Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operational Research

Nowcasting GDP using a nonlinear mixed-frequency factor approach

Menno Smit (580595)

Supervisor:	P. A. Opschoor, PhD candidate
Second assessor:	dr. P. Vallarino
Date final version:	26th May 2025

Abstract

Nowcasting quarterly GDP offers insight into the current economic situation, which is essential as first estimates from official institutions are released at least a month after the quarter's end. For this purpose, we introduce a nonlinear mixed-frequency factor approach. Monthly factors, derived from an expansive set of macroeconomic indicators, are obtained using Sigmoid and RBF kernel principal component analysis (kPCA). Up to six factors, along with up to six factor lags, up to one lag of GDP, and an intercept are modelled using (i) Factor-Bridge and (ii) Factor-MIDAS regressions, leading to a variety of nonlinear models. These models are compared to their linear counterparts, which use principal component analysis (PCA) or sparse principal component analysis (SPCA) for factors instead. The data consists of quarterly vintage data for US GDP and monthly vintage data from the FRED-MD dataset. From January 2010 until March 2024, the nowcasting accuracy of the different models is evaluated and compared. This is done specifically for different conjunctural periods including a relatively stable period, and the COVID-19 financial crisis. Moreover, the effect of data availability within the quarter on the different nowcasts is being investigated. Finally, the robustness of the results to data revisions is examined. The results show that in several scenarios, nonlinear models increase nowcasting accuracy and lead to a decreased mean squared prediction error (MSPE) of up to 50%. In other situations, conventional linear methods outperform. It is concluded that future research could further examine the optimal conditions for kPCA in nowcasting.

1 Introduction

In macroeconomics, the practice of making short-horizon predictions, or *nowcasting*, has gained effectiveness with the growing amount of economic data. Nowcasts give insight into the current status of *flow statistics*, which measure the flow of a quantity over a specific period. One important flow statistic is the *gross domestic product (GDP)*, which theoretically measures the amount of economic value added over a period. In practice, it can be seen as an indicator of conjunctural status, which is measured on quarterly and yearly basis. For the US GDP in particular, a first official estimate is usually released one month after the quarter's end, and new estimates and revisions are published months later when more data becomes available (Bureau of Economic Analysis, 2023). This makes the GDP figures backward-looking, and not reflecting the current status of the economy. With this lack, the challenge of making accurate GDP nowcasts found its way into literature studies.

For many parties, the current status of GDP is highly relevant. One could think of the great increase in data-driven decisions that are made nowadays. These decisions often affect myriad people, as is the case, for instance, in governments. Also, central banks such as the Federal Reserve Board (FED) or European Central Bank (ECB) base their policies on these figures, which highly impact the economy and therefore the overall well-being of civilians. As for instance utilised by Bernanke, Boivin and Eliasz (2005), making decisions based on forecasts of GDP can be highly valuable and informative in these situations, as they provide a reflection of current GDP. Additionally, commercial banks, which are a crucial economic pillar of the financial system, need to be able to make proper decisions, decisions which incorporate proper safety margin for the level of risk at hand. This means, for instance, in times of economic uncertainty, keeping sufficient liquidity on the balance sheet. Finally, many individuals make

daily decisions based on these figures, since consumer confidence often drives their consumption, saving, and investment behaviour and is inherit with their own safety.

This research adds to the topic of nowcasting by attempting to nowcast the US GDP using a nonlinear mixed-frequency factor approach. In particular, it examines whether using nonlinear factors, extracted using *kernel principal component analysis (kPCA)*, provide enhanced nowcasting accuracy compared to conventional linear principal components used as factors.

With the increase of data availability, econometric models have gained more potential of capturing the true pattern of the economy using more sophisticated methods suitable for big data. In particular, there has been much attention for *artificial intelligence (AI)* methods in the literature, which can explain data nonlinearly to the extent where they risk the danger of 'overfitting' the training data, which leads to decreased out-of-sample accuracy. Nti, Adekoya and Weyori (2020) show in an extensive study evaluating different *machine learning (ML)* models on their predictive performance in large datasets that *deep learning (DL)* provides the best results for high-dimensional scalar predictions. Its limitation, however, is its inability to deal with high dimensionality (Kapetanios, Papailias et al., 2018). Using a large set of macroeconomic variables to make predictions based on, often even up to 250 macroeconomic indicators, can provide significant benefit due to the nature of macroeconomic variables being that they provide comprehensive information (Stock & Watson, 2002b). This makes DL methods suboptimal for the task. Also standard statistical methods such as *Ordinary Least Squares (OLS)* fail in this high-dimensional context, since they require significantly more observations than regressors to make reasonable estimates of parameters, previously called the 'curse of dimensionality', as for instance discussed by Fan and Li (2001). In macroeconomic context, traditional statistical methods such as *Factor models* manage to capture most of the variation in the regressors set in a few latent factors, allowing them to outperform many AI models as was shown by Kim and Swanson (2018). This is possibly due to their underlying assumption, which is that a few latent factors can explain most of the common patterns in a set of variables, being closely related to empirics. Therefore, the relaxation of the assumption when shifting to AI-based models shifts the use of information partly to uncover this relationship, rather than on more specific information extraction.

Macroeconomic factor models often use *principal component analysis (PCA)* to extract factors. These linear *principal components (PCs)* explain most of the variation in the regressors set by forming orthogonal linear functions of the regressors, maximising their variance explained. Their assumed linear functional form is exactly where their limitation lies: when explanatory variables contain nonlinear common patterns, these are not captured by the PCs, making them suboptimal as factors. As an alternative, Schölkopf, Smola and Müller (1997) introduced the concept of nonlinear *kernel principal components (kPCs)*, with their later work illustrating the use of these 'kPCs' in practice (Schölkopf, Smola & Müller, 1998). Kutateladze (2022) as one of the first uses the kPCs in the macroeconomic context, using monthly-based kPCs to forecast several monthly-based macroeconomic processes. He shows that kernel principal component analysis is a consistent estimator for all kernels with finite feature spaces and for several kernels containing infinite feature spaces, with in particular the *Radial Basis Function (RBF)* and *Sigmoid* kernels. In addition, he shows that these two kPCs nest linear PCs for certain hyper-

parameter restrictions, allowing kPCA to approximate PCA in short prediction horizons, for which the PCA often shows dominant results (Stock & Watson, 2002a). Also, he shows the kPCA's allowance for nonlinearity provides significant benefit for longer-horizon predictions, from up to three months in the future and more. However, what to the best of my knowledge has not yet been investigated is the use of monthly kPCs in the setting of mixed-frequency nowcasting, which could uncover nonlinear patterns in the data that allow for high predictability of quarterly GDP. The question this paper therefore examines is as follows:

Do nonlinear monthly factors enhance nowcasting accuracy of quarterly GDP?

The paper investigates this question by first replicating the results from Kutateladze (2022) and subsequently utilising monthly kPCs in nowcasting setting. The effectiveness of the kPCs is compared to linear PCs as a benchmark. This entails additionally to standard PCA-extracted factors, *sparse principal components (SPCs)* as a second linear reference. The factors are combined with the (i) *factor-Bridge* and (ii) *factor-MIDAS* regressions, of whom the accuracy in nowcasting is evaluated from January 2010 until March 2024. In addition, it is examined how different conjunctural periods, such as the COVID-19 financial crisis, influence the result. Also, this paper investigates how data availability associated with the month in the quarter influence the nowcasting accuracy, as typically more information should lead to enhanced results. Finally, it examines the robustness of the results to data revisions. The data that is used contains monthly vintage data by McCracken and Ng (2016), which is the same data used by Kutateladze (2022) for comparison ways. Furthermore, quarterly vintage GDP data is obtained from the Federal Reserve Bank of Philadelphia (2024).

The results show that kPCA can reduce the mean squared prediction error (MSPE) for a subset of scenarios. While the combination between Bridge and PCA has the lowest MSPE for the total out-of-sample period and for the first and third months of quarters, and while the combination of PCA and MIDAS Beta dominates during the COVID-19 crisis, the combination of Sigmoid kPCA and MIDAS Exponential Almon performs best over the full sample, excluding COVID-19. Additionally, the combination between RBF kPCA and MIDAS Beta performs best during second months of quarters. Also, for the sub-sample 2015-2018, which represents a calm period within the out-of-sample, the combination of Sigmoid kPCA and Bridge leads to a 50% lower MSPE compared to its best linear competitor. The previous findings suggesting that kPCA leads to improved nowcasting accuracy within a subset of scenarios in the nowcasting setting.

The paper is structured in the following way. Section 2 discusses the existing literature of nowcasting, of factor models, and of the kernel trick, and ties the different knots together. Then, Section 3 discusses the data, which requires extra precision when nowcasting. Section 4 follows with a detailed description of all methods, and mathematically clarifies each concept. The outcome of the research, which is discussed via various tables and figures, is then discussed in Section 5. Finally, Section 6 concludes on the paper and discusses limitations and possible angles for future research.

2 Literature review

'The prediction of the present, the very near future and the very recent past' is how the ECB defines nowcasting (Bańbura, Giannone & Reichlin, 2010). Nowcasting originally has its roots in meteorologic applications. In these applications, the goal of making short-term weather predictions using past and current available information, which arose as a challenge somewhere around the 1970s, was addressed in possibly the first book about nowcasting by Browning (1982).

Application of nowcasting in macroeconomic context arose later, arguably with the paper by Stock and Watson (1989), in which economic indicators as regressors on either monthly or quarterly basis were used to make inferences about macroeconomic statistics such as employment and GDP, which have the same frequency. The invention of Sims (1980), who introduced the *vector autoregressive model (VAR)*, added significant value to macroeconomic forecasting literature, as many macroeconomic time series models could be built as a corollary. Following upon this, factor models began being used frequently for macroeconomic predictions in the early 2000s, in particular after the release by Stock and Watson (2002a), who dove specifically into *dynamic factor models (DFMs)* and using *principal component analysis (PCA)* to form them. Although different methods have also been used in recent years for macroeconomic forecasts, the factor models remain, as Goulet Coulombe, Leroux, Stevanovic and Surprenant (2022) point out, 'the best regularisation'. Also, they claim that nonlinearities are the 'true game changer' for macroeconomic predictions, which brings us to the next point.

