ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Bachelor Thesis Econometrics and Operations Research

Title of thesis: Forecast Selection for Combination with Integer Programming: a Robustness Check with a Focus on the Use of Different Estimation Windows

Name of student: Cynthia Yang
Student ID number: 405574

Name of Supervisor: dr. W. Wang
Name of Second assessor: Prof. dr. P.H.B.F. Franses

July 2, 2017

Abstract

In this paper, we investigate the robustness of a forecast selection algorithm in which integer programming is used to select forecasts for averaging, given an estimated covariance matrix, instead of averaging over every available forecast. While numerous empirical studies on forecast combination methods fail to consistently outperform simple averaging, we are interested in whether this particular method succeeds in doing so. Using forecasts of real GDP growth and unemployment from the European Central Bank Survey of Professional Forecasters, we apply the algorithm in combination with different estimation windows. Besides considering a single expanding and a single rolling window, we use pseudo-out-of-sample cross-validation to determine optimal window size. We also combine different estimation windows using both equal weights and weights based on the pseudo-out-of-sample performance. The findings of our paper reveal that the algorithm is not robust to the estimation window used in the presence of data instability, and that improvements in forecast accuracy can be made by taking a careful look at pseudo-out-of-sample performance. Furthermore, the overall performance of the algorithm is not consistent across different panels and different horizons, as a consequence of bias in the estimation of the covariance matrix. When there is no significant indication of estimation error, the algorithm shows promising results relative to simple averaging.

1 Context and background

Forecasts and expectations play a crucial role in how the economy functions. For key macroeconomic variables that drive policy and decision making, there are often multiple forecasts of the same event available. Since the widely cited work of Bates and Granger (1969), a considerable number of both theoretical and empirical studies has supported the usefulness of combining individual forecasts. The forecasts being combined can either be subjective, provided for example by forecasters participating in surveys, or else provided by various quantitative models. Still, it remains unclear which combination method is best to use. Empirical studies show that simple averaging performs quite well in practice, relative to other approaches that rely on estimated combination weights. For example, Stock and Watson (2004) apply several forms of forecast combination methods on output growth for 7 countries and find that the simplest methods were the most stable and best-performing in terms of mean squared error (MSE). Genre et al. (2013) look at combining forecasts using data from the European Central Bank (ECB) Survey of Professional Forecasters (SPF), and conclude that there is no reason to replace equal weighting as the headline indicator to summarise the forecasts in the ECB SPF.

No method in previous literature consistently outperforms the equal weighted combination, both across variables and at different horizons. This phenomenon is often referred to as the "forecast combination puzzle". In order to find an explanation for this puzzle, Claeskens et al. (2016) analyse the properties of a combined forecast theoretically. The authors conclude that weight estimation may substantially affect the variance of the forecasts and that this effect is also likely to be larger for weights based on estimated covariances. However, Matsypura et al. (2017) propose an algorithm in which integer programming is used to select forecasts for averaging, given an estimated covariance matrix, instead of averaging over every available forecast. Based on an application to ECB SPF data on quarterly rates of GDP growth and unemployment, the authors conclude that this method gives improved accuracy over simple averaging and other common forecast combination methods. A somewhat remarkable result, in particular when considering the findings of Genre et al. (2013) and Claeskens et al. (2016).

Additionally, looking at data on GDP growth and unemployment, there have been large fluctuations over the years, indicating unpredictable behaviour. These large fluctuations, also known as structural breaks, are widely recognised as an important source of forecast failure in macroeconomics. We notice that the methods employed by Matsypura et al. (2017) are limited to an expanding window, using all available data. Rossi and Inoue (2012) point out that in the forecasting literature, reporting empirical results only for a single estimation window raises concerns, as statistical significance may differ across different windows due to fluctuations in the data. The main concern is that satisfactory results may be obtained by chance, or perhaps only the results for a successful window size are presented. In the same vein, it is unclear whether the empirical conclusions of Matsypura et al. (2017) are specific to the estimation window used. Therefore, in this paper, we investigate the robustness of the proposed algorithm with a focus on the use of different estimation windows. Our findings show that for most of the data sets, the performance varies with the estimation window used. The results reveal the importance of choice of estimation window in the presence of data instability. We find that the robustness of the algorithm for a specific data set can be investigated prior to forecasting out-of-sample. Accordingly, improvements in forecast accuracy can be made by selecting the estimation window based on pseudo-out-of-sample performance.

The structure of this paper is as follows. First, in Sect. 2, we give a brief outline of related work and our goals of research. In Sect. 3, we describe the forecast selection algorithm proposed by Matsypura et al. (2017). Next, in Sect. 4, we describe the data that we operate on and the way we deal with non-response, followed by the specification of our methodology in Sect. 5. In Sect. 6 we evaluate the different methods and we discuss the results. Sect. 7 concludes.

2 Review of relevant literature and goals of research

To deal with the issue of structural breaks, a conventional approach is to estimate the break points and then to base forecasts on the post-break observations. However, Pesaran and Timmermann (2007) reveal that this approach is not always optimal when the objective is to optimise forecasting performance. In addition, it is often the case that the time and size of the structural break are still uncertain. The authors propose a range of alternative methods that can be implemented without relying on exact information about breaks. They propose selecting the window size by searching across different starting points using pseudo-out-of-sample cross-validation, or combining forecasts from the same model but computed over various estimation window sizes. Using Monte Carlo simulation they show that the methods work well in comparison with methods that do not take the possible presence of breaks into account. Furthermore, they find that in the absence of breaks, using an expanding window generates the lowest out-of-sample MSE-value.

Other recent work on forecasting in the presence of breaks extend the research of Pesaran and Timmermann (2007). For example, Pesaran and Pick (2011) provide empirical evidence that averaging across estimation windows also works well in practice. Clark and McCracken (2009) show that combining expanding and rolling windows can provide improvements in forecast accuracy relative to using either of the two. However, literature on combining forecasts both across models and across estimation windows is limited. As an example, Assenmacher-Wesche and Pesaran (2008) find that averaging over estimation windows is at least as effective as, and even complements averaging over a class of related models in improving forecast precision. In the same vein, Pesaran et al. (2009) find that averaging global vector autoregressive forecasts over both different model specifications and different estimation windows gives results that outperform forecasts based on individual models.

The findings of these papers motivate us to develop similar methods that may be implemented in combination with the forecast selection algorithm proposed by Matsypura et al. (2017), using survey data from the ECB SPF. Pesaran and Timmermann (2007) and Pesaran and Pick (2011) among others demonstrate that incorporating information on break dates when combining estimation windows does not necessarily improve and may even harm the forecasting performance. Furthermore, little is known about the exact way structural breaks are reflected in the individual survey forecasts. Therefore, we choose to focus on developing methods that do not rely on exact knowledge about structural breaks. Hence, we indirectly exploit the trade-off between the bias and forecast error variance.

Our objective is to evaluate the robustness of the out-of-sample forecasting performance of the forecast selection algorithm to window size and choice. We do this by considering both expanding and rolling windows, and we use pseudo-out-of-sample cross-validation to determine optimal window size. We are also interested in whether combining different estimation windows can improve the forecasting performance, either by using equal weights, or weights based on the pseudo-out-of-sample performance. The use of pseudo-out-of-sample evaluation of forecasts for different starting points allows us to analyse different estimation windows prior to forecasting out-of-sample. This is especially useful as previous literature has shown that the ideal method may vary across different variables and different forecast horizons. By looking at the individual survey forecasts, it remains unclear whether using a single window or a combination of different windows is more appropriate.

Additionally, due to free entry and exit of forecasters in the SPF panel, there is the issue of missing data, while the forecast selection algorithm proposed by Matsypura et al. (2017) relies on an estimate of the full covariance matrix of forecast errors. Rather than imputing the missing data, the authors suggest using pairwise observations to construct the covariance matrix, which results in an unbiased estimate, as long as the observations are missing at random. However, if the observations are not missing at random, the estimate may be biased, with the consequence

of biased forecasts as well. The authors fail to confront this problem, and also in other literature, there is no consensus on an ideal way of dealing with forecast bias caused by poor estimation of the covariance matrix.

While combining different estimation windows may partially account for the forecast bias, in this paper, we investigate whether a proper bias correction can be made to the forecasts obtained using the forecast selection algorithm. We follow a method proposed by Capistrán and Timmermann (2009), where equal-weighted forecasts are adjusted based on a least squares regression. With an application to both simulated and empirical data, the authors show that their method has good overall performance and can be extended to other applications if the sample size permits the estimation. Instead of applying the bias adjustment to one-quarter-ahead or one-year-ahead equal-weighted forecasts, we apply it to one-year-ahead or two-year-ahead forecasts obtained using the forecast selection algorithm.

