross-validation and nonparametric k nearest neighbor estimation. *Econometrics Journal 9,* 448–71.

Nowman, B. and B. Saltoglu (2003). Continuous time and nonparametric modelling of U.S. interest rate models. *International Review of Financial Analysis 12,* 25-34.

Peligrad, M. and S.A. Utev (1997). Central limit theorem for Linear Processes. *The Annals of Probability 25,* 443-56.

Pena, D., G.C. Tiao and R.S. Tsay (2003). *A course in time series analysis.* Wiley Series in Probability and Statistics: New York.

Prakasa Rao, B.L.S. (1983). *Nonparametric functional estimation.* Orlando FL, Academic press.

Rünstler, G. and F. Sedillot (2003). Short-term estimates of Euro area real GDP by means of monthly data. *European Central Bank WP, No 276.*

Silverman, B.W. (1986). *Density estimation for statistics and data Analysis.* Chapmann and Hall: London.

Stone, C. (1977). Consistent non parametric regression. *Annals of Statistics 5,* 595-645.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Annals of Statistics 12,* 917-26.

Tkacz, G. and S. Hu (1999). Forecasting GDP growth using artificial neural networks. *Bank of Canada WP, No 99.*

Tong, H. (1990). *Non-linear time series: a dynamical system approach.* Oxford University Press: Oxford.

Yakowitz, S. (1987). Nearest neighbors method for time series analysis. *Journal of Time Series Analysis 8,* 235-47.

Yatchew, A.J. (1998). Nonparametric regression techniques in economics. *Journal of Economic Literature 36,* 669-721.

5 Proofs of Theorem (2.1) and Corollary(2.1)

We start giving the proof of theorem (2.1). To prove this theorem, we first establish a lemma which provides (2.7).

Lemma 5.1. *Under the hypotheses of theorem (2.1), either the estimate $m_n(\underline{x})$ is asymptotically unbiased or*

$$E[m_n(\underline{x})] = m(\underline{x}) + O(n^{-\beta}) \tag{5.1}$$

with $\beta = \frac{(1-Q)p}{d}$.

Proof 5.1. *We denote R_n the distance between $\underline{x}$ and the $k(n)^{th}$ nearest neighbor of $\underline{x}$ and $B(\underline{x}, r_0) = \{z \in \mathbb{R}^d, \|\underline{x} - z\| \leq r_0\}$ the ball centered at $\underline{x}$ with radius $r_0 > 0$. To be sure that the $k(n)$ observations fall in the ball $B(\underline{x}, r)$, we specify r. Since the function $h(.)$ is $p-$continuously differentiable, for a given i the probability q_i of an observation $\underline{x}_i$ to fall in $B(\underline{x}, r)$ is:*

$$\begin{aligned} q_i &= P(\underline{x}_i \in B(\underline{x}, r)) && (5.2)\\ &= \int_{B(\underline{x},r)} h(\underline{x}_i) d\underline{x}_i = h(\underline{x}). \int_{B(\underline{x},r)} d\underline{x}_i + \int_{B(\underline{x},r)} (h(\underline{x}_i) - h(\underline{x})) d\underline{x}_i && (5.3)\\ &= h(\underline{x}) c r^d + o(r^d), && (5.4) \end{aligned}$$

where c is the volume of the unit ball and $\underline{x} = dx_1 dx_2 \cdots dx_d$. Thus, $q_i - q_j = o(r^d)$ for all $i \neq j$. We consider now the kth-NN vectors $\underline{x}_{(k)}$ and we denote q the probability that they are in the ball $B(\underline{x}, r)$, $q = P(\underline{x}_{(k)} \in B(\underline{x}, r))$, then :

$$q_i = q + o(r^d). \tag{5.5}$$

Being given $N(r, n)$, the number of observations falling in the ball $B(\underline{x}, r)$, for a given $r > 0$, we characterize r satisfying that $k(n)$ observations fall in $B(\underline{x}, r)$. We proceed as follows: we denote $\mathcal{S}_i^n$ all non ordered combinations of the $i-$uple indices from n indices, then:

$$\begin{aligned} E[N(r,n)] &= \sum_{i=0}^{n-d} i P(N(r,n) = i) = \sum_{i=0}^{n-d} i \sum_{(j_1,\cdots,j_i) \in \mathcal{S}_i^n} \prod_{j=j_1}^{j_i} q_j \prod_{\substack{\ell=1 \\ \ell \notin \{j_1,\cdots,j_i\}}}^{n-d} (1 - q_\ell) \\ &\geq \sum_{i=0}^{n} i \sum_{(j_1,\cdots,j_i) \in \mathcal{S}_i^n} \underline{q}^i (1-\overline{q})^{n-d-i} = \sum_{i=0}^{n} i \begin{pmatrix} n \\ i \end{pmatrix} \underline{q}^i (1-\overline{q})^{n-d-i} \qquad (5.6) \\ &= \underline{q}(n-d)(1 + \underline{q} - \overline{q})^{n-d}, \end{aligned}$$

where $\underline{q}$ and $\overline{q}$ are respectively the smallest and largest probabilities q_i $i = 1, \cdots, n-d$. Thus, we obtain a lower bound for $E[N(r, n)]$. If $E[N(r, n)] = k(n)$, using (5.4) - (5.6), we obtain:

$$r \leq \left(\frac{k(n)}{(n-d)} \right)^{\frac{1}{d}} D(\underline{x}), \tag{5.7}$$

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

with $D(\underline{x}) = \left(\frac{1}{h(\underline{x})c}\right)^{\frac{1}{d}}$.

Now, using the relationship (2.1), we have:

$$E[m_n(\underline{x})] = \sum_{i \in N(\underline{x})} E[w(\underline{x} - \underline{X}_{(i)})Y_i], \tag{5.8}$$

where $Y_i = X_{(i)+1}$. *We can remark that* $E[w(\underline{x}-\underline{X}_{(i)})Y_i] = \int_{\mathbb{R}^d}\int_{\mathbb{R}} w(\underline{x}-\underline{x}_i)yf(y,\underline{x}_i)d\underline{x}_i dy$. *Since* $f(y,\underline{x}_i) = f(y \mid \underline{x}_i)h(\underline{x}_i)$, *then we get* $E[w(\underline{x} - \underline{X}_{(i)})Y_i] = \int_{\mathbb{R}^d}\int_{\mathbb{R}} w(\underline{x} - \underline{x}_i)yf(y \mid \underline{x}_i)h(\underline{x}_i)d\underline{x}_i dy$. *Thus, as soon as the weighting function* $w(\cdot)$ *is vanishing outside the ball* $B(\underline{x}, r)$*:*

$$E[w(\underline{x} - \underline{X}_{(i)})Y_i] = \int_{B(\underline{x},r)} w(\underline{x} - \underline{x}_i)\left(\int_{\mathbb{R}} yf(y \mid \underline{x}_i)dy\right)h(\underline{x}_i)d\underline{x}_i \tag{5.9}$$

$$= \int_{B(\underline{x},r)} w(\underline{x} - \underline{x}_i)m(\underline{x}_i)h(\underline{x}_i)d\underline{x}_i. \tag{5.10}$$

To compute the bias we need to evaluate: $E[m_n(\underline{x})] - m(\underline{x})$. *We begin to evaluate :*

$$\sum_{i \in N(\underline{x})} \int_{B(\underline{x},r)} w(\underline{x} - \underline{x}_i)m(\underline{x})g(\underline{x}_i)d\underline{x}_i = m(\underline{x})E[\sum_{i \in N(\underline{x})} w(\underline{x} - \underline{X}_{(i)})] = m(\underline{x}). \tag{5.11}$$

Then,

$$E[m_n(\underline{x})] - m(\underline{x}) = \sum_{i \in N(\underline{x})} \int_{B(\underline{x},r)} w(\underline{x} - \underline{x}_i)(m(\underline{x}_i) - m(\underline{x}))h(\underline{x}_i)d\underline{x}_i. \tag{5.12}$$