The kernel trick, which stems originally from the 1960s (Aizerman, 1964), can be used to model nonlinearities in a computationally efficient way, while maintaining the same feature extraction as *neural network (NN) methods* when appropriate kernels such as Sigmoid or RBF are used (Schölkopf et al., 1997). This clearly shows their high relevance in macroeconomic context, since the result by Nti et al. (2020) shows that DL methods, which is a variation of NN, are the most potent AI methods in big data scalar predictions. As Kapetanios et al. (2018) show, DL methods are not potent when not enough observations exist compared to the number of regressors, which is often the case in macroeconomic high-dimensional data settings. The kernel method involves using a kernel function to calculate the similarity between two observations in a higher-dimensional feature space without explicitly having to calculate the feature space itself. This is convenient as higher dimensions allow more features per observations to be taken into account. This means that original observations can be distinguished linearly in higher-dimensional settings, which would require a nonlinear distinguish function in the original data space. In the context of predicting monthly macroeconomic processes, Kutateladze (2022) showed that for monthly Sigmoid and RBF kPCs in a factor-augmented model that 'one of these two approaches dominates the others in nearly 95% of the cases considered.' In addition, he concludes that the gains of using monthly kPCs is not significant compared to regular PCA for the shortest prediction horizons.

In the field of nowcasting, where originally single frequency models were the standard (Stock & Watson, 1989), higher frequency data was until long left unexploited. It was noted later, however, that using mixed-frequency regressors could enhance nowcasting accuracy through effective modelling and that the recent information could have significant predictive power for nowcasting (Ghysels, Sinko & Valkanov, 2007). The modelling of mixed-frequency data also

comes with a challenge, since classical regression techniques such as the VAR model, which is still often used in *time series analysis* (Hamilton, 2020), can suffer in this context due to not adequately dealing with the difference in variance of regressors. This is because higher-frequency variables typically contain more noise than lower-frequency ones. An additional complication with increasing the frequency is the number of variables having missed observations, which, in the context of nowcasting when variable releases are asymmetrically, introduces the *ragged edge* problem (Jazwinski, 1970).

A few solutions for the first problem were formed throughout the years. A first solution is the use of 'aggregation', which was first investigated in the context of *Bridge equations (BE)*, with one early implementation by Parigi and Schlitzer (1995) for Italian monetary policy. Later, Baffigi, Golinelli and Parigi (2004) showed via a more comprehensive empirical analysis that the BE significantly outperformed compared to existing models at the time for many different economic geographic systems. One theoretical limitation of the BE models, however, is that they aggregate information, which could result in the fact that 'a lot of potentially useful information might be destroyed, and mis-specification inserted in the model' (Foroni & Marcellino, 2013). A model that was created to avoid this theoretical limitation and allows for more flexibility in data handling is the *mixed-data sampling (MIDAS)* model, with it's first uses in nowcasting by Ghysels, Santa-Clara and Valkanov (2004). One relevant method called *Factor-MIDAS* by Marcellino and Schumacher (2010) incorporates the factor approach into the MIDAS structure, making it a relevant method in macroeconomic nowcasting context.

For the second problem, a few approaches have been used in the past to deal with the ragged edge problem, as discussed by Marcellino and Schumacher (2010). A first approach is using *data filtering techniques* to create a stable dataset. Firstly, Altissimo, Cristadoro, Forni, Lippi and Veronese (2010) uses *vertical alignment* in combination with a DFM model to obtain economic indicators for estimating GDP. Secondly, the *expectation-maximisation (EM)* algorithm fills in likely values for missing observations based on an iterative PCA estimation procedure (Dempster, Laird & Rubin, 1977). A second approach estimates an initial 'hidden factor state', which it updates with each piece of known data, which is referred to as the Kalman Filter (Kalman, 1960). This approach is efficient and consistent (Jazwinski, 1970), but does require heavy parametric assumptions (Bai, Ghysels & Wright, 2013). Also, it requires the exact representation of the data space, which is infeasible for infinite-dimensional data spaces.

Bringing everything together, the significant results by Kutateladze (2022) for macroeconomic predictions, and the progress made in adequate model-building for mixed-frequency nowcasting, pulls our attention to the use of monthly kPCs to nowcast quarterly GDP. As Kutateladze (2022) showed, monthly kPCs do not enhance nowcasting accuracy compared to standard linear PCs in single-frequency single-horizon monthly predictions. This raises the question whether monthly kPCs may be able to uncover a monthly nonlinearity pattern that can be used for mixed-frequency nowcasting of quarterly GDP.

We hypothesise that kPCA will enhance nowcasting accuracy in all different model combinations compared to linear PCA and SPC due to its flexibility to nonlinearities in the data. In Factor-Bridge, the advantage might be reduced compared to Factor-MIDAS due to MIDAS's modelling flexibility which might keep more nonlinearity in the factors intact. Within financial

crises, like COVID-19's financial crisis in 2020–2021, we expect kPCA to have a larger benefit as crises might contain additional nonlinearities compared to calmer conjunctural periods. This last intuition stems from the findings of Goulet Coulombe et al. (2022), who show that nonlinear methods like the kernel Ridge regression and random forest model lead to most improvements compared to linear variants during economic instabilities when looking at medium to high forecast horizons.

3 Data

The *data gathering process* is of utmost importance for the analysis since it drives the results of the paper. It consists of both selecting data, transforming data, dealing with missed observations, and potentially removing outliers from the dataset, which will now be discussed in detail.

3.1 Regressor dataset

The regressor data is that of McCracken and Ng (2016), which is commonly referred to as the *FRED-MD* data set. Their dataset contains monthly observations of up to 136 macroeconomic variables, starting from January 1959 up to March 2024, at the moment of writing. The dataset is used more frequently for macroeconomic research since it contains many possible explanatory variables for different economic processes. Furthermore, it is updated frequently and managed with care, such that variables are replaced or calculated in revised ways to make them most relevant. Specifically, the FRED-MD contains monthly vintage data from August 1998 onwards. Vintage data is especially useful in the nowcasting setting, since a revised dataset from later periods creates an information bias, which results from not fairly incorporating actual knowns at the nowcasting time (Giannone, Reichlin & Small, 2008). The data vintages in FRED-MD contain several missing observations, which is natural in the nowcasting setting due to differing release dates. Table 5, in Appendix B, contains the different variables in the dataset and their release lags, as we found when we looked closely at the different vintage datasets. In this context, we refer to *immediate availability* whenever the variable is publicly available within the next month's vintage data, as all variables in the dataset require at least some time to calculate.

3.2 Dependent variable dataset

The dependent variable in our research is the GDP. Vintage data for GDP from November 1965 up to May 2024, at the moment of writing, can be found at Federal Reserve Bank of Philadelphia (2024). All vintages start from January 1947 and end in their release month. Lagged GDP is most cases immediate availability within the vintage data of the month after. In rare cases vintage data contains a missing observation at the end of the vintage, which we imputed using an AR(1) model that takes the last known previous value.

3.3 Data transformation and outlier removal

The monthly data is transformed as suggested by McCracken and Ng (2016). This is based on more frequently used transformations for certain types of macroeconomic variables, accounting

for any non-stationarity in them. The GDP is transformed using the log difference, which approximates the growth rate, similarly as discussed by McCracken and Ng (2020).

No outliers have been removed from the dataset. Primarily, outliers can be the most informative observations when looking at which models most adequately capture the true underlying pattern of the data. Also, outliers might contain nonlinearity, which could give kPCA an edge over PCA. A robustness check in which the COVID-19 financial crisis is excluded is performed later to evaluate the impact of these potential outliers.

3.4 Ragged edge

Missing observations in both monthly and quarterly data are common in the nowcasting setting, due to the unevenly distributed release times, which is referred to as the ragged edge setting. The ragged edge setting, when not accounted for, causes the failure of estimation methods and the inability to make a nowcast at each moment during the quarter for the next GDP figure.

In literature, several methods to address this problem have been used, which were mentioned in Section 2. This study applies the EM algorithm to impute missing data in the vintages.

Firstly, it allows the use of kPCA as a last step on the imputed data, keeping intact their nonlinearity. This is useful as the other methods use an initial factor estimation on fully known data, and linearly update the factors, which can hide the nonlinearity given by kPCA. Secondly, if one would attempt to update them nonlinearly using the higher-dimensional feature matrix, this would require its explicit formula, which is exactly what the kernel trick avoids, and would require another 'kernel trick' for updating. Thirdly, the EM algorithm uses PCA estimation to predict macroeconomic variables over short horizons, which Kutateladze (2022) showed to be potent. Finally, the EM algorithm allows many factors to capture significant trends that take into account the recent past of the variables when predicting their current status.

The implementation we use is the algorithm by Bennett (2023), for which we chose a maximum number of 12 factors, which we based on the square root of factor rule (Jolliffe, 2002). This rule of thumb is used to determine how many factors to use based on the number of variables in the dataset. Later, the results will show whether nowcasts driven on more imputations will suffer in nowcasting accuracy or not.

4 Research methodology

The methodology consists of several parts. Initially, it discusses methodology related to factor models and linear PCs. Subsequently, it reviews the kernel trick and its application in the domain of PCA, which resulted in kPCA. After which, it covers the (i) factor-Bridge and (ii) factor-MIDAS regressions, for which the different principal components can be used. Then, notation for the different resulting models is introduced. Finally, the research setup and estimation procedure are discussed.

4.1 Factor models

Often in time series analysis, when one tries explaining a *dependent variable* y, of dimension $T \times 1$, using a *regressor set* X, of dimension $T \times K$, it is unclear which variables X should consist

of as adequate predictors. Particularly in macroeconomics, where many different variables exist and where many have unique explanatory power, it is important to account for most of the variance in these variables while at the same time limiting the number of regressors K relative to the number of observations T. This is important as high-dimensional models often overfit the training data. Factor models use a limited number of factors to explain most of the variance in the regressor set. These factors can subsequently be used as the new regressors.

Static factor models, in which factor loadings are assumed to be constant over time, can be mathematically written as

$$X = \underset{(T\times R)}{F}\underset{(R\times K)}{\Lambda'} + \underset{(T\times K)}{e},$$

in which R the number of factors, in which F is the *factor matrix* containing the common factors over time, and in which Λ represents the *factor loading matrix*.

When using PCA, the factors are extracted by minimising $_{F,\Lambda}\|X - F\Lambda'\|_F^2$, under the identification constraints $\Lambda'\Lambda = I_R$ and $F'F$ diagonal (Kutateladze, 2022), in which $\|.\|_F^2$ denotes the *Frobenius norm*. The resulting factors are the linear PCs, and the factor loadings are the eigenvectors of the covariance matrix of X.

