Mixed-Frequency Modeling and Economic Forecasting

Clément Marsilli

Université de Franche-Comté
Laboratoire de Mathématiques de Besançon
Ecole doctorale Carnot-Pasteur

Mixed-frequency modeling and economic forecasting

Thèse de Doctorat en Mathématiques Appliquées

Clément Marsilli

sous la direction de Juan-Pablo Ortega

Thèse soutenue le 6 mai 2014, à Besançon,
devant un jury composé de

Luc Bauwens (rapporteur),
Professeur à l'Université Catholique de Louvain,

Frédérique Bec (rapporteur),
Professeur à l'Université de Cergy-Pontoise,

Yacouba Boubacar Maïnassara (examinateur),
Maître de Conférences à l'Université de Franche-Comté,

Laurent Ferrara (co-encadrant),
Chef du Service de Macroéconomie Internationale à la Banque de France,
Professeur associé à l'Université Paris Ouest-Nanterre La Défense,

Sébastien Laurent (président du jury),
Professeur à l'Université d'Aix-Marseille,

Juan-Pablo Ortega (directeur de thèse),
Chargé de Recherche CNRS.

Remerciements

Juan-Pablo, Laurent, je tiens tout d'abord à vous remercier. Vous m'avez véritablement guidé et soutenu durant ces trois années. J'ai grâce à vous, à votre disponibilité, et à vos conseils, pu mener cette thèse à bien. J'espère avoir été digne de la confiance que vous m'avez toujours témoignée, et que nous aurons encore, à l'avenir, l'occasion de travailler ensemble.

Je tiens également à remercier les membres du jury: Luc Bauwens et Frédérique Bec pour leurs commentaires et leurs conseils en tant que rapporteurs de cette thèse, Sébastien Laurent qui me fait l'honneur de présider ce comité, et Yacouba Boubacar Mainassara qui a également accepté de participer à cette soutenance.

La Banque de France m'a fourni un formidable environnement de travail, je tiens à remercier Olivier de Bandt, Olivier Vigna et Bertrand Pluyaud qui m'ont accueilli, il y a quatre ans de cela, au Service du Diagnostic Conjoncturel, à remercier Matthieu Bussière et Bruno Cabrillac qui m'ont permis d'effectuer mes travaux de thèse au sein de la formidable équipe du Service de Macroéconomie International, le célèbre SEMSI. Un grand merci à tous mes collègues, Sophie, Claude, Valérie, Daniele, Catherine, Julia, Guilia, Alessandro, Valérie, Simona, Murielle, Yannick, Éliane, Céline, Camille, Soledad, Nicolas et Pascal.

Je remercie également mes condisciples et amis, Gong, Keyou, Mohammed, Pauline,

Nicolas, Lilia, Dilyara, Majdi, Ludovic, Marie-Louise, Margarita, Henri, Thibaud, Alessandra, Miklos, Simon; Mathieu, Julie, Lyudmila, Dima, Arnaud, Thomas, Yiqiao, ainsi que mes collègues bisontins, matheux devant l'Éternel, Céline, Michel, Aude et Charlotte, Alexis, Ibrahim, Cyrille, Xiao, et les autres membres du labo.

Enfin je remercie ma Clémentine, ma Sœur, Maman, Papa, Maxime, ma famille creusoise, parisienne et morberande.

Contents

Introduction

Tout ce qui était n'est plus ; tout ce qui sera n'est pas encore.
Ne cherchez pas ailleurs le secret de nos maux.

Alfred de Musset,
La Confession d'un enfant du siècle.

De la modélisation multi-fréquentielle pour la prévision économique

La crise financière mondiale, la crise des dettes souveraines, les récessions qu'ont endurées et qu'endurent aujourd'hui encore, en ce début d'année 2014, nombre de pays parmi les plus riches, témoignent de la difficulté d'anticiper les fluctuations économiques, même à des horizons proches. La prévision économique à court terme à l'échelle macroéconomique dans un ensemble globalisé et interdépendant, est un exercice aussi complexe qu'essentiel pour la définition de la politique économique et

monétaire contemporaine. En effet, économistes, politiques, banquiers, journalistes, citoyens, employés et employeurs, consommateurs, producteurs et investisseurs, tous scrutent les conditions économiques actuelles, anticipées, espérées, prédites ou prévues, et adaptent en conséquence leurs comportements, politiques et décisions. Ainsi, la publication trimestrielle des chiffres du taux de croissance du Produit Intérieur Brut (PIB), qui représente l'évolution de l'ensemble de la valeur ajoutée qu'une économie produit durant une certaine période, tel que défini par la comptabilité nationale, attise les passions et anime les débats. Bien que le PIB fasse l'objet de critiques, il constitue aujourd'hui l'indicateur privilégié de la santé économique d'un pays et concentre à ce titre l'intérêt premier des économistes et prévisionnistes. Les recherches effectuées ces trois dernières années dans le cadre de la thèse de doctorat qui est présentée dans ce manuscrit se sont attachées à étudier, analyser et développer des modélisations de prévision. Nous identifions dans cette introduction les problématiques qui ont structuré nos travaux.

Notre travail de thèse, qui débuta en 2011, intervient à la suite de la crise financière 2008-2009. Cette période, dont il fut espéré qu'elle soit celle de la reprise économique après la récession, vit finalement poindre une nouvelle crise européenne. Le sauvetage de la Grèce, la création de différents mécanismes européens (FESF puis MES), les aides financières dispensées aux pays en difficultés en contrepartie de lourds plans de rigueur, le pacte de stabilité renforcé, l'usage de politiques monétaires non conventionnelles par la Banque Centrale Européenne : en Europe les évènements s'enchainèrent et la croissance, bien que redevenue positive durant quelques trimestres en 2009, resta faible et inégale jusqu'en 2014. Ces épisodes de récession mirent à mal les méthodes de modélisations classiques qui n'arrivaient pas à anticiper de manière précise les fluctuations proches des taux de croissance. Les modèles économétriques de prévisions reposent, en effet, généralement sur des régressions qui cherchent à expliquer et à prédire une série d'intérêt, dans notre cas la croissance économique, par un ensemble d'informations contemporaines et passées. Les données disponibles sur lesquelles baser une analyse prédictive n'ont jamais été aussi importantes mais toutes ne sont pas considérer de manière équivalente. Les statistiques de l'industrie, de l'emploi, les enquêtes d'opinion, les prix de matières premières, d'actions, d'obligations cotées en temps *quasi* continu, les indicateurs du marché immobilier, sont autant de variables explicatives et de prédicteurs potentiels de la croissance économique d'un pays. Discerner l'information du bruit revient parfois à séparer le bon grain de l'ivraie. On remarquera toutefois que les séries temporelles issues de l'économie réelle et financière ne présentent pas les mêmes caractéristiques, tant au niveau de leur fréquence d'échantillonnage que de leur apport prédictif. Se posent subséquemment des questions quant à l'utilisation de

ces données : quelle agrégation temporelle est la plus judicieuse ? Quels indicateurs sont à considérer ? Comment spécifier un modèle de prévision ? Quel horizon de prévision est le plus adéquat ? Et si d'aventure, la modélisation que nous proposions était empiriquement convenable, quelle interprétation économique donner à ses résultats ? Quelle confiance devons-nous attribuer à ses prévisions ? Nos travaux de thèse entendent apporter quelques éléments de réponse à ces questionnements.

Un ensemble d'informations hétérogène

L'ensemble des données économiques sur lequel baser un modèle de prévision est vaste. La collecte des statistiques mesurant l'évolution des différents marchés de l'économie est en effet institutionnalisée depuis longtemps dans de nombreux pays. Afin d'identifier les variables explicatives essentielles à toute modélisation prédictive, nous définirons tout d'abord trois grandes familles d'indicateurs économiques. (i) Tout d'abord, les variables "réelles" (appelées *hard data*): il s'agit d'indicateurs mensuels de l'activité économique tels que les statistiques d'emploi, données de consommation, indices de production, etc. Elles reflètent de manière effective l'activité mesurée de l'économie, et entrent, de manière indirecte, dans le calcul de la croissance. Ces variables sont donc par nature coïncidentes, en termes d'horizon optimal de prévision. (ii) Les séries dites *softs* : ces données sont issues d'enquêtes de conjoncture, ou d'opinion, réalisées mensuellement auprès des consommateurs ou des chefs d'entreprise. Celles-ci entendent mesurer les conditions économiques actuelles voire les anticipations des différents agents (propension à consommer, nouvelles commandes industrielles, etc.). Celles-ci ont généralement tendance à être avancées par rapport au cycle économique. (iii) Enfin, les variables financières, et plus généralement liées aux marchés financiers, parmi lesquelles sont concernées : indices boursiers, obligations, prix de matières premières, taux d'intérêt, taux de change, etc. Les variables financières sont généralement absentes des modélisations économétriques. En effet, celles-ci sont intrinsèquement volatiles, il est donc difficile de discerner leur véritable apport prédictif d'un point de vue macroéconomique. De plus, la fréquence temporelle des données financières est presque continue et donc difficile à appréhender dans le cadre d'une modélisation trimestrielle. Le lien entre sphère financière et économie réelle est toutefois patent, en particulier durant la période récente: les sévères récessions que de nombreux pays ont subies en 2009 sont en effet perçues comme les répercussions directes de la crise financière et bancaire mondiale de 2008. Certains travaux de la littérature suggèrent que ces séries financières peuvent être utiles lorsqu'elles sont correctement utilisées. La volatilité des variables financières est notamment une mesure évidente de l'incertitude

économique. La volatilité est une notion statistique bien précise, il s'agit d'une mesure de l'amplitude des variations des signaux financiers, celle-ci n'est malheureusement pas observable ou mesurable, elle est donc sujette à l'estimation d'un modèle. Nos travaux considèrent la volatilité financière comme un indicateur prédictif des fluctuations économiques futures.

Le choix des variables à privilégier dans le cadre de prévisions macroéconomiques est délicat et dépend fortement de l'horizon de prévision. Il existe différentes méthodes permettant de réduire l'ensemble d'information aux variables possédant une véritable significativité statistique. La sparsité, qui caractérise un ensemble peuplé majoritairement de zéros, revêt dans ce cadre une importance tout particulière. Lorsque l'un des éléments est nul, il n'est, par essence, pas statistiquement significatif. Un ensemble sparse réalise donc indirectement une sélection des variables les plus importantes. Cette technique est notamment envisagée dans le cadre de l'un de nos projets de recherche sur la sélection de modèle pour la prévision économique.

Une temporalité particulière

La croissance, s'entend le taux de croissance du PIB réel d'un pays, est généralement issue des comptes nationaux calculés par l'institut statistique du pays concerné et est traditionnellement dévoilée trimestriellement. En France, le chiffre rendu public par l'INSEE est un résultat comptable reposant sur les données de consommation, d'investissement, de variations de stocks, d'exportations et importations, et représentant la production de valeur ajoutée durant la période. Sa publication intervient de manière retardée par rapport au trimestre en question et fait, de plus, l'objet de révisions successives, ne délivrant un résultat définitif que plusieurs années plus tard. En France, le chiffre de croissance est connu environ un mois et demi après la fin du trimestre en question (*e.g.* mi-mai pour le 1er trimestre). Ce décalage dans le temps n'a, en particulier, permis d'identifier la récession française débutée en mars 2008, qu'à partir de novembre 2008. On mesure ainsi la nécessité d'anticiper précisément les fluctuations à un horizon très court. Il ne s'agit plus de prévoir la période future mais bien la période actuelle. En effet, les délais de parution et les révisions successives des séries économiques ont même contraint les prévisionnistes à envisager des analyses prospectives à des horizons intra-période, appelées *nowcasting* ; e.g. prévoir la croissance du premier trimestre au mois de janvier. De telles modélisations sont définies de manière à mobiliser l'information contemporaine disponible. Il convient de noter que les données, issues des trois familles que nous avons décrites précédemment, à partir desquelles nous

souhaitons construire une méthodologie prédictive sont certes nombreuses et probablement informatives mais présentent des fréquences d'échantillonnages bien différentes. Leur utilisation requiert donc l'élaboration de modélisations multi-fréquentielles adaptées. Ce schéma temporel particulier ne doit pour autant pas représenter un obstacle à la modélisation mais bien constituer l'une de ses caractéristiques fondamentales. Dompter cette temporalité constitue un véritable enjeu pour les économistes et une gageure de notre travail de recherche.

Les approches théoriques ou empiriques que nous présentons sont généralement basées sur des méthodologies de régressions linéaires classiques à ceci près que les variables incorporées sont de fréquences différentes, à savoir que notre série temporelle d'intérêt est observée à fréquence basse (trimestrielle) et que les variables explicatives sont échantillonnées à fréquence haute (journalière ou mensuelle). Nous raisonnons dans un premier temps de manière contemporaine, la notion de prévision n'intervient qu'une fois le modèle établi. La situation est la suivante : pour expliquer une donnée trimestrielle (par exemple celle du 1er trimestre disponible fin mars), on considère par exemple les 4 dernières données d'une variable réelle mensuelle (mars, février, janvier et décembre), et les 6 derniers mois de données d'une variable financière journalière (au moins $20 \times 6 = 120$ données journalières d'octobre à mars). L'idée première serait de pondérer chacune de ces valeurs par un coefficient que l'on estimerait, cette stratégie est inenvisageable pour un problème de grande dimension : l'utilisation des deux variables explicatives précédentes impliquerait l'estimation d'au moins 124 paramètres. Il s'agit d'un problème récurrent avec des échantillons finis. La modélisation que nous proposons cherche à concilier le mélange des fréquences d'échantillonnage et la parcimonie nécessaire à son estimation. Dans la droite ligne de la littérature des modèles à retards échelonnés, dont les travaux de Shirley Almon sont, selon nous, l'une des pierres angulaires, la méthodologie MIDAS (de l'anglais *Mixed Data Sampling*), développée et popularisée par Eric Ghysels et ses coauteurs, entend conjurer les divers problèmes de modélisation que nous avons énumérés. L'idée est simple et pour autant innovante: les poids présents devant chacune des données à haute fréquence, permettant l'agrégation temporelle, sont liés par une fonction. Il ne s'agit donc plus d'estimer chacun de ces poids mais les paramètres de cette fonction. Le problème d'optimisation sous-jacent voit dès lors sa dimension réduite. Une telle méthodologie s'adapte ainsi automatiquement aux données qu'elle entend modéliser et informe, par la forme de la fonction, de la capacité prédictive de celles-ci.

La thèse qui est présentée dans ce manuscrit entend étudier la modélisation temporelle pour la prévision macroéconomique. Nous évoquons à cet effet dans un premier chapitre les éléments fondamentaux de l'économétrie. Nous abordons en effet

la modélisation multi-fréquentielle telle qu'elle a été conçue et pensée, à la frontière entre modèles à retards échelonnés et agrégation temporelle. Le deuxième chapitre regroupe les résultats de deux travaux empiriques qui illustrent l'intérêt économique de telles modélisations. Le premier montre l'apport prédictif macroéconomique que constitue l'utilisation de la volatilité des variables financières en période de retournement conjoncturel. Le second traite de l'estimation en temps réel de la croissance mondiale annuelle. Le troisième chapitre s'étend ensuite sur l'analyse bayésienne de ces modèles à fréquences multiples. L'inférence bayésienne qui repose sur la déduction de la probabilité d'un évènement à partir des probabilités d'autres évènements déjà connues est en effet au coeur de nos recherches. Nous explorons notamment par ce biais de nouvelles méthodes d'estimation, de sélection de modèles, et nous présentons un travail empirique issu de l'adjonction d'une volatilité stochastique à notre modèle. Enfin, le quatrième chapitre apparait comme une conséquence et une suite logique des travaux menés jusqu'alors. Il s'agit d'une étude des techniques de sélection de variables à fréquences multiples dans l'optique d'améliorer la capacité prédictive de nos modélisations. Diverses méthodologies sont à cet égard développées, leurs aptitudes empiriques sont comparées, et certains faits stylisés sont esquissés. Cette dernière étude achève ainsi cette thèse, mais ouvre de nombreuses perspectives et présage de nouvelles recherches.

On mixed-frequency modeling and economic forecasting

Economic downturn and recession that many countries experienced in the wake of the global financial crisis demonstrate how important but difficult it is to forecast economic fluctuations, especially within a short time horizon. Economists, politicians, bankers, journalists, employees and employers, consumers, producers and investors, all tried hard to decipher economic conditions - both current and anticipated - in order to best adjust their economic behavior. Thus, the quarterly publication of the Gross Domestic Product (GDP), which represents the evolution of the aggregate values added within a given time period always attracts public attentions and incites debates. Notwithstanding subject to criticisms, the GDP constitutes a main indicator of economic health of a country and as such constitutes the primary interest of economists and forecasters. The doctoral dissertation conducted over the past three years studies, analyses and develops models for macroeconomic forecasting. This introduction aims at identifying the main issues that have motivated our research works.

Our doctoral research began in 2011 in the wake of the 2008-2009 financial crisis. By then, the world economy went through a "double dip" instead of heading towards a recovery path as many economists had previously forecast. Due to the unexpected shock of the European sovereign debt crisis, the economic outlook in Europe was especially unpredictable and has been weak even until today. A number of audacious actions have been, however, taken to put the European economies back on track: bailout of Greece, creation of the European crisis mechanism (EFSF then ESM), financial support provided to countries under financial strain with heavy austerity measures counterpart, reinforcement of the European financial stability pact, the use of unconventional monetary policies by the European Central Bank. Recent economic events have cast serious doublets on standard methods of economic prediction which failed to provide accurate economic snapshot and forecast for policymaking in time of crisis. In fact, econometric forecasting models are usually based on regressions that seek to explain and predict a variable of interest - the economic growth rate in our studies - through a range of contemporaneous and historical information. However, the volume of available data for economic forecasting is huge. An important achievement would be to determine the more relevant indicators from: industrial figures, employment statistics, opinion surveys, prices of commodities, stocks, bonds quoted in *quasi*-continuous time, indicators of real estate market, etc. These time series can potentially be explanatory variables to predict economic growth but they also can contain noise. To

only extract relevant information from these variables is as difficult as to separate the wheat from the chaff. It can also be noticed that time series coming from real and financial economy do not have the same characteristics, both in terms of sampling frequency and predictive power. To most efficiently these time series of different frequency for our forecasting models requires us to think about the following questions: what is the most adequate temporal aggregation strategy? Which indicators should be considered? How to specify a forecasting model? Which forecasting horizon to be consider? What is the economic interpretation of forecasting results? What confidence level should be used? Our thesis aims to provide some answers to these questions.

A set of heterogeneous information

The set of information coming from economic activity is vast and disparate. The collection of data measuring the performance in every market of the economy has indeed been institutionalized for many years in many countries. In order to identify variables which are essential to predictive modeling, we first define three major types of economic indicators. (i) *Real* data (also called *hard* data): monthly indicators of economic activity, such as employment figures, consumption data, production indices, etc.. They reflect realized economic activity and are used to calculate economic growth rate. These variables are coincident with the business cycle in terms of optimal forecast horizon. (ii) *Soft* data: These data usually from opinion surveys conducted in a monthly basis with consumers or producers. The soft data measure economic agents' perception of current economic conditions or their expectations (propensity to consume, new orders, etc.). These variables generally lead economic cycles. (iii) Financial data are more generally related to financial markets: stock indexes, bonds, commodity prices, interest rates, exchange rates, etc.. The financial data are generally less frequently used in macroeconomic forecasting models. Indeed, financial data are inherently volatile and therefore it is difficult to discern their real predictive input from a macroeconomic point of view. In addition, the temporal frequency of financial data is almost continuous and hence difficult to use in the context of a quarterly modeling. However, the macro-financial linkages are obvious, especially during the recent period: the severe recessions that many countries has experienced since 2009 are indeed seen as the direct result of the global financial and banking crisis of 2008. Moreover, the volatility of financial variables - magnitude of changes in financial signals - is often considered as a measure of economic uncertainty. Unfortunately, the volatility is not observable or measurable, it requires the specification of an adequate modeling. Our works consider financial volatility as a predictor of future economic fluctuations.

For macroeconomic forecasting, how to choose explanatory variables from the above mentioned three types of data, is tricky and heavily depends on the forecast horizon. There is a dilemma regarding this choice: on the one hand we want to include as much explanatory variables as possible to increase the predictive power of the model; on the other hand, facing the limited sample size, we need to care about model sparsity. There exist various methods to reduce the information set with respect to statistical significance. The sparsity, that characterizes a set populated primarily with zeros, is an important concept which indirectly carries a selection of the most important variables. In fact, when one element in the set is equal to zero, this means that it is not statistically significant. This technique is used in one of our research projects on model selection for economic forecasting.

A problem of temporality

The economic growth, *i.e.* the growth rate of real GDP of a country, is related to national accounts and hence usually computed in a quarterly basis by national statistical institutes. In France, economic growth released by INSEE is calculated using data on consumption, investment, changes in inventories, exports and imports. It represents the sum of the values added during a certain period. Thea data release of the GDP growth in a given quarter is lagged and subject to *ex-post* revisions until the final evaluation is provided several years later. In France, the first release of the GDP growth estimate occurs about a month and a half after the end of that quarter (*e.g.* mid-May for the first quarter). For instance, the recession started in March 2008 in France while the GDP data only confirmed that in November. This situation proves the necessity to precisely anticipate economic fluctuations even at very short horizons. It requires consideration of the current period as specific range of prediction. Indeed, due to the publication delay and the subsequent revisions, forecasters have developed intra-period forecasting techniques. This is called *nowcasting*, namely to predict the growth rate in first quarter in January. Such models are designed to take into account available contemporaneous information. However, notice that the real, soft and financial data presented above are usually sampled at various frequencies. To fully explore information contained in these data requires developing models compatible with mixed-frequency framework. Nevertheless, this specific temporal pattern must not be an obstacle to the modeling but rather constitutes one of its fundamental characteristics. Dealing with this temporality issue is a real challenge for economists and a motivation of our research.

Theoretical and empirical approaches that we propose are generally based on standard linear regressions methodologies using variables with different frequencies: while the target variable is observed at low frequency (usually quarterly), explanatory variables are sampled at higher frequencies (daily or monthly). Here is a concrete example of our general forecast methodology: to explain the first quarter GDP growth rate at the end of March, we shall consider any real variables available since December of last year (March, February, January and December) as well as financial variables available since last October(financial data are sampled at daily frequency thus have at least 20 x 6 = 120 data points). Parametrizing linear regression in such context would imply the use, and the estimation of at least 124 parameters, that can be a problem in finite samples. Therefore short-term forecasting models should both allow the use of mixed frequency data and parsimony. In line with the literature of distributed lag models of which Shirley Almon's work is, in my opinion, a cornerstone, the Mixed-Data Sampling (MIDAS) methodology has been developed and popularized by Eric Ghysels and his coauthors in the last decade to suit these purposes. In fact, the underlying idea is simple but innovative; weight coefficients that allow the temporal aggregation of high-frequency data rely on a function with a small number of parameters. Those are estimated as part of the whole optimization process. Such data-driven methodology exploits the predictive ability of our set of information to improve empirical results.

The PhD thesis dissertation aims at investigating mixed-frequency modeling for macroeconomic forecasting. The first chapter is dedicated to time series econometrics within a mixed-frequency framework. In particular, we examine distributed lag models and temporal aggregation schemes and introduce the MIDAS methodology. The second chapter contains two empirical works. The first study sheds light on macro-financial linkages by assessing the leading role of the daily financial volatility in macroeconomic prediction during the Great Recession. The second proposes a real-time monthly indicator of global economic outlook using nowcasting methodology. Both studies illustrate the macroeconomic forecasting power of mixed-frequency models. The third chapter extends mixed-frequency model into a Bayesian framework. Indeed, the Bayesian inference, which is basically based on the deduction of the probability of an event using the probabilities of other events, is of a particular interest for our research. We explore model selection through Bayesian methods of estimation and present an empirical study using a stochastic volatility augmented MIDAS model. The fourth chapter focuses on variable selection techniques in mixed-frequency models for short-term forecasting. We address the selection issue by developing MIDAS-based dimension reduction techniques and introducing two novel approaches using either a method of penalized variable selection or a Bayesian stochastic search variable selection. These

features integrate a cross-validation procedure that allows automatic in-sample selection based on recent forecasting performances. Our model succeeds in constructing an objective variable selection with broad applicability.

Chapter 1

Economic modeling with mixed-frequency data

The *Great Recession* has led many governments in developed economies to strongly adapt their monetary and fiscal policies, especially by using unconventional policy tools. Those events have therefore heightened the interest of economists in understanding and above all in preventing downturns and recessions. Practitioners and forecasters have especially emphasized the necessary re-assessment of the usual econometric models and hence their ability to really anticipate the business cycle. Ng and Wright (2013) particularly emphasized that "*this Great Recession is important not only because of its impact on the economic well-being of consumers and firms, but also because it once again led econometricians and macroeconomists to question the adequacy of their analysis.* [...] *It has involved a full-blown financial crisis that brought the role of financial markets back to center-stage of business cycle analysis*". In this thesis we propose to study the econometrics of short term macroeconomic forecasting and more specifically focus on mixed-frequency modeling involving financial data. In

this respect, we will introduce the main concepts, consider the literature, develop some new tools and discuss their efficiency with several empirical exercises. A key element in macroeconomics modeling is the time structure that makes the whole economy consistent. For instance, we can frequently observe a time lag between a policy decision or an economic event and their relative effects in the national accounts main aggregates data. That can be due to different causes: the structure of the economy, the international outlook, or perhaps just the availability of data. In this context, taming this specific temporal scheme is the essence of any macroeconomic forecasting model.

In order to track the variable of interest $\{y_t\}_t$, we base our regression model on one or several variables supposed to be informative and relevant, that we denote $\{x_\tau\}_\tau$. An increase of the industrial production ($x_\tau = IP_\tau$) inevitably leads to an expansion of the Gross Domestic Product ($y_t = GDP_t$). Yet, the effect of such increase may be either lagged, gradually diffused or inexistent depending on other factors (economic structure). Moreover, the facts that various economic time series are sampled at different frequencies prevents an efficient use of available data. We propose in this chapter to discuss those issues by reviewing the econometrics of mixed-frequency data.

1.1 Distributed lag models

In this section we discuss the econometrics of distributed lag models. We first consider that $\{y_t\}_t$ and $\{x_\tau\}_\tau$ are two stationary time series processes observed at different frequencies. The gap between those two frequencies is defined using the coefficient κ such as $t = \tau\kappa$; thus, $\kappa<1$, x_τ is observed at an higher frequency than $\{y_t\}_t$. Furthermore we define $x_t^\kappa := x_{\tau\kappa}$. For instance, if we consider the quarterly GDP growth as the dependent variable $\{y_t\}_t$ and one monthly explanatory variable $\{x_\tau\}_\tau$, we get a frequency gap coefficient κ equal to $1/3$.

For all t, we assume that y_t is not only determined by x_t^κ but also by a weighted sum of the K past values of the explanatory variable: $\{x_{t-(K-1)}^\kappa, \dots, x_{t-1}^\kappa, x_t^\kappa\}$. In such case, we obtain a regression model that involves *distributed lags* which is of the form

$$y_t = \sum_{k=0}^{K-1} w_k \, x_{t-k}^\kappa + e_t, \tag{1.1}$$

where we assume that every element of x_t is independent of the stochastic process e_t.

Considering a finite number of lagged x_t requires the choice of a lag length K and sets $w_k = 0$ when $k > K$. The length of the lag may be known, defined via a cross validation procedure, or assumed to be infinite. Notice that when K is too large, least squares do not allow estimation of the equation (1.1). An infinite number of parameters in the model would require further assumptions to allow a tractable specification.

In contrast, the fact that K is greater than one relies on a more general and underlying hypothesis. Indeed, it is assumed that the impact of the dependent variable to the independent variable does not only affect a single point in time but it may be distributed over a certain number K of future points in time. Although this *lagged effects* assumption seems reasonable since the equation mixes sampling frequencies, it involves strong economic postulates on the structure of the lag effect. In macroeconomics, its shape can particularly have a direct interpretation: it can represent an institutional constraint, a habit persistence or an expectational effect and hence may be considered as a real stylized effect. Finally, it can be noticed that the frequency of the explanatory variable does not really matter in this regression equation (1.1). That can be either greater, lower or equal to the standard imposed by the variable of interest y_t. In this respect, the use of κ as the fraction of time associated to the possible temporal gap is more informative than mathematically necessary. We will focus on temporal aggregation methods and their econometric consequences later on in this chapter.

Shirley Almon reviewed in her thesis defended at Harvard in 1964 the main contributions of this field, especially Koyck (1954), Eisner (1960) and Solow (1960), and developed new techniques in econometrics of distributed lag. Almon (1965) published a part of her PhD thesis in *Econometrica* in which she proposed a polynomial distributed lag structure that would become the main reference in distributed lags:

$$w_k = \sum_{j=0}^{p} \theta_j k^j, \quad \text{where } k = 0, \dots, K-1, \tag{Almon}$$

As long as the polynomial order p is substantially lower than K, this method constitutes a parsimonious approximate fit to least squares estimates, whose estimation is simple. The polynomial Almon allows the use of an intercept coefficient θ_0. Assuming that the e_t is a normally distributed white noise process $\mathcal{WN}(0, \sigma^2)$, we have

$Y = X\beta + e = XQ\,\theta + e$ where

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}, \quad X = \begin{pmatrix} x_K & x_{K-1} & \cdots & x_1 \\ \vdots & \vdots & & \vdots \\ x_T & x_{T-1} & \cdots & x_{T-K+1} \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix},$$

$$\text{and } Q = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 2 & 2^2 & \cdots & 2^p \\ \vdots & \vdots & \ddots & \vdots \\ K & K^2 & \cdots & K^p \end{pmatrix}. \tag{1.2}$$

By applying least squares on the Almon polynomial as defined by (1.2), we obtain the estimator $\hat{\theta} = ((XQ)'(XQ))^{-1}(XQ)'Y$, and as a consequence the estimator $\hat{\beta} = Q\hat{\theta}$ which follows the normal distribution $\mathcal{N}\left(\beta, \sigma^2 Q\left((XQ)'(XQ)\right)^{-1}Q'\right)$. It has been emphasized by Shiller (1973) that the Almon distributed lag technique relies "*on a polynomial of known degree, over a known interval, but with unknown coefficients*". Indeed, this method could be plagued by model misspecifications regarding its polynomial degree (usually small) and its projection space (usually $\mathbb{R}^+$). Yet, Teräsvirta (1980) showed that those exogenous prior specifications can sometimes improve estimation accuracy, specially when facing small samples, large model error variance, and smooth lag function. That is rather common when dealing with macroeconomic issues.

Many papers have later on studied statistical specifications of the Almon lag technique, we refer among others to Schmidt and Waud (1973) and to Schmidt and Sickles (1975). Shiller (1973) proposed a Bayesian way to introduce a finite distributed lag model with stochastic coefficients. He suggested to use priors on linear combinations of parameters to overlook erratic shapes like tidy smooth curves produced by polynomial lag models, especially the Almon form. Dhrymes (1971) and Sims (1974) provided surveys on distributed lag models. The book of Judge et al. (1985) also represented an important contribution to the literature (chapters 9 and 10 are dedicated to finite and infinite distributed lags). Recently, Ghysels et al. (2002) have conceived up-to-date distributed lag models by developing the MIDAS standing for MIxed DAta Samping which can be viewed as a generalization of the Almon technique. Those will be introduced in Section 1.3 and studied in detail in this thesis.

1.2 Temporal aggregation

Difference in sampling frequency emerges as a recurrent problem in the context of short term forecasting. Since we focus on predicting economic growth which is usually quarterly (or yearly), we obviously base the analysis on pertinent variables regardless of their sampling frequencies. Most series about real economic activity, as the production indices, opinion surveys, data on new orders, job statistics and price indices are monthly sampled. The situation is displayed in the Figure 1.1.

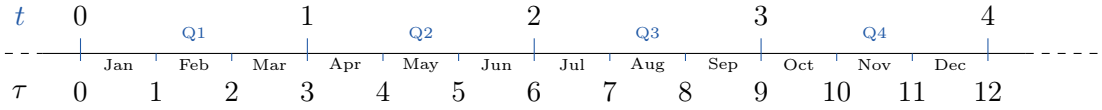


Figure 1.1: Temporal scheme

We define t as the time index for the lowest sampling frequency (quarterly for example) and τ as the highest frequency (monthly in this setup 1.1). Thus, the issue of the underlying temporal aggregation of the distributed lag matter arises. In fact when dealing simultaneously with monthly and quarterly variables, a standard way to proceed is to temporally aggregate the high frequency variable in order to assess dependence between variables sampled at the same frequency. This strategy transforms a high frequency series $\{x_\tau\}_\tau$ into a low frequency process $\{\mathsf{x}_t\}_t$ using an adapted aggregation method. In order to comply with macroeconomic conditions, we assume that data are available at the earliest at the end of the month depending on the publication schedule.

1.2.1 Aggregation schemes and bridge equation

The standard aggregation methods are stock or flow depending on the nature of the variable. From those two aggregating schemes, we derive five cases:

(i) Stock aggregation is defined as

$$\mathsf{x}_t = \{\mathsf{x}_t\}_t = \{x_{\tau\kappa}\}_\tau, \tag{1.3}$$

In the case described by Figure (1.1) in which we aggregate monthly data to obtain a quarterly series, κ is equal to $1/3$. Stock aggregation retained only one

third of the elements in x_τ. This technique particularly suits data series in level, for instance expressed in dollars (GDP, money stock, etc.), in terms of people (population, census, etc.) or in goods.

(ii) Flow aggregation is of the form

$$\mathsf{x}_t = \{\mathsf{x}_t\}_t = \left\{ \sum_{i=0}^{\kappa-1} x_{\tau\kappa+i} \right\}_\tau , \tag{1.4}$$

When the elements of x are flow variables, the aggregated process is made of (partial) sums of the disaggregated ones. Flow magnitudes include spending, saving, or income. Stock and Flow aggregation techniques are the two main ways of thinking about temporal aggregation. The next two methods are directly derived from flow aggregation.

(iii) Averaging relies on a particular flow scheme defined as

$$\mathsf{x}_t = \{\mathsf{x}_t\}_t = \left\{ \sum_{i=0}^{\kappa-1} \frac{1}{\kappa} x_{\tau\kappa+i} \right\}_\tau , \tag{1.5}$$

This technique corresponds to an average sampling. This scheme is for instance adapted to rate and index series (like opinion surveys on household spending or business climate).

(iv) Weighted averaging is straightforwardly defined as

$$\mathsf{x}_t = \{\mathsf{x}_t\}_t = \left\{ \sum_{i=0}^{\kappa-1} \frac{w_i}{\kappa} x_{\tau\kappa+i} \right\}_\tau , \tag{1.6}$$

where $w_1 + \ldots + w_\kappa = 1$. This strategy can be assimilated to a distributed lag method as defined in (1.1) with $K = \kappa$.

(v) Mariano and Murasawa (2003) argued that, in the case of the temporal aggregation of the quarterly GDP, this time series is a geometric mean of latent monthly random sequence. They therefore proposed a specific weighting scheme based on

three-period differences as follows

$$\begin{aligned}\mathsf{x}_t &= \frac{1}{3}\left(x_t + x_{t-1/\kappa} + x_{t-2/\kappa}\right) + \frac{1}{3}\left(x_{t-1/\kappa} + x_{t-2/\kappa} + x_{t-3/\kappa}\right) \\ &\quad + \frac{1}{3}\left(x_{t-2/\kappa} + x_{t-3/\kappa} + x_{t-4/\kappa}\right) \\ &= \frac{1}{3}x_t + \frac{2}{3}x_{t-1/\kappa} + x_{t-2/\kappa} + \frac{2}{3}x_{t-3/\kappa} + \frac{1}{3}x_{t-4/\kappa}\end{aligned}$$

This strategy relies on some strong economic assumptions and provide a specific aggregation scheme that should be used carefully.

Then, in a second step, the aggregated time series are simply incorporated into a regression model (or even a distributed lag model). This *bridge* equation links the low-frequency variable y_t and the time-aggregated variable x_t as follows:

$$y_t = \beta_0 + \beta_1 \mathsf{x}_t + e_t \tag{1.7}$$

Yet, in the context of macroeconomic forecasting, last values of the time-aggregated process may be unobserved. Short-term forecasting method usually involves all the available information and thus faces "*the real-time data flow*" described by Banbura et al. (2012). In fact, the time of the last available observation can differ from series to series. This important feature of real time analysis is due to publication delays. That is referred to as the *ragged-edge* of the information set. As an example, in France, the first quarterly GDP estimate is released approximately 45 days after the end of quarter. This situation is roughly the same in others countries ($\sim$ 30 days for US and UK, $\sim$ 45 days for Germany, $\sim$ 50 days for Japan). Moreover, this permanent lag emphasizes the importance of short term forecasting especially in terms of policy implication. That *vicious circle* situation involves predictive models which can deal with this feature. We will see that there exist different ways to make forecasts in this context.

1.2.2 Direct vs. iterated approach

We distinguish two predicting approaches: the direct forecasting method which consists in a horizon-set estimated regression and the iterated multistep method, which iterates forward to the horizon a one-period ahead forecasts. Those two schools of though are opposed. While iterated forecasts are theoretically more efficient if the one-period ahead model is well specified, direct forecasts are more robust to model

misspecification. In this respect, Chevillon (2007) proposes an overview of the whole literature on this topic; another interesting work is the paper of Marcellino et al. (2006) that focuses on the empirical comparison of both methods for macroeconomic forecasting purposes.

We define h the forecasting horizon in the lower frequency time units (the same as t) and denote the h ahead prediction by $t+h|t$. As proved in Hamilton (1994), the forecast of y_t at $t+h$ that minimizes the Mean Square Error is the expectation of y_{t+h} conditional on $\mathcal{F}_t$:

$$\hat{y}_{t+h|t} = \mathrm{E}[y_{t+h}|\mathcal{F}_t] \tag{1.8}$$

where the σ-field $\mathcal{F}_t$ represents all the information available at time t.

The relationship of interest at the h^{th} horizon described in (1.8) is the Direct multi step method. It involves a direct estimate of the horizon specific model. By contrast, the iterated procedure relies on multiperiod ahead one-step estimation defined as follows:

$$\begin{aligned}
\hat{y}_{t+1|t} &= \mathrm{E}[y_{t+1}|\mathcal{F}_t] \\
\hat{y}_{t+2|t} &= \mathrm{E}[y_{t+2}|\mathcal{F}_{t+1}] \\
&\vdots \\
\hat{y}_{t+h|t} &= \mathrm{E}[y_{t+h}|\mathcal{F}_{t+h-1}]
\end{aligned} \tag{1.9}$$

Despite the fact that those iterated single-period models can not be too badly misspecified because of the proximity to the forecasting target, this strategy gradually leads to an increase in the variance compared to the direct multi-step estimator of the forecast. Those methods have each their advantages and disadvantages depending on the horizon and the stochastic properties of the data. In our works, we will consider the direct multistep procedure, the iterated method will be explored in future research.