The equation (5.12) holds because $\sum_{i \in N(\underline{x})} \int_{B(\underline{x},r)} w(\underline{x} - \underline{x}_i) = 1$, *(Assumption (iv) in Theorem 3.1). Then,*

$$|E[m_n(\underline{x})] - m(\underline{x})| \leq \sum_{i \in N(\underline{x})} \int_{B(\underline{x},r)} w(\underline{x} - \underline{x}_i)a \left\|\underline{x}_i - \underline{x}\right\|^p g(\underline{x}_i)d\underline{x}_i. \tag{5.13}$$

We get this last expression since a is a known constant and $m(.)$ *is* $p-$*continuously differentiable. This last inequality implies that:*

$$|E[m_n(\underline{x})] - m(\underline{x})| \leq ar^p E[\sum_{i \in N(\underline{x})} w(\underline{x} - \underline{X}_{(i)})]. \tag{5.14}$$

The previous relationship holds because $\left\|\underline{x}_i - \underline{x}\right\|^p < r^p$, *as soon as* $\underline{x}_i \in B(\underline{x}, r)$. *Now, both can be considered:*

1. *When* r *is very small, than the bias is negligible and* $E[m_n(\underline{x})] = m(\underline{x})$.

2. *If the bias is not negligible, using (5.7) and (5.14), we get:*

$$|E[m_n(\underline{x})] - m(\underline{x})| \leq a\left(\frac{k(n)}{(n-d)}\right)^{\frac{p}{d}} D(\underline{x})^p. \tag{5.15}$$

If we choose $k(n)$ integer part of n^Q, and knowing that $\frac{k}{n-d} \sim \frac{k}{n}$, then $|E[m_n(\underline{x})] - m(\underline{x})| = O(n^{-\beta})$ with $\beta = \frac{(1-Q)p}{d}$.

The proof of lemma 5.1 is complete.

Now, we prove the theorem 2.1.

Proof 5.2. 1. *We begin to establish the relationship (2.7). In the following, we denote $Y_i = X_{(i)+1}$. We rewrite the left part of (2.7) as:*

$$E[(m_n(\underline{x}) - m(\underline{x}))^2] = Var(m_n(\underline{x})) + (E[m_n(\underline{x})] - m(\underline{x}))^2. \tag{5.16}$$

We first compute the variance of $m_n(\underline{x})$, considering 2 cases:

a) First case: The weights w_i, $i = 1, ..., k$, are independent of $\{X_n\}$. In that case the variance of $m_n(x)$ is equal to:

$$Var(m_n(\underline{x})) = A + B, \tag{5.17}$$

where $A = \sum_{i=1}^{k(n)} w_i^2 Var(Y_i)$ and $B = \sum_{i=1}^{k(n)} \sum_{j \neq i} w_i w_j cov(Y_i, Y_j)$. Using the assumption (ii) of theorem 3.1, we get $|B| \leq \sum_{i=1}^{k(n)} \sum_{j \neq i} |cov(Y_i, Y_j)|$. This last term is negligible due to Yakowitz' result (1987) on the sum of covariances. Now , $A = \frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n)w_i)^2 (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2)$. Using the fact that the weights are decreasing with respect to the chosen distance, $w_k \leq \cdots \leq w_1$, we get:

$$\frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n)w_k)^2 (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2) \leq A \leq \frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n)w_1)^2 (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2). \tag{5.18}$$

As soon as $k(n) \to \infty$ the product $k(n)w_i$ converges to one in case of uniform weights, and can be bounded for exponential weights for all i and for all n, thus there exist two positive constants c_0 and c_1 such that (5.18) becomes :

$$\frac{c_1^2}{k(n)^2} \sum_{i=1}^{k(n)} (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2) \leq A \leq \frac{c_0^2}{k(n)^2} \sum_{i=1}^{k(n)} (v(\underline{x}) + (E[Y_i] - m(\underline{x}))^2). \tag{5.19}$$