In the case of SPC, when there is need for a more sparse factor loading matrix, there is an additional $L1$ penalty term placed on the factor loadings in the objective function, in which under the same identification criteria $_{F,\Lambda}\left\{\|X - F\Lambda'\|_F^2 + \lambda\|\Lambda\|_1\right\}$ is minimised. In the expression, $\lambda \geq 0$ can be used as a hyperparameter to decide on the sparsity of the matrix Λ. A sparse factor loading matrix can be preferably when one wants focus on only the most key patterns in the data.

4.2 Kernel principal component analysis

The kPCs are derived by applying the kernel trick on the regressor set X. In the next two sections, first the kernel trick will be reviewed and subsequently its application in PCA regard.

4.2.1 Kernel trick

The kernel trick is about measuring similarity between observations of X in its higher dimensional *feature space* $\phi(X)$, in which $\phi(X) = \begin{bmatrix}\phi(X_1), & \phi(X_2), & \cdots & \phi(X_T)\end{bmatrix}'$ is of dimension $T \times M$, and M is the number of features per observation, which can be up to unlimitedly large (Aizerman, 1964). In this higher-dimensional space, linear distinction between observations is possible, which would not have been possible in their original observational space. The similarities between data points i and j, for which $i, j \in \{1, 2, \ldots, T\}$, are measured by kernel functions $K_{i,j} = k(X_i, X_j) = \phi(X_i)'\phi(X_j)$, of which the result can be put in the symmetric Gram matrix K as an entry. The 'kernel trick' refers to being able to use the higher dimension for linear distinction while only having to use a kernel function rather than having to know the exact representation of the *feature mapping* $\phi(.)$. This makes it computationally rewarding.

4.2.2 Kernel trick applied for nonlinear dimension reduction

Schölkopf et al. (1997) were the first to introduce kPCA as a nonlinear dimension reduction technique. As pointed out by Kutateladze (2022), kPCA's first step involves forming a linear

factor model for the higher dimension feature space, or

$$\underset{(T\times M)}{\phi(X)} = \underset{(T\times R)}{F_\phi} \underset{(R\times M)}{\Lambda'_\phi} + \underset{(T\times M)}{e_\phi},$$

in which F_ϕ is the factor matrix of $\phi(X)$, in which Λ_ϕ the associated factor loading matrix, and e_ϕ the corresponding error. Applying PCA on the transformed space requires the covariance matrix of $\phi(X)$, which requires the explicit formula of $\phi(X)$, 'which is generally unknown for interesting problems' (Kutateladze, 2022). In other cases it is computationally infeasible as the resulting covariance matrix is $(M \times M)$ dimensional. However, by noticing that we can write the eigenvector as a linear function of features, we can apply PCA as

$$\underset{(T\times 1)}{\hat{F}_\phi^{(r)}} = \underset{(T\times 1)}{F^{(r)}}(\phi(X)) = \phi(X) \underset{(M\times 1)}{V^{(r)}} = \phi(X)\phi(X)' \underset{(T\times 1)}{A^{(r)}} = \underset{(T\times T)}{K} A^{(r)},$$

in which $\hat{F}_\phi^{(r)}$ is the r'th kPC, for which $r \in \{1, 2, \ldots, R\}$, in which $F^{(r)}(\phi(X))$ denotes the r'th linear PC with respect to the feature matrix, in which $V^{(r)}$ denotes the r'th eigenvector of the feature matrix, and in which $A^{(r)}$ is the r'th eigenvector of the Gram matrix K (Kutateladze, 2022). As one can see, the Gram matrix is used to calculate the nonlinear factors without needing the exact formula for $\phi(.)$. Kutateladze (2022) showed that the resulting factors are consistent when $M < \infty$. Similarly he showed that when $M \to \infty$ that under several assumptions the factors are consistent too, which is the case for the the Sigmoid and RBF kernels. For these specific kernels, Kutateladze (2022) showed that their kPCs nest the linear PCs for low hyperparameter values. This is particularly useful as linear PCs perform particularly well in short horizon predictions (Stock & Watson, 2002a), which can be useful when nowcasting GDP close to the end of the quarter.

4.2.3 Sigmoid and RBF kernels

Within this study, both the Sigmoid and RBF are used to extract the kPCs. This idea stems from the results form Kutateladze (2022), in which he shows that these kernels can perform significantly better than linear factors when making macroeconomic predictions. Furthermore, these kernel functions can extract features similar to those extracted by DL methods (Schölkopf et al., 1997), which perform particularly well in high-dimensional prediction contexts (Nti et al., 2020).

The Sigmoid kernel is defined as $k(X_i, X_j) = \tanh(c_0 + \gamma X_i' X_j)$, in which the hyperparameter $\gamma > 0$ tunes the severity of reaction of the function to differences between X_i and X_j, $i, j \in \{1, 2, \ldots, T\}$. A higher value of γ essentially creates more nonlinearity, while a lower γ does the opposite. The parameter c_0 is a constant, and is for general purposes set to zero, as for our study.

The RBF kernel is defined as $k(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|^2)$, in which a higher value of $\gamma > 0$ makes the normal distribution associated with this kernel less spread, in this case also adding more nonlinearity.

For both kernels, the r'th kPC converges towards the r'th linear PC as the hyperparameter approaches zero, which can be mathematically written as $\lim_{\gamma \to 0} F^{(r)}(\phi(X)) = F^{(r)}(X)$, in

which both X and $\phi(X)$ are standardised.

The optimal hyperparameters for both models are selected based on a grid search by means of cross-validation over the last two comparable months. This choice firstly stems from the motivation of 'fair validation', in which only validations based on months in previous quarters are fair. Secondly, it takes into account similar data availability. For these previous months, data is imputed such that similar conditions as the current month are obtained.

The grids used are equivalent to those by Kutateladze (2022):

- For the Sigmoid kernel: $[10^{-6}, 10^{-5.5}, 10^{-5}, 10^{-4.5}, 10^{-4}, 10^{-3.5}, 10^{-3}, 10^{-2.5}, 10^{-2}]$.
- For the RBF kernel: $[10^{-6}, 10^{-5.5}, 10^{-5}, 10^{-4.5}, 10^{-4}, 10^{-3.5}, 10^{-3}]$.

4.3 Mixed-frequency estimation methods

The second step of the model-building procedure is the mixed-frequency estimation. PCs are obtained based on monthly data and form monthly factors. At the same time, the GDP is on quarterly basis, which requires us to model this mixed-frequency setting accordingly. Although several approaches have been used in the literature, two main variants are by means of factor-Bridge and factor-MIDAS, described in upcoming sections. The final goal of both is to obtain quarterly factors and quarterly factor lags, which can be used in a standard OLS regression.

Additional regressors can include an intercept and GDP lags, of which the latter can be useful to exploit the leftover autocorrelation when log-differenced GDP, which approximates GDP growth rate, is not fully stationary. This brings us to the final regression:

$$y_t = \alpha + \sum_{p=1}^{P} \gamma_p y_{t-p} + \sum_{k=1}^{K} \sum_{h=0}^{H} \beta_{kh} F_{k,t-h} + \epsilon_t. \tag{1}$$

In the equation, y_t is the growth rate of GDP at time t, which is the dependent variable. Furthermore, the constant in the model is denoted by α. Then, y_{t-p} is the p'th lag of GDP growth and γ_p is its coefficient, for which $p \in \{1, 2, \ldots, P\}$, and in which P is the number of lags of GDP growth. The k'th factor, lag h, is denoted by $F_{k,t-h}$ and β_{kh} is its coefficient, in which $k \in \{1, 2, \ldots, K\}$, $h \in \{0, 1, \ldots, H\}$, and in which K the number of factors and H the number of factor lags. Finally, ϵ_t is the leftover error of the model at time t. Since in the model all components are on quarterly basis, standard OLS can be used to estimate the coefficients, which serves as our standard regression method. Alternatively, the Ridge regression can be used, for which its motivation will be discussed in Section 4.5.

The optimal parameters are chosen by minimising the *Bayesian Information Criteria (BIC)* for $K \in \{1, 2, \ldots, 6\}$, $H \in \{1, 2, \ldots, 6\}$, and for $P \in \{0, 1\}$, which is apart from the choice of P similar to Kutateladze (2022), who chooses $P \in \{1, 2, \ldots, 6\}$ instead. The BIC is known for its severe penalty on additional regressors, which serves our purposes as sparse models in literature are known for their predictive ability.

4.3.1 Factor-Bridge

Bridge Equations, originally introduced by Baffigi et al. (2004), involve aggregating *high-frequency variables* to obtain *low-frequency* regressors. These are then applied in a linear regression, as

used by Giannone et al. (2008) and Ferrara and Simoni (2023).

Although more variants exist, one typical aggregation strategy involves using simple average to make the time transition, which for a regressor k means

$$x_{k,q} = \frac{1}{M} \sum_{m=1}^{M} x_{k,m},$$

in which $x_{k,q}$ is the regressor at higher frequency and $x_{k,m}$ at lower frequency.

In our case, M is the number of months per quarter, meaning $M = 3$. Furthermore, $m \in \{1, 2, 3\}$ is the month in the quarter $q \in \{1, 2, 3, 4\}$. An example relevant for our studies would be that the value of a regressor for the first quarter is constructed using the average of the first three months in the year.

Factor-Bridge replaces these averaged regressors by averaged factors. Additional regressors often include an intercept and possibly lags of GDP growth, making it a *Bridge-AR* model. For this study, only up to a single lag of GDP has been chosen, which leads to limited efficiency loss (Higgins, 2014).

4.3.2 Factor-MIDAS

The MIDAS was originally introduced by Ghysels, Santa-Clara and Valkanov (2005), and evolved from the need for a model that does not have to aggregate data like Bridge equations, which causes loss of information. Rather, it optimises the way in which the data is transformed to the lower frequency per regressor. MIDAS has many forms, but our focus is on the Exponential Almon and Beta versions. The Exponential Almon Exponentially gives less weight to more distant lags, which can be useful when recency matters, making it particularly useful for nowcasting. The Beta polynomial is known for being robust to the data that is used, allowing for many different shapes for different data (Ghysels et al., 2007). We hypothesise that this can be useful when data contains nonlinearity and takes on extreme forms. The latter is the case in kPCs when using high hyperparameter values. Also, the Beta polynomial is known to be good for smaller models, which are often good for forecasting matters.