1.2.3 Forecasting with model-based aggregated data

In a multifrequency setup, the iterated multistep forecasting method is a two steps procedure that requires first or model-based temporal aggregated series to be plugged then into a bridge equation to obtain the forecasts. In fact, the explanatory variables are first expanded using generally forecasting equation associated to linear model (AR, ARMA or even vector autoregressive models in the multivariate case) to fill the missing

(unobserved) values of the period. The temporal aggregation of ARMA processes has been widely studied in the literature forecasting literature, we refer to Abraham (1982) and Lütkepohl's work (see Lütkepohl, 2007). Formally, assuming that the process $\{x_\tau\}_\tau$ of high frequency variables has a VAR(r) structure such that $x_\tau = \nu + A_1 x_{\tau-1} + \cdots + A_r x_{\tau-r} + \eta_\tau$, we forecast $\hat{x}_t = \hat{x}_{\tau/\kappa}$ using the direct procedure (1.8) such as follows

$$\hat{x}_t = \mathrm{E}[x_t|\mathcal{F}_{t-1}] = \mathrm{E}[x_t|\{x_s|s < t\}]. \tag{1.10}$$

Then, data are aggregated to the lowest frequency, in order to obtain conditions to the linear regression, namely the bridge equation, from which one gets forecasts. Thus, assuming that the aggregation is just a flow aggregation, we have $\mathsf{x}_t = \frac{1}{3}x_{t-2} + \frac{1}{3}x_{t-1} + \frac{1}{3}\hat{x}_t$. Then, using the direct forecast specification within bridge equation (1.7), that forwardly gives $\hat{y}_{t+h|t} = \hat{\beta}_0 + \hat{\beta}_1 \mathsf{x}_t$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are OLS estimates.

Nevertheless, Marcellino (1999) shown that the use of temporally aggregated data can lead to a loss of economic properties, such as Granger-causality, structural invariance or cointegration. Grigoryeva and Ortega (2012) have proposed a large review of finite sample forecasting methods and have developed a new "hybrid" scheme for the forecasting of temporal aggregates coming from ARMA processes. It consists in using high frequency data for estimating the model and in making prediction based on the corresponding aggregated model and data. The model parameters are therefore estimated using all the information available with the bigger sample size provided by the disaggregated data. In many cases, aggregation based approaches lead unfortunately to a loss of information due to either exogenous specifications far from data reality or errors accumulation (from the multistep procedure).

1.3 MIDAS regression models

Gathering both distributed lags and temporal aggregation literature and introducing some new generalization features, Ghysels and his coauthors have developed the MIDAS (Mixed Data Sampling) regression model. The MIDAS aims at accommodating the temporal aggregation by using a specific class of time series models that involves parsimony and flexibility. Derived from the distributed lag models technique, this novel econometric tool is based on both a regression structure and a weight function which tracks the high frequency lags of the explanatory variables.

In the same context as the equation (1.1), the MIDAS aims at explaining y_t using the lags of the explanatory variable x_t^κ sampled at the frequency $t\kappa$; it can be written as follows:

$$y_t = \beta_0 + \beta_1 \, m_K(\theta, L) \, x_t^\kappa + \varepsilon_t. \qquad \text{(MIDAS)}$$

We notice that the MIDAS combines usual linear regression features with an aggregation structure defined by the function m_K. Ghysels et al. (2002) and Ghysels et al. (2007) showed that both the intercept β_0 and the variable coefficient β_1 can easily hold some helpful empirical interpretations. The idiosyncratic term ε_t stands for the residuals. We will now focus on the MIDAS kernel function m_K, on its specifications and its estimation.

1.3.1 Almon function and weighting scheme

The kernel function m_K is specify with respect to a parameter θ and to the past values of x_t^κ. We define the lag operator as $L^k x_t^\kappa = L^k x_{\tau\kappa} = x_{t/\kappa-k}$. The number of lags K is exogenous; as we already discussed in the previous sections, the choice of K may be statistically tested or empirically assessed. The parameters family $\{\beta_0, \beta_1, \theta\}$ is estimated (estimation techniques will be discussed later on). However, it can be noted that the presence of the β_1 coefficient implies that the function m_K provides normalized weights for the K past values of x_t. We define:

$$m_K(\theta, L) := \sum_{k=0}^{K-1} \frac{\varphi(k, \theta)}{\sum_{l=0}^{K-1} \varphi(l, \theta)} L^k. \qquad \text{(Weigth function)}$$

This expression of the Weigth function is the common form of the MIDAS as it has been popularized over the last decade. Many parametrizations of this weight function have been proposed depending on the number of coefficients or the shape of the function φ. Models for mixed-frequency data has been recently reviewed by Foroni and Marcellino (2013b). One can notice that the Almon expression, that combines equations (1.1) with the Almon form, is a specific case of the MIDAS that can be written as:

$$\beta_1 \, m_K(\theta, L) = \sum_{k=0}^{K-1} \left(\sum_j \theta_k k^j \right) L^k \tag{1.11}$$

The recent MIDAS literature initiated by Ghysels et al. (2002) has preferred non-linear expressions for the weight function including mainly two forms: the Beta lag and the exponential Almon lag function. Those are defined below:

– The normalized Beta probability density function is defined as follows:

$$\varphi(k, \theta) = \varphi_K(k, \theta_1, \theta_2) = \frac{\frac{k}{K}^{\theta_1 - 1} \left(1 - \frac{k}{K}\right)^{\theta_2 - 1} \Gamma(\theta_1 + \theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)} \tag{1.12}$$

where $\Gamma(\theta) = \int_0^\infty \mathrm{e}^{-x} x^{\theta-1} \mathrm{d}x$. The size of the polynomial p is defined with respect to both regression performances and parsimony. Note that the Beta lag form allows interesting features depending on its specification. For instance, restricting the argument size of the function (1.12) to a unique parameter θ_1 amounts to imposing decreasing weights values. This weighting scheme which incorporates only one hyper parameter θ_1 is of the form

$$\varphi(k, \theta) = \varphi_K(k, \theta_1) = \theta_1 (1 - k)^{\theta_1 - 1} \tag{1.13}$$

In terms of economic interpretation, downwards sloping property can represent a desirable feature especially in a direct multistep forecasting setup.

– Another popular expression of the MIDAS weight function is the exponential Almon lag form, which can be written as:

$$\varphi(k, \theta) = \varphi(k, \theta_1, \ldots, \theta_p) = \exp\left(\sum_{j=1}^{p} \theta_j k^j \right) \tag{1.14}$$

That formula is derived from the Almon function in a straight forward way. Using exponential function forces the weights to be positive (see Judge et al., 1985). The exponential Almon function is specified in the literature with two parameters ($p = 2$ in equation (1.14)): $\varphi(k, \theta_1, \theta_2) = \exp\left(\theta_1 k + \theta_2 k^2\right)$.

Those two forms provide a flexible and parsimonious data-driven weights scheme that involves a small set of parameters and hence is fully adapted to small samples. Figure 1.2 exhibits different shapes of the exponential Almon lag weight function with respect of the choice of its two parameters $\theta = (\theta_1, \theta_2)$.

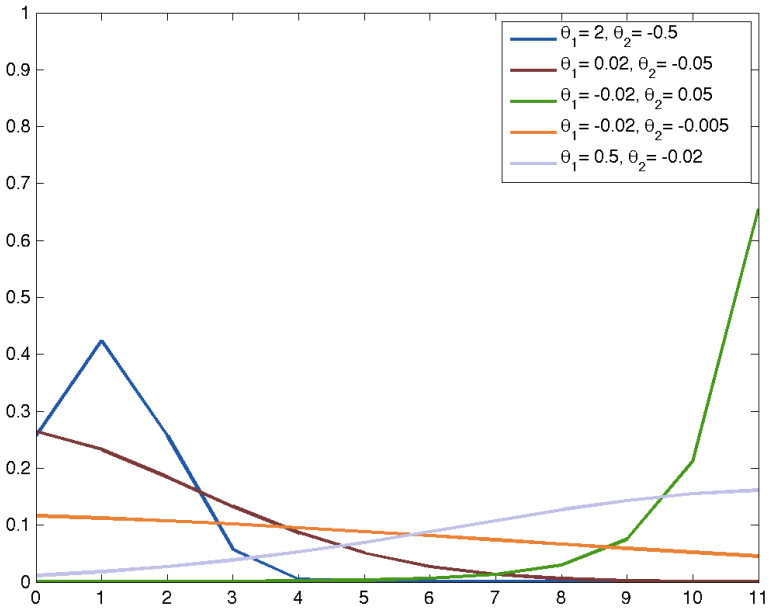


FIGURE 1.2: Exponential Almon lag weighting structure of the MIDAS with K=12

Kvedaras and Zemlys (2012) have proposed a test for the evaluation of statistical acceptability of a functional constraint which is imposed on parameters in the MIDAS regression. Andreou et al. (2010) have also put forward a test to examine if equal weights in aggregating time series are suitable in a mixed-frequency regression model.

1.3.2 THE MIDAS NLS ESTIMATOR

We assume that the disturbance term ε_t is normally distributed with density given by

$$F(\varepsilon_t) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma^2}\right) \tag{1.15}$$

Henceforth, we denote by ϕ the family of the unknown parameters, i.e. $\phi = \{\beta_0, \beta_1, \theta, \sigma\}$ and we define $\mathcal{X}_t(\phi) = \mathcal{X}(\phi, x_t) = \beta_0 + \beta_1 \, m_K(\theta, L) \, x_t^{\kappa}$. Assuming that the sample size

is T, for all $t = 0, \ldots, T$, the conditional probability distribution of y_t is therefore given by:

$$\begin{aligned} F(y_t|x_t;\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{t=1}^{T} \frac{y_t - (\beta_0 + \beta_1\, m_K(\theta, L)\, x_t)}{2\sigma^2}\right), \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{t=1}^{T} \frac{y_t - \mathcal{X}_t(\phi)}{2\sigma^2}\right) \end{aligned} \tag{1.16}$$

The log-likelihood function can be written as:

$$\ln F(Y|\phi) = \sum_{t=1}^{T} \ln F(y_t|x_t;\phi) = \frac{1}{2}\ln 2\pi - \frac{T}{2}\ln\sigma^2 - \frac{T}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - \mathcal{X}_t(\phi)\right)^2, \tag{1.17}$$

which is maximized with respect to ϕ. However, in the context of the non linear regression model, it can be noticed that the min/max-imization problem (like the Newton-Raphson algorithm) are simplified by expressing $\hat{\sigma}^2$ as a function of $\hat{\beta}$ and $\hat{\theta}$. That is achieved by solving the first order condition for $\hat{\sigma}^2$ which has the solution:

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \mathcal{X}_t(\hat{\phi})\right)^2. \tag{1.18}$$

Thus maximising the log-likelihood function leads to redefining the unknown parameter vector as $\phi = \{\beta_0, \beta_1, \theta\} \equiv \{\beta, \theta\}$.

This likelihood is maximized when the sum of squared residuals $S(\phi) = \left(y_t - \mathcal{X}_t(\phi)\right)^2$ is minimized:

$$\hat{\phi} = \arg\min_{\phi} S(\phi) \tag{1.19}$$

Then, differentiating $S(\phi)$,

$$\frac{\partial S(\phi)}{\partial \phi} = \frac{\partial \left(y_t - \mathcal{X}_t(\phi)\right)^2}{\partial \phi} = -2\,\left(y_t - \mathcal{X}_t(\phi)\right)^2\,\frac{\partial\left(y_t - \mathcal{X}_t(\phi)\right)}{\partial \phi} \tag{1.20}$$

Setting the partial derivatives equal to 0 produces equations that determine the regression coefficients. There is no closed-form solution to the nonlinear least squares problem. We use instead numerical algorithms to find parameters value which minimize 1.19[1.1]. Nevertheless the nonlinear least square estimator has some asymptotic

[1.1] Gradient formula in the case of the exponential Almon lag function is provided in Appendix A.

properties. Assuming that the gradient of $\nabla\mathcal{X}_t(\phi) = \left[\frac{\partial \mathcal{X}_t(\phi)}{\partial \phi}\right]$ exists, the MIDAS nonlinear least square estimator $\hat{\phi}$ is asymptotically[1.2] normal:

$$\sqrt{T}(\hat{\phi} - \phi) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \, \mathrm{E}[\nabla\mathcal{X}_t(\phi)\, \nabla\mathcal{X}_t(\phi)']^{-1}\right) \tag{1.21}$$

This result has been rigorously proved by Jennrich (1969)[1.3]. We refer to Judge et al. (1985), Chapter 6, for further details in nonlinear statistical models. In fact, MIDAS regression models are usually estimated using standard iterative optimization. The nonlinear specifications φ require numerical optimization to determine solutions (*e.g.* Levenberg-Marquardt algorithm or any other gradient descent method).

It has been proved that misspecifications due to a flat temporal aggregation lead to an inconsistency of the estimator. Andreou et al. (2010) have studied the asymptotic properties of the MIDAS nonlinear least squares estimator. They proposed to decompose the conditional mean of the MIDAS regression to assess the consequences of flat temporal aggregation. Following their prescriptions, we derive from the MIDAS equation a sum of two terms: an aggregated term based on flat weights and a non linear term which higher order differences of the high frequency process:

$$y_t = \beta \, \mathsf{x}_t^l + \beta \, \mathsf{x}_t^{nl}(\theta) + u_t \tag{1.22}$$

where the first term is the averaging aggregation such as defined in (1.5): $\mathsf{x}_t^l = \sum_{k=0}^{K-1} \frac{1}{K} x_{t-k}$. The second one x_t^{nl} is defined as the difference between the MIDAS structure weights and the flat weights, and hence it depends on the hyperparameter θ. It is of the following form:

$$\mathsf{x}_t^{nl}(\theta) = m_K(\theta, L)\, x_t - \mathsf{x}_t^l = \sum_{k=0}^{K-1} \left(\frac{\varphi(k,\theta)}{\sum_{l=0}^{K-1} \varphi(l,\theta)} - \frac{1}{K} \right) x_{t-k}. \tag{1.23}$$

The nonlinearity of the $\mathsf{x}_t^{nl}(\theta)$ term is due to the nonlinear weighting scheme of MIDAS regression model according to the form of the function φ. Furthermore, using this simple framework, Andreou et al. (2010) exhibit that the traditional least squares estimator that involves regression models with a flat aggregation scheme is inconsistent. They showed that its asymptotic bias depends on the matrix of coefficients of the regression (1.23) of the nonlinear ("omitted") term on the linearly aggregated co-

[1.2] Convergence in distribution.

[1.3] Three assumptions are necessary to prove the asymptotic normality of the nonlinear least squared estimator: (i) the parameter space is compact (closed and bounded) and ϕ is its interior point, (ii) the function $S(\phi)$ is continuous in ϕ, then last (iii) $\operatorname{p\,lim} T^{-1} S(\phi)$ exists, is non-stochastic, and its convergence is uniform in ϕ.

variates x_t^l. In their paper, they also provide some theoretical results on asymptotic and finite sample properties in difference cases of regression models. They especially focus on the slope coefficient, denoted by β_1 in our setting. They especially derived the formula (1.21) to give in general the asymptotic variance of the NLS estimator of β_1 (the slope coefficient):

$$A\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2 \, \mathrm{E}[\mathsf{x}_\theta^2] \, \mathrm{Var}[\mathsf{x}_\theta]}{\mathrm{Var}[\mathsf{x}_\theta] \, (\mathrm{E}[\mathsf{x}_\theta^2]\mathrm{E}[\mathsf{x}(\theta)]^2 - \mathrm{E}[\mathsf{x}(\theta)\mathsf{x}_\theta]^2) - (\mathrm{E}[\mathsf{x}_\theta^2]\mathrm{E}[\mathsf{x}(\theta)] - \mathrm{E}[\mathsf{x}(\theta)\mathsf{x}_\theta]\mathrm{E}[\mathsf{x}_\theta])^2} \tag{1.24}$$

where $\mathsf{x}(\theta) = m_K(\theta, L)\, x_t^\kappa$ to simplify notation and $\mathsf{x}_\theta = \nabla \mathsf{x}_t^{nl}(\theta)$.

We will assess those results using some Monte Carlo simulations on a Data Generating Process (DGP) and real economic data in the next two subsections.

Monte Carlo simulations

We now examine the properties of the MIDAS NLS estimator in the regression model using a Monte Carlo analysis. Our simulation design uses the following Data Generating Process (DGP) of a MIDAS regression model:

$$y_t = \beta_0 + \beta_1 \, m_K(\theta_1, \theta_2, L)\, x_t^\kappa + \varepsilon_t, \tag{1.25}$$

where we define the followings specifications:

- the innovations are i.i.d. and normally distributed: $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, with $\sigma^2 = 0.1$
- the parameter values are $\beta_0 = 0.5$, $\beta_1 = 4$, $\theta_1 = 2$ and $\theta_2 = -0.5$.
- the MIDAS kernel is based on an exponential Almon lag polynomial which we define as

$$m_K(\theta_1, \theta_2, L) = \sum_{k=0}^{K-1} w_k(\theta) L^k = \sum_{k=0}^{K-1} \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{l=0}^{K-1} \exp(\theta_1 l + \theta_2 l^2)} L^k \tag{1.26}$$

- the covariate $x_t^\kappa = x_{t/\kappa} = x_\tau$ is an AR(1) process: $x_\tau^\kappa = c + \lambda x_\tau^\kappa + \eta_t$ with $c = 0.3$, $\lambda = -0.8$ and $\eta_\tau \sim \mathcal{N}(0, \sigma_e^2)$ with $\sigma_e^2 = 0.8$.
- we consider four aggregation horizons $\kappa = \{1/3, 1/12, 1/22\}$ which correspond to the four aggregation schemes: the month-to-quarter, the month-to-year, and the day-to-month, respectively. We also define four sample size $T = \{30, 50, 100, 500\}$ and $K = \kappa$.

We compare the estimation results with the theoretical results of Andreou et al. (2010) to assess the real scope of the asymptotic properties for various sample sizes. In their paper, Andreou et al. (2010) have proved that the asymptotic variance of $\hat{\beta}_1$ is such as described in (1.24); under conditional homoskedasticity when considering a MIDAS regression model with x_t^κ is an AR(1) process (as described in our simulation process) the asymptotic variance of $\hat{\beta}_1$ is of the form

$$A\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2 \mathrm{E}[\mathsf{x}_\theta^2]}{\mathrm{Var}[\mathsf{x}_\theta]\mathrm{E}[\mathsf{x}_\theta^2] - \mathrm{E}[\mathsf{x}(\theta)\mathsf{x}_\theta]^2} \tag{1.27}$$

where

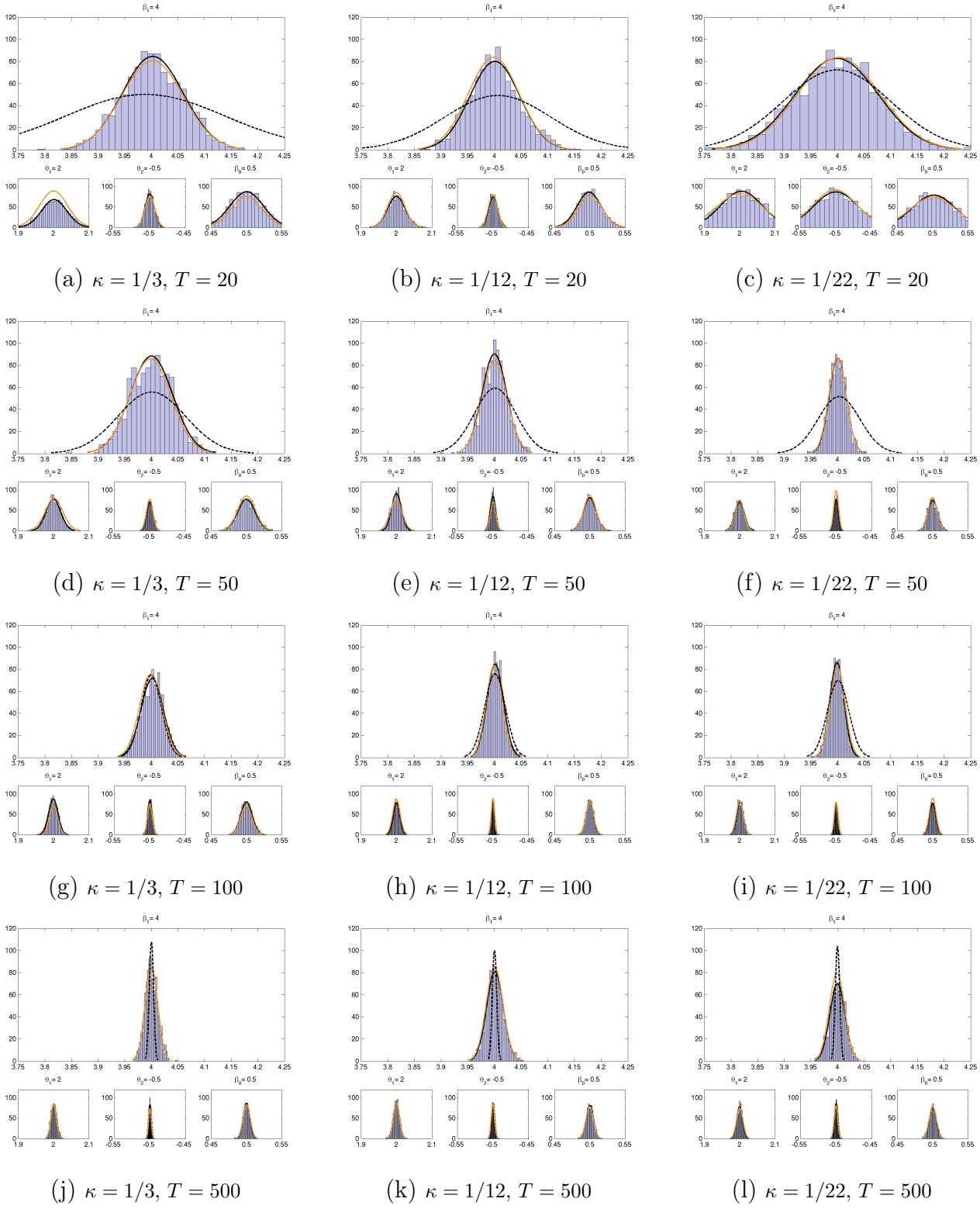
$$\mathrm{E}[\mathsf{x}_\theta^2] = \frac{\sigma_e^2}{1-\lambda^2}\left(\sum_{k=1}^{\kappa-1}\frac{\partial w_k(\theta)}{\partial\theta}(1-\lambda^{\kappa-k})\right)^2 + \sigma_e^2\sum_{k=1}^{\kappa-1}\left(\sum_{l=0}^{\kappa-k-1}\frac{\partial w_{l+1}(\theta)}{\partial\theta}\lambda^{\kappa-k-(l+1)}\right)^2,$$

$$\mathrm{Var}[\mathsf{x}(\theta)] = \frac{\sigma_e^2}{1-\lambda^2}\left(1-\sum_{k=1}^{\kappa-1}w_k(\theta)(1-\lambda^{\kappa-k})\right)^2 + \sigma_e^2\sum_{k=1}^{\kappa-1}\left(\sum_{l=0}^{\kappa-k-1}w_{l+1}(\theta)\lambda^{\kappa-k-(l+1)}\right)^2,$$

$$\begin{aligned}\mathrm{E}[\mathsf{x}(\theta)\mathsf{x}_\theta] = {} & \frac{\sigma_e^2}{1-\lambda^2}\left(\sum_{k=1}^{\kappa-1}\frac{\partial w_k(\theta)}{\partial\theta}(1-\lambda^{\kappa-k})\right)\times\left(\sum_{k=1}^{\kappa-1}w_k(\theta)(1-\lambda^{\kappa-k})-1\right) \\ & + \sigma_e^2\sum_{k=1}^{\kappa-1}\left(\left(\sum_{l=0}^{\kappa-k-1}\frac{\partial w_{l+1}(\theta)}{\partial\theta}\lambda^{\kappa-k-(l+1)}\right)\times\left(\sum_{l=0}^{\kappa-k-1}w_{l+1}(\theta)\lambda^{\kappa-k-(l+1)}\right)\right).\end{aligned}$$

Note that the asymptotic distribution of the estimator $\hat{\beta}_1$ given by (1.21) and specified by the asymptotic variance (1.27) for the AR(1) regressor also depends on the sample size T.

We also use the nonparametric bootstrapping technique put forward by Efron (1979). It involves the random resampling, with replacement, of elements from the original data to generate a replicate data vector with the same size. The distribution of those replicates around the observed data is a correct approximation of the distribution of observed data sets on the true that generates the data sets but is unknown. These replicates are then used as the series of interest so that it allows parameter re-estimation and hence sample a parameter distribution. Bootstrap can be interpreted as a measure of the repeatability of the estimates. Using this framework, we run 1000 simulations of the DGP (1.25) that we compare with both the theoretical asymptotic density and the bootstrapping distribution. The results are exhibited in Figure 1.3.



(a) $\kappa = 1/3$, $T = 20$ (b) $\kappa = 1/12$, $T = 20$ (c) $\kappa = 1/22$, $T = 20$

(d) $\kappa = 1/3$, $T = 50$ (e) $\kappa = 1/12$, $T = 50$ (f) $\kappa = 1/22$, $T = 50$

(g) $\kappa = 1/3$, $T = 100$ (h) $\kappa = 1/12$, $T = 100$ (i) $\kappa = 1/22$, $T = 100$

(j) $\kappa = 1/3$, $T = 500$ (k) $\kappa = 1/12$, $T = 500$ (l) $\kappa = 1/22$, $T = 500$

FIGURE 1.3: Distributions of simulated, asymptotic and bootstrapped NLS estimators across different samples sizes T and different temporal lag horizon κ.
The dashed black line is the theoretical asymptotic distribution, the solid orange line is the simulated draw distribution, the histogram represents the bootstrapping sample and the solid black line is its normal distribution fit.

Several conclusions can be drawn from these simulations. First, the MIDAS NLS estimator is unbiased and provides errorless figures even for the smallest sample size ($T = 50$). Note also that the temporal gap κ does not really influence the convergence of the estimator in terms of mean but it impacts its variance. In fact, the MIDAS NLS estimator is more efficient for both high frequency processes (and high aggregation horizon $1/\kappa$) and obviously with large sample size. Second, we observe that bootstrapping distribution may sometimes be shifted from the simulation distribution. While the simulation distribution relies on 1000 simulated estimates, the bootstrapping distribution is only based on one model and hence is centered on its NLS estimate. Nevertheless, we note that both distributions are similar in variance. That proves that our bootstrapping procedure is suitable for making inference on the MIDAS structure. Finally, simulation graphs also show that theoretical asymptotic estimator of β_1 is obviously not adapted to small size.

It can also be noticed that, the theoretical estimation is too restrictive than the reality in terms of variance even in the case of large sample size. That means that convergence towards this theoretical asymptotic variance requires a very large sample size and hence is not reliable when dealing with macroeconomic time series.

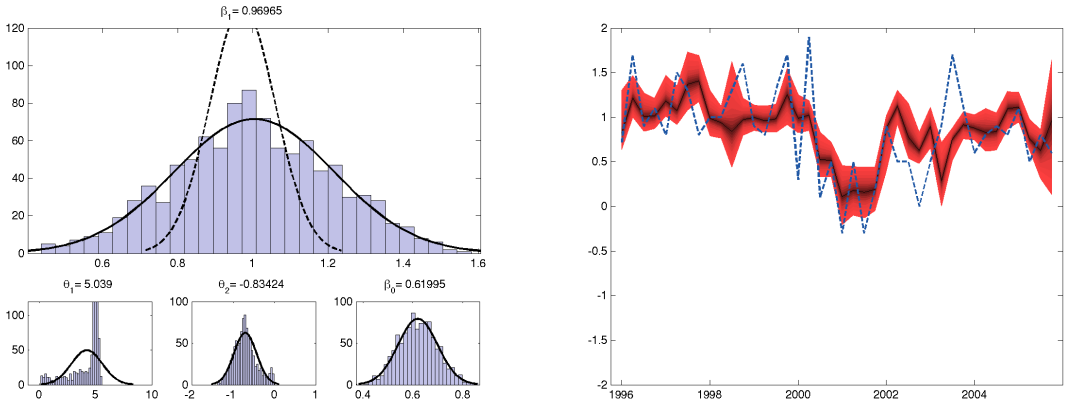
Empirical assessment

We have seen that the MIDAS NLS estimator provides pertinent results on simulated data and is well adapted to underlying aggregation issues within regression models. Let now see its efficiency on real economic data. In this respect, we build a standard macroeconomic framework that explains the US quarterly GDP using the US monthly IPI; the model equation is the following:

$$GDP_t^Q = \beta_0 + \beta_1 \, m_K(\theta_1, \theta_2, L) \, IPI_t^M + \varepsilon_t \tag{1.28}$$

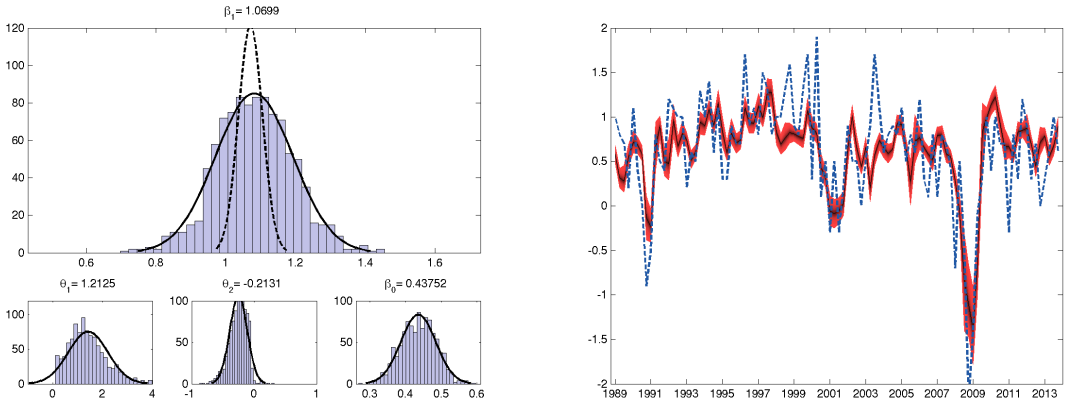
where $M = \kappa = 1/3$ and the parameter family $\{\beta_0, \beta_1, \theta_1, \theta_2\}$ is estimated using the MIDAS NLS method. The lag length coefficient is exogenously defined as $K = 10$. We assess our model on two sample sizes: $T = 40$, 10 years of data corresponding to the period from 1996:q1 to 2005:q5, and $T = 100$ (25 years) from 1989:q1 to 2013:q4. We use a bootstrapping approach to draw the parameters distribution and hence to define their confidence intervals[1.4]. We compare those results with the asymptotical distribution of MIDAS NLS estimator (as given in (1.21) and in (1.24)). The results are presented in Table 1.1 and displayed in Figures 1.4 and 1.5.

[1.4] We run some test for residuals autocorrelation (Ljung-Box Q-test, sample autocorrelation function plots, etc.).



(a) (b)

FIGURE 1.4: Regression results over the period 1996:q1-2005:q4 ($T = 40$), MSE=0.17.



(a) (b)

FIGURE 1.5: Regression results over the period 1989:q1-2013:q4 ($T = 100$), MSE=0.19. On the left-hand side (1.4a and 1.5a), the dashed black line is the theoretical asymptotic distribution, the histogram represents the simulated draws and the solid black line is its normal distribution fit. On the right-hand side (1.4b and 1.5b), the dashed blue line is the observed values of GDP_t while the red fan chart depicts the bootstrapped distribution of the fitted series $\widehat{GDP}_t$.

Sample size	$\hat{\beta}_1$	Bootstrap CI	Asymptotic CI
$T = 40$ (1996:q1-2005:q4)	0.97	[0.64;1.33]	[0.90;1.04]
$T = 100$ (1989:q1-2013:q4)	1.07	[0.91;1.27]	[1.01;1.12]

TABLE 1.1: Slope coefficient estimates and its relative confidence interval (CI) at 95% for two different sample sizes.

It turns out that estimation accuracy is obviously related to the sample size, nevertheless it seems that the model fits the data quite well across periods (see Figures 1.4b and 1.5b). Moreover, it can be noticed that the theoretical variance is smaller than the empirical variance coming from the bootstrapping estimates, as already observed in the Monte Carlo simulations. This theoretical formula is not adapted to the empirical macroeconomic issues which we deal with. In contrast, bootstrapping has provide useful inferences on MIDAS regression parameters.

1.3.3 Various specifications of MIDAS regression models

There is a growing literature using MIDAS models in order to deal with multi frequencies and hence facing particular economic or forecasting issues. In this respect, various extensions or specifications have been recently introduced. We especially refer to the thesis of Foroni (2012) in which she addresses different issues related to the use of mixed-frequency data.

Multiple explanatory variables

When working with economic issues, we face a data-rich environment. As we already discussed, forecasting macro-aggregates involves the use of both explanatory variables coming from various sectors of activity regardless of their sampling frequency and the technical specifications they require. The MIDAS regression models can be extended to the multivariate setting in order to provide a flexible and reasonable framework for this problem. Considering n explanatory variables each sampled at its own frequency κ_i, the MIDAS regression model becomes:

$$y_t = \beta_0 + \sum_{i=0}^{n} \beta_i \, m_{K_i}(\theta_i, L) \, x_{t,i}^{\kappa_i} + \varepsilon_t. \tag{1.29}$$

The number of lags K_i involved in the regression may also be different for each covariate. To explain quarterly series, a daily explanatory variable would normally include more lags than monthly series. Thus, defining the $T \times (n+1)$ matrix of explanatory variables $\mathsf{X}(\theta)$ as follows

$$\mathsf{X}(\theta) = \begin{pmatrix} 1 & m^{K_1}(\theta_1, L)\, x_{1,1}^{\kappa_1} & \cdots & m^{K_n}(\theta_n, L)\, x_{1,n}^{\kappa_n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & m^{K_1}(\theta_1, L)\, x_{T,1}^{\kappa_1} & \cdots & m^{K_n}(\theta_n, L)\, x_{T,n}^{\kappa_n} \end{pmatrix}, \tag{1.30}$$

we can write the MIDAS regression in matrix form such that

$$Y = \mathsf{X}(\theta)\beta + \varepsilon \tag{1.31}$$

where we have parameter vectors $\beta = (\beta_0, \beta_1, \ldots, \beta_n)'$ and $\theta = (\theta_1, \ldots, \theta_n)$, with respect to the number of θ_i, $i = 1, \ldots, n$, used in the weight function, and the residual vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$.

Unrestricted MIDAS model

Foroni et al. (2013) have proposed a new parametrization for the MIDAS that relies on a linearization of the distributed lag function called unrestricted MIDAS (U-MIDAS). That can be assimilated to the expression (1.1) where all the parameters are estimated using OLS. The U-MIDAS model is based on a linear lag polynomial that can be written as:

$$c(L)\mathsf{y}_\tau = \delta(L)x_{\tau-1}^{\kappa} + \epsilon_\tau, \tag{1.32}$$

where $c(L) = (1 - c_1L^1 - \ldots - c_cL^c)$, $\delta(L) = \left(\delta_0 + \delta_1L^1 + \delta_dL^d\right)$ and where $\{\mathsf{y}_\tau\}_\tau$ is the disaggregated process $\{y_t\}_t$ (the underlying des-aggregation scheme is supposed to be known, as well as the polynomial orders c and d[1.5]). Then, forecasts are obtained using a form of direct estimation approach (see Foroni et al. (2013) for details).

Empirically, it has been showed that this technique works quite well as long as the gap between the low and the high frequency is not too large; typically month to quarter or quarter to year. In such cases, one can indeed easily consider the U-MIDAS as a powerful macroeconomic model and therefore as a competitive benchmark for forecasting purposes.

Measuring volatility with MIDAS

Mixed-frequency setting is also of particular interest in the context of volatility modeling. Ghysels et al. (2005) have proposed a MIDAS estimator of volatility computed as a *realized volatility*; that is, by using a weighted average of lagged squared daily returns as a proxy of monthly conditional variance. In the wake of French et al. (1987), Andersen, Bollerslev or Shephard have popularized the use of data sampled at a higher frequency to calculate an ex-post measure of volatility. Here, the sample variance (or

[1.5] In practise, those could be defined using information criteria.

realized volatility) of the previous month of returns $\{r_\tau\}_\tau$:

$$RV_t = \left(\sum_{d=0}^{D-1} r_{\tau-d}^2 \right)^{1/2} \tag{1.33}$$

where D is the number of data of the considered period (the number of days in a month is usually set at $N = 22$). In their paper, Ghysels et al. (2005) proposed a risk return tradeoff model that involves a MIDAS setup derived from the (1.33) such that

$$R_{t+1} = \mu + \gamma RV_t^{midas} + \nu_{t+1} \tag{1.34}$$

where $\nu_t \sim iid\mathcal{N}(0, \sigma_n^2 u)$ and $RV_t^{midas} = \text{Var}[R_{t+1}|\mathcal{F}_t]$ for which they assume that

$$RV_t^{midas} = D \sum_{d=0}^{D-1} \frac{\varphi(d, \theta_1, \theta_2)}{\sum_{l=0}^{D-1} \varphi(l, \theta_1, \theta_2)} \, r_{\tau-d}^2 \tag{1.35}$$

They estimate the parameters θ jointly with μ and γ via QMLE.

Furthermore, Engle et al. (2006) introduced GARCH MIDAS models that combine GARCH process with MIDAS polynomial structure. Then, Colacito et al. (2011) have recently extended the idea of component models for volatility in introducing DCC MIDAS models in order to capture volatility dynamics within a MIDAS structure.

Extensions of the MIDAS

Different extensions of the MIDAS model have been developed in the literature. A first extension is introducing an autoregressive term in the MIDAS regression framework. In the macroeconomic forecasting context, adding AR element usually improves the forecasting accuracy especially when considering macroeconomic aggregates like the GDP (see Stock and Watson (2002)). However, Clements and Galvão (2009) highlighted that this strategy could lead to a misspecification due to a seasonal response of $\{y_t\}_t$ to $\{x_t\}_t$. They suggest to introduce the AR term as a common factor to avoid this inconvenience.