where $v(\underline{x})=Var(X_{n+1} \mid \underline{X}_n = \underline{x})$. *Using the assumption (iv) of Theorem 3.1, we remark that* $E[Y_i] = E[m_n(\underline{x})]$. *Now, if* $k(n) = [n^Q]$, *then when* $n \to \infty$, $A = O(n^{-Q})$ *from lemma 5.1. It follows that the relationship (5.17) becomes:*

$$Var(m_n(\underline{x})) = O(n^{-Q}), \tag{5.20}$$

and

$$(E[m_n(\underline{x}) - m(\underline{x})])^2 = O(n^{-2\beta}). \tag{5.21}$$

Plugging equations (5.20) and (5.21) inside equation (5.16), we get $2\beta = Q$ *or* $Q = \frac{2p}{2p+d}$ *and the proof is complete.*

b) Second case: the weights w_i, $i = 1, ..., k$, *depend on* $\{X_n\}$. *We use again the relationship (5.17) with* $A = \sum_{i=1}^{k(n)} Var(w(x - \underline{X}_{(i)})Y_i)$ *and* $B = \sum_{i=1}^{k(n)} \sum_{j \neq i} cov(w(x - \underline{X}_{(i)})Y_i, w(x - \underline{X}_{(j)})Y_j)$. *Remarking that* $\{w(x-\underline{X}_{(j)})Y_j\}$ *are* ϕ*-mixing since* $\{X_j\}$ *and* $\{Y_j\}$ *are* ϕ*-mixing, then* B *is negligible from Yakowitz' result (1987). Also, we remark that* $A = \sum_{i=1}^{k(n)} (E[(w(x-\underline{X}_{(i)})Y_i)^2] - (E[w(x - \underline{X}_{(i)})Y_i])^2)$, *then*

$$A = \sum_{i=1}^{k(n)} [\int_{\mathbb{R}^d} \int_{\mathbb{R}} w(\underline{x} - \underline{x}_i)^2 y_i^2 f(y_i, \underline{x}_i) d\underline{x}_i dy_i - (\int_{\mathbb{R}^d} \int_{\mathbb{R}} w(\underline{x} - \underline{x}_i) y_i f(y_i, \underline{x}_i) d\underline{x}_i dy_i)^2]. \tag{5.22}$$

When k *increases, the weights* w_i *decrease, and* $k(n)w_i \sim \gamma$ *where* γ *is a constant, then*

$$A = \frac{\gamma^2}{k(n)^2} \sum_{i=1}^{k(n)} [\int_{\mathbb{R}^d} \int_{\mathbb{R}} y_i^2 f(y_i, \underline{x}_i) d\underline{x}_i dy_i - (\int_{\mathbb{R}^d} \int_{\mathbb{R}} y_i f(y_i, \underline{x}_i) d\underline{x}_i dy_i)^2] \tag{5.23}$$

$$= \frac{\gamma^2}{k(n)^2} \sum_{i=1}^{k(n)} (E[Y_i^2] - E[Y_i]^2). \tag{5.24}$$

As the time series (X_n) *is assumed to be stationary and with* $Y_i = X_{(i)+1}$, *then (5.24) is equivalent to* $A = \frac{\gamma^2}{k(n)}(E[X_1^2] - E[X_1]^2)$ *and* $A = \frac{\gamma^2}{k(n)} Var(X_1)$, *and finally expression (5.17) is :*

$$Var(m_n(\underline{x})) = \frac{\gamma^2}{k(n)} Var(X_1). \tag{5.25}$$

Moreover, when we take $k(n) = n^Q$, *thus (5.25) becomes:*

$$Var(m_n(\underline{x})) = O(n^{-Q}). \tag{5.26}$$

Plugging equations (5.26) and (5.21) in equation (5.16), we get $2\beta = Q$, *and* $Q = \frac{2p}{2p+d}$, *and the proof is complete.*

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

2. *We prove now the asymptotic normality. We assume that the variance* $\sigma_n = var[m_n(\underline{x})]$ *exists and is non null, thus:*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{k(n)} \frac{w_i Y_i - Ew_i Y_i}{\sigma_n}. \tag{5.27}$$