3 Forecast selection algorithm

Let us denote the variable of interest as $y \in \mathbb{R}$ and the n forecasts of this variable as $\hat{\boldsymbol{y}} = (\hat{y}_1 \hat{y}_2 \dots \hat{y}_n)' \in \mathbb{R}^n$. The errors of the individual forecasts are $\boldsymbol{e} = \boldsymbol{\iota} y - \hat{\boldsymbol{y}} = (e_1 e_2 \dots e_n)'$, where $\boldsymbol{\iota}$ is an n-vector of ones. Forecast errors are typically assumed to have expectation $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$ and finite covariance $\mathrm{E}(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{\Sigma}$. We construct a combination forecast by introducing a vector of weights $\boldsymbol{w} = (w_1 w_2 \dots w_n)' \in \mathbb{R}^n$. Assuming that the individual forecasts are unbiased, the condition that the weights sum to unity is generally imposed with the result that the constructed forecasts remain unbiased. A combination forecast formed with these weights is then $\hat{y}^c = \boldsymbol{w}'\hat{\boldsymbol{y}}$, whose error $e^c = y - \hat{y}^c$ has expectation $\mathrm{E}(e^c) = 0$ and variance $\mathrm{Var}(e^c) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$. Minimising the variance under the constraint $\boldsymbol{w}'\boldsymbol{\iota} = 1$ yields the weights of the optimal combination in terms of MSE:

$$\boldsymbol{w}^{opt} = \underset{\boldsymbol{w}'\boldsymbol{\iota}=1}{\arg\min}(\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}) = (\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\iota})^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\iota}.$$

The optimal combination forecast is then $\hat{y}^{opt} = (\boldsymbol{w}^{opt})'\hat{\boldsymbol{y}}$ with error variance $\mathrm{Var}(e^{opt}) = (\boldsymbol{\iota}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\iota})^{-1}$. For a combination forecast formed by simple averaging, a vector of equal weights is required:

$$\boldsymbol{w}^{avg} = \frac{1}{n}\boldsymbol{\iota}.$$

The corresponding combination forecast is again simply $\hat{y}^{avg} = (\boldsymbol{w}^{avg})'\hat{\boldsymbol{y}}$ with the error variance $\mathrm{Var}(e^{avg}) = n^{-2}\boldsymbol{\iota}'\boldsymbol{\Sigma}\boldsymbol{\iota}$. Naturally, $\mathrm{Var}(e^{avg}) \geq \mathrm{Var}(e^{opt})$, and only in very specific cases $\boldsymbol{w}^{opt}$ reduces to $\boldsymbol{w}^{avg}$ (Timmermann, 2006).

In order to obtain the optimal combination, Matsypura et al. (2017) suggest using integer programming rather than brute force to select forecasts prior to averaging. That is, certain weights are equal to each other and sum to one and the remaining weights are equal to zero. Formally, the set of all weights with such properties is

$$\mathcal{W} = \left\{\boldsymbol{w} \mid w_i = 0 \,\forall\, i \in \mathcal{S}_1, w_i = \left|\mathcal{S}_2\right|^{-1} \forall\, i \in \mathcal{S}_2,\, \mathcal{S}_1 \cup \mathcal{S}_2 = \{1, 2, \dots, n\}, \mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset\right\}.$$

Here, $\mathcal{S}_1, \mathcal{S}_2 \subseteq \{1, 2, \dots, n\}$ are sets of forecast indices, and $|\mathcal{S}_2|$ denotes the cardinality of $\mathcal{S}_2$. The forecasts whose indices are in $\mathcal{S}_2$ are selected for averaging, whereas those in $\mathcal{S}_1$ are not selected. By construction, the weights always sum to one. The number of elements in $\mathcal{W}$ is precisely $2^n - 1$, containing also simple averaging. That is, $\boldsymbol{w}^{avg} \in \mathcal{W}$, with $\mathcal{S}_1 = \emptyset$ and $S_2 = \{1, 2, \dots, n\}$. In order to formulate the problem in a way that optimising over it is a 0-1 integer programming problem, we introduce the binary variable $\tilde{w} \in \{0, 1\}^n$. Furthermore, in order to ensure convexity in the objective function, it is necessary to fix the number of non-zero

weights to a non-negative integer $k \in \{1, 2, \ldots, n\}$. The subset of $\mathcal{W}$ that has all elements having k non-zero weights can be written as

$$\mathcal{W}_k = \left\{ \boldsymbol{w} = \frac{\tilde{\boldsymbol{w}}}{k} \,\middle|\, \tilde{\boldsymbol{w}}'\boldsymbol{\iota} = k, \tilde{\boldsymbol{w}} \in \{0,1\}^n \right\}.$$

By dividing $\tilde{\boldsymbol{w}}$ by k, a weight vector $\boldsymbol{w}$ is obtained, where some elements are equal and sum to one and others are zero. $\tilde{\boldsymbol{w}}'\boldsymbol{\iota}$ is a count of the number of non-zero elements in $\tilde{\boldsymbol{w}}$, and the condition $\tilde{\boldsymbol{w}}'\boldsymbol{\iota} \geq 1$ ensures that k forecasts are selected for combination. We can see that $\tilde{\boldsymbol{w}} = \boldsymbol{\iota}$ results in simple averaging, as all forecasts are selected for averaging.

Finding the optimal solution with k equal weights is equivalent to solving Problem P1, with $\boldsymbol{\Sigma}$ the $n \times n$ positive-definite covariance matrix of forecast errors and $\tilde{\boldsymbol{w}}$ the vector of binary variables.

$$\begin{aligned} \text{min.} \quad & k^{-2}\tilde{\boldsymbol{w}}'\boldsymbol{\Sigma}\tilde{\boldsymbol{w}} \\ \text{s.t.} \quad & \tilde{\boldsymbol{w}}'\boldsymbol{\iota} \geq 1 \\ & \tilde{\boldsymbol{w}} \in \{0,1\}^n \end{aligned} \tag{P1}$$

Within the objective function, the term k^{-2} is necessary to ensure that the objective is evaluated at $\boldsymbol{w}$ and not at $\tilde{\boldsymbol{w}}$. Problem P1 is now formulated as a tractable convex optimization problem and therefore it can be solved using a general-purpose integer programming solver. By solving the problem n times for all $k \in \{1, 2, \ldots, n\}$ and selecting the best solution, we obtain the optimal solution on the set $\mathcal{W} = \cup_{k=1}^{n} \mathcal{W}_k$.

Algorithm 1 shows a natural algorithm for solving the forecast selection problem, as proposed by Matsypura et al. (2017), where the incumbent solution from a given k is only updated if the solution for that k is better. Matsypura et al. (2017) show that this problem can be solved to optimality for relatively high dimensions in reasonable time. The algorithm takes $\boldsymbol{\Sigma}$, an $n \times n$ positive-definite covariance matrix of forecast errors, as a single input argument, and returns a single output argument $\boldsymbol{w}^*$, an $n \times 1$ weight vector. Furthermore, the algorithm requires setting parameter ϵ to a sufficiently small constant. In our application, we set $\epsilon = 10^{-5}$. We implement the algorithm in `R` (R Core Team, 2017) using `Gurobi` (Gurobi Optimization Inc., 2017) as the solver.

Algorithm 1 Forecast selection algorithm

procedure ForecastSelection($\boldsymbol{\Sigma}$)
  $f^* \leftarrow \inf$ ▷ initialise best objective value to infinity
  for $k \in \{1, 2, \ldots, n\}$ **do** ▷ loop through all k
    min $\tilde{f}(\tilde{\boldsymbol{w}}) = k^{-2}\tilde{\boldsymbol{w}}'\boldsymbol{\Sigma}\tilde{\boldsymbol{w}}$ s.t. $\tilde{\boldsymbol{w}}'\boldsymbol{\iota} = k, \tilde{\boldsymbol{w}} \in \{0,1\}^n$ ▷ find solution to Problem P1
    if $f^* - \tilde{f}^* \geq \epsilon$ **then** ▷ check quality of current solution
      $f^* \leftarrow \tilde{f}^*$ ▷ update best objective if improvement
      $\boldsymbol{w}^* \leftarrow k^{-1}\tilde{\boldsymbol{w}}^*$ ▷ update best solution if improvement
    end if
  end for
  return $\boldsymbol{w}^*$ ▷ return optimal solution
end procedure

4 Data source

4.1 Description of the data

In this paper, we focus on survey data from the ECB SPF. The ECB has been conducting the SPF at a quarterly frequency since the launch of the euro currency in January 1999. The survey

participants are experts affiliated with financial and non-financial European institutions. They are asked to provide both point and density forecasts for several variables including rates of real GDP growth and unemployment in the euro area. As we aim to further investigate the results from Matsypura et al. (2017), we use data on these two variables over the entire period up to the first quarter of 2017. We focus on the one- and two-year-ahead forecast horizons and we obtain the data from `http://www.ecb.europa.eu/stats/prices/indic/forecast`[1]. For GDP growth, there are 71 and 67 observations for the one- and two-year-ahead horizon, respectively. For unemployment, there are 70 and 66 observations for the one- and two-year-ahead horizon, respectively, instead. While growth is observed quarterly, unemployment is observed on a monthly basis. Therefore, in the survey for the latter variable, forecasts are provided for February for Q1, May for Q2, August for Q3, and November for Q4.

For observations of the actual outcomes of each of the variables, we use data released by the ECB. However, the ECB alters the estimate of the outcomes through data revisions, with different releases referred to as vintages, which raises the question of which vintage to use. Genre et al. (2013) find that the relative performance of different combinations appears insensitive to the vintage used, and therefore we choose to use the most recent one. We obtain the Q1 2017 vintage for GDP growth and the February 2017 vintage for unemployment from `http://sdw.ecb.europa.eu`[2].

