Forecasting the Finnish Producer Price Index Using High-Dimensional Data

Joackim Raikamo
University of Helsinki
Faculty of Social Sciences
Economics
Master's Thesis
May 2022

Abstract

Faculty: Faculty of Social Sciences

Degree programme: Master's Programme in Economics

Study track: General

Author: Joackim Raikamo

Title: Forecasting the Finnish Producer Price Index Using High-Dimensional Data

Level: Master's Thesis

Month and year: May 2022

Number of pages: 82+4

Keywords: high-dimensional forecasting, dynamic factor model, penalized regression, regularized regression, producer price index

Supervisor or supervisors: Mika Meitz

Where deposited: Helsinki University Library

Additional information:

Abstract:

Producing timely information regarding the current and future state of the economy is important for the practice of economic policy: the delay between the implementation of policy measures and the emergence of their effects is typically considerable, which creates a need to anticipate developments in macroeconomic variables. The producer price index is one such variable: producer price indices are used to track changes in the general price level of goods produced within an economy from the point-of-view of producers, which makes them prominent indicators of inflationary pressures and business cycle conditions. The principal objective of this thesis is to investigate whether the Finnish Producer Price Index for Manufactured Goods could be reliably forecasted in the short run using large sets of external predictors.

Increasing the number of predictors exposes standard forecasting methods to inaccuracies and makes their application outright infeasible once the number of variables exceeds the number of observations available for the estimation of the forecasting model. Various alternative methods have been proposed to counter this issue. This thesis provides a broad overview of these methods as well as other relevant issues pertaining to the forecasting macroeconomic variables.

Given that no single framework has proven to dominate others in practical applications, a selection of methods has been chosen for the empirical section of this thesis. These methods represent two different approaches to high-dimensional forecasting: dynamic factor models and penalized regressions. The effectiveness of dynamic factor models is based on the assumption that relevant information contained in high-dimensional data can be summarized using only relatively few underlying factors, the estimates of which can, in turn, be used for forecasting. The solution offered by penalized regressions, on the other hand, is based on striking a balance between the bias and variance of the forecasts. Out of the broader class of penalized methods, four different variations will be utilized in this thesis: the Ridge, Lasso, Elastic Net, and Adaptive Lasso.

The empirical performance of the methods will be assessed by conducting a simulated out-of-sample forecasting experiment, in which a series of consecutive forecasts are estimated for the target variable using historical data. These forecasts are, in turn, compared to their realized counterparts. The objective of the experimental arrangement is to produce representative information regarding the empirical accuracy of the respective forecasting models by emulating circumstances faced in real-time forecasting: only information that would have been available at the time is used to produce each forecast. The set of predictors used in the experiment is composed of monthly economic time series collected from a variety of sources.

Based on the forecasting experiment, the benefit of the high-dimensional models in terms of average forecasting accuracy turns out to be only marginal in comparison to a univariate autoregressive benchmark at the one-, two-, and three-month horizons. Moreover, the differences among the respective high-dimensional methods are found to be insignificant. On the other hand, more favorable results are achieved by using relatively timely market-based variables to predict the concurrent rather than strictly future values of the index. In this case, the penalized models perform particularly well. The results indicate that leveraging the advantage in publication lag enjoyed by external predictors for the purpose of contemporaneous prediction, or nowcasting, could represent the most potential for predicting the producer price index.

Tiivistelmä

Tiedekunta: Valtiotieteellinen tiedekunta

Koulutusohjelma: Taloustieteen maisteriohjelma

Opintosuunta: Yleinen

Tekijä: Joackim Raikamo

Työn nimi: Teollisuuden tuottajahintaindeksin ennustaminen suuriulotteisen aineiston avulla

Työn laji: Maisterintutkielma

Kuukausi ja vuosi: Toukokuu 2022

Sivumäärä: 82+4

Avainsanat: suuriulotteinen ennustaminen, dynaaminen faktorimalli, regularisoitu regressio, tuottajahintaindeksi

Ohjaaja tai ohjaajat: Mika Meitz

Säilytyspaikka: Helsingin yliopiston kirjasto

Muita tietoja:

Tiivistelmä:

Kansantalouden nykyistä ja tulevaa tilaa koskevan ajankohtaisen tiedon tuottaminen on tärkeää käytännön talouspolitiikan näkökulmasta: politiikkatoimien toimeenpanon ja niiden vaikutusten ilmenemisen välillä on tyypillisesti merkittäviä viiveitä, mikä luo tarpeen ennakoida kokonaistaloudellisten suureiden kehitystä. Tuottajahintaindeksi on yksi tällainen makrotaloudellinen suure: tuottajahintaindeksien avulla pyritään seuraamaan kansantaloudessa tuotettujen hyödykkeiden yleisen hintatason muutoksia tuottajien näkökulmasta, mikä tekee niistä varteenotettavan inflaatiopaineen ja suhdanneolojen indikaattorin. Tämän tutkielman pääasiallisena tavoitteena on selvittää mahdollisuuksia kotimaisen teollisuuden tuottajahintaindeksin luotettavaan ennustamiseen lyhyellä aikavälillä hyödyntäen suurta ulkoisten ennustavien muuttujien joukkoa.

Ennustavien muuttujien lukumäärän kasvattaminen altistaa tavanomaiset ennustamismenetelmät epätarkkuuksille ja tekee niiden soveltamisen suoranaisen mahdottomaksi, kun muuttujien määrä ylittää mallin estimoimiseen käytettävissä olevien havaintojen lukumäärän. Ratkaisuksi tähän ongelmaan on ehdotettu lukuisia vaihtoehtoisia menetelmiä. Tutkielma tarjoaa laajan yleiskatsauksen näihin menetelmiin sekä muihin makrotaloudellisten muuttujien ennustamisen kannalta oleellisiin näkökohtiin.

Koska yksikään vaihtoehtoisista menetelmistä ei ole osoittautunut käytännön sovelluksissa yksiselitteisesti muita paremmaksi, tutkielman empiiriseen osuuteen on valittu sovellettavaksi menetelmiä, jotka edustavat kahta keskenään erityyppistä lähestymistapaa suuriulotteiseen ennustamiseen: dynaamisia faktorimalleja ja regularisoituja regressioita. Dynaamisten faktorimallien vaikuttavuus perustuu oletukseen, jonka mukaan suuriulotteisen aineiston sisältämä oleellinen informaatio voidaan tiivistää huomattavasti pienempään joukkoon taustalla vaikuttavia muuttujia, faktoreita, joiden estimaatteja voidaan käyttää edelleen ennustamiseen. Regularisoitujen regressioiden tarjoama ratkaisu taas perustuu ennusteeseen liittyvän harhan ja varianssin tasapainottamiseen. Laajempaan regularisoitujen regressioiden luokkaan kuuluvista menetelmistä tutkielmassa on käytössä neljä eri muunnosta: ridge, lasso, elastinen verkko ja adaptiivinen lasso.

Menetelmien empiiristä suorituskykyä arvioidaan toteuttamalla simuloitu otoksen ulkopuolinen ennustekoe, jossa kohdemuuttujalle estimoidaan historiallisen aineiston avulla sarja peräkkäisiä ennusteita verrattavaksi vastaavan ajanjakson toteutuneisiin arvoihin. Koejärjestelyn tavoitteena on tuottaa edustavaa tietoa ennustemallien tarkkuudesta jäljittelemällä tosiaikaisen ennustamisen olosuhteita: kunkin ennusteen tuottamiseksi hyödynnetään ainoastaan informaatiota, joka olisi ollut käytettävissä ennusteen laadinta-ajankohtana. Kokeessa käytettävien ennustavien muuttujien joukko koostuu eri lähteistä kerättyjen taloudellisten muuttujien kuukausittaisista aikasarjoista.

Ennustekokeen perusteella suuriulotteisten mallien etu keskimääräisessä ennustetarkkuudessa yksinkertaiseen autoregressiiviseen verrokkimalliin verrattuna osoittautuu ainoastaan marginaaliseksi yhden, kahden ja kolmen kuukauden päähän tähtäävillä ennustehorisonteilla. Myöskään käytettyjen suuriulotteisten menetelmien kesken ei havaita merkittäviä eroja ennustetarkkuudessa. Suotuisampia tuloksia saavutetaan sen sijaan käyttämällä suhteellisen nopeasti saataville tulevien markkinamuuttujien havaintoja indeksin samanaikaisten arvojen ennustamiseen tulevien arvojen sijaan. Tässä tapauksessa erityisesti regularisoidut mallit esiintyvät edukseen. Tulokset antavat osviittaa, että varteenotettavimmat mahdollisuudet tuottajahintaindeksin ennakoimiseen voisivat perustua ulkoisten muuttujien julkaisuviiveeseen liittyvän edun hyödyntämiseen indeksin samanaikaisessa ennustamisessa.

Table of Contents

1 Introduction

Being able to peer into the future is a notion that has without a doubt compelled man from the beginning of times. In macroeconomics, foresight is especially called for from a policy perspective, as the effects of policy measures usually manifest themselves only after considerable lags following implementation. This notion has given rise to an active field of research focused on developing methods for forecasting various macroeconomic variables.

The *producer price index* (PPI) is one such variable. The Finnish Producer Price Index for Manufactured Goods provides a monthly measure of the aggregate price-level of goods produced within Finland from the point-of-view of producers. The PPI has many applications. In a macroeconomic context, the aggregate index provides a useful indicator for policy makers for anticipating developments in overall economic circumstance. For example, producer prices can serve as an early indicator for inflation, reflecting cost-push forces stemming from the prices of domestically produced goods. In addition, changes in producer prices may signal turning points in the business cycle. Thus, being able to produce timely information regarding the evolution of producer prices is of great practical relevance.

Despite being an already relatively fast indicator of economic circumstances, the publication of the Finnish PPI lags behind the events that it intends to capture by nearly one full month, which is due to the meticulous data collecting process that underlies the index. This raises a question: would it be possible to augment the information content of the monthly publication by providing reliable predictions of the current and near-future values of the index? This is the main question that the present study seeks to answer. Specifically, the intention is to find out if any data available in advance of the publication date of the month's index could be useful in this regard. This poses a challenge, as the prices of the hundreds of individual goods that underlie the PPI could be affected by equally numerous external determinants, and it is not a priori clear which predictors out of a large number of candidates should be included in a model attempting to forecast the aggregate index. Therefore, a forecasting model able to accommodate large numbers of predictor variables is required.

While traditional statistical forecasting techniques begin to struggle as the number of predictors grows, the increasing availability of ever larger datasets in forecasting has spurred the development of various alternative methods specifically designed to cope with such data. The literature reports of numerous successful empirical applications of different methodologies in

forecasting macroeconomic variables, but no single framework has, so far, been found to dominate others across different applications. Owing to this, it is reasonable to consider some alternatives for the purpose of forecasting the PPI. To narrow down the range of candidates, two classes of forecasting methodologies, which have exhibited promising results in the literature, will be considered: *dynamic factor models* and *penalized regressions*.

By now, factor models represent an established way of dealing with high-dimensional data in macroeconomic forecasting. The appeal of dynamic factor models lies in their ability to summarize data in large panels of time series by extracting hypothetical factors that commonly drive the individual observable series. Over time, a number of different methods have been proposed for the purpose of factor estimation. In this study, the method based on principal components as proposed by Stock and Watson (2002a, 2002b) will be utilized.

Penalized regressions, on the other hand, represent a more recent approach to time series forecasting. Instead of summarizing data from large panels of time series, their strength lies in the ability to emphasize particularly relevant individual variables from large sets of candidates and construct forecasts based on the principle of bias-variance trade-off. The general class of penalized regressions encompasses a range of different methods. The specific variants considered in this study are the Ridge, Lasso, Elastic Net and Adaptive Lasso.

While somewhat arbitrary, the choice of methods is necessary in order to keep the scope of the study at a manageable level. However, these choices can be motivated in terms of the prevalence and empirical performance of the chosen approaches as well as their complementary features. For one, dynamic factor models are chosen because they represent an established approach to forecasting macroeconomic variables using large datasets. The dynamic factor model and its extensions have not only been shown to provide generally superior forecasting performance over the earlier parsimonious models, but they have also managed persist in the face of subsequent competing approaches. The penalized regression methods represent an emerging approach to macroeconomic forecasting: despite finding extensive application in general statistics before, they have only relatively recently begun to receive wider attention in time series applications.

The complementary nature of the two methodologies arises from the unknown characteristics of the data generating process of the forecasted variable. The effectiveness of the factor models relies on the existence of truly common factors shared by the target and predictors and the

failure of this assumption has been found to be detrimental to its forecasting capability (Boivin and Ng 2006). Penalized regressions, on the other hand, have been observed to be more robust to such specific assumptions and can leverage relevant information contained in individual predictors in the absence of common processes (Smeekes and Wijler 2018).

Additionally, some penalized methods may serve to supplement a particular deficiency of the factor models, which is the fact that the estimated factors provide little interpretability and tractability with regards to individual predictor variables. That is, while the factor models summarize information from the complete dataset, they are largely opaque with regards to the role of individual predictors. On the other hand, the penalized methods considered here work by singling out individual predictors based on their relevance to the forecast target, with most of them being capable of explicit variable selection and enhanced interpretability.

The aforementioned variable selection property of the penalized regression models gives rise to a final argument in favor of the particular choice of methods: in addition to representing independent approaches to forecasting by themselves, features of the dynamic factor and shrinkage methods can be combined to give rise to further means of leveraging information contained in the data. In particular, the penalized methods can be used to alleviate the sensitivity of the factor forecasts to mismatches between the set of predictors and the target variable by selecting a subset of the predictors prior to factor extraction (Bai and Ng 2008).

While the primary aim of this study is to assess the possibility of predicting the Finnish PPI in the short run, this setting also enables the study to contribute to the existing literature by providing evidence on the relative performance of the aforementioned high-dimensional forecasting approaches as a secondary objective. In addition, using recent data up to November 2021 also allows the study to touch upon another interesting and highly topical matter: the effect of the COVID-19 pandemic, which began in early 2020, on the forecasting performance of the high-dimensional models. The Finnish PPI has been markedly affected by the pandemic, which makes it a prime candidate to facilitate such an examination: while the early onset of the crisis was met with a sharp decline in aggregate prices, this fall was quickly countered by an incline of unprecedented magnitude, which is still ongoing as of the time of writing (see Figure 1). Such extraordinary events present a challenge to any forecasting procedure even in the short term and providing evidence regarding the performance of the models under such adverse conditions can be seen as another valuable addition to the forecasting literature.

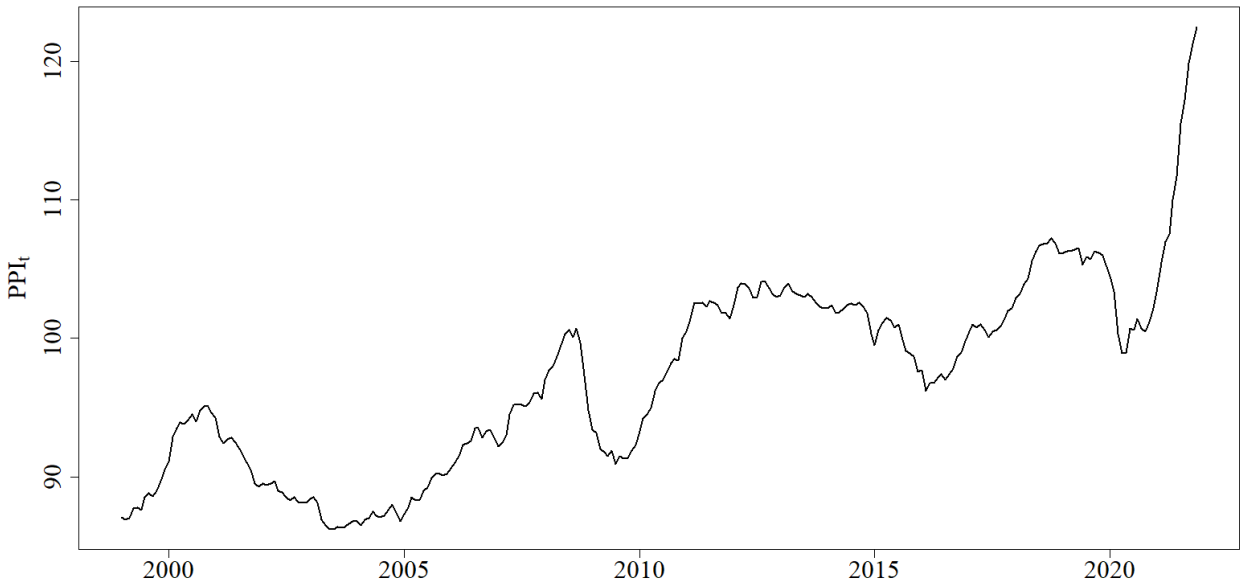

Figure 1. *Evolution of the Finnish Producer Price Index for Manufactured Goods over time.*

The structure of the study will be as follows. The next section provides a brief overview of producer price indices in Finland. In the third section, a general review of methodologies and other issues pertaining to macroeconomic forecasting will be presented, with an emphasis on methods applicable to high-dimensional data, dynamic factor models and penalized regressions in particular. The fourth section provides an account of the empirical methodology employed in this study: this includes an outline of the simulated out-of-sample forecasting experiment, which is used to assess the empirical performance of the different forecasting methods, as well as a detailed exposition of the forecasting approaches themselves. The fifth section presents the data used in the forecasting experiment. Results of the empirical out-of-sample experiment are presented and discussed in the sixth section, while the seventh section concludes.

2 Overview of the Finnish Producer Price Index

The purpose of the Producer Price Index (PPI) is to track aggregate developments in the prices of goods manufactured within a given economic area. As the name implies, the PPI is a closely related to the consumer price index (CPI) in concept: while the latter is used to measure changes in the general price level from the point-of-view of consumers, the latter measures inflation from the perspective of the producers of goods. In Finland, various producer price indices are compiled by Statistics Finland, the national statistical authority, on a monthly basis. The most prominent among these is the Producer Price Index for Manufactured Goods (*Teollisuuden tuottajahintaindeksi*), which measures the prices of industrial goods manufactured in Finland

for either local consumption or for export. Other related price indices include the Export and Import Price Indices, which measure the prices of exported and imported goods from the points-of-view of domestic producers and importers, respectively. It is noteworthy that the aforementioned indices are exclusively concerned with the prices of manufactured goods. The aggregate prices of services are tracked separately by the Producer Price Indices for Services, which consider the prices of services offered to both businesses and consumers alike on a quarterly basis. The Price Index for Manufactured Goods, which will be specifically referred to as the Finnish PPI throughout the paper, is the primary measure of interest in the present study. A brief overview of this index, based on Statistics Finland (2020), will be provided next.

The PPI has many general applications. Perhaps most prominent is its use as a deflator in various aggregate output statistics, where the price index is used to convert nominal quantities into real quantities. Owing to its relatively fast publication schedule, the PPI is also useful as an indicator of macroeconomic circumstances: systematic changes in producer prices may anticipate changes in consumer price inflation as well as signal changes in business cycle conditions. More business-oriented applications include the possibility of using good-specific subindices of the PPI as reference terms in long-term contracts. Tying the contract price to a reference index insures both suppliers and buyers to unexpected changes in the general price level relevant to the object of the contract.

As an aggregate price index, the Finnish PPI is constructed based on the prices of individual goods. Goods featured in the index are classified using the European Classification of Products by Activity (CPA), which assigns individual goods to categories based on common attributes among the goods. The CPA features a hierarchical structure, which allows goods to be aggregated at different levels of detail. The basic units of the Finnish PPI are goods categories identified at the most accurate level of the classification: the main index is formed from subindices corresponding to each category according to the Laspeyres index formula.

The PPI considers a majority of goods production in Finland: the featured categories are chosen in the order of their prevalence in terms of annual production, so that their combined production covers around 80% of the value of annual industrial production at the national level. As of 2018, the PPI includes over 600 individual goods categories. The share of annual production also determines the weight of each goods category in the calculation of the overall index. To maintain the correspondence of the price index with the structure of industrial production over time, the weights require regular updating. Prior to 2019, the weights were updated every five

years to coincide with the change of the base year of the index. Since then, a more frequent updating regime has been adopted, in which the weights as well as the composition of the index is updated annually to improve its representativeness. Owing to this, the base year of the index also implicitly changes annually, while the explicit base year of the index is still changed only quinquennially. To maintain commonality across years, the index is formed by chaining together these annual indices, that is, by extrapolating changes in the updated index to the earlier one (for details, see, Statistics Finland 2020, 25–26).

The price index pertaining to each individual goods category is based on a representative sample of prices of goods belonging to a given category. The price information is primarily collected directly from the manufacturers themselves and are tracked over time using monthly electronic surveys. That is, the same firms report the prices for the same products, usually each month. As of 2018, the sample underlying the PPI comprises the prices of over 3000 individual products, which are aggregated to form good-specific indices and further aggregates.

The monthly data collection cycle usually starts at the end of the month to which the price observations pertain to and concludes around three weeks later. During this time, the gradually accumulating responses are screened both automatically and manually to detect potential errors and anomalies. Following this, the finished data is used to calculate the indices, which are published around the 24th day of the following month.

3 Macroeconomic Forecasting

3.1 Standard Methods and the Curse of Dimensionality

The need to deduce the developments of tomorrow and beyond by using information available today has given rise to a broad field of research within macroeconomics, which has produced a range of alternative forecasting methods. Different types of linear regression models have traditionally been the workhorses of data driven forecasting. The most basic representatives of these are *autoregressive moving average* (ARMA) models, which rely solely on the past information of the variable to be forecasted, the *target variable*, to predict its future values. The dynamic process of the target is modeled as a linear combination of past observations and shocks occurring in the series. The *autoregressive* (AR) model is a special case and further simplification of the former, which encompasses only a regression of the future value of the target variable on its preciding values. While the obvious appeal of such univariate models lies in their simplicity of application and modest data requirements, they have also been found to provide

highly competitive forecasting accuracy in empirical applications. (Elliot and Timmermann 2016, chap. 7)

It is, however, not unlikely that forecasts could benefit from the inclusion of information in addition to that contained in the target variable itself. The *vector autoregressive* (VAR) model provides a convenient way to consider multiple variables in a forecasting application. The VAR effectively generalizes the AR model into a multivariate setting by considering multiple endogenous variables that are allowed to simultaneously interact with one another over time. This enables the model to trace out complicated relationships, both direct and indirect, between the variables considered. Thus, the VAR model can effectively, given the appropriate choice of variables, model the whole dynamic process of the data. Owing to this and its relative simplicity of application, the VAR model has enjoyed great popularity in forecasting and general econometric modelling. (Elliot and Timmermann 2016, chap. 9)

While VAR models provide a flexible multivariate framework, its use in macroeconomic forecasting may be limited by the nature of macroeconomic data, which is usually sampled at a relatively low frequencies of monthly or quarterly rates. As a result, the number of historical observations available for the estimation of the model is inherently limited, which gives rise to the so-called *curse of dimensionality*. The premise of the curse lies in the lack of degrees of freedom, which results when there are few observations available relative to the number of unknown parameters in a model requiring estimation. Owing to the curse of dimensionality, coefficient estimates in standard models are characterized by high sampling variance, which leads to erratic coefficient estimates and, ultimately, to poor forecasting performance. Moreover, increasing the number of parameters beyond the number of observations makes the estimation of models outright infeasible using standard statistical methods, such as ordinary least squares (OLS).

The most obvious way to alleviate the problem of dimensionality is by selecting a subset of the data of a size that can be processed by means of regular estimation methods. One way is to consult economic theory to choose variables that should play a relevant role in the data generating process. While, the choices of variables can be accomplished heuristically, more systematic ways of extracting the most relevant variables have also been developed to aid in the process. A traditional example of the latter is the so called general-to-specific approach, which involves gradually pruning down the dataset of predictor variables based on their explanatory power over the target variable (Elliot and Timmermann 2016, 93–96). Typically, the process

starts from a general model, that is, one with all possible explanatory variables, from which variables are gradually removed one at a time according to their explanatory power over the dependent variable. The process continues until a variable is encountered, whose explanatory power exceeds a preset threshold. The specific-to-general approach reverses the order of the process, starting from a model of just a single variable with variables added until the explanatory power of the last variable falls below the threshold. The latter approach is generally most useful in especially high-dimensional applications, where the initial general model may be infeasible to estimate (Lütkepohl 2007).

3.2 High-Dimensional Methods

3.2.1 Using Factors to Summarize Information

Owing to the availability of ever-increasing numbers of economic data during the latter part of the last century, a need was recognized for methods that could accommodate large numbers of variables to take advantage of their full information content in time series modelling and forecasting.