The MIDAS transformation from higher to lower frequency allows for more flexibility compared to simple average and can do this by means of a polynomial. Mathematically, for a regressor k, this can be written as

$$x_{k,q} = \sum_{m=1}^{M} c(m, \theta_k) L^{m-1} x_{k,m_3},$$

in which $\sum_{m=1}^{M} c(m, \theta_k) L^{m-1}$ is the *weight function* or *polynomial* of the regressor k, and in which $c(m, \theta_k)$ is the *transformation coefficient* for lag m. Furthermore, L is the *monthly lag operator* which transforms a regressor into its value m months ago. Finally, x_{k,m_3} is the regressor during the third month of the quarter. Similarly as for Bridge: $m \in \{1, 2, 3\}$, $q \in \{1, 2, 3, 4\}$, and $M = 3$.

By allowing different polynomial parameters per regressor, we allow for different optimal transformations per regressor, which can be effective when regressors capture different patterns,

in our case to model the different patterns each principal component captures.

Marcellino and Schumacher (2010) describe *Factor-basic MIDAS*, which uses the *Exponential Almon lag polynomial*, and defines the transformation coefficients as

$$c(m, \theta_k) = \frac{\exp(\theta_{k1}(m-1) + \theta_{k2}(m-1)^2)}{\sum_{m=1}^{M} \exp(\theta_{k1}(m-1) + \theta_{k2}(m-1)^2)},$$

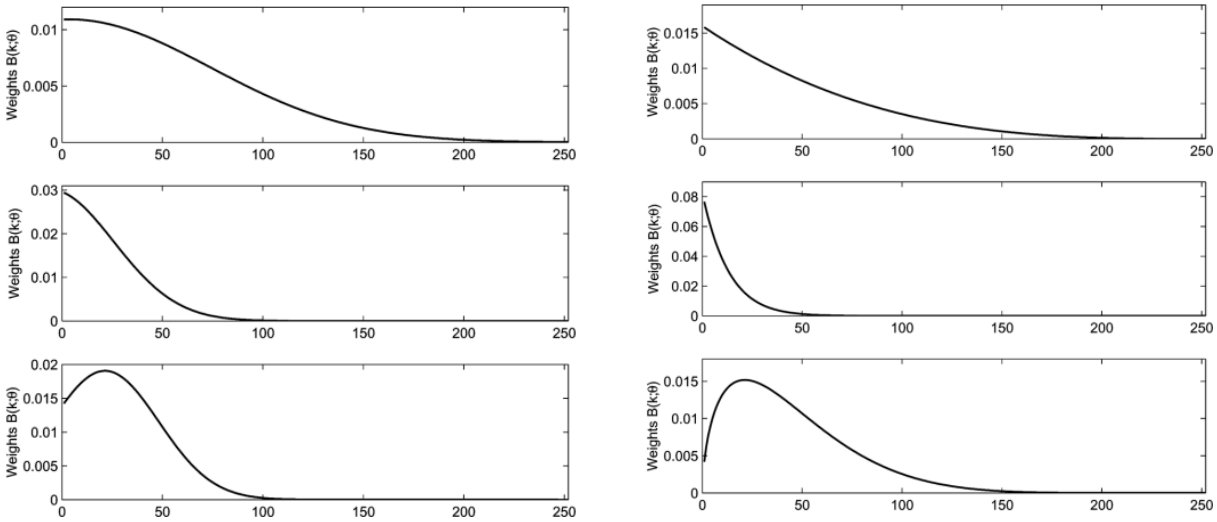
in which we have added the subscript k to emphasise the regressor dependency of the optimal theta parameters. The starting values of θ_k are set as $(\theta_{k,1}, \theta_{k,2})' = (-1, 0)' \quad \forall k \in \{1, 2, \ldots, K\}$. This is commonly used for Exponential Almon, as from that initialisation many different optima can be reached during the *nonlinear least squares (NLS)* optimisation process (Kvedaras & Zemlys-Balevicius, 2013).

The *Beta lag polynomial* defines the transformation coefficients as

$$c(m, \theta_k) = \frac{((m-1)/M)^{\theta_{k1}-1}(1-(m-1)/M)^{\theta_{k2}-1}}{\sum_{m=0}^{M}((m-1)/P)^{\theta_{k1}-1}(1-(m-1)/M)^{\theta_{k2}-1}}.$$

The starting weights $(\theta_{k,1}, \theta_{k,2})'$ are set to $(1, 5)'$. These values give access to many different transformations. An alternative is the use of the starting values $(2, 3)'$. This, however, led within our studies to negligible differences in optima, indicating that our results are robust to the starting weights. One possible shape the Beta polynomial can obtain is an equal weight distribution over the different months, which is similar to Bridge and therefore makes Bridge nest MIDAS-b.

Note: The different shapes arise from the different choices for θ, which have been chosen in the figures to show a variety of different shapes, and of which we decided to exclude the absolute numbers as the figures are only for illustrative purposes. The horizontal axis in this context denotes the number of daily lags of up to 252, which means that when shifting to up to three monthly lags, the shapes could react differently to the choices of the parameters. The vertical axis shows different weights, of whom the exacts are as described by Ghysels et al. (2007). One not shown case is the equal weight distribution that the Beta polynomial can obtain.



(a) Exponential Almon (Ghysels et al., 2007). (b) Beta (Ghysels et al., 2007).

Figure 1: Different possible shapes of the Exponential Almon and Beta polynomials

For both polynomials, NLS is used to find the optimal theta values after initialisation. NLS

repeatedly minimises the *sum of squared residuals (SSR)*, given the current theta estimates, using a standard OLS regression, and uses the *Jacobian matrix* to find theta values that are likely directions in which the SSR can be minimised further. It does this until convergence. The resulting transformation from monthly to quarterly frequency is optimal for the training sample and can have many different shapes, as illustrated by Figure 1.

Within our studies, when using the MIDAS regression together with kPCA, convergence issues occur in rare cases when high hyperparameter values are used for the kernel function, as this leads to extreme values within the kernel factors. One can solve this problem by using a small $L2$ penalty term on the values of θ_k, $\forall k \in \{1, 2, ..., K\}$, which avoids unlimited increment in the values of θ.

Finally, the transition from MIDAS to Factor-MIDAS is made by replacing the regressors by factors and applying the polynomial scheme to the high-frequency factors instead. Then, when adding lags of GDP, the MIDAS is typically referred to as *MIDAS-AR*. Our study only uses up to one lag of GDP, as Marcellino and Schumacher (2010) show that zero or one lag empirically leads to the highest nowcasting accuracy.

The program this study uses for MIDAS was an extended version we built based on well-known Python libraries, which we extended by allowing up to K high-frequency regressors, as of which the details are in Appendix C.

4.4 Resulting models

The resulting models consists of two sub-parts: firstly, the set of *optional factor extraction methods* E, which we define as $E = \{\text{PCA, SPC, Sigmoid, RBF}\}$, and secondly the set of *optional regression methods* R, which we define as $R = \{\text{Bridge, MIDAS-b, MIDAS-e}\}$. In the latter, MIDAS-b abbreviates Factor-MIDAS Beta, and MIDAS-e abbreviates Factor-MIDAS Exponential Almon. These together result in the set of *model combinations* M, which we define as $M = E \times R$, in which $\times$ denotes the Cartesian product of the two. In total, this leads to $|M| = 12$ different models, of which half contain nonlinear factors. As illustration to our notation, the combination between PCA and Bridge we would refer to as model (PCA, Bridge).

4.5 Research Setup and Estimation Procedure

Obtaining results requires several choices to be made. This includes the implementation of a step-by-step estimation procedure, addressing overfitting, and selecting the software and programs, which are briefly discussed here.

Firstly, the training data is from January 1980 until December 2010, and updates using an expanding window. Then, monthly out-of-sample forecasts are made for January 2010 until March 2024, which is approximately a train-to-test split of 65/35. This choice is made based on the choice of Kutateladze (2022), who uses 10-year monthly training data to estimate the kPCs, involving a rolling window. This translates to thirty years of monthly data when aggregating monthly to quarterly data for mixed-frequency regressions. Also, it is based on Marcellino and Schumacher (2010) who uses an expanding window as this uses latest information but also perseveres long-term structural movements of GDP. In particular, the financial crisis of 2008-2009 is included in the training data which might train the kPCs to adequately deal with the

COVID-19 financial crisis.

The forecast construction in each iteration is as will be explained now. Firstly, the vintage data of the month is accessed, transformed, and imputed, including for all unknown months in the current quarter. The data is standardised, after which the PCs and their lags are extracted. The optimal model combination is chosen for Equation 1, for factor-Bridge, for factor-MIDAS-b, and for factor-MIDAS-e. Finally, using the optimal models, forecasts are made, who is compared to the last GDP vintage data.

After having obtained predictions over the full sample, the MSPE is calculated by averaging the sum of the squared errors. Also the Diebold-Mariano (DM) Test is calculated, which tests whether H_0 of equal nowcasting accuracy between models can be significantly rejected.

Within our studies, overfitting is one thing to look out for. Firstly, kPCA uses hyperparameters that need to be tuned such that the factors do not overfit the training data, which we solved by using 'fair validation' as explained earlier. Secondly, MIDAS regressions are prone to overfitting as their flexible data transformations can overfit the training data, while not being optimal for the out-of-sample, which we examined the severity of by means of a Ridge regression. Using the Ridge regression as an alternative to OLS, which in our case meant an additional penalty on the factors $\lambda \sum_{k=1}^{K} \sum_{h=0}^{H} \beta_{kh}^2$, for which $\lambda > 0$, and in which β_{kh} is the coefficient associated with regressor k, lag h, also counteracts multi-collinearity, and consequently reduces the total variance of predictions, at the cost of a small bias. Especially within large factor models that contain many factors and lags, this can be useful, as illustrated in the results section.

Lastly, the program technicalities which were used to obtain the results can be found in Appendix C and are mostly relevant for reproduction purposes.

5 Results

First the main results will be discussed, followed by an in-depth analysis of kPCA's behaviour in nowcasting setting and several robustness analyses.

5.1 The main results

After having implemented the estimation procedure, the main results in Table 1 were obtained.