Another interesting extension has been proposed by Marcellino and Schumacher (2010), they have developed a new approach that combines factor models with the MIDAS framework. In particular, this Factor-Augmented MIDAS model will be described and exploited for different empirical studies in this dissertation in Chapters 2 and 4. Some other models incorporate regime changes in the parameters; we refer to the Markov switching MIDAS model proposed by Guérin and Marcellino (2013) or to the

smooth transition MIDAS models introduced by Galvão (2013). Recently, the MIDAS framework has been extended to multivariate specifications, we refer especially to the mixed-frequency vector autoregressive models developed in Foroni et al. (2014) and to the recent paper of Götz and Hecq (2014) on causality.

A review of all these techniques has been proposed for nowcasting purposes by Foroni and Marcellino (2013a); Sestieri (2014) has discussed their paper and highlighted the main results. In their work, Claudia Foroni and Massimiliano Marcellino conclude that: (i) Although no model significantly outperforms the others, it seems that both bridge models and AR-MIDAS models tend to improve the nowcasting performances. (ii) Pooling information using factor models improves the nowcasting accuracy. It can be noted that their results are obtained using both single indicator models and forecast combinations within each class of models. That point is of interest in the next Chapters.

Chapter 2

Macroeconomic forecasting with mixed-frequency data

2.1 Financial volatility as a macroeconomic leading indicator

This section is based on the paper entitled "Forecasting growth during the Great Recession: is financial volatility the missing ingredient?", written with Laurent Ferrara and Juan-Pablo Ortega and published in Economic Modelling, no. 36(C) (January 2014).

In the wake of the financial and banking crisis, virtually all industrialized countries experienced a very severe economic recession during the years 2008 and 2009, generally referred to as the Great Recession. This recession has shed light on the necessary reassessment of the contribution of financial markets to the economic cycles. There is

a huge volume of work in the literature that underlines the leading role of financial variables in the forecasting of macroeconomic fluctuations. For example, Kilian (2008) reviewed the impact of energy prices shocks, especially oil prices, on macroeconomic fluctuations; Hamilton (2003) put forward a non-linear Markov-Switching model to predict the US Gross Domestic Product (GDP) growth rate using oil prices. Stock and Watson (2003) have proposed a review on the role of asset prices for predicting the GDP, while Claessens et al. (2012) have empirically assessed interactions between financial and business cycles. Recently, Ferrara and Marsilli (2013) have evaluated the predictive power of some major financial variables to anticipate GDP growth in Euro Area countries during the Great Recession[2.1].

Nevertheless, there are only very few studies in the literature dealing with the impact of financial volatility on macroeconomic fluctuations. Among the rare existing references, Hamilton and Lin (1996) have shown evidence of relationships between stock market volatility and US industrial production through non-linear Markov-Switching modeling; Ahn and Lee (2006) have estimated bi-variate VAR models with GARCH errors for both industrial production and stock indices in five industrialized countries. Chauvet et al. (2012) have recently analyzed the predictive ability of stocks and bonds volatilities over the Great Recession using a monthly aggregated factor. Indeed, they estimate a monthly volatility common factor based on realized volatility measures for stock and bond markets. They show that this volatility factor largely explains macroeconomic variable during the 2007-2009 recession, both in-sample and out-of-sample.

In this section, we aim at assessing the impact of financial volatility on output growth in three advanced economies (US, UK, and France) using a MIDAS model capable of putting together daily and monthly sampled explanatory variables in order to predict the quarterly GDP growth rate; this approach has been explained in detail in Chapter 1. We use two well-known daily sampled financial ingredients, namely, commodity and stock prices, combined with a monthly industrial production index to empirically show the gain in prediction performance for various forecasting horizons, when daily financial volatility is included in the mixed-frequency models. Our study provides conclusive empirical proof that this approach increases the predictive accuracy during a period that includes the last Great Recession for the three considered countries.

[2.1] This work, entitled *Financial variables as leading indicators of GDP growth: Evidence from a MIDAS approach during the Great Recession* written with Laurent Ferrara, is presented in Appendix B.

2.1.1 Financial volatility and real economic activity

We aim at assessing the predictive content of the daily volatility of financial variables regarding the gross domestic product (GDP) using the MIDAS approach introduced in Ghysels et al. (2004). This forecasting strategy allows the use of explaining variables sampled at different frequencies avoiding at the same time the loss of information associated to data temporal aggregation; this is achieved by exploiting parsimoniously parametrized weight functions that specify the importance of each covariate along their past in an economically reasonable fashion. A major motivation for exploring this scheme is the well known fact that hard data, generally sampled with monthly frequency, convey additional information to anticipate the GDP that is, in turn, quarterly measured. Using the MIDAS approach we will go a step further and will incorporate in the forecasting setup a combination of monthly and daily sampled covariates. This approach has already been studied by Andreou et al. (2013) who show the pertinence, from the point of view of increase in the forecasting power, of combining monthly macroeconomic indicators with daily financial explaining data. The GDP prediction proposed in their work is constructed via the weighted combination of a number of individual MIDAS based forecasts obtained by using a single financial covariate at a time. The authors have indeed used an important financial dataset in order to construct a rich family of separate MIDAS forecasts; their combination yields satisfactory results and shows the predictive relevance of daily information in the macroeconomic context. Our work can be seen as an extension, focusing on the financial volatility as predictor of the real GDP growth during the Great Recession.

Let y_t^Q be a quarterly sampled stationary variable that we aim at predicting, X_t^M is a vector of N_M stationary monthly quoted variables, and X_t^D is a vector of N_D stationary daily variables. We use the multivariate MIDAS model we introduced in (1.29) enabling the mixing of daily and monthly information:

$$y_t^Q = \alpha + \sum_{i=1}^{N_D} \beta_i \, m_{K_D}(\theta_i) \, X_{i,t}^D + \sum_{j=1}^{N_M} \gamma_j \, m_{K_M}(\omega_j) \, X_{j,t}^M + \phi \, y_{t-1}^Q + \varepsilon_t^Q, \qquad (2.1)$$

where ε_t^Q is a white noise process with constant variance and α, β, θ, γ, ω are the regression parameters to be estimated. We also include a first order autoregressive term in the expression (2.1) as it has been showed that it generally improves forecasting accuracy based on leading indicators (see for example Stock and Watson (2003)). The m_K function in equation (2.1) prescribes the polynomial weights that allow the frequency mixing. In this respect, we use a Beta restricted function which we defined in (1.13). While other weight function specifications often employed in the literature

like the exponential Almon form, relies on the use of at least two parameters, the Beta restricted function involves only one parameter. Additionally it imposes decreasing weight values which is a desirable feature in view of the direct multistep forecasting setup that we adopt later on in the empirical application that we will carry out later on.

As one of the main objectives of our work consists in providing evidence of the macroeconomic predictive content of financial volatilities, a crucial issue is the estimation of volatility. Given that volatility is not directly observable, several methods have been developed in the literature to estimate it. The most straightforward approach to this problem relies in the use of the absolute value of the returns as a proxy for volatility; unfortunately, the results obtained this way are generally very noisy (see Andersen and Bollerslev, 1998). This difficulty can be partially fixed by using an average of this noisy proxy over a given period; this method yields one of the most widely used notion of volatility, namely the *realized* volatility (as described in (1.33) and used, for example, in Chauvet et al., 2012). Since our goal is working with daily financial volatility, the realized approach would require intraday data whose availability may be an issue and that, additionally, requires a delicate handling (overnight effects, price misrecordings, etc). An alternative convenient approach appears to be the volatility filtered out of a GARCH-type parametric family (see Engle, 1982 and Bollerslev, 1986). The AR(p)-GARCH(r,s) specification is given by

$$\begin{cases} r_t^D &= \psi_0 + \psi_1\, r_{t-1}^D + \cdots + \psi_p\, r_{t-p}^D + w_t, \\ w_t &= v_t^D\, \eta_t, \\ (v_t^D)^2 &= c + \sum_{i=1}^{\mathrm{r}} a_i\, w_{t-i}^2 + \sum_{j=1}^{\mathrm{s}} b_j\, (v_{t-j}^D)^2, \end{cases} \tag{2.2}$$

where ψ_0 is a constant, $\psi = (\psi_1, \ldots, \psi_p)$ is a p-vector of autoregressive coefficients and $\{\eta_t\}_t \sim \mathrm{WN}(0,1)$. In order to ensure the existence of a unique stationary solution and the positivity of the volatility, we assume that $a_i > 0$, $b_j \geq 0$ and $\sum_{i=1}^{\mathrm{r}} a_i + \sum_{j=1}^{\mathrm{s}} b_j < 1$. Estimated daily volatilities $\{\hat{v}_t^D\}_t$ stemming from equation (2.2) will be considered as explanatory variables of the macroeconomic fluctuations using the MIDAS regression equation (2.1), with $X_{i,t}^D = \hat{v}_{i,t}^D$.

Finally, when using general regression models for forecasting purposes at a given horizon $h > 0$, forecasters can either predict covariates or implement direct multi-step forecasting (see Section 1.2.2 and Chevillon (2007) for a review). The idea behind direct multi-step forecasting is that the potential impact of specification errors on the one-step-ahead model can be reduced by using the same horizon both for estimation

and for forecasting at the expense of estimating a specific model for each forecasting horizon. In our work we adopt the direct multi-step forecasting and assume that the predictor $y^Q_{t+h|t}$ of the GDP quarterly growth rate, for any forecasting horizon h, is given by

$$y^Q_{t+h|t} = \hat{\alpha}^{(h)} + \sum_{i=1}^{N_D} \hat{\beta}_i^{(h)}\; m_{K_D}(\hat{\theta}_i^{(h)})\, \hat{v}^D_{i,t} + \sum_{j=1}^{N_M} \hat{\gamma}_j\; m_{K_M}(\hat{\omega}_j^{(h)})\, X^M_{j,t} \; + \hat{\phi}^{(h)}\, y^Q_t, \quad (2.3)$$

where $\left\{\hat{\alpha}^{(h)}, \hat{\beta}_1^{(h)}, \ldots, \hat{\beta}_{N_D}^{(h)}, \hat{\theta}_1^{(h)}, \ldots, \hat{\theta}_{N_D}^{(h)}, \hat{\gamma}_1^{(h)}, \ldots, \hat{\gamma}_{N_M}^{(h)}, \hat{\omega}_1^{(h)}, \ldots, \hat{\omega}_{N_M}^{(h)}, \phi^{(h)}\right\}$ are the non-linear least squares estimates ($2N_D + 2N_M + 2$ parameters need to be estimated).

2.1.2 Empirical results

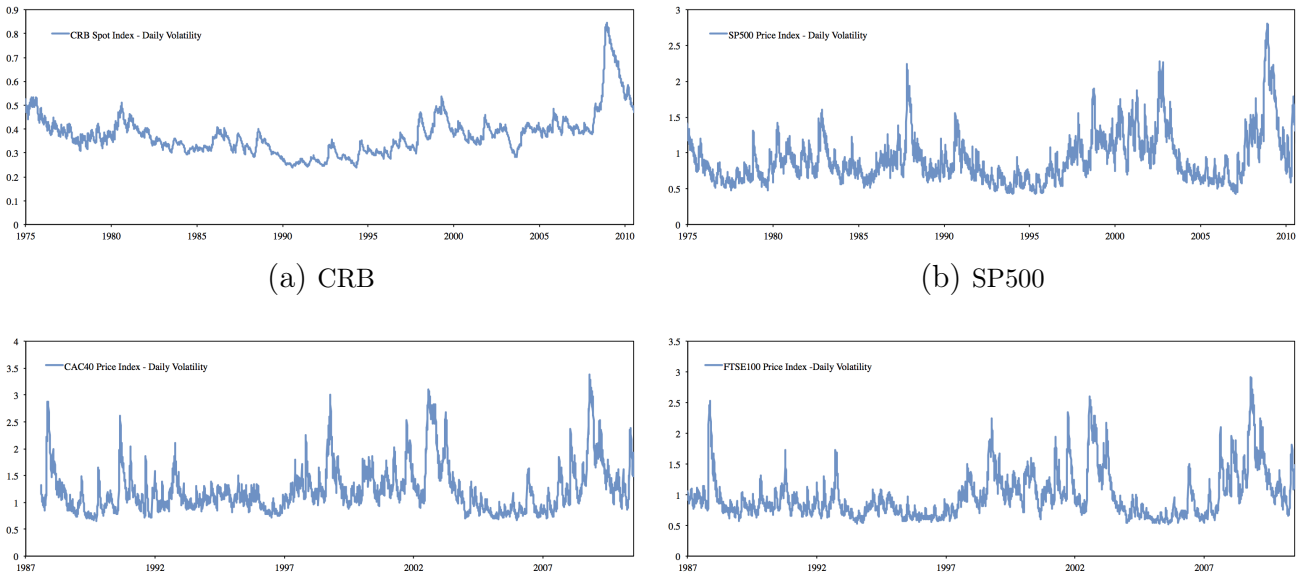
In this section we focus on the GDP growth prediction. We implement the model previously introduced in equation (2.3) in order to assess the forecasting ability of the volatility of two financial variables, namely commodity and stock prices (i.e. $N_D = 2$) in comparison with the monthly industrial production (i.e. $N_M = 1$). The variable that we want to predict is the quarterly growth rate of the real GDP (expressed in percentage and denoted GDP_t) of three countries: US, France, and the UK, as released by the corresponding national offices of statistics in July 2013. Details concerning sources and the datasets are given in Table 2.1[2.2]

Quarterly output		
GDP	Real US GDP growth (*Bureau of Economic Analysis*)	1976q1:2010q4
	French GDP growth (*INSEE*)	1988q1:2010q4
	UK GDP growth (*Office for National Statistics*)	1988q1:2010q4
Daily volatilities		
CRB	CRB spot price index (*Commodity Research Bureau*)	01jan1964:31dec2010
SP	S&P500 index (*Standard & Poors*)	01jan1964:31dec2010
	CAC40 index (*Euronext Paris*)	03aug1987:31dec2010
	FTSE100 index (*FTSE*)	01jan1987:31dec2010
Monthly series		
IPI	US industrial production index manuf. (*Fed. Reserve*)	jan1976:dec2010
	French industrial production index manuf.(*INSEE*)	jan1988:dec2010
	UK industrial production index manuf. (*ONS*)	jan1988:dec2010

Table 2.1: Description of indicators and covariates

[2.2] All the data have been downloaded from *Datastream*.

We consider as explanatory daily variables the CRB index of commodity prices and the main national stock price indices of those three countries, namely the S&P500, the CAC40, and the FTSE100, that we denote generically as CRB_t and SP_t. Note that stock prices for France and the UK begin later than for the US (1987 instead of 1975). For all daily returns of financial variables, we estimate their volatility on the available sample by using a AR(1)-GARCH(1,1) specification as in the equation (2.2). The model orders have been selected using the Bayes Information Criterion (BIC). Since we are using a standard maximum likelihood estimator for the GARCH process and not a robust one (see for example, Charles and Darné, 2005, or Carnero et al., 2012) we have smoothed out outliers from all returns via a 99.5% Winsorization (for instance the *Black Monday* outlier in the S&P500 series occurring October 19$^{\text{th}}$, 1987). Estimates of daily volatility for both variables, denoted $\{\hat{v}^D_{t,CRB}\}_t$ and $\{\hat{v}^D_{t,SP}\}_t$, are presented in Figure 2.1.



(a) CRB (b) SP500

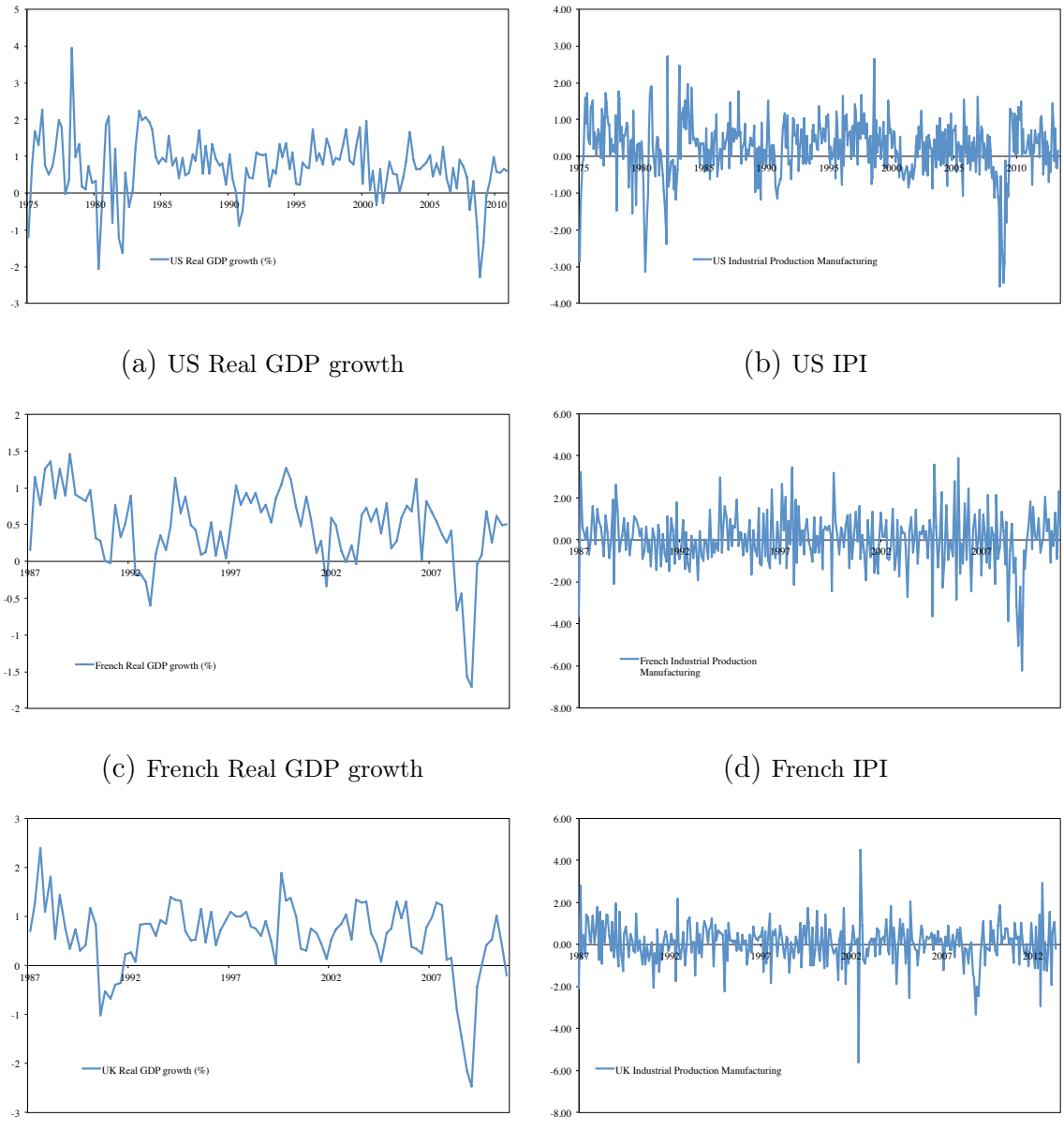
(c) CAC40 (d) FTSE100

FIGURE 2.1: Volatilities estimated using GARCH models.

It is worth noting that, as it usually happens with financial time series, periods of high volatility are clustered in time; nevertheless, the high volatility clusters do not occur at the same time for both time series. The volatility of stock prices presents a huge peak during the recent financial crisis, as well as several smaller peaks related to specific events (Asian crisis, burst of the internet bubble, *etc.*). Since we focus on the post 1973 oil crisis period (from 1975 to 2010), the commodity volatility exhibits only one main peak related to the recent financial crisis. Some specific events also drive commodity volatility dynamics such as the second oil shock in the early 1980s or the

Asian crisis in the late 1990s. The information conveyed by both volatilities does not seem redundant in spite of a recent increase in their correlation (see *e.g.* Creti et al., 2012) and both variables are potentially useful in explaining GDP growth.

As monthly explanatory variable in (2.3) we use the Industrial Production manufacturing Index (IPI) that is well known by practitioners to be informative about the evolution of macroeconomic variables in general and as to the dynamics of the GDP growth in particular. Those series for the three countries are represented in Figure 2.2.



(a) US Real GDP growth

(b) US IPI

(c) French Real GDP growth

(d) French IPI

(e) UK Real GDP growth

(f) UK IPI

FIGURE 2.2: GDP and IPI growth rates for the US, France, and the UK.

In our study we carry out parameters estimation using the periods 1976q1-2006q4 for the US and 1988q1-2006q4 for France and the UK, then we implement an out-of-sample experience over the period 2007q1 - 2010q4 that includes the Great Recession. Concerning the forecasting experiment, given that financial data are always available the last working day of any given month, we suppose that forecasts for a specific quarter are computed at the end of each month, for 12 horizons that range from $h = 0$ (nowcasts computed at the end of the last month of the reference quarter) to $h = 11/3$ (forecasts computed 11 months before the end of the reference quarter). For any time t the MIDAS regression optimally takes advantage of the fluctuations of the last $K_M = 6$ for the monthly series IPI_t and $K_D = 90$ for financial covariates $\hat{v}^D_{t,CRB}$ and $\hat{v}^D_{t,SP}$. As we have chosen a direct multi-step forecasting approach, model parameters are estimated separately for each prediction horizon h, as in equation (2.3).

In a first step, we assess the specific impact of both financial volatilities on the GDP growth process through a standard MIDAS model that relates daily variables with a quarterly variable. Thus, the first model that we estimate, denoted **Model $\mathbf{M_d}$**, contains as regressors the daily volatilities of both financial series, namely $\hat{v}^D_{t,CRB}$ and $\hat{v}^D_{t,SP}$:

$$GDP_{t+h|t} = \hat{\alpha}^{(h)} + \hat{\beta}_1^{(h)}\; m_{K_D}(\hat{\theta}_1^{(h)})\, \hat{v}^D_{t,CRB} + \hat{\beta}_2^{(h)}\; m_{K_d}(\hat{\theta}_2^{(h)})\, \hat{v}^D_{t,SP} + \hat{\phi}^{(h)}\, GDP_t. \tag{M_d}$$

The second model, denoted **Model $\mathbf{M_m}$**, contains only as regressors the monthly growth rate of the IPI:

$$GDP_{t+h|t} = \hat{\alpha}^{(h)} + \hat{\gamma}^{(h)}\; m_{K_M}(\hat{\omega}^{(h)})\, IPI_t + \hat{\phi}^{(h)}\, GDP_t. \tag{M_m}$$

Explaining GDP growth using industrial production is standard in the empirical literature on short-term macroeconomic forecasting, especially when using bridge equations (see for example Diron, 2008, or Barhoumi et al., 2012). However, the monthly IPI series is generally aggregated before using it in quarterly equations. Here, by using a standard MIDAS equation, we allow for different weights concerning the contribution of monthly IPI to GDP growth, adding thus more flexibility to the model.

The third model, denoted **Model $\mathbf{M_{dm}}$**, contains as regressors both daily volatilities and the monthly IPI:

$$\begin{aligned} GDP_{t+h|t} = \hat{\alpha}^{(h)} &+ \hat{\beta}_1^{(h)}\; m_{K_D}(\hat{\theta}_1^{(h)})\, \hat{v}^D_{t,CRB} + \hat{\beta}_2^{(h)}\; m_{K_D}(\hat{\theta}_2^{(h)})\, \hat{v}^D_{t,SP} \\ &+ \hat{\gamma}^{(h)}\; m_{K_M}(\hat{\omega}^{(h)})\, IPI_t + \hat{\phi}^{(h)}\, GDP_t. \end{aligned} \tag{M_dm}$$

To assess the forecasting accuracy of each model, we compute the Root Mean Square Forecasting Errors (RMSFE), for all forecasting horizons h, based on differences between realized values GDP_{t+h} and forecasted values $GDP_{t+h|t}$ on the 16 point forecasts over the 4 years out of sample from 2007q1 to 2010q4. In order to have a measure of the real predictive ability of financial market volatility, we also provide forecasting results using a simple autoregressive model AR(1) as a benchmark:

$$GDP_{t+h|t} = \hat{\alpha}^{(h)} + \hat{\phi}^{(h)} GDP_t. \tag{AR}$$

We note that an autoregressive element has always been added in the three models to play the role of control variable and to potentially improve the prediction. Comparing the obtained results with those using Model (AR) help us measuring the real contribution of the explanatory variables and financial data in particular.

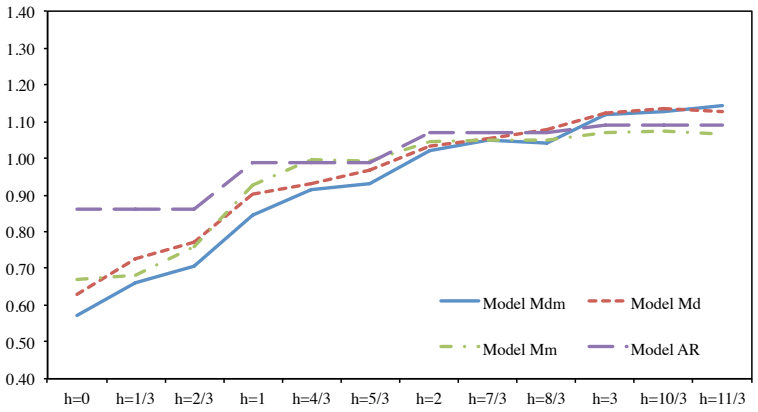
For each model (**Model $\mathbf{M_d}$**, **Model $\mathbf{M_m}$**, **Model $\mathbf{M_{dm}}$**) and each forecast horizon h, RMSFE(h) values are presented in Table 2.2. In addition, RMSFE(h) values, for h ranging from zero to 11/3, are also plotted in Figure 2.3a for US, in Figure 2.3b for France and in Figure 2.3c for the UK.

	Forecasting horizons h											
	0	1/3	2/3	1	4/3	5/3	2	7/3	8/3	3	10/3	11/3
RMSFE(h) for the US												
Model M_d	0.63	0.73	0.77	0.90	0.93	0.97	1.03	1.05	1.08	1.12	1.13	1.12
Model M_m	0.67	0.68	0.76	0.93	1.00	0.99	1.05	1.05	1.05	1.07	1.07	1.07
Model M_{dm}	0.57	0.66	0.71	0.84	0.92	0.93	1.02	1.05	1.04	1.12	1.13	1.14
Model AR	0.86	0.86	0.86	0.99	0.99	0.99	1.07	1.07	1.07	1.09	1.09	1.09
RMSFE(h) for France												
Model M_d	0.61	0.62	0.62	0.69	0.69	0.70	0.82	0.82	0.80	0.98	1.02	0.99
Model M_m	0.51	0.53	0.53	0.70	0.73	0.72	0.78	0.77	0.79	0.88	0.83	0.84
Model M_{dm}	0.48	0.51	0.50	0.61	0.65	0.68	0.80	0.81	0.80	0.89	0.81	0.87
Model AR	0.62	0.62	0.62	0.73	0.73	0.73	0.82	0.82	0.82	0.86	0.86	0.86
RMSFE(h) for the UK												
Model M_d	0.74	0.76	0.78	0.99	1.01	1.04	1.19	1.20	1.22	1.34	1.35	1.37
Model M_m	0.84	0.85	0.91	1.04	1.10	1.23	1.30	1.30	1.27	1.29	1.32	1.31
Model M_{dm}	0.71	0.72	0.80	0.97	1.02	1.04	1.14	1.17	1.16	1.34	1.32	1.32
Model AR	1.10	1.10	1.10	1.29	1.29	1.29	1.36	1.36	1.36	1.37	1.37	1.37

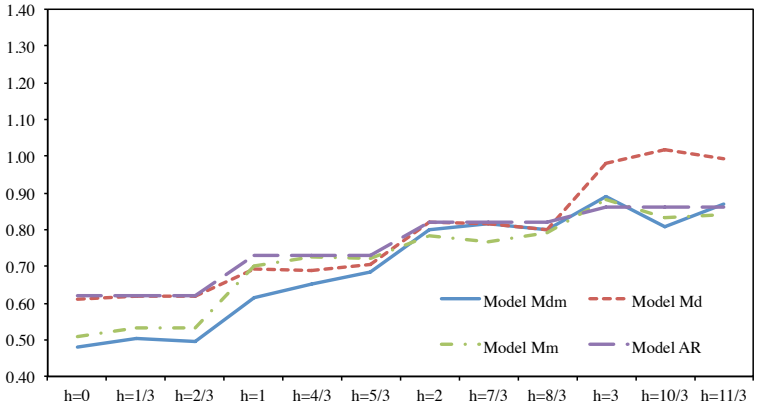
TABLE 2.2: RMSFE(h) for quarterly GDP growth. The forecasting horizon h is in quarters.

As expected, RMSFE(h) for all models and all countries decrease when h tends to

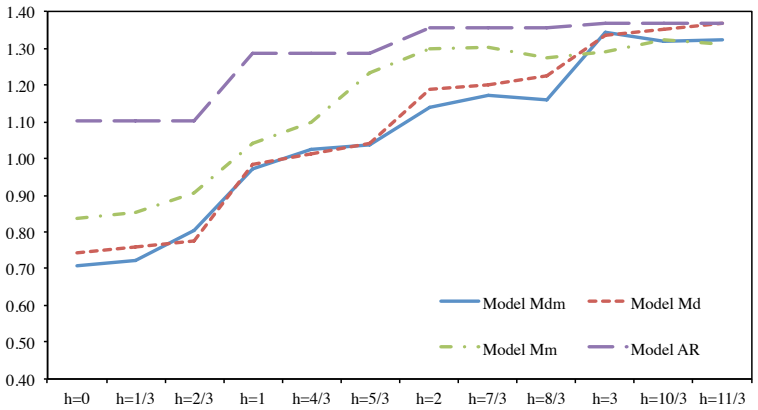
zero, reflecting the use of an information set of increasing size. Indeed, RMSFE(h) are more than halved when h goes from 8/3 to zero. Especially, when $2/3 \leq h \leq 4/3$, we observe a strong negative slope, visible for all models. This is due to the integration of the newly available GDP growth figure of the previous quarter.



(a) RMSFE(h) for US



(b) RMSFE(h) for France



(c) RMSFE(h) for UK

FIGURE 2.3: RMSFE(h)

An analysis of Figure 2.3a, Figure 2.3b, and Figure 2.3c shows that the **Model** $\mathbf{M_{dm}}$ based on daily financial volatilities and monthly IPI unanimously provides the best results for all horizons h for the three economies analyzed. This result proves in a robust fashion that combining information coming from both macroeconomic and financial sources appears to be a good strategy when forecasting GDP. We are in agreement in this point with much of the literature on macroeconomic forecasting and nowcasting that underlines the usefulness of either combining information (for example through dynamic factor models, see e.g. Giannone et al. (2008)) or combining forecasts (see *e.g.* Timmermann (2006)). In fact, the forecasting gain obtained by using the **Model** $\mathbf{M_{dm}}$ becomes important already for $h \leq 1$. We also note that between $h = 4/3$ and $h = 7/3$, the contribution of financial volatilities to the forecasting accuracy is remarkable, specially for the US and the UK economies, as RMSFE(h) stemming from **Model** $\mathbf{M_{dm}}$ and **Model** $\mathbf{M_{d}}$ are almost similar. This result is interesting for practitioners in the sense that using industrial production to predict GDP with a lead of four to seven months does not appear useful; only financial volatilities help in this range of horizons. Nevertheless, we note that the forecasting results for the UK are leaded by the financial **Model** $\mathbf{M_{d}}$ while it appears that the **Model** $\mathbf{M_{m}}$ does not really contribute, not even for short term horizons, to the predictive accuracy of the combined **Model** $\mathbf{M_{dm}}$. These results suggest that financial variables play an important role in forecasting the real UK economy. This has often been underlined in the literature; we refer, among others, to Simpson et al. (2001).

When we are close to the target date, that is during the quarter before the release (i.e. $h \leq 1$), the IPI tends to increase its impact on the forecast in particular in the case of the US and France. This stylized fact has been also observed in empirical papers pointing out the increasing role of hard variables on macroeconomic forecasts when we are close to the release date, while financial variables have a stronger impact for longer horizons (we refer for example to Angelini et al., 2011). Our study shows that the information contained on the industrial output series cannot replace the one associated to financial volatility; both sources of information are playing an important role, but at various horizons.

2.2 Nowcasting the world growth

This section is based on the working paper entitled "Nowcasting Global Growth", written with Laurent Ferrara. It proposes new models to forecast the current state of the global economy. This research is currently used for policy analysis and decision-making at the Banque de France.

Assessing world economic growth in real-time is a key point for macroeconomists in charge of monitoring global economic issues but also a real challenge for econometricians. There is currently no global statistical institute in charge of providing official quarterly national accounts at a global level, in spite of recent efforts in this direction coordinated by international institutions. In this respect, the OECD now releases real GDP growth rate figures for the G20 aggregate on a quarterly basis, based on a common work with several other institutions, such as IMF, BIS, ECB or Eurostat, within the framework of the G20 Data Gaps Initiative[2.3].

This G20 GDP has the great advantage of being sampled on a quarterly basis, but presents the drawbacks of (i) starting only in 2002 which somewhat limits the econometric analysis and (ii) focusing only on G20 countries leaving aside around 15% of world GDP. In addition, GDP figures are released around 70 days after the end of the quarter. Another well known reference among macroeconomists is the IMF that provides global estimates that are considered by experts in the field as benchmark figures when aiming at monitoring the world economy. The time series of the annual global growth, as provided by the IMF in the April 2014 World Economic Outlook (WEO hereafter), is presented in Figure 2.4, from 1995 to 2013.

In this study, we will consider this IMF-WEO series as the definitive estimates[2.4]. We clearly see that the world economic growth has been strongly affected by the Great Recession in 2009, reaching its lowest level since the start of the series. We also observe a sharp increase in growth since the early 2000s, due to some emerging countries, like China in particular. Since the bounce-back in 2010, it seems that the world economy was rather sluggish, showing a marked deceleration.

[2.3] For further details see OECD website.
[2.4] The IMF WEO estimates and projections account for 90 percent of the world purchasing-power-parity weights and are available in the IMF website.

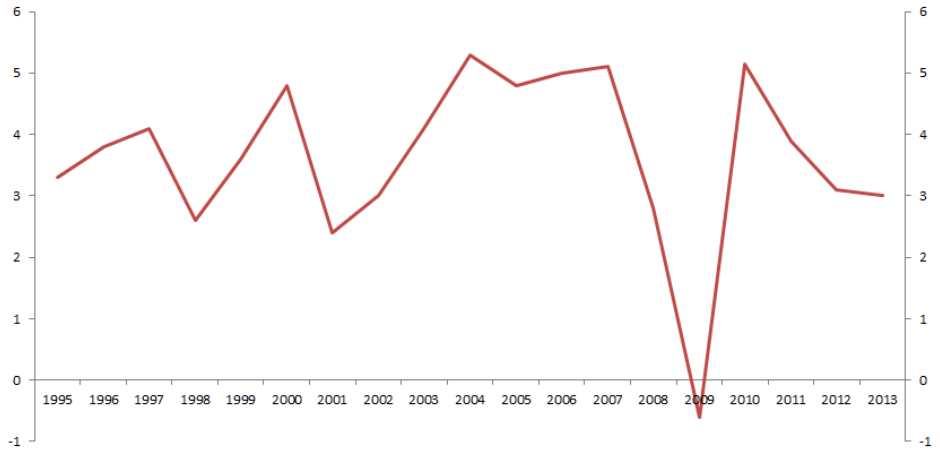


FIGURE 2.4: Annual global growth estimates (source: WEO-IMF, April 2014).

In fact, each time the IMF-WEO is published, the IMF releases estimates of annual global growth for the past years but also for the current year (i.e. nowcasts) and the two upcoming years (i.e. forecasts). The WEO is released two times per year (usually in April and October) and two other WEO updates also come in January and July, but with much less details. Thus, it turns out that four nowcasts of the world economic growth rate for the current year are available.

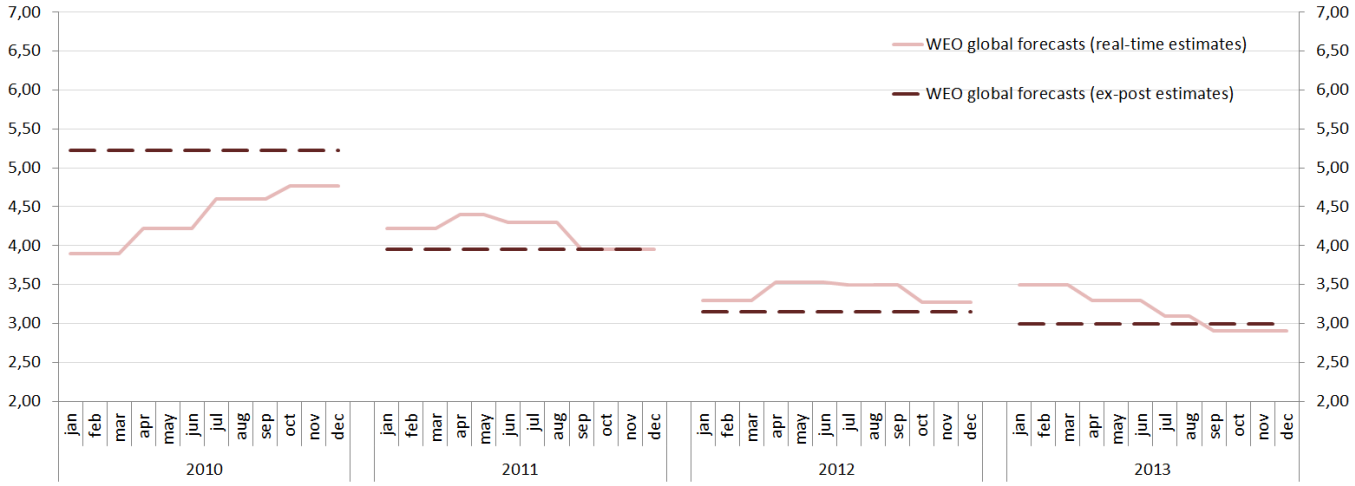


FIGURE 2.5: WEO Global growth nowcasts and final estimates over the sluggish recovery period, as provided in April 2014.

In Figure 2.5, we present the evolution of IMF-WEO nowcasts for global growth over the period 2010-2013, as well as the definitive figures stemming from the April 2014 WEO release. It is noteworthy that there a clear bias at the beginning of each year, then the nowcasts tend to slowly converge to the realized growth rate. However, this

bias does not appear to be systematic. Indeed, it turns out that in 2010, the IMF started by largely underestimating global growth: while the final growth is estimated at 5.2%, the first nowcasts were slightly above 3% in the wake of the Great Recession that affected simultaneaously all the countries. Then the first revision of the year that came with the April 2010 WEO nowcast led to an upward shift by around one percentage point. In opposition, the IMF tended to overestimate growth in their nowcasts in 2011 and 2012. Those optimistic forecasts are partly related to the higher than expected fiscal multipliers, especially in the euro area in 2011 and 2012 as acknowledged by Blanchard and Leigh (2013). In fact, fiscal consolidation programs implemented in the main advanced countries strongly weighed on growth, at least much more than expected by standard macroeconomic models. In addition, it is likely that some confidence effects, often neglected in forecasting models, were at play during this specific period of time, acting as a drag on growth, especially on investment.