The dynamic factor model is perhaps the earliest approach to data reduction in time series to gain wide acceptance and one that remains prominent to this day. The basic assumption underlying the dynamic factor model is that observable economic variables relevant to the forecast, both the target and predictors, are driven by a considerably smaller number of unobservable underlying variables, the factors, which evolve over time according to their own dynamic process. Under this assumption, the observable series are merely measurable manifestations this process. It follows that the ideal way to analyze a given series would be to model them by using the factors directly, instead of the multitude of variables that only act as proxies to the underlying processes. Moreover, given that the factors should number few relative to the observable variables, their use in modelling would circumvent the curse of dimensionality that inhibits the use of standard estimation techniques.

The principal challenge pertaining to the use of common factors is the fact that they cannot be observed directly. They can, however, be estimated, and to this end various techniques have been proposed over the years. Foremost among these is the method proposed by Stock and Watson (2002a, 2002b), which is based on *principal components analysis*. Specifically, the factor estimates are recovered as the largest principal components extracted from the panel of predictor variables. The use of principal components as a tool for dimension reduction traces its

roots to cross-sectional applications. In these applications the efficacy of principal components for estimating the underlying factors had traditionally relied on the assumption that the observable variables were both cross-sectionally and serially uncorrelated, giving rise to so-called exact factor models. While this assumption can be supported in cross-sectional settings, it is hardly the case in time series contexts. However, Stock and Watson (2002a) show that the principal components methodology can be used to construct consistent estimators for the underlying factors even in the presence of serial correlation and weak cross-correlation among the observable variables. In this case, the factor model is said to be approximate as opposed to the exact non-dependent models. This consistency result is obtained as the both the number of predictor series used for estimation and the number of observations tend to infinity. Owing to this result in double asymptotics, the principal components approach can be thought of as effectively turning the curse of dimensionality into an outright blessing. That is, the estimates of the underlying factors and, by extension, the accuracy of forecasts based on them could actually be improved by expanding the set of predictors. Moreover, Stock and Watson (2002a) show that, under general conditions, the estimated factors can be used in a forecasting equation in place of the actual unobserved factors to yield consistent estimates using standard methods.

Since this inaugural work, interest in the dynamic factor framework has spurred the development of various alternative methods to estimate the factors for use in time series forecasting. These include the dynamic principal components estimator of Forni et al. (2005), the maximum likelihood estimator of Doz, Giannone, and Reichlin (2011), among others. While the alternative methods have been shown to provide some gains in forecasting accuracy over the static principal components estimator (Forni et al. 2005; Eickmeier and Ziegler 2008), the practical benefits of the more elaborate methods appear to be highly dependent on specific application and are typically marginal (see, e.g., Smeekes and Wijler 2018).

The ability of the factor framework to provide improvements in forecasting accuracy relative to low-dimensional methods has been showcased using real data in a number of empirical experiments. These include the inaugural papers by Stock and Watson (2002a, 2002b), in which the authors apply variations of the factor model to forecasting different US macroeconomic variables. Others include the study of Artis, Banerjee, and Marcellino (2005) on forecasting UK variables and Wang (2009), who compares the performance of factor and structural DSGE models in forecasting US output and inflation, only to name a few. Smeekes and Wijler (2018)

provide a more recent comparison of a range of different factor estimation methods, as well as other forecasting methods, in forecasting various US macroeconomic variables.

While the reported results are generally favourable to the factor models, there is also a great deal of variation involved: depending on the target variable as well as other circumstances, such as the country or period considered, the benefits of factor models can range from considerable to non-existent. This issue is conveniently showcased in the meta-study of Eickmeier and Ziegler (2008), in which the results of a large number of both published and unpublished empirical studies examining the performance of factor models in forecasting output and inflation are summarized. Their findings suggest that the factor models outperform low-dimensional benchmarks on average in terms of both target variables, but only somewhat marginally so, with both inferior and superior outcomes commonly encountered. Moreover, they find evidence of systematic variation in the geographic performance of the models. For example, the factor models appear to provide better output forecasts for the US than the euro area, while opposite applies for inflation forecasts.

The aforementioned results highlight the fact that, despite its potential, there is no guarantee that the factor framework can yield improvements over simpler alternatives in forecasting. While forecast failure may be caused by the fact that a given set of predictors is simply not informative regarding the future developments of certain target variables, another possibility is that the factor approach itself may be ill suited to extract relevant information from the predictors in some circumstances.

Specifically, the factor approach has been found to be sensitive to the composition of the data used for forecasting. While asymptotic results imply that increasing the number of predictors can be used to improve factor estimation, the practical utility of this notion with regards to forecasting performance relies profoundly on the assumption that the predictors and the target variable are indeed determined by a *common* underlying factor process. Furthermore, even if the common factor assumption is satisfied, the contribution of additional predictors also depends on their quality both in absolute terms and relative to other predictors. Specifically, predictors may be noisy, in the sense that the variation in their observed series is dominated by a random error term, rather than the common factor process, or they can be cross-correlated, in the sense that the errors of individual predictors are highly correlated with each other. Boivin and Ng (2006) show that adding such poor predictors to the set used for factor extraction can actually deteriorate the accuracy of the estimates. Moreover, they show that the forecasting

performance of the factor model can be poor if the set of predictors is dominated by variables that do not share the common factors with the target variable. The authors refer to this issue as *oversampling*. These results highlight the sensitivity of the factor framework to assumptions regarding model variables and their interrelations.

To counter the issues above in practice, Boivin and Ng (2006) propose various procedures to aid in the preselection of variables prior to factor extraction and show they can be useful in improving the forecasting accuracy of the factor model. These conclusions are also echoed by Caggiano, Kapetanios, and Labhard (2011). An alternative way of dealing with oversampling in particular is proposed by Bai and Ng (2008), who show that the performance of factor forecasts could be improved by preselecting predictors either by an algorithmic selection procedure based on regressions between the target and individual variables or by using variable selection based on penalized regression methods. Similar conclusions using these so-called *targeted predictors*, are reached by Eickmeier and Ng (2011), although some, for example, Medeiros and Vasconcelos (2016), have encountered less conclusive evidence.

Another option is to consider alternative forecasting methods altogether. The continued proliferation of large economic datasets has given rise to a number of competing methods capable of leveraging high-dimensional information. One such alternative, which has been found to be more robust than the factor models to features of the underlying data, is the group of penalized regressions.

3.2.2 Leveraging the Bias-Variance Trade-Off Using Penalized Methods

Various machine learning methods have been prominently utilized for prediction in non-dependent settings for a long time. Although the definition is generally vague, machine learning can be thought of as encompassing methods that rely on numerical procedures to solve a given optimization problem. This in contrast to, say, regular OLS estimation, which, owing to the existence of an explicit closed form solution solves the linear optimization problem by analytical means rather than procedural trial and error. Owing to pioneering empirical results and some theoretical advances, machine learning methods have relatively recently begun to make their way into economic forecasting applications as well (Masini, Medeiros, and Mendes 2021).

Machine learning approaches can be further divided into groups of linear and nonlinear methods. So-called penalized regression methods, also known as *regularized* regressions, are the foremost representatives of the class of linear methods. The underlying model pertaining to

these methods is a linear regression, which characterizes a linear relationship between a number of independent covariates and the target variable. Thus, it is worth underlining that penalized regressions do not by themselves entail a distinctive model, but rather provide an alternative method for the estimation of linear models with distinct benefits in comparison to standard methods.

While both penalized and OLS methods obtain coefficient estimates by minimizing the sum of squared residuals of the fitted model, the former methods perform optimization subject to a restriction on the regression coefficients. This is accomplished by way of including an additional term in the objective function, which levies a penalty on the loss function for large coefficient values. The presence of this *penalty function* causes the coefficient estimates produced by the penalized methods to be closer to zero, or shrunk relative to the corresponding OLS estimates. Owing to this, the penalized methods are also commonly referred to as *shrinkage* methods. While shrinkage introduces bias into the estimator, it also reduces its variance relative to the OLS estimator. That is, while the predictions produced by penalized methods may be slightly off target on average, the dispersion of the predictions tends to be smaller, which will, ideally, result in overall smaller prediction errors. This feature represents the main attraction of penalized methods with regards to high-dimensional prediction and balancing the trade-off between bias and variance is crucial to extract the most benefit out of the methods.

Different variations of the general penalized regression framework are further distinguished by the form of the penalty function that is used to constrain the linear optimization problem. Numerous alternatives have been suggested over the years to tackle specific estimation challenges and some of the most prominent methods applied to macroeconomic forecasting will be briefly discussed here. Two of the most fundamental representatives are the *Ridge* and *Lasso* estimators. The Ridge estimator, first proposed by Hoerl and Kennard (1970) was originally developed specifically to counter the detrimental effects of collinear covariates in linear regressions. While the Ridge can shrink the coefficients of each covariate arbitrarily close to zero, it cannot, notwithstanding pure chance, set them to exactly zero, thus making it incapable of outright variable selection even if some explanatory variables were truly irrelevant. The Lasso, short for 'Least Absolute Shrinkage and Selection Operator' (Tibshirani 1996) was developed specifically to address this deficiency. As its name implies, the Lasso can simultaneously perform variable selection in addition to shrinkage by truncating the coefficient estimates of irrelevant variables at exactly zero, thus being able to identify sparsity among the predictors. While

it is not clear a priori which methodology provides better results in term of pure predictive performance, the variable selection property offers a distinctive advantage in terms of interpretability.

Ideally, a process combining model selection and estimation should not only be able to recognize the correct variables but also estimate their coefficients effectively. In this case, efficiency means that the coefficient estimates of the process should converge to the OLS estimates recovered from a regression that involves only the relevant variables. This dual requirement is referred to as the *oracle property*. Recognizing the deficiency of the regular Lasso in this regard, Zou (2006) proposed the *Adaptive Lasso* to provide a penalized estimator with the oracle property. The Adaptive Lasso essentially augments the Lasso penalty by applying initial weights to the coefficients in the penalty function, which allows for varying amounts of shrinkage to be applied to different variables to enhance estimation results.

The *Elastic Net* was proposed by Zou and Hastie (2005) to address another shortcoming of the Lasso regression, namely, the fact that despite being able to handle sets of covariates larger than the number of observations, the latter can assign non-zero values only up to a number of variables equal to the number of observations. The Elastic Net circumvents this restriction by combining the penalties of the Lasso and Ridge estimators. Adaptive extensions, which combine features of the Adaptive Lasso and Elastic Net, have also been explored in the literature (e.g., Smeekes and Wijler 2018).

The *Group Lasso* is yet another modification of the basic Lasso. As the name implies, the group Lasso penalty, first proposed by Yuan and Lin (2006), encourages the selection of groups of related variables to either be included or excluded from the regression together. The variables are assigned to groups prior to estimation and can be based on prior knowledge such as economical relationship. The main attraction of the Group Lasso penalty is to enhance interpretability of the estimated model in comparison to the regular Lasso, which may exhibit inconsistent variable selection behavior within groups composed of similar predictors (Li and Chen 2014).

Theoretical research on the properties of penalized regressions in time series contexts is, so far, sparse and covers mostly only a few specific methods, with most results pertaining to the model selection consistency and oracle properties of the Lasso type estimators, the Lasso and Adaptive Lasso, in particular. A recent summary of these results is provided by Masini, Medeiros, and Mendes (2021). In a nutshell, research so far indicates that the favorable

properties of the Lasso and Adaptive Lasso in independent settings tend to carry over to time series applications as well.

Among the most substantial results with respect to high-dimensional forecasting applications are presented by Medeiros and Mendes (2016), who examine the properties of the Adaptive Lasso in estimating very general time series regressions, including models featuring numerous exogenous variables as well as their respective lags. Specifically, they show that the estimator retains the oracle property even as the number of candidate variables exceeds the number of observations, and the predictor series are characterized by conditionally heteroscedastic, highly correlated and non-normal error terms. They furthermore reaffirm their analytical results for finite samples by conducting a Monte Carlo experiment, in which estimation performance is tested on different time series data generating processes and sample sizes.

The lack of theoretical foundations has not, however, hindered the empirical application of penalized regressions in macroeconomic forecasting. Several studies have so far reported encouraging results regarding the forecasting prowess of penalized methods in combination with high-dimensional data. Given the prevalent status of the factor model, many studies also provide direct comparisons of the respective methods in different contexts. These include Eickmeier and Ng (2011), who apply the Ridge and Elastic Net to forecast output in New Zealand. Examples of forecasting various US macroeconomic variables include Li and Chen (2014), who use the Lasso, Elastic Net and Group Lasso; Medeiros and Mendes (2016) and Medeiros and Vasconcelos (2016) who apply versions of Lasso and Adaptive Lasso; and Smeekes and Wijler (2018), who utilize the Ridge, Lasso and Elastic Net, as well as adaptive variations of the latter two. Overall, the results indicate that the penalized methods, in general, offer a viable alternative to the factor models, with performance usually more or less on par with the latter. There is however some variation in the results, with the factor models providing superior results for some combinations of target and forecast horizon, while penalized regressions prove superior for others. Moreover, there is also similar variation among the different penalized methods.

While the literature provides little in the way of conclusive evidence in favor of either penalized or factor models, some of them manage to highlight the specific advantages of penalized methods with respect to factor models and where they might arise from. Eickmeier and Ng (2011) examine the effects of using different sets of predictors, either national, international variables or both, for forecasting output in New Zealand. They find that the performance of factor

forecasts generally deteriorates once the set of predictors is augmented with the international variables. This implies that the signal of the more relevant national variables may be crowded out by the less relevant international variables in factor estimation, which is consistent with the notion of oversampling raised by Boivin and Ng (2006). In contrast, the penalized methods considered, the Ridge and Elastic Net regressions, appear to be far less affected by the addition of potentially irrelevant data.

Smeekes and Wijler (2018) explicitly shed light on the implications of different data generating processes on the performance of factor models and penalized regressions respectively. As the data generating processes of empirical variables are, without exception, unknown, the authors conduct a series of Monte Carlo experiments, in which the respective methods are used to forecast simulated time series based on either an underlying factor process or a sparse subset of the observable variables directly. Their results indicate that, while the factor models have a slight edge in forecasting the former types of processes, the penalized models have a considerable advantage in forecasting the latter. Their subsequent empirical assessment of the methods in forecasting different US macroeconomic variables is, however, not as conclusive. In particular, they note that factor models tend to generally dominate penalized models for forecasting real variables, but the field is more even for nominal variables.

Overall, the studies discussed up to this point reflect the general difficulty of drawing conclusions regarding the superiority or inferiority of any particular forecasting approaches based on empirical assessments: the relative performance of given approaches depends profoundly on the features of the research design, with different predictors, targets and overall circumstances often leading to different outcomes regarding forecasting performance. This is most evident in large-scale forecasting experiments, such as the formidable study of Kim and Swanson (2014) that features dozens of different forecasting specifications, including factor models and penalized regressions, as well as other methods and their combinations. With eleven target variables considered in total, the choice of optimal model differs considerably between targets as well as across different forecasting horizons. Results such as this underline the fact that finding an optimal method among many candidates is a fundamentally empirical question that can generally be answered only on a case-by-case basis.

The empirical literature results does, however, indicate that both factor models and penalized regressions offer viable alternatives for forecasting with high-dimensional data. In addition, they indicate that these methods could complement each other by providing supplementary

features to account for the uncertainty involved in the data generating process of the target variable. For these reasons, the factor model and a selection of different penalized methods have been chosen as the basis of the present study. While this choice is necessary for practical reasons, it inevitably leaves some prominent alternative methods outside of consideration. The remainder of this subsection is devoted to a brief overview of some of these methods.

3.2.3 Other High-Dimensional Methods

The class of *nonlinear machine learning* methods encompasses a range of distinctive approaches that aim to estimate and approximate nonlinear functions by minimizing a loss function by way of different sequential algorithms. As the name implies, the main attraction of these methods lies in their ability to account for nonlinear relationships among predictor and target variables, which sets them apart from the penalized methods, whose alternatives for fitting the observed data are strictly limited to the linear domain. Specifically, this means that the response of the target variable to a change in one variable can depend on the values of other predictors, which is not an unlikely possibility in either cross-sectional or time series contexts. Specific methods commonly applied to time series forecasting include neural networks and different ensemble methods based on regression trees. A thorough treatment of these and other machine learning methods can be found in Hastie, Tibshirani, and Friedman (2009). Recent examples of the application of these methods to macroeconomic forecasting includes Maehashi and Shintani (2020), who apply nonlinear algortihms, along with penalized regressions and factor models to a selection of Japanese macroeconomic variables. While the results vary by target variable, the authors find that, in general, the nonlinear methods tend to be particularly beneficial at longer forecasting horizons.

The idea of *combining*, or *pooling* individual forecasts to improve accuracy has been an enduring idea in the literature. The foremost theoretical advantage of forecast combinations lies in the notion that averaging over a number of unbiased forecasts can be used to reduce the variance of the overall forecast relative to any single forecasts. Another potential benefit arises from the fact that combinations may be more robust to structural shocks occurring in the target series than any of the underlying individual forecast. Due to structural shocks, different individual models may provide superior forecasts during different periods. However, because it is only possible to identify such models after the fact, combinations of numerous candidate models may be useful in providing robustness for structural shocks in advance. To date, numerous different methods have been devised to this end, ranging from taking simple means or medians of

to applying time-varying weights based on past forecasting performance to combine individual models. (Elliot and Timmermann 2016, chap. 14)

Forecast combinations can be applied regardless of the type of model underlying the individual component forecasts. For this reason, it is not unusual that combinations are included in many comprehensive empirical experiments, as they are straightforward to implement once there is already a considerable number of individual forecasts available. While combining models that are already quite complicated and able to aggregate high dimensional data by themselves may seem superfluous, they have been, nonetheless, found to occasionally improve results in these cases as well (e.g, Li and Chen 2014; Kim and Swanson 2014).

Perhaps more organic applications of forecast combinations to the high-dimensional setting are those, in which large numbers of individual low dimensional models are aggregated to leverage expansive datasets. An example is found in Stock and Watson (2004), who apply different techniques to combine the forecasts of individual single predictor autoregressive distributed lag (ARDL) regressions. The ARDL is an extension of the basic AR framework, where lags of exogenous variables are included in the regression along with the lags of the target variable. They find that such combinations tend to outperform competing dynamic factor models in a forecasting exercise considering US output.

A distinct high-dimensional forecasting approach implicitly encompassing forecast combinations is the *Complete Subset Regression* (CSR) proposed by Elliott, Gargano, and Timmermann (2013). The basic idea behind CSR is to regress the target variable individually on every possible subset of a given size extracted from a set of predictors in turn. The ultimate forecast is obtained by averaging over each of the subset forecasts. While leveraging fairly basic linear regression methods at the grassroot level, the repetitive algorithmic nature and ensuing computational demand makes CSR essentially another representative of machine learning methods. The drawback of this approach is that, even if the size of the subset is limited, the number of possible combinations and, hence, number of underlying forecast models increases rapidly with the number of candidate predictors, which makes the procedure computationally demanding and even practically infeasible for particularly large datasets unless some preselection of the candidate predictors is carried out in advance. Favorable results regarding the empirical performance of the CSR approach in macroeconomic forecasting have been reported by Elliott, Gargano, and Timmermann (2015) for US variables, and Garcia, Medeiros, and Vasconcelos (2017) for Brazilian inflation.

The idea of using *Bayesian estimation* to overcome the limitations of the VAR framework in high-dimensional applications has been suggested by Bańbura, Giannone, and Reichlin (2010). They show that by imposing shrinkage on the model coefficients by way of choosing a suitable prior distribution, which reduces the coefficients of irrelevant variables to zero, the VAR framework can be extended to take advantage of as many as hundreds of endogenous variables along with their lags. Favorable forecasting results have also been reported by Koop (2013), who finds that different variations of the Bayesian VAR framework to be generally superior to competing factor models in forecasting US macroeconomic variables. On the other hand, Carriero, Galvão, and Kapetanios (2019) find that large-scale Bayesian VARs are on average outdone by both factor models as well as combinations of simple single predictor models in an output and inflation forecasting exercise encompassing multiple countries.

3.3 Additional Considerations for Forecasting

3.3.1 Nowcasting

Traditionally, forecasting has been predominantly concerned with predicting the strictly future values of given variables. Relatively recently, *nowcasting*, the objective of inferring the *current* values of variables has also received increasing attention in the macroeconomic forecasting literature. While assessing the current state of an economic variable may seem like a trivial task, the relevance of this issue can be justified based on the typically lengthy publication lags involved in producing macroeconomic time series. This issue was already alluded to in the case of the Finnish PPI, which is published close to a month following the end of the relevant time-period, but the issue is more pronounced for other variables such as the GDP, which are usually published at quarterly frequencies and with lags of several months. Such a considerable gap between the start of the period of interest and the eventual publication of the data opens up the possibility of using other, more frequent and more timely sources of information regarding the current period to gain an indication of the concurrent evolution of the variable of interest.

Although nothing prevents the use of regular forecasting models, there are some practical issues specific to nowcasting, which have warranted the development of more specialized methods specifically for the task. These challenges include missing observations at the end of the panel of predictors, which are due to disparate publication lags among the variables, and fitting together data observed at widely different frequencies (Elliot and Timmermann 2016, 498–499). The former issue, commonly referred to as the *ragged edge* problem, causes difficulty for

typical time series models, as they require the observations of different variables to be temporally aligned and absent of missing observations. One solution to the latter problem is proposed by Giannone, Reichlin, and Small (2008), who use methods based on Kalman filtering to estimate dynamic factors based on partially observed time series. They show that their approach is able to provide increasingly accurate nowcasts of the US GDP as monthly data gradually accumulates during the quarter.

The second issue arises from the fact that timelier predictors are typically observed at higher frequencies than the slow-moving target variable, and one might wish to include lags of the predictors in the model. While the low-frequency target can be regressed on the lags of the high-frequency predictors directly, this may require the estimation of a large number of parameters in case the temporal mismatch between the variables is great, which can, again, lead to degrees of freedom problems. For example, every lag of a quarterly target variable corresponds to three lags of a monthly predictor, and the contrast is even starker for combinations of monthly and daily variables. One solution to this problem is the *mixed data sampling regression* (MIDAS) originally proposed by Ghysels, Sinko, and Valkanov (2007). The idea of the MIDAS approach is to compress information contained in the lags of high frequency variables by approximating the lag structure with a small-order polynomial function, which reduces the number of parameters requiring estimation compared to estimating the coefficients of the high frequency lags directly. This makes the MIDAS regression essentially another form of dimension reduction, albeit from a slightly different viewpoint. Applications of the MIDAS in nowcasting include Heinisch and Scheufele (2017), who use weighted combinations of a number of single equation MIDAS and ARDL models based on monthly and quarterly predictors, respectively, to predict German output.

3.3.2 Estimation Window

A practical question concerning forecasting is, how much data should be used to estimate the models. The intuitively obvious answer to the question would be ‘as much as possible,’ but the matter is complicated by the potential instability of time series processes over time. Most frequently, either a *rolling* or *recursive* estimation window is used in the empirical literature. In the rolling window approach, the size of the estimation sample is kept constant by removing earlier observations as the *forecast origin*, the point at which a given forecasts is made, moves forward. In the recursive, or expanding window setup, the number of observations used for

estimation increases over time, as all previous observations are preserved in the estimation sample, while new ones are added.

There are theoretical advantages to both approaches. The rolling window can be thought as being more immune to errors arising from structural breaks in the series as the weight of earlier observations is gradually diminished by the rolling window. In the absence of structural breaks, the recursive window approach should yield benefits arising from improved estimation performance over time as the sample size grows. However, even in the presence of structural breaks, the benefits arising from consistency are not unlikely to outweigh the effect of improved resilience provided by the rolling window. (Elliot and Timmermann 2016, 376–379)

Owing to these considerations the recursive window has traditionally been favored in macroeconomic forecasting, while the rolling window has been more prevalent in finance, where relevant data is typically reported at a much higher frequency (Clark and McCracken 2009). However, more recently, the rolling windows have become more popular in macroeconomics as well. In practice, most empirical studies in macroeconomic forecasting tend to resort to either approach without much elaboration. Moreover, in the case of rolling windows, the length of the window is usually set to a somewhat arbitrary number of in-sample periods. For example, Smeekes and Wijler (2018) adopt a rolling window approach with 10 years' worth of observations for their monthly forecasting experiment. The formidable forecasting experiment of Kim and Swanson (2014) touches upon the subject of windowing schemes more explicitly, as the authors compare the performance of different forecast models across both rolling and recursive estimation windows. Their results indicate that, although the recursive window scheme prevails on average, neither approach clearly dominates the other.

Other, less frequently used alternatives include using exponentially declining weights to gradually discount past observations as the forecast origin moves forward. In contrast to the strictly discrete nature of the rolling and recursive window schemes, all observations are retained in the sample, but more recent observations have more influence over the outcome of estimation (Elliot and Timmermann 2016, 379). More elaborate alternatives include using combinations of rolling and recursive estimation windows. One example of such an approach is the method proposed by Clark and McCracken (2009), in which a convex combination of models estimated from both rolling and recursive windows is constructed based on an estimated structural break in the time series.