The results show that over the whole sample (PCA, Bridge) has the lowest MSPE, closely followed by (Sigmoid, Bridge) and (RBF, Bridge), for which the DM-test's null hypothesis of equal nowcasting accuracy is not rejected for up to 60%, and 25% significance levels, respectively. The DM-test's null on 5% significance is rejected for five out of nine other model combinations, including (Sigmoid, MIDAS-e) and (RBF, MIDAS-e). The Bridge regression, when combined with PCA, Sigmoid kPCA, or RBF kPCA leads to a higher nowcasting accuracy when combined with MIDAS-b or MIDAS-e, which could originate from MIDAS overfitting the data in these combinations. On the contrary, combinations involving SPC benefit from using MIDAS as opposed to Bridge. Within MIDAS-b and MIDAS-e, the kPCA-based models lead to a higher nowcasting accuracy than the linear PCA- and SPC-based models, which in the case of kPCA RBF led to rejection of the DM-test's null on 5% significance. These results signal that, within these combinations, the additional nonlinearity is in fact useful.

Table 1: The nowcasting accuracy for the different model combinations: over the full sample, over each month in the quarter individually, during the COVID-19 financial crisis, and over the full-sample excluding COVID

	Bridge				***MIDAS-b***				***MIDAS-e***			
	PCA	***SPC***	***Sigmoid***	**RBF**	***PCA***	***SPC***	***Sigmoid***	**RBF**	***PCA***	***SPC***	***Sigmoid***	**RBF**
Total MSPE	0.113*	1.705	0.130	0.144	0.331	0.365	0.263	0.190	0.375	0.502	0.371	0.300
PCA DM-test	N.A.	0.249	0.621	0.289	N.A.	0.561	0.275	**0.049**	N.A.	0.389	0.832	**0.003**
Best Model DM-test	N.A.	0.249	0.621	0.289	**0.029**	**0.024**	0.102	0.313	**0.019**	0.064	**0.024**	**0.043**
First Month MSPE	0.098*	0.137	0.131	0.130	0.369	0.350	0.371	0.289	0.362	0.342	0.389	0.306
PCA DM-test	N.A.	0.246	0.589	0.249	N.A.	0.070	0.942	0.224	N.A.	**0.010**	0.562	0.130
Best Model DM-test	N.A.	0.246	0.589	0.249	0.186	0.206	0.183	0.269	0.182	0.217	0.188	0.269
Second Month MSPE	0.201	4.425	0.208	0.224	0.384	0.372	0.178	0.142*	0.375	0.362	0.362	0.300
PCA DM-test	N.A.	0.308	0.925	0.771	N.A.	0.746	0.087	**0.039**	N.A.	0.173	0.255	0.052
Best Model DM-test	0.313	0.285	0.240	0.451	**0.039**	**0.045**	0.624	N.A.	**0.032**	0.096	**0.044**	0.089
Third Month MSPE	0.040*	0.590	0.049	0.075	0.239	0.371	0.237	0.137	0.388	0.807	0.362	0.291
PCA DM-test	N.A.	0.284	0.113	0.063	N.A.	0.438	0.983	0.337	N.A.	0.349	0.134	0.094
Best Model DM-test	N.A.	0.284	0.113	0.063	0.202	0.257	0.149	0.101	0.148	0.195	0.164	0.171
COVID MSPE	0.078	1.694	0.932	0.100	0.031*	0.351	0.240	0.161	0.348	0.492	0.359	0.282
PCA DM-test	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
Best Model DM-test	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.
COVID Excluded MSPE	0.813	1.699	0.875	1.011	0.743	0.630	0.717	0.791	0.922	0.672	0.621*	0.647
PCA DM-test	N.A.	**1.9e-05**	0.460	0.080	N.A.	**0.047**	0.761	0.516	N.A.	**7.4e-04**	**2.9e-05**	**3.7e-04**
Best Model DM-test	0.102	**1.8e-06**	0.053	**5.6e-04**	0.108	0.773	0.440	0.139	**3.7e-04**	0.689	N.A.	0.577

Note: The nowcasting accuracy is measured in MSPE, of which each entry in the table is the model's relative MSPE to the AR(1) model. The model with the lowest MSPE in each category is denoted by an asterisk. For each model combination, the p-value of the Diebold Mariano (DM) test when comparing it to the benchmark PCA is denoted in the second rows of the categories, and to the lowest MSPE-model in the third rows. Bold in these rows indicates significant rejection of the null hypothesis under 5% significance. The COVID-19 financial crisis ranges from 2020-2021. N.A. denotes Not Applicable.

The results show similar patterns when looking at the first and third months in quarters. The best-performing model combination is (PCA, Bridge), closely followed by (Sigmoid, Bridge) and (RBF, Bridge). In this case, (PCA, Bridge), when compared to other models, never leads to rejection of the DM-test's null on 5% significance.

Different conclusions hold for the second months in the quarters, as (RBF, MIDAS-b) has the lowest MSPE, followed by (Sigmoid, MIDAS-b). The former outperforms all linear benchmarks by at least 30%. The DM-test, when comparing the models to (RBF, MIDAS-b), in four out of eleven cases rejects the null of equal nowcasting accuracy on 5% significance. Similar to previous results, PCA performs best within Bridge, while kPCA Sigmoid and RBF within MIDAS-b and MIDAS-e.

Comparing the MSPE per month: the prediction made in the last months of the quarter does lead to the lowest MSPE compared to the other two months in most scenarios, which is similar to the intuition of more information leading to better nowcasts. Surprisingly, the first-month MSPE follows in second place and closely follows the MSPE of the third month. The MSPE of the second month seems to stay behind in MSPE compared to the other months, suggesting that the additional information from the first month leads to reduced nowcasts.

During the COVID-19 financial crisis, in 2020–2021, (PCA, MIDAS-b) has the lowest MSPE, followed by (PCA, Bridge) and (PCA, RBF). Whether it has significantly better nowcasting accuracy is, however, unknown. This is because the number of observations is only twelve, which is not enough to fairly calculate the DM-test as its power is too low, which can lead to a spurious finding.

When looking at the full sample and excluding the COVID-19 financial crisis, the (Sigmoid, MIDAS-e) combination achieves the lowest MSPE. Via the DM-test, for three out of eleven models the null is rejected at 5%, and for seven out of eleven models at 15% significance, including (PCA, Bridge). These results suggest that the nonlinearity might be more useful in calmer periods when making nowcasts, which is contrary to our earlier hypothesis and to the findings of Goulet Coulombe et al. (2022).

Looking at all the different scenarios discussed, kPCs can lead to a lower MSPE within second months in quarters, over the full sample with the COVID-19 financial crisis excluded, and within MIDAS regressions. Furthermore, Bridge models in most scenarios lead to a lower MSPE compared to MIDAS, which might signal that there is overfitting present within MIDAS. Finally, since predictions are not necessarily better in later months of quarters, we can conclude that the EM algorithm imputations have limited effect on the nowcasting accuracy.

Based on Figure 2, which shows the actual growth of GDP versus its best-performing nowcasts, we additionally conclude that Bridge-based models generally overestimate shocks, while the opposite can be said for MIDAS-based models. This is most clear during the COVID-19 financial crisis, but generally also holds when closely examining sub-periods in the graph. A forecast combination between linear Bridge and nonlinear MIDAS might therefore be useful to get the best of both worlds and could be examined in future research.

For the remainder of the analysis, SPC will be excluded as it led in no scenario to the most optimal model, which leaves PCA as the only linear reference.

Note: Optimal nowcasts refer to nowcasts made by the models that achieve the lowest MSPE in Table 1. The GDP has a step shape, as during each nowcast made within the quarter, the actual GDP was imputed for comparison. Finally, the GDP growth rate was estimated using log-differenced GDP.

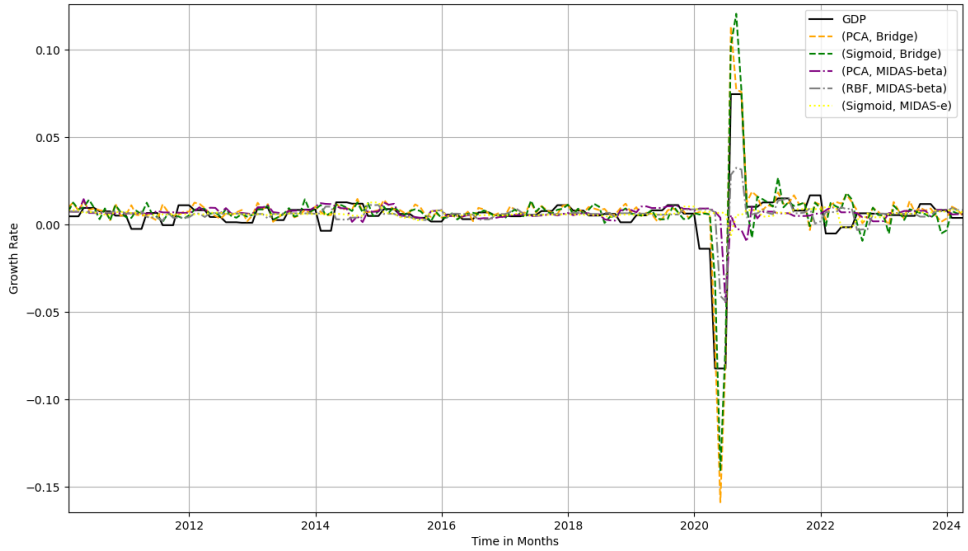


Figure 2: GDP growth rate and its optimal nowcasts

5.2 Behaviour of kPCA in nowcasting

An interesting statistic is the optimal amount of nonlinearity for the kPCs within the nowcasting setting. Table 2 shows the different hyperparameters of the grid and their number of picks over the full sample and over the COVID-19 financial crisis, respectively.

Interestingly, for Sigmoid, across the full sample and all different regression methods, the highest hyperparameter value is most often selected, indicating that adding a high dose of nonlinearity is optimal for the validation sample. The second most often selected hyperparameter is the lowest value of the grid, which approximates PCA-factors. It is also interesting to see that the optimisation process has a tendency to choose the maxima and minima, while it rarely chooses $\gamma = 10e{-}05$ for example. For the COVID-19 financial crisis, all three different regression models react very differently: while for Bridge there is a tendency to choose nonlinearity in the middle of the grid, for MIDAS-b there is a preference for high non-nonlinearity, and for MIDAS-e low nonlinearity.