The main issue with those IMF-WEO nowcasts is that they reflect an annual growth rate and are released at a quarterly frequency, while obviously economists have at disposal a large set of information on the world economy available on a higher frequency. For example, for many countries, practitioners have access to a large volume of data from opinion surveys of households and businessmen as well as various series on prices (equity prices, housing prices, etc.) and real activity, such as the industrial production index (IPI), household consumption, unemployment rate, etc. Some recent papers have tackled this issue by considering various approaches. For example, Golinelli and Parigi (2013) have developed several bridge models to forecast quarterly world GDP growth rates based on monthly indicators for many countries. Rossiter (2010) takes a similar approach but only considers PMI indicators to explain global variables. Matheson (2011) estimates some dynamic factor models for a large panel of countries and then aggregates forecasts in order to get estimates of the global growth. Drechsel et al. (2014) also show that adding monthly leading global indicators (such as OECD composite leading indicators) to the IMF-WEO forecasts, through bridge equations, lead to accuracy improvements in some cases.

Against this background, when aiming at nowcasting global growth on a high frequency basis, let's say monthly, then one faces two major issues namely (i) a data-rich environment and (ii) a discrepancy between annual GDP figures, on the one hand, and monthly information, on the other hand. In recent years, those two issues have been tackled by econometricians. First, a number of econometric methods have been proposed in the literature enabling to deal with such data-rich environments. Among the different methodologies, dynamic factor models have grown significantly in popularity since the early 2000s and the seminal papers of Forni et al. (2000), Forni et al. (2003)

and Stock and Watson (2002). These models can be used to summarise the information contained in large datasets into a small number of factors common to those variables and have proved very useful in macroeconomic analysis and forecasting in a data-rich environment (see among others Giannone et al. (2008)). Second, when dealing with variables sampled at various frequencies (e.g. annual GDP and monthly financial information), the mixed data sampling (MIDAS hereafter) approach put forward by Ghysels and his co-authors has led to many interesting results in macroeconomic applications (see Ghysels et al. (2007)). Especially in the forecasting framework, several empirical papers have shown the ability of financial information to predict macroeconomic fluctuations; we refer for example to Clements and Galvão (2008) or Ferrara et al. (2014) for the US of Ferrara and Marsilli (2013) for the euro area (see Section 2.1 or Appendix B). Combining dynamic factor models and a MIDAS approach into a Factor-Augmented MIDAS (FA-MIDAS) model has been put forward by Marcellino and Schumacher (2010) when dealing with the German economy. This latter approach is convenient as the two main stylized facts, namely large databases and mixed frequencies, can be accounted for by the FA-MIDAS model.

We implement the FA-MIDAS approach in order to nowcast the global GDP growth rate on a monthly basis, starting from a large database of macroeconomic indicators for several advanced and emerging countries. We compare our results with IMF-WEO nowcasts during the recovery from the Great Recession and we empirically show that our approach is able to better reflect global economic conditions, by reducing mean squared-errors, at least at the beginning of each year, when fewer information is available.

2.2.1 The econometric framework

The econometric methodology implemented in this study builds on the FA-MIDAS approach put forward by Marcellino and Schumacher (2010). In this approach, the information contained in the large database of monthly macro-variables is summarized into few underlying factors, supposed to represent the common evolution of all the series. Then we assume that the annual world GDP growth rate can be explained by a MIDAS regression enabling to explain this low frequency variable by exogenous monthly variables, without any aggregation procedure and within a parsimonious framework.

To exploit a large database including various variables for different countries of the world economy, we implement first a factor analysis that reduces the dimension of the problem. Thus, assume the $1 \times n$ time vector of monthly macroeconomic variables, X_τ,

can be represented as the sum of two mutually orthogonal unobservable components: the common component χ_τ and the idiosyncratic component ξ_τ. For a given month τ, the static factor model is defined by

$$X_\tau = \Lambda f_\tau + \xi_\tau, \tag{2.4}$$

where $X_\tau = (x_{\tau 1} \ \dots \ x_{\tau n})'$ has zero mean and covariance matrix $\Gamma(0)$, Λ is the loading matrix such that $\Lambda = (\lambda_1 \ \dots \ \lambda_n)'$, the common components $\chi_\tau = \Lambda f_\tau$ are driven by a small number r of factors f_τ common to all the variables in the model such that $f_\tau = (f_{\tau 1} \ \dots \ f_{\tau r})'$, and $\xi_\tau = (\xi_{\tau 1} \ \dots \ \xi_{\tau n})'$ is a vector of n idiosyncratic mutually uncorrelated components, driven by variable-specific shocks.

Once the r common monthly factors from the original database have been extracted, we relate them to the annual global growth y_t sampled on a yearly frequency described by the index t. Thus, we observe m times the explanatory factor over the period $[t-1, t]$ which corresponds to $[\tau/\mathsf{m}-1, \tau/\mathsf{m}]$ where $\mathsf{m} = 12$. The standard multivariate MIDAS regression for explaining a stationary low-frequency variable y_t, augmented with a first order autoregressive component, is given by:

$$y_t = \beta_0 + \sum_{i=1}^{r} \beta_i \ m_K(\theta_i, L) \ \hat{f}_{i,t}^{(\mathsf{m})} + \lambda y_{t-1} + \varepsilon_t, \tag{2.5}$$

where $f_{i,t}^{(\mathsf{m})} = f_{i,\tau}$ is one of the exogenous stationary common factor sampled at a monthly frequency. The MIDAS function $m_K(\theta, L)$ controls the polynomial weights that allows the frequency mixing. Indeed, the MIDAS specification consists in smoothing the K past values of $f_t^{(\mathsf{m})}$ on which the regression is based. As in Ghysels et al. (2002), we implement the one parameter Beta lag polynomial such as

$$m_K(\theta, L) = \sum_{k=1}^{K} \frac{\theta k (1-k)^{\theta-1}}{\sum_{l=1}^{K} \theta l (1-l)^{\theta-1}} L^{(k-1)} \tag{2.6}$$

where L is the lag operator applied on the high frequency variable $x_t^{(\mathsf{m})}$ such that $L^s x_t^{(\mathsf{m})} = x_{\tau-s}$. In our setup we assume that the annual global growth is only influenced by the information conveyed by the last $K = 15$ values of the monthly factor $f_t^{(\mathsf{m})}$; the windows size K being exogenous. It can also be noticed that the parameter θ is part of the estimation problem. Other parameterizations of the weight function can be used, but we choose (2.6) since it constitutes a parsimonious and reasonable restriction for which the weights are always positive.

Parameter estimation of this model described by equations (2.4) and (2.5) is carried

in two steps. First, factors f_t are estimated using the static principal component analysis (see Stock and Watson, 2002). An eigenvalue decomposition of the estimated covariance matrix $\widehat{\Gamma}_0 = T^{-1}\sum_{t=1}^{T} X_t X_t'$ provides the $n \times r$ eigenvector matrix $\hat{S} = (\hat{S}_1 \ \ldots \ \hat{S}_r)$ containing the eigenvectors $\hat{S}_i$ corresponding to the r largest eigenvalues for $i = 1, \ldots, r$. The factor estimates are the first r principal components of X_t defined as $\hat{f}_t = \hat{S}' X_t$. Then, the MIDAS equation is estimated using standard non-linear least squares, assuming factors are known.

A tricky question arising within this kind of framework is related to the number of factors r to include in the equation (2.5). Several statistical tests are available in the econometric literature. In the forecasting framework, it turns out that some of them lead to more accurate forecasts, as shown in Barhoumi et al. (2013). Alessi et al. (2010) have suggested an information criterion based on Bai and Ng (2002) to determine the number of factors r in the context of an static factor analysis. This criterion can be written as:

$$\mathrm{IC}_p^T(r) = \log V(r, f) + c.r.p(n, T), \tag{2.7}$$

where $p(\cdot)$ is a penalty function defined as: $p(n,T) = \frac{n+T}{nT}\log\frac{nT}{n+T}$, and $V(\cdot)$ is a goodness-of-fit measurement based on sum of squared errors such as:

$$V(r, f) = (nT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{n}\left(X_t - \Lambda\hat{f}_t\right)^2 \tag{2.8}$$

which depends on the estimates of the static factors and on the number r of those factors. Following Alessi et al. (2010) and according to our modelling specifications, we set the exogenous parameters $c = 2$ and $r_{max} = 5$. The estimated number of factors r^* is defined as the one that minimises the criterium (2.7), as follows:

$$r^* = \underset{0 \le r \le r_{max}}{\arg\min}\ \mathrm{IC}_p^T(r). \tag{2.9}$$

The selected number of factors are therefore empirically used in (2.10) for global growth nowcasting purposes. The monthly nowcast of the annual global growth $\hat{y}_{t+1|t+1-h}$ is defined as the conditional expectation of y_t at a given month of the current year. For all forecasting horizon $h < \mathsf{m}$, the nowcasting estimate is computed using the following Factor-Augmented MIDAS equation:

$$\hat{y}_{t+1|t+1-h}(h) = \hat{\beta}_0(h) + \sum_{i=1}^{r^*}\hat{\beta}_i(h)\ m_K(\hat{\theta}_i(h), L)\ \hat{f}_{i,t+1-h}^{(\mathsf{m})} + \hat{\lambda}(h) y_t \tag{2.10}$$

where h is the forecasting horizon expressed in terms of the high frequency ranging from $h = 0$ months (corresponding to December) to $h = 11$ months (for January's forecasts). The equation (2.10) characterizes predictions of the current period involving new intermediary data of the explanatory variables using an update of the factors estimation $\hat{f}^{(\mathsf{m})}_{i,t+h}$. Besides that, the MIDAS parameters also are re-estimated at each horizon h via the non-linear least squares method. It is noteworthy that we allow parameters to depend on the forecasting horizon h.

2.2.2 Empirical Results

Database

Our methodology is based on a large data set gathering economic indicators from 37 countries, both advanced and emergingas described below.

- Advanced economies: France, Germany, Italia, Spain, Netherlands, United Kingdom, United States, Japan, Canada, Sweden, Switzerland, Norway, Denmark.
- Emerging Asia: China, India, Indonesia, South Korea, Taiwan, Thailand, Hong Kong, Malaysia, Singapore.
- Latin America: Brazil, Argentina, Mexico, Colombia.
- Europe: Poland, Czech Republic, Romania, Hungary, Latvia, Lituania, Bulgaria.
- Rest of the world: Russia, Turkey, South Africa, Saudi Arabia.

We can notice that the share of those countries is more than 80% of the world GDP as computed by the IMF WEO. From those 37 economies, we choose monthly variables suppose to convey useful information to assess short-term fluctuations of economic activity. Thus for each country we select a set of real variables (industrial production, household consumption, retail sales, new car registrations, etc.), financial variables (exchange rate, stock market indexes, interest rates, etc.) and household confidence index. The exhaustive list is the following:

- Real economic conditions: Housing, Car registrations, Retail sales, Employment, Industrial production index, Unemployment rate, Producer price index, Consumer price index.
- Financial Series: Exchange rate, Money supply M2, Main national stock market index, 10 years government bond interest rate, 3 months interbank interest rate.

- Survey: Household confidence index.
- Overall indicators: Oil price (Brent, WTI and Dubai), Baltic dry index, Import and export price (CPB), Energy price (HWWI), VIX index (CBOE).

Constraints are imposed on this choice, in the sense that we aim at having a similar set for each country and that we want to start our analysis in the early nineties. In addition, we augment this database using global indicators of trade, commodity prices, financial uncertainty, ... Overall we get a sample of $n = 392$ monthly variables. This database possesses the great advantage of being rapidly updated. All series are monthly and are expressed in difference or log-difference; the financial ones are sampled as the monthly average of daily quotes, and transformed in log-returns. The problem of ragged-edge series and unbalanced database is solved here by using the last available data as the contemporaneous one. This approach is referred to as the *realignement strategy* in the empirical litterature (see, for example, Marcellino and Schumacher, 2010).

Nowcasting methodology

Using principal component analysis defined in equation (2.4), we extract one monthly factor that describes variability of the whole dataset. The various implemented tests on the number of factors to select led us to choose $r = 1$. The estimated factor is displayed and compared to yearly global growth WEO estimates in Figure 2.6. It is noteworthy that this first factor seems to follow quite closely global growth fluctuations, in spite of some deviations during specific periods of time. The idea is now to formally relate this estimated factor to the global growth through the MIDAS equation (2.5). The targeted variable is the world GDP growth rate provided by the IMF in its April 2014 WEO and presented in Figure 2.4.

In a first step, we carry out an in-sample analysis over the period from January 1995 to December 2009. Knowing that financial data are available the last working day of the month, we suppose that nowcasts for a given month are computed at the end of each month, for 12 horizons ranging from $h = 0$ (nowcasts computed the last month of the reference year) to $h = 11$ (nowcasts computed 11 months before the end of the reference year). For each date t, the MIDAS regression optimally exploits the monthly fluctuations of the last $K = 15$ data of the $f_t^{(\mathsf{m})}$ series using the weight polynomial, given in equation (2.6). Estimated weights are presented in Figure 2.7. The shape of the weights is in line with what we could expect according to the forecasting horizon. Indeed for long horizons, the shape gives a non-null value to all the weights until

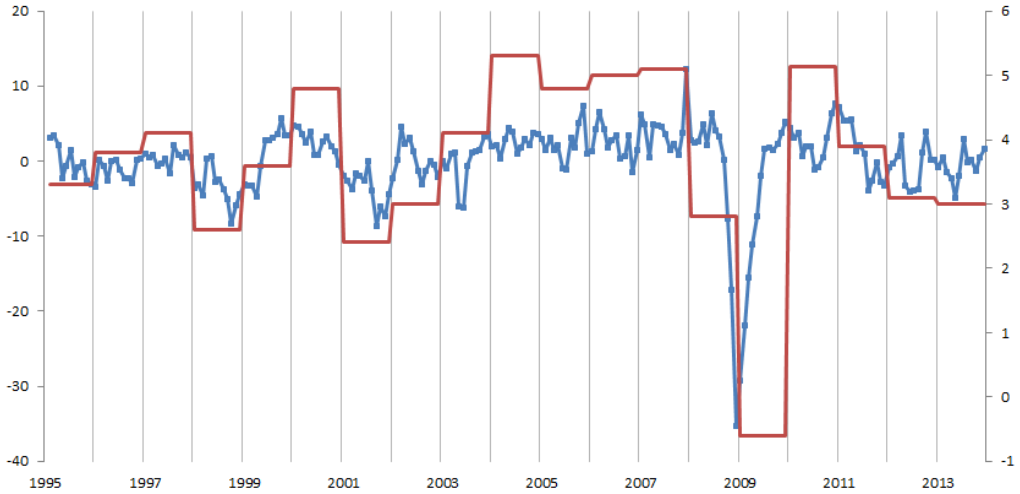


FIGURE 2.6: Explanatory factor vs. WEO Global growth estimates.

$K = 15$. When the horizon shorten (e.g. for $h = 6$), the shape is more peaked and the maximum value is reached for $k = 2$. Finally when $h = 0$, that is when the nowcast is made in December of the current year, the mass is mainly concentrated in $k = 0$ and the function rapidly decreases.

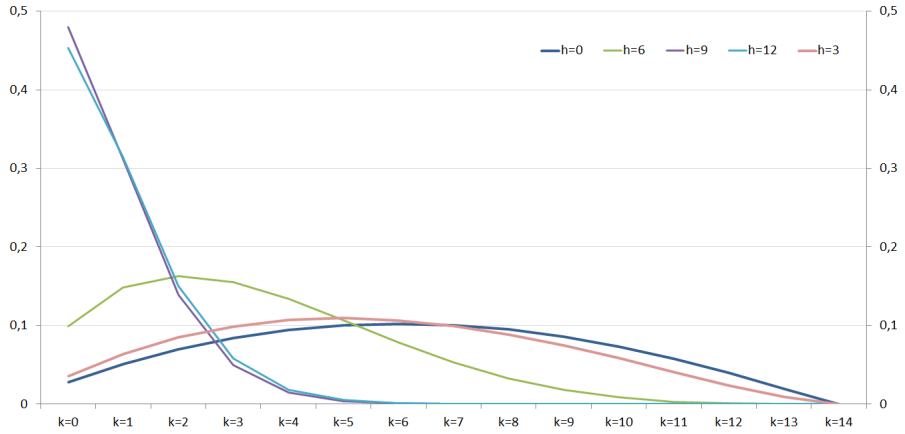


FIGURE 2.7: In-sample MIDAS weight functions with respect to the forecasting horizon h.

In a second step, we implement a quasi-real-time experience over the post-crisis period from January 2010 to December 2013. For each month, we estimate the global growth of the current year and we compare it with the real-time estimates stemming from the four IMF-WEO reports released per year. In practice, we do not re-estimate all the parameters each month, but instead we use the parameters estimated using the information until December of the previous year. Empirical results are presented in Figure 2.8, as well as final estimates as released with the April 2014 IMF-WEO. As expected, real-time estimates tend to convergence to the final figures, which is con-

sistent with the fact that more information leads to more accurate estimates. Our estimate evolves with the monthly flow of conjunctural information that we received within the year, while the IMF-WEO estimates is more related to the release of quarterly national accounts. We note that in 2010, the IMF-WEO largely underevaluated the bounce-back in world GDP growth, especially at the beginning of the year, while our nowcast fluctuated around the final figure. We also note that since 2011, both estimates were generally revised in the same direction.

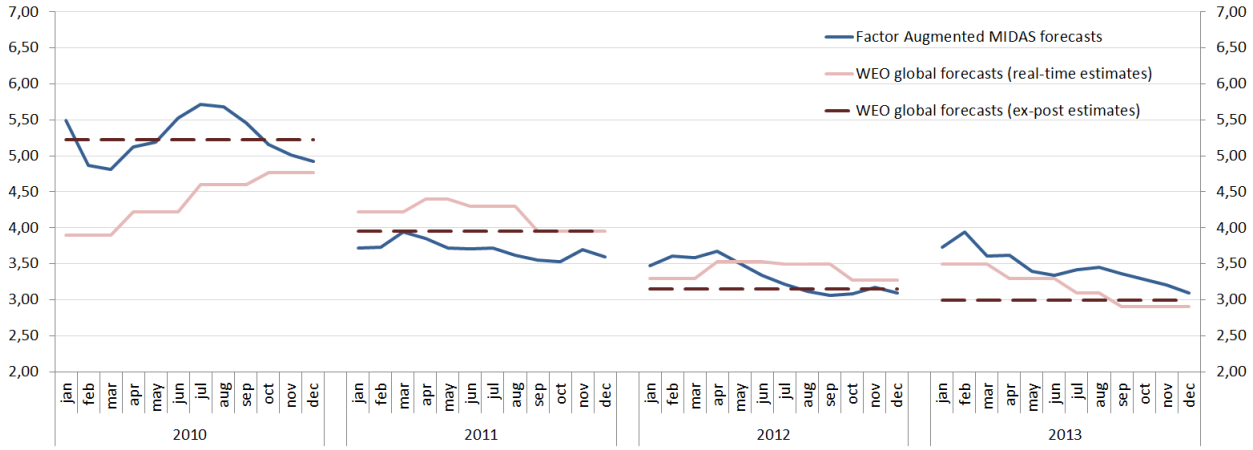


FIGURE 2.8: WEO vs FAMIDAS nowcasts over the period 2010-2013

In addition to nowcasts, we also develop non-parametric bootstrapping technique *a la* Efron (1979) in the MIDAS regression context to get confidence intervals around nowcasts and hence a measure of the uncertainty. The methodology involves random resampling, with replacement, of elements from the original data to generate a replicate data vector of similar size[2.5]. This kind of approach has been already used by Aastveit et al. (2014) for density forecasts and by Clements and Galvão (2008) for significance tests. The 90% confidence intervals are exhibited in Figure 2.9. Eyeballing the figure suggests that the uncertainty was shifted downward since the year 2010 and seems to remain broadly constant. But some periods of time present larger confidence intervals, sometimes with an asymmetry pointing out that risks are tilted to the downside (or to the upside).

In order to evaluate the accuracy of our approach, we compute the squared errors of the nowcasts stemming from both the WEO and the FA-MIDAS model. Monthly averages of squared errors over the period 2010-2013 are showed in Figure 2.10. Overall, the FA-MIDAS model provides more accurate nowcasts over the year and are equivalent

[2.5] Our bootstrapping methodology has been explicitly described and empirically assessed in Section 1.3.2.

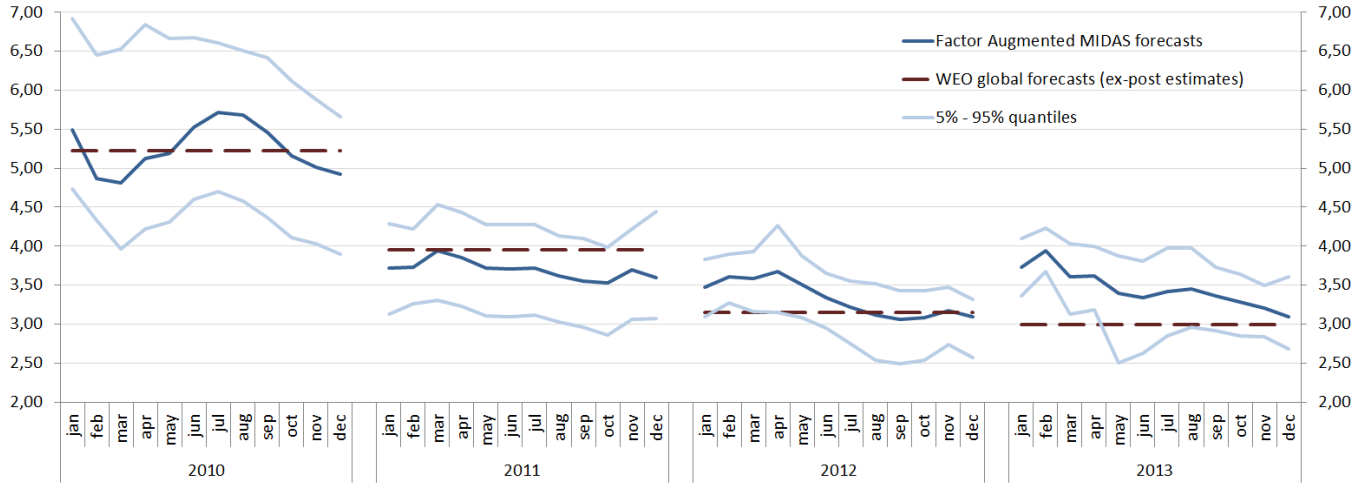


FIGURE 2.9: FAMIDAS nowcasts with confidence interval (5%-95%) over the period 2010-2013

to the WEO forecasts by the end of the year, from October to December. In fact, we notice that the forecasting gain obtained by using the FA-MIDAS is particularly important from 12 to 4 months ahead, that is from January to September. Indeed, at the the beginning of the year, the information available to the WEO update of January is rather scarce. Also, when economists are working on the preparation of the April WEO, they do not have at hand the realized GDP for the first quarter of the current year. Similarly, the release of the second quarter of GDP growth occurs well after the July update. Consequently, it seems that our tool could consitute a nice complement to the WEO estimates for economists interested in monitoring the world economic growth.

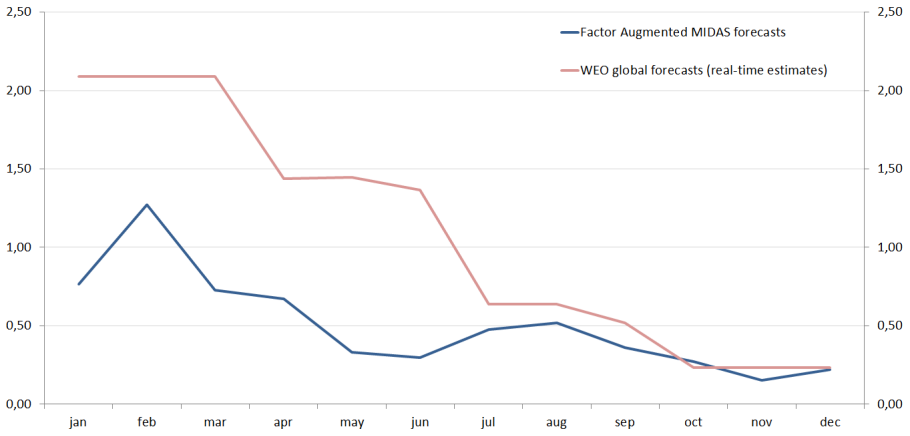


FIGURE 2.10: Monthly averages of Mean Square Errors of WEO and FA-MIDAS nowcasts

Conclusion

In this section, we put forward a new tool in order to nowcast the global economic growth in real-time. We implement a Factor Augmented MIDAS approach enabling to explain the annual global growth by a large database of monthly variables. The targeted variable is the annual global growth estimated by the IMF in its World Economic Outlook assessment. It turns out that our tool is able to efficiently track on a high frequency the global growth. Especially nowcasts are much more accurate at the beginning of the year when fewer information is available. This tool could be fruitfully used by macroeconomists to monitor global economic developments, in addition to the IMF-WEO estimates.

Chapter 3

Bayesian inference on MIDAS model

Treating the information regardless of the sampling frequency leads to obtaining empirical gains in terms of forecasting accuracy and to capturing relevant unknown data features. The MIDAS-based modeling approach we implement in Chapter 2 allowed us, for instance, to point out the predictive power of financial volatility to anticipate GDP growth during the specific period of the global economic recession. This chapter aims at developing a Bayesian approach in the context of MIDAS regression problems (Section 3.1) in order to provide a flexible framework to investigate some well-known macroeconomic features. More specifically, we extend the MIDAS regression model by allowing for stochastic volatility in the data (Section 3.2).

3.1 Bayesian MIDAS model

As we have seen in Chapter 1, the nonlinear MIDAS estimator is consistent. That means that the average value of the sample estimates converges to the unknown value of the parameter as the sample size increases. However, a bootstrapping exercise pointed out a difference in variance: the variance of the bootstrapped NLS estimator remained higher than the theoretical asymptotic estimate, meaning that convergence requires a very large sample size. Introducing a Bayesian procedure may allow us to learn more about values and distributions of MIDAS parameters. Indeed, the Bayesian approach has been applied successfully to a wide range of econometrics problems. The works of Tsurumi, Park, Gao, Lahiri, and Zellner showed the good performance of various Bayesian estimation procedures by contrast with that of leading non-Bayesian estimation methods. The Bayesian approach is particularly able to deal with common problems of nonlinear models due to either the flatness of the likelihood function or the existence of local minima.

Rodriguez and Puggioni (2010) have recently exploited the Bayesian model selection approach to estimate MIDAS parameters. They particularly investigated the problem of collinearity of intraperiod observations by combining lag size specified-model (each MIDAS model corresponds to a specific lag size K) with a Bayesian model selection strategy. In their framework, the issue relative to the nonlinear weight function of the MIDAS model is avoided. In fact they rather consider a linear distributed lag model jointly with a factor analysis to deal with parameter proliferation. Recently Carriero et al. (2012) and Marcellino et al. (2013) have also exploited Bayesian methods to estimate mixed frequency model with stochastic volatility developed for forecasting purposes. Their approach are most closely related to the U-MIDAS specification of Foroni et al. (2013) described in Section 1.3.3.

In this section, we introduce the Bayesian MIDAS model by using a general framework for model comparison. We suppose that each model corresponds to a particular choice of variables. Seeking the most relevant model, Zellner (1971) proposed to compare hypotheses using Bayesian analysis and the ratio of posterior probabilities associated with each of the hypotheses tested. In this context, the Bayesian strategy we advocate has some theoretical and practical advantages for predicting purposes. First, Bayesian techniques for model selection allow multiple comparisons and hence let us explore full model space efficiently. Second, it is straightforwardly related to Bayesian model averaging that provides the optimal forecasting procedure according to Palm and Zellner (1992). That topic will be of interest of future empirical research. We

discuss prior specifications and precisely investigate the estimation procedure using a Markov Chain Monte Carlo method. We specially develop a generic and flexible algorithm based on a combined use of the Metropolis Hastings and the Gibbs sampler algorithms. We illustrate our approach by proposing a simulation exercise comparable to that in Chapter 1, and empirically assess it in the second Section of the Chapter.

3.1.1 Bayesian setup

The Bayesian learning model is based on the Bayes theorem that relies on initial information regarding possible values of the parameters summarized in a *prior* probability density function and on the likelihood function that represents the current dataset information. We combine those elements using the Bayes theorem in order to obtain a *posterior* distribution for the parameters that contains both the prior and the sample information. We use this strategy in the context of MIDAS models by setting a generic model review framework.

Bayesian model selection

As we have seen in the previous two empirical examples in Chapter 2, the quality of any forecasting methodology depends on the choice of the explanatory variables. Any underlying technical specifications involving mixed frequency modeling should provide a way to assess predictive content of the data set. Our Bayesian analysis relies on a generic framework allowing model comparison[3.1]. We consider a collection of models $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_R\}$. Each of those models corresponds to a specific subset of the n variables collection. Thus, we have $R = 2^n$ and $\mathcal{M} = \{\mathcal{M}_r,\ r \in \{1, \ldots, 2^n\}\}$. We define $\xi_i^{(r)}$ as an dummy indicator of the presence of the i^{th} β coefficient in the subset $\beta_{(r)}$ of the model $\mathcal{M}_r$:

$$\xi_i^{(r)} = \begin{cases} 1 & \text{if } \beta_i \in \beta_{(r)} \\ 0 & \text{otherwise} \end{cases}$$

The Bayesian approach to model selection requires computing the probability $\pi(\mathcal{M}_r|Y)$ meaning that *the model $\mathcal{M}_r$ is the correct model, given the data.* For example, in the case where we only have two models, $\pi(\mathcal{M}_1|Y) = 1 - \pi(\mathcal{M}_2|Y)$. We compute those

[3.1] We particularly focus on the issue of variable and model selection in the context of mixed-frequency models in Chapter 4.

probabilities using the Bayes formula:

$$\pi(\mathcal{M}_r|Y) = \frac{F(Y|\mathcal{M}_r) \times \pi(\mathcal{M}_r)}{F(Y)}, \tag{3.1}$$

where $F(Y|\mathcal{M}_r)$ is the marginal likelihood, $\pi(\mathcal{M}_r)$ is the prior of the model $\mathcal{M}_r$ and $F(Y)$ is the full unconditional likelihood, a constant that normalizes the posterior distribution and that can be ignored for convenience since it does not depend on the model specification. Thus we write the posterior as:

$$\pi(\mathcal{M}_r|y) \propto F(Y|\mathcal{M}_r) \times \pi(\mathcal{M}_r). \tag{3.2}$$

Since we assume equal probabilities $\pi(\mathcal{M}_r)$ for all r meaning that we do not discriminate any model, the probabilities $\pi(\mathcal{M}_r)$ can be ignored in (3.2), and therefore we have: $\pi(\mathcal{M}_r|y) \propto F(y|\mathcal{M}_r)$. Another approach would be to compute the posterior distribution in (3.2) using Bernoulli distributions as prior probabilities $\pi(\mathcal{M}_1), \ldots, \pi(\mathcal{M}_{2^n})$ on each model. That implies the use of an exogenous parameter $\eta \in [0, 1]$ which set independently $\pi(\xi_i) = \eta^{\xi_i}(1-\eta)^{1-\xi_i}$ for $i = 1, \ldots n$. Let denote Ξ_r the number of non zero $\xi^{(r)}$, i.e. the number of variables involved in the model $\mathcal{M}_r$. Hence, we have

$$\pi(\mathcal{M}_r|\eta) = \eta^{\Xi_r}(1-\eta)^{n-\Xi_r}. \tag{3.3}$$

While Bernoulli priors *treat all variables equally*, the parameter η controls the sparsity in the model. The variable η can be involved in the Bayesian estimation using a hyperprior $\pi(\eta)$. We use the Beta distribution, $\eta \sim \mathcal{B}e(\tilde{a}, \tilde{b})$, where $\frac{\tilde{a}}{\tilde{a}+\tilde{b}}$ is an *a priori* on the proportion of selected variables in the subset. That strategy specifying Bernoulli type priors is widely used in Bayesian model selection, we refer to the stochastic search variable selection technique put forward by George and McCulloch (1993) and developed within a MIDAS model for forecasting purposes in Chapter 4.

In the case of Bayesian Model Averaging, usually referred to as BMA, model uncertainty is taken into account using a weighted averaging across all possible models according to their posterior probability. As a remark, we also note that ratios of posterior probabilities rewritten as posterior odds, $\frac{\pi(\mathcal{M}_{s_1}|y)}{\pi(\mathcal{M}_{s_2}|y)}$, summarized the evidence in favor of one model through the Bayes factor $\frac{\pi(y|\mathcal{M}_{s_1})}{\pi(y|\mathcal{M}_{s_2})}$ (see for example Draper (1995) for a review). Jeffreys (1961) proposed some guidelines for the interpretation of such pairwise comparisons. However, it can be noticed that these Bayesian selection approaches (BMA, Bayes factors, posterior odds ratios, etc.) require the marginal likelihood $F(Y|\mathcal{M}_r)$. Since working with nonstandard distribution or with flat prior relative to the likelihood, integrating marginal calculation in the Monte Carlo is gen-

erally unstable and inaccurate. In this context, several methods have been developed to estimate marginal likelihood (see Gelfand and Dey, 1994 or Chib, 1995). Those techniques are of particular interest for forecasting purposes and will be examined in further research.

The marginal likelihood can be interpreted as the expected value of the likelihood function with respect to the prior distribution. Thus, let $\phi_{(r)}$ be the parameters vector of the model $\mathcal{M}_r$, we integrate the marginal likelihood $F(Y|\mathcal{M}_r)$ with respect to its parameter $\phi_{(r)}$:

$$\begin{aligned} F(Y|\mathcal{M}_r) &= \int F(Y, \phi_{(r)}|\mathcal{M}_r)\, \mathrm{d}\phi_{(r)} \\ &= \int F(Y|\phi_{(r)}, \mathcal{M}_r)\, \pi(\phi_{(r)}|\mathcal{M}_r)\, \mathrm{d}\phi_{(r)} \end{aligned}$$

Thus, the marginal likelihood is given by:

$$F(Y|\mathcal{M}_r) = \frac{F(Y|\mathcal{M}_r, \phi_{(r)})\ \pi(\phi_{(r)}|\mathcal{M}_r)}{\pi(\phi_{(r)}|Y, \mathcal{M}_r)} \tag{3.4}$$

MIDAS specifications

In Chapter 1, we defined the standard MIDAS model and introduced the case of multiple explanatory variables whose equation can be written in the matrix form as $Y = \mathsf{X}(\theta)\beta + \varepsilon$, where the matrix $\mathsf{X}(\theta)$ is denoted in (1.30) and the residuals $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_T)'$ are independent and assumed to be Gaussian: $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. The regression model depends on the parameters family $\phi = \{\beta, \theta, \sigma^2\}$.

According to Bayes' formula (3.1), the model selection relies on an estimate of the posterior ordinate $\pi(\phi_{(r)}|Y, \mathcal{M}_r)$. Since we now focus on parameter estimates, the posterior $\pi(\phi_{(r)}|Y, \mathcal{M}_r)$ is denoted $\pi(\beta, \theta, \sigma^2|Y)$ to simplify the notation. The parameter family $\{\beta, \theta, \sigma^2\}$ obviously corresponds to $\mathcal{M}_r$ specifications. Thus, the posterior can be written as

$$\pi(\beta, \theta, \sigma^2|Y) \propto F(Y|\beta, \theta, \sigma^2) \times \pi(\beta, \theta, \sigma^2). \tag{3.5}$$

The marginal likelihood is required for model selection as prescribed in (3.4). We ignore it for the time being. The likelihood function has already been described in

(1.16) and can be written as follows:

$$\begin{aligned} F(Y|\beta,\theta,\sigma^2) &\equiv F(Y|\mathcal{M}_r,\phi_{(r)}) \\ &= \frac{1}{(2\pi\sigma^2)^{T/2}}\exp\left(-\frac{1}{2\sigma^2}\left(Y-\mathsf{X}(\theta)\beta\right)'\left(Y-\mathsf{X}(\theta)\beta\right)\right) \\ &= \mathcal{N}(Y-X(\theta)\beta,\sigma^2 I_T) \end{aligned}$$

Assuming that the parameters β, θ, and σ^2 are unknown, we define priors for each of these parameters. Considering the MIDAS model as a standard linear regression model that involves a nonlinear kernel which enables mixing frequencies, we use conjugate priors which provide a posterior distribution coming from the same family as the prior for the linear parameters. In fact, we suppose that priors for the regression parameters β are normally distributed and use Gamma prior for the MIDAS lag polynomial coefficients both θ_1 and θ_2 as suggested by Ghysels (2012). More specifically, regarding prior specifications for θ, we use the Jeffreys prior which is defined by:

$$\pi(\theta) \propto \sqrt{\det \mathcal{I}(\theta)}. \tag{3.6}$$

The Jeffreys prior is a non-informative prior distribution proportional to the square root of the determinant of the Fisher information $\mathcal{I}(\theta)$[3.2]. That specification yields a non tractable posterior distribution which requires the use of a Monte Carlo Markov Chain (MCMC) method to be sampled. That is described in the following section.

3.1.2 Estimation using MCMC

The full set of conditional distributions is not available, a direct sampling of the posterior distribution is therefore unachievable. In order to estimate the MIDAS model, we implement a Gibbs sampler with respect to specific features of the mixed data sampling framework. In fact, the algorithm relies on a few steps which successively draw β and σ^2 from the Normal-Inverse Gamma prior, and θ from a candidate generating density using an independence chain Metropolis-Hastings (iMH) algorithm.

iMH within Gibbs sampler

Given initial values for all unknown parameters, the algorithm iteratively updates their values by drawing from their conditional distribution and hence constructing a

[3.2] While $\ln F(Y|\beta,\theta,\sigma^2)$ is twice differentiable with respect to θ.

Markov chain with an invariant distribution. The algorithm is constructed as follows:

1. Initialize β, θ and σ^2.

2. Sample $\beta|\theta, Y$ from $\mathcal{N}(\hat{\beta}, \hat{B})$

$$\pi(\beta|\theta, y) \propto \exp\left(-\frac{1}{2}(\beta - \hat{\beta})'\hat{B}^{-1}(\beta - \hat{\beta})\right) \tag{3.7}$$

   where $\hat{B} = \left(B_0^{-1} + \frac{\mathsf{X}(\theta)'\mathsf{X}(\theta)}{\sigma^2}\right)^{-1}$, and $\hat{b} = \hat{B}\left(B_0^{-1}b_0 + \frac{\mathsf{X}(\theta)'y}{\sigma^2}\right)$.