3.3.3 Contemporaneous Aggregation

A further possibility for forecasting arises from the fact that most macroeconomic time series are the result of some form of aggregation. That is, the ultimate series are the result of numerous sub-series summed together using some weighting scheme or the other. The situation with price indices is obvious: the Finnish PPI, for example, considers the price changes of hundreds of individual products, each of which occupies its own sub-index. These sub-indices are then aggregated together into gradually more general and larger indices using weights that reflect their prevalence in total industrial production, ultimately producing the main index.

This notion raises a question: instead of forecasting the aggregate series directly, could forecasting performance be improved by forecasting individual components of the aggregate individually and combining the forecasts afterwards using the known contemporary weights? The procedure described above is commonly referred to as *contemporaneous aggregation*. While forecasting a potentially large number of series individually is obviously more arduous than simply forecasting the aggregate in one fell swoop, the approach has some intuitively appealing characteristics. First, it is possible that the individual components of an aggregate series are uncorrelated and treating the aggregate as monolithic could leave relevant internal dynamics unaccounted for. Second, especially for various indices, the weighting schemes of the aggregate variables are subject to periodic updates. This constitutes an effective change in the data generating process of the aggregate, which is known in advance and could therefore be accounted for in forecasting. While the incremental changes are usually small, given the usually long spans of data used for estimation, they could have a non-negligible effect on forecast accuracy. Finally, contemporaneous aggregation can be thought of as a special case of forecast combinations, which implies that similar benefits in forecasting accuracy could also be attained. (Hubrich 2005)

Despite being outwardly appealing, empirical studies of on contemporaneous aggregation have produced predominantly unsatisfactory results. Hubrich (2005) finds that contemporaneous aggregation produces inferior results in comparison to direct forecasts for euro area inflation. While forecasting core inflation, that is, inflation excluding unprocessed food and energy products, benefits more, the improvements are only marginal. On the other hand, Moser, Rumler, and Scharler (2007) uncover somewhat better results for Austrian inflation, but these gains are nonetheless statistically insignificant. Similar conclusions are reached by Hendry and Hubrich (2011), who find that contemporaneous aggregation fails to improve upon forecasts of US

inflation. Instead, they suggest that using the component sub-indices as predictors for the aggregate variable in a dynamic factor framework could represent a more viable means of leveraging disaggregate information in forecasting.

4 Methodology

This section details the methodology applied in this study. The first subsection describes the out-of-sample forecasting experiment, which provides the general framework to assess the performance of different forecasting approaches. The second subsection provides a detailed description of these forecasting models.

4.1 Real Time Out-of-Sample Forecasting Experiment

The primary aim of this study is to find out, which, if any, of the models considered could be used to produce accurate forecasts of the Finnish PPI in the short term. This is foremost an empirical question, which will be investigated by way of a real-time out-of-sample forecasting experiment. Out-of-sample forecasting experiments are commonly used to assess the relative performance of different forecasting methods. The central idea in such experiments is to use historical data to simulate a sample of forecasts for the target variable over a given timespan in a realistic manner. That is, for each individual forecast, only information that would have been available at the forecast origin will be used to construct each forecast. Given the resulting sample of forecasts, the performance of different models with regards to each other is determined using measures of average accuracy.

In this study, the main out-of-sample period ranges from December 2011 to November 2021, yielding a total of 120 monthly point forecasts of the Finnish PPI for each combination of model and forecast horizon. In addition, some alternative out-of-sample periods will be considered to gauge the effects of certain extraordinary events on forecasting accuracy, the COVID-19 pandemic in particular.

To reflect a realistic setting, each of the models will be specified anew at each forecast origin to fully leverage the gradual accumulation of data. This entails not only the estimation of the forecast equation itself, which gives rise to new coefficient estimates, but also the tuning of additional parameters in the respective models. Furthermore, estimation will be based on the recursive, or expanding window approach. That is, all historical observations will be conserved in the estimation sample for consecutive periods.

It is worth noting that, while this experiment aims to reflect real time forecasting, it is not necessarily precisely so due to two reasons. First, statistics are sometimes subject to post-publication revisions. While the historical datasets used here contain only the most up to date versions of the respective time series, these may be different from the data that was initially published and would have been used in for an actual forecast. The second reason is related to the timing of publication of the data. That is, the predictors have been chosen based on their current publication time relative to the PPI, so that the relevant predictors would be available at the time of publication of the PPI. This timing may be subject to change, and information that would be suitably available at this point, may not have been so in the past or may not be so in the future. For these reasons, experiments such as this are usually referred to as *pseudo* real-time experiments. In this case, however, the concerns regarding revisions and timing should be small, as a majority of the predictors used are market-based variables, which are rarely subject to revisions and are available in a timely manner relative to the PPI.

The fact that the aggregate PPI is composed of a number of subindices based on different product categories opens up the possibility of constructing forecasts of the overall index from forecasts of individual subindices, that is, by contemporaneous aggregation. While contemporaneous aggregation could provide some benefits by accounting for the changing weights used to construct the main index over time, considering the generally underwhelming results with respect to contemporaneous aggregation discussed in the previous section, the added complexity involved in this approach is unlikely to be justified and, thus, will not be pursued in this study. Instead, the forecasts will be based on the aggregate index directly.

4.1.1 Measuring Forecasting Accuracy

Quantifying the forecasting performance of different models in forecast experiments is usually based on measures of aggregate error calculated over the sample of out-of-sample forecasts. While a multitude of different measures have been proposed to emphasize different aspects of forecasting accuracy (see, eg., Hyndman and Koehler, 2006), by far the most common ones used in macroeconomic forecasting studies are various scale dependent error measures. The most prominent of these are the *mean absolute error* (MAE), *mean squared error* (MSE) and *root mean squared error* (RMSE). Each of these aggregate errors measures the average deviation of the forecast from the realized value of the target variable. In the present study, the MAE and RMSE will be utilized.

Consider an observation of the target variable y_t at time t and the respective forecast from a given model $\hat{y}_t$. The forecast error at time t is then defined as $e_t = \hat{y}_t - y_t$. The out-of-sample experiment produces a sample of forecast errors $e = (e_1, \dots, e_T)'$ for the particular model. The MAE is defined as the average of absolute forecast errors:

$$MAE = T^{-1} \sum_{t=1}^{T} |e_t|.$$

Correspondingly, the RMSE is defined as the square root of the average squared forecast errors, that is:

$$RMSE = \sqrt{MSE} = \sqrt{T^{-1} \sum_{t=1}^{T} e_t^2}.$$

While both MAE and RMSE essentially measure the dispersion of the forecasts from the realized values, there is a subtle difference between the measures. Due to the squaring of individual errors, the RMSE gives a stronger emphasis to large deviations compared to the MAE. That is, the RMSE punishes relatively larger errors more than small ones (Verbeek 2017, 83). This is obviously a feature shared with the MSE, which is possibly the most commonly encountered measure of error in the forecasting literature. In this case, RMSE is chosen over the latter, because it is reported on the same scale as the target variable, which makes the results more straightforward to interpret.

4.1.2 Statistical Significance of Forecasting Results

Differences in observed forecast accuracy using a finite sample can occur due to sampling variation instead of true differences in underlying performance. Tests of statistical significance can be used to lend further credence to conclusions based on these observations. The most commonly used means of assessing relative forecasting performance in the literature is the test of equal predictive accuracy proposed by Diebold and Mariano (1995). The Diebold-Mariano (DM) test provides a pairwise comparison of two series of forecasts based solely on the differences in their squared forecast errors. Thus, the DM test is agnostic with respect to the specific features of the models: only information on individual prediction errors is required to conduct the test. While this makes the test easy to apply, it also follows that one must take care in interpreting the results. As noted by Diebold (2015), the DM test is fundamentally a test of the accuracy of forecasts, not of forecast models. As such, the test answers the question of whether the historical performance of the respective forecasts would have been significantly different from

each other, which does not guarantee that the relative performance will necessary carry over to the future. Despite this shortcoming, the DM test will be utilized in this study due to its prevalence in macroeconomic forecasting literature.

The specific version of the DM test used here includes the small sample modification proposed by Harvey, Leybourne, and Newbold (1997). Consider two series of T h-step-ahead forecast errors from different models, $e_1 = (e_{1,1}, \dots e_{1,T})'$ and $e_2 = (e_{2,1}, \dots e_{2,T})'$ respectively. Furthermore, defining the squared forecasting error differential as $d_t = e_{1,t}^2 - e_{2,t}^2$ and the sample mean of the error differential as $\bar{d}$, the standard DM test statistic is expressed as

$$DM = \frac{\bar{d}}{T^{-1}\sqrt{\hat{\gamma}_0 + 2\sum_{k=1}^{h-1}\hat{\gamma}_k}},$$

where $\hat{\gamma}_k$ is the kth sample autocovariance of the error differential d_t:

$$\hat{\gamma}_k = T^{-1}\sum_{t=k+1}^{T}(d_t - \bar{d})(d_{t-k} - \bar{d}).$$

Under the null hypothesis of equal predictive ability, the original DM statistic follows a standard normal distribution. Noting that the standard test tends to reject the null too often in small samples, Harvey, Leybourne, and Newbold (1997) suggest a refinement of the following form to the statistic:

$$DM_{HLN} = DM\sqrt{\frac{T + 1 - 2h + T^{-1}h(h-1)}{T}}.$$

Given this rescaling, the critical values of the test are obtained from a Student's t distribution with $T - 1$ degrees of freedom.

4.2 Forecasting Models

In this section, a closer examination of the specific forecasting models utilized in this study will be presented. As noted above, the main models of interest encompass two distinctive approaches to forecasting with high-dimensional data: dynamic factor models and penalized regressions. The dynamic factor model used is based on the approach proposed by Stock and Watson (2002a, 2002b), in which static estimates of the underlying dynamic factors are recovered using principal components analysis. For penalized regressions, the four variations to be considered are the Ridge, Lasso, Elastic Net, and Adaptive Lasso estimators. In addition, a

hybrid model combining the variable selection properties of the Lasso estimator and the ability of the factor models to summarize information in large datasets, the targeted predictors approach, as originally proposed by Bai and Ng (2008) will be applied.

Finally, two simple *benchmark models* will be applied to act as yardsticks for the high-dimensional models. These are the random walk and univariate autoregressive models.

4.2.1 Dynamic Factor Model

The forecasts of the dynamic factor model used here are based on the following forecast equation:

$$\hat{y}_{T+h}^{h} = \hat{\mu}^{h} + \sum_{p=1}^{P} \hat{\alpha}_{p}^{h} y_{T+1-p} + \sum_{r=1}^{R} \hat{\beta}_{r}^{h} \hat{F}_{r,T}. \tag{1}$$

Here, $\hat{y}_{T+h}^{h}$ is the h-step-ahead forecast of the target variable and $\hat{F}_{r,T}$ is an estimate of the rth static factor, whereas $\hat{\mu}^{h}$ is the estimated intercept of the model, and $\hat{\alpha}_{p}^{h}$ and $\hat{\beta}_{r}^{h}$ are coefficient estimates pertaining to the pth lag of the target variable and rth factor estimate respectively.

A notable feature of the forecast equation is the way in which multistep forecasts are obtained. Specifically, a distinct model is estimated for each forecast horizon individually, which is emphasized by the superscript h in the parameters of the model. This approach is referred to as *direct* multistep forecasting, which is in contrast to the *iterative* approach typically used in AR and VAR models, where the one-step-ahead forecasts are used iteratively as inputs to the model to construct forecasts for subsequent horizons.

Another noteworthy point is that the factors enter in the forecast model only in the form of estimates owing to their de facto unobservable nature. While the forecasting model itself is estimated by OLS, this step is preceded by the estimation of the factors. Before addressing the issue of factor estimation, it is instructive to demonstrate the general assumptions regarding the dynamic factor model in closer detail. The following exposition is based on Elliot and Timmermann (2016, chap. 10). The central assumption underlying the factor approach is that the data generating process of each observable, including the target variable and each predictor $x_{i,t}$, is characterized as a combination of a factor-driven *common component* and an *idiosyncratic component* as follows:

$$x_{i,t} = \lambda_{i}(L)' f_{t} + \varepsilon_{i,t}. \tag{2}$$

Here, $\lambda_i(L)$ is a Q-dimensional vector of lag polynomials and f_t is a Q-dimensional vector of dynamic factors, which together form the common component of $x_{i,t}$. The common component is so named because it transmits the influence of the common factors to each observable variable. The extent to which the factors affect each observable variable is determined by the variable-specific *dynamic factor loadings* contained in $\lambda_i(L)$. On the other hand, $\varepsilon_{i,t}$ is a random innovation term, which represents the idiosyncratic component. While the idiosyncratic component is unique to each variable, it can be cross-correlated with the idiosyncratic components of other variables. Moreover, it can also be serially correlated.

The dynamic factors themselves are determined by a vector autoregressive process of the following form:

$$f_t = \Psi(L) f_{t-1} + \eta_t.$$

Here, $\Psi(L)$ is $Q \times Q$ matrix of lag polynomials, which determine the interactions of the factors and their lags among themselves, whereas η_t is a vector of Q random innovations. Under these assumptions, the dynamics of the observable variables ultimately arise from two primary sources: the idiosyncratic components of the variables themselves and the innovations of the factor process. Under these assumptions, the observable variable can alternatively be expressed in terms of lagged values of itself and the dynamic factors:

$$x_{i,t} = \alpha(L) x_{i,t-1} + \beta_f(L)' f_{t-1} + u_{i,t}. \tag{3}$$

Here, $u_{i,t}$ is another idiosyncratic term that, this time, collects the effects of all of the unpredictable shocks affecting $x_{i,t}$ that arise from both $\varepsilon_{i,t}$ and η_t (for details, see, Elliot and Timmermann 2016, 222). The above equation casts the factor process in a form that could already be used for forecasting. However, the inclusion of lagged factors prevents the use of the principal components approach, which is a fundamentally cross-sectional procedure, to estimate the factors. As long as the lag-order of the dynamic factors is finite, the dynamic factor process in equation (2) can equivalently be expressed in the following, so-called static form to facilitate estimation by principal components analysis:

$$x_{i,t} = \Lambda_i' F_t + \varepsilon_{i,t}.$$

Here, F_t is an r-dimensional vector of static factors and Λ_i is the corresponding vector of static factor loadings. $F_t = (f_t, \dots, f_{t-p})'$ contains the current and lagged values of the dynamic factors in stacked vector form. That is, the vector contains all past and present values of the dynamic

factors that have a contemporaneous effect on the observable variables. Given this alternative representation, equation (3) can now also be expressed in terms of static factors as

$$x_{i,t} = \alpha(L)x_{i,t-1} + \beta' F_t + u_{i,t}.$$

This gives rise to an estimable forecast equation corresponding to equation (1). While this discussion justifies the use of common factors for forecasting, the obstacle posed by the unobservability of the common factors remains. The estimation of these factors will be discussed next.

4.2.1.1 Estimation of the Factors Using Principal Components

Estimation of the unobservable factors will be carried out using the principal components method proposed by Stock and Watson (2002a, 2002b), which recovers estimates of the R static factors, or the space spanned by them, as the R largest principal components of the panel of the predictor variables. As shown by Stock and Watson (2002a), these estimates can be used in place of the true, unobserved factors to produce asymptotically equivalent forecasting results.

Although not strictly necessary, the standardization of the predictors prior to factor extraction by principal components has been found to be beneficial in empirical applications (Elliot and Timmermann 2016, 232). This approach will be followed in this study as well. Specifically, standardization entails transforming each observable variable to have zero mean and unit variance. That is, $x_{i,t} = (\tilde{x}_{i,t} - \bar{x}_i)/\hat{\sigma}_i$, where $\tilde{x}_{i,t}$ is the original value of the series at time t, and $\bar{x}_i$ and $\hat{\sigma}_i$ are the sample mean and standard error of the series calculated up to time t, respectively.

The static presentation of the factor process can be compactly expressed for the panel of observable variables $X_t = (x_{1,t}, \dots, x_{N,t})'$ as

$$X_t = \Lambda F_t + \varepsilon_t,$$

where Λ is an $N \times R$ matrix containing the static factor loadings for each individual variable, while ε_t is a vector containing the N respective idiosyncratic components. The principal components estimator aims to recover the factors and their loadings from the set of observable variables by solving the least squares optimization problem

$$\min_{F_1,\dots,F_T,\Lambda} \frac{1}{NT} \sum_{t=1}^{T} (X_t - \Lambda F_t)'(X_t - \Lambda F_t) \tag{4}$$

subject to certain normalizations regarding the factor loadings and the covariance matrix of the factors. Specifically, the covariance matrix of the factors, $\Sigma_{FF} = E(F_t F_t')$, is assumed be diagonal and $N^{-1}\Lambda'\Lambda = I_R$. The need for these restrictions arises from the fact that both Λ and F_t are unobserved, which means that extracting either one is not possible without imposing further restrictions. Even in this case, arbitrary normalizations can be used only to identify the space spanned by the factors, that is, some linear combinations of the underlying factors, which depend on the normalization used, instead of the factors themselves. However, for the purpose of forecasting, this fundamental restriction is of no concern and the aforementioned normalizations provide a 'mathematically convenient' way to extract the relevant information from the factors. (Stock and Watson 2016, 425–426)

The solution to (4) subject to the aforementioned normalizations can be expressed as,

$$\hat{F}_t = N^{-1}\hat{\Lambda}' X_t,$$

where the estimate of the factor loadings, $\hat{\Lambda}$ is composed of the eigenvectors pertaining to the R largest eigenvalues of the sample covariance matrix of the observed variables, which, given the standardized predictors, equals $\hat{\Sigma}_{XX} = T^{-1}\sum_{t=1}^{T} X_t X_t'$. That is,

$$\hat{\Lambda} = [m_1, \dots, m_R],$$

where m_r is the eigenvector corresponding to the rth largest eigenvalue of $\hat{\Sigma}_{XX}$. Thus, the vector of static factor estimates $\hat{F}_t$ are the R largest principal components of X_t.

4.2.1.2 Model Selection

The central model selection issue pertaining to dynamic factor models is the determination of the number of factor estimates to be included in the forecast equation. Using the principal components procedure, factors estimates are extracted in the order of their relevance to the overall variation in the set of predictors: while the first factor estimate explains the largest individual fraction of variation, the explanatory power of subsequent factors gradually declines. The number of factors estimates that the principal components process extracts can at most equal the total number of variables N. Including too many factors would obviously defeat the purpose of the dimension reduction procedure altogether, while including too few would run the risk of neglecting potentially relevant information contained in the omitted factors. Thus, to identify the true number of static factors R, additional methods need to be employed.

Over time, a number of different approaches have been proposed to this end, ranging from visual inspection of so-called scree plots, which depict the contributions of successive factors to the overall variation of the set of predictors, to different information criteria and tests (Stock and Watson 2016, 435). Arguably the most established tools used in the empirical forecasting literature are the information criteria proposed by Bai and Ng (2002). Each criterion is optimized by choosing a number of factors that minimizes the sum of squared residuals resulting from fitting the observable variables to the estimated factors subject to a penalty factor that accounts for both the number of variables and number of observations used in estimation. The authors show that the information criteria consistently estimate the number of static factors and demonstrate favorable small sample behavior in a Monte Carlo experiment.

Among the more recent additions to the set of factor selection tools is the eigenvalue ratio test proposed by Ahn and Horenstein (2013). In this case, the optimal number of factors k^* chosen is simply the one that maximizes the ratio of two adjacent ordered eigenvalues of the covariance matrix of the predictors. That is,

$$k^* = \underset{k}{\operatorname{argmax}} \left(\frac{\mu_k}{\mu_{k+1}} \right),$$

where μ_k is the kth largest eigenvalue of Σ_{XX} and $k \in (1,2, \dots, k_{max})$. Despite its overwhelming simplicity, the authors show that the eigenvalue ratio test provides an equally consistent estimate of the true number of factors as well as providing generally superior small sample performance with respect to a number of alternative methods based on Monte Carlo simulations. Furthermore, their results indicate that the eigenvalue ratio test is less sensitive to the arbitrary choice of k_{max}, the upper bound for the number of factors, than the Bai and Ng (2002) information criteria.

The Ahn-Horenstein eigenvalue ratio test is chosen as the factor selection method applied in this study based on the aforementioned results and, equally importantly, owing to the lesser computational burden involved in its application compared to the alternative information criteria, which is no trivial consideration, given that the number of factors must be estimated individually for each combination of forecast origin, horizon and model. As the eigenvalues of the covariance matrix of the predictors are already derived during principal components estimation, the application of the eigenvalue ratio test is particularly straightforward.

With the number of factors decided upon, another specification choice concerns the lag order P of the AR component in (1). Following existing literature, lag selection will be conducted by using standard information criteria. In the empirical experiments, the maximum number of lags will be set to six. Commonly used information criteria include the Akaike and Bayesian information criteria (AIC and BIC, respectively). Based on the results of Marcellino, Stock, and Watson (2006) regarding macroeconomic forecasting applications, the AIC will be used here.

4.2.1.3 Targeted Predictors

Given the practical concerns raised over the sensitivity of the factor forecasts to potentially irrelevant variables in the set of predictors (Boivin and Ng 2006), a modification of the basic dynamic factor approach will also be considered in this study. Specifically, this extension is a variant of the targeted predictors approach proposed by Bai and Ng (2008). In this approach, variable selection techniques are used to select a subset of the predictors, which are, in turn, used to estimate the factors that are ultimately used for forecasting. This preliminary step aims to remedy the potential shortcoming of the basic approach that it takes no account of the relationship between the target variable and predictors in factor estimation, which can lead to a mismatch between the estimated factors and the forecasted variable, or oversampling. In an empirical evaluation, the authors find that factor forecasts based on a subset of no more than thirty predictors as selected by penalized regressions or the closely related LARS algorithm tend to improve upon factor forecasts based on the complete set predictors. Similarly encouraging results have also been reported by Eickmeier and Ng (2011).

Based on these results, a similar arrangement will be tried out in this study. Specifically, the Lasso, a variant of the penalized regression methods, will be initially used to determine a subset of 40 predictors for each forecast individually. Variable selection is carried out by adjusting the shrinkage applied by the Lasso regression to produce a model featuring exactly the desired number of nonzero coefficients and selecting the corresponding predictors into the subset used for factor estimation. Thus, only the variable selection features of the Lasso will be utilized in this application. Furthermore, the size of the subset will remain constant, but its composition may change from period to period. Apart from the initial screening of the predictors, model selection and estimation will be carried out in a manner no different from the regular dynamic factor model as described above.

4.2.2 Penalized Regressions

While a number of different penalized methods will be considered here, they all share a common forecasting equation. At the h period forecast horizon, it encompasses a linear projection of the h-step-ahead target variable onto the N predictor variables and an AR component encompassing lagged values of the target itself as follows:

$$\hat{y}^h{}_{T+h} = \hat{\mu}^h + \sum_{p=1}^{P} \hat{\alpha}_p^h y_{T+1-p} + \sum_{n=1}^{N} \hat{\beta}_n^h x_{n,T}.$$

That is, the forecast equation assumes the general form of an AR model with external variables specification, which is not unlike the forecast equation of the dynamic factor model earlier. Consequently, multistep forecasts of the target variable are obtained by way of the direct approach in this case as well. The only differences with respect to the factor model are in the predictors used: the forecast equation now includes each of the N standardized predictor variables $x_{i,t}$ directly instead of the factors extracted from them. To simplify notation, explicit reference to the lagged values of the target variable will be excluded from this point on. This can be done without loss of generality because the coefficients of these lags will be estimated by the penalized methods impartially with the other predictors. Thus, the past values of the target variable can be thought of as being included among the set of exogenous variables, represented by some individual $x_{i,t}$.

While the forecast equation is a rather straightforward linear regression affair, its estimation using regular means becomes increasingly imprecise as the number of predictors increases relative to the number of available observations, and outright infeasible once the regressors outnumber observations. Penalized regressions remedy this problem by imposing a restriction on the coefficient vector. The distinguishing characteristics of penalized regressions is best illustrated by the constrained optimization problem by which the estimated coefficient vector is obtained:

$$\hat{\beta}^h = \underset{\beta^h}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_{t+h}^h - \mu^h - \sum_{n=1}^{N} \beta_n^h x_{n,t} \right)^2$$

$$s.t.\ p(\beta^h) \leq s.$$

Here, $\beta^h = (\beta_1^h, \dots, \beta_N^h)'$ is vector containing the N regression coefficients and $p(\beta^h)$ is the positive real-valued penalty function, which is increasing in the norm of the coefficient vector. In

general, a norm is some measure of the distance of the coefficient vector from the origin of the coefficient space. On the other hand, s is a positive constant, which determines an upper bound for the penalty. Thus, in the restricted optimization problem, the residual sum of squares is minimized by any vector of coefficients, whose penalty lies within the boundary imposed by s. It follows that for values of s large enough, the optimization problem becomes an unrestricted one with a solution equivalent to the OLS optimization problem. Conversely, for a value of s small enough, the solution of the problem can only be obtained by way of a vector whose norm is reduced relative to the OLS solution, which gives rise to the bias-variance trade-off that underlies the utility of the penalized methods.