For the RBF kernel, over the full sample, results differ across the Bridge and MIDAS regressions. While for Bridge there is a preference for a combination between low nonlinearity ($\gamma = 10e{-}06$), and high nonlinearity ($\gamma = 10e{-}03, \gamma = 10e{-}03.5$), for the MIDAS methods there is clear preference for high nonlinearity. This might be due to the ability of MIDAS to more effectively retain the nonlinearity when moving down the frequency, as we hypothesised earlier. Within the COVID-19 financial crisis, the results seem quite scattered. The only thing standing out is the preference for keeping MIDAS-e's nonlinearity in the middle.

Another interesting insight is the resulting models that lead to the different nowcasts, which Table 3 shows for both the full out-of-sample and COVID-19 financial crisis. One can first conclude that the models are by no means sparse but contain many regressors. Secondly, the same models are often chosen, as the number of occurrences is high. Finally, the models chosen in the COVID-19 financial crisis are generally sparser when compared to the full sample.

Table 2: The amount of nonlinearity added to the kPCA-factors

	Sigmoid			**RBF**		
	Bridge	***MIDAS-b***	***MIDAS-e***	**Bridge**	***MIDAS-b***	***MIDAS-e***
γ = *10e-06*	41, 2	30, 0	26, 7	68, 4	21, 0	30, 1
γ = *10e-05.5*	0, 0	16, 0	12, 0	1, 0	5, 0	11, 0
γ = *10e-05*	0, 0	9, 0	6, 0	2, 1	11, 0	6, 0
γ = *10e-04.5*	10, 3	7, 1	9, 0	11, 0	7, 0	23, 11
γ = *10e-04*	10, 5	17, 1	9, 3	15, 4	23, 2	4, 0
γ = *10e-03.5*	11, 1	9, 0	8, 0	33, 3	32, 8	37, 0
γ = *10e-03*	14, 1	11, 0	20, 2	41, 0	72, 2	60, 0
γ = *10e-02.5*	21, 0	9, 1	23, 0	-	-	-
γ = *10e-02*	64, 0	63, 9	58, 0	-	-	-
Total	171, 12	171, 12	171, 12	171, 12	171, 12	171, 12

Note: The first and second numbers represent the number of occurrences that each hyperparameter value is being selected for the full out-of-sample and for the COVID-19 financial crisis, containing 171, and 12 observations respectively. Important to recognise is that the smaller the value of γ, the more kPCA converges towards PCA, using less linearity for the factors, and vice versa. The empty values in the table arise from RBF having a less broad grid compared to Sigmoid.

Table 3: Optimal model compositions

	MO	**O**	**SMO**	**O**	**TMO**	**O**	**CMO**	**O**
(PCA, Bridge)	(6,3,0)	94/171	(3,5,0)	39/171	(6,2,0)	21/171	(6,3,0)	12/12
(PCA, MIDAS-b)	(6,3,1)	79/171	(3,4,0)	20/171	(5,3,0)	17/171	(3,4,0)	4/12
(PCA, MIDAS-e)	(6,4,1)	57/171	(2,6,0)	23/171	(4,6,1)	13/171	(2,5,0)	8/12
(Sigmoid, Bridge)	(6,3,0)	73/171	(3,5,0)	59/171	(6,2,0)	10/171	(6,3,1)	6/12
(Sigmoid, MIDAS-b)	(6,3,1)	42/171	(3,5,0)	33/171	(6,4,1)	20/171	(5,2,1)	6/12
(Sigmoid, MIDAS-e)	(6,3,0)	55/171	(3,5,0)	30/171	(6,3,1)	10/171	(5,2,1)	8/12
(RBF, Bridge)	(6,3,0)	61/171	(3,5,0)	33/171	(4,4,0)	17/171	(4,4,0)	7/12
(RBF, MIDAS-b)	(6,4,0)	44/171	(6,3,0)	24/171	(6,3,1)	22/171	(6,2,0)	10/12
(RBF, MIDAS-e)	(6,3,0)	55/171	(6,4,1)	24/171	(6,4,0)	18/171	(6,4,1)	12/12

Note: The models are chosen as optimal under the BIC. Each combination between brackets is firstly the number of PC factors, secondly their number of lags, and finally whether a lag of GDP is adopted. The occurrences of each combination are counted over the whole sample, and for the COVID-19 financial crisis, which ranges from 2020-2021. These periods contain 171 and 12 monthly observations, and 57 and 4 quarterly observations, respectively. MO denotes most occurring, O occurrences, SMO second most occurring, TMO third most occurring, and CMO COVID-19 most occurring.

5.3 Robustness checks

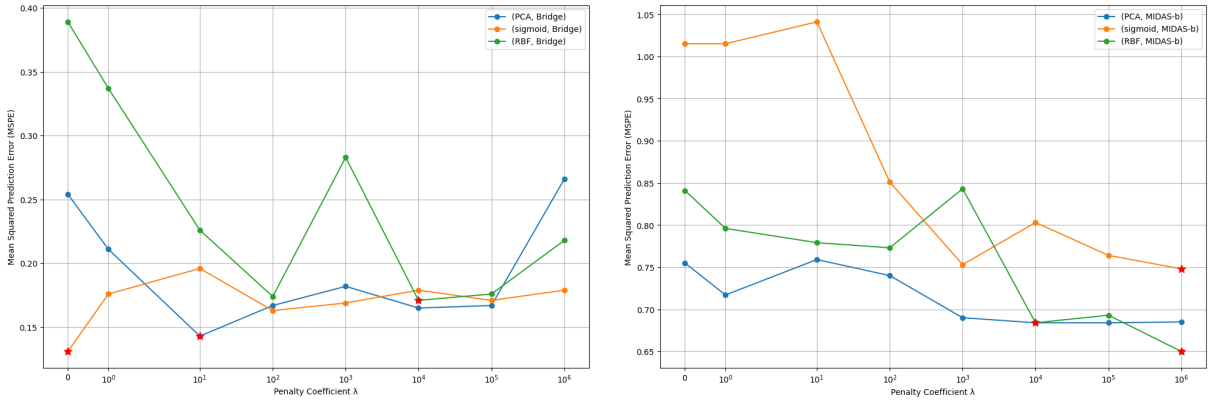
The previous result showed that the optimal model compositions contain many regressors. For OLS, this could lead to overfitting, as sparse models generally perform better in out-of-sample forecasting. At the same time, the BIC contains the highest penalty for the number of regressors of any criteria. This suggests that many factors contain predictive power within this context. Penalising the factor coefficients via a Ridge regression can be useful to decrease the nowcasting variance, and counteract multi-collinearity within the factors, at the cost of a small bias. Especially for MIDAS, in which overfitted data transformations can make it sub-optimal compared to Bridge, this could counteract overfitting.

Implementing the Ridge regression led to Figures 3, 4 and 5, of whom Figure 4 and Figure 5

in Appendix A. For a calm period and dynamic period, which were selected based on the GDP growth's amplitude, in Figure 2, the dynamics in MSPE were investigated. For the calm period the period 2015-2018 was selected, and for the dynamic period the COVID-19 financial crisis from 2020-2021. As these periods were ex-post selected, for different periods the analysis could be repeated, to examine whether comparable results hold.

Looking at bridge during the calm period, Figure 3a shows that for the OLS benchmark, the MSPE of (Sigmoid, Bridge) is close to 50% lower compared to (PCA, Bridge), suggesting that the nonlinear (Sigmoid, Bridge) does particularly well over this period compared to linear (PCA, Bridge). This suggests that nonlinearity can be useful over certain periods, while the opposite is true for others. Also, the figure shows that the additional Ridge penalty does in many cases lead to reduced MSPE, which is similar for the financial crisis in Figure 5a.

Note: The calm period is 2015-2018. The vertical axes are the relative MSPE compared to the AR(1) predictions. The optimum penalty for each method is indicated with a red star.



(a) Bridge (b) MIDAS-b

Figure 3: MSPE for Bridge and MIDAS-b with different Ridge penalties during the calm period

Looking at MIDAS-b during the calm period, as in Figure 3b, one can conclude that a significant penalty term is optimal for all factors, and that it can lead to a decrease in MSPE from 10% to 25% depending on the exact factor-extraction method. Similar results hold when looking at MIDAS-e in Figure 4a. During the COVID-19 financial crisis, for MIDAS-b and MIDAS-e, an additional penalty achieves the exact opposite result, as in Figure 5b, and 5c.

A final robustness analysis, evaluating the effect of revisions on the main results, led to Table 4. For this analysis, the last vintage dataset was cut off, and past values were removed according to Table 5 in Appendix B using a random generator to match the actual vintage data.

Table 4: Robustness analysis comparing the effect of vintage data on nowcasting accuracy

	Bridge	**MIDAS-b**	**MIDAS-e**
PCA	0.113, 0.181	0.331, 0.332	0.375, 0.373
Sigmoid	0.130, 0.188	0.263, 0.289	0.371, 0.459
RBF	0.144, 0.397	0.190, 0.299	0.300, 0.327

Note: The MSPE is relative to a AR(1) vintage-based nowcasts. The left number represents the results when using intermediate vintage data, while the right one represents the last vintage data, which includes all revisions.

Surprisingly, the results show that last monthly vintage data, which contains more revisions

of data, leads to decreased nowcasting accuracy for most model combinations, of which the result is most significant for kPCA. One reason could be that the earlier vintage data have additional variables that are not included in the last vintage dataset. Another could be that the random generator chooses more elements to zero than the actual data on average, leading to more imputations by the EM algorithm. Nonetheless, using vintage data for future research is important as the effect is not uniform across different methods.

6 Conclusion

The challenge of making high accuracy nowcasts for quarterly GDP emerged from the lack of timely estimates until at least one month after the quarter's end.