3. Sample $\sigma^2 \sim \mathcal{IG}(s_n, S_n)$

   where $s_n = s_0 + \frac{T-1}{2}$, and $S_n = \frac{1}{2}(Y - \mathsf{X}(\theta)\beta)'(Y - \mathsf{X}(\theta)\beta)$.

4. Sample $\theta|\beta, \sigma^2, Y$ using independence chain Metropolis-Hasting step within the Gibbs sampler.

   The acceptance probability $\tilde{\alpha}$ to change to the new value θ^{new} drawn from the candidate density, determines whether the chain moves from areas of low posterior probability to high. It can be written as:

$$\tilde{\alpha} = \min\left[\frac{\pi(\theta = \theta^{\text{new}}|y)}{\tilde{\iota}(\theta = \theta^{\text{new}})}\frac{\tilde{\iota}(\theta = \theta^{\text{old}})}{\pi(\theta = \theta^{\text{old}}|y)}, 1\right].$$

   To define the candidate generating density $\tilde{\iota}$, we use an approximation based on the asymptotic normality of the maximum likelihood estimator $\hat{\theta}_{ML}$, and on its asymptotic variance-covariance matrix $var(\hat{\theta}_{ML}) = \mathcal{I}(\theta)^{-1}$. We compute the Fisher information matrix $\mathcal{I}(\theta) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\log f(Y|\beta, \theta, h)\right)$, using numerical differentiation procedures to obtain the approximate variance: $\widehat{var(\hat{\theta}_{ML})}$. Thus, we set the candidate generating density as $\iota(\theta) = f_T(\theta|\hat{\theta}_{ML}, \widehat{var(\hat{\theta}_{ML})})$ since we approximate the posterior by a multivariate normal distribution with mean $\hat{\theta}_{ML}$ and covariance matrix $\widehat{var(\hat{\theta}_{ML})}$.

   Draw $u \sim \mathcal{U}(0, 1)$. If $u < \alpha$, retain the new candidate by setting $\theta = \theta^{\text{new}}$, otherwise $\theta = \theta^{\text{old}}$.

5. Repeat J times steps 2 to 4.

Repeating a certain number of times the steps 2, 3 and 4 yields the chain to converge to a steady state. The algorithm generates a sample $\{\beta^{(j)}, \theta^{(j)}, \sigma^{2(j)}\}_{j=1}^{J}$ which is a sample of the posterior distribution.

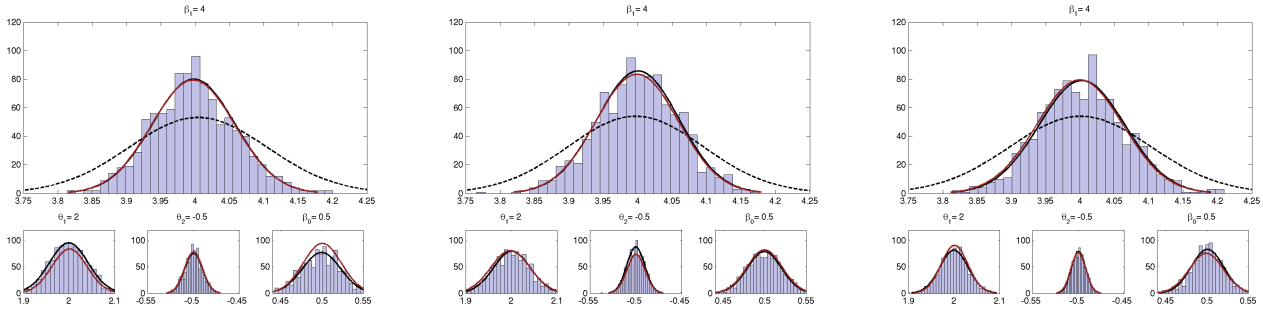
The algorithm can be easily extended to the Bayesian model selection approach we previously set by incorporating a model probability sample step and possible hyper-parameters sample steps. Then the marginal likelihood can be calculated from the sample of the posterior distribution using, for instance, the Chib and Greenberg (1995) method. Recent works have proposed new algorithms automatically generating the proposal of the Metropolis and running multiple chains in parallel. We especially refer to the DEMC algorithm of ter Braak and Vrugt (2008) and to the DREAM algorithm of Vrugt et al. (2009). Those can be adapted to the MIDAS estimation problem and will be the subject of future research.

Monte Carlo simulations

We assess our Bayesian MIDAS approach via a simulation study and an empirical application similar to those done in Section 1.3.2. For this purpose, we use a DGP based on a univariate MIDAS regression model (1.25) where the dependent variable is an AR(1) process, the weight function is the exponential Almon lag polynomial (1.26), and the innovations are normally distributed. In Chapter 1, we compared the NLS MIDAS estimates with the asymptotic estimator using 1000 simulated models. Here, we focus on one DGP[3.3] for which we compare both empirical parameters distributions obtained using either the posterior distribution sample of the Bayesian estimation approach described above, or the bootstrapping procedure of the nonlinear least square estimate, with the asymptotical distribution of the NLS estimate given by Andreou et al. (2010) and described in 1.3.2. Those are presented in Figure 3.1.

We observe that both Bayesian and bootstrapped distributions are basically equivalent in terms of variance of the coefficient β_1. Nevertheless, we notice that the variance of the bootstrapping parameter estimates for θ_1, θ_2 and β_0 are generally smaller than the variance of their respective Bayesian estimates. The relative efficiency of the two estimators improves as T and κ increase. We also note that the convergence of both Bayesian and bootstrapping distributions towards the asymptotic estimator density necessitates a sample of considerable size ($T > 500$). A similar behavior have already been observed in the previous simulation exercise. We now implement this exercise in an empirical assessment based on macroeconomic data.

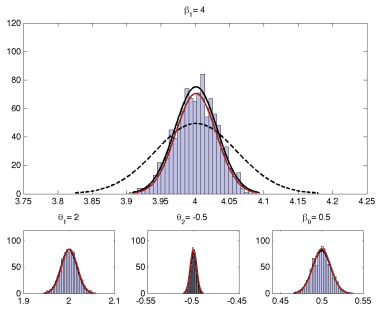
[3.3] Model specifications and parameters value are the same than those used in Section 1.3.2.



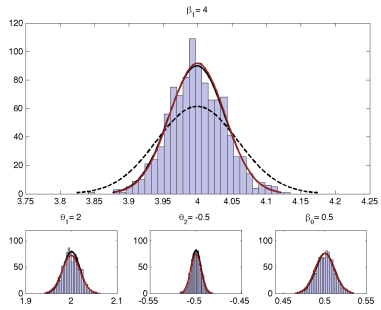
(a) $\kappa = 1/3$, $T = 20$

(b) $\kappa = 1/12$, $T = 20$

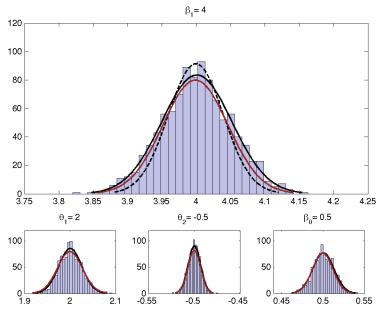
(c) $\kappa = 1/22$, $T = 20$



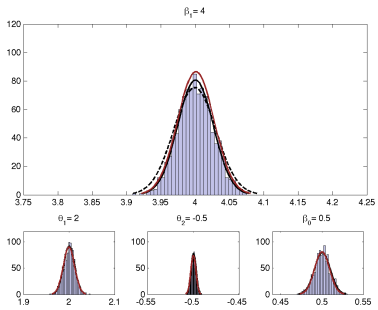
(d) $\kappa = 1/3$, $T = 50$



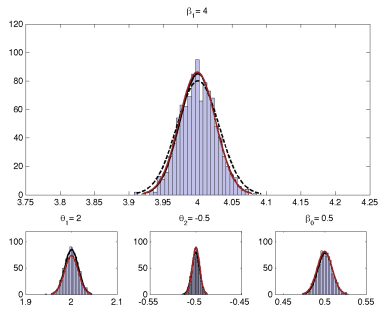
(e) $\kappa = 1/12$, $T = 50$



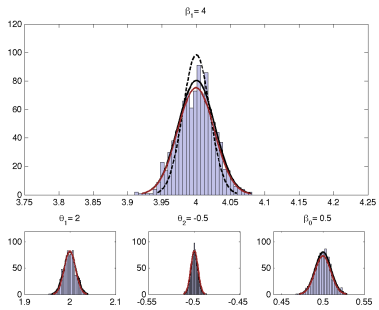
(f) $\kappa = 1/22$, $T = 50$



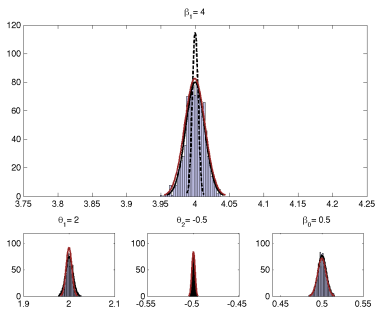
(g) $\kappa = 1/3$, $T = 100$



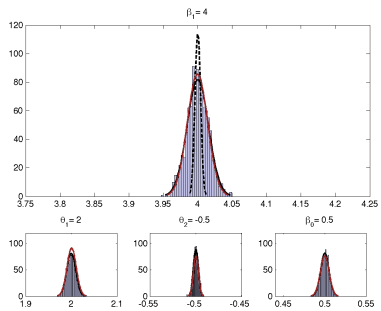
(h) $\kappa = 1/12$, $T = 100$



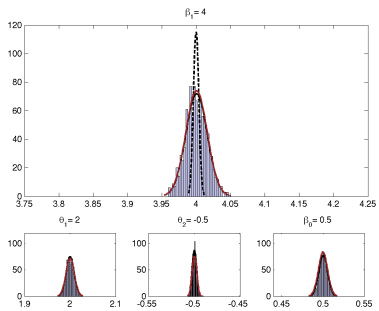
(i) $\kappa = 1/22$, $T = 100$



(j) $\kappa = 1/3$, $T = 500$



(k) $\kappa = 1/12$, $T = 500$



(l) $\kappa = 1/22$, $T = 500$

FIGURE 3.1: Bayesian, bootstrapping and asymptotical distributions of MIDAS estimates across different sample sizes T and different aggregation horizons κ.
The histogram depicts the bootstrapping draws and the solid black line is its normal distribution fit. The solid red line is the normal density fit of the Bayesian posterior distribution sample, the dashed black line represents the asymptotical distribution of the NLS estimator.

Empirical assessment

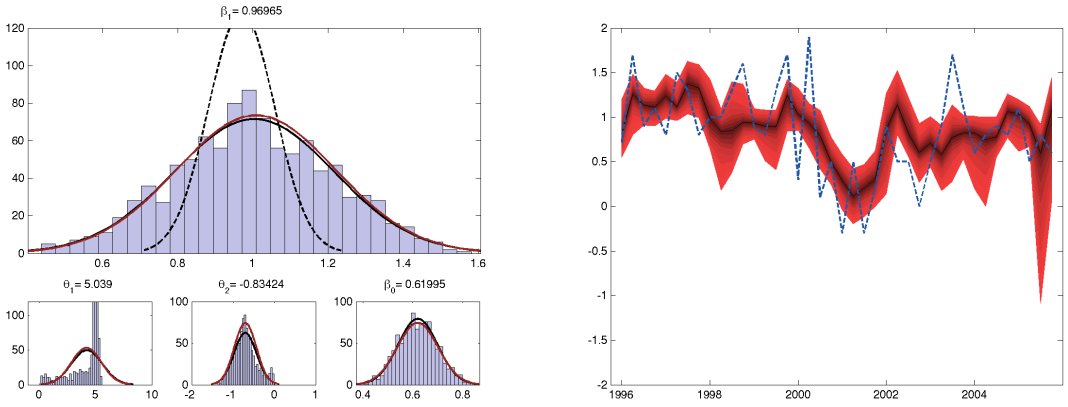
We evaluate the Bayesian estimation approach within a well-known macroeconomic regression problem involving data sampled at various frequencies; that is explaining the US quarterly GDP using the US monthly IPI. The model equation is given in (1.28) and the design of this empirical exercise is described in Section 1.3.2. The Bayesian MIDAS model is compared with both the (*frequentist*) standard approach, whose parameters are bootstrapped, and the theoretical asymptotic distribution of the NLS estimator. We use two different sample sizes $T = 40$ and $T = 100$ which correspond, respectively, to the period from 1996:q1 to 2005:q5, and 1989:q1 to 2013:q4. The results are displayed in Figures 3.2 and 3.3.

We notice that the Bayesian MIDAS model provides a very good fit of the GDP growth rate (MSE are 0.16 and 0.19 for the small and the large sample, respectively) and catches well downturns of the recent Great Recession period. Despite slight differences in the parameters distributions (the contrast between the standard NLS approach and the Bayesian estimation can be noticed in the left-hand side Figures 3.2a and 3.3a), these in-sample results are comparable, and quite similar in terms of accuracy, to those obtained in Chapter 1 with the standard NLS approach. We compute confidence intervals of parameter β_1 according to the three estimation methods we put forward; those are summarized in Table 3.1.

Sample size	$\hat{\beta}_1$	Bayesian CI	Bootstrap CI	Asymptotic CI
$T = 40$ (1996:q1-2005:q4)	0.97	[0.72; 1.21]	[0.65;1.34]	[0.90;1.04]
$T = 100$ (1989:q1-2013:q4)	1.07	[0.91; 1.22]	[0.91;1.27]	[1.01;1.12]

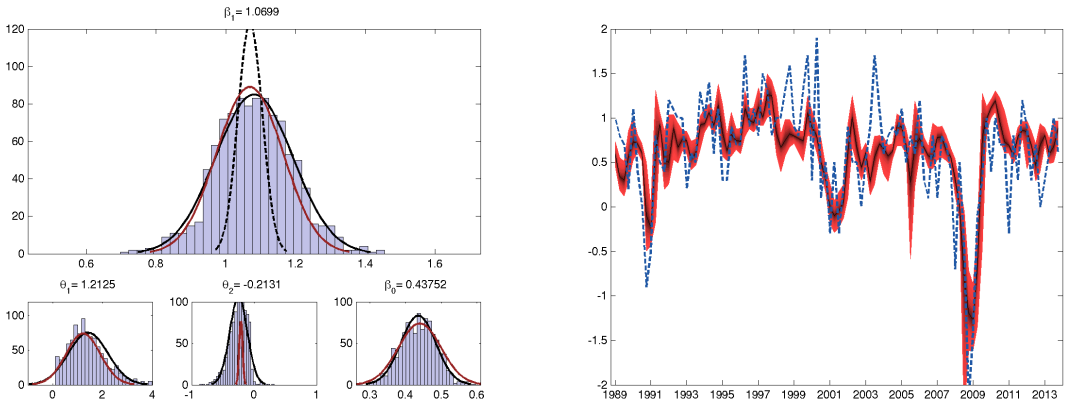
Table 3.1: Slope coefficient estimates and its relative confidence interval (CI) at 95% for two different sample sizes.

We note that the above results are consistent with those obtained in Chapter 1. As we have seen in the simulation exercise, the Bayesian estimation corresponds to the bootstrapping results: the median of the Bayesian posterior distribution of β_1 is 0.96 while the NLS estimate is 0.97. Moreover, we see that the two confidence intervals are almost similar. Those results are coherent with the statistical nature of the bootstrapping method. In fact, as described by Hastie et al. (2009), "*the bootstrap distribution represents an (approximate) nonparametric, noninformative posterior distribution for our parameter. But this bootstrap distribution is obtained painlessly – without having to formally specify a prior and without having to sample from the posterior distribution*". Our exercise empirically illustrates this fact.



(a) (b)

FIGURE 3.2: Regression results over the period 1996:q1-2005:q4 ($T = 40$), MSE=0.16.



(a) (b)

FIGURE 3.3: Regression results over the period 1989:q1-2013:q4 ($T = 100$), MSE=0.19.
On the left-hand side (3.2a and 3.3a), the dashed black line corresponds to the theoretical asymptotic distribution, the histogram represents the bootstrapping draws and the solid black line is its normal distribution fit while the solid red line is the normal density fit of the Bayesian posterior distribution sample. On the right-hand side (3.2b and 3.3b), the dashed blue line is the observed values of GDP_t while the red fan chart depicts the Bayesian distribution of the fitted series $\widehat{GDP}_t$.

3.2 A MIXED-FREQUENCY MODEL WITH STOCHASTIC VOLATILITY

This section is based on a work-in-progress with Laurent Ferrara and Massimiliano Marcellino, entitled "A mixed-frequency model with stochastic volatility".

Since the onset of the Great Recession, financial variables have been largely reconsidered in econometric models, mainly to explain growth fluctuations. Due to the inherent nature of the data involved in such empirical analyses – growth being measured using quarterly GDP growth rates while financial information being available on a high-frequency basis – several mixed-frequencies econometric models have been developed in the literature to account for this stylized fact. Some applications have showed that an extended MIDAS approach using daily financial variables increases the predictive accuracy of quarterly GDP growth in various industrialized countries (see e.g. Andreou et al., 2013 or Ferrara et al., 2014[3.4]). Some recent works also put forward nonlinear extensions of the MIDAS framework to account for an asymmetric macro-financial relationships along the business cycle (see e.g. Guérin and Marcellino, 2013)

Many recent research papers point out the interest of heteroscedatic models to account for the volatility of financial variables. For example, Clark (2011) has recently showed that compared to models with constant variances, models that include stochastic volatility improve real time density forecasts from a VAR approach. In the context of mixed-frequency models, Carriero et al. (2012), using an unrestricted MIDAS model Foroni and Marcellino (2013b), and Marcellino et al. (2013), using a dynamic factor analysis, have also showed that the stochastic volatility specification is useful to obtain reliable density forecasts and is comparable to other usual forecasting models in terms of point forecasts.

In this section we put forward an extended MIDAS model that integrates stochastic volatility (MIDAS-SV hereafter) enabling to account for financial volatility. The model extends both Andreou et al. (2013), by allowing for stochastic volatility, and Carriero et al. (2012), by allowing for a standard MIDAS specification. The proposed framework is convenient since it limits parameter proliferation and allows modeling even in the presence of high frequency mismatch between the dependent and the explanatory variables. The Bayesian approach we developed simultaneously estimates

[3.4] Section 2.1 is based on this research paper.

all the parameters of this new MIDAS-SV model[3.5]. We assess its performance in a empirical exercise that evaluates the role of daily financial variables for nowcasting quarterly US GDP growth.

3.2.1 The MIDAS-SV model

We base our analysis on the standard MIDAS regression model using the notation that we defined in Chapters 1 and 2. We extend the standard MIDAS model by relaxing the homoscedastic hypothesis. We put forward a MIDAS regression model augmented with a stochastic volatility specification, that can be written as:

$$y_t = \beta_0 + \beta_1 \, m_K(\theta, L) \, x_t^{\kappa} + v_t, \tag{MIDAS}$$
$$v_t \sim \mathcal{N}(0, \mathsf{h}_t) \tag{SV}$$

$$\text{where } \ln \mathsf{h}_t = \lambda_0 + \lambda_1 \ln \mathsf{h}_{t-1} + \eta_t \text{ and } \eta_t \sim \mathcal{N}(0, \psi^2). \tag{3.8}$$

The model given in those three previous equations is referred to as the MIDAS-SV model. Against this background, the disturbance v_t of the standard MIDAS regression model follows a Gaussian distribution whose variance is time-varying. Especially, the log-volatility, $\ln \mathsf{h}_t$, is commonly assumed evolving as a random walk process that may be viewed as a limiting case of an AR(1) process where $\lambda_0 = 0$ and $\lambda_1 = 1$ in equation (3.8). This type of volatility specification has been popularized in macro-econometrics by Cogley and Sargent (2005) and Primiceri (2005), see also Clark (2011). Recently Carriero et al. (2012) and Marcellino et al. (2013) have used mixed frequency stochastic volatility models for forecasting purposes.

Parameter estimation

In order to estimate the MIDAS-SV model, we adopt a Bayesian approach in which we specify the prior density $\pi(\phi)$ for the vector of all unknown parameters $\phi = \{\beta, \theta, \lambda, \psi^2\}$ where $\beta = \{\beta_0, \beta_1\}$ and $\lambda = \{\lambda_0, \lambda_1\}$, the likelihood function $F(Y|\phi)$ and the posterior distribution which, according to the Bayes formula, is given by:

$$\pi(\phi|Y) = \frac{F(Y|\phi) \times \pi(\phi)}{F(Y)}.$$

[3.5] Our formulation also allows the regression coefficients to be time-varying.

Since we do not focus on model selection in our work, the marginal likelihood $f(Y)$ does not really play an important role. Thus, under specific priors for ϕ, we can write the posterior distribution as:

$$\pi(\phi|Y) \propto F(Y|\phi) \times \pi(\phi)$$

We proceed using Markov Chain Monte Carlo techniques as suggested by Jacquier et al. (1994) and Kim et al. (1998). We implement a multi-block Gibbs sampler such as developed in Section 3.1.2 to obtain posterior estimates under the assumption of conjugate priors for the regression parameters. In fact, we suppose that priors for the regression parameters β are normally distributed and use Gamma prior for the MIDAS lag polynomial coefficients both θ_1 and θ_2 as suggested by Ghysels (2012). Similarly we choose priors for λ and ψ^2 from the normal-inverse gamma family. We use a hybrid algorithm by combining Metropolis Hastings steps with a Gibbs sampler (the same procedure has been used by Clark, 2011 or Nakajima, 2011), and by adapting it to both MIDAS and stochastic volatility prescriptions. This accept-reject step lies on a candidate from a proposal density and iteratively draws the posterior of both θ in the MIDAS coefficient block and h_t in the stochastic volatilty block. The algorithm is constructed as follows:

1. Initialize β, θ, h_1, λ and ψ^2.

2. Sample $\beta|\theta, \mathsf{h}_t, \lambda, y$ from $\mathcal{N}(\hat{\beta}, \hat{B})$

$$\pi(b|\theta, \mathsf{h}_t, y) \propto \exp\left(-\frac{1}{2}(\beta - \hat{\beta})'\hat{B}^{-1}(\beta - \hat{\beta})\right) \tag{3.9}$$

   where $\hat{B} = \left(B_0^{-1} + \frac{\mathsf{X}(\theta)'\mathsf{X}(\theta)}{\exp(\mathsf{h})}\right)^{-1}$, and $\hat{b} = \hat{B}\left(B_0^{-1}b_0 + \frac{\mathsf{X}(\theta)'y}{\exp(\mathsf{h})}\right)$.

3. Sample $\lambda|\beta, \theta, \mathsf{h}_t, \psi^2, y_t$.

   We have assumed that $\mathsf{h}_t|\psi^2 \sim \mathcal{N}(\lambda_0 + \lambda_1 h_{t-1}, \psi^2)$. Then the conditional distribution of h_t can be written as, for all t :

$$\pi(\mathsf{h_t}, \lambda, \psi^2) = \pi(\mathsf{h_t}|\lambda, \psi^2) \times \pi(\lambda|\psi^2) \times \pi(\psi^2) \tag{3.10}$$

$$= \prod_{\tau=2}^{t-1} \pi(\mathsf{h}_{\mathsf{t}\tau}|\lambda, \psi^2) \tag{3.11}$$

   where the prior density of $\{\mathsf{h}_1, \lambda, \psi^2\}$ follows a normal inverse gamma distribution.

4. Sample $\psi^2|\beta, \theta, \mathsf{h}_t, Y$ from $\mathcal{IG}(\frac{s}{2}, \frac{d}{2})$.

   Since we have the draw of h_t, we compute the residuals η_t and we draw samples from the inverse Gamma distribution where the scale parameter is $s = s_0 + \sum_{t=1}^{T-1} \eta_t^2$ and degrees of freedom are $d = T + \eta_0$.

5. Sample θ using another independence chain Metropolis-Hasting algorithm as described in Section 3.1.2.

6. Repeat J times steps 2 to 5.

3.2.2 An empirical example on US data

As an illustration of the MIDAS-SV model put forward in the previous section, we focus on nowcasting US GDP growth rate, using two different model specifications. We aim at describing the relationship between the US economic output growth financial markets evolutions, as measured by daily log-returns of the S&P500. The model is thus given by:

$$GDP_{t+h}^{Q} = \gamma_0 + \gamma_1 \, m_{K_D=100}(\theta_1, L) \, SP_t^D + v_t, \tag{M1}$$

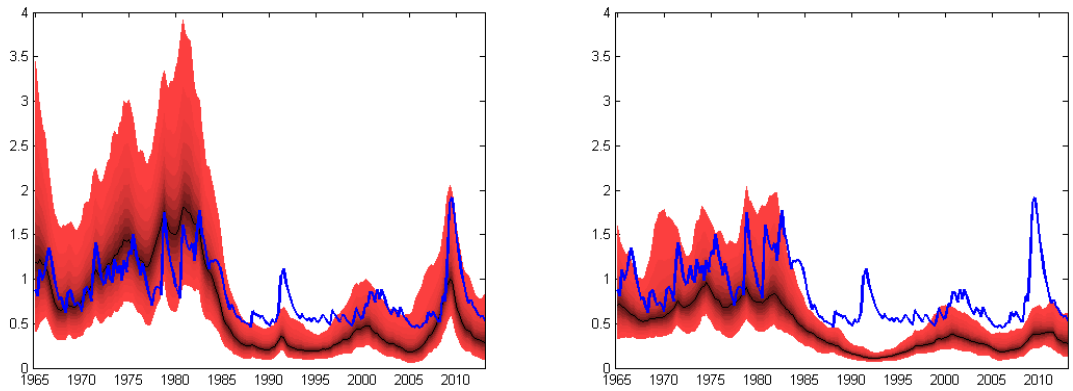
where $h < 1$ is the nowcasting horizon. We propose a second model which incorporates both an autoregressive term and the *PMI* which is a very important sentiment reading for the US economy on a monthly basis. This second model, denoted M2, involved the same intraperiod forecasting horizon h and is given by:

$$\begin{aligned} GDP_{t+h}^{Q} = \beta_0 + \beta_1 \, m_{K_D=100}(\theta_1, L) \, SP_t^D + \beta_2 \, m_{K_M=7}(\theta_2, L) \, PMI_t^M \\ + \beta_3 GDP_t^Q + v_t, \end{aligned} \tag{M2}$$

It is important to note that the error term v_t follows equations (SV) and (3.8) in both M1 and M2 models.

Empirical stochastic volatility

Bayesian parameter estimation over the whole sample 1964q4-2012q4 is carried out using the methodology presented in the previous section. The estimated stochastic volatility, $\ln(\hat{\mathsf{h}}_t)$, of both models are presented in Figure 3.4.



(a) M1 model (b) M2 model

FIGURE 3.4: Empirical stochastic volatility in both nowcasting MIDAS models
Red fan charts represent stochastic volatility distributions in both considered models. Conditional variance of GDP measured using a GARCH(1,1) model is displayed with the solid blue line.

We clearly observe a downward shift in volatility regime starting from the mid-eighties, well documented in the empirical macroeconomics literature and referred to as the Great Moderation period (see e.g. Perez-Quiros and Timmermann, 2001). Since then we observe some spikes of relative amplitude in the stochastic volatility that correspond to US economic recessions, meaning that macroeconomic volatility increases during those specific phases. It is noteworthy that the increase due to the recent worldwide Great Recession in 2008-09 is of limited magnitude when compared with previous observed levels of volatility, especially during the two world oil shocks in 1974-75 and 1981. Interestingly, we also observe that the uncertainty surrounding the stochastic volatility estimates, $\ln(\hat{\mathsf{h}}_t)$, has been also strongly reduced starting from the beginning of the Great Moderation period. This uncertainty is measured by the confidence interval using a standard 90% level, stemming from the estimation step, as can be seen on Figure 3.4a. When comparing the median of both stochastic volatility distributions, displayed by the solid black line, they are relatively similar in expansion period since the Great Moderation period, and substantially differ during recessions[3.6]. In model M1, if we interpret this stochastic volatility measure on the residuals as the macroeconomic conditional variance that cannot be explained by financial information, this leads to conclude that, during the last Great Recession, a large fraction of the macroeconomic variance was driven by financial volatility. This conclusion is also supported by the fact that the stochastic volatility during the years 2011-12 went back to pre-recession levels.

[3.6] The NBER reports three recession periods since 1985: in the early 1990's, in the early 2000's, and the Great Recession from beginning-2008 to mid-2009

Those empirical stochastic volatilities can be compared to the conditional variance of GDP measured using a GARCH(1,1) model (represented by the solid blue line in Figure 3.4) that roughly corresponds to the macroeconomic volatility. This narrower uncertainty around this macroeconomic volatility partly reflects a stronger long-run interaction between macroeconomic and financial areas. It is striking to note that during the Great Recession period, in spite of a increase in the volatility, though limited, the uncertainty around the volatility does not present any dramatic changes. These results are consistent with comparable estimates reported, for example, in Carriero et al. (2012).

Nowcasting

In order to assess the predictive power of the MIDAS-SV model, we provide GDP growth nowcasts over the period from 1988q1 to 2012q4 that includes three recession periods. Using a recursive window framework, nowcasts are updated daily using the specifications of the model M2. The nowcasts are presented in Figure 3.5.

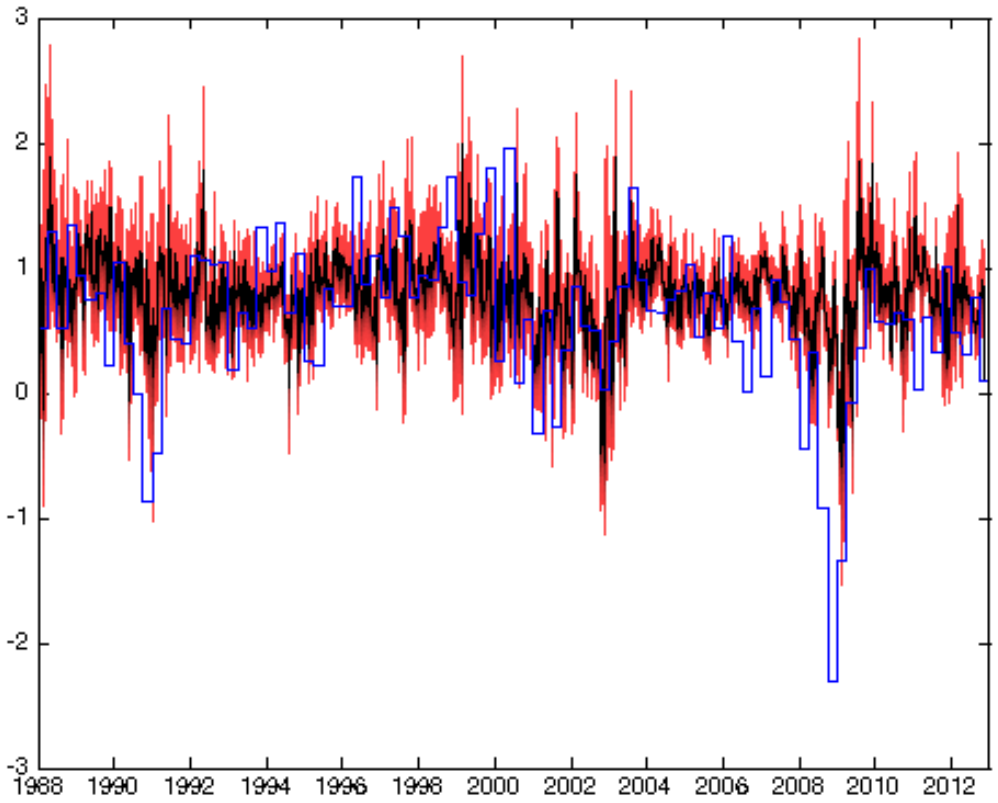


Figure 3.5: Nowcasts of the US GDP growth over the period from 1988q1 to 2012q4. The red fan chart corresponds to the distribution of daily nowcasts. The solid blue line is the series of realized values of the US GDP growth rate.

The nowcasts are daily updated with respect to the last available data, that explains the volatile nature of these forecasts. Nevertheless, the model roughly catches the overall evolutions of the economic activity. To really assess the quality of the forecasts, and given the shape of the stochastic volatility we described above, we separately compute Mean Square Forecasting Errors for periods of expansion and recession. We use the recession dates provided by the NBER for this purpose. The results are presented in Figure 3.6.

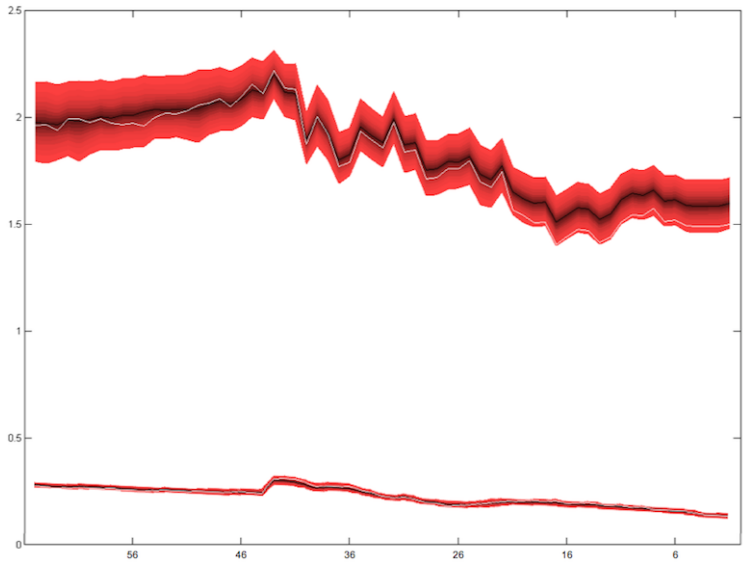


FIGURE 3.6: RMSFE during expansions and recessions over the period from 1988q1 to 2012q4.
Fan charts represent the distribution of the intra-quarter MSFE over recessions at the top and expansions at the bottom. The horizontal axis corresponds to the forecasting horizon in working days.

There is a huge difference in terms of forecasting accuracy between the periods of expansion and recession. While relative errors remain between 2 at the beginning of the current quarter and 1.6 at the end in times of recession, MSFE are between 0.3 and 0.1 in period of expansion. This regression model seems not to be appropriate to anticipate downturns. That suggests the use of regime switching or time-varying parameters in the model specifications. The Bayesian framework we developed can easily be extended in order to incorporate these features in future research (that involves some new hyper-parameter specifications and algorithm steps). We also notice that both MSFE distributions do not uniformly decrease with the forecasting horizon. We observe a slight improvement at the 2-month ahead forecasting horizon corresponding to 44 working days due to the release of the GDP_{t-1}.

This study develops further our work on financial volatility in Section 2.1 and provides

new results on the relationship between the US real activity and the financial markets. In particular, empirical results prove that the economic growth volatility is a stochastic process that can be explained to a certain extent by financial variables. Furthermore, the Bayesian MIDAS model we put forward is a general framework which has broad applicability.

Chapter 4

Variable selection in predictive mixed-frequency models

Short-term analysis aims at providing forecasts based on all the available information and it usually requires the use of data sampled at different frequencies. The Great Recession experienced by the main industrialized countries during the period 2008-2009 in the wake of the American subprimes crisis has encouraged many forecasters to reconsider their model specifications, especially regarding the interactions between financial and macroeconomic variables. In this respect, Foroni and Marcellino (2013b) and Banbura et al. (2012) have recently reviewed the existing mixed-frequency models designed for handling immediate past data (usually referred to as *ragged-edge* data) and nowcasting.

As we have seen in the previous chapters, the MIDAS method method allows us to explain a low frequency variable by using exogenous variables sampled at higher frequencies without resorting to any aggregation procedure. That is particularly suitable in macroeconomic forecasting and in capturing early signals of turning points using mul-

tifrequency explanatory variables. The main empirical use of MIDAS is the prediction quarterly GDP fluctuations using either monthly real economic data or daily financial series; Andreou et al. (2013) and Ferrara et al. (2014) showed that this combination of information significantly improves the prediction results. Furthermore, many works prove that an appropriate selection of explanatory variables, regardless of their sampling frequencies, has a major impact on the performance of this forecasting method. Therefore, an important achievement would be to determine the more relevant indicators from the huge volume of available data and thus improving the forecasting accuracy; separating the wheat from the chaff. In macroeconomic forecasting, empirical models are generally based on dimension reduction methods coming from either variable or model selection. The difference between these two close schemes is mainly methodological: variable selection aims at an a priori determination of the relevant predictors, while model selection provides an algorithmic approach to combine models which are typically univariate. The main goal of this chapter is introducing and comparing various variable selection methods within the mixed-frequency framework for macroeconomic forecasting. We will show that the use of well targeted predictors significantly improves the quality of forecasts. All schemes proposed have their grounds on either variable or model selection to tackle the so-called *curse of dimensionality* within a MIDAS forecasting framework.

A large family of widely used techniques in the literature for economic forecasting is based on principal component analysis and factor models. We refer, among others, to Forni et al. (2000) or Stock and Watson (2002). In the context of mixed-frequency models, Marcellino and Schumacher (2010) have put forward a dynamic factor MIDAS model (FAMIDAS) as a way to tackle the lack of parsimony associated to the profusion of covariates. FAMIDAS is a method to incorporate in a MIDAS framework standard tools of factor analysis that usually produce very good results for short-term forecasting (see Giannone et al., 2008 or Barhoumi et al., 2010) and hence represents a competitive benchmark when we compare the performance of different models. As an alternative to principal components analysis De Mol et al. (2008) have suggested the use of Bayesian regressions or penalized regressions (especially LASSO method[4.1] introduced by Tibshirani, 1996) as a dimension reduction technique. Other approaches for using mixed-frequency data are the bridge models introduced, for instance, by Barhoumi et al. (2008) for forecasting purposes. In Bencivelli et al. (2012) bridge based techniques are put together with Bayesian Model Averaging (BMA) to combine predictions coming from various model settings. The literature shows that forecast

[4.1] LASSO stands for Least Absolute Shrinkage and Selection Operator and is described in detail in Section 4.1.1.

combinations and, in particular, model averaging like BMA, yield good forecasting results. Indeed, Palm and Zellner (1992) claim that "*in some instances it is sensible to use a simple average of individual forecasts*". Rodriguez and Puggioni (2010) have recently adapted Bayesian approaches to estimate MIDAS models for forecasting exercises. In their paper, BMA provides a way to estimate the weights applied on the explanatory variables. Unfortunately, searching the best model using this approach involves maximizing marginal likelihoods and hence requires in general the assessment of the 2^n different combinations of models which may prove to be numerically expensive. Another technique available in the literature that extends the Bayesian selection analysis to stochastic search relies on the mixture of priors on regressor coefficients with spike and slab components. In this respect, we refer to Mitchell and Beauchamp (1988) for Dirac point mass spikes or to George and McCulloch (1993) for absolutely continuous spikes. These techniques have been widely exploited in econometrics; see, for example, Korobilis (2013) for an empirical application to the prediction of economic growth or Kaufmann and Schumacher (2012) for finding sparsity on factor models. The recent paper by Scott and Varian (2013) also considers spike and slab regression for variable selection in the nowcasting of economic time series.