4.2 Dealing with non-response

Approximately 100 forecasters participate in the survey. However, as free entry and exit of forecasters in the panel is allowed, the SPF suffers from non-response. Since the forecast selection algorithm relies on historical data for estimation, it is necessary to filter out forecasters who respond infrequently or who only joined the panel recently. Following Matsypura et al. (2017), we choose to filter out all forecasters who fail to respond for 24 periods (6 years) or more, requiring a response rate of at least 75 percent, approximately. In most periods, this results in around 25 remaining forecasts for combination in each panel. These are plotted in Fig. 1. We can see that overall, the one-year-ahead forecasts appear to capture the actual outcomes better than the two-year-ahead forecasts. We can also see that the overall forecast accuracy varies throughout the years, with lower accuracy around major events that affected the EU economy, such as the early 2000s recession, the global financial crisis of 2007-2008 and the European debt from 2009 onwards. For summary statistics of each of the data sets after filtering, see Appendix A.

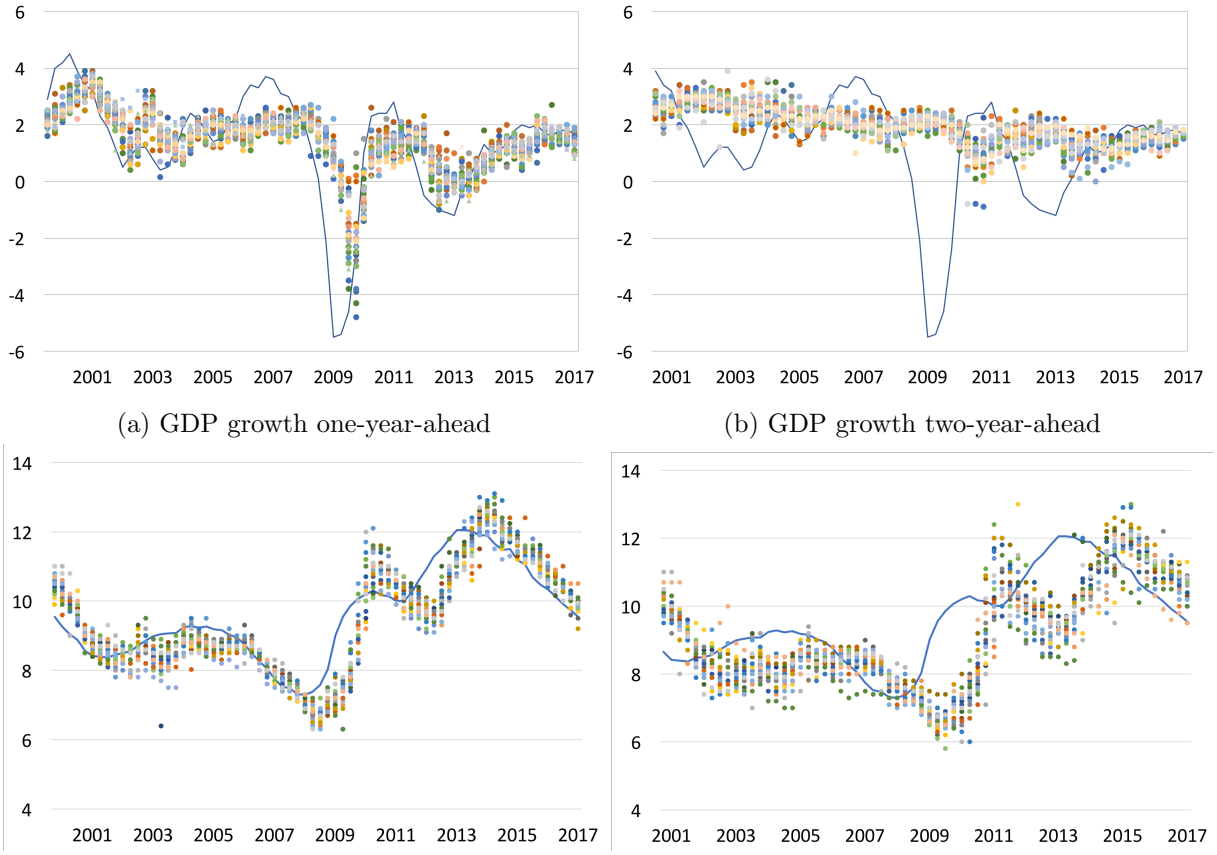
Since there are still missing observations post-filtering, the forecast selection algorithm that relies on an estimate of the covariance matrix of forecast errors is not straightforward to implement, because the standard sample covariance matrix requires a complete set of observations. Furthermore, it is ineffective to select forecasters that do not respond to the survey in the period for which we construct a forecast combination. Therefore, we resort to estimating the covariance matrix using pairwise observations for all forecasters that are available in the given period. That is, the matrix is constructed element-by-element as follows. Let $\mathcal{N}$ be the set of forecasters that respond in the given period and let $\mathcal{T}_i$ be the set of the periods in which the ith forecaster has responded to the survey, with $i \in \mathcal{N}$ and $\mathcal{T} \subseteq \{1, 2, \ldots, T\}$, where T is the period of the latest available observation. Additionally, let e_{it} denote the ith forecasters forecast error for the time period $t \in \{1, 2, ..., T\}$. A typical element of the covariance matrix is then computed as

$$\hat{\sigma}_{ij} = \begin{cases} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} e_{it}^2 & \text{if } i = j \\ \frac{1}{|\mathcal{T}_i \cap \mathcal{T}_j|} \sum_{t \in \mathcal{T}_i \cap \mathcal{T}_j} e_{it} e_{jt} & \text{if } i \neq j \\ 0 & \text{if } \mathcal{T}_i \cap \mathcal{T}_j = \emptyset. \end{cases}$$

[1]The data used in this paper was downloaded on May 14, 2017.
[2]The data used in this paper was downloaded on May 14, 2017.

The resulting matrix is not necessarily positive-definite, while all covariance matrices in the population are positive-definite, and therefore we use the `nearPD` function from `Matrix` (Bates and Maechler, 2017) in `R` to find the nearest valid covariance matrix.

Capistrán and Timmermann (2009) note that, by construction, an estimate of the covariance matrix may be biased in case entry and exit of forecasters is not at random, or when a pair of forecasters does not have any overlapping data. As previously described, a biased estimate of the covariance matrix may in turn result in biased forecasts through the forecast selection algorithm. In this paper, we perform pseudo-out-of-sample tests for forecast bias and we investigate whether an appropriate bias correction can be made to the out-of-sample forecasts. This procedure is further described in Sect. 5.1.



(a) GDP growth one-year-ahead

(b) GDP growth two-year-ahead

(c) Unemployment one-year-ahead

(d) Unemployment two-year-ahead

Figure 1: Individual quarterly forecasts against actual outcomes

5 Research method

Let $\mathbf{\Sigma}_{m,t}$ denote the $n \times n$ positive-definite covariance matrix of forecast errors based on an estimation window from period m to period t, and let $\boldsymbol{w}^*_{m,t}$ denote the corresponding vector of weights, obtained from Algorithm 1. Then the corresponding h-step ahead combination forecast is calculated as follows

$$\hat{y}^c_{m,t+h|t} = \boldsymbol{w}^*_{m,t}{}' \hat{\boldsymbol{y}}_{t+h}.$$

In order to evaluate the forecasting methods, we consider the most recent 16 periods (4 years) as the test data, representing approximately 25 percent of the available data. Denoting T_0 as the initial period in the testing set, this means $T_0 = T - 15$. For our data from the ECB SPF, the testing set is period 2013Q2 to period 2017Q1 for GDP growth, and period 2013May to period

2017Feb for unemployment. The remaining data forms the training set. Specifically, when we are forecasting period $t+h$, the training set consists of period 1 to period t. We define the estimation window as the part of the training set that we use to estimate the covariance matrix of forecast errors. In the case of an expanding window, this includes the complete training set. In the case of a rolling window this includes period $t-(T_0-h-1)$ to period t, with a fixed window size of T_0-h periods. Besides considering these commonly used estimation windows, we develop a set of methods to either select an estimation window or to combine different windows. In this section, we first describe the way we deal with forecast bias. We then describe the different methods used for window selection and window combination. Finally, we describe the criteria used to evaluate the different methods.

5.1 Bias correction

Due to data limitations, we investigate whether an appropriate bias correction can be made to the out-of-sample forecasts using only a single expanding window. In order to perform pseudo-out-of-sample tests for forecast bias prior to constructing out-of-sample forecasts, we reserve the last $\tilde{v}_1$ observations of the training data up to T_0. That is, periods $T_0-\tilde{v}_1$ to T_0-1. Using an expanding window including all remaining training data, we construct forecasts using the forecast selection algorithm for these periods. With the resulting set of forecasts, we first test the null hypothesis of no forecast bias, that is, $H_0: E(e^c_{t+h|t})=0$, by regressing the forecast errors on a constant as follows

$$e^c_{t+h|t} = \beta_0 + u_t.$$

In case the mean forecast error, β_0, is significantly different from zero, we can reject the null hypothesis of unbiased forecasts. Next, we estimate the Mincer and Zarnowitz (MZ) regression proposed by Mincer and Zarnowitz (1969) using the same set of forecasts, defined as

$$y_{t+h} = \beta_0 + \beta_1 \hat{y}^c_{t+h|t} + \eta_{t+h},$$

with the null hypothesis of unforecastable forecast errors, that is, $H_0: \beta_0 = 0, \beta_1 = 1$. If we can reject the null hypothesis of unbiased forecasts based on either of these tests, we adjust the out-of-sample forecasts in the testing set from T_0 to T as follows

$$\tilde{y}^c_{t+h|t} = \hat{\beta}_{0,t} + \hat{\beta}_{1,t}\hat{y}^c_{t+h|t}, \tag{1}$$

where we estimate the parameters $\beta_{0,t}$ and $\beta_{1,t}$ through least squares regression using an expanding window from $T_0-\tilde{v}_1$ to t.