This paper investigated whether nonlinear monthly factors, when combined with the (i) factor-Bridge and (ii) factor-MIDAS regressions, could outperform compared to linear factors as a benchmark. Using quarterly vintage data (Federal Reserve Bank of Philadelphia, 2024) and a wide range of monthly macroeconomic vintage data (McCracken & Ng, 2016), it was shown that applying these methods over the out-of-sample period 2010-2024 improved nowcasting accuracy in several scenarios. This leads us to answer our research question

'Do nonlinear monthly factors enhance nowcasting accuracy of quarterly GDP?'

with a circumstantial yes. When excluding the COVID-19 financial crisis from the full sample, this led to a lower MSPE compared to all linear methods. Furthermore, when applying the kPCA to a relatively calm period, it resulted in at least a 50% lower MSPE compared to linear factor models. Finally, during the second month of the quarter, when predictions were found to be least efficient, the kPCA performed best, beating all linear models by at least 30%. In other scenarios, the linear variants were dominant, with especially the combination between Bridge and PCA providing good overall estimates. One last finding of this research is that it is important to use vintage data, as not doing so non-uniformly affects the MSPE across different methods, making comparisons unrealistic.

For future research, it is interesting to see under which exact conditions kPCA improves nowcasting accuracy compared to PCA, as this research showed that both methods have their moments of shine. Additionally, one could then examine whether forecast combinations between linear Bridge and nonlinear MIDAS can enhance nowcasting accuracy as our studies showed that both combinations have opposite biases. A second recommendation is extending the number of lags of GDP, as large factor models proved to be optimal within our studies, suggesting that more lags can enhance predictions too. Thirdly, one could turn attention to the Kalman filter and investigate whether one could adopt this nonlinearly using another 'kernel trick'. Then, one could adopt MIDAS to use a single polynomial scheme per factor and its lags, as assuming they should be transformed similarly can be useful to tackle overfitting and is likely due to factors capturing one specific aspect of data. Lastly, different MIDAS models could be investigated, as its literature is vast and expanding (Ghysels, Kvedaras & Zemlys-Balevičius, 2020).

References

Aizerman, A. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, *25*, 821–837.

Altissimo, F., Cristadoro, R., Forni, M., Lippi, M. & Veronese, G. (2010). New Eurocoin: tracking economic growth in real time. *The Review of Economics and Statistics*, *92*(4), 1024–1034.

Baffigi, A., Golinelli, R. & Parigi, G. (2004). Bridge models to forecast the Euro area GDP. *International Journal of Forecasting*, *20*, 447–460.

Bai, J., Ghysels, E. & Wright, J. H. (2013). State space models and MIDAS regressions. *Econometric Reviews*, *32*(7), 779–813.

Bańbura, M., Giannone, D. & Reichlin, L. (2010). Nowcasting. *Social Science Research Network*(1275), 1–40.

Bennett, C. Y. K. (2023). Factor-based imputation for missing data. *GitHub*.

Bernanke, B. S., Boivin, J. & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, *120*(1), 387–422.

Browning. (1982). Nowcasting. *Michigan: Academic Press*, 1–256.

Bureau of Economic Analysis. (2023). Reliability of the initial estimates of gross domestic product and gross domestic income. *US Department of Commerce*.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Federal Reserve Bank of Philadelphia. (2024). Monthly vintages (billions of real dollars, seasonally adjusted). *Real-Time Data Set: Full-Time Series History*.

Ferrara, L. & Simoni, A. (2023). When are Google data useful to nowcast GDP? An approach via preselection and shrinkage. *Journal of Business & Economic Statistics*, *41*(4), 1188–1202.

Foroni, C. & Marcellino, M. G. (2013). A survey of econometric methods for mixed-frequency data. *Social Science Research Network*, 1–45.

Ghysels, E., Kvedaras, V. & Zemlys-Balevičius, V. (2020). Mixed data sampling (MIDAS) regression models. *Handbook of Statistics*, *42*, 117–153.

Ghysels, E., Santa-Clara, P. & Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. *University of California: eScholarship*, 1–32.

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

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

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

Goulet Coulombe, P., Leroux, M., Stevanovic, D. & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5),

920–964.

Hamilton, J. D. (2020). Time series analysis. *Princeton University Press*, 1–816.

Higgins, P. C. (2014). GDP now: A model for GDP 'nowcasting'. *Social Science Research Network*, 1–86.

Jazwinski, A. H. (1970). Stochastic processes and filtering theory. *New York: Academic Press*, 1-400.

Jolliffe, I. T. (2002). Principal component analysis for special types of data. *Springer Series in Statistics*(13), 338-272.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kapetanios, G., Papailias, F. et al. (2018). Big data & macroeconomic nowcasting: Methodological review. *Economic Statistics Centre of Excellence*, *1-77*.

Kim, H. H. & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339–354.

Kutateladze, V. (2022). The kernel trick for nonlinear factor modeling. *International Journal of Forecasting*, *38*(1), 165–177.

Kvedaras, V. & Zemlys-Balevicius, V. (2013). midasr: Mixed data sampling regression. *GitHub*.

Marcellino, M. & 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.

McCracken, M. & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589.

McCracken, M. & Ng, S. (2020). Fred-qd: A quarterly database for macroeconomic research. *NBER Working Paper Series*, *1*(26872), 1–52.

Nti, I. K., Adekoya, A. F. & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, *53*(4), 3007–3057.

Parigi, G. & Schlitzer, G. (1995). Quarterly forecasts of the Italian business cycle by means of monthly economic indicators. *Journal of Forecasting*, *14*(2), 117–141.

Sapphire. (2020). midas_pro: Python version of mixed data sampling (MIDAS) regression. *GitHub*.

Schölkopf, B., Smola, A. & Müller, K.-R. (1997). Kernel principal component analysis. *Artificial Neural Networks - ICANN '97*, 583–588.

Schölkopf, B., Smola, A. & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*(5), 1299–1319.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, 1–48.

Stock, J. H. & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *The University of Chicago Press Journals*, *4*, 351–394.

Stock, J. H. & Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*(460), 1167–

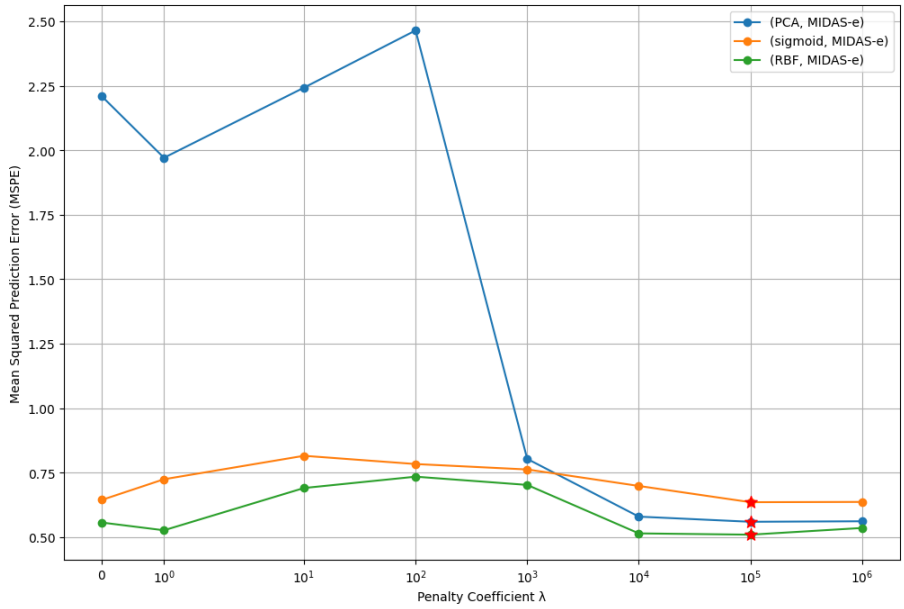
1179.

Stock, J. H. & Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indices. *Journal of Business and Economic Statistics*, *20*, 147–162.

Zuskin, Y. (2020). Mixed data sampling (MIDAS) modeling in python. *GitHub*.

Appendix A: Figures

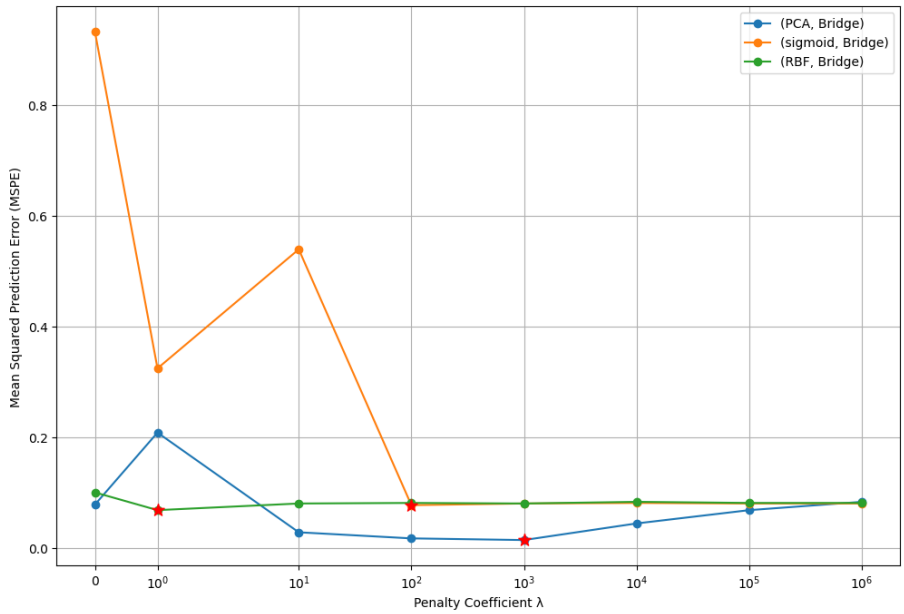
Note: The calm period is 2015-2018. The vertical axes are the relative MSPE compared to the AR(1) predictions. The optimum penalty for each method is signified with a red star.