In this chapter, we will focus on four different dimension reduction techniques that we combine with the MIDAS regression structure. More specifically, we introduce two new methods: (i) the LASSO augmented MIDAS model and (ii) the Bayesian MIDAS model with stochastic search variable selection. These novel strategies are then compared with (iii) the Factor Augmented MIDAS model, and (iv) a forecasts combination technique of univariate MIDAS based predictions. In these four approaches, the selection is carried out in-sample using a cross-validation procedure based on recent forecasting performances. We empirically assess the different selection methods by comparing point forecasts and prediction errors on the US GDP growth from 2000 to 2013. Our empirical results allow us to draw several important conclusions: first, we show that adequate variable selection significantly improves forecasting performances for all phases of the business cycle observed. Second, we observe that the two novel techniques developed succeeded in identifying early signals of the Great Recession from 3 to 6 months in advance while two other models were unable to capture the downturn. Third, the set of chosen predictors determined by the proposed variable/model selection procedure reflects the varying nature of the economic outlook.

The chapter is structured as follows: Section 4.1 describes the novel variable selection techniques that we develop, namely the LASSO augmented MIDAS model and the Bayesian MIDAS with stochastic search technique. In Section 4.2 we introduce the predictive cross-validation selection strategy. In Section 4.3, we empirically show how

the proposed selection methods of explanatory variables out of a universe of well-known economic variables can significantly improve short-term forecasts of US GDP.

4.1 Variable selection within MIDAS framework

4.1.1 The LASSO augmented MIDAS model

LASSO (Least Absolute Shrinkage and Selection Operator) has been introduced by Tibshirani (1996) as a covariate selection method in a linear regression setup. LASSO operates by penalizing the optimization problem associated to the regression with a term that involves the ℓ_1-norm of the coefficients. It belongs to the family of penalized regression model which amounts to performing least squares with some additional constraints on the coefficients, the ℓ_1-norm in the case of LASSO. Ng (2012) have shown that LASSO tends to have a lower misspecification risk in forecasting models when compared with usual information criteria. In the econometrics setup Bai and Ng (2008) and Schumacher (2010) have proposed to forecast economic series by using a combination of factor analysis with a LARS (see Efron et al., 2004) implementation of LASSO.

To be more specific, the LASSO takes advantage of the sparsifying properties of the ℓ_1-norm when solving the penalized optimization problem,

$$\begin{aligned}\widehat{b} &= \arg\min_b \sum_t \left(y_t - b_0 - \sum_i b_i \, x_{t,i} \right)^2 + \lambda_{\text{lasso}} \sum_i |b_i| \\ &= \arg\min_b ||Y - Xb||_2^2 + \lambda_{\text{lasso}} \, ||b||_1,\end{aligned} \tag{4.1}$$

where y_t is the dependent variable, x_t is the vector of covariates, b is the vector containing the regression parameters, and λ_{lasso} is the exogenous parameter which controls the strength of the LASSO penalization. The LASSO method does indeed reduce the dimension of the explanatory matrix X by driving non informative β_i elements to zero. Increasing $\lambda_{\text{lasso}} \in \mathbb{R}^+$ brings gradually elements of the β vector to zero, hence selecting relevant explanatory variables. The choice of the exogenous parameter λ_{lasso} that determines the number of covariates that are eliminated is essential and therefore a key issue that we will address later on via cross-validation.

Ridge regression is another popular penalized optimization scheme which, as opposed

to the ℓ_1 penalty of LASSO, is based on a ℓ_2-norm penalty. Figure 4.1 illustrates the underlying principle of both techniques in the case of a multivariate regression model with two variables: b_1 and b_2. The LASSO is on the left, and the ridge regression on the right.

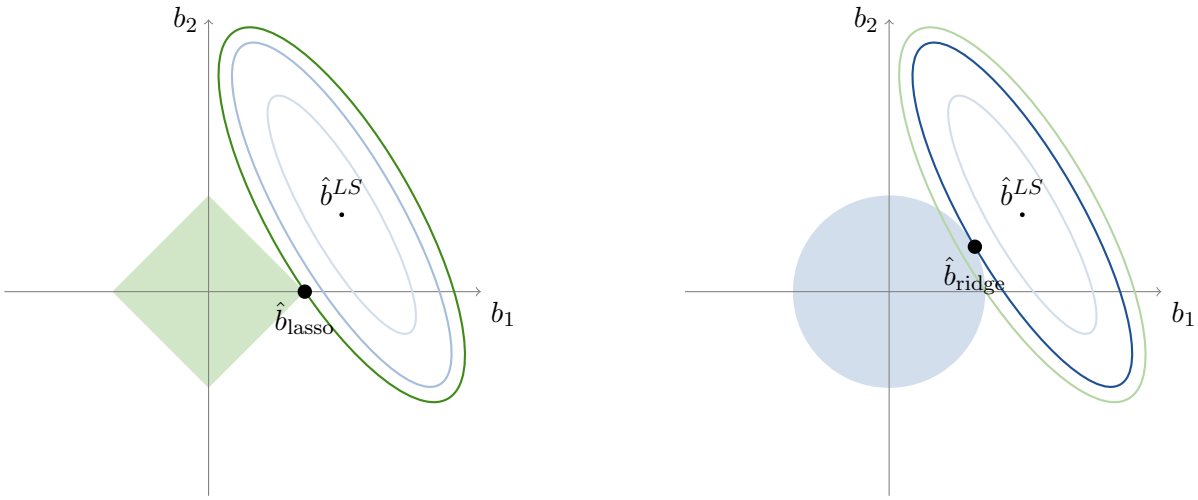


FIGURE 4.1: Penalized least squares estimate for the ℓ_1-norm (green) and the ℓ_2-norm (blue)

The ellipses around the least square estimator, $\hat{b}^{LS}$ represent the level sets of the squared error function $\|Y - Xb\|_2^2$ and the light colored areas correspond to balls of the ℓ_1 and ℓ_2 norms. In view of expression (4.1), the solution of the optimization problem that we are interested in takes place at the points in which both surfaces are tangent. The geometry of the problems makes that in the $\hat{b}^{\ell_1}$ case, the solution is generically located at the vertices of the ℓ_1-balls and hence the LASSO penalized solutions have entries equal to zero. The ridge based solutions are generally not located at that kind of specific points and are hence not necessarily sparse.

We put forward an extension of the LASSO model to the nonlinear MIDAS regression context by proposing the following optimization problem:

$$\begin{aligned}[\,\hat{\beta}, \hat{\theta}\,] &= \arg\min_{\beta,\theta} \sum_t \left(y_t - \beta_0 - \sum_{i=1}^{n} \beta_i \; m_{K_i}(\theta_i) \, x_{t,i}^{\kappa_i} \right)^2 + \lambda \sum_i |\beta_i| \\ &= \arg\min_{\beta,\theta} \|Y - \mathsf{X}(\theta)\,\beta\|_2^2 + \lambda \, \|\beta\|_1, \end{aligned} \tag{4.2}$$

where the matrix $\mathsf{X}(\theta)$ contains the MIDAS specifications that we previously described in (1.30),

As we have seen in the linear case in Figure 4.1, the ℓ_1 penalization on the β parameters implies a selection of the most relevant predictors. The number of covariates eliminated can be chosen by tuning the value of the exogenous parameter λ which controls the size of the constraint involved by the ℓ_1 penalty. A technical complication in solving (4.2) via any gradient descent method arises due to the non-smooth nature of the ℓ_1 norm. We overcome this difficulty using a local regularization technique due to Nesterov (2005). We start by noting that the ℓ_1 norm can be expressed using the function g defined as:

$$g(\beta) = \|\beta\|_1 = \max_{\|\gamma\|_\infty \leqslant 1} \gamma'\beta.$$

Then, we define the function g_μ such that $g_\mu \longrightarrow g$ with respect to $\mu \to 0$ and $\mu > 0$. We have:

$$g_\mu(\beta) := \max_{\|\gamma\|_\infty \leqslant 1} \gamma'\beta - \frac{\mu}{2}\|\gamma\|_2^2,$$

The Nesterov regularization technique consists of replacing the norm $g(\beta) = \|\beta\|_1$ by $g_\mu(\beta)$ with μ small. The advantage of proceeding in this fashion is that the function g_μ is obviously smooth with a gradient $\nabla g_\mu(\beta)$ whose components are given by

$$\nabla_i g_\mu(\beta) = \begin{cases} \text{sign}(\beta_i) & \text{if } |\beta_i| > \mu, \\ \dfrac{1}{\mu}\beta_i & \text{if } |\beta_i| < \mu. \end{cases}$$

As opposed to other standard iterative variable selection techniques, the combination of LASSO with the MIDAS regression presents the advantage of being a one-step procedure. This feature affects directly the numerical effort involved in its implementation, where the most expensive step will be the determination of the penalization strength λ. This parameter will be selected using what we call later on a predictive cross-validation method.

4.1.2 Bayesian variable selection in MIDAS models

Another approach that we explore in order to define the relevant subset of variables which should be included in the final regression model is a specific Bayesian variable selection technique that relies on *spike and slab priors* (see George and McCulloch

(1993)). This stochastic variable selection strategy is an a alternative to other usual Bayesian constructions that involve the comparison of all 2^n possible models , where n is the number of explanatory covariates under consideration. The approach that we propose yields a hierarchy on the covariates with respect to posterior distributions and relative inclusion probabilities. Kaufmann and Schumacher (2012) have recently used this technique to find relevant variables in sparse factor models.

The model selection relies on drawing the posterior ordinate using the Bayes formula. Indeed, we assume that residuals of the MIDAS regression model follow a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. Thus, the conditional likelihood function of the MIDAS model under study has the following form:

$$f(Y|\beta, \theta, \sigma) = \frac{1}{(2\pi\sigma)^{T/2}} \exp\left[\frac{1}{2\sigma}(Y - \mathsf{X}(\theta)\beta)'(Y - \mathsf{X}(\theta)\beta)\right], \tag{4.3}$$

where $(Y - \mathsf{X}(\theta)\beta)$ stands for the matrix expression of the MIDAS regression (see equation (4.2)).

Bayesian approaches have been rarely used in the context of MIDAS regression model; the main reference in this direction is Rodriguez and Puggioni (2010) where the authors focus not on variable selection but on the number of temporal lags used in the regression. In this context, they use an exponential Almon weight function combined with linear methods that are reminiscent of the U-MIDAS scheme of Foroni et al. (2013).

In the Bayesian framework, model parameters are derived from the posterior density which is, according to the Bayes formula, proportional to the likelihood times the prior, as described in (3.5). We extend the Bayesian MIDAS framework we developed in Chapter 3 to the variable selection purpose. We choose specific priors that will help us in determining whether a variable should be included or not. Indeed, we work with the spike and slab priors technique introduced by Mitchell and Beauchamp (1988) that constraints regressor coefficients to be zero (coefficient drawn from the "spike" prior) or not (drawn from the flat distribution: the "slab" prior). More specifically, we adopt a generalization of this method due to George and McCulloch (1993) that is usually referred to as Stochastic Search Variable Selection (SSVS) that takes as prior the following mixture of two normal distributions:

$$\beta_i|h_i \sim h_i\mathcal{N}(0, \varphi^2) + (1 - h_i)\mathcal{N}(0, c\,\varphi^2), \tag{4.4}$$

where c is a small positive number ($c << 1$), φ^2 sufficiently large, and h_i is the binary

random variable defined by

$$h_i = \begin{cases} 1 & \text{with} \quad \pi(h_i = 1) = \omega_i, \\ c & \text{with} \quad \pi(h_i = c) = 1 - \omega_i, \end{cases} \tag{4.5}$$

which allows the switching from a density concentrated around zero to another one with larger variance. When $h_i = 1$, β_i exhibits flat distribution and we can therefore consider that the covariate $x_{t,i}^{\kappa_i}$ that goes with it should be included in the model; conversely, when $h_i = c$, the density of the coefficient is concentrated around the zero value. Figure 4.2 illustrates this mixture of normal distributions.

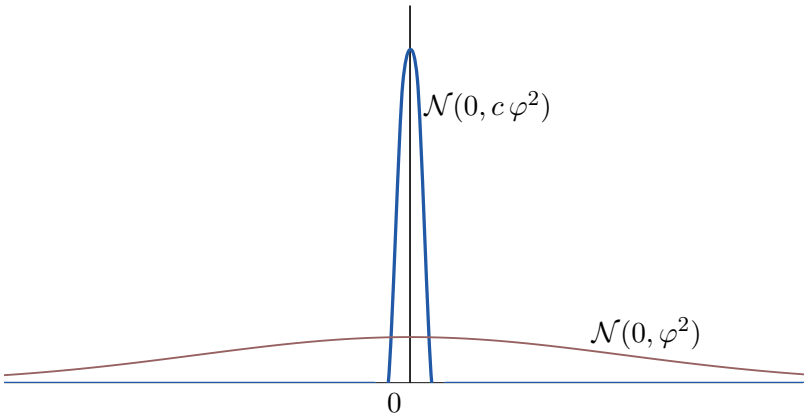


FIGURE 4.2: Mixture of a slab (red) and a spike (blue) normal distributions

We infer that, when $h_i = c$, the corresponding variable should not be taken into account as a regressor. Consequently, formulas (4.4) and (4.5) can be interpreted by saying that ω_i is the prior probability that $x_{t,i}^{\kappa_i}$ should be kept as a explanatory variable. This particular feature of the SSVS method has been often reviewed in the literature. In particular, more complex choices of prior can be made: for example, George and McCulloch (1997) defined an hierarchical prior for the inclusion probability using a beta distribution and Yuan and Lin (2005) have preferred the definition of a hierarchical Bayes formulation to show that it can be related to the LASSO estimator. Other references are Ishwaran and Rao (2005) or Malsiner-Walli and Wagner (2011).

In addition to the prior for β specified in (4.4) we analogously need to specify the prior for the residual variance σ. We choose for this purpose the inverse gamma distribution. We subsequently proceed by implementing a Gibbs sampler to generate a ergodic Markov chain in which all parameters $(h, \omega, \beta, \theta, \sigma^2)$ are embedded. In the particular case of the MIDAS parameter θ, we use an Independence Chain Metropolis Hastings algorithm (iMH) within the Gibbs sampler to draw the posterior conditional distribution of θ. The candidate posterior distribution chosen in the iMH is a normal

distribution whose mean and covariance matrix are approximated using the maximum likelihood estimator $\hat{\theta}_{ML}$. Details on the algorithm are provided in Appendix. The algorithm converges relatively fast to a steady state of the Markov chain and the distribution obtained is an approximation of the posterior distribution which informs us about the selection that can be carried out in terms of the probabilities ω_i.

Finally, we establish a probability threshold $\Omega \in [0, 1]$ via a predictive cross-validation technique based on forecasting performance such that when $0 < \omega_i < \Omega < 1$ we will consider that the relative predictor $x_{t,i}^{\kappa_i}$ should not be included in the model. We emphasize that in the same vein as the LASSO approach, the stochastic search variable selection yields a one-step estimation and selection procedure.

4.2 Predictive cross-validation

In this section, we propose a cross-validation method in order to determine model specifications that possess the best predictive power. In this respect, we assume that the selection is updated according to its predictive error. We investigate four families of forecasting models that we implement using the proposal predictive cross-validation: the LASSO augmented MIDAS, the Bayesian-MIDAS Stochastic Search, the FAMIDAS, and the forecast combination of univariate MIDAS regressions.

We start by defining the variable ξ_i that will be used as an indicator that determines whether the i^{th} variable must be taken into account or not in the model, that is,

$$\xi_i = \begin{cases} 1, & \text{if } x_{i,t}^{\kappa_i} \text{ is selected to be present in the model,} \\ 0, & \text{otherwise.} \end{cases}$$

We now rewrite the MIDAS model using the ξ_i variables and the direct multistep forecasting framework at the horizon h:

$$\hat{y}_{t+h|t} = \hat{\beta}_0 + \sum_{i=1}^{n} \xi_i \, \hat{\beta}_i \; m_{K_i}(\hat{\theta}_i, L) \, x_{t,i}^{\kappa_i}, \tag{4.6}$$

where $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n, \hat{\theta}_1, \ldots, \hat{\theta}_n)$ are parameter estimates usually obtained by either non-linear least squares or maximum likelihood methods.

It can be noticed that the effective size of the explanatory subset is determined by

the ξ variables that take a non-zero value and that will be chosen using either the LASSO or the Bayesian SSVS based procedures adapted to the MIDAS context that we presented in the previous section. In order to evaluate their performance we will take as a benchmark the Factor Augmented MIDAS (FAMIDAS) model put forward by Marcellino and Schumacher (2010). This approach is based on the combined use of two techniques: first, pooling information from blocks of covariates that share the same frequency into a certain number of factors, and second, tracking the dependent variable with a MIDAS regression model by incorporating these factors as explanatory variables.

Another benchmark that we also consider is forecast combination. Indeed, there is a growing volume of literature that shows that combining forecasts provides particularly competitive results in prediction tasks; we refer to the very complete survey of Timmermann (2006). This technique has also been implemented in the MIDAS context by Andreou et al. (2013) using a very rich financial data set.

It can also be noticed also that the direct multi-step forecasting strategy that we adopt provides parameter estimates $\beta^{(h)}$ and $\theta^{(h)}$ that depend on the prediction horizon h which is given at the lowest frequency time units. The selection parameters $\lambda_t^{(h)}$ for the LASSO approach and $\Omega_t^{(h)}$ for the Bayesian SSVS model are time dependent because their choice is based on the use of a recursive window framework over the whole out-of-sample period. Notice that since the selection variables ξ_i depend on $\lambda_t^{(h)}$ or $\Omega_t^{(h)}$ they are hence also time dependent. Note also that the FAMIDAS model and the forecast combination also involve time-varying specifications using the predictive cross-validation. Model settings are described below:

(i) In the case of the LASSO, we have the following forecasting equation:

$$\hat{y}_{t+h|t}(\lambda_t^{(h)}) = \hat{\beta}_0^{(h)} + \sum_{i=1}^{n} \xi_i(\lambda_t^{(h)})\, \hat{\beta}_i^{(h)}\; m_{K_i}(\hat{\theta}_i^{(h)}, L)\, x_{t,i}^{\kappa_i}. \qquad \text{(LASSO-MIDAS)}$$

As opposed to Tibshirani (1996), our goal is not to recover an underlying sparsity in the coefficients vector β but to use the penalty to reduce the covariates cardinality. The question that arises in this context is the selection of the optimal strength of the ℓ_1 penalty that ensures a favorable forecasting performance. Our cross-validation procedure follows those prescriptions: for a given value $\lambda > 0$, we set its corresponding selection by estimating the equation (4.2), and we forecast $\hat{y}_{t|t-h}(\lambda)$ such as defined in the LASSO-MIDAS equation. Then, repeating that for a range of λ, we determine $\lambda_t^\star$ as the one which minimizes the following

forecasting residual at time t:

$$\lambda_t^{\star(h)} = \arg\min_{\lambda} \sum_{\mathsf{t}=t-d}^{t} \delta^{t-\mathsf{t}} \left(y_{\mathsf{t}} - \hat{y}_{\mathsf{t}|\mathsf{t}-h}(\lambda_{\mathsf{t}}^{(h)}) \right)^2, \tag{4.7}$$

where we set $\delta = 0.8$ in order to be coherent with our wish to involve a decrease on the MSFE weight with respect to the historical performance, which allows to comparatively promote recent forecasting accuracies.

(ii) The Bayesian Stochastic Search variable selection combined with the MIDAS forecasting model is given by:

$$\hat{y}_{t+h|t}(\Omega_t^{(h)}) = \hat{\beta}_0^{(h)} + \sum_{i=1}^{n} \xi_i(\Omega_t^{(h)})\, \hat{\beta}_i^{(h)}\; m_{K_i}(\hat{\theta}_i^{(h)}, L)\, x_{t,i}^{\kappa_i}. \quad \text{(BAYESIAN-MIDAS)}$$

The posterior probability ω_i as described in (4.4) and in (4.5) specifies the probability that β_i has not been drawn from the spike prior, namely the probability to include it in the model. The issue that arises in this case is to choose a threshold $\Omega^\star \in [0, 1]$ below which variables are simply removed. Following exactly the same procedure than in the LASSO case, we forecast $\hat{y}_{t|t-h}(\Omega)$ according to its relative set of selected variables. Then, we set $\Omega_t^\star$ as the minimum argument of the square error for the period t:

$$\Omega_t^{\star(h)} = \arg\min_{\Omega} \sum_{\mathsf{t}=t-d}^{t} \delta^{t-\mathsf{t}} \left(y_{\mathsf{t}} - \hat{y}_{\mathsf{t}|\mathsf{t}-h}(\Omega_t^{(h)}) \right)^2.$$

(iii) The FAMIDAS model is based on a factor structure assumption for the explanatory variables matrix, that can be described as follows:

$$X_\tau = \Lambda F_\tau + \eta_\tau,$$

where τ is given in one of the higher frequencies (daily or monthly in our case). The components of the factors vector are denoted as $F_\tau = (f_{1,\tau}, \ldots, f_{r,\tau})$. This approach consists of using the standard MIDAS technique with the r first estimated principal factors that are employed as explanatory variables[4.2]. The model

[4.2] This strategy has been used in Section 2.2 for nowcasting the global output growth

is given by

$$\hat{y}_{t+h|t}(r^{(h)}) = \hat{\beta}_0^{(h)} + \sum_{i=1}^{r^{(h)}} \hat{\beta}_i^{(h)} m_{K_i}(\hat{\theta}_i^{(h)}, L)\, \hat{f}_{t,i}^{\kappa_i} \qquad \text{(FAMIDAS)}$$

Since factors are linearly uncorrelated, the size of the factor vector, r, can be determined with a statistical hypothesis test, as proposed by Bai and Ng (2008). In our study, we propose to define $r^\star$ depending on the forecasting performances. In that case, the parameter we focus on is the number of factors to include in the final model.

$$r_t^{\star(h)} = \arg\min_{r} \sum_{\mathsf{t}=t-d}^{t} \delta^{t-\mathsf{t}} \left(y_{\mathsf{t}} - \hat{y}_{\mathsf{t}|\mathsf{t}-h}(r_t^{(h)}) \right)^2$$

Note that factors can only represent a family of variables sampled at the same frequency. Since we mix daily and monthly predictors, we define $r = (r^D, r^M)$, where $r^D = \{0, 1\}$ and $r^M = \{0, 1, 2\}$.

(iv) Combining forecasts is often considered as a good alternative to model selection. Formally, we compute n individual forecasts respectively based on the i^{th} variable of the entire set, as follows:

$$\hat{y}_{t+h|t,i} = \hat{\beta}_0^{(h)} + \hat{\beta}_i^{(h)}\ m_{K_i}(\hat{\theta}_i^{(h)}, L)\, x_{t,i}^{\kappa_i}. \qquad (4.8)$$

The combination is then made using a weighted average of the individual forecasts (4.8), thus it can be written as follows:

$$\hat{y}_{t+h|t}(w_t^{(h)}) = \sum_{i=1}^{n} w_{t,i}^{(h)}\, \hat{y}_{t+h|t,i} \qquad \text{(COMBINATION)}$$

The forecast relies on the vector of the time-varying combination weights $w_{i,t}^{(h)}$ which can be estimated using several methods; Stock and Watson (2008) show some of those techniques. In this paper, we determine using an equivalent procedure than others selection methods to fairly compare all models. This model relies on the vector of $w_{i,t}^{\star}$ that weights the individual forecasts, see (4.8). Those are given as follows:

$$w_{t,i}^{\star(h)} = \frac{\mu_{t,i}^{-a}}{\sum_{j=1}^{n} \mu_{t,j}^{-a}} \quad \text{where } \mu_{t,i} = \sum_{\mathsf{t}=t-d}^{t} \delta^{t-\mathsf{t}} \left(y_t - \hat{y}_{\mathsf{t}|\mathsf{t}-h,i} \right)^2, \text{ and } a = 2.$$

In these four models, the predictive cross-validation is based on forecasting performances over the d previous quarters. Notice that the value of d would have different meanings, e.g. $d = 1$ tells that we only base the analysis on the last period whereas $d = 20$ represents the selection that gave best results over the last 5 years. Furthermore, instead of the usual MSFE (Mean Squared Forecasting Error), we prefer focusing on an discounted version of this criterion such as Andreou et al. (2013) used in their paper. That metric promotes recent performances by weighting squared residuals according to their historical records. Concerning the pseudo out-of-sample period, we opt for an intermediate parametrization which corresponds to forecasting performances over the last year, i.e. $d = 4$.

Using this cross-validation procedure on previous quarters preceding the forecasting stage $t+h$ within the recursive window framework that we describe above, the selection is updated every period of the out-of-sample. This technique leads to an automated model selection procedure that improves the selection of the leading indicators and yields greater efficiency in their use.

4.3 An assessment based on macroeconomic forecasting

4.3.1 Empirical exercise on US data

We assess the performance of the four models that we have presented above using a forecasting exercise on US GDP data over the period 2000q1-2012q4 while the full sample covers a longer period going from 1964q3 to 2012q4. In this forecasting exercise, we focus on predicting the quarterly US Gross Domestic Product using a set of 24 variables which includes monthly real indicators and daily financial variables. More specifically, the dataset incorporates a daily spread rate and three financial times series. Our set also includes seventeen monthly indicators related to the real US economy and coming from "soft" and "hard" data (production index, housing statistics, unemployment rate, opinion survey, etc.). An entire description of the dataset is available in Table 4.1.

The Great Recession has shed light on the necessary re-assessment of the contribution of financial markets to the economic cycles. There is a huge volume of work in the

Daily series		
10y-3m	Spread rate: 10y Treasury Rate - 3m Treasury Bill	daily Δ
CRB	CRB Spot index, commodities price index	daily $\Delta \log$
DJ	Dow Jones industrial share price index	daily $\Delta \log$
SP500	S&P500 index	daily $\Delta \log$
CRBvolat	CRB Spot index, commodities price index	daily volatility (see 2.2)
DJvolat	Dow Jones industrial share price index	daily volatility (see 2.2)
SP500volat	S&P500 index	daily volatility (see 2.2)
Monthly series		
AAA	Moody Yield on Seasoned Corporate Bonds AAA	monthly $\Delta \log$
AMBSL	St Louis Adjusted Monetary Base	monthly $\Delta \log$
BAA	Moody Yield on Seasoned Corporate Bonds BAA	monthly $\Delta \log$
BusLoans	Commercial and Industrial Loans at Commercial banks	monthly $\Delta \log$
CPI	Consumer Price Index for all Urban Consumers: All items	monthly $\Delta \log$
Curr	Currency component of M1	monthly $\Delta \log$
DSPIC	Real Disposable Personal Income	monthly $\Delta \log$
Housing	New privately owned housing units started	monthly $\Delta \log$
IPI	Industrial Production Index	monthly $\Delta \log$
Loans	Loans and leases in bank credit, all commercial banks	monthly $\Delta \log$
M2	M2 money stock	monthly $\Delta \log$
Oil	Spot oil price: WTI	monthly $\Delta \log$
PCE	Personal Consumption Expenditures	monthly $\Delta \log$
PMI	ISM manufacturing survey: PMI composite index	monthly level
PPI	Producer Price Index: all commodities	monthly $\Delta \log$
TotalSL	Total consumer credit owned and securitized outstanding	monthly $\Delta \log$
Unemploy.	Unemployment rate	monthly Δ

TABLE 4.1: US data set from 1964:1 to 2012:4

literature that underlines the leading role of financial variables in the forecasting of macroeconomic fluctuations. Recently, Chauvet et al. (2012) and Ferrara et al. (2014) have even shown that daily volatility of financial time series series have a significant forecasting power concerning US growth. We particularly focus on this topic in Section 2.1. Using variable selection models within the predictive cross-validation that we have put forward, we evaluate whether both returns and volatility of financial time series should be included in the model specifications to forecast US GDP growth. Given that volatility is not directly observable, several methods have been developed in the literature to estimate it. Following Ferrara et al. (2014), we use a GARCH model on whitened and winsorized daily financial series, as described in Section 2.1.1 in the equation 2.2 Estimated daily volatilities are considered as explanatory variables of the US macroeconomic fluctuations.

4.3.2 Forecasting results

From 2000 to 2013, the US economy experienced different phases of the business cycle. In 2008, in the wake of the financial crisis, the United States entered a severe recession, referred to as the Great Recession. The recovery since 2009 was weak and growth remained uneven. Our approach allows to set the horizon at which leading indicators have early information and can send warnings about turning point. In this respect, we assess the MIDAS-based models presented in the previous sections, by splitting our sample in three parts: Early 2000's (from 2000q1 to 2007q2), Great Recession (from 2007q3 to 2009q4), and Recovery (from 2010q1 to 2012q4). Table 4.2 reports the Mean Squared Forecasting Errors (MSFE) in these three periods. Moreover, in order to assess the predictive gain of selecting variables, we also report results from a MIDAS model that makes use of the full set of variables. Point forecasts and model inclusion for all horizons are exhibited in the Appendix D. Results of the forecast comparison exercise for GDP growth are discussed below.

	2000q1-2012q4 Full sample	2000q1-07q2 Early 2000's	2007q3-09q4 Great Recession	2010q1-12q4 Recovery
$h = 0$ (Nowcasting)				
LASSO-MIDAS	0,34	0,33	0,45	0,19
BAYESIAN-MIDAS	0,32	0,29	0,52	0,20
FAMIDAS	0,33	0,28	0,73	0,14
COMBINATION	0,38	0,33	0,62	0,12
MIDAS	0,43	0,52	0,54	0,23
$h = 3$				
LASSO-MIDAS	0,37	0,32	0,79	0,13
BAYESIAN-MIDAS	0,40	0,37	0,79	0,15
FAMIDAS	0,40	0,27	1,14	0,12
COMBINATION	0,42	0,33	0,94	0,20
MIDAS	0,51	0,55	0,84	0,23
$h = 6$				
LASSO-MIDAS	0,52	0,34	1,24	0,18
BAYESIAN-MIDAS	0,48	0,37	1,14	0,19
FAMIDAS	0,46	0,30	1,39	0,13
COMBINATION	0,42	0,30	0,99	0,27
MIDAS	0,62	0,49	1,49	0,29
$h = 9$				
LASSO-MIDAS	0,55	0,35	1,53	0,23
BAYESIAN-MIDAS	0,47	0,31	1,19	0,21
FAMIDAS	0,52	0,33	1,65	0,15
COMBINATION	0,42	0,32	1,07	0,16
MIDAS	0,58	0,52	1,45	0,29
$h = 12$				
LASSO-MIDAS	0,54	0,31	1,61	0,22
BAYESIAN-MIDAS	0,66	0,38	2,10	0,18
FAMIDAS	0,52	0,29	1,76	0,15
COMBINATION	0,44	0,30	1,12	0,25
MIDAS	0,68	0,46	1,85	0,26

TABLE 4.2: MSFE (Mean Squared Forecasting Errors)

Nowcasting

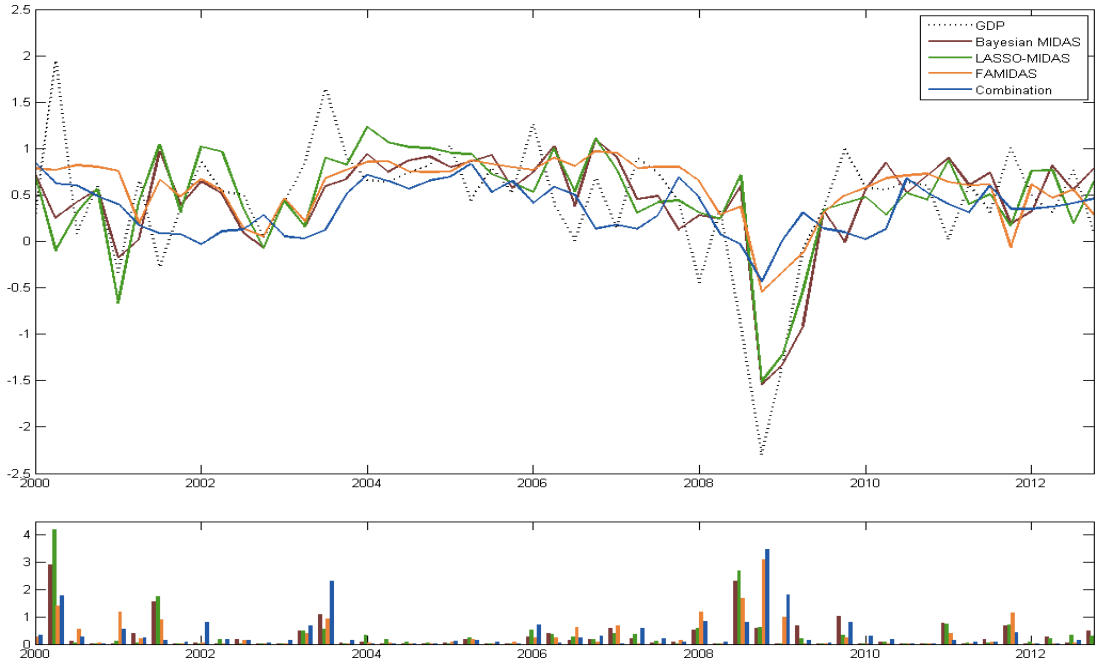


Figure 4.3: Point forecasts (top) and squared errors (bottom) for $h = 0$

For short horizons, the indicators chosen by both variable selection techniques, the LASSO-MIDAS and the BAYESIAN-MIDAS, are primarily related to the real economic activity (production, labor market, housing, consumption). This stylized fact has been observed in empirical papers pointing out the increasing role of hard indicators on macroeconomic forecasts when we are close to the release date. In addition, at this horizon, the financial volatility of the S&P500 was among the best predictors (always included in both predictor set).

Best performances in nowcasting the Great Recession were provided by both variable selection methods. Those indicate that financial instability, especially observed via volatility variables and commodity price indices, triggered confusion and fear among consumers and firms. Lower confidence and lower stock price leads to a net decrease in consumption in that period and hence in GDP growth. These findings are in agreement with the COMBINATION model showed that the IPI and the ISM PMI survey are particularly important during this period.

The 3-month to 6-month ahead horizons

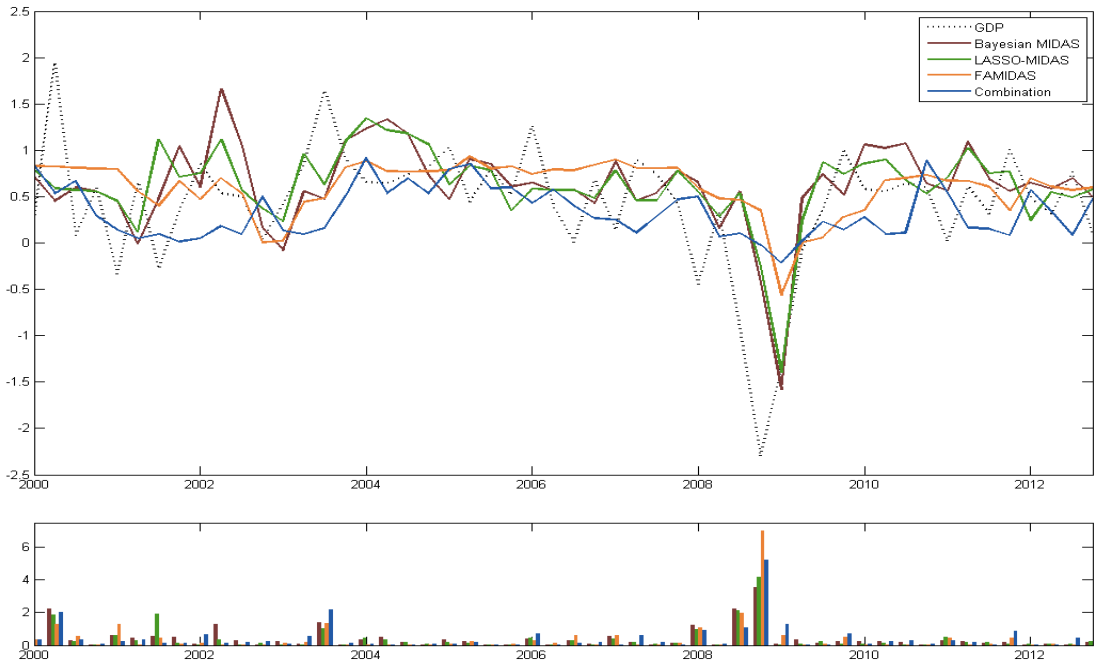


Figure 4.4: Point forecasts (top) and squared errors (bottom) for $h = 3$.

For the 6-month-ahead horizon, financial variables emerge as the most useful indicators. Rate spreads and stock price volatility dominate the top ranks in model inclusion. A key difference between pure nowcasting for 0-month-ahead and 3-month-ahead forecasts is that for the latter horizon, IPI variables are not very prominent. This result is interesting for practitioners in the sense that using the industrial production index for this horizon does not appear useful. We also note that the LASSO-MIDAS tends to select variables that are not encompassed by other indicators. In fact, the LASSO would prefer substitution in spite of complementarity that could be involved by the BAYESIAN-MIDAS shrinkage and the forecast COMBINATION.

Regarding performances in predicting the Great Recession, these four models have captured early warnings from 6 to 3 months ahead. By the end of 2007, serious short-term risks were looming: uncertainty on financial markets (captured by stock prices volatility), bank loan contraction, rising interest rates. We note that model inclusion of those indicators in both variable selection and an increasing weight in the combination. From early 2010 to the end of 2011, while financial indices remain high, the recovery was slower than expected, and it was referred to in the literature as *Sluggish Recovery*[4.3]. Both FAMIDAS and COMBINATION models show this disconnection

[4.3] We especially focus on this post-crisis period at the worldwide level in Section 2.2.

by either not including anymore the daily factors or by reducing their weights in the regression. Finally, by the end of 2011, in the wake of the sovereign debt crisis, some financial indicators (spread rate, corporate bonds and stocks volatility) were again chosen.

The 9-month to 12-month ahead horizons

For the 9-month ahead horizon, we note that financial volatility indicators already played an important role in forecasting, especially over the Early 2000's and the Recovery, as already noticed for the 6-month-ahead forecasts. In addition for this horizon we get complementary information from money related variables such currency component of M1, M2, and St Louis monetary base. We also find that the inflation rate (CPI) is chosen by both variable selection and highlighted as one the main indicator in the combination model for both 9-month and 12-month ahead horizons. Findings during the Great Recession period should be interpreted with care since forecasting errors are really high. Indeed, from 12 to 9 months ahead, it turns out that the four models provided flat predictions, and hence did not yield informative contents to anticipate the crisis.