An alternative way of expressing the constrained optimization problem is in the following Lagrangian form, in which the penalty function is nested in the optimization equation directly (Hastie, Tibshirani, and Friedman 2009, 63):

$$\hat{\beta}^h = \underset{\beta^h}{\operatorname{argmin}} \left(\sum_{t=1}^{T} \left(y_{t+h}^h - \mu^h - \sum\nolimits_{n=1}^{N} \beta_n^h x_{n,t} \right)^2 + \lambda p(\beta^h) \right). \tag{5}$$

Here, the penalty function is scaled by the non-negative parameter λ, which is known as the complexity parameter. The complexity parameter has a direct correspondence with s, although inversely so: the larger λ is, the stronger the shrinkage applied to the coefficient vector is, and vice versa. The latter point is easily seen by setting λ to zero in (5), which makes the penalty function vanish from the optimization problem altogether and gives rise to the typical OLS optimization problem. The Lagrangian form of the problem can be intuitively interpreted as balancing model fit and coefficient magnitude: if increasing a single coefficient value improves the fit of the model, but by less than its marginal effect on the size of the penalty scaled by λ, the loss function increases in value. This can be contrasted to the solution of regular OLS optimization, in which the higher value would be deemed more optimal given the improvement in fit, no matter how small it is. This also makes explicit the role of λ, a higher value thereof imposing a stronger emphasis to the penalty with respect to any improvement in model fit. Thus, finding an optimal value for the complexity parameter is key to leveraging the bias-variance trade-off of the penalized methods. Optimization of the complexity parameter is foremost an empirical question that depends on the characteristics of the data and is usually accomplished by way of computational methods. The precise means to this end will be discussed in detail further on.

Like the factor models, shrinkage methods generally require the predictors to be standardized prior to application. In this case, standardization of the series is performed for two reasons. First of all, the mechanism of shrinkage is based on rationing the magnitude of the regression coefficients. If the scales of the predictor variables vary, the estimated coefficient of a given predictor reflects not only its ability to explain variation in the target variable, but also its scale relative to the other predictors. By extension, this means that the amount of shrinkage applied to each variable also depends on its scale, not its relevance, which can lead to excessive shrinkage of relevant variables and inadequate shrinkage of irrelevant ones. Standardizing each predictor to have unit variance rectifies this problem. Second, the demeaning of the predictors is performed to conserve the intercept term $\hat{\mu}^h$ in the optimization problem from unnecessary shrinkage. This is accomplished by fitting the restricted model based on the demeaned target and predictors without an intercept. The sample mean is then added to both sides of the resulting estimated regression equation to produce a forecast equation of the original target variable. (Hastie, Tibshirani, and Friedman 2009, 63–64)

The inclusion of the penalty function imposes practical limitations to the estimation of the penalized regressions: with the exception of the Ridge regression, the solutions to the constrained optimization problem generally have no closed form analytical solutions (Hastie, Tibshirani, and Friedman 2009, 68). Thus, to find the minimum of the restricted sum of squares, computational optimization algorithms are required, which makes the penalized regressions essentially machine learning methods, despite their outwardly similarity to regular regression problems. The software used to apply the penalized regressions methods in this study, the R package glmnet, utilizes the so-called *coordinate descent* algorithm to solve the optimization problem. Coordinate descent minimizes the objective function by adjusting a single coefficient at a time while keeping the rest of the parameters fixed in the interim. Upon encountering a minimum along the path of the single parameter, the algorithm switches to another parameter and adjusts it to further minimize the objective. The process continues until altering any parameter no longer decreases the value of the objective function relative to a convergence criterion, indicating that a local minimum has been encountered. As the penalized regression problems considered here are all convex problems (Hastie, Tibshirani, and Friedman 2009, 92), any local minimum encountered is also a global minimum, which makes application of coordinate descent feasible.

With the general structure of the penalized regression problem explained, a closer examination of the specific members of the family of methods to be considered in the present study and their features will be carried out next, starting with the Ridge regression.

4.2.2.1 Ridge

The progenitor of the penalized regressions methods is the Ridge regression originally proposed by Hoerl and Kennard (1970) as a means of dealing with highly correlated regressors in cross-sectional contexts. The specific form of the penalty function of the Ridge regression is as follows:

$$p(\beta) = \sum_{n=1}^{N} \beta_n^2.$$

That is, the penalty function equals the sum of squares of each individual coefficient, which gives rise to the penalty function measured as the Euclidian norm, also known as the l_2 norm, of the parameter vector: the further away the vector is from the origin of the coefficient space, the larger a value the penalty function yields. Due to the characteristics of the penalty function, the Ridge regression effectively limits the estimated coefficient vector to an N-dimensional ball around the origin, with the complexity parameter controlling for the radius of the ball.

The interaction between the Ridge penalty and the general optimization problem can be illustrated in a simple two-variable case in frame (a) of Figure 2. The restriction imposed by the Ridge penalty is represented by the gray disc covering the intersection of the coefficient axes: this area contains the set of possible solutions to the restricted optimization problem. The unrestricted OLS solution to the problem, represented in the figure by β^{OLS}, resides in the upper right quadrant of the diagram. The ellipses around the unrestricted solution represent contours lines, along which the residual sum of squares of the model given a combination of coefficients is equal: the closer to the OLS solution the ellipses are, the smaller the corresponding sum of squared residuals is. The solution to the Ridge problem is found at a point where a contour line lays tangent to selection boundary, which are the coefficient values associated with the smallest residuals that comply with the restriction.

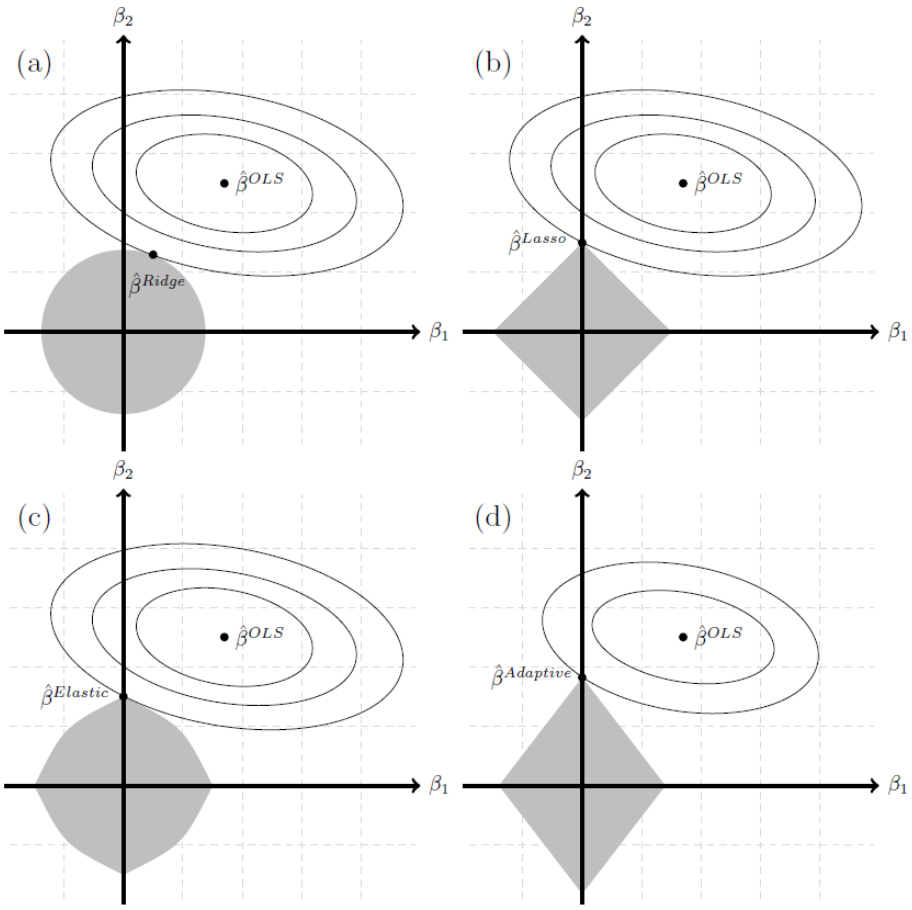

Figure 2. *Illustration of the selection regions of the Ridge (a), Lasso (b), Elastic Net (c), and Adaptive Lasso (d) regressions in two-dimensional space.*

As the penalty function is differentiable, unlike the other shrinkage methods considered here, the Ridge estimator also has a closed form solution, which can be derived from the Lagrangian form of the optimization problem in a manner similar to the regular OLS problem by deriving the normal equations of the optimization problem. In matrix notation, the general form of the Ridge estimator is as follows:

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'y.$$

Owing to the restriction, the Ridge estimator can also be defined uniquely in case where X is not of full column rank. This includes not only situations where there are perfectly correlated variables in X but also situations where there are more columns than rows, that is, more variables than observations, which makes the approach applicable for high-dimensional problems. Another consequence of the continuous penalty function is that while the regularization reduces individual coefficient values arbitrarily close to zero, they are run down to exactly zero only by

coincident. This can be seen in panel (a) of Figure 2: a given coefficient is reduced to zero only if the residual contour happens to lay tangent to the circular selection boundary at exactly the point of an axis. This is the reason why the Ridge regression cannot perform outright variable selection in the sense of discarding irrelevant variables completely.

4.2.2.2 Lasso

The Lasso regression, first proposed by Tibshirani (1996), seeks to combine both shrinkage and variable selection into a single estimator in a setting, where the true underlying data generating process may be characterized by sparsity in the candidate variables. The Lasso estimator seeks to identify the relevant variables, and is characterized by the following penalty function:

$$p(\beta) = \sum_{n=1}^{N} |\beta_n|.$$

That is, the Lasso penalty measures the sum of the absolute values of individual parameters, another measure of vector length, which is also known as the l_1 norm. In contrast to the Ridge penalty, the Lasso penalty constraints the coefficient vector into an N-dimensional diamond around the origin, with the constraint boundaries characterized by discontinuities along the parameter axis, which can give rise to the possibility of zero coefficients for certain variables. This feature enables the Lasso to perform both explicit variable selection as well as estimation and shrinkage at the same time.

The variable selection feature of the Lasso is illustrated in panel (b) of Figure 2, where selection region of the model is again represented by the grey area surrounding the origin of the two-dimensional coefficient space. In contrast to the Ridge regression, the Lasso encourages restricted solutions at the corners of its diamond shaped selection boundary, which reside along the axis of the coefficient space, giving rise to sparse solutions. Such a situation is illustrated in the figure, where, presented with the same optimization problem as the Ridge regression, the Lasso sets the coefficient estimate of the first variable to exactly zero, while giving the second variable a non-zero value.

4.2.2.3 Elastic Net

While the ability of the Lasso estimator to impose sparsity on the coefficient space makes it an appealing alternative to the Ridge regression, it is known to struggle in the presence of highly

correlated variables: while the Lasso simply chooses one of the correlated in variables in an indiscriminate manner, discarding the other, the Ridge tends to shrink their coefficients towards each other Hastie, Tibshirani, and Friedman (2009, 663). Furthermore, the variable selection feature of the Lasso is limited by the number of observations: while the Lasso can handle estimation based on predictor sets larger than the number of observations, it is unable to assign non-zero coefficients in excess of the number of observations. To overcome these deficiencies, the Elastic Net was proposed as a hybrid of the Ridge and Lasso approaches by Zou and Hastie (2005). The penalty function of the elastic net can be expressed as a convex combination of the Ridge and Lasso penalties:

$$p(\beta) = (1-\alpha)\sum_{n=1}^{N}\beta_n^2 + \alpha\sum_{n=1}^{N}|\beta_n|.$$

Here, the l_2 penalty of the Ridge averages the coefficients of correlated variables while the l_1 penalty of the Lasso handles variable selection. The parameter $\alpha \in [0,1]$ determines the weight of the respective penalties relative to each other. It is evident that both the Ridge and Lasso regressions can also be thought of as special cases of the Elastic Net, with the former arising from setting $\alpha = 0$ and the latter from setting $\alpha = 1$.

In general, the addition of α represents another parameter, whose value must be decided upon prior to the estimation of the final model. Finding an optimal value with regards to forecasting performance is an empirical matter. In the forthcoming study, the optimal α will be defined for each forecast origin along with λ.

To gain intuition, the nature of the Elastic Net restriction can once again be illustrated in two-dimensional space as in panel (c) of Figure 2: the selection region of the Elastic Net regression assumes the shape of a diamond like the Lasso, but with sides curving outward. Owing to his curvature, the Elastic Net encourages non-zero coefficient solutions more frequently than the Lasso. The extent of the curvature is dictated by the relative weight parameter α in the penalty function: the smaller α is, the more curvature the selection boundary will exhibit, giving rise to less sparsity in the estimates.

4.2.2.4 Adaptive Lasso

The Adaptive Lasso, proposed by Zou (2006), was motivated to provide a shrinkage estimator with the oracle property. While the Lasso has been found to be effective in performing variable selection under certain conditions, its estimation properties are generally lacking: especially

when the true coefficient values of some variables are large relative to others, the Lasso tends to underestimate them by applying excessive amounts of shrinkage, which may lead to poor predictive performance.

The Adaptive Lasso is distinguished from the Lasso by the addition of variable-specific penalty weights in the penalty function, which allow for varying amounts of shrinkage to be applied to individual coefficients:

$$p(\beta) = \sum_{n=1}^{N} w_n |\beta_n|.$$

The purpose of the penalty weight w_n is to impose additional shrinkage on irrelevant variables. Less relevant variables receive a higher additional penalty, which causes their corresponding coefficients estimates to be shrunk more relative to their more relevant peers. The penalty weight can be further broken down as follows:

$$w_n = \frac{1}{\left|\beta_n^{init}\right|^{\gamma}}.$$

Here, ${\beta_n}^{init}$ is a consistent coefficient estimate pertaining to x_n, which has been recovered from some initial regression of the complete model. That is, the larger the initial coefficient is, the more relevant the variable is initially deemed to be, and the smaller the penalty weight is in turn. While Zou (2006) suggests the use of OLS coefficients to obtain initial estimates for the weights, he notes that the assumption of consistency can be relaxed to utilize other estimates without sacrificing the oracle property. This notion becomes especially relevant for cases in which OLS estimation becomes infeasible due to the number of predictors exceeding the number of observations. As an alternative, any of the single stage penalized regressions discussed up to this point, the Ridge, Lasso or Elastic Net, can be used for this purpose in place of the OLS estimates (see, e.g., Medeiros and Mendes 2016; Medeiros and Vasconcelos 2016).

In this study, three variants of the Adaptive Lasso will be considered, with the initial weights determined by way of either OLS, Ridge, or Lasso. Even if the use of the penalized regressions weights was not strictly necessary in this application from the point-of-view of feasibility, the experiment could nonetheless provide meaningful results regarding the effects of the different initial estimates on the performance of the Adaptive Lasso.

The parameter γ can be further used to adjust the magnitude of the additional penalty. Although a value of $\gamma = 1$ is commonly used, it can also be considered as an additional tuning parameter

that can be optimized along with λ in a manner similar to α in the Elastic Net. Given that the forecasting performance of the Adaptive Lasso have been found to be very sensitive to the choice of γ in empirical applications (Medeiros and Vasconcelos 2016), this is the approach that will be followed in this study.

The features of the Adaptive Lasso estimator are illustrated in panel (d) of Figure 2: the selection boundary of the Adaptive Lasso can be presented as an elongated diamond, which distinguishes it from the symmetrical diamond shaped boundary of the plain Lasso. This illustrates the ability of the Adaptive Lasso to discriminate coefficients according to the initial penalty weight. For any given coefficient value, the variable with the larger initial weight will be penalized more relative to the other in order to comply with the penalty restriction. In this case, β_1 has an implicitly larger initial penalty, owing to which its coefficient holds more weight in the penalty function. This is reflected by the corners of the boundary along the vertical axis being closer to the origin relative to the vertical axis. For β_2, on the other hand, the Adaptive Lasso applies less shrinkage than the plain Lasso solution, yielding a larger coefficient estimate.

4.2.2.5 Parameter Tuning

The central model selection issue pertaining to penalized regressions is the choice of hyperparameters, which require optimization prior to the implementation of the methods. While the complexity parameter λ is common to all methods, the Elastic Net and Adaptive Lasso require the additional the choice of the α and γ parameters respectively. Furthermore, in the Adaptive Lasso, λ must be tuned twice in case penalized methods are utilized in the initial stage to obtain the penalty weights. Let us therefore define θ as the vector of tuning parameters pertaining to each method with $\theta = \lambda$ for the Ridge and Lasso, $\theta = (\lambda, \alpha)'$ for the Elastic Net and $\theta = (\lambda, \gamma)'$ for the Adaptive Lasso.

An optimal parameter choice would be characterized as one that minimizes the prediction error of the resulting model. Determining such a value using only past data, however, is not straightforward as basing the choice on the in-sample fit of the model will simply result in the OLS solution with no shrinkage. A common way to circumvent this problem is by using separate parts of the in-sample data for estimation and assessment of the model to obtain optimal parameter values. This approach is generally referred to as *cross-validation*, of which there are different ways to conduct in practice.

In cross-sectional contexts, the most prominent method for parameter tuning in penalized regressions, and machine learning models in general, has traditionally been by using *K-fold* cross-validation (Hastie, Tibshirani, and Friedman 2009, 241). In K-fold cross validation, observations in the sample set are randomly partitioned into K equal length subsets called folds. That is, the observations are mixed so that each individual observation is placed into one of the folds regardless of its original placement among the observations. Following partitioning, an out-of-sample prediction is generated for every observation in the sample set by using a model fitted solely on observations beyond the fold that it itself occupies. Specifically, given an observation in, say, fold k, the prediction is generated by a model, which has been fitted using only observations from the other $K-1$ folds. This procedure is repeated for each observation to obtain a sample of out-of-sample predictions and consequent prediction errors, the size of which equals the number of observations. The whole process is subsequently repeated using different values of the tuning parameters. The parameter value or vector that yields the smallest mean error is chosen for the model.

To illustrate the application of K-fold cross-validation in the present context, consider an individual observation $\{y_{t+h}^h, x_{1,t}, \dots, x_{N,t}\}$ out of a total sample of T observations consisting of predictor variables and the h-step-ahead target variable of the forecast at time t. Following Hastie, Tibshirani, and Friedman (2009, 242), let κ be an indexing function, which assigns each observation to one of the K folds, with $\kappa(t) = k$. Furthermore, let $\hat{f}^{-k}(x_t, \theta)$ be a forecast based on the model fitted using observations contained in the folds *other than* k and a candidate vector of tuning parameters θ. Given a mean squared error loss function, the optimal choice of tuning parameters can be presented as

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} (y_{t+h}^h - \hat{f}^{-\kappa(t)}(x_t, \theta))^2.$$

The main attraction of K-fold cross validation arises from the fact that it enables information in the sample to be used in an efficient manner: each observation in the sample is used in each iteration of the fitting process either to fit the model or to test its accuracy. Running through each of the K folds in turn yields a sample of prediction errors equal to the number of observations in the sample. At the same time, each iteration of the models is fitted using all of the information in the remaining $K-1$ folds, which serves to control for the effects of sampling variation in estimation. Combined, these features should provide for a robust choice of tuning parameters.

The choice of K is a key consideration. In general, K-fold cross-validation can be thought of as producing an estimate of the true out-of-sample error of the model. Increasing K up to the number of observations decreases the bias of the estimate while increasing its variance at the same time, another example of the by-now familiar tradeoff. The choice of K also involves a consideration of computational cost, as a total of K estimations of the model must be conducted for each iteration of the vector of candidate parameters. A common compromise in the literature is to set the number of folds to a relatively low number of five or ten. (Hastie, Tibshirani, and Friedman 2009, 242–243)

The effective use of available data is obviously a favorable proposition in general, and especially so with regards to macroeconomic time series. However, the use of regular K-fold cross-validation is not as straightforward for time series as it is for independently distributed data: due to the serial dependence between consecutive observations in time series, the assumption of independent and identically distributed observations critically underlying the theoretical efficacy of randomized cross-validation is violated. An additional tribulation presents itself in the form of changes occurring in the data generating process of the target variable over time due to structural breaks or trends, for example, which will go unaccounted for if the ordering of the data is mixed indiscriminately. For this reason, regular K-fold cross validation could lead to inferior model selection in time series applications. (Bergmeir and Benítez 2012)

To account for these possibilities, the traditional manner of assessing model adequacy for time series is by respecting the temporal ordering of the data by fitting the candidate models only on a subset of earlier data and testing their performance on a latter part reserved solely for this purpose. While this approach, sometimes referred to as *last block validation*, could be effectively used to solve the issues related to the temporal dimension of the data, this benefit is counterweighted by the fact that both the number of prediction samples that can be produced from a given number of observations as well as the number of observations that can be used to estimate the candidate models at each iteration is inherently limited compared to K-fold cross-validation. This can, in turn, expose model selection to errors arising from sampling variation rather than temporal considerations.

Fortunately, the theoretical challenges related to the use of cross-validation may be of only limited concern in empirical applications, especially when utilizing stationary time series, such as the ones considered in this study. To shed light on these practical implications, Bergmeir and Benítez (2012) employ different model selection procedures, including regular K-fold cross

validation and variants thereof, as well as last block techniques, to optimize various univariate time series models based on both traditional estimation methods and machine learning approaches. Their results indicate that using cross-validation techniques on stationary time series data tends to produce superior outcomes compared to the last block validation methods, which implies that the gains in the efficient use of the data outweigh any potential disadvantages due to temporal dependencies among the observations.

In particular, Bergmeir and Benítez (2012) recommend the use of so-called *blocked* K-fold cross validation in time series applications. In blocked cross-validation, the sample observations are still grouped into K subsets, but, instead of random partitioning, the equal length blocks are formed from consecutive observations. The reasoning is that blocked folds limit the temporal dependencies among the folds. For example, if the observations contain lagged values of the predictor variables, the dependencies among the blocks are limited to the boundaries between the consecutive blocks, where, depending on the number of lags employed, the folds can include the same observations of individual variables. In regular K-fold cross-validation, on the other hand, the dependencies can spread out across a number of individual folds as consecutive observations can end up in any of them.

Beyond the way in which the folds are formed, the model selection procedure of blocked cross-validation is identical to regular K-fold cross-validation: each fold is used for testing and the others for fitting, in turn. Thus, blocked cross-validation does not solve the potential of problems arising from structural changes in the data generating process, as the ordering of the observations is still mixed for the estimation, albeit to a lesser extent. Nonetheless, this should be a minor problem if the data used is truly stationary, which is a notion that Bergmeir and Benítez (2012) stress upon application of the method to time series data.

Based on these considerations, parameter tuning of the penalized models in this study will be performed by way of blocked K-fold cross-validation. Furthermore, cross-validation will be applied using ten folds, which is a common choice in the applied literature. While the folds are by default set to equal length, in the case of uneven division, the excess observations will be distributed one at a time to successive folds, starting from the first one.

4.2.3 Benchmark Models

Forecasts are bound to contain some amount of error and quantifying this error is essential to assessing their performance. To provide context for the measured errors, it is customary to

employ benchmark models in forecasting studies, against which the proposed forecasting methods are compared to. The benchmarks are usually based on simple models that represent a bare minimum of forecasting performance that the more elaborate models should exceed in order for their application to be justified. Here, two benchmarks often utilized in the macroeconomic forecasting literature will be utilized: the random walk (RW) and univariate autoregressive (AR) models, both of which rely solely on the past observations of the target variable to generate predictions.

The RW forecast offers the simplest possible starting point. The intuition behind the RW forecast is based on the notion that the target series is in fact produced by an integrated time series process whose next value equals the sum of the last value and a mean zero random error. As such, the best one can do to forecast the series is to assume that its next observations will be equal to its last one observed. That is, for any given horizon, the forecast is simply

$$\hat{y}^{h}{}_{t+h} = y_t.$$

For obvious reasons, the random walk forecast is commonly referred to as the naïve forecast.

In terms of complexity, the AR model is a step beyond the RW in that it attempts to model the dynamic process of the target variable based on a number of previous values of the series to produce forecasts. Regarding multistep forecasts, the AR model allows for two possibilities: either direct or iterative forecasts. The direct approach is the one used for the main models as outlined above, where each h-step-ahead value of the target is regressed individually on the observed past values. In the iterative approach, the one-step-ahead forecasts are substituted into the regression to move the horizon forward.

The relative performance of the direct and indirect methods in empirical settings have been studied by Marcellino, Stock, and Watson (2006), who find that the iterative approach generally yields the most accurate results in macroeconomic forecasting. Forecasting price series, however, appears to represent an exception to this rule, with direct forecasts proving superior over the iterative approach in the short term and providing only slightly inferior results in the longer term. Due to this, and to allow straightforward comparison with the high-dimensional models, the direct approach is adopted for the AR benchmark here as well. Therefore, the forecast equation assumes the following form:

$$\hat{y}^{h}_{t+h} = \hat{\mu}^{h} + \sum_{p=1}^{P} \hat{\alpha}_p y_{t+1-p}.$$

Regarding model specification, the AR model requires the selection of appropriate lag length for the target variable, which is commonly obtained using information criteria. Like the AR component in the dynamic factor model, lag selection will be guided by use of the AIC at each forecast origin. Estimation of the AR model is carried out using OLS.