To establish the asymptotic normality of $m_n(\underline{x})$*, we distinguish 3-step corresponding to three different weighting functions.*

i) The weights are uniform, $w_i = \frac{1}{k(n)}$*, then (5.27) becomes:*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{k(n)} \frac{1}{k(n)} Z_i, \tag{5.28}$$

where $Z_i = \frac{Y_i - EY_i}{\sigma_n}$*. The asymptotic normality of (5.28) is obtained using theorem 2.2 in Peligrad and Utev (1997) . To compute the variance, we follow Yakowitz's work (1987).* $var(m_n(\underline{x})) = \frac{1}{k(n)^2} var(\sum_{i=1}^{k(n)} Y_i) = \frac{var(Y|\underline{X}=\underline{x})}{k(n)}$*, then (5.28) becomes,*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sqrt{n^Q} \sum_{i=1}^{k(n)} \frac{w_i Y_i - Ew_i Y_i}{\sigma}, \tag{5.29}$$

when $k(n) = [n^Q]$ *and* $\sigma^2 = var(Y \mid \underline{X} = \underline{x})$*, and then proof is finished.*

ii) The weights w_i *are real numbers and do not depend on* $(X_n)_n$*, then*

$$\frac{m_n(\underline{x}) - Em_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{k(n)} w_i Z_i, \tag{5.30}$$

where $Z_i = \frac{Y_i - EY_i}{\sigma_n}$*. Now, applying again the theorem 2.2 in Peligrad and Utev (1997), we get the asymptotic normality remarking that* $E[\sum_{i=1}^{k(n)} w_i Z_i] = 0$ *and* $Var[\sum_{i=1}^{k(n)} w_i Z_i] = 1$*. To compute* $\sigma_n^2 = Var[m_n(x)]$*, we use the stationarity of* $(X_n)_n$*, thus:*

$$Var[m_n(\underline{x})] = \sum_{i=1}^{k(n)} w_i^2 Var[Y_i] = \sum_{i=1}^{k(n)} w_i^2 [Var[Y_{n+1}|\underline{X}_n = \underline{x}] + B^2],$$

where B *is given in lemma 3.1. Remarking that* $\frac{1}{k(n)^2} \sum_{i=1}^{k(n)} (k(n)w_i)^2 < \infty$*, then* $\sum_{i=1}^{k(n)} w_i^2 < \infty$ *and*

$$Var[m_n(\underline{x})] = [Var[Y_i|\underline{X}_i = \underline{x}] + B^2] \sum_{i=1}^{k(n)} w_i^2.$$

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

As soon as $\sum_{i=1}^{k(n)} w_i^2 \sim \frac{\gamma^2}{k(n)}$, *and* $k(n) = [n^Q]$, *we get the result.*

iii) Finally, we assume that $w_i = \frac{w(\underline{x}-\underline{X}_{(i)})}{\sum_{i=1}^{K} w(\underline{x}-\underline{X}_{(i)})}$ *where* $w(.)$ *is a given function. In that latter case, the weights depend on the process* $(X_n)_n$. *In the following, we denote by* $N(i)$ *the order of the* i^{th} *neighbor. We rewrite the neighbor indices in an increasing order such that* $M(1) = min\{N(i), 1 \leq i \leq K\}$ *and* $M(k) = min\{N(i) \notin \{M(j), \forall j < k\}, 1 \leq i \leq K\}$ *for* $2 \leq k \leq K$ *and* $K = k(n)$ *is the number of neighbors. We introduce a real triangular sequence* $\{\alpha_{Ki}, 1 \leq i \leq K$ *and* $\alpha_{Ki} \neq 0 \ \forall i\}$ *such that*

$$\underset{K}{Sup} \sum_{i=1}^{K} \alpha_{Ki}^2 < \infty \quad and \quad \max_{1 \leq i \leq K} |\alpha_{Ki}| \underset{n \to \infty}{\longrightarrow} 0. \tag{5.31}$$