Summary

According to our results about the forecast comparison exercise for GDP growth, four main conclusions can be drawn. First, it can be noticed that over the whole sample, our four MIDAS based models outperform the full MIDAS models and we also observe that forecasting errors for all models decrease when the forecasting horizon tends to zero. Second, results significantly differ depending on the period. More specifically, early warnings of the "Great Recession" were clearly identify from 3 to 6 months before it happened. In fact, most models that we have studied tend to perform best with short horizons although in some cases the performance extends to three or four quarters. Third, a few economic stylized facts have been summarized concerning the set of predictors and the forecasting horizon. The set of chosen indicators includes reasonable variables from an economic point of view and reflects both their intrinsic leading features and the time varying nature of the economic outlook. Fourth, our forecasting exercise on GDP growth proves that pooling indicators ability provides very reliable models. Observed individually over their respective primary horizons, some of indicators would already give very good results, grouping yields even better performances and minimum forecast errors.

Conclusion

In short-term forecasting, it is essential to take into account all available information on the current state of the economic activity. Yet, the fact that various time series are sampled at different frequencies prevents an efficient use of available data. In this respect, the Mixed Data Sampling (MIDAS) model has proved to outperform existing tools by combining data series of different frequencies. This thesis aims at investigating macro-financial linkages by assessing the leading role of major financial variables. We propose different studies based on MIDAS models focusing on the prediction of the economic growth during the last decade, and especially over the Great Recession period. In particular, we put forward mixed-frequency models that integrate both daily and monthly explanatory variables to forecast the quarterly GDP growth. We argue that adding daily financial returns and volatilities increases the forecasting accuracy in comparison with benchmark models that include hard data of the real economic activity as an explanatory variable (see Chapter 2).

In the context of MIDAS regression problems we also developed a Bayesian framework in order to provide a flexible approach to allow for stochastic volatility in the economic growth data (see Chapter 3). The empirical results suggest extending the MIDAS regression model to regime switching or time varying features. From the perspective of prediction, these nonlinear specifications would be particularly useful to anticipate turning points and foresee downturns. This Bayesian estimation procedure is based

on a Metropolis-Hastings algorithm and constitutes a generic approach that may be adapted to those various augmented MIDAS models. Predicting US recessions using financial indicators already is of particular interest of a new research project for which we develop a Probit MIDAS model.

In Chapter 4 we develop four tools to identify leading indicators of the US GDP growth, regardless of their sampling frequency, using an automatic model selection procedure based on recent best performances. More specifically, we introduce a LASSO augmented MIDAS model and a Bayesian MIDAS Stochastic Search Variable Selection that we compare with the Factor Augmented MIDAS model, and the combination forecast technique of univariate MIDAS models. Those are specified using a predictive cross-validation methodology relying on a recursive window and on a set of economic series with respect to their forecasting ability. These dimension reduction methods go beyond point forecast and highlight the leading role of some indicators in macroeconomics. Our findings particularly emphasize the role of daily financial information in predicting GDP anew and showed that combining daily and monthly data significantly increases the forecasting accuracy. The question we address focuses on variable selection in predictive mixed-frequency models. Forecasting GDP is only one of many examples where our methods can be applied. These approaches have broad applicability and indeed can be of general interest in many other macroeconomic applications.

Appendices

A Gradient of the exponential Almon function

Considering an exponential Almon lag weight function $\varphi(k, \theta_1, \theta_2) = \exp(\theta_1 k + \theta_2 k^2)$, the gradient of the MIDAS function $\nabla m_K(\theta_1, \theta_2, L)$ is composed of the two following elements:

$$\begin{aligned}\frac{\partial m_K(\theta_1, \theta_2, L)}{\partial \theta_1} = \sum_{k=0}^{K-1} \Bigg(& k \exp\left(\theta_1 k + \theta_2 k^2\right) \left(\sum_{l=0}^{K-1} \exp\left(\theta_1 l + \theta_2 l^2\right) \right) \\ & - \exp\left(\theta_1 k + \theta_2 k^2\right) \left(\sum_{l=0}^{K-1} l \exp\left(\theta_1 l + \theta_2 l^2\right) \right) \Bigg) \\ & \times \frac{1}{\left(\sum_{l=0}^{K-1} \exp\left(\theta_1 l + \theta_2 l^2\right) \right)^2} L^k\end{aligned}$$

$$\frac{\partial m_K(\theta_1,\theta_2,L)}{\partial \theta_2} = \sum_{k=0}^{K-1}\left(k^2\,\exp\left(\theta_1 k+\theta_2 k^2\right)\left(\sum_{l=0}^{K-1}\exp\left(\theta_1 l+\theta_2 l^2\right)\right)\right.$$
$$\left.-\exp\left(\theta_1 k+\theta_2 k^2\right)\left(\sum_{l=0}^{K-1} l^2\,\exp\left(\theta_1 l+\theta_2 l^2\right)\right)\right)$$
$$\times\frac{1}{\left(\sum_{l=0}^{K-1}\exp\left(\theta_1 l+\theta_2 l^2\right)\right)^2}\,L^k$$

It can be noticed that the gradient of the exponential Almon lag weight function exists for all θ_1 and θ_2.

B Financial variables for forecasting growth in the Euro Area during the Great Recession

This section is based on the paper entitled "Financial variables as leading indicators of GDP growth: evidence from a MIDAS approach during the Great Recession", written with Laurent Ferrara and published in Applied Economics Letters, vol. 20(3) (February 2013).

In the wake of the financial and banking crisis, most of all industrialized countries have experienced a very severe economic recession during the years 2008 and 2009, sometimes referred to as the Great Recession. This Great Recession has emphasized the necessary re-assessment of financial markets in their ability to anticipate the business cycle. Regarding the role of financial market variables, there is a huge literature pointing out the leading property of those series to forecast macroeconomic fluctuations (see a review in Stock and Watson (2003). For example, Kilian (2008) reviewed the impact of energy prices shocks, especially oil prices, on macroeconomic fluctuations and Hamilton (2003) put forward a non-linear Markov-Switching model to predict US GDP growth rate through oil prices. The term spread has also been widely considered in empirical approaches to assess in a quantitative manner future GDP growth, we refer among others to Estrella et al. (2003) for the US and to Duarte et al. (2005) or Bellégo and Ferrara (2012) for the euro area. When dealing with variables sampled at various frequencies (quarterly GDP and monthly financial information), the MIDAS approach put forward by Ghysels and his co-authors has proved to be a useful

tool. Especially in the forecasting framework, several empirical papers have shown the ability of financial information to predict macroeconomic fluctuations; we refer for example to Clements and Galvão (2008) for the US or Marcellino and Schumacher (2010) for Germany. We assess the impact of financial returns as leading indicators for GDP growth for the four main euro area countries (Germany, France, Italy and Spain), as well as for the euro area as a whole. We carry out a forecasting analysis, over the period ranging from 2007q1 to 2009q4, focused on three well-known financial variables, namely oil prices, stock prices, and spread between long and short interest rates, for several forecasting horizons. The MIDAS enables to use variables of various frequencies in a single univariate model. Especially a MIDAS regression allows to explain a low frequency variables by exogenous variables of higher frequency, without any aggregation procedure and within a parsimonious framework; econometric details has been provided in Chapter 1.

We use here univariate MIDAS regressions designed to accommodate direct multi-step forecasting. To predict the quarterly GDP growth GDP_t^Q, we base our regression model on an monthly explanatory variable x_t^M and on a first order autoregressive component. The forecasting equation is given by:

$$GDP_{t+h|t}^Q = \hat{\beta}_0^{(h)} + \hat{\beta}_1^{(h)}\ m_K(\hat{\theta}^{(h)}, L)\, x_t^M + \hat{\lambda}^{(h)}\, GDP_t^Q \tag{B.1}$$

where m_K is the Weigth function and h is the quarterly forecasting horizon.We implement forecasts for quarterly GDP growth rates for each of the main euro area countries (Germany, France, Italy and Spain), as well as for the euro area as a whole, starting from the same set of explanatory variables, namely oil prices, stock prices and the spread between long and short-term interest rates. The financial time series are detailed in Table 4.3.

Real output		
GDP	GDP growth in France, Germany, Italy, Spain, and euro area (resp. *INSEE, DeStatis, Istat, INE, Eurostat*)	Quarterly growth rate
Financial series		
Stocks	CAC40, DAX, FTSE MIB, IBEX35, and DJ EuroStoxx50 indices (*Bloomberg*)	Monthly Δlog
Oil	Oil price quoted at New York Mercantile Exchange (*Bloomberg*)	Monthly Δlog
Spread	Term spread: 10 years Government bond - 3 months interbank rate (Euribor 3m) (National Central Banks and *ECB*)	Monthly Δ
Benchmark		
ESI	*Economic Sentiment Indicator* in France, Germany, Italy, Spain, and euro area (*Eurostat*)	Monthly Δ

TABLE 4.3: Description of variables

The output growth measure considered in this study is the quarterly growth rate of chain-linked Gross Domestic Product as released by the national instituts of statistics of the four countries, namely: INSEE (France), DeStatis (Germany), Istat (Italy), and INE (Spain) and by Eurostat for the euro area at mid-July 2011. We carry out an in-sample analysis over the period 1990q1-2006q4, then we implement a quasi-real-time experience over the crisis period from 2007q1 to 2009q4. Knowing that financial data are available the last working day of the month, we suppose that forecasts for a given quarter are computed at the end of each month, for 12 horizons ranging from $h = 0$ (nowcasts computed at the end of the last month of the reference quarter) to $h = 11/3$ (forecasts computed 11 months before the end of the reference quarter). For each date t, the MIDAS regression optimally exploits the monthly fluctuations of the last $K = 10$ data of the x_t^M series using the Weigth function. For each of the five economies (France, Germany, Italy, Spain, and euro area), we specify three univariate MIDAS regressions based on the three financial variable returns. The direct multi-step forecasting approach used in our work allows parameter estimation using an OLS method and unconstrained Levenberg-Marquardt algorithm on Matlab, for each horizon from $h = 0$ to $h = 11/3$. In order to evaluate the accuracy of those forecasts, we compare them with those stemming from a benchmark MIDAS model based on the Economic Sentiment Index as leading indicator (noted *ESI MIDAS*), a key opinion survey variable to predict output growth, see for example Mourougane and Roma (2003) or Ferrara (2007). As a comparative measure, we present in table 4.4 the ratios of Root Mean Squared Forecasting Errors (RMSFEs) of GDP growth between the *Financial MIDAS* models and the benchmark *ESI MIDAS* model for each h horizon defined by:

$$r^{(h)} = \frac{\text{RMSFE}^{(h)}_{\text{Financial MIDAS}}}{\text{RMSFE}^{(h)}_{\text{ESI MIDAS}}}$$

For a given horizon h, when the ratio $r^{(h)}$ is lower than one, it means that the MIDAS model based on a given financial variable outperforms the benchmark *ESI MIDAS* model and the opposite prevails when the ratio is greater than one (see results in table 4.4).

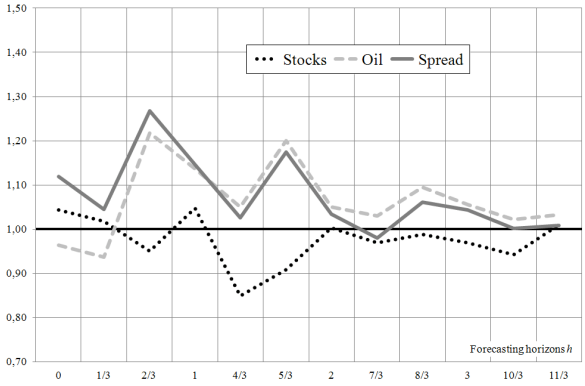
Starting from the results presented in table 4.4 and figure 4.5, we can draw below some conclusions that seem useful for practitioners. First, for all five economies, it turns out that financial MIDAS models are able to improve the benchmark *ESI MIDAS* model for at least one forecasting horizon, although the gain is not uniform through various horizons. In general, the optimal forecast horizon lies between 3 and 5 months, depending on the country: over 4 months for Italy, from 4 to 5 months for Germany

	Forecasting horizons h											
	0	1/3	2/3	1	4/3	5/3	2	7/3	8/3	3	10/3	11/3
France												
Stocks	1,04	1,02	0,95	1,05	0,85	0,91	1,00	0,97	0,99	0,97	0,94	1,01
Oil	0,96	0,94	1,22	1,14	1,05	1,20	1,05	1,03	1,09	1,06	1,02	1,03
Spread	1,12	1,04	1,27	1,15	1,03	1,17	1,03	0,98	1,06	1,04	1,00	1,01
Germany												
Stocks	1,23	1,23	1,09	1,11	0,97	1,06	1,02	1,01	1,01	0,97	0,98	1,01
Oil	1,21	1,19	1,18	1,24	0,99	0,97	1,01	1,03	1,05	1,01	0,98	1,03
Spread	1,17	1,20	1,20	1,23	1,13	1,06	1,02	1,00	1,02	1,00	1,00	1,03
Italy												
Stocks	0,92	0,92	1,02	1,05	0,87	1,01	1,05	1,06	1,07	1,11	1,04	1,04
Oil	0,95	0,95	1,19	1,23	1,02	1,15	1,09	1,18	1,17	1,24	1,14	1,09
Spread	1,00	0,95	1,13	1,23	1,02	1,12	1,12	1,11	1,15	1,19	1,08	1,06
Spain												
Stocks	1,33	1,46	0,96	1,00	0,94	1,11	1,13	1,14	1,18	1,21	1,21	1,33
Oil	1,44	1,41	0,87	0,93	0,87	1,17	1,25	1,26	1,21	1,23	1,23	1,36
Spread	1,38	1,36	0,94	1,14	1,04	1,24	1,25	1,22	1,26	1,28	1,29	1,45
Euro Area												
Stocks	0,91	0,87	1,03	1,05	0,95	0,98	0,98	0,97	0,98	0,97	0,96	0,98
Oil	1,00	0,96	1,21	1,19	1,10	1,06	1,05	1,05	1,11	1,04	1,02	1,02
Spread	1,02	0,99	1,20	1,18	1,11	1,11	1,09	1,02	1,11	1,06	0,98	1,03

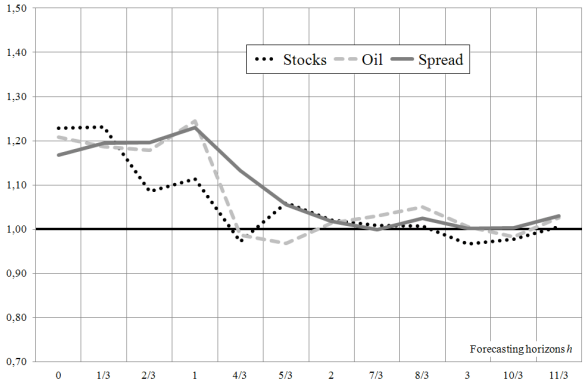
TABLE 4.4: Ratio $r^{(h)}$ of RMSFE for the five economies and the three financial variables

and France, and 2 to 4 months for Spain. This horizon is often encountered as the optimal horizon for financial variables in the empirical literature that deals with the linkages between financial and macroeconomic variables.

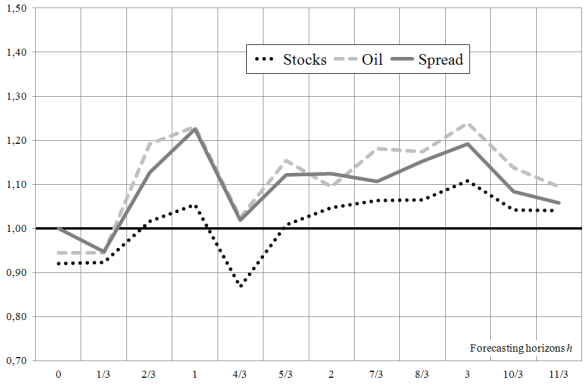
When comparing financial variables according to their forecasting power, it turns out that stock prices generally seem to be the most informative variable in terms of predicting output growth, specially for France, Italy, and the Euro Area. In fact, this Great Recession was initiated by a turmoil on financial markets, equity prices having experienced large falls. Thus this does not seem surprising that stock prices possess a predictive power over macroeconomic evolutions from 2007 to 2009. In opposition, the term spread does not clearly improve forecasting results from the ESI, for all five economic areas. It turns out that this variable is a reliable predictor of turning points in the business cycles, as advocated in many papers (see for example Rudebusch and Williams (2009)), but does not appear as precise predictor of quantitative GDP growth. The same remark holds for oil prices. Indeed, in general, oil prices do not help to improve forecasts by comparison with the ESI, except in the case of Spain for which the oil price variable gives significant better results than the ones obtained from



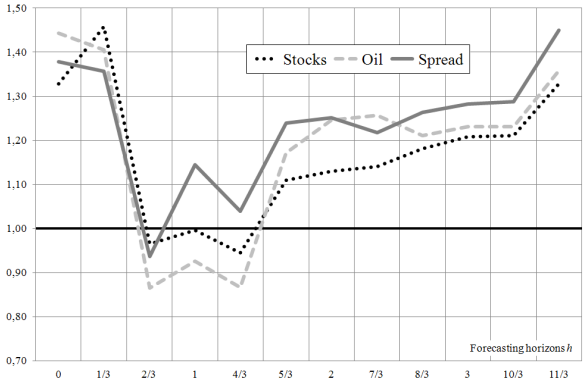
(a) France



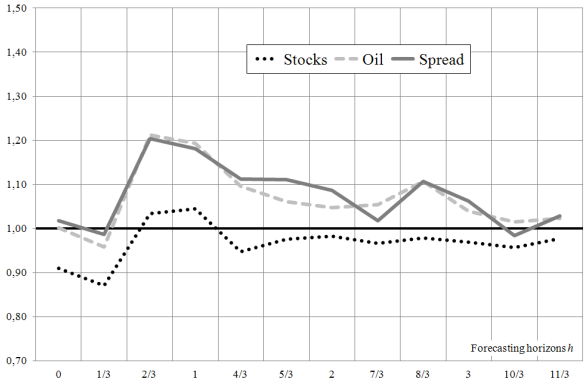
(b) Germany



(c) Italy



(d) Spain



(e) Euro Area

FIGURE 4.5: Evolution of RMSFE ratios $r(h)$ with respect to h

ESI, for forecast horizons ranging from 2 to 4 months (ratios lower than 0.90). Note also that oil prices present a ratio lower than one for Italy and France, for a very short term horizon (the last two months).

Lastly, when comparing countries, it turns out that for France, stock prices provide significantly better results in forecasting GDP than the ESI, for almost each forecast horizon, which is a remarkable result. Also for the Euro Area as a whole, stock prices

lead to an almost systematic gain by comparison with the ESI, except for the 2-month and 3-month horizons. As regards other countries, it is striking to note that those financial variables do not help to improve German GDP forecasts by comparison with the ESI. This result is interesting for practitioners in charge of short-term forecasting for Germany. This means that variables reflecting macroeconomic fundamentals, like for example industrial production index, are likely much more important to get accurate macroeconomic forecasts than financial market variables. For Spain, only oil prices possess a leading pattern between 2 and 4 months. For Italy, we observe a gain as regards stock prices for $h = 0, 1$ and also the ratio is lower than 0.90 for $h = 4$.

C Indepence Metropolis Hastings algorithm for Stochastic Search Variable Selection

To estimate the SSVS-MIDAS model, we implement a Gibbs sampler with respect to specific features due to the mixed data sampling framework[4.4]. The algorithm relies on a few steps which successively sampling h from the spike and slab prior, the hyperparameter ω from a Beta distribution, β and σ from the usual Normal-Inverse Gamma prior, and θ from a candidate generating density using an Independence chain Metropolis-Hastings algorithm. Given initial values for all unknown parameters, the algorithm iteratively updates their values by sampling from their conditional distribution and hence constructing a Markov chain with an invariant distribution.

The algorithm is constructed as follows:

1. Sample $h_i,\ \forall i = 1, \ldots, n$,
   $\pi(h_i|\beta_i, \omega_i) = (1-\omega)\pi(\beta_i; 0, c\varphi^2)I_{\{h_i=c\}} + \omega\pi(\beta_i; 0, \varphi^2)I_{\{h_i=1\}}$,
2. Sample ω from $\mathcal{B}(c_0 + n_1, d_0 + n - n_1)$,
   where $n_1 = \sum_i \mathrm{I}_{\{h_i=1\}}$
3. Sample $\beta_i \sim \mathcal{N}(a_n, A_n)$
   where $A_n^{-1} = \frac{1}{\sigma}\mathsf{X}(\theta)'\mathsf{X}(\theta) + D^{-1}$, $a_n = A_n\frac{\mathsf{X}(\theta)Y}{\sigma}$, and $D = \mathrm{diag}(\phi^2 h_i)$
4. Sample $\sigma \sim \mathcal{IG}(s_n, S_n)$
   where $s_n = s_0 + \frac{T-1}{2}$, and $S_n = \frac{1}{2}(Y - \mathsf{X}(\theta)\beta)'(Y - \mathsf{X}(\theta)\beta)$

[4.4] Comparable algorithms have already been developed in Chapter 3.

5. Sampling θ using an independence chain Metropolis-Hasting algorithm.
   The acceptance probability α to change to the new value θ^{new} drawn from the candidate density, determines whether the chain moves from areas of low posterior probability to high. The acceptance ratio, has been already described in Section 3.1.2.

Repeating 25000 times these 5 steps yields the chain to converge to a steady state. The posterior distribution will allow us to determine the selection with respect to ω. The MATLAB code is available on my website: www.seltenhut.com/clement.marsilli.

D Variable selection

The following figures indicate the variable selection as described by the variable ξ in equation (4.6) with respect to t and according to either the LASSO-MIDAS model or the BAYESIAN-MIDAS model. We also point out the number of factors involved in the FAMIDAS model and the weights of individual predictions in the forecast COMBINATION model.

D.1 Results for $h = 0$

Bayesian-MIDAS ($h = 0$)

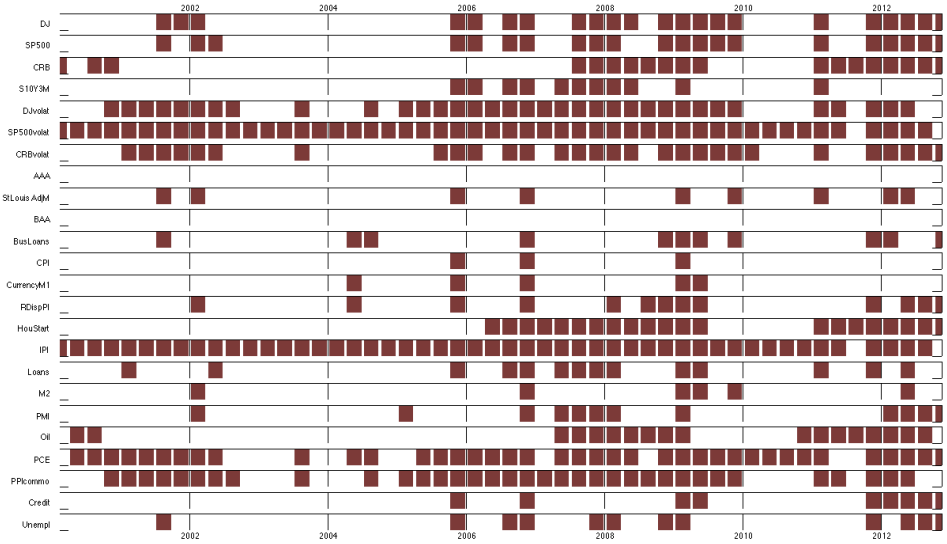


Figure 4.6: Variable selection from 2000q1 to 2012 q4 with the Bayesian-MIDAS model

LASSO-MIDAS ($h = 0$)

Figure 4.7: Variable selection from 2000q1 to 2012 q4 with the LASSO-MIDAS model

FAMIDAS ($h = 0$)

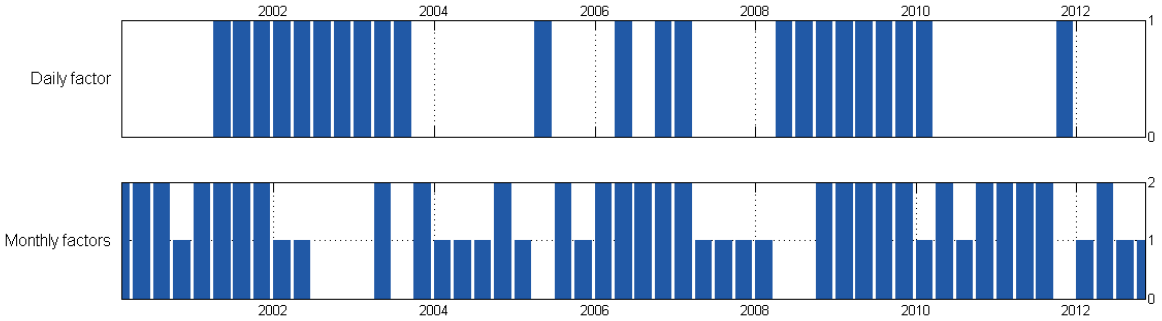


FIGURE 4.8: Variable selection from 2000q1 to 2012 q4 with the FAMIDAS model

FORECAST COMBINATIONS ($h = 0$)

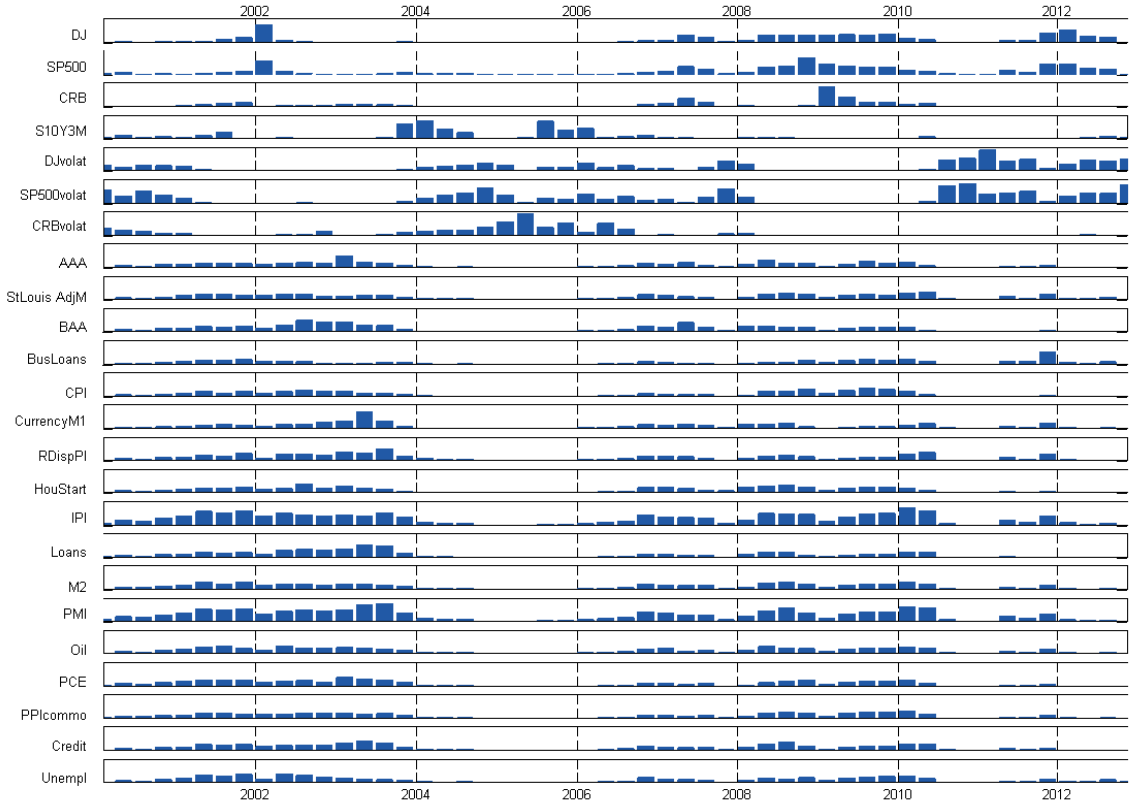


FIGURE 4.9: Weights for each variable of the combination from 2000q1 to 2012 q4

D.2 Results for $h = 3$

Bayesian-MIDAS ($h = 3$)

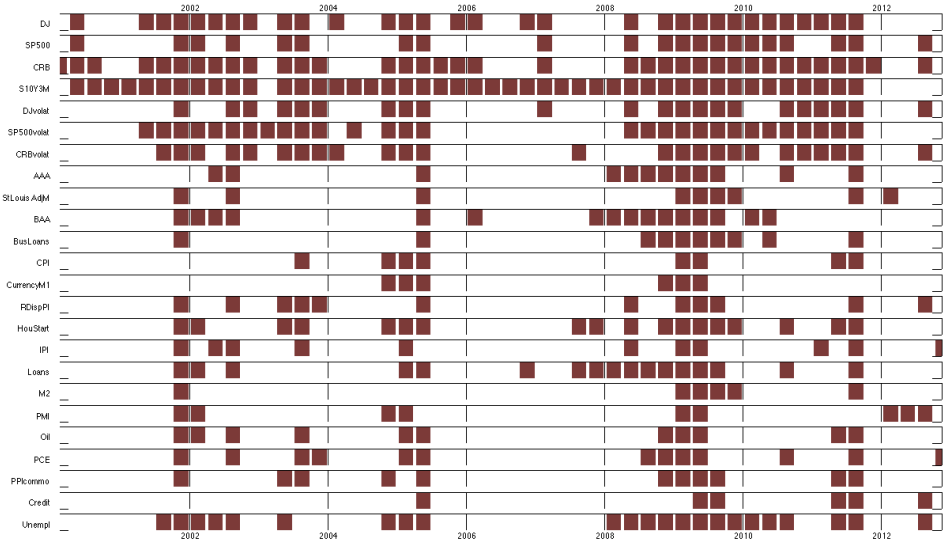


Figure 4.10: Variable selection from 2000q1 to 2012 q4 with the Bayesian-MIDAS model

LASSO-MIDAS ($h = 3$)

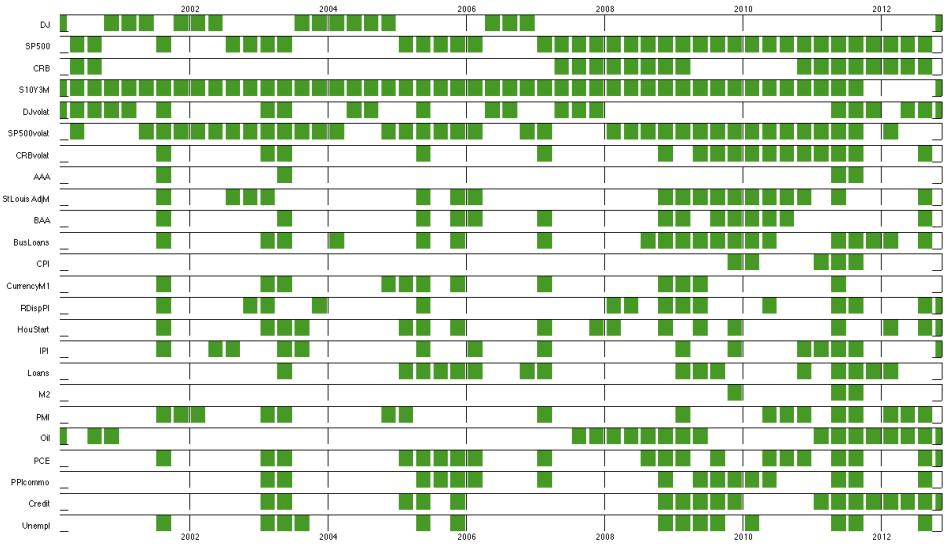


Figure 4.11: Variable selection from 2000q1 to 2012 q4 with the LASSO-MIDAS model

FAMIDAS ($h = 3$)

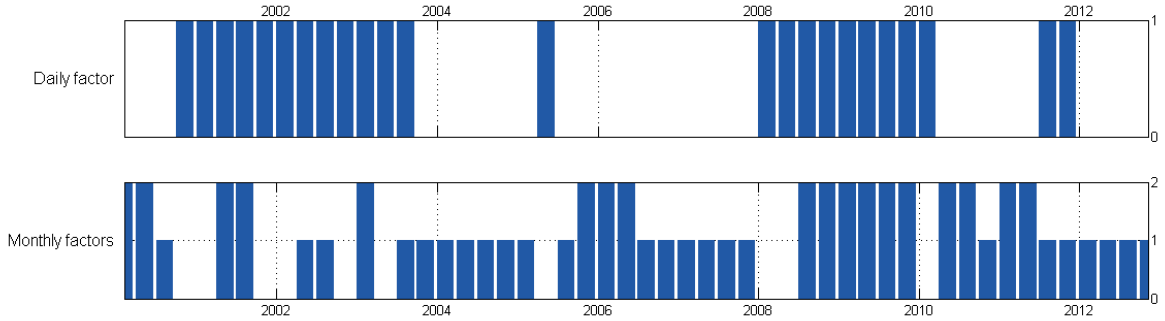


FIGURE 4.12: Variable selection from 2000q1 to 2012 q4 with the FAMIDAS model

FORECAST COMBINATIONS ($h = 3$)

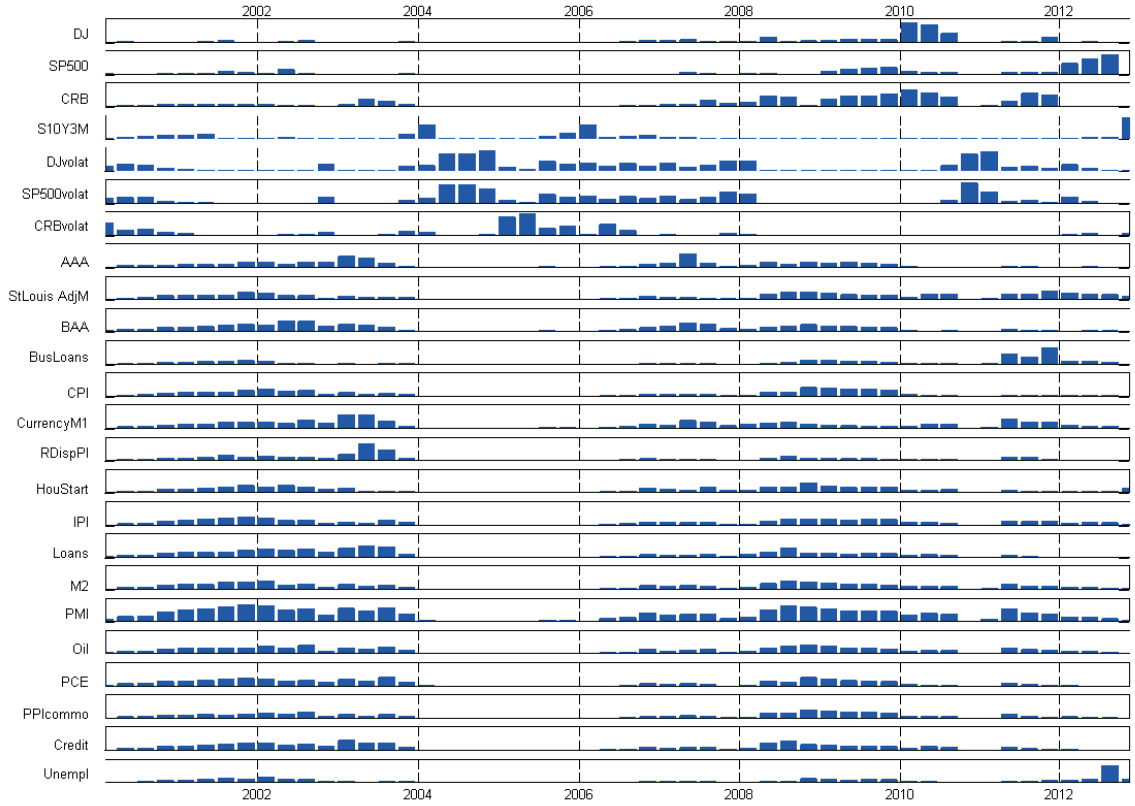


FIGURE 4.13: Weights for each variable of the combination from 2000q1 to 2012 q4

D.3 Results for $h = 6$

Bayesian-MIDAS ($h = 6$)

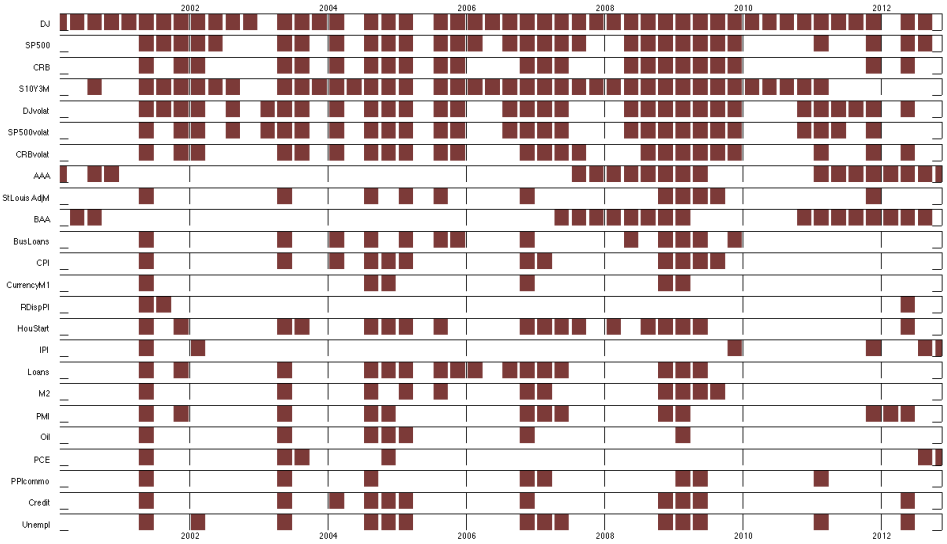


Figure 4.14: Variable selection from 2000q1 to 2012 q4 with the Bayesian-MIDAS model

LASSO-MIDAS ($h = 6$)