5 Data and Variables

This section presents the data used in the simulated forecasting experiment. The first subsection provides an overview of the predictors used, while the second one discusses the transformations applied to the predictors in order to induce stationarity necessary for the application of the forecasting models.

5.1 Predictors

Producer prices are bound to be affected by a variety of factors at any given point in time. For example, business cycle dynamics are likely play a role via shocks to aggregate demand, while on the supply side, shocks to the costs of inputs, be it raw or intermediate materials, energy, or labor, are bound to affect producer prices.

Some of these factors are more difficult to measure and quantify in practice than others. The former category includes features of the market, such as the extent of competition and demand for individual products. While not directly observable, various proxies could be used to encompass these features of the data generating process. However, even if relevant data were to exist, the issue of timing presents a considerable challenge to the application of such data in a near-term forecast setting. For example, some real variables, such as the volume index of industrial output, which is a measure of real industrial production activity in Finland reported at a monthly frequency, could provide a compelling means to measure the demand side factors affecting producer prices in the short run, and such measures have been commonly used in forecasting studies. In reality, however, even though the volume index of industrial output is a fast indicator of business cycle conditions, it still lags behind the PPI by nearly one full month. Thus, to forecast producer prices one month ahead, one could only use the data on the volume from two months before. Considering that prices are generally faster to adjust than real quantities, it is unlikely that a shock to the real index would affect prices only after a period of two months. The issue raised above applies to a range of potential candidate predictors, even more severely so to many others, as the PPI itself is a fast indicator to begin with. Thus, to respect the pretense

of realism, care must be taken in the choice of predictors for the forecasting experiment. Here, the choice of data is based on a cut-off point relative to the publication of the PPI. The idea is that only variables whose latest observations become available reasonably early with respect to the PPI could be used to produce forecasts of the following months index.

Despite imposing a significant restriction, there is still data that manages to fulfil these criteria. Most prominently, various market data are usually quoted at a daily frequencies and monthly aggregates thereof are therefore expeditiously available following the end of the respective month. The downside of such market data is that it mostly limited to nominal quantities of one form or another. Regardless, it is plausible that such data could provide early indication for developments in producer prices. For example, commodities prices could provide an indicator for supply side effects, whereas stock market indices could signal broader changes in the market environment. Whether the relationships between these variables and the PPI exist and are consistent enough to be captured by the methods considered here is the empirical question this study attempts to answer.

With the considerations above accounted for, it is time to present an overview of the data employed in this study. All series utilized are reported at a monthly frequency and span the time frame from January 2000 to November 2021. A detailed list of all individual variables is provided in Table 4 of the appendix.

Starting with the PPI itself, in addition to the past values of the main index, the set of predictors includes derivative price indices based on the Main Industrial Groupings (MIG) classification. Using the MIG classification, products are divided into five categories according to their end use: intermediate goods, capital goods, consumer durables, consumer non-durables and energy goods. A similar treatment is used for the Import Price Index (IPI), which is published in conjunction with the PPI. While IPI indices are reported for each of the five MIG categories, the PPI excludes the energy category due to its negligible role in Finnish output. In addition to the producer price indices, data sourced from Statistics Finland includes the Finnish market spot price for electricity.

International commodities prices, sourced from the Primary Commodity Prices database of the International Monetary Fund (IMF), represent the largest single group of predictors, comprising a total of 54 individual series. These commodities include raw materials, such as various metals and energy products as well as different foodstuffs. The IMF dataset contains prices of

commodities reported from different markets around the world and often those marketplaces are far from the Finnish producers relevant to our forecasting exercise. While this geographical disparity could be a cause of concern, these concerns should be somewhat alleviated by the nature of commodities: by definition, commodities are goods that are universally used, easy to transport and of standardized quality. Thus, owing to arbitrage, the prices reported in one market should reflect global prices to a reasonable degree. While the original commodities prices are reported in terms of US dollars, the series have been transformed into euros using contemporary exchange rates to reflect prices potentially faced by Finnish producers.

The next category of predictors contains stock market indices. These include data from the OMX Helsinki, Euro Stoxx and Standard and Poor's indices, which track the market prices of stocks of Finnish, European and American publicly traded companies respectively. In addition to the main indices, the predictors include industry specific subindices for the Finnish and European markets, which encompass the stock prices of companies active in different fields such as manufacturing and finance. All data are monthly averages of daily quotations and have been sourced from the Bank of Finland (BoF) and European Central Bank (ECB).

While a majority of Finnish foreign trade nowadays occurs within the common currency area, a number of exchange rates of additional prominent trade partners are included in the set of candidate predictors. These include the British pound, the Swedish krone, the Norwegian krone, the Russian ruble, the US dollar, the Japanese yen, and the Chinese yuan. The exchange rates are monthly averages expressed in terms of the foreign currency price of the euro and sourced from the ECB.

Various interest rates are also considered. These include the Euribor rates at maturities ranging from one to twelve months as well as a selection of government bond yields. The latter group includes yields for Finnish and US government bonds as well as the euro area benchmark bonds. All rates are sourced from the ECB, with the exception of the Finnish bond yields, which are provided by the BoF.

The final category of predictors consists of business climate indicators published by the European Commission. These indicators are based on surveys conducted at a monthly frequency on a sample of firms in member countries of the European Union. The questions in the survey concern different aspects of the operational environment of the firms, both present and future, including current stocks of finished products and expectations regarding developments in sales

prices of their products in the near future. Individual answers are aggregated to form numerical indicators for each specific question and country. Here, the results concerning Finnish companies in particular as well as responses averaged over all euro area countries are considered.

5.2 Transformations

Stationarity, precisely weak or covariance stationarity in the sense that the mean and autocovariance of a given series remain stable over time, is a perquisite for the application of most econometric time series models, including the ones considered here. The stationarity of both target and predictor series was assessed primarily by means of the augmented Dickey-Fuller (ADF) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests, as well as visual inspection of the series as necessary. Each predictor series was tested in original levels, first differences and second differences, in turn, in order to determine the suitable level of differentiation required to induce stationarity. Series featuring only non-negative values, that is, all series except for the interest rates and business climate indicators, were additionally tested in log-levels and the respective log-differences. Both tests were specified with constant terms and, to ensure robustness, varying numbers of lags in the test regression, with up to 16 lags considered at most for both tests. The detailed test results are available upon request from the author.

With few exceptions, the ADF null hypothesis of unit root could not be rejected for the untransformed or log-transformed predictor series with the results corroborated by the KPPS test, which indicates the rejection of the null hypothesis of stationarity for most of the same series. The most notable exception is the group of business climate indicators, which are scaled to vary between a fixed interval by design.

First differencing either the raw or log-transformed series changes the situation, with both tests indicating stationarity for a vast majority of the predictors: the ADF test rejects the null hypothesis of unit root for nearly all predictors decisively at the usual confidence levels, with the exception of the PPI series for capital goods and consumer durables. However, the results of the KPSS test provide somewhat contradictory results, with the test indicating a rejection of the stationary null for some additional series as well. These include both the PPI and IPI series for capital goods and consumer durables as well as some stock market indices. Visual inspection of these particular series in first-differenced form reveals that all of them generally revolve consistently around their respective long-run means, but occasionally exhibit large but short-lived deviations, which may explain the results of the KPSS test. Furthermore, plotting the

autocorrelation functions of the series indicates little persistence in the series, which provides further evidence in favor of the assumption of stationarity.

While the results of both tests indicate that a second round of differencing unambiguously takes care of the problem of non-stationarity for the remaining predictor series as well, to retain commonality among the different price and stock market index series and given the generally inconclusive evidence in favor of non-stationarity, all of the aforementioned predictors were retained in first-differenced form for the forthcoming analysis. To summarize, all prices, price indices and stock market indices were transformed into first log-differences, and interest rates into first differences, while the business climate indicators were left without transformation.

The last order of business in the present section pertains to the definition of the target variable. The approach to multistep-ahead targets adopted here is based on the Stock and Watson (2002b), which has been utilized by a number of subsequent studies (eg., Bai and Ng 2008; Kim and Swanson 2014). Given the results discussed above, that is, treating the PPI as a series integrated of order one, the *h*-step-ahead target for the direct forecasts is defined as follows:

$$y_{t+h}^{h} = h^{-1}(\log(PPI_{t+h}) - \log(PPI_t)) = h^{-1}\log\left(\frac{PPI_{t+h}}{PPI_t}\right).$$

That is, instead of predicting isolated month-on-month changes multiple periods ahead, the multiperiod dependent variable is the average of cumulative changes of the logarithm of the PPI during the *h* periods following *t*. An advantage of this approach is that the multiperiod forecast can be straightforwardly transformed back into levels without needing to explicitly construct the forecasts of the intervening periods. On the other hand, a side-effect of these transformation is that the each multiperiod target variable effectively represent a unique series, as the differences are calculated with respect to different observations depending on the horizon considered. This is illustrated in Figure 3, which exhibits how averaging over multiple periods of monthly changes smooths the respective series relative to the one-month series.

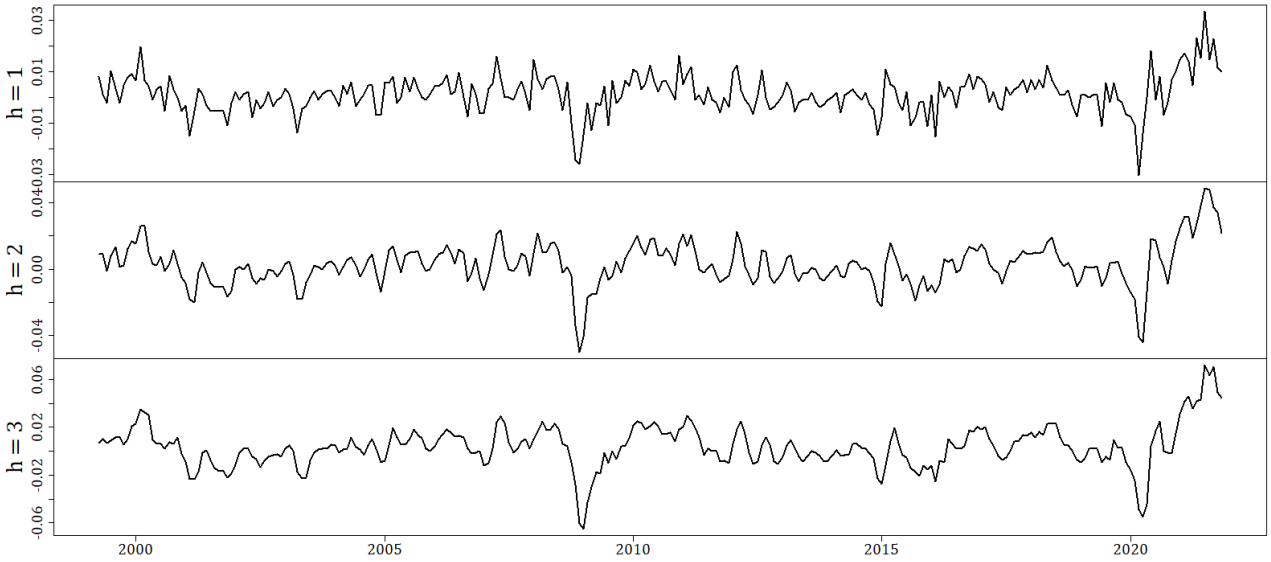

Figure 3. *Evolution of the target series for different forecast horizons.*

Despite these changes, ADF and KPSS tests indicate that the series pertaining to the longer horizons retain the stationarity of the month-on-month first-differenced PPI series. Furthermore, while the target variable depends on the scope of the forecast, the predictor variables are shared commonly by all horizons. This applies also to the lagged observations of the PPI itself, which are always included as month-on-month differences when used as predictors.

6 Empirical Results and Discussion

This section presents and discusses the results of the comparative out-of-sample forecasting experiment on the Finnish PPI. The section is further divided into three subsections corresponding to separate experiments. The first subsection details the results of outright forecast simulation, in which only data pertaining to the previous period are used to predict the forthcoming values of the PPI at the one-, two- and three-month horizons. In the second subsection, the forecasting performance of the models over time and during alternative out-of-sample periods will be examined. In the third subsection, a separate experiment will be carried out to assess whether the advantage in publication lag of certain market variables present among the predictors could be leveraged by using them to predict the concurrent values of the PPI instead of future ones. To reflect the use of contemporary data for prediction, the last experiment will be referred to as the *nowcasting experiment*.

6.1 Forecasting Experiment

The results of the out-of-sample forecasting experiment over the main sample period, ranging from December 2011 to November 2021, are summarized in Table 1. For each model, the mean absolute errors and root mean squared errors (MAE and RMSE, respectively) are reported on the columns for each of the three forecast horizons. Specifically, each respective error is reported as a *ratio* relative to the aggregate error of the main benchmark, the AR model, to enable straightforward comparisons. With the benchmark error occupying the denominator of the relative error, a value of less than one indicates that the forecast produced by the competing model is, on average, more accurate than the AR model. A value of over one implies the opposite. For the AR model itself, the actual aggregate errors are reported in the table instead. Furthermore, the best results in terms of both error measures and for each forecast horizon are reported in bold.

In addition to the relative errors, Table 1 reports the p-value of the Diebold-Mariano test, again relative to the AR benchmark for each model. The test can be specified with either one-sided or two-sided alternative hypothesis. In this case, the one-sided alternative is used, as the main question of interest is specifically whether or not the competing models are useful in yielding improvements in forecasting performance. Thus, the alternative hypothesis states that the forecasts produced by the high-dimensional models are more accurate than those produced by the benchmark.

The first two rows of table are occupied by the benchmarks, the AR and RW models respectively. The benchmarks are followed by the dynamic factor models (DFM). In addition to the factor model, in which the optimal number of factors is determined individually for each forecast origin by way of the eigenvalue ratio test, models with fixed numbers of factors are included in the experiment as well. While the model with dynamic factor selection is the one with most practical relevance, as the optimal number of factors is otherwise difficult to determine in advance, the fixed factor variations are included in order to assess the adequacy of the factor selection methodology as well as to provide a broader perspective on the potential of the factor approach. To this end, the results for a total of five fixed factor models, with the number of factors ranging from one to nine, will be presented. Experiments with further numbers of factors were also conducted, but these provided no additional benefit and, thus, their results will not be presented. The parenthesized expression in the model label indicates whether the choice

of static factors included is fixed ($R = 1,3,5,7,9$) or determined by the Ahn-Horenstein eigenvalue ration test (AH). Maximum lag-order of the AR component is set to six.

Next in line are the factor models estimated using targeted predictors (T-DFM), where the static factors have been extracted from a subset of 40 predictors variables chosen by an initial Lasso estimator on the target variable at each forecast horizon. Apart from the different set of predictors, the application of the models is identical to the regular factor models. Furthermore, results are reported similarly for models featuring both fixed and dynamically selected numbers of factors.

The final group of models encompasses the penalized methods: the Ridge, Lasso, Elastic Net and Adaptive Lasso (Ada-Lasso). The Adaptive Lasso is, furthermore, present in three variations, which are distinguished by the method used to determine the penalty weights in the regression: as indicated by the parenthesized expression in the label, either OLS, Ridge or Lasso regression is used for this purpose.

For the Elastic Net, the parameter α is optimized at each forecast origin along with lambda over a sequence of five candidate values ($\alpha \in \{0.1,0.3,0.5,0.7,0.9\}$). Similarly, the weight parameter γ in the Adaptive Lasso models is optimized over three candidate values ($\gamma \in \{0.5,1,2\}$). In keeping with the factor models, each penalized model includes six lags of the target as potential predictors.

Overall, the results indicate that, while some of the high-dimensional models would have, on average, offered improvements over the AR benchmark, the aggregate forecast errors are generally very close to the benchmark and any gains in performance arising from these models tend to be very modest. This is also reflected in the fact that none of these improvements, at any horizon, are statistically significant at the common confidence levels in terms of the Diebold-Mariano test, as indicated by the p-values reported in the table.

Table 1

Results of the out-of-sample forecasting experiment for 2011M12–2021M11.

	h = 1			h = 2			h = 3		
Model	MAE	RMSE	p-value	MAE	RMSE	p-value	MAE	RMSE	p-value
AR	*0.00493*	*0.00694*		*0.00445*	***0.00611***		*0.00441*	*0.00604*	
RW	1.222	1.145	0.983	1.224	1.183	0.949	1.242	1.249	0.956
DFM (AH)	0.968	0.980	0.260	0.978	1.002	0.529	0.976	1.013	0.630
DFM (R=1)	0.981	**0.975**	0.179	0.977	1.007	0.586	0.979	1.018	0.658
DFM (R=3)	0.968	0.980	0.260	0.978	1.002	0.529	0.976	1.013	0.630
DFM (R=5)	**0.966**	0.983	0.297	**0.963**	1.004	0.540	0.983	1.023	0.711
DFM (R=7)	0.971	0.987	0.342	0.970	1.017	0.670	0.986	1.024	0.728
DFM (R=9)	0.981	0.994	0.437	0.966	1.023	0.719	0.967	1.016	0.670
T-DFM (AH)	1.014	1.003	0.541	0.988	1.009	0.611	0.979	1.044	0.795
T-DFM (R=1)	1.019	1.005	0.557	0.994	1.010	0.633	0.979	1.034	0.736
T-DFM (R=3)	0.971	0.993	0.417	0.978	1.012	0.656	0.977	1.039	0.787
T-DFM (R=5)	0.988	1.005	0.557	0.976	1.028	0.750	0.981	1.042	0.782
T-DFM (R=7)	0.999	1.010	0.610	0.978	1.030	0.730	0.988	1.040	0.769
T-DFM (R=9)	1.016	1.007	0.575	1.008	1.049	0.829	1.000	1.046	0.814
Ridge	0.977	1.009	0.606	1.018	1.052	0.891	**0.972**	1.023	0.716
Lasso	0.995	1.004	0.549	1.032	1.045	0.848	0.980	0.994	0.427
Elastic Net	0.991	1.005	0.570	1.015	1.032	0.802	0.980	1.009	0.604
Ada-Lasso (OLS)	1.059	1.031	0.779	1.147	1.129	0.984	1.210	1.160	0.983
Ada-Lasso (Ridge)	1.003	1.005	0.568	1.052	1.059	0.917	0.991	0.985	0.321
Ada-Lasso (Lasso)	1.028	1.013	0.665	1.046	1.033	0.758	0.978	**0.973**	0.222

Note: Mean errors are reported relative to the AR forecast. P-value reported for the Diebold-Mariano test with alternative hypothesis of superior forecast accuracy with respect to the AR forecast.

At the one-month horizon, the plain dynamic factor models stand out from the set of candidates: all of these models manage to produce lower mean errors than the AR model. In particular, the models with fixed numbers of one and five factors prevail over all competing models in terms of MAE and RMSE, respectively. However, even for these models, the improvements in forecasting accuracy are very marginal: the MAE of the five-factor model, for example, is only some 3.4 percent lower than that of the simple benchmark, whereas the RMSE of the one-factor model is, equivalently, no more than 2.5 percent smaller. These observations extend to further horizons as well, with similarly modest improvements over the AR benchmarks in terms of MAE. Moreover, in terms of RMSE, even this meager advantage is all but lost, with the sample errors equal to or even slightly in excess of the benchmark.

Regarding model selection, the Ahn-Horenstein test sets the number of factors consistently to three for each of the forecast origins over the whole out-of-sample period, which explains the ostensibly identical results between the dynamically optimized model and the corresponding fixed factor model. While the relatively favorable results of this model would imply that the test

does a good job at estimating the adequate number of factors, the differences among the factor models are overall not profound enough to make such conclusions, as none of them are far from each other in terms of accuracy.

In general, the forecasts of the dynamic factor models tend to follow the AR forecast very closely. This is illustrated in Figure 4 below, in which the series of forecasts from the five-factor model and the AR model are plotted against the realized values of the first-differenced log-PPI series for the one- and three-month horizons. As the results are qualitatively similar for the other factor models as well as the second forecast horizon, the plots for these will be omitted for the sake of brevity.

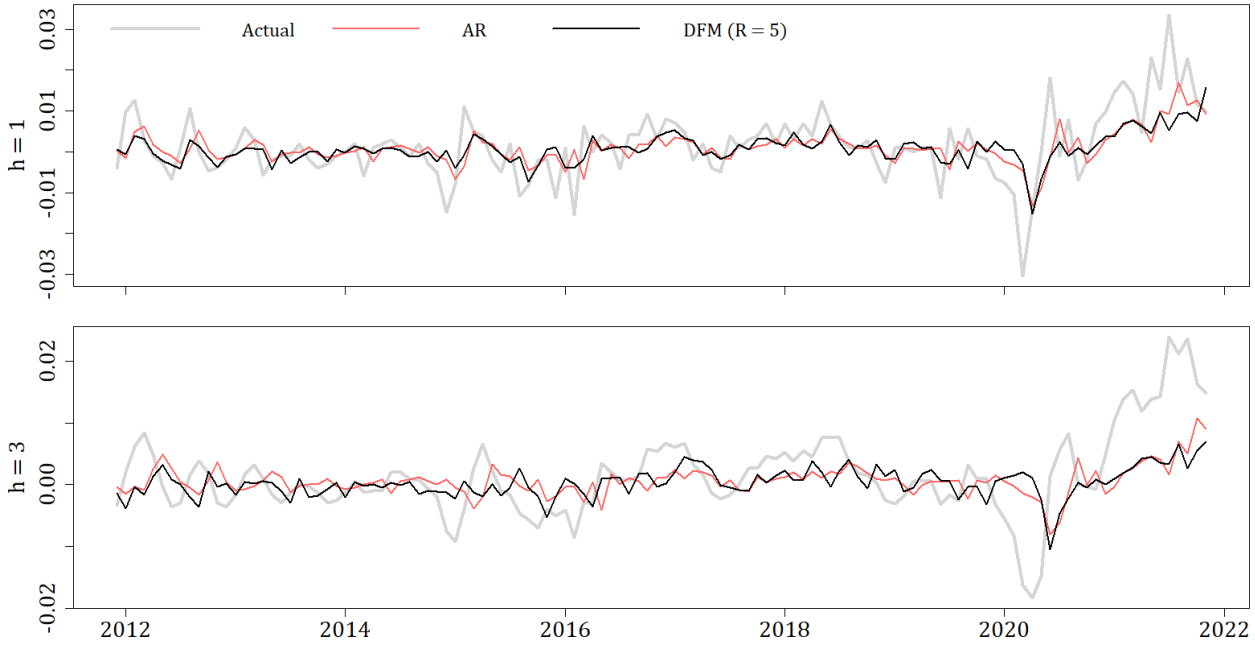

Figure 4. *Comparison of the AR and five-factor DFM forecasts against realized PPI at the one- and three-month horizons.*

While both forecasts roughly follow the realized target series, they exhibit difficulty in accounting for large unforeseen shocks in the series at both horizons. This is especially evident in the latter part of the sample, which is characterized by particularly large deviations from the long-run mean in the post 2020 period. This is a situation, where one might expect a model relying on information other than the history of the target to prevail, but this does not seem to be the case. Rather, both factor and AR forecasts exhibit behavior that can be characterized as reactive rather than proactive. That is, preceding shocks appear to be extrapolated into following forecasts with a lag, which implies that the additional factors, or the variables that they are based on, have little predictive power over the PPI even in the short run. The reactive behavior is

further emphasized at longer horizons, where the effects of the large shocks appear in the forecasts with gradually increasing delays. The three-month horizon is also characterized by a noticeably widening gap between the forecasts and the realized values, which is especially evident towards the end of the sample.

The targeted predictors appear to provide little assistance to the factor models. On the contrary, the factor models appear to benefit more from the quantity rather than the quality of the data, as evidenced by the superior results of the regular models, although this is most noticeable at the one-month horizon. The earlier observations regarding the model selection capacity of the Ahn-Horenstein test carry over to this case as well, with the resulting model yielding mediocre, but not considerably worse results on average compared to the fixed factor alternatives.

Turning our attention to the penalized methods, the Ridge, Lasso and Elastic Net models manage to beat the AR forecast at the one-month-horizon in terms of MAE, albeit with even smaller margins than the factor models. The Ridge forecast achieves accuracy closest to the latter with a meager 2.3 percent improvement over the benchmark, but others fare even worse, with each of the Adaptive Lasso models providing outright inferior performance. Out of the latter, the model with OLS-derived penalty weights is, by far, the most unsuccessful of all the high-dimensional models across all horizons. In terms of RMSE, none of the models manage to beat the AR at the one- or two-month horizons. The situation is curiously reversed for the three-month horizon, with the Adaptive Lasso model achieving best RMSE accuracy out of all models, although marginally so.