Now using the sequences $M(j), j = 1, \cdots, K$ *and* $(\alpha_{Ki}), 1 \leq i \leq K$, *we can rewrite (5.27) as:*

$$\frac{m_n(\underline{x}) - E m_n(\underline{x})}{\sigma_n} = \sum_{i=1}^{K} \alpha_{Ki} S_i, \tag{5.32}$$

with $S_i = \frac{w_{M(i)} X_{M(i)+1} - E w_{M(i)} X_{M(i)+1}}{\alpha_{Ki} \sigma_n}$. *The sequence* (S_i^2) *is uniformly integrable and* S_i *is function only of* $(X_j, j \leq M(i)+1)$, *thus if we denote* $\mathcal{F}_i$, $\mathcal{G}_i$, $\mathcal{F}_i^j$ *and* $\mathcal{G}_i^j$, *the sigma algebras generated by* $\{X_r\}_{r \leq i}$, $\{S_r\}_{r \leq i}$, $\{X_r\}_{r=i}^j$ *and* $\{S_r\}_{r=i}^j$ *respectively, then* $S_i \in \mathcal{F}_{M(i)+1}$, *and* $\mathcal{G}_i \subset \mathcal{F}_{M(i)+1}$. *For a given integer* ℓ, *we have also* $\mathcal{G}_{n+\ell}^{\infty} \subseteq \mathcal{F}_{n+M(\ell)+1}^{\infty}$ *since* $M(1) < M(1)+1 \leq M(2) < \cdots \leq M(n+\ell) < M(n+\ell)+1 \leq M(n+\ell+1)$. *Then:*

$$\underset{\ell}{sup} \underset{A \in \mathcal{G}_1^{\ell}, B \in \mathcal{G}_{n+\ell}^{\infty}, P(A) \neq 0}{Sup} |P(B \mid A) - P(B)| \leq \underset{\ell}{sup} \underset{A \in \mathcal{F}_1^{M(\ell)+1}, B \in \mathcal{F}_{n+M(\ell)+1}^{\infty}, P(A) \neq 0}{Sup} |P(B \mid A) - P(B)|. \tag{5.33}$$

Under the $\phi-$*mixing assumption on* $(X_n)_n$, *the right hand part of the expression (5.33) tends to zero as* $n \to \infty$ *and the lelf hand part of (5.33) converges to zero, hence the sequence* (S_i) *is* $\phi-$*mixing. Moreover:*

$$S_i \quad is \quad centered \quad and \quad var(\sum_{i=1}^{K} \alpha_{Ki} S_i) = var(\frac{m_n(\underline{x})}{\sigma_n}) = 1. \tag{5.34}$$

Then using expressions (5.31) - (5.34) and the theorem 2.2 in Peligrad and Utev (1997), we get:

$$\frac{m_n(\underline{x}) - E m_n(\underline{x})}{\sigma_n} \to_{\mathcal{D}} \mathcal{N}(0,1) \tag{5.35}$$

The variance of $m_n(x)$ *is given by the relation (5.25). The proof of the theorem 3.1 is complete.*

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

We provide now the proof of Corollary 3.1.

Proof 5.3 (Proof of corollary 2.1)**.** *From theorem 2.1, a confidence interval, for a given α can be computed, and has the expression:*

$$-z_{1-\frac{\alpha}{2}} \leq \frac{m_n(\underline{x}) - Em_n(\underline{x})}{\hat{\sigma}_n} \leq z_{1-\frac{\alpha}{2}} \tag{5.36}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ quantile of Student law. Previoulsy, we have established that the estimate $m_n(\underline{x})$ can be biased, thus the relationship (5.36) becomes:

$$m_n(\underline{x}) + B - \hat{\sigma}_n z_{1-\frac{\alpha}{2}} \leq m(\underline{x}) \leq m_n(\underline{x}) + B + \hat{\sigma}_n z_{1-\frac{\alpha}{2}} \tag{5.37}$$

When the bias is negligible, the corollary is established. If this bias is not negligible, we can bound it. The bound is obtained using (5.7) and (5.38):

$$B = O\left(\left(\frac{k(n)}{(n-d)\hat{h}(\underline{x})c}\right)^{\frac{p}{d}}\right) \tag{5.38}$$

with $c = \frac{\pi^{d/2}}{\Gamma((d+2)/2)}$, $\hat{h}(\underline{x})$ being an estimate of the density $h(\underline{x})$. Introducing this bound in (5.37) completes the proof.