Note that the choice for an expanding window with starting point $T_0-\tilde{v}_1$ is based on a trade-off between the number of observations used to construct forecasts $\hat{y}^c_{t+h|t}$ for the periods $T_0-\tilde{v}_1$ to T_0-1 with the forecast selection algorithm and the number of constructed forecasts, $\tilde{v}_1$, used to estimate the parameters $\beta_{0,t}$ and $\beta_{1,t}$ in Eq. 1.

5.2 Window selection

For the selection of an optimal window size, we use a cross-validation approach following a method proposed by Pesaran and Timmermann (2007). This approach reserves the last $\tilde{v}_2$ observations of the training data up until period t for a pseudo-out-of-sample estimation exercise and chooses the starting point m for the estimation window that generates the smallest root MSE (RMSE) value on this sample. The idea is that we choose the starting point in such a way that it would also give optimal out-of-sample forecasts in the RMSE sense. We assume that a minimum of $\underline{v}$ observations is needed to select forecasters using Algorithm 1, which means that window sizes smaller then $\underline{v}$ are not considered. Therefore, for h-step-ahead forecasts,

$\tilde{v}_2 + h + \underline{v}$ data points are required to adopt this method. For each potential starting point of the estimation window, m, the recursive pseudo-out-of-sample RMSE value is computed as

$$RMSE(m|t, \tilde{v}_2) = \sqrt{\tilde{v}_2^{-1} \sum_{\tau=t-\tilde{v}_2-h}^{t-h} (y_{m,\tau+h|\tau} - \hat{y}^c_{m,\tau+h|\tau})^2}. \tag{2}$$

The optimal starting point is then determined from

$$m^*(t, \tilde{v}_2, \underline{v}) = \underset{m=1,\ldots,t-\tilde{v}_2-h-\underline{v}}{\arg\min} RMSE(m|t, \tilde{v}_2), \tag{3}$$

with the corresponding forecast for period $t + h$ computed as

$$\hat{y}^c_{t+h|t}(m^*) = \hat{y}^c_{m^*,t+h|t}. \tag{4}$$

Here, the selection of $\underline{v}$ and $\tilde{v}_2$ is also based on a trade-off. If $\underline{v}$ is set too short, the forecast selection algorithm may not be robust against the potential influence of extreme forecast errors. Similarly, if $\tilde{v}_2$ is set too short, then the performance of the forecast selection algorithm relative to other combination methods may be affected too greatly by random variations. Alternatively, if $\tilde{v}_2$ is set too large, then starting points m with good performance earlier in the sample may more likely be selected than those that perform better closer to t.

5.3 Window combination

Since the size of our training set is relatively small, we develop two alternative methods to deal with uncertainty over the selection of $\underline{v}$ and $\tilde{v}_2$ in Eq. 3. Instead of selecting a single estimation window, we combine different estimation windows. We do this by giving each window size a weight proportional to the inverse of the associated out-of-sample RMSE-values from Eq. 2 raised to a power q. This approach builds on methods that are often seen in the forecast combination literature, with weights proportional to some measure of historical performance, for example in Stock and Watson (1998). Pesaran and Timmermann (2007) also note that combining different estimation windows instead of selecting a single one can be useful in case the breaks in the data are relatively small.

We use the resulting weights to construct forecasts in two different ways. The first method is more straightforward, with the weighted average forecast given by

$$\hat{y}^c_{t+h|t}(\underline{v}, \tilde{v}_2) = \frac{\sum_{m=1}^{t-\tilde{v}_2-h-\underline{v}} \hat{y}^c_{m,t+h|t} (RMSE(m|t, \tilde{v}_2))^{-q}}{\sum_{m=1}^{t-\tilde{v}_2-h-\underline{v}} (RMSE(m|t, \tilde{v}_2))^{-q}}. \tag{5}$$

Besides setting weights equal to the inverse of the MSE with $q = 2$, we also consider $q = 3$ and $q = 1$, allowing for more and less variable weights, respectively. Additionally, we consider $q = 0$, assigning equal weights to all estimation windows, a commonly used approach when combining different windows. This is equivalent to averaging over all $m \leq t - \underline{v} - \tilde{v}_2$:

$$\hat{y}^c_{t+h|t}(\underline{v}, \tilde{v}_2) = \frac{1}{t - \tilde{v}_2 - h - \underline{v}} \sum_{m=1}^{t-\tilde{v}_2-h-\underline{v}} \hat{y}^c_{m,t+h|t}. \tag{6}$$

For the second method, we average over the covariance matrices estimated using the different windows and use the resulting covariance matrix to select forecasts for combination. That is, the weighted average covariance matrix is calulated as follows

$$\mathbf{\Sigma}^c(t, \underline{v}, \tilde{v}_2) = \frac{\sum_{m=1}^{t-\tilde{v}_2-h-\underline{v}} \mathbf{\Sigma}_{m,t}(RMSE(m|t, \tilde{v}_2))^{-q}}{\sum_{m=1}^{t-\tilde{v}_2-h-\underline{v}} (RMSE(m|t, \tilde{v}_2))^{-q}}, \tag{7}$$

with corresponding weight vector $\boldsymbol{w}^{c*}$ from the forecast selection algorithm. Again, we consider $q = 0, 1, 2, 3$, and the corresponding forecast for period $t + h$ is computed as

$$\hat{y}^c_{t+h|t}(\mathbf{\Sigma}^c) = \boldsymbol{w}^{c*\prime}\hat{\boldsymbol{y}}_{t+h}. \tag{8}$$

5.4 Evaluation criteria

While we are mainly interested in finding optimal forecasts in terms of RMSE, another commonly used measure of forecast accuracy in the forecasting literature is the mean absolute error (MAE). Therefore, in order to evaluate the out-of-sample performance of the different methods, we consider both square and absolute loss using RMSE and MAE, respectively. The forecast error of an h-step ahead forecast is defined as $e^c_{t+h|t} = \hat{y}^c_{t+h|t} - y_{t+h}$. Then the RMSE and MAE are defined as follows.

$$RMSE_c = \sqrt{\frac{1}{T - T_0} \sum_{t=T_0-h}^{T-h} e^2_{t+h}}$$

$$MAE_c = \frac{1}{T - T_0} \sum_{t=T_0-h}^{T-h} |e^c_{t+h}|$$

We observe these measures of loss relative to simple averaging over the filtered panel, which we use as the benchmark, and therefore, we define the relative RMSE and the relative MAE as follows.

$$Rel.\, RMSE_c = \frac{RMSE_c}{RMSE_{avg(filt)}}$$

$$Rel.\, MAE_c = \frac{MAE_c}{MAE_{avg(filt)}}$$

A value below one indicates higher forecasting accuracy relative to simple averaging, and a value greater than one indicates lower forecasting accuracy relative to simple averaging.

In addition to these measures, we analyse the forecasting performance using the Diebold-Mariano (DM) test proposed by Diebold and Mariano (1995). For the DM test, we define the loss differential between the two forecasts as $d_{t+h} = (e^c_{t+h|t})^2 - (e^{avg(filt)}_{t+h})^2$, since we are interested in minimizing the RMSE. The two forecasts have equal accuracy if and only if the loss differential has zero expectation for all t. That is, we test $H_0 : E(d_{t+h}) = 0\,\forall t$ against $H_a : E(d_{t+h}) \neq 0$. The DM test statistic is given by

$$DM = \frac{\overline{d}}{\sqrt{\frac{1}{T-T_0} 2\pi \hat{f}_d(0)}} \sim N(0, 1),$$

where $\overline{d} = \frac{1}{T-T_0+1} \sum_{t=T_0-h}^{T-h} d_{t+h|t}$ is the sample mean loss differential, and $\hat{f}_d(0)$ is a consistent estimate of $f_d(0)$ defined as

$$\hat{f}_d(0) = \frac{1}{2\pi} \sum_{l=-(T-T_0)}^{T-T_0} I\left(\frac{l}{h-1}\right) \hat{\gamma}_d(l),$$

with

$$I\left(\frac{l}{h-1}\right) = \begin{cases} 1 & \text{for } |\frac{k}{h-1}| \leq 1 \\ 0 & \text{otherwise} \end{cases}.$$

and

$$\hat{\gamma}_d(l) = \frac{1}{T - T_0 + 1} \sum_{t=|l|+T_0-h}^{T-h} (d_{t+h} - \overline{d})(d_{t+h-|l|} - \overline{d}).$$

As the size of our testing set is relatively small, we make a bias correction to the DM test statistic as suggested by Harvey et al. (1997). The corrected statistic is defined as

$$DM^* = \sqrt{\frac{T - T_0 + 2 - 2h + (T - T_0 + 1)^{-1} h(h-1)}{T - T_0 + 1}} DM,$$

and follows a Student's t-distribution with $T - T_0$ degrees of freedom instead of the standard normal distribution.