Overall, despite somewhat inferior average accuracy, the performance of the penalized models is not considerably different from the factor models. To illustrate, Figure 5 plots the MAE best penalized forecast, the Ridge, against the realized values of the PPI and the AR benchmark at the one- and three-month horizons. As with the various factor models, the penalized forecasts exhibit largely similar behavior among themselves and for this reason it suffices to examine only the Ridge forecast in more detail. The plot reveals a familiar pattern: like the factor forecast, the Ridge forecast tends to follow the AR forecast very closely. This notion extends to the fact that the penalized forecast shares the deficiency of the factor models, in that they appear to be no better equipped to capture large changes in producer price inflation in advance.

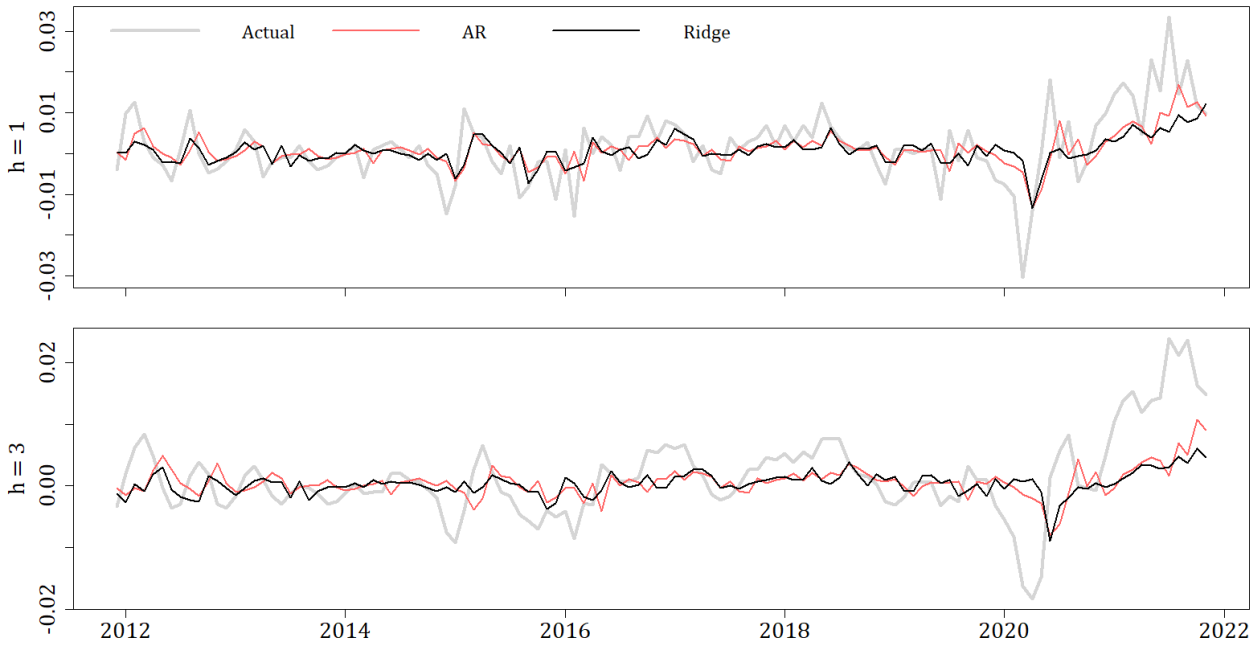

Figure 5. *Comparison of the AR and Ridge forecasts against the realized PPI at the one- and three-month horizons.*

Based on the results of the out-of-sample experiment, it can, by now, be stated with a high level of confidence that the added complexity involved in applying the high-dimensional models cannot be justified in terms of forecasting performance. While some average improvements are observed, they are not substantial enough to be meaningful in practice. Furthermore, they are not substantial enough to rule out the possibility of pure chance as the cause of the results, as implied by the p-values of the Diebold-Mariano tests reported in Table 1, which indicate the failure to reject the null hypothesis of equal or worse forecasting performance on the part of the high-dimensional models.

That being said, the results are consistent with earlier empirical experiments in forecasting the PPI using high-dimensional methods. The large-scale forecasting experiment of Kim and Swanson (2014) includes the US PPI among other target variables. Moreover, their candidate methods include most of the models considered in this study, alongside several others. Their results indicate that the high-dimensional methods tend to generally produce results very close to the AR model, with accuracy in terms of MSE usually no more than a few percent away from the benchmark, either above or below, in the short term. Their results also indicate similar relative performance between the penalized and factor forecasts, with the factor models generally providing marginally superior accuracy. These results are echoed by Smeekes and Wijler (2018), who also forecast the US PPI among other variables using different variations of dynamic factor models and penalized regressions. Their results exhibit similarly small advantages

as well as disadvantages for the high-dimensional models. Furthermore, the improvements encountered in neither study are, apart from few exceptions, statistically significant despite the relatively longer out-of-sample periods utilized in these experiments. Thus, these results provide some context to the ones presented here, as they give testimony to the general difficulty of forecasting the PPI.

The most obvious candidate to explain the lackluster performance of the high-dimensional models is that the information contained in the set of predictors is simply not informative regarding the future values of the PPI. Another possibility is that forecasting performance may be inhibited by structural instability in the target series. That is, the forecasting models estimated on a sample of past data may become detached from the future values of the target variable due to shifts in the underlying data generating process. While the overall results of the experiment, as reported in Table 1, represent only a single summary of average forecasting performance over a relatively long timespan, it is possible that the performance of individual models may have changed over time due to such changes in the forecasting environment. In particular, the onset of the global pandemic, which caused a shock in the global markets for goods by abruptly shifting demand and disrupting established production chains, is not unlikely to represent such a structural shock, which may have broken down relationships between the target and predictors that were prevalent in the past.

The resilience of high-dimensional forecasting methods to the pandemic shock has been recently studied by Goulet Coulombe, Marcellino, and Stevanović (2021). With an emphasis on machine learning methods, they examine the performance of different approaches in forecasting a selection of UK macroeconomic variables, including the local PPI, during different subperiods ranging from 2008 up to late 2020. They find that the pandemic had a considerable detrimental effect on forecasting accuracy for a majority of the models and target variables, and that different models tend to prevail depending on the time-period considered. For the PPI specifically, they find that the Elastic Net and Lasso tend to generally work well at the one-month-horizon during the pre-pandemic period, but their relative accuracy deteriorates during the pandemic. On the other hand, a variant of the nonlinear Random Forest model provides considerable improvements over the AR benchmark during the latter period, but this advantage is counterweighted by considerably inferior performance during the earlier period.

Given the general notion of time-dependence in forecasting performance, the accuracy of the Finnish PPI forecasts over time will be examined more closely next.

6.2 Forecasting Accuracy Over Time

Figure 6 plots the evolution of forecast errors for the AR benchmark and the five-factor dynamic factor model and the Ridge regressions, the best performers of their respective classes, over the main out-of-sample period. Results are reported solely for the one-month horizon, as this is where the forecasts appear to yield the most benefit in absolute terms. The errors are presented in terms of RMSE and MAE calculated over a rolling window of 12 months for illustrative purposes. Due to this, the graphs exhibit some occasionally abrupt ebbs and flows, which are caused by large individual errors entering and leaving the rolling window. Despite this, the plots provide a general idea of the systematic variation in the accuracy exhibited by the respective forecasts over time.

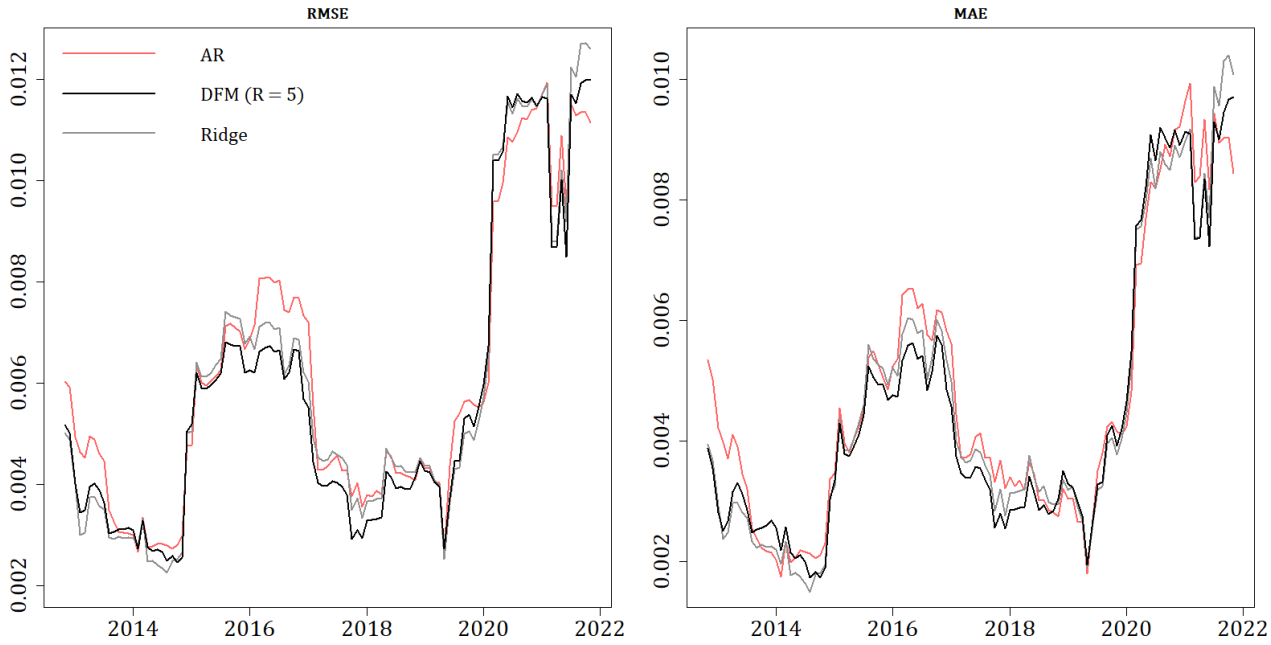

Figure 6. *Evolution of mean errors calculated over a 12-month rolling window for select forecasts at the one-month horizon.*

The figure reveals some interesting insights. First of all, the mean errors exhibit considerable variation over time. For example, the period between late 2014 and early 2017 is distinguished by particularly large errors, while during the three-year period that follows, the errors settle to a level of some one third lower than before. After phase of relative calmness, the errors hike to unprecedented levels following early 2020, nearly tripling in magnitude. While the reason for the earlier incline in errors is ambiguous, the latest hike clearly coincides with the beginning of the COVID-19 pandemic.

Second, while the errors of different forecasts tend to generally move closely together, there is an inkling of heterogeneity in the responses over time. Especially during the earlier and middle parts of the out-of-sample period, the high-dimensional models appear to have a slight edge on the benchmark in terms accuracy, whereas the situation is somewhat reversed towards the end of the period.

To illustrate this further, Figure 7 plots the evolution of the cumulative errors of the respective models *relative* to the benchmark. That is, at each point in time, the vertical axis measures the relative mean errors for each model accumulated up to a given point in time, with a value of less than one indicating superior accuracy with respect to the benchmark as in Table 1. It follows that the rightmost points in the plots correspond directly to the values reported in the main results. On the other hand, earlier points correspond to a situation where the out-of-sample period had been cut short of its full length. The plot excludes cumulative errors prior to 2014, as they exhibit considerable volatility owing to the small number of accumulated observations.

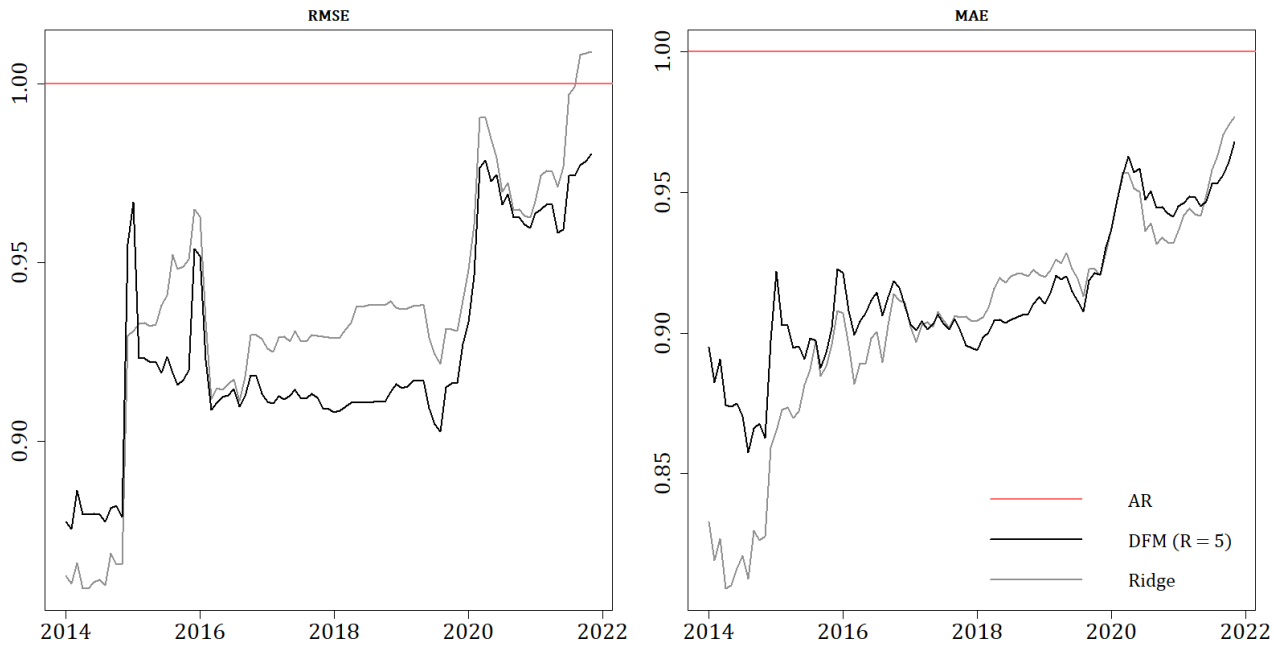

Figure 7. *Cumulative errors of select high-dimensional forecasts relative to the AR forecast at the one-month horizon.*

Figure 7 offers a contrast to the overall results regarding the relative performance of the high-dimensional models. For most of the early out-of-sample period the advantage of the high-dimensional models is far greater than implied by the ultimate results. That is, the external predictors appear to offer a more meaningful advantage during the earlier period. It is only towards the very end of the out-of-sample period that the benchmark begins to catch up with the

high-dimensional models. This is consistent with the observation of Goulet Coulombe, Marcellino, and Stevanović (2021), who attest to the relative disadvantage of high-dimensional linear models following the beginning of the pandemic.

While the historical perspective presents the high-dimensional models in a slightly more favorable light than before, they should not be taken as particular evidence in contradiction to the overall results discussed earlier. The results of Table 1 still stand as resilience to structural shocks is a feature that an adequate forecasting model should possess. Rather, the volatility of the cumulative errors should be interpreted as a cautionary example of the limitations involved in using the out-of-sample experimental procedure to determine the ranking or absolute accuracy of different forecasting models. For instance, had the present analysis been conducted, say, two years earlier, using data up to November 2019 instead of November 2021, the results of the forecasting experiment would have appeared considerably more favorable to the high-dimensional models. In that case, average gains in forecasting accuracy of the five-factor model relative to the AR model would have been no less than 9.3 and 9.6 percent in terms of MAE and RMSE, respectively.

Detailed out-of-sample results for all models during the sub-periods both preceding and following the onset of the pandemic are reported in Table 2 for the one-month-ahead forecasts. The results reaffirm the notion that the high-dimensional models are at a considerable disadvantage during the pandemic, while exhibiting far greater, even statistically significant performance gains with respect to the AR between December 2011 and November 2019.

While even these gains would have hardly been enough to justify the practical utility of the high-dimensional models, they would have implied considerable gains in forecasting accuracy in comparison to the earlier studies on forecasting the US PPI, which were discussed above. It is however noteworthy, that while those results are not burdened by the recent pandemic, they encompass another event with potentially devastating effects for forecasting performance, the global financial crisis of 2008, which the out-of-sample period of our main experiment excludes. Owing to this, the comparison above may not be a fair one.

To provide a more meaningful comparison to the previous literature as well as to shed further light on effects of the global economic calamity on forecasting performance, the out-of-sample experiment was repeated for an alternative time-period. To include the abnormal negative shock in producer prices caused by the climax of the financial crisis in late 2008 and to exclude

the volatility caused by the pandemic, the sample of forecasts for this experiment considers the period ranging from April 2008 to November 2019 for a total of 140 out-of-sample periods. Apart from the time period, the forecasting experiment is carried out identically to the main experiment, albeit only for the one-month horizon. All previous models are included except for the Adaptive Lasso with OLS penalty weights, whose estimation becomes infeasible owing to the low number of observations available for the earliest out-of-sample forecasts. By extension, the following results also serve to showcase the performance of the models in a situation where the number of variables exceeds the number of observations.

Table 2

Results of forecasting experiments for alternative out-of-sample periods at the one-month horizon.

	2008M04–2019M11			2011M12–2019M11			2019M12–2021M11		
Model	MAE	RMSE	p-value	MAE	RMSE	p-value	MAE	RMSE	p-value
AR	*0.00453*	*0.00609*		*0.00396*	*0.00533*		***0.00881***	***0.01128***	
RW	1.214	1.170	0.997	1.268	1.238	0.999	1.138	1.054	0.679
DFM (AH)	0.949	0.918	0.033	0.921	0.916	0.032	1.052	1.035	0.796
DFM (R=1)	0.944	**0.901**	0.023	0.932	0.923	0.035	1.068	1.019	0.701
DFM (R=3)	0.955	0.925	0.051	0.921	0.916	0.032	1.052	1.035	0.796
DFM (R=5)	0.952	0.925	0.047	**0.907**	**0.904**	0.019	1.072	1.049	0.859
DFM (R=7)	0.954	0.946	0.122	**0.907**	0.908	0.029	1.085	1.052	0.881
DFM (R=9)	0.947	0.957	0.233	0.910	0.917	0.052	1.108	1.059	0.868
T-DFM (AH)	0.963	0.923	0.052	0.965	0.939	0.122	1.103	1.057	0.918
T-DFM (R=1)	0.959	0.920	0.045	0.966	0.940	0.125	1.114	1.059	0.923
T-DFM (R=3)	0.924	0.909	0.031	0.917	0.925	0.070	1.068	1.050	0.860
T-DFM (R=5)	0.928	0.910	0.031	0.921	0.928	0.072	1.109	1.069	0.906
T-DFM (R=7)	0.935	0.917	0.061	0.929	0.931	0.081	1.124	1.076	0.928
T-DFM (R=9)	0.963	0.935	0.101	0.983	0.954	0.193	1.075	1.052	0.830
Ridge	**0.921**	0.912	0.003	0.920	0.931	0.041	1.078	1.074	0.921
Lasso	0.940	0.925	0.010	0.956	0.950	0.112	1.067	1.049	0.885
Elastic Net	0.934	0.925	0.008	0.952	0.952	0.115	1.062	1.051	0.871
Ada-Lasso (Ridge)	0.936	0.916	0.039	0.976	0.973	0.282	1.053	1.033	0.802
Ada-Lasso (Lasso)	0.969	0.939	0.134	1.010	0.991	0.430	1.060	1.033	0.801

Note: See notes for Table 1.

The results for this alternative forecast simulation are also reported in Table 2. Despite being burdened by the financial crisis, the accuracy of all models is considerably improved relative to the benchmark in terms of MAE and especially in terms of RMSE, with all models yielding superior results. This applies especially to the penalized models, which provide accuracy on par or in excess of the factor models, although the differences are still rather marginal. Furthermore,

the targeted predictors appear to be somewhat more helpful during this time period. Most of the improvements are also statistically significant relative to the AR benchmark.

The differences are even more striking when compared to the earlier forecasting results regarding the US PPI. At the one-month horizon, the greatest improvements relative to the AR benchmark reported by Kim and Swanson (2014) and Smeekes and Wijler (2018) are equal to around 9 and 7 percent respectively in terms of MSE. On the other hand, the RMSE best model here, the fixed single-factor model, yields an improvement of no less than 18.8 percent once converted to MSE.

Although one should exercise caution in comparing these results, as both of the aforementioned studies consider very different out-of-sample periods, the improvements here are, nonetheless, considerable. In this sense, a more direct comparison is afforded by Goulet Coulombe et al. (2021) who also consider the specific timespan between 2008 and 2019 as one of their alternative out-of-sample periods. During this period, the best one-month forecasts for the UK PPI are obtained by the Elastic Net, which yields an MSE improvement of only 5 percent over their AR benchmark.

Apart from different time periods, there are two main alternatives to explain these gaps in forecasting accuracy. First, the Finnish PPI may be fundamentally easier for to predict using high-dimensional models, for some reason or the other. Geographical variation in forecasting results is not an uncommon occurrence in the empirical literature (see, e.g., Eickmeier and Ziegler 2008). While this is a feature of the research design that is beyond the control of the researcher, the choice of data, which is the other potential reason, is not. The aforementioned studies rely on somewhat generic sets of predictors, which feature a wide array of different macroeconomic and financial variables. It is possible that the set of predictors used here, despite its limitations, may lend itself better to forecasting the PPI in particular. Specifically, the most distinguishing feature of the present dataset is the prevalence of highly disaggregated commodity prices, which could explain the improvements in accuracy. To this end, the explicit variable selection features of the Lasso-based models can be used to shed light on the relevance of individual predictors in the overall forecasting results.

Figure 8 presents the number of times that 30 of the variables most frequently selected by Ridge-based Adaptive Lasso were featured in the forecasts. The shorthands used in the figure for individual predictors are explained in Table 4 of the appendix. It turns out, that several

commodity prices are prominently represented: out of the 30 most frequent predictors, 18 are commodity prices, with the price of copper featured in each of the 140 samples.

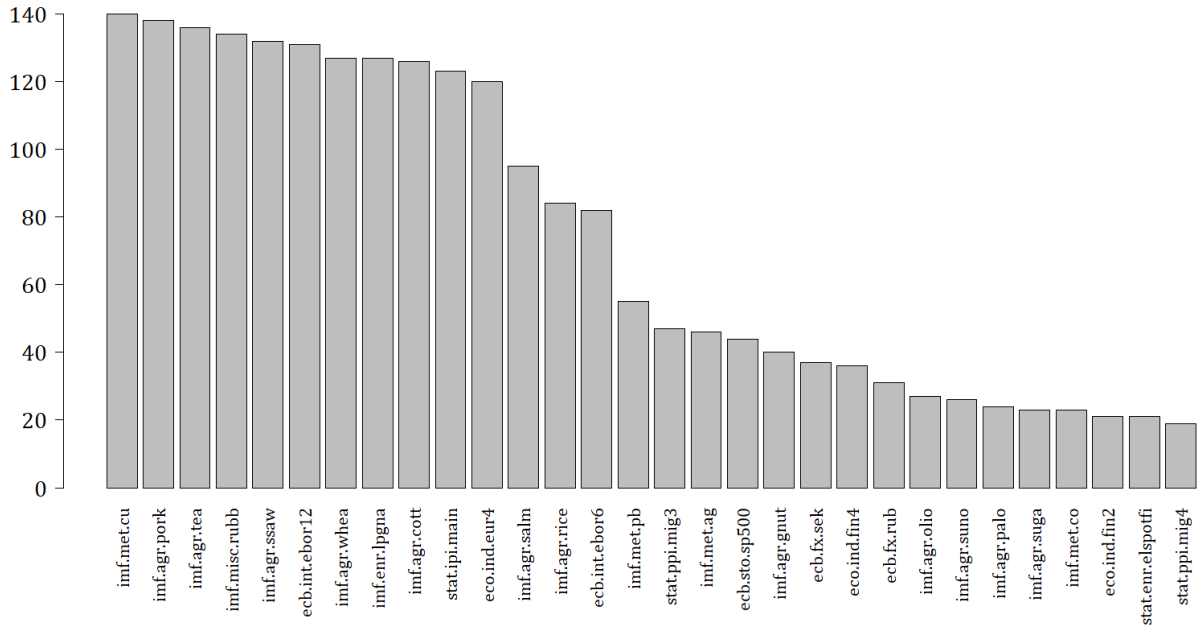

Figure 8. *Variables most frequently selected by the Adaptive Lasso (Ridge) model over the 2008M04–2019M11 out-of-sample-period.*

While not conclusive nor particularly formal, these results suggest that there may indeed be some information contained in the international commodity prices, that is particularly relevant for forecasting the aggregate producer prices. While the other studies also feature some commodity prices as predictors, they are far fewer especially relative to the other predictors included. Given that the factor models especially have been shown to be sensitive to the composition of the set of predictors (Boivin and Ng 2006), it would be interesting to examine, if a dataset featuring commodity prices more prominently, such as the one here, would be useful for forecasting the producer prices of other countries as well.