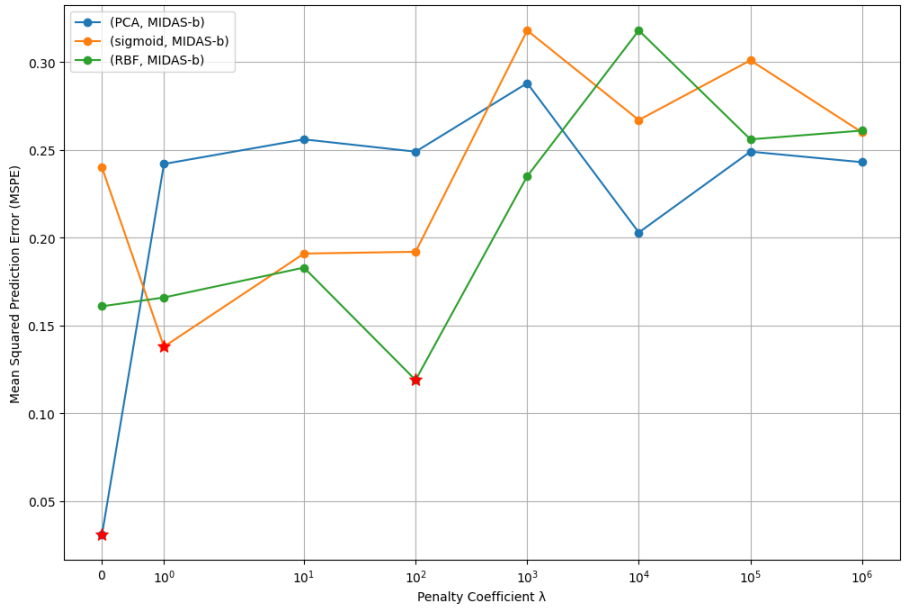
(a) MIDAS-e

Figure 4: MSPE for different Ridge regressions during the calm period

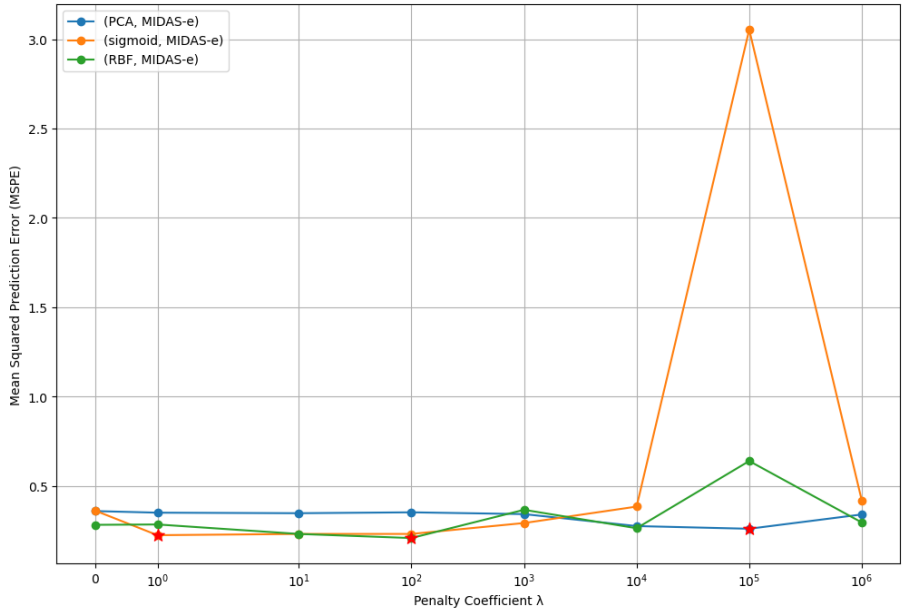
Note: The COVID financial crisis is 2020-2021. The vertical axes are the relative MSPE compared to the AR(1) predictions. The optimum penalty for each method is signified with a red star.



(a) Bridge



(b) MIDAS-b



(c) MIDAS-e

Figure 5: MSPE for different Ridge regressions during the COVID financial crisis

Appendix B: Macroeconomic variables

Table 5: Contents and publication lags of FRED-MD

Group	Variable Name	Notes
Immediately Available		
CPI	*All Items*	
	All items less food	
	All items less medical care	
	All items less shelter	Seasonally adjusted since April 2017
	Apparel	
	Commodities	
	Durables	Seasonally adjusted since April 2017
	Medical Care	
	Services	
	Transportation	
Commodities	*Crude Oil, spliced WTI and Cushing*	
	Crude Materials	Replaced since March 2016
	Finished Consumer Goods	Replaced since March 2016
	Finished Goods	Replaced since March 2016
	Intermediate Materials	Replaced since March 2016
	Metals and metal products	
Employment	*Civilian Employment*	
	Civilian Labor Force	
	Civilian Unemployment Rate	
	Civilians Unemployed - Less Than 5 Weeks	
	Civilians Unemployed - 15 Weeks & Over	
	Civilians Unemployed for 15-26 Weeks	
	Civilians Unemployed for 27 Weeks and Over	
	Civilians Unemployed for 5-14 Weeks	
Financial Rates	*Effective Federal Funds Rate*	
	Moody's Aaa Corporate Bond Minus FEDFUNDS	

Table 5: Contents and publication lags of FRED-MD
(continued)

Group	Variable Name	Notes
	Moody's Seasoned Aaa Corporate Bond Yield	
	Moody's Seasoned Baa Corporate Bond Yield	
	S&P's Common Stock Price Index: Composite	
	CBOE S&P 100 Volatility Index (VXO)	Replaced since December 2021
Housing	*Housing Starts: Midwest*	
	Housing Starts: Northeast	
	Housing Starts: South	
	Housing Starts: Total New Privately Owned	
	Housing Starts: West	
	New Private Housing Permits (SAAR)	
	New Private Housing Permits, Midwest (SAAR)	
	New Private Housing Permits, Northeast (SAAR)	
	New Private Housing Permits, South (SAAR)	
	New Private Housing Permits, West (SAAR)	
IP	*Business Equipment*	
	Consumer Goods	
	Durable Consumer Goods	
	Durable Materials	
	Final Products (Market Group)	
	Final Products and Non-industrial Supplies	
	Fuels	
	Index	
	Manufacturing (SIC)	
	Materials	
	Non-durable Consumer Goods	
	Non-durable Materials	
	Residential Utilities	

Table 5: Contents and publication lags of FRED-MD
(continued)

Group	Variable Name	Notes
ISM	*Manufacturing: Employment Index*	Removed since June 2016
	Prices Index	Removed since June 2016
	Production Index	Removed since June 2016
	Inventories Index	Removed since June 2016
	New Orders Index	Removed since June 2016
	PMI Composite Index	Removed since June 2016
	Supplier Deliveries Index	Removed since June 2016
Monetary	*M1 Money Stock*	
	M2 Money Stock	
	Reserves Of Depository Institutions	
	St. Louis Adjusted Monetary Base	Replaced since January 2020
	Total Reserves of Depository Institutions	
FX Rates	*Switzerland / U.S. Foreign Exchange Rate*	
	Trade Weighted U.S. Dollar Index: Major Currencies	Replaced since April 2020
	U.S. / U.K. Foreign Exchange Rate	
	Japan / U.S. Foreign Exchange Rate	
	Canada / U.S. Foreign Exchange Rate	
L&R	*Real Estate Loans at All Commercial Banks*	
	Initial Claims	Seasonal adjustment since Augustus 2015, new calculation since April 2017
Orders	*New Orders for Non-defense Capital Goods*	
	Unfilled Orders for Durable Goods	
Retail	*Retail and Food Services Sales*	
Sentiment	*University of Michigan: Consumer Sentiment Index*	
Employees	*All Employees: Construction*	
	Durable goods	
	Financial Activities	

Table 5: Contents and publication lags of FRED-MD
(continued)

Group	Variable Name	Notes
	Goods-Producing Industries	
	Government	
	Manufacturing	
	Mining and Logging: Mining	
	Non-durable goods (Employees)	
	Retail Trade	
	Service-Providing Industries	
	Total non-farm	
	Trade, Transportation & Utilities	
	Wholesale Trade	
Available After Zero or One Months		
PCE	*Personal Cons. Exp: Durable goods*	
	Non-durable goods (PCE)	
	Services	
	Chain Index	
Inventories	*Total Business Inventories*	
	Total Business: Inventories to Sales Ratio	
Credit	*Total Non-revolving Credit*	
	Non-revolving consumer credit to Personal Income	
Available After One Month		
Orders	*New Orders for Consumer Goods*	
Sales	*Real Manu. and Trade Industries Sales*	
Available After One or Two Months		
RPI	*Real Personal Income*	
	Real personal consumption expenditures	

Table 5: Contents and publication lags of FRED-MD
(continued)

Group	Variable Name	Notes
	Real personal income ex transfer receipts	
S&P	*S&P's Composite Common Stock: Dividend Yield*	
Loans	*Total Consumer Loans Owned by Finance Companies* *Total Consumer Loans and Nonrevolving Loans Owned by Finance Companies*	

Available After Zero to Eight Months

Help Wanted	*Help-Wanted Index for United States*	Calculation changed Sep 2017
	Ratio of Help Wanted/No. Unemployed	Calculation changed Sep 2017
S&P	*S&P's Composite Common Stock: Price-Earnings Ratio*	

Not Available

GDP	*Real Gross Domestic Product (GDP)*	Vintages removed Apr 2024

Within the table several abbreviations are used. In order: CPI abbreviates Consumer Price Index, IP abbreviates Industrial Production, ISM abbreviates Institute for Supply Management, FX Rates abbreviates Foreign Exchange Rates, L&R abbreviates Loans and Reserves, PCE abbreviates Personal Consumption Expenditures, OSI abbreviates Orders Sales Inventories, RPI abbreviates Real Personal Income, and S&P abbreviates Standard and Poor.

Appendix C: Program Technicalities

Python and R were used to build the nowcasting program this study utilises. The program consists of many functions, of whom a few inspired by others, for which we would like to mention a few references. Firstly, it uses the data imputation function by Bennett (2023), secondly it is inspired by the main program of Kutateladze (2022) with heavy modifications to make it viable for mixed-frequency nowcasting, thirdly it uses an expanded version of earlier MIDAS packages, which can be relevant for future research within the Python environment, and will now therefore be shortly clarified.

The MIDAS package this study utilises is expanded based on the works of Zuskin (2020), which created a MIDAS for a single high-frequency regressor, single lag of the dependent variable, and an intercept. Furthermore, it is carried by Sapphire (2020), which made the package suitable for an additional high-frequency regressor. Expanding their logicality in how more regressors are added in the MIDAS, made us fabricate a generalised MIDAS which can include up to K high-frequency factors, an intercept and one lag of the dependent variable. In addition, our version has implemented the restrictions on θ_k, as discussed in Section 4.3.2, making it suitable for cases under which similar numerical instability might occur. In the future it might be interesting to expand the package to include up to P lags of the low-frequency dependent variable as well.

The final program was run in Google Colab, using the standard CPU, which had significant improvement over local computer run-times, and improved on total RAM, which is especially useful in high-dimensional data analysis and machine learning. For replication matters, the total run time to obtain the main results in Table 1, was up to eighteen hours.