6 Results

In Table 1, the p-values of the tests for forecast bias using $\tilde{v}_1 = 32$ are presented. Based on these tests, we can conclude for both the one-year-ahead and the two-year-ahead forecasts of unemployment rate that there is a significant indication of forecast bias. Therefore, we apply a bias correction to the expanding window unemployment forecasts according to Eq. 1. The resulting forecasts using $\tilde{v}_1 = 32$ are presented in Table 2 along with the corresponding forecasts without bias adjustment. We can see that for one-year-ahead unemployment forecasts the bias adjustment improves the results, but this is not the case for two-year-ahead forecasts. The difference is possibly the consequence of the larger horizon limiting the number of observations available for the estimation of the bias. This suggests that the bias adjustment can only be applied if the sample size and the forecast horizon allow for it. For a more precise evaluation of the different window selection and combination methods, we leave out the bias adjustment in the rest of this section.

Table 1: p-values of tests for forecast bias

	Regression on constant	MZ regression
GDP growth one-year-ahead	0.2941	0.3213
GDP growth two-year-ahead	0.0620	0.0512
Unemployment one-year-ahead	0.0999	0.0322
Unemployment two-year-ahead	0.0294	0.0004

Tests are based on forecasts computed over the $\tilde{v}_1 = 32$ periods from Q2 2005 to Q1 2013 for GDP growth and from May 2005 to February 2013 for unemployment using an expanding window. A value below 0.05 implies significant indication of forecast bias.

In Fig. 2, for all four data sets, the pseudo-out-of-sample RMSE values obtained with an expanding window are plotted for different starting points. The starting points considered are based on $\tilde{v}_2 = 16$ and $\underline{v} = 8$. While less starting points are considered when a rolling window is used, the pseudo-out-of-sample RMSE values for the individual starting points are the same (see Appendix B). Considering all graphs are scaled identically in Fig. 2, we can clearly see two important features. First of all, it appears that across all four data sets, the variation in forecasting accuracy for different starting points is quite limited. Especially for the majority of the two-year-ahead GDP growth forecasts and the one-year-ahead unemployment forecasts, the forecasting accuracy appears to be relatively stable across different starting points. This

suggests that breaks in the data are not well-defined, and an expanding window or combining different estimation windows is more appropriate than selecting starting points using cross-validation.

A second feature is that the RMSE is higher for earlier sets of pseudo-out-of-sample periods used to evaluate the forecasts, and this is more pronounced for GDP growth than for unemployment. This implies that for GDP growth, there is more variation in forecast accuracy of the individual survey forecasts among different sets of pseudo-out-of-sample periods. Moreover, for all two-year-ahead forecasts the RMSE is higher than for one-year-ahead forecasts, suggesting that there is more uncertainty due to the longer forecast horizon. Looking at the data, we find that the earliest set of pseudo-out-of-sample periods used to evaluate the forecasts is from August 2008 to May 2012 for one-year-ahead forecasts and from August 2007 to May 2011 for two-year-ahead forecasts, whereas the latest set of pseudo-out-of-sample periods is from May 2012 to February 2016 for one-year-ahead forecasts and from May 2011 to February 2015 for two-year-ahead forecasts. In Fig. 1 we could see that all forecasters performed poorly throughout the financial crisis of 2007-2008, especially for two-year-ahead GDP growth. This explains the higher RMSE for earlier sets of pseudo-out-of-sample periods.

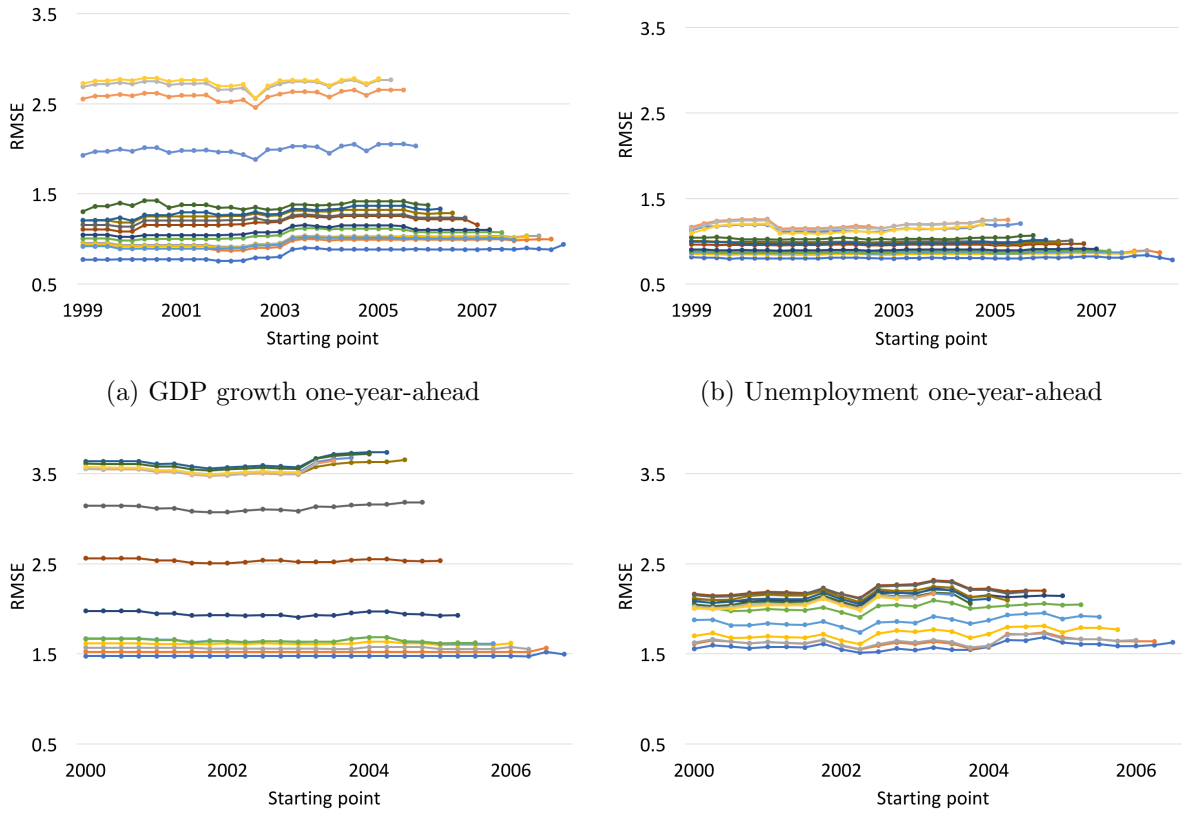
Table 2: Results of bias-adjusted unemployment forecasts using an expanding window

One-year-ahead forecast horizon		
	Rel. RMSE	Rel. MAE
Expanding window	1.0840	1.0615
Expanding window (bias-adjusted)	1.0350	0.8960
Simple averaging (filtered): 1-y.a. RMSE = 0.4972%, 1-y.a. MAE = 0.4769%		
Two-year-ahead forecast horizon		
	Rel. RMSE	Rel. MAE
Expanding window	1.0344	1.0527
Expanding window (bias-adjusted)	1.2241	1.1388
Simple averaging (filtered): 2-y.a. RMSE = 1.0352%, 2-y.a. MAE = 1.2312%		

Forecasts are computed over the 16 periods from from May 2013 to February 2017 using an expanding window and the bias adjustment is based on $\tilde{v}_1 = 32$. Relative RMSE and MAE are reported with simple averaging (filtered) as the benchmark. A value below one indicates superior accuracy to the benchmark, and a value greater than one indicates inferior accuracy. The RMSE and MAE for simple averaging (filtered) are reported at the bottom.

In Table 3, we can see exactly for each out-of-sample period, which starting point is selected based on the RMSE value for the pseudo-out-of-sample forecasts, both in the case of an expanding window and a rolling window. With an expanding window, we can see that across all data sets, leaving out the earliest few observations in the data for estimation appears to give a lower RMSE value in most of the cases. Particularly, for two-year-ahead unemployment forecasts, February 2003 is consistently selected as the optimal starting point. Selecting starting points using cross-validation with an expanding window for this data set is therefore equivalent to directly using an expanding window with February 2003 as the starting point. This pattern could be explained by the fact that the forecast accuracy is more volatile throughout the early 2000s recession, as we could see in Fig. 1. The results for the other three data sets are more variable, suggesting that the selection may be affected by minor variation of the RMSE value. That is, a starting point could be selected based on a pseudo-out-of-sample RMSE value that is only slightly lower than the ones given by other starting points. This would result in a situation where the selected starting point is less likely also the optimal starting point for the out-of-sample forecast. Especially for two-year-ahead GDP growth forecasts, the selected starting points are quite different from each other. Hence, in line with the interpretation of Fig. 2, this suggests that selecting a single starting point using cross-validation is not appropriate for this data set.