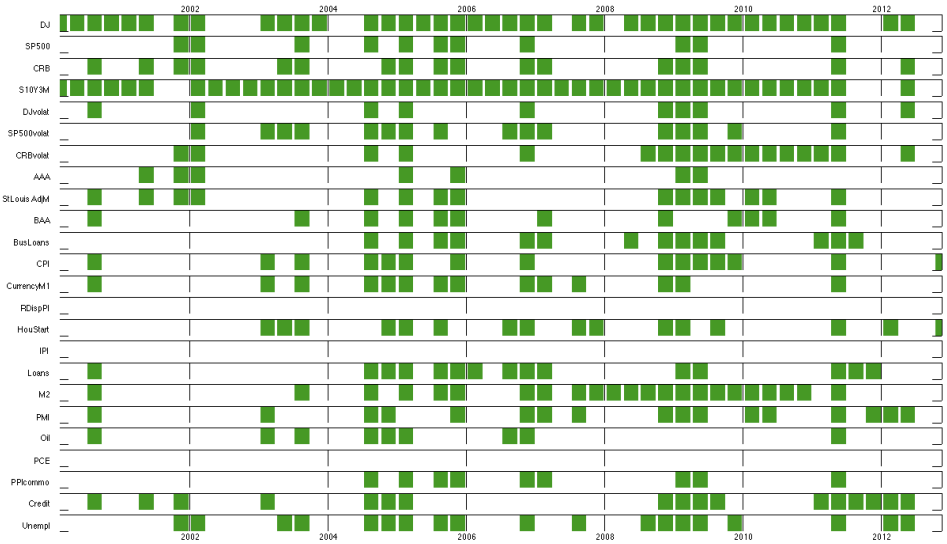


Figure 4.15: Variable selection from 2000q1 to 2012 q4 with the LASSO-MIDAS model

FAMIDAS ($h = 6$)

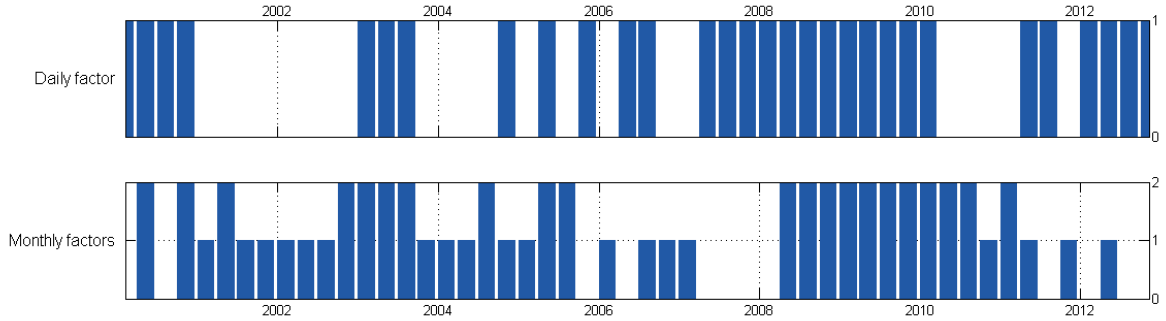


FIGURE 4.16: Variable selection from 2000q1 to 2012 q4 with the FAMIDAS model

FORECAST COMBINATIONS ($h = 6$)

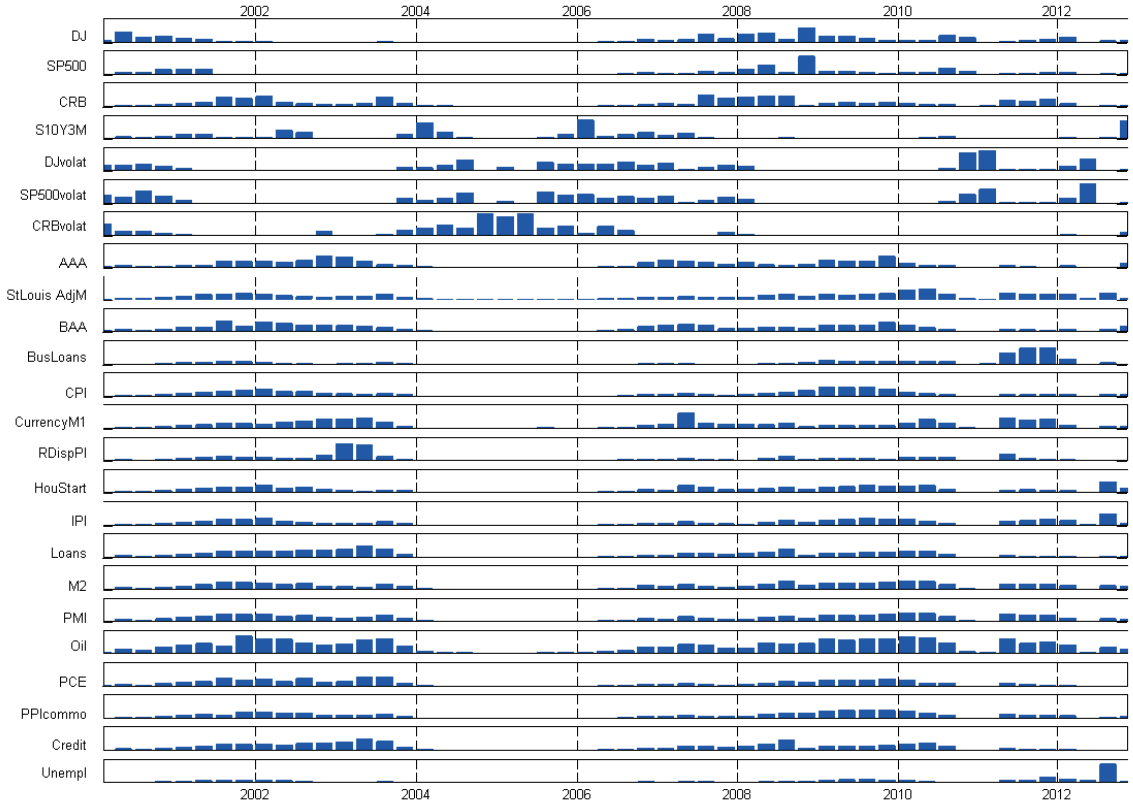


FIGURE 4.17: Weights for each variable of the combination from 2000q1 to 2012 q4

D.4 Results for $h = 9$

Bayesian-MIDAS ($h = 9$)

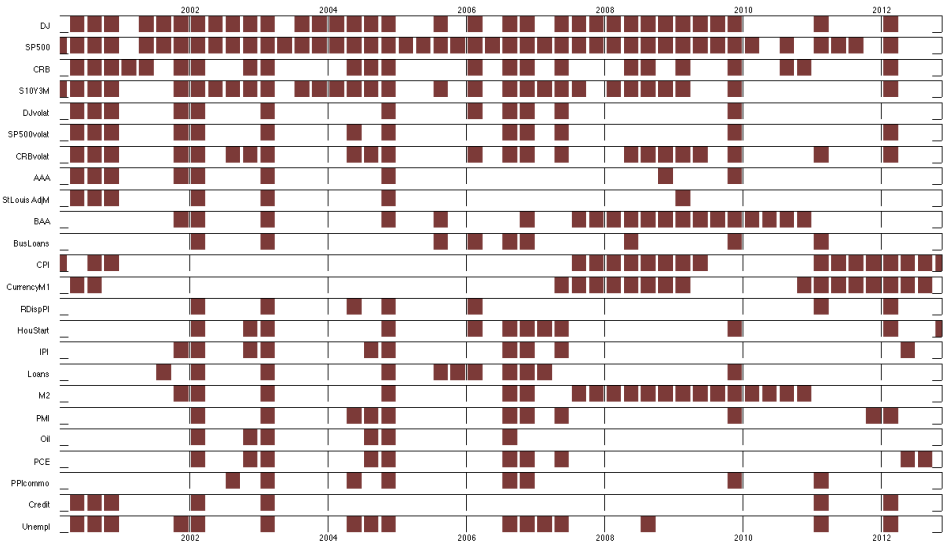


Figure 4.18: Variable selection from 2000q1 to 2012 q4 with the Bayesian-MIDAS model

LASSO-MIDAS ($h = 9$)

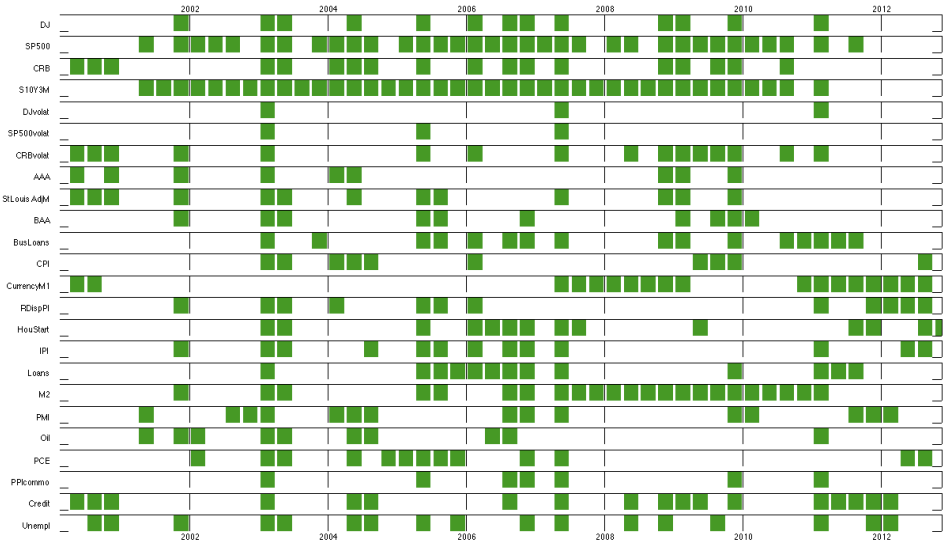


Figure 4.19: Variable selection from 2000q1 to 2012 q4 with the LASSO-MIDAS model

FAMIDAS ($h = 9$)

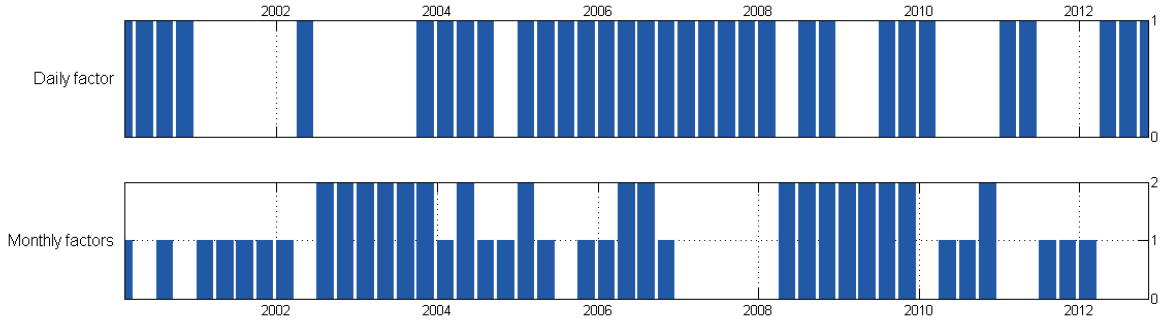


FIGURE 4.20: Variable selection from 2000q1 to 2012 q4 with the FAMIDAS model

FORECAST COMBINATIONS ($h = 9$)

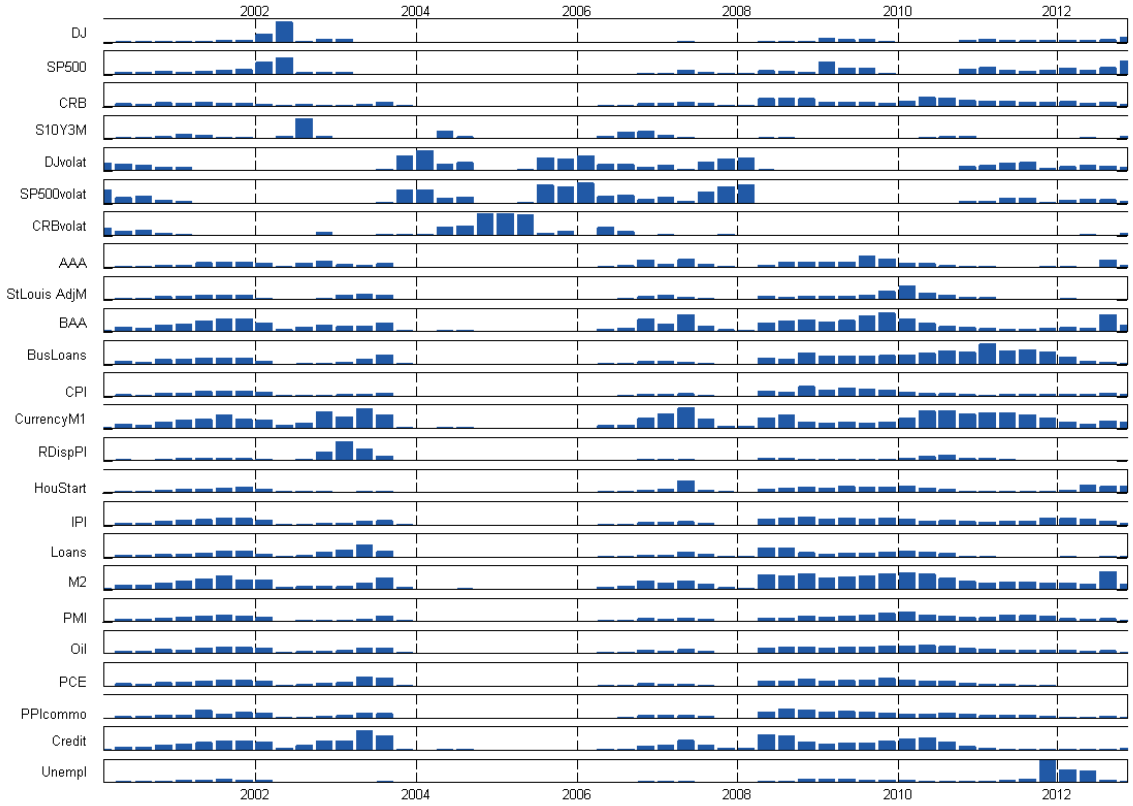


FIGURE 4.21: Weights for each variable of the combination from 2000q1 to 2012 q4

D.5 Results for $h = 12$

Bayesian-MIDAS ($h = 12$)

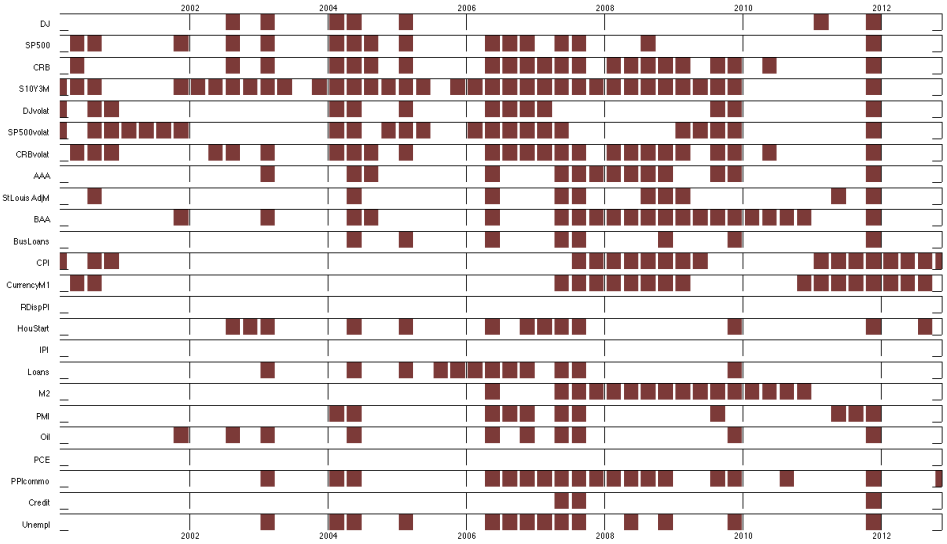


Figure 4.22: Variable selection from 2000q1 to 2012 q4 with the Bayesian-MIDAS model

LASSO-MIDAS ($h = 12$)

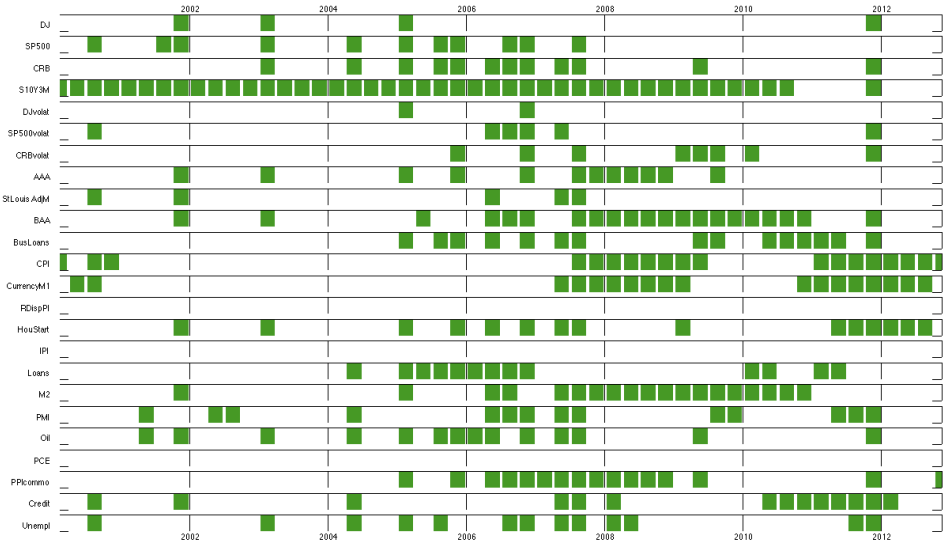


Figure 4.23: Variable selection from 2000q1 to 2012 q4 with the LASSO-MIDAS model

FAMIDAS ($h = 12$)

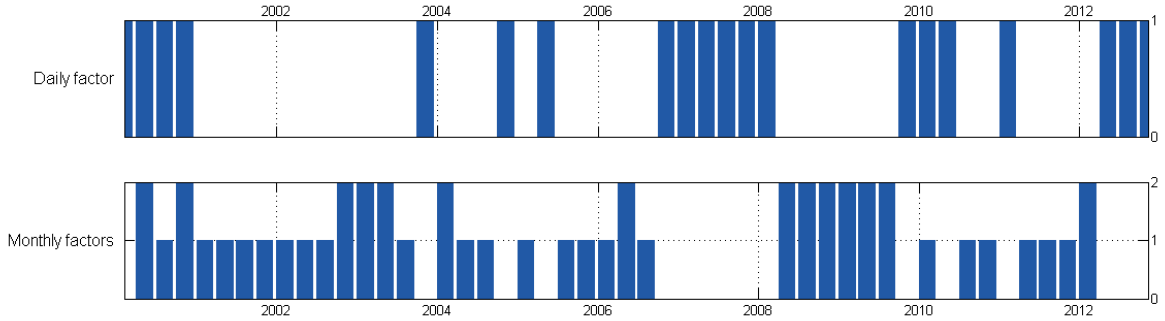


FIGURE 4.24: Variable selection from 2000q1 to 2012 q4 with the FAMIDAS model

FORECAST COMBINATIONS ($h = 12$)

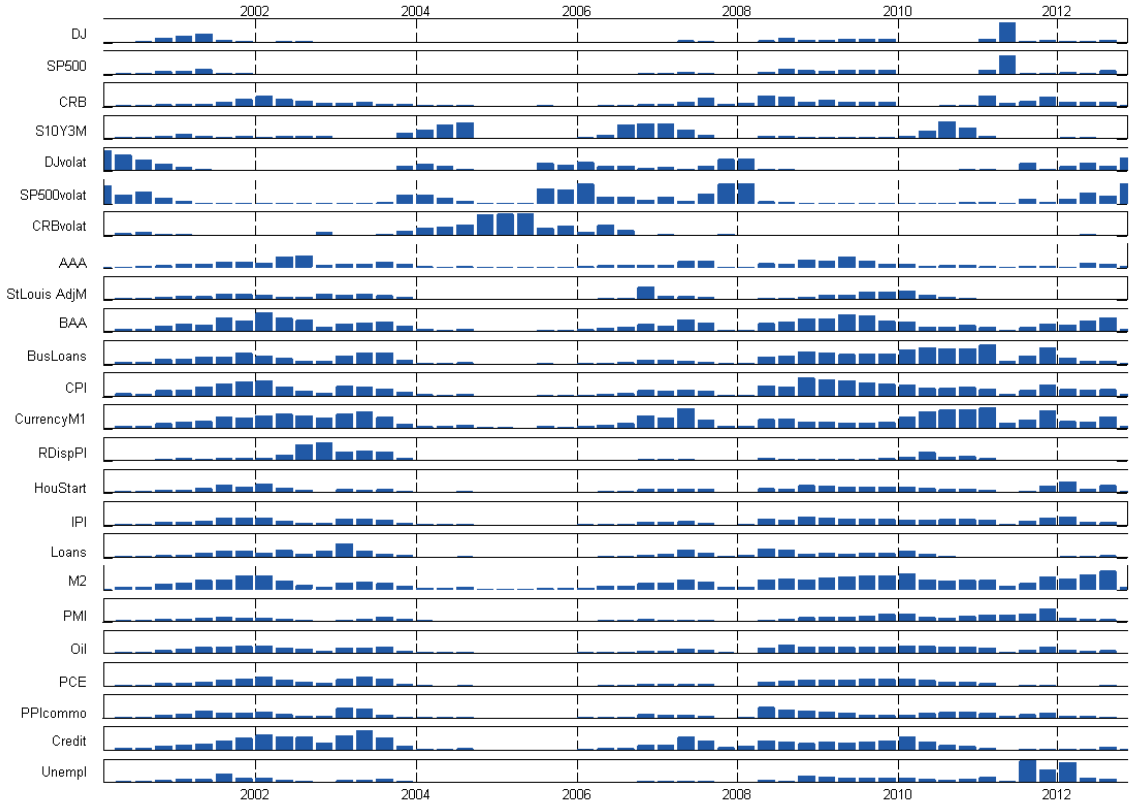


FIGURE 4.25: Weights for each variable of the combination from 2000q1 to 2012 q4

Bibliography

Aastveit, K. A., Claudia Foroni, C., and Ravazzolo, F. (2014). Density forecasts with MIDAS models.

Abraham, B. (1982). Temporal Aggregation and Time Series. *International Statistical Review / Revue Internationale de Statistique*, 50(3):285–291.

Ahn, E. S. and Lee, J. M. (2006). Volatility relationship between stock performance and real output. *Applied Financial Economics*, 16(11):777–784.

Alessi, L., Barigozzi, M., and Capasso, M. (2010). Improved penalization for determining the number of factors in approximate factor models. *Statistics & Probability Letters*, 80(23-24):1806–1813.

Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica*, 33(1):178–196.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158(2):246–261.

Andreou, E., Ghysels, E., and Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business and Economic Statistics*, 31(2):240–251.

Angelini, E., Camba-Mendez, G., Giannone, D., Reichlin, L., and Rünstler, G. (2011). Short-term forecasts of euro area GDP growth. *The Econometrics Journal*, 14(1):C25—-C44.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317.

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2012). Now-Casting and the Real-Time Data Flow.

Barhoumi, K., Benk, S., Cristadoro, R., Reijer, A. D., Jakaitiene, A., Jelonek, P., Rua, A., Rünstler, G., Ruth, K., Nieuwenhuyze, C. V., and Jakaitiene, A. (2008). Short-term forecasting of GDP using large monthly datasets: a pseudo real-time forecast evaluation exercise. Working paper, European Central Bank.

Barhoumi, K., Darné, O., and Ferrara, L. (2010). Are disaggregate data useful for factor analysis in forecasting French GDP? *Journal of Forecasting*, 29(1-2):132–144.

Barhoumi, K., Darné, O., and Ferrara, L. (2013). Testing the Number of Factors: An Empirical Assessment for a Forecasting Purpose. *Oxford Bulletin of Economics and Statistics*, 75(1):64–79.

Barhoumi, K., Darné, O., Ferrara, L., and Pluyaud, B. (2012). Monthly GDP orecasting using bridge models: Application for the French economy. *Bulletin of Economic Research*, forthcomin.

Bellégo, C. and Ferrara, L. (2012). Macro-financial linkages and business cycles: A factor-augmented probit approach. *Economic Modelling*, 29(5):1793–1797.

Bencivelli, L., Marcellino, M., and Moretti, G. (2012). Selecting predictors by using Bayesian model averaging in bridge models.

Blanchard, O. J. and Leigh, D. (2013). Growth Forecast Errors and Fiscal Multipliers. *IMF Working Papers*.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Carnero, M. A., Peña, D., and Ruiz, E. (2012). Estimating GARCH volatility in the presence of outliers. *Economics Letters*, 114(1):86–90.

Carriero, A., Clark, T. E., and Marcellino, M. (2012). Real-time nowcasting with a Bayesian mixed frequency model with stochastic volatility.

Charles, A. and Darné, O. (2005). Outliers and GARCH models in financial data. *Economics Letters*, 86(3):347–352.

Chauvet, M., Senyuz, Z., and Yoldas, E. (2012). What does financial volatility tell us about macroeconomic fluctuations? Working paper, Federal Reserve Board.

Chevillon, G. (2007). Direct multi-step estimation and forecasting. *Journal of Economic Surveys*, 21(4):746–785.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Claessens, S., Kose, M. A., and Terrones, M. E. (2012). How do business and financial cycles interact? *Journal of International Economics*, 87(1):178–190.

Clark, T. E. (2011). Real-Time Density Forecasts From Bayesian Vector Autoregressions With Stochastic Volatility. *Journal of Business & Economic Statistics*, 29(3):327–341.

Clements, M. P. and Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data. *Journal of Business and Economic Statistics*, 26(4):546–554.

Clements, M. P. and Galvão, A. B. (2009). Forecasting US output growth using Leading Indicators: An appraisal using MIDAS models. *Journal of Applied Econometrics*, 24(7):1187– 1206.

Cogley, T. and Sargent, T. J. (2005). Drift and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S. *Review of Economic Dynamics*, 8(2):262–302.

Colacito, R., Engle, R. F., and Ghysels, E. (2011). A component model for dynamic correlations. *Journal of Econometrics*, 164(1):45–59.

Creti, A., Joëts, M., and Mignon, V. (2012). On the links between stock and commodity markets' volatility.

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328.

Dhrymes, P. J. (1971). *Distributed lags: problems of estimation and formulation.* Holden-Day, 1st editio edition.

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(5):371–390.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society*, 57(1):44–97.

Drechsel, K., Giesen, S., and Lindner, A. (2014). Outperforming IMF Forecasts by the Use of Leading Indicators. IWH Discussion Papers 4, Halle Institute for Economic Research.

Duarte, A., Venetis, I. A., and Paya, I. (2005). Predicting real growth and the probability of recession in the Euro area using the yield spread. *International Journal of Forecasting*, 21(2):261–277.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.

Eisner, R. (1960). A Distributed Lag Investment Function. *Econometrica*, 28(1):1–29.

Engle, R. F. (1982). Autoregressive Conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):pp. 987–1007.

Engle, R. F., Ghysels, E., and Sohn, B. (2006). On the Economic Sources of Stock Market Volatility.

Estrella, A., Rodrigues, A. P., and Schich, S. (2003). How Stable is the Predictive Power of the Yield Curve? Evidence from Germany and the United States. *Review of Economics and Statistics*, 85(3):629–644.

Ferrara, L. and Marsilli, C. (2013). Financial variables as leading indicators of GDP growth: Evidence from a MIDAS approach during the Great Recession. *Applied Economics Letters*, 20(3):233–237.

Ferrara, L., Marsilli, C., and Ortega, J.-P. (2014). Forecasting growth during the Great Recession: is financial volatility the missing ingredient? *Economic Modelling*, 36:44–50.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The Generalized Dynamic-Factor Model: Identification And Estimation. *The Review of Economics and Statistics*, 82(4):540–554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2003). Do financial variables help forecasting inflation and real activity in the euro area? *Journal of Monetary Economics*, 50(6):1243–1255.

Foroni, C. (2012). *Econometric Models for Mixed-Frequency Data*. PhD thesis, European University Institute.

Foroni, C., Ghysels, E., and Marcellino, M. (2014). Mixed-frequency vector autoregressive models. *Advances in Econometrics*.

Foroni, C. and Marcellino, M. (2013a). A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. *International Journal of Forecasting*, (2010):1–15.

Foroni, C. and Marcellino, M. (2013b). A survey of Econometrics methods for mixed frequency data. Working Paper 02, European University Institute.

Foroni, C., Marcellino, M., and Schumacher, C. (2013). U-MIDAS: MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society - Series A*, forthcomin.

French, K. R., Schwert, G. W., and Stambaugh, R. F. (1987). Expected stock returns and volatility. *Journal of Financial Economics*, 19:3–29.

Galvão, A. B. (2013). Changes in predictive ability with mixed frequency data. *International Journal of Forecasting*, 29(3):395–410.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society*, 56(3):501–514.

George, E. I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Ghysels, E. (2012). Mixed Frequency Vector Autoregressive Models.

Ghysels, E., Santa-clara, P., and Valkanov, R. (2002). The MIDAS Touch: Mixed Data Sampling Regression Models.

Ghysels, E., Santa-clara, P., and Valkanov, R. (2004). The MIDAS touch : Mixed data sampling regression models. Technical Report 919, mimeo.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2005). There is a risk-return trade-off after all. *Journal of Financial Economics*, 76(3):509–548.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, 26(1):53–90.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676.

Golinelli, R. and Parigi, G. (2013). Tracking world trade and GDP in real time. Temi di discussione (Economic working papers) 920, Bank of Italy, Economic Research and International Relations Area.

Götz, T. B. and Hecq, A. (2014). Nowcasting causality in mixed frequency vector autoregressive models. *Economics Letters*, 122(1):74–78.

Grigoryeva, L. and Ortega, J.-P. (2012). Finite Sample Forecasting with Estimated Temporally Aggregated Linear Processes. *SSRN Electronic Journal*, pages 1–41.

Guérin, P. and Marcellino, M. (2013). Markov-Switching MIDAS Models. *Journal of Business & Economic Statistics*, 31(1):45–56.

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press.

Hamilton, J. D. (2003). What is an oil shock? *Journal of Econometrics*, 113(2):363–398.

Hamilton, J. D. and Lin, G. (1996). Stock market volatility and the business cycle. *Journal of Applied Econometrics*, 11(5):573–593.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second EditionÂă(Google eBook)*. Springer.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994). Bayesian Analysis of Stochastic Volatility Models. *Journal of Business & Economic Statistics*, 12(4):371–389.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Clarendon Press, 3rd editio edition.

Jennrich, R. I. (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Judge, G. G., Griffiths, W. E., Hill, R. C., Lütkepohl, H., and Lee, T.-C. (1985). *The Theory and Practice of Econometrics (Wiley Series in Probability and Statistics)*. Wiley.

Kaufmann, S. and Schumacher, C. (2012). Finding relevant variables in sparse Bayesian factor models: Economic applications and simulation results. Technical Report 29, Deutsche Bundesbank Discussion Paper.

Kilian, L. (2008). The economic effects of energy price shocks. *Journal of Economic Literature*, 46(4):871–909.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies*, 65(3):361–393.

Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting*, 29(1):43–59.

Koyck, L. (1954). *Distributed Lags and Investment Analysis*. Amsterdam: North-Holland Publishing Compagny.

Kvedaras, V. and Zemlys, V. (2012). Testing the functional constraints on parameters in regressions with variables of different frequency. *Economics Letters*, 116(2):250–254.

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer.

Malsiner-Walli, G. and Wagner, H. (2011). Comparing Spike and Slab Priors for Bayesian Variable Selection. *Austrian Journal of Statistics*, 40(4):241–264.

Marcellino, M. (1999). Some Consequences of Temporal Aggregation in Empirical Analysis. *Journal of Business & Economic Statistics*, 17(1):129–136.

Marcellino, M., Porqueddu, M., and Venditti, F. (2013). Short-term GDP forecasting with a mixed frequency dynamic factor model with stochastic volatility. Technical report, Banca d'Italia.

Marcellino, M. and Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics*, 72(4):518–550.

Marcellino, M., Stock, J. H., and Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, 135(1-2):499–526.

Mariano, R. S. and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, 18(4):427–443.

Matheson, T. (2011). New Indicators for Tracking Growth in Real Time. IMF Working Papers 11/43, International Monetary Fund.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Nakajima, J. (2011). Time-Varying Parameter VAR Model with Stochastic Volatility : An Overview of Methodology. *Monetary and Economic Studies*, 29:107–142.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152.

Ng, S. (2012). Variable Selection in Predictive Regressions.

Ng, S. and Wright, J. H. (2013). Facts and Challenges from the Great Recession for Forecasting and Macroeconomic Modeling. *Journal of Economic Literature*, 51(4):1120–1154.

Palm, F. C. and Zellner, A. (1992). To combine or not to combine? issues of combining forecasts. *Journal of Forecasting*, 11(8):687–701.

Perez-Quiros, G. and Timmermann, A. (2001). Business cycle asymmetries in stock returns: Evidence from higher order moments and conditional densities. *Journal of Econometrics*, 103(1-2):259–306.

Primiceri, G. E. (2005). Time Varying Structural Vector Autoregressions and Monetary Policy. *Review of Economic Studies*, 72(3):821–852.

Rodriguez, A. and Puggioni, G. (2010). Mixed frequency models: Bayesian approaches to estimation and prediction. *International Journal of Forecasting*, 26(2):293–311.

Rossiter, J. (2010). Nowcasting the Global Economy. Discussion Paper 12, Bank of Canada.

Rudebusch, G. D. and Williams, J. C. (2009). Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve. *Journal of Business and Economic Statistics*, 27(4):492–503.

Schmidt, P. and Sickles, R. (1975). On the efficiency of the Almon lag technique. *International Economic Review*, 16(3):792–795.

Schmidt, P. and Waud, R. (1973). The Almon lag technique and the monetary versus fiscal policy debate. *Journal of American Stastistical Association*, 68(1):11–19.

Schumacher, C. (2010). Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters*, 107(2):95–98.

Scott, S. L. and Varian, H. R. (2013). Bayesian variable selection for nowcasting economic time series. Technical Report 19567, NBER Working Papers.

Sestieri, G. (2014). Comments on "A comparison of mixed frequency approaches for nowcasting euro area macroeconomic aggregates". *International Journal of Forecasting*, forthcomin.

Shiller, R. J. (1973). A Distributed Lag Estimator Derived from Smoothness Priors. *Econometrica*, 41(4):775–788.

Simpson, P. W., Osborn, D. R., and Sensier, M. (2001). Forecasting UK industrial production over the business cycle. *Journal of Forecasting*, 20(6):405–424.

Sims, C. A. (1974). Distributed Lags. In Intriligator, M. D. and Kendrick, D. A., editors, *Frontiers of Quantative Economics II*. Amsterdam: North-Holland.

Solow, R. M. . (1960). On a Family of Lag Distributions. *Econometrica*, 28(2):393–406.

Stock, J. H. and Watson, M. W. (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97:1167–1179.

Stock, J. H. and Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, 41(3):788–829.

Stock, J. H. and Watson, M. W. (2008). Phillips Curve Inflation Forecasts. Technical Report 14322, NBER Working Papers.

ter Braak, C. J. F. and Vrugt, J. a. (2008). Differential Evolution Markov Chain with snooker updater andÂăfewer chains. *Statistics and Computing*, 18(4):435–446.

Teräsvirta, T. (1980). The polynomial distributed lag revisited. *Empirical Economics*, 5:69–81.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288.

Timmermann, A. (2006). Forecast combinations. In Elliott, G., Granger, C. W., and Timmermann, A., editors, *Handbook of Economic Forecasting*, chapter 4. Elsevier.

Vrugt, J. A., ter Braak, C., Diks, C., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009). Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling. *International Journal of Nonlinear Sciences and Numerical Simulation*, 10:273 – 290.

Yuan, M. and Lin, Y. (2005). Efficient Empirical Bayes Variable Selection and Estimation in Linear Models. *Journal of the American Statistical Association*, 100(472):1215–1225.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Wiley.

Index

Résumé

La prévision macroéconomique à court terme est un exercice aussi complexe qu'essentiel pour la définition de la politique économique et monétaire. Les crises financières récentes ainsi que les récessions qu'ont endurées et qu'endurent aujourd'hui encore, en ce début d'année 2014, nombre de pays parmi les plus riches, témoignent de la difficulté d'anticiper les fluctuations économiques, même à des horizons proches. Les recherches effectuées dans le cadre de la thèse de doctorat qui est présentée dans ce manuscrit se sont attachées à étudier, analyser et développer des modélisations pour la prévision de croissance économique. L'ensemble d'informations à partir duquel construire une méthodologie prédictive est vaste mais également hétérogène. Celle-ci doit en effet concilier le mélange des fréquences d'échantillonnage des données et la parcimonie nécessaire à son estimation. Nous évoquons à cet effet dans un premier chapitre les éléments économétriques fondamentaux de la modélisation multi-fréquentielle. Le deuxième chapitre illustre l'apport prédictif macroéconomique que constitue l'utilisation de la volatilité des variables financières en période de retournement conjoncturel. Le troisième chapitre s'étend ensuite sur l'inférence bayésienne et nous présentons par ce biais un travail empirique issu de l'adjonction d'une volatilité stochastique à notre modèle. Enfin, le quatrième chapitre propose une étude des techniques de sélection de variables à fréquence multiple dans l'optique d'améliorer la capacité prédictive de nos modélisations. Diverses méthodologies sont à cet égard développées, leurs aptitudes empiriques sont comparées, et certains faits stylisés sont esquissés.

Mots-clés

Économétrie, série temporelles, modèle de prévisions, macroéconomie internationale, modélisation multi-fréquentielle, MIDAS.

Abstract

Economic downturn and recession that many countries experienced in the wake of the global financial crisis demonstrate how important but difficult it is to forecast macroeconomic fluctuations, especially within a short time horizon. The doctoral dissertation studies, analyses and develops models for economic growth forecasting. The set of information coming from economic activity is vast and disparate. In fact, time series coming from real and financial economy do not have the same characteristics, both in terms of sampling frequency and predictive power. Therefore short-term forecasting models should both allow the use of mixed-frequency data and parsimony. The first chapter is dedicated to time series econometrics within a mixed-frequency framework. The second chapter contains two empirical works that sheds light on macro-financial linkages by assessing the leading role of the daily financial volatility in macroeconomic prediction during the Great Recession. The third chapter extends mixed-frequency model into a Bayesian framework and presents an empirical study using a stochastic volatility augmented mixed data sampling model. The fourth chapter focuses on variable selection techniques in mixed-frequency models for short-term forecasting. We address the selection issue by developing mixed-frequency-based dimension reduction techniques in a cross-validation procedure that allows automatic in-sample selection based on recent forecasting performances. Our model succeeds in constructing an objective variable selection with broad applicability.

Keywords

Econometrics, time series, forecasting, international macroeconomics, mixed-frequency models, MIDAS.

Classification JEL

C53, E37.

Classification AMS

91B, 62J, 62M10, 62P25.