6 ANNEX

6.1 Euro Area Monthly Indicators

We provide the list of the monthly economic indicators used in this study for the computation of the GDP using the bridge equations.

6.2 The bridge equation

We introduce now the bridge equations as they are listed in Diron (2008). Let us define Y_t as: $Y_t = (\log GDP_t - \log GDP_{t-1}) \times 100$, where GDP_t is the GDP at time t. The final GDP Y_t is the mean of the eight values computed below.

1. EQ1. $Y_t^1 = a_0^1 + a_1^1(\log I_t^1 - \log I_{t-1}^1) + a_2^1(\log I_t^2 - \log I_{t-1}^2) + a_3^1 I_{t-1}^3 + \varepsilon_t$.
2. EQ2.

$$Y_t^2 = a_0^2 + a_1^2(\log I_t^1 - \log I_{t-1}^1) + a_2^2(\log I_t^2 - \log I_{t-1}^2) + a_3^2(\log I_t^4 - \log I_{t-1}^4) + a_4^2(\log I_t^5 - \log I_{t-1}^5) + \varepsilon_t.$$

Documents de Travail du Centre d'Economie de la Sorbonne - 2009.50 (Version révisée)

Short Notation	Notation	Indicator Names	Sources	Period
I^1	IPI	Industrial Production Index	Eurostat	1990-2007
I^2	CTRP	Industrial Production Index in Construction	Eurostat	1990-2007
I^3	SER-CONF	Confidence Indicator in Services	European Commission	1995-2007
I^4	RS	Retail sales	Eurostat	1990-2007
I^5	CARS	New passenger registrations	Eurostat	1990-2007
I^6	MAN-CONF	Confidence Indicator in Industry	European Commission	1990-2007
I^7	ESI	European economic sentiment index	European Commission	1990-2007
I^8	CONS-CONF	Consumers Confidence Indicator	European Commission	1990-2007
I^9	RT-CONF	Confidence Indicator in retail trade	European Commission	1990-2007
I^{10}	EER	Effective exchange rate	Banque de France	1990-2007
I^{11}	PIR	Deflated EuroStock Index	Eurostat	1990-2007
I^{12}	OECD-CLI	OECD Composite Leading Indicator, trend restored	OECD	1990-2007
I^{13}	EUROCOIN	EuroCoin indicator	Bank of Italy	1999-2007

Table 2: Summary of the thirteen economic indicators of Euro area used in the eight GDP bridge equations.

3. EQ3. $Y_t^3 = a_0^3 + a_1^3 I_t^7 + a_2^3 I_{t-1}^7 + \varepsilon_t$.

4. EQ4. $Y_t^4 = a_0^4 + a_1^4 (I_t^6 - I_{t-1}^6) + a_2^4 I_t^3 + \varepsilon_t$.

5. EQ5. $Y_t^5 = a_0^5 + a_1^5 (I_t^6 - I_{t-1}^6) + a_2^5 I_t^9 + a_3^5 I_t^8 + \varepsilon_t$.

6. EQ6. $Y_t^6 = a_0^6 + a_1^6 (\log I_{t-2}^{10} - \log I_{t-3}^{10}) + a_2^6 (\log I_{t-1}^{11} - \log I_{t-2}^{11}) + \varepsilon_t$.

7. EQ7. $Y_t^7 = a_0^7 + a_1^7 (\log I_t^{12} - \log I_{t-1}^{12}) + a_2^7 (\log I_{t-2}^{12} - \log I_{t-3}^{12}) + a_3^7 Y_{t-1}^7 + \varepsilon_t$,and

8. EQ8. $Y_t^8 = a_0^8 + a_1^8 I_t^{13} + \varepsilon_t$.