With a rolling window, by construction, several of the starting points that are selected when an expanding window is used, are no longer considered. This is especially the case for one-year-ahead GDP growth forecasts, where most of the time the earliest considered starting point is selected instead. For two-year-ahead unemployment forecasts, we can see that when February 2003 is no longer considered as a starting point, the selection is shifted to the fall of 2004. This suggests that the use of a rolling window is not appropriate for these two data sets as valuable information may be omitted, not only when selecting a single estimation window, but also when combining different estimation windows. For two-year-ahead GDP growth forecasts the selection is not different when a rolling window is used instead, and for one-year-ahead unemployment forecasts the difference is minor. However, this pattern may appear by chance, as we speculate that the selection for these two data sets is not accurate. Therefore, it is not necessarily the case that a rolling window is more appropriate.



(a) GDP growth one-year-ahead

(b) Unemployment one-year-ahead

(c) GDP growth two-year-ahead

(d) Unemployment two-year-ahead

Forecasts are computed pseudo-out-of-sample for $\tilde{v}_2 = 16$ periods using estimation windows based on $\underline{v} = 8$, and evaluated following Eq. 2. The shortest line corresponds to the earliest set of $\tilde{v}_2 = 16$ pseudo-out-of-sample periods with a smaller number of feasible starting points and the longest line corresponds to the latest set of $\tilde{v}_2 = 16$ pseudo-out-of-sample periods with a larger number of starting points.

Figure 2: Pseudo-out-of-sample RMSE for different starting points using an expanding window

Table 3: Selected starting points using cross-validation

One-year-ahead forecast horizon						
Out-of-sample	Pseudo-out-of-sample		GDP growth		Unemployment	
	From	To	Expanding	Rolling	Expanding	Rolling
Feb 2017	May 2012	Feb 2016	Aug 2002	May 2003	May 2009	May 2009
Nov 2016	Feb 2012	Nov 2015	Aug 2002	Feb 2003	Aug 2000	Nov 2003
Aug 2016	Nov 2011	Aug 2015	Aug 2002	Nov 2002	Aug 2003	Aug 2003
May 2016	Aug 2011	May 2015	May 2002	Aug 2002	Aug 2003	Aug 2003
Feb 2016	May 2011	Feb 2015	May 2000	May 2002	May 2003	May 2003
Nov 2015	Feb 2011	Nov 2014	May 2000	Feb 2002	May 2003	May 2003
Aug 2015	Nov 2010	Aug 2014	May 2000	Nov 2001	May 2003	May 2003
May 2015	Aug 2010	May 2014	May 2000	Aug 2001	May 2003	May 2003
Feb 2015	May 2010	Feb 2014	May 2000	May 2003	May 2003	May 2003
Nov 2014	Feb 2010	Nov 2013	May 2000	Feb 2001	May 2003	May 2003
Aug 2014	Nov 2009	Aug 2013	Aug 2000	Nov 2000	May 2003	May 2003
May 2014	Aug 2009	May 2013	Aug 1999	May 2003	Aug 2001	Aug 2001
Feb 2014	May 2009	Feb 2013	Feb 2003	Feb 2003	Aug 2001	Aug 2001
Nov 2013	Feb 2009	Nov 2012	Feb 2003	Feb 2003	Aug 2001	Aug 2001
Aug 2013	Nov 2008	Aug 2012	Feb 2003	Feb 2003	Aug 2001	Aug 2001
May 2013	Aug 2008	May 2012	Feb 2003	Feb 2003	May 2002	May 2002
Two-year-ahead forecast horizon						
Out-of-sample	Pseudo-out-of-sample		GDP growth		Unemployment	
	From	To	Expanding	Rolling	Expanding	Rolling
Feb 2017	May 2011	Feb 2015	Nov 2006	Nov 2006	Feb 2003	Nov 2004
Nov 2016	Feb 2011	Nov 2014	Nov 2006	Nov 2006	Feb 2003	Aug 2004
Aug 2016	Nov 2010	Aug 2014	Nov 2006	Nov 2006	Feb 2003	Aug 2004
May 2016	Aug 2010	May 2014	Aug 2005	Aug 2005	Feb 2003	Aug 2004
Feb 2016	May 2010	Feb 2014	Aug 2005	Aug 2005	Feb 2003	Aug 2004
Nov 2015	Feb 2010	Nov 2013	Aug 2005	Aug 2005	Feb 2003	Aug 2004
Aug 2015	Nov 2009	Aug 2013	Aug 2003	Aug 2003	Feb 2003	Feb 2003
May 2015	Aug 2009	May 2013	Aug 2002	Aug 2002	Feb 2003	Feb 2003
Feb 2015	May 2009	Feb 2013	Aug 2002	Aug 2002	Feb 2003	Feb 2003
Nov 2014	Feb 2009	Nov 2012	Aug 2002	Aug 2002	Feb 2003	Feb 2003
Aug 2014	Nov 2008	Aug 2012	May 2002	May 2002	Feb 2003	Feb 2003
May 2014	Aug 2008	May 2012	May 2002	May 2002	Feb 2003	Feb 2003
Feb 2014	May 2008	Feb 2012	May 2002	May 2002	Feb 2003	Feb 2003
Nov 2013	Feb 2008	Nov 2011	May 2002	May 2002	Feb 2003	Feb 2003
Aug 2013	Nov 2007	Aug 2011	May 2002	May 2002	Feb 2003	Feb 2003
May 2013	Aug 2007	May 2011	May 2002	May 2002	Feb 2003	Feb 2003

For each out-of-sample period, the starting point that gives the lowest RMSE value for the pseudo-out-sample forecasts is selected. The different starting points are based on $\tilde{v}_2 = 16$ and $\underline{v} = 8$.

The forecasts for GDP growth and unemployment using different estimation windows are presented in Table 5 and Table 6, respectively. We find that the use of a weighted average of the covariance matrices for different estimation windows does not result in a difference in forecasts selected for $q = 0, 1, 2, 3$ in Eq. 7. This implies that a minor variation in the estimated covariance matrix does not affect the forecast selection algorithm, and therefore we also report the results for $q = 10, 20$. Each method considered is ranked by its performance under both RMSE and MAE. For GDP growth, almost all different window types result in improved forecasts relative to simple averaging, in terms of RMSE and MAE. In the best cases, the forecast selection algorithm yields RMSE and MAE improvements over the benchmark of 13.63 and 10.65 percent, respectively, for one-year-ahead forecasts, and even 21.65 and 16.76 percent, respectively, for two-year-ahead forecasts. However, for unemployment, the opposite holds. Even in the best cases, the algorithm yields RMSE and MAE loss over the benchmark of 4.94 and 1.78 percent, respectively, for one-year-ahead forecasts, and 1.66 and 1.00 percent, respectively, for two-year-ahead forecasts. These results are in line with the outcomes of the tests for forecast bias in Table 1. That is, unbiased GDP growth forecasts give improvements in accuracy over simple averaging, whereas biased unemployment forecasts give relatively lower accuracy.

An important result is that compared to Matsypura et al. (2017), our findings are very different. Whereas Matsypura et al. (2017) find improvements in accuracy over simple averaging across all data sets using an expanding window, we only find improvements for GDP growth forecasts. This difference is caused by a minor change in the data used, as we included observations that recently became available and our expanding window therefore also includes a few additional observations. This already indicates that the forecast selection algorithm is not robust to the estimation window used. Additionally, looking at the DM test statistics adjusted for sample size in Table 4, we can see that not a single value has a absolute value larger than the critical value 2.131. This means that there is no significant indication that any method performs better or worse compared to simple averaging, despite the differences in RMSE and MAE. However, the RMSE and MAE values do give us a good indication of the performance of the forecast selection algorithm for different estimation windows. Therefore, in the rest of this section, we look at these values in Table 5 and Table 6 to evaluate the different methods for each data set, in order to further understand the performance of the forecast selection algorithm.

Table 4: DM test statistics for all forecasts

	GDP growth		Unemployment	
	1-y.a.	2-y.a.	1-y.a.	2-y.a.
Expanding window	0.4964	0.5629	-1.0053	-0.5329
Rolling window	-0.0190	0.5421	-1.0249	-0.5366
Cross-validation exp.	0.1532	0.5421	-0.7087	-0.2083
Cross-validation rol.	-0.1359	0.5421	-0.7375	-0.7806
Weighted forecast exp. q=0	0.8964	0.5490	-0.9402	-0.6643
Weighted forecast exp. q=1	0.8962	0.5490	-0.9397	-0.6638
Weighted forecast exp. q=2	0.8956	0.5490	-0.9393	-0.6632
Weighted forecast exp. q=3	0.8945	0.5490	-0.9389	-0.6626
Weighted forecast rol. q=0	0.8585	0.5451	-0.9337	-0.6948
Weighted forecast rol. q=1	0.8573	0.5451	-0.9330	-0.6948
Weighted forecast rol. q=2	0.8561	0.5451	-0.9323	-0.6948
Weighted forecast rol. q=3	0.8547	0.5452	-0.9316	-0.6947
Weighted cov.mat exp. q=0,1,2,3	0.8638	0.5421	-1.0492	-0.1703
Weighted cov.mat exp. q=10	0.9386	0.5421	-1.0492	-0.1703
Weighted cov.mat exp. q=20	1.0405	0.5421	-1.0253	-0.1703
Weighted cov.mat rol q=0,1,2,3	0.8638	0.5421	-0.9102	-0.5920
Weighted cov.mat rol. q=10,20	0.8638	0.5421	-0.9102	-0.5533

Forecasts are computed over the 16 periods from from May 2013 to February 2017 using different estimation windows with $\tilde{v}_2 = 16$ and $\underline{v} = 8$. A value below the critical value of -2.131 indicates significant worse forecasting performance than simple averaging, and a value greater than the critical value of 2.131 indicates significant better forecasting performance than simple averaging.