The detrimental effects of the pandemic cannot, however, be escaped. While the historical experiment is of academic interest, more recent data reveals the sensitivity of high-dimensional methods to structural shocks, which undermines their utility in practical use.

6.3 Nowcasting Experiment

The analysis so far has indicated that, the set of predictors considered provides limited predictive capacity over the PPI in the short run. There is, however, one advantage that the predictors possess, whose potential for forecasting remains to be unexplored. That is the fact that a

majority of the predictors considered here represent market variables that are quoted at daily frequencies, owing to which information regarding these variables usually becomes available with little delay. Even if these variables appear to provide little predictive power over the forthcoming months' PPI, there remains a possibility that they may comove with aggregate producer prices. In this case, the temporal advantage of the market variables relative to the PPI could be used to give leading indication of the latter by using contemporaneous observations instead of past ones for prediction. This possibility will be explored next in a separate out-of-sample experiment.

The following simulation has been conducted by including only the market variables in the set of predictors and by aligning their values with the concurrent values of PPI in the forecasting equation. That is, the set of predictors excludes the non-market variables altogether to produce a panel of 87 potential external predictors. These variables are elaborated in Table 4 of the appendix. Apart from this, the experimental procedure is unaltered from the earlier simulations.

A practical limitation of the present application is that the data used represents monthly averages of the market variables, which become available only after the end of the respective month. Thus, the window of opportunity for nowcasting is inherently limited as the PPI itself is published only mere weeks later than this. While this short lead on the actual PPI inevitably limits utility of these nowcast, the upcoming experiment could, nonetheless, provide some meaningful insights into the relationship between the predictors and the target variable. Thus, more so than a potential tool for practical nowcasting, the following experiment should be considered as a sort of proof-of-concept for the viability of using contemporaneous market information in predicting the PPI. Specifically, given that the monthly averaged data used here is based mostly on daily data, this underlying data could be leveraged directly to provide timelier nowcasts in further applications.

The results of the nowcasting experiment are presented in Table 3. The AR model is, once again, used as the main benchmark, against which the performance of the high-dimensional models is compared to in terms of aggregate errors and the Diebold-Mariano test. As the AR and RW forecasts utilize no external predictors, their results are identical to the earlier one-month-ahead forecasts but are nonetheless repeated here for convenience. The results of Table 3 offer a striking contrast to the earlier forecast experiments. Using contemporaneous information, all of the high-dimensional models not only manage to surpass the accuracy of the AR benchmark on

average, but this time around the improvements are also statistically significant: the one-sided null hypothesis of superior forecasting performance of the AR model can be rejected comfortably at the one percent confidence level for all high-dimensional models.

Table 3

Results of the out-of-sample nowcasting experiment for 2011M12–2021M11.

Model	MAE	RMSE	p-value
AR	*0.00493*	*0.00694*	
RW	1.222	1.145	0.983
DFM (AH)	0.871	0.831	0.004
DFM (R=1)	0.870	0.831	0.004
DFM (R=3)	0.885	0.853	0.001
DFM (R=5)	0.889	0.863	0.007
DFM (R=7)	0.819	0.800	0.003
DFM (R=9)	0.795	0.790	0.003
T-DFM (AH)	0.823	0.808	0.002
T-DFM (R=1)	0.833	0.815	0.002
T-DFM (R=3)	0.814	0.791	0.001
T-DFM (R=5)	0.749	0.764	0.001
T-DFM (R=7)	0.736	0.765	0.002
T-DFM (R=9)	0.720	**0.739**	0.001
Ridge	0.764	0.758	0.001
Lasso	0.726	0.747	0.000
Elastic Net	0.734	0.753	0.001
Ada-Lasso (OLS)	0.743	0.762	0.001
Ada-Lasso (Ridge)	**0.699**	0.740	0.001
Ada-Lasso (Lasso)	0.717	0.752	0.001

Note: See notes for Table 1.

The basic factor models yield improvements in ranging between 11 and 20 percent over the AR benchmark depending on the number of factors used. In contrast to the forecasting experiment, best performance is now obtained by using as many as nine of the estimated factors. On the other hand, the Ahn-Horenstein eigenvalue ratio test suggest using no more than two factors for any given forecast, with two and one factors favored over the earlier and latter parts of the period respectively. This implies that the common processes that characterize the complete set of predictors is not a particularly good counterpart for the PPI in this case.

The results of the hybrid models provide additional evidence in favor of this conclusion: in yet another interesting contrast to the earlier forecasting results, the factor models based on targeted predictors appear to provide systematic improvements compared to the pure factor

models. This is consistent with the results of Boivin and Ng (2006), who show that augmenting factor models with unnecessary variables can have a detrimental effect on forecasting accuracy. Moreover, the results are in line with those of Bai and Ng (2008), who suggest that the accuracy of factor models may be, conversely, improved upon by preselecting the variables prior to factor estimation by using penalized regressions. Despite using targeted subsets of the predictors, the eigenvalue ratio test, again, appears to provide a rather poor fit to the target variable by underestimating the number factors: much like in the case of the pure factor model, the test indicates using either one or two factors for every forecast, while the best results are obtained using as many as nine.

The overall greatest gains in predictive performance are, however, offered by the penalized regressions. While the penalized models were at a disadvantage compared to the factor models in the forecasting scenario, this time, the situation has been decisively turned in favor of the former, with the penalized models exhibiting superior performance over all of the pure factor models. Moreover, the rankings among the penalized models appear to be distinctively reversed as well. The Ridge regression, which was the best-performing penalized model in the strict forecasting application, now occupies the bottom of the bunch, whereas the most promising results are exhibited by the Adaptive Lasso models. Out of these, the variant utilizing the Ridge regression to estimate the adaptive weights achieves greatest accuracy overall in terms of MAE, with an impressive improvement of 30 percent with respect to the univariate benchmark. Regarding the performance among the different Adaptive Lasso models, these results also provide some further indication that using shrinkage models instead of OLS to estimate the initial penalty weights for the regression may be advantageous in high-dimensional applications even if sufficient numbers of observations were available to make the latter option feasible.

A further advantage of the penalized models relative to the factor models showcased here is the fact that the former tend to work well 'out-of-the-box.' That is, the models specified at the forecast origin using contemporary information deliver good results. The best factor models, on the other hand, utilize fixed numbers of factors, which means that they could be identified only after the fact. For this reason, the most justified reference models to the penalized models would arguably be the factor models based on the Ahn-Horenstein test since their specification is equally based on prior information. In this sense, the tendency of the eigenvalue test to underestimate the number of factors with respect to forecasting performance widens the gap

between the respective approaches even further. This observation corresponds to earlier results in the literature regarding the use of information criteria to determine factor numbers: for example, Boivin and Ng (2006) point out that the information criteria of Bai and Ng (2002) also tend to underestimate number of factors in empirical forecasting applications.

While nowcasting studies predominantly concentrate on predicting particularly low-frequency variables, such as GDP, and, to our knowledge, no published studies deal with nowcasting aggregate producer prices in particular, general parallels can be drawn between the existing literature and the results encountered here. Specifically, the results here are consistent with a number of studies that attest to the usefulness of contemporary information in predicting macroeconomic variables (e.g., Giannone et al., 2008; Heinisch and Scheufele, 2017). In particular, Monteforte and Moretti (2013) nowcast the euro area CPI inflation using a number of market-based predictors, including interest rates and commodity prices. Given the nature of the target variables and the set of predictors, their study can be considered somewhat more closely related to the present one, although they utilize the daily observations of the predictors directly in a forecast model encompassing the MIDAS framework. Their results indicate that the forecast model augmented with daily market data provides a nearly 20 percent improvement in forecasting accuracy compared to an AR benchmark, which is similar to the results obtained here for the PPI using monthly averages.

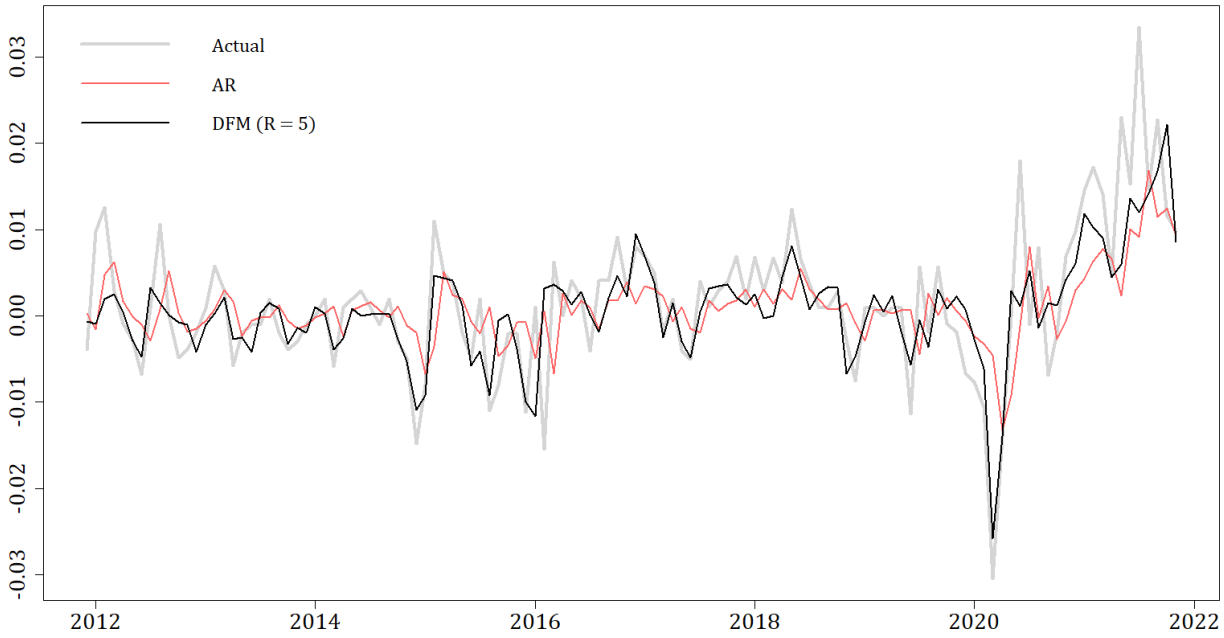

Figure 9. *Comparison of the AR forecasts and the Adaptive Lasso (Ridge) nowcasts against the realized PPI.*

To illustrate the beneficial role of contemporary data over time, Figure 9 plots the series of out-of-sample nowcasts for the MAE best Adaptive Lasso next to the PPI and the one-month ahead forecasts of the AR model. Compared to the high-dimensional forecasts, the advantage of the additional information contained in the exogenous predictors is more evident in the series of nowcasts. Instead of slavishly following the AR forecast, the nowcast is noticeably more independent and manages to capture some significant shocks to the PPI inflation impressively well in advance instead of the delayed responses that were characteristic of the forecasts examined earlier. Notable examples of such successes include the abrupt drops in producer price inflation that occurred in late 2014 and early 2020, which the Adaptive Lasso manages to account for nearly perfectly in terms of both timing and magnitude. That being said, there are still phases when the nowcasts are noticeably detached from the realized values. This is most evident around the start of 2016 and following early 2020. That is, even contemporary data has trouble keeping up with the upheaval caused by the pandemic, although to an ostensibly lesser extent than the strict forecasts.

The tendencies of the AR and Adaptive Lasso predictions are further illustrated in the scatterplots of Figure 10, in which the realized values of the PPI are plotted against the forecasts and nowcasts of the respective models. In both panels, the dashed line is the 45-degree line, with points lying below and above the line indicating over- and undershooting of the individual forecasts respectively. The solid line is fitted based on the observations. While the differences between the models are not enormous, the observation pairs of the Adaptive Lasso nowcast tend to reside closer to the 45-degree line and are more evenly spread out along the horizonal axis, whereas the AR forecasts are visibly more concentrated around the center of the plot. This illustrates the notion that a major part of the relative advantage of the Adaptive Lasso nowcast arises from its ability to account particularly well for extreme values of the target variable. The fitted lines, on the other hand, indicate that both models exhibit a similar tendency to underestimate large positive, and overestimate large negative values of the target on average respectively.

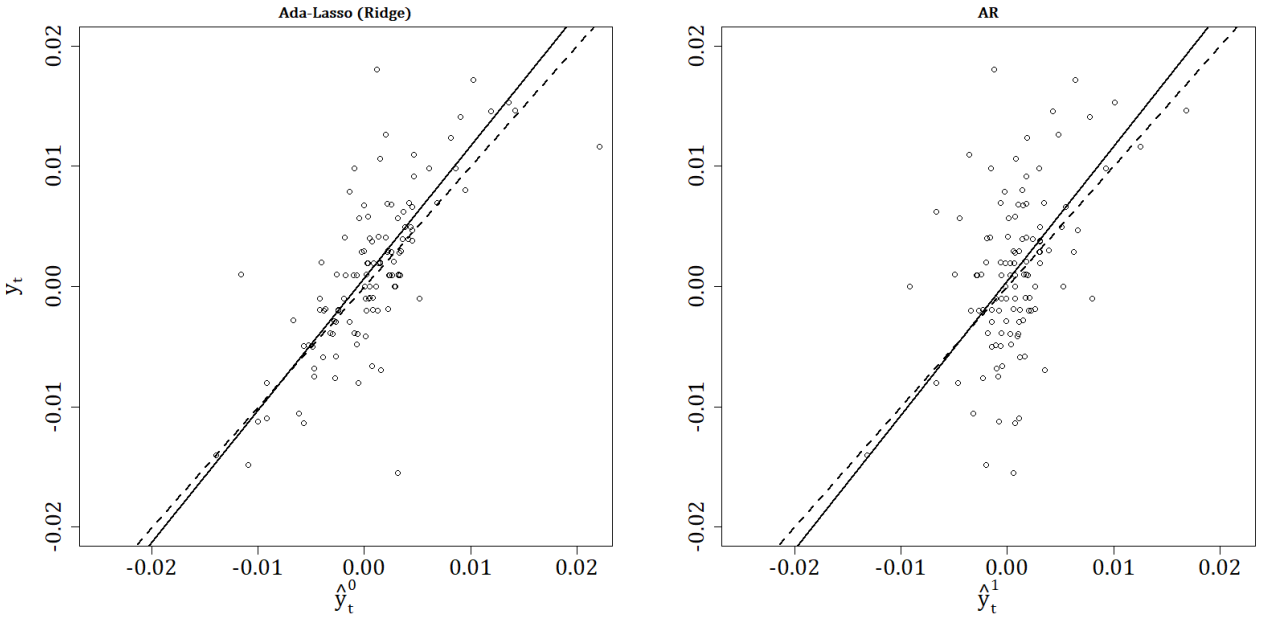

Figure 10. *Scatterplots of the AR forecasts and Adaptive Lasso (Ridge) nowcasts against the realized values of the PPI. The dashed line is the 45-degree line, whereas the solid line is fitted on the observations.*

In analogy to the forecasting experiment, it is of interest to briefly examine the evolution of the relative errors of some of the models over time. The models included are the best performers of the factor and penalized classes, the nine-factor model and the Ridge-based Adaptive Lasso model, respectively. In addition, the targeted predictor factor model with factor selection by the eigenvalue ratio test is included, as, based on the discussion above regarding the model selection issues of the factor models, it can arguably be considered a more genuine reflection of the performance of the factor models in practice.

Figure 11 plots the cumulative RMSE and MAE of these models relative to the AR benchmark. The contrast to the forecast experiment is quite striking: while the forecast models exhibited gradually deteriorating relative accuracy towards the end of the out-of-sample period, the cumulative mean errors of the nowcasts remain remarkably stable for most of its duration, especially in terms of MAE. In the case of the nine-factor model, the cumulative errors even exhibit a clear declining tendency. However, apart from the absolute beginning of the observation period, the Adaptive Lasso consistently and considerably dominates all competing models throughout the out-of-sample period.

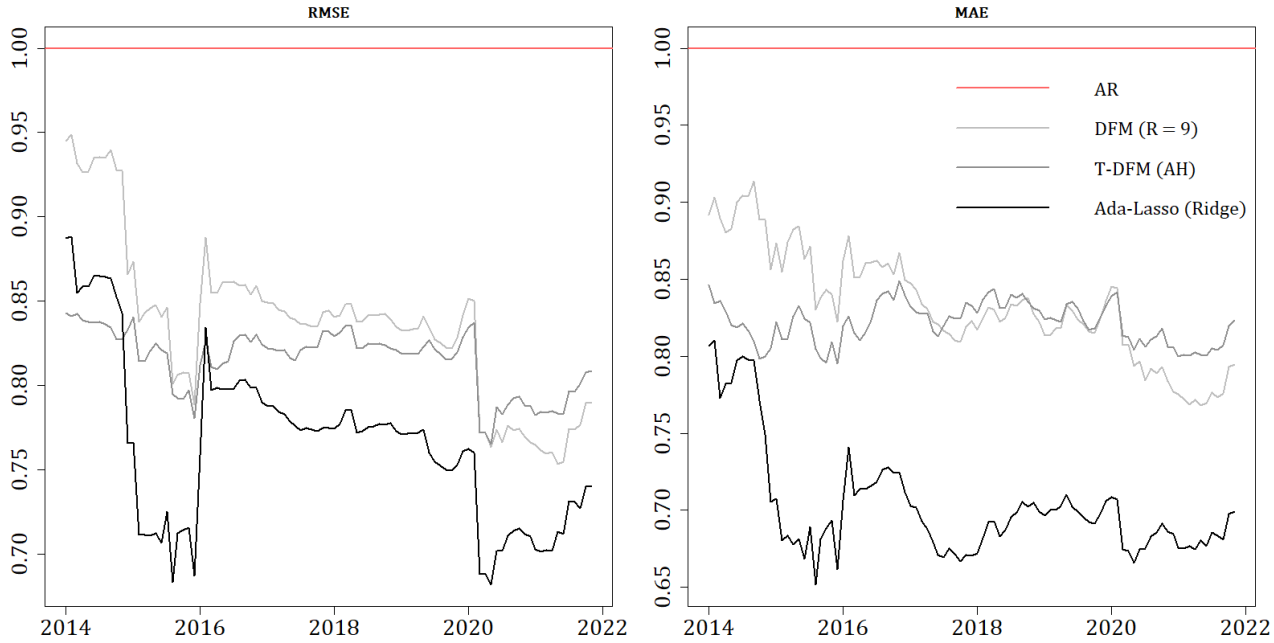

Figure 11. *Cumulative errors of select high-dimensional nowcasts relative to the AR forecast.*

The superior performance of the Lasso based models relative to the pure factor models, in particular, implies that the relationship between the set of predictors and the PPI may be characterized by sparsity. That is, instead of common factors aggregated over the entire set of predictors, only a small subset thereof is likely to be relevant to predicting the target, which is a scenario where the Lasso-based models would be expected to fare particularly well (Smeekes and Wijler 2018). This is particularly interesting given the opposite results of the strict forecasting experiment earlier, which implied information contained in the entire set of predictors as a whole to be somewhat more effective than individual predictors.

The variable selection capability of the Lasso-based models can once again be used to investigate this issue in more detail. To this end, Figure 12 presents an illustration of the variables selected by the Ridge-based Adaptive Lasso model over time for each of the 120 out-of-sample periods. With each vertical block of the grid corresponding to a single nowcast, each colored rectangle represents the estimated coefficient assigned to the corresponding predictor, which are listed along the vertical axis of the plot. Furthermore, the color of the rectangle indicates the magnitude of the respective coefficient: the darker the color, the larger the coefficient is in absolute value, with white indicating a zero coefficient for the specific forecast and predictor. As all predictors are scaled to have unit variance prior to estimation, the magnitude of the coefficient can be interpreted as being roughly indicative of the influence of a given variable for the nowcast of the target variable. To enhance interpretability, the figure includes only the

variables that have been assigned a nonzero coefficient at least once during the forecasting experiment. Thus, it excludes 48 predictors out of a total of 93, which were not utilized by the forecast model at any point in time.

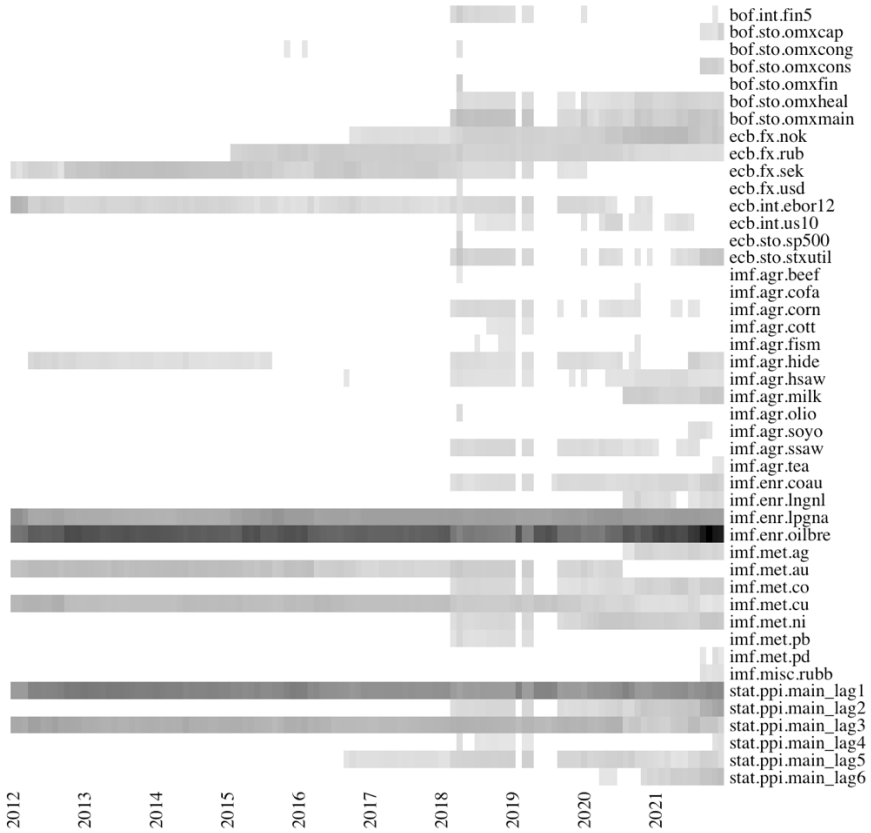

Figure 12. *Variable selection and coefficient magnitude of the Adaptive Lasso (Ridge) nowcast over time.*

The figure provides evidence in favor of the sparsity hypothesis: especially for the earlier part of the out-of-sample period, the Adaptive Lasso employs particularly few predictors, with no more than eleven variables utilized for any given nowcast. While model complexity increases towards the end of the out-of-sample period, no more than 28 predictors are utilized for any single forecast. The variation in the number of nonzero coefficients over time is caused by the complexity parameter, which is adjusted at each horizon based on the cross-validation procedure, the results of which are, in turn, affected by the characteristics of the data observed up to any given time-point. The sudden and conspicuous increase in the number of nonzero coefficients following early 2018 is puzzling but reflects the workings of this in-sample optimization process.

Even though the number of nonzero coefficients varies, the choices of variables remain mostly consistent over time, with the same predictors usually chosen for consecutive forecasts. In

particular, out of the 93 predictors, five are picked by the Adaptive Lasso for every one of the 120 nowcasts. These include the prices of crude oil, liquified petroleum gas, and copper, as well as the first and third lags of the PPI itself. Furthermore, as indicated by the colors of the respective entries in Figure 12, these predictors also tend to be the ones that exert the most influence over the target variable, given their estimated coefficients. Other frequently featured variables include the price of gold, the 12-month Euribor interest rate, the foreign exchange rates of the Russian ruble, the Swedish krone, and the Norwegian krone.

A noteworthy feature is that many of these variables seem to relate to the costs of energy either directly or indirectly. In addition to oil and gas prices, the Norwegian and Russian exchange rates could also be related to fluctuations in energy prices as these countries are major exporters of fossil fuels and changes in their exchange rates may reflect changes in the costs faced by importers of energy. Energy prices, oil in particular, play a key role throughout the production chain by affecting the costs faced by producers both directly and indirectly. In particular, the Finnish PPI features the producer prices of various oil-refined products, such as gasoline, prominently, with a combined weight of some 7.5 percent in the overall index as of 2021. As these prices are likely to be both volatile and highly responsive to changes in the prices of crude oil, it is not unlikely that at least a part of the contemporaneous predictive capability over the aggregate index arises from these subindices in particular.