As expected, there is not a single method that consistently outperforms other methods across all four data sets. For one-year-ahead GDP growth forecasts, we can see that combining different estimation windows gives higher accuracy than using a single estimation window. This is in line with Fig 2a, where we could see that while there is variation in the pseudo-out-of-sample RMSE values, in most of the cases there is not a single starting point that clearly gives a lower value. However, there is a difference between using a single expanding window and using a single rolling window. Whereas using a single expanding window, either by including all available data or by selecting starting points with cross-validation, gives an improvement in accuracy over simple averaging in terms of RMSE, using a single rolling window yields relative RMSE loss. In line with the results in Table 3, we thus find that a rolling window is not appropriate. An interesting feature is that while combining different estimation windows in general appears to work well, using a weighted average covariance matrix gives the highest accuracy. This suggests that for this data set, the individual covariance matrices can be estimated relatively accurately, and

combining covariance matrices can give a further improvement in accuracy. Considering the outcomes of the tests for forecast bias in Table 1, this is not a surprising result. Overall, the results imply that for this data set, while with most of the methods there is an improvement in forecast accuracy over simple averaging, the performance of the forecast selection algorithm still varies with the estimation window used.

In agreement with our expectations based on Fig. 2c and Table 3, we find that for two-year-ahead GDP growth forecasts, using an expanding window gives both the lowest RMSE value and the lowest MAE value. However, the different methods result in relative RMSE and relative MAE values that are all quite close to each other, with biggest differences of only 0.70 and 3.78 percent, respectively. Quite frequently, the same forecasts are selected for each out-of-sample period, indicating that there are forecasters in the panel that consistently provide relatively more accurate forecasts. These results imply that for this data set, the forecast selection algorithm is relatively robust to the estimation window used.

Table 5: Results of real GDP growth forecasts using different estimation windows

One-year-ahead forecast horizon			
	Rel. RMSE		Rel. MAE
Weighted cov.mat. exp. q=20	0.8637	Weighted cov.mat. exp. q=20	0.8935
Weighted cov.mat. exp. q=10	0.8661	Weighted cov.mat. exp. q=10	0.9073
Weighted cov.mat. exp.&rol. q=0,1,2,3	0.8854	Weighted cov.mat. exp.&rol. q=0,1,2,3	0.9451
Weighted cov.mat. rol. q=10,20	0.8854	Weighted cov.mat. rol. q=10,20	0.9451
Weighted forecast rol. q=0	0.8950	**Expanding window**	0.9470
Weighted forecast rol. q=1	0.8956	Weighted forecast rol. q=0	0.9504
Weighted forecast rol. q=2	0.8962	Weighted forecast rol. q=1	0.9509
Weighted forecast rol. q=3	0.8968	Weighted forecast rol. q=2	0.9515
Weighted forecast exp. q=0	0.9042	Weighted forecast exp. q=0	0.9518
Weighted forecast exp. q=1	0.9053	Weighted forecast rol. q=3	0.9521
Weighted forecast exp. q=2	0.9065	Weighted forecast exp. q=1	0.9522
Weighted forecast exp. q=3	0.9077	Weighted forecast exp. q=2	0.9525
Expanding window	0.9510	Weighted forecast exp. q=3	0.9529
Cross-validation exp.	0.9879	Cross-validation exp.	1.0092
Rolling window	1.0023	Rolling window	1.0573
Cross-validation rol.	1.0137	Cross-validation rol.	1.0711
Simple averaging (filtered): 1-y.a. RMSE = 0.4423%, 1-y.a. MAE = 0.3313%			
Two-year-ahead forecast horizon			
	Rel. RMSE		Rel. MAE
Expanding window	0.7835	**Expanding window**	0.8324
Weighted forecast exp. q=0,1,2,3	0.7882	Weighted forecast exp. q=0,1,2,3	0.8632
Weighted forecast rol. q=0,1,2,3	0.7895	Weighted forecast rol. q=0,1,2,3	0.8673
Rolling window	0.7905	Rolling window	0.8702
Cross-validation exp.&rol.	0.7905	Cross-validation exp.&rol.	0.8702
Weighted cov.mat. exp.&rol. q=0,1,2,3,10,20	0.7905	Weighted cov.mat. exp.&rol. q=0,1,2,3,10,20	0.8702
Simple averaging (filtered): 2-y.a. RMSE = 0.6073%, 2-y.a. MAE = 0.4159%			

Forecasts are computed over the 16 periods from Q2 2013 to Q1 2017 using different estimation windows with $\tilde{v}_2 = 16$ and $\underline{v} = 8$. Relative RMSE and MAE are reported. The benchmark is simple averaging (filtered), and the results are ranked. A value below one indicates higher forecasting accuracy relative to simple averaging, and a value greater than one indicates lower forecasting accuracy relative to simple averaging. The RMSE and MAE for simple averaging (filtered) are reported at the bottom.

Somewhat surprisingly, for one-year-ahead unemployment forecasts, we find that the cross-validation methods give the lowest RMSE and MAE values. Looking at the graph in Fig. 2b, we can see that for the earliest four out-of-sample periods, the earliest few starting points appear to give higher pseudo-out-of-sample RMSE values. The cross-validation methods therefore do not select these starting points, resulting in a small gain in accuracy compared to other methods. Accordingly, the use of a single expanding or rolling window gives the lowest accuracy, as for both of these methods the concerning starting points for the earliest four out-of-sample periods are still used. While the window combination methods also use these starting points, the differences between different starting points are accounted for. This explains why combining estimation windows results in slightly higher accuracy compared to the use of a single expanding or rolling window. Overall, we can see that for this data set, the performance of the forecast selection

algorithm only varies a little with the estimation window used. However, improvements in forecast accuracy can be made by taking a good look at the pseudo-out-of-sample performance.

For two-year-ahead unemployment forecasts, we can see that in line with our interpretation of Table 3, the cross-validation method using an expanding window has relatively good performance. For this data set, this method is equivalent to using an expanding window with starting point February 2003 instead of November 2000. However, the difference with using November 2000 as the starting point is only 1.78 percent, indicating that the difference in estimation accuracy is quite small. The weighted average covariance matrix using an expanding window, with the same forecasts selected for all q considered, also has relatively good performance. This suggests that this method perhaps removes part of the bias in the estimated covariance matrix, considering the outcomes of the tests for forecast bias in Table 1. This would also explain the difference in performance between this method and the other window combination methods. Overall, the performance of the forecast selection algorithm varies with the estimation window used. Also for this data set, we find that improvements in forecast accuracy can be made by taking pseudo-out-of-sample performance into account.

Table 6: Results of unemployment forecasts using different estimation windows

One-year-ahead forecast horizon			
	Rel. RMSE		Rel. MAE
Cross-validation exp.	1.0494	Cross-validation exp.	1.0178
Cross-validation rol.	1.0512	Cross-validation rol.	1.0211
Weighted cov.mat. rol. q=0,1,2,3,10,20	1.0673	Weighted cov.mat. rol. q=0,1,2,3,10,20	1.0495
Weighted forecast rol. q=3	1.0716	Weighted forecast rol. q=3	1.0566
Weighted forecast rol. q=2	1.0717	Weighted forecast rol. q=2	1.0568
Weighted forecast rol. q=1	1.0718	Weighted forecast exp. q=3	1.0568
Weighted forecast rol. q=0	1.0719	Weighted forecast rol. q=1	1.0569
Weighted forecast exp. q=3	1.0720	Weighted forecast exp. q=2	1.0570
Weighted forecast exp. q=2	1.0721	Weighted forecast rol. q=0	1.0571
Weighted forecast exp. q=1	1.0722	Weighted forecast exp. q=1	1.0571
Weighted forecast exp. q=0	1.0722	Weighted forecast exp. q=0	1.0573
Weighted cov.mat. exp. q=20	1.0778	Weighted cov.mat. exp. q=20	1.0581
Weighted cov.mat. exp. q=0,1,2,3,10	1.0806	**Expanding window**	1.0615
Expanding window	1.0840	Weighted cov.mat. exp. q=0,1,2,3,10	1.0626
Rolling window	1.0857	Rolling window	1.0647
Simple averaging (filtered): 1-y.a. RMSE = 0.4972%, 1-y.a. MAE = 0.4769%			
Two-year-ahead forecast horizon			
	Rel. RMSE		Rel. MAE
Cross-validation exp.	1.0166	Weighted cov.mat. exp. q=0,1,2,3,10,20	1.0100
Weighted cov.mat. exp. q=0,1,2,3,10,20	1.0130	Cross-validation exp.	1.0144
Expanding window	1.0344	**Expanding window**	1.0527
Weighted cov.mat. rol. q=10,20	1.0424	Cross-validation rol.	1.0562
Cross-validation rol.	1.0435	Weighted cov.mat. rol. q=10,20	1.0617
Weighted forecast exp. q=0	1.0462	Weighted forecast exp. q=0	1.0688
Weighted forecast exp. q=1	1.0463	Weighted forecast exp. q=1	1.0688
Weighted forecast exp. q=2	1.0463	Weighted forecast exp. q=2	1.0688
Weighted forecast exp. q=3	1.0464	Weighted forecast exp. q=3	1.0688
Weighted cov.mat. rol. q=0,1,2,3	1.0491	Weighted forecast rol. q=0	1.0717
Weighted forecast rol. q=0	1.0514	Weighted forecast rol. q=1	1.0718
Weighted forecast rol. q=1	1.0516	Weighted forecast rol. q=2	1.0720
Weighted forecast rol. q=2	1.0517	Weighted forecast rol. q=3	1.0721
Weighted forecast rol. q=3	1.0518	Rolling window	1.0753
Rolling window	1.0658	Weighted cov.mat. rol. q=0,1,2,3	1.0799
Simple averaging (filtered): 2-y.a. RMSE = 1.0352%, 2-y.a. MAE = 1.2312%			