Despite the encouraging results presented above, the caveat of the timing of the data must be borne in mind. Even though the nowcasts of the mean error best Adaptive Lasso model would have been only 0.34 percentage points away from the realized values of monthly PPI inflation on average, the fact that the data used becomes available only weeks prior to the actual target variable reduces the practical value of this result. Nonetheless, the results can be considered as evidence of the predictive capacity of contemporaneous market data over the Finnish PPI. A potential way to leverage this capacity in a more meaningful way would be by using the higher frequency data underlying the monthly averages used here. This could be accomplished by using mixed frequency frameworks, such as the MIDAS regression, the capabilities of which have been demonstrated in the nowcasting literature (e.g, Monteforte and Moretti 2013; Heinisch and Scheufele 2017). Using daily data would allow truly intra-monthly nowcasts to be constructed using information accumulated up to the nowcast origin. Whether or not such data would be useful in providing early indication of contemporaneous producer price inflation is a topic that could be explored in further research.

While collecting daily data on all of the market predictors is potentially an unreasonably arduous task, the results of the Adaptive Lasso nowcast, which implies sparsity in the set of predictors, indicate that favorable results could be obtained by using far fever predictors. As an illustration of this possibility, one final out-of-sample nowcasting experiment was conducted using only the seven external predictors selected by the Adaptive Lasso during the earliest part of the out-of-sample period in addition to six lags of the target variable. Given the reduced number of predictors, the forecast equation was estimated by means of standard OLS. On average, this parsimonious model produces an MAE of 0.0035 and an RMSE of 0.0051, indicating accuracy roughly on par with or even slightly in excess of the Adaptive Lasso over the main out-of-sample period. This result indicates that high-dimensional data may not in itself be strictly necessary to produce favorable results in prediction. Rather than deriving each nowcast from the entire dataset, a more efficient way to utilize high-dimensional data could be to screen the data infrequently for relevant predictors, and to use this subset, in turn, to produce the actual forecasts in the intervening periods. This would alleviate the need to collect data on potentially irrelevant predictors for every forecast, thus reducing the burden pertaining to data collection, which is a relevant concern in practical applications.

7 Conclusions

The objective of this study has been to examine the possibility of predicting the Finnish Producer Price Index for Manufactured Goods, a monthly indicator of aggregate producer price inflation, in the short term. Specifically, the intention has been to find out whether particularly large sets of predictive variables could be useful in this regard. To this end, a dataset consisting of a total of 114 predictors, including market prices of commodities and financial market variables, was utilized.

Different forecasting models designed specifically to accommodate such high-dimensional sets of predictors were considered. These models fall under two distinct classes of methodologies: models based on dynamic factors and models based on penalized regressions. While both classes of methods have been found to be effective in leveraging information contained in high-dimensional data for forecasting macroeconomic variables, their contrasting approaches to dimensional reduction could offer complementary features to account for uncertainties involved in the data generating process of the target variable. The performance of these forecasting approaches relative to each other is also a topic, upon which this study sought to shed light on.

The empirical performance of the respective models was assessed by way of a pseudo real-time out-of-sample forecasting experiment, in which historical data is used to simulate a sample of forecasts in a realistic manner. The main out-of-sample period ranged from December 2011 to November 2021. In addition, the performance of the forecasts was examined over alternative periods to gauge the effects of the COVID-19 pandemic on predictive accuracy.

At forecast horizons ranging from one to three months, the high-dimensional models are found to provide only limited advantages compared to a univariate benchmark model over the main out-of-sample period. The dynamic factor models generally yield the most accurate average forecasts, while the penalized models fare only slightly worse. An examination of the evolution of forecast errors over time reveals that the high-dimensional models would have performed somewhat better during the period preceding the outbreak of the global pandemic in early 2020. A separate out-of-sample experiment excluding the pandemic period exhibits improved results, especially relative to earlier studies considering the forecasting of the PPI in other countries. However, the observed sensitivity of the high-dimensional forecasts to structural shocks, as evidenced by the deterioration of their performance following the onset of the pandemic, undermines their utility in practical applications.

In addition to using strictly past observations of the predictors to forecast the PPI, a separate experiment was conducted to explore whether contemporary values of a subset of the predictors could be used to provide more meaningful gains in predictive accuracy. This experiment could be justified by the temporal advantage in availability enjoyed by certain market-based predictors relative to the PPI. The results of this nowcasting experiment indicate that contemporary information could indeed provide far greater benefits in predictive accuracy compared to earlier vintages of data. The experiment also revealed more substantial differences among the classes of methodologies considered, with penalized methods, the Adaptive Lasso in particular, providing superior performance over the factor models. Although the best of the tested factor models are not far behind, the scale is tipped further in favor of the penalized models owing to model selection issues that burden the former approach. Specifically, the eigenvalue ratio test utilized here tends to underestimate the optimal number of factors, which inhibits the applicability of the factor models to some extent.

The success of the Adaptive Lasso model, especially, indicates that the greatest gains could be achieved by using only a sparse subset of the predictors rather than information summarized over the whole set of predictors: predictors related to the price of energy, in particular, appear

to be relevant as contemporaneous indicators of aggregate producer prices. That is, rather than the high-dimensional data itself, being able to identify relevant variables from among the large set of candidates appears the be the key to obtaining optimal predictive performance in this case.

While promising, a major caveat of the latter results is the fact that they were achieved using data averaged over the whole month: despite being available relatively quickly, the monthly data leads the PPI itself only by mere weeks, which gives limited scope for conducting nowcasting in practice. Regardless, the favorable results indicate that the very near-term prediction of the Finnish PPI could hold potential worthy of further investigation. Specifically, daily data on the market variables could be used to move the forecast origin further back to produce more timely, truly intra-monthly nowcasts of the index. Exploring the potential of such higher frequency data for nowcasting the Finnish PPI could be a viable subject for further research on the topic.

References

Ahn, Seung C., and Alex R. Horenstein. 2013. 'Eigenvalue Ratio Test For The Number of Factors'. *Econometrica* 81 (3): 1203–27. https://doi.org/10.3982/ECTA8968.

Artis, Michael J., Anindya Banerjee, and Massimiliano Marcellino. 2005. 'Factor Forecasts for the UK'. *Journal of Forecasting* 24 (4): 279–98. https://doi.org/10.1002/for.957.

Bai, Jushan, and Serena Ng. 2002. 'Determining the Number of Factors in Approximate Factor Models'. *Econometrica* 70 (1): 191–221. https://doi.org/10.1111/1468-0262.00273.

———. 2008. 'Forecasting Economic Time Series Using Targeted Predictors'. *Journal of Econometrics* 146 (2): 304–17. https://doi.org/10.1016/j.jeconom.2008.08.010.

Bańbura, Marta, Domenico Giannone, and Lucrezia Reichlin. 2010. 'Large Bayesian Vector Auto Regressions'. *Journal of Applied Econometrics* 25 (1): 71–92. https://doi.org/10.1002/jae.1137.

Bergmeir, Christoph, and José M. Benítez. 2012. 'On the Use of Cross-Validation for Time Series Predictor Evaluation'. *Information Sciences* 191: 192–213. https://doi.org/10.1016/j.ins.2011.12.028.

Boivin, Jean, and Serena Ng. 2006. 'Are More Data Always Better for Factor Analysis?' *Journal of Econometrics* 132 (1): 169–94. https://doi.org/10.1016/j.jeconom.2005.01.027.

Caggiano, Giovanni, George Kapetanios, and Vincent Labhard. 2011. 'Are More Data Always Better for Factor Analysis? Results for the Euro Area, the Six Largest Euro Area Countries and the UK'. *Journal of Forecasting* 30 (8): 736–52. https://doi.org/10.1002/for.1208.

Carriero, Andrea, Ana Beatriz Galvão, and George Kapetanios. 2019. 'A Comprehensive Evaluation of Macroeconomic Forecasting Methods'. *International Journal of Forecasting* 35 (4): 1226–39. https://doi.org/10.1016/j.ijforecast.2019.02.007.

Clark, Todd E., and Michael W. McCracken. 2009. 'Improving Forecast Accuracy by Combining Recursive Forecasts'. *International Economic Review* 50 (2): 363–95. https://doi.org/10.1111/j.1468-2354.2009.00533.x.

Diebold, Francis X. 2015. 'Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests'. *Journal of Business & Economic Statistics* 33 (1): 1–1. https://doi.org/10.1080/07350015.2014.983236.

Diebold, Francis X., and Roberto S. Mariano. 1995. 'Comparing Predictive Accuracy'. *Journal of Business & Economic Statistics* 13 (3): 253-. https://doi.org/10.2307/1392185.

Doz, Catherine, Domenico Giannone, and Lucrezia Reichlin. 2011. 'A Two-Step Estimator for Large Approximate Dynamic Factor Models Based on Kalman Filtering'. *Journal of Econometrics* 164 (1): 188–205. https://doi.org/10.1016/j.jeconom.2011.02.012.

Eickmeier, Sandra, and Tim Ng. 2011. 'Forecasting National Activity Using Lots of International Predictors: An Application to New Zealand'. *International Journal of Forecasting* 27 (2): 496–511. https://doi.org/10.1016/j.ijforecast.2009.10.011.

Eickmeier, Sandra, and Christina Ziegler. 2008. 'How Successful Are Dynamic Factor Models at Forecasting Output and Inflation? A Meta-Analytic Approach'. *Journal of Forecasting* 27 (3): 237–65. https://doi.org/10.1002/for.1056.

Elliot, Graham, and Allan Timmermann. 2016. *Economic Forecasting*. Princeton, New Jersey: Princeton University Press.

Elliott, Graham, Antonio Gargano, and Allan Timmermann. 2013. 'Complete Subset Regressions'. *Journal of Econometrics* 177 (2): 357–73. https://doi.org/10.1016/j.jeconom.2013.04.017.

———. 2015. 'Complete Subset Regressions with Large-Dimensional Sets of Predictors'. *Journal of Economic Dynamics & Control* 54: 86–110. https://doi.org/10.1016/j.jedc.2015.03.004.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. 2005. 'The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting'. *Journal of the American Statistical Association* 100 (471): 830–40. https://doi.org/10.1198/016214504000002050.

Garcia, Márcio GP, Marcelo C. Medeiros, and Gabriel FR Vasconcelos. 2017. 'Real-Time Inflation Forecasting with High-Dimensional Models: The Case of Brazil'. *International Journal of Forecasting* 33 (3): 679–93. https://doi.org/10.1016/j.ijforecast.2017.02.002.

Ghysels, Eric, Arthur Sinko, and Rossen Valkanov. 2007. 'MIDAS Regressions: Further Results and New Directions'. *Econometric Reviews* 26 (1): 53–90. https://doi.org/10.1080/07474930600972467.

Giannone, Domenico, Lucrezia Reichlin, and David Small. 2008. 'Nowcasting: The Real-Time Informational Content of Macroeconomic Data'. *Journal of Monetary Economics* 55 (4): 665–76. https://doi.org/10.1016/j.jmoneco.2008.05.010.

Goulet Coulombe, Philippe, Massimiliano Marcellino, and Dalibor Stevanović. 2021. 'Can Machine Learning Catch the COVID-19 Recession'. *National Institute Economic Review* 256: 71–109. https://doi.org/10.1017/nie.2021.10.

Harvey, David, Stephen Leybourne, and Paul Newbold. 1997. 'Testing the Equality of Prediction Mean Squared Errors'. *International Journal of Forecasting* 13 (2): 281–91. https://doi.org/10.1016/S0169-2070(96)00719-4.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. Springer.

Heinisch, Katja, and Rolf Scheufele. 2017. 'Bottom-up or Direct? Forecasting German GDP in a Data-Rich Environment'. *Empirical Economics* 54 (2): 705–45. https://doi.org/10.1007/s00181-016-1218-x.

Hendry, David F., and Kirstin Hubrich. 2011. 'Combining Disaggregate Forecasts or Combining Disaggregate Information to Forecast an Aggregate'. *Journal of Business & Economic Statistics* 29 (2): 216–27. https://doi.org/10.1198/jbes.2009.07112.

Hoerl, Arthur E., and Robert W. Kennard. 1970. 'Ridge Regression: Biased Estimation for Nonorthogonal Problems'. *Technometrics* 12 (1): 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Hubrich, Kirstin. 2005. 'Forecasting Euro Area Inflation: Does Aggregating Forecasts by HICP Component Improve Forecast Accuracy?' *International Journal of Forecasting* 21 (1): 119–36. https://doi.org/10.1016/j.ijforecast.2004.04.005.

Kim, Hyun Hak, and Norman R. Swanson. 2014. 'Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence'. *Journal of Econometrics* 178: 352–67. https://doi.org/10.1016/j.jeconom.2013.08.033.

Koop, Gary M. 2013. 'Forecasting with Medium and Large Bayesian VARS'. *Journal of Applied Econometrics (Chichester, England)* 28 (2): 177–203. https://doi.org/10.1002/jae.1270.

Li, Jiahan, and Weiye Chen. 2014. 'Forecasting Macroeconomic Time Series: LASSO-Based Approaches and Their Forecast Combinations with Dynamic Factor Models'. *International Journal of Forecasting* 30 (4): 996–1015. https://doi.org/10.1016/j.ijforecast.2014.03.016.

Lütkepohl, Helmut. 2007. 'General-to-Specific or Specific-to-General Modelling? An Opinion on Current Econometric Terminology'. *Journal of Econometrics* 136 (1): 319–24. https://doi.org/10.1016/j.jeconom.2005.11.014.

Maehashi, Kohei, and Mototsugu Shintani. 2020. 'Macroeconomic Forecasting Using Factor Models and Machine Learning: An Application to Japan'. *Journal of the Japanese and International Economies* 58: 101104-. https://doi.org/10.1016/j.jjie.2020.101104.

Marcellino, Massimiliano, James H. Stock, and Mark W. Watson. 2006. 'A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series'. *Journal of Econometrics* 135 (1): 499–526. https://doi.org/10.1016/j.jeconom.2005.07.020.

Masini, Ricardo P., Marcelo C. Medeiros, and Eduardo F. Mendes. 2021. 'Machine Learning Advances for Time Series Forecasting'. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12429.

Medeiros, Marcelo C., and Eduardo F. Mendes. 2016. 'L1 -Regularization of High-Dimensional Time-Series Models with Non-Gaussian and Heteroskedastic Errors'. *Journal of Econometrics* 191 (1): 255-. https://doi.org/10.1016/j.jeconom.2015.10.011.

Medeiros, Marcelo C., and Gabriel FR Vasconcelos. 2016. 'Forecasting Macroeconomic Variables in Data-Rich Environments'. *Economics Letters* 138: 50–52. https://doi.org/10.1016/j.econlet.2015.11.017.

Monteforte, Libero, and Gianluca Moretti. 2013. 'Real-Time Forecasts of Inflation: The Role of Financial Variables'. *Journal of Forecasting* 32 (1): 51–61. https://doi.org/10.1002/for.1250.

Moser, Gabriel, Fabio Rumler, and Johann Scharler. 2007. 'Forecasting Austrian Inflation'. *Economic Modelling* 24 (3): 470–80. https://doi.org/10.1016/j.econmod.2006.10.003.

Smeekes, Stephan, and Etienne Wijler. 2018. 'Macroeconomic Forecasting Using Penalized Regression Methods'. *International Journal of Forecasting* 34 (3): 408–30. https://doi.org/10.1016/j.ijforecast.2018.01.001.

Statistics Finland. 2020. 'Producer Price Indices for Manufactured Products and Services 2015=100: Handbook for Users'. Statistics Finland.

Stock, James H., and Mark W. Watson. 2002a. 'Forecasting Using Principal Components From a Large Number of Predictors'. *Journal of the American Statistical Association* 97 (460): 1167–79. https://doi.org/10.1198/016214502388618960.

———. 2002b. ‘Macroeconomic Forecasting Using Diffusion Indexes’. *Journal of Business & Economic Statistics* 20 (2): 147–62. https://doi.org/10.1198/073500102317351921.

———. 2004. ‘Combination Forecasts of Output Growth in a Seven-Country Data Set’. *Journal of Forecasting* 23 (6): 405–30. https://doi.org/10.1002/for.928.

———. 2016. ‘Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics’. In *Handbook of Macroeconomics 2A*, 415–526.

Tibshirani, Robert. 1996. ‘Regression Shrinkage and Selection via the Lasso’. *Journal of the Royal Statistical Society*, Series B, Methodological, 58 (1): 267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Verbeek, Marno. 2017. ‘A Guide to Modern Econometrics’. Hoboken, New Jersey: Wiley.

Wang, Mu-Chun. 2009. ‘Comparing the DSGE Model with the Factor Model: An out-of-Sample Forecasting Experiment’. *Journal of Forecasting* 28 (2): 167–82. https://doi.org/10.1002/for.1081.

Yuan, Ming, and Yi Lin. 2006. ‘Model Selection and Estimation in Regression with Grouped Variables’. *Journal of the Royal Statistical Society*, Series B, Statistical Methodology, 68 (1): 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x.

Zou, Hui. 2006. ‘The Adaptive Lasso and Its Oracle Properties’. *Journal of the American Statistical Association* 101 (476): 1418–29. https://doi.org/10.1198/016214506000000735.

Zou, Hui, and Trevor Hastie. 2005. ‘Regularization and Variable Selection via the Elastic Net’. *Journal of the Royal Statistical Society*, Series B, Statistical methodology, 67 (2): 301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix

Table 4 lists the predictor variables used in the forecasting experiments. The data were sourced from Statistics Finland (StatFin), the International Monetary Fund (IMF), the Bank of Finland (BoF), the European Central Bank (ECB) and the European Commission (EC). The transformation codes correspond to levels (1), first differences (2), and first log-differences (3). The variables are organized into groups according to their type. The groups of variables used in the nowcasting experiment are indicated by an asterisk (*) next to the group name.

Table 4

List of variables.

Name	Description	Source	Trans.
Producer Price Indices			
stat.ppi.main	PPI Main Index	StatFin	3
stat.ppi.mig1	PPI Intermediate Products (MIG1)	StatFin	3
stat.ppi.mig2	PPI Capital Goods (MIG2)	StatFin	3
stat.ppi.mig3	PPI Consumer Durables (MIG3)	StatFin	3
stat.ppi.mig4	PPI Consumer Non-Durables (MIG4)	StatFin	3
Import Price Indices			
stat.ipi.main	IPI Main Index	StatFin	3
stat.ipi.mig1	IPI Intermediate Products (MIG1)	StatFin	3
stat.ipi.mig2	IPI Capital Goods (MIG2)	StatFin	3
stat.ipi.mig3	IPI Consumer Durables (MIG3)	StatFin	3
stat.ipi.mig4	IPI Consumer Non-Durables (MIG4)	StatFin	3
stat.ipi.mig5	IPI Energy Products (MIG5)	StatFin	3
Energy Prices			
stat.enr.spotfi	Electricity Spot Price, Finland	StatFin	3
*Commodity Prices**			
imf.met.al	Aluminum	IMF	3
imf.met.cu	Copper	IMF	3
imf.met.pb	Lead	IMF	3
imf.met.ni	Nickel	IMF	3
imf.met.tin	Tin	IMF	3
imf.met.zn	Zinc	IMF	3
imf.met.co	Cobalt	IMF	3
imf.met.pd	Palladium	IMF	3
imf.met.pl	Platinum	IMF	3
imf.met.au	Gold	IMF	3
imf.met.ag	Silver	IMF	3
imf.met.u	Uranium	IMF	3
imf.enr.coau	Coal, Australia	IMF	3

Table 4 (continued)

Name	Description	Source	Trans.
imf.enr.cosa	Coal, South Africa	IMF	3
imf.enr.lngnl	Natural Gas, Netherlands	IMF	3
imf.enr.oilbre	Crude Oil, Brent	IMF	3
imf.enr.lpgna	Propane	IMF	3
imf.mis.rubb	Rubber	IMF	3
imf.mis.urea	Urea Granular	IMF	3
imf.agr.bana	Banana	IMF	3
imf.agr.barl	Barley	IMF	3
imf.agr.beef	Beef	IMF	3
imf.agr.coco	Cocoa	IMF	3
imf.agr.cofa	Coffee, Arabica	IMF	3
imf.agr.cofr	Coffee, Robusta	IMF	3
imf.agr.raps	Rapeseed Oil	IMF	3
imf.agr.fism	Fishmeal	IMF	3
imf.agr.gnut	Groundnuts	IMF	3
imf.agr.hide	Cow Hides	IMF	3
imf.agr.lamb	Lamb	IMF	3
imf.agr.cott	Cotton	IMF	3
imf.agr.slog	Soft Logs	IMF	3
imf.agr.hlog	Hard Logs	IMF	3
imf.agr.corn	Corn	IMF	3
imf.agr.olio	Olive Oil	IMF	3
imf.agr.palo	Palm Oil	IMF	3
imf.agr.pork	Pork	IMF	3
imf.agr.chick	Poultry	IMF	3
imf.agr.rice	Rice	IMF	3
imf.agr.salm	Salmon	IMF	3
imf.agr.hsaw	Hard Sawnwood	IMF	3
imf.agr.ssaw	Soft Sawnwood	IMF	3
imf.agr.shri	Shrimp	IMF	3
imf.agr.soym	Soybean Meal	IMF	3
imf.agr.soyo	Soybean Oil	IMF	3
imf.agr.soyb	Soybean	IMF	3
imf.agr.suno	Sunflower Oil	IMF	3
imf.agr.suga	Sugar	IMF	3
imf.agr.tea	Tea	IMF	3
imf.agr.whea	Wheat	IMF	3
imf.agr.sorg	Sorghum	IMF	3
imf.agr.wooc	Wool, Coarse	IMF	3
imf.agr.woof	Wool, Fine	IMF	3
*Stock Market Indices**			
bof.sto.omxmain	OMX Helsinki	BoF	3
bof.sto.omxcap	OMX Helsinki CAP	BoF	3
bof.sto.omxmat	OMX Helsinki Basic Materials	BoF	3

Table 4 (continued)

Name	Description	Source	Trans.
bof.sto.omxind	OMX Helsinki Industrials	BoF	3
bof.sto.omxcong	OMX Helsinki Consumer Goods	BoF	3
bof.sto.omxcons	OMX Helsinki Consumer Services	BoF	3
bof.sto.omxheal	OMX Helsinki Health	BoF	3
bof.sto.omxfin	OMX Helsinki Financials	BoF	3
ecb.sto.stx50	Euro STOXX 50	ECB	3
ecb.sto.stxmain	Euro STOXX	ECB	3
ecb.sto.stxmat	Euro Stoxx Basic Materials	ECB	3
ecb.sto.stxfin	Euro Stoxx Financials	ECB	3
ecb.sto.stxtec	Euro Stoxx Technology	ECB	3
ecb.sto.stxind	Euro Stoxx Industrial	ECB	3
ecb.sto.stxtele	Euro Stoxx Telecommunications	ECB	3
ecb.sto.stxutil	Euro Stoxx Utilities	ECB	3
ecb.sto.sp500	Standard & Poor's 500	ECB	3
*Exchange Rates**			
ecb.fx.usd	EUR/USD	ECB	3
ecb.fx.gbp	EUR/RUB	ECB	3
ecb.fx.gbp	EUR/GBP	ECB	3
ecb.fx.cny	EUR/CNY	ECB	3
ecb.fx.sek	EUR/SEK	ECB	3
ecb.fx.nok	EUR/NOK	ECB	3
ecb.fx.jpy	EUR/JPY	ECB	3
*Interest Rates**			
ecb.int.ebor1	Euribor 1 Month	ECB	2
ecb.int.ebor3	Euribor 3 Months	ECB	2
ecb.int.ebor6	Euribor 6 Months	ECB	2
ecb.int.ebor12	Euribor 12 Months	ECB	2
ecb.int.euro5	EA Govt. Benchmark Bond Yield, 5 Years	ECB	2
ecb.int.euro10	EA Govt. Benchmark Bond Yield, 10 Years	ECB	2
ecb.int.us10	US Govt. Bond Yield, 10 Years	ECB	2
bof.int.fin5	Finnish Govt. Bond Yield, 5 Years	BoF	2
bof.int.fin10	Finnish Govt. Bond Yield, 10 Years	BoF	2
Business Surveys			
eco.ind.eurcof	Confidence Indicator, Euro Area	EC	1
eco.ind.eur1	Production Trend Observed, Euro Area	EC	1
eco.ind.eur2	Order Book Levels, Euro Area	EC	1
eco.ind.eur3	Export Order Book Levels, Euro Area	EC	1
eco.ind.eur4	Stocks of Finished Products, Euro Area	EC	1
eco.ind.eur5	Production Expectations, Euro Area	EC	1
eco.ind.eur6	Selling Price Expectations, Euro Area	EC	1
eco.ind.eur7	Employment Expectations, Euro Area	EC	1
eco.ind.fincof	Confidence Indicator, Finland	EC	1
eco.ind.fin1	Production Trend Observed, Finland	EC	1
eco.ind.fin2	Order Book Levels, Finland	EC	1

Table 4 (continued)

Name	Description	Source	Trans.
eco.ind.fin3	Export Order Book Levels, Finland	EC	1
eco.ind.fin4	Stocks of Finished Products, Finland	EC	1
eco.ind.fin5	Production Expectations, Finland	EC	1
eco.ind.fin6	Selling Price Expectations, Finland	EC	1
eco.ind.fin7	Employment Expectations, Finland	EC	1