Forecasts are computed over the 16 periods from from May 2013 to February 2017 using different estimation windows with $\tilde{v}_2 = 16$ and $\underline{v} = 8$. Relative RMSE and MAE are reported. The benchmark is simple averaging (filtered), and the results are ranked. A value below one indicates superior accuracy to the benchmark, and a value greater than one indicates inferior accuracy. The RMSE and MAE for simple averaging (filtered) are reported at the bottom.

7 Conclusion

In this paper, we investigated the robustness of the forecast selection algorithm proposed by Matsypura et al. (2017), with a focus on the use of different estimation windows. Matsypura et al. (2017) propose an algorithm in which integer programming is used to select forecasts for averaging, given an estimated covariance matrix, instead of averaging over every available forecast. While numerous empirical studies on forecast combination methods fail to consistently outperform simple averaging, the authors conclude based on an application to ECB SPF data, that this particular method does succeed in doing so. A quick look at the methodology raises doubts about the robustness of the method, as results are reported only for a single expanding window including all available data.

Using ECB SPF data on rates of GDP growth and unemployment, both one-year-ahead and two-year-ahead, we applied the forecast selection algorithm in combination with different estimation windows. Besides considering a single expanding and a single rolling window, we used pseudo-out-of-sample cross-validation to determine optimal window size. We also combined different estimation windows using both equal weights and weights based on the pseudo-out-of-sample performance. For most of the data sets, the performance of the forecast selection algorithm varied with the estimation window used. We find that the robustness of the algorithm for a specific data set can be investigated prior to forecasting out-of-sample. Accordingly, improvements in forecast accuracy can be made by selecting the estimation window based on pseudo-out-of-sample performance.

Furthermore, we find that the overall performance of the forecast selection algorithm is not consistent across different panels and different horizons. This is in line with previous forecasting literature, and the pattern can easily be explained. Due to free exit and entry of forecasters in the survey panel, there is the issue of missing data, resulting in the need for omitting data and even after filtering, a poor estimation of the covariance matrix. As the forecast selection algorithm relies on an estimate of the full covariance matrix of forecast errors, a bias in this matrix can lead to biased forecasts constructed. The forecasts that did not appear to be biased based on pseudo-out-of-sample analysis, had big improvements in out-of-sample forecast accuracy relative to simple averaging, in terms of RMSE and MAE. The forecasts that did appear to be biased had relatively worse out-of-sample performance. However, we find that if the sample size and the forecast horizon allow for it, a simple bias-adjustment can be applied.

The methods developed in this paper can be extended to other forecast combination methods. The selection of the estimation window used is often overlooked, and usually only results for a single estimation window are reported. The findings of our paper show the importance of choice of estimation window in the presence of data instability, and that improvements in forecast accuracy can be made by taking a careful look at pseudo-out-of-sample performance. In future work, the issue of bias in the estimated covariance matrix should be further investigated. Especially when the matrix is used for forecast combination weights estimation, a bias could have a large impact on the outcomes.

References

Assenmacher-Wesche, K. and Pesaran, M. H. (2008). Forecasting the swiss economy using vecx models: An exercise in forecast combination across models and observation windows. *National Institute Economic Review*, 203(1):91–108.

Bates, D. and Maechler, M. (2017). Matrix: sparse and dense matrix classes and methods. *R package version 1.2-10, URL http://cran. r-project. org/package= Matrix*.

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468.

Capistrán, C. and Timmermann, A. (2009). Forecast combination with entry and exit of experts. *Journal of Business & Economic Statistics*, 27(4):428–440.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

Clark, T. E. and McCracken, M. W. (2009). Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, 50(2):363–395.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, pages 253–263.

Genre, V., Kenny, G., Meyler, A., and Timmermann, A. (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, 29(1):108–121.

Gurobi Optimization Inc. (2017). *Gurobi Optimizer Reference Manual*.

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2):281–291.

Matsypura, D., Thompson, R., and Vasnev, A. (2017). Optimal selection of expert forecasts with integer programming.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER.

Pesaran, M. H. and Pick, A. (2011). Forecast combination across estimation windows. *Journal of Business & Economic Statistics*, 29(2):307–318.

Pesaran, M. H., Schuermann, T., and Smith, L. V. (2009). Forecasting economic and financial variables with global vars. *International Journal of Forecasting*, 25(4):642–675.

Pesaran, M. H. and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, 137(1):134–161.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rossi, B. and Inoue, A. (2012). Out-of-sample forecast tests robust to the choice of window size. *Journal of Business & Economic Statistics*, 30(3):432–453.

Stock, J. H. and Watson, M. W. (1998). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. Technical report, National Bureau of Economic Research.

Stock, J. H. and Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, 1:135–196.

Appendix

A Summary statistics of the filtered ECB SPF data

Table 1: Summary statistics of filtered ECB SPF data on GDP growth one-year-ahead

Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.	Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.
1999Q3	24	2.10	0.06	1.60	2.50	2.90	2008Q3	27	1.73	0.13	0.90	2.60	0.10
1999Q4	27	2.24	0.11	1.70	3.10	4.00	2008Q4	26	1.33	0.12	0.60	1.90	-2.10
2000Q1	23	2.48	0.08	2.10	3.30	4.20	2009Q1	25	0.92	0.22	0.10	2.20	-5.50
2000Q2	25	2.88	0.06	2.40	3.40	4.50	2009Q2	27	-0.04	0.19	-1.00	0.70	-5.40
2000Q3	26	3.05	0.08	2.20	3.70	3.90	2009Q3	28	-1.93	1.04	-3.80	0.10	-4.60
2000Q4	27	3.26	0.10	2.50	3.90	3.40	2009Q4	28	-2.13	1.78	-4.80	1.00	-2.40
2001Q1	23	3.39	0.07	3.00	3.90	3.20	2010Q1	27	-0.44	0.40	-1.40	1.50	1.00
2001Q2	30	3.08	0.08	2.50	3.60	2.30	2010Q2	26	1.06	0.26	0.20	2.60	2.30
2001Q3	29	2.67	0.05	2.10	3.10	1.90	2010Q3	29	1.07	0.17	0.20	1.80	2.40
2001Q4	27	2.30	0.15	1.10	2.80	1.20	2010Q4	25	1.34	0.12	0.70	2.20	2.40
2002Q1	25	1.92	0.23	1.00	2.90	0.50	2011Q1	27	1.30	0.15	0.50	2.10	2.80
2002Q2	24	1.29	0.30	0.40	2.60	0.90	2011Q2	29	1.43	0.13	0.70	2.10	1.80
2002Q3	27	1.54	0.30	0.60	3.20	1.20	2011Q3	28	1.45	0.08	0.80	2.10	1.40
2002Q4	28	2.32	0.22	1.30	3.20	1.20	2011Q4	27	1.57	0.04	1.10	1.90	0.50
2003Q1	24	2.42	0.23	1.10	3.20	0.80	2012Q1	26	1.50	0.14	0.70	2.30	-0.50
2003Q2	28	1.61	0.22	0.15	2.40	0.40	2012Q2	30	0.62	0.15	-0.30	1.20	-0.80
2003Q3	28	1.46	0.10	0.60	2.25	0.50	2012Q3	28	0.05	0.37	-1.00	1.20	-1.00
2003Q4	26	1.28	0.14	0.75	2.10	1.10	2012Q4	26	0.19	0.23	-0.70	1.70	-1.10
2004Q1	23	1.22	0.26	0.60	2.20	1.90	2013Q1	23	0.01	0.08	-0.40	1.00	-1.20
2004Q2	31	1.53	0.05	1.00	2.00	2.40	2013Q2	26	0.10	0.16	-0.50	1.20	-0.40
2004Q3	26	1.92	0.04	1.50	2.30	2.20	2013Q3	24	0.02	0.17	-0.70	1.00	0.10
2004Q4	31	1.87	0.07	1.30	2.50	1.80	2013Q4	25	0.33	0.10	-0.30	0.80	0.70
2005Q1	29	2.03	0.11	1.20	2.50	1.40	2014Q1	21	0.58	0.14	0.00	1.80	1.30
2005Q2	31	1.98	0.09	1.20	2.50	1.50	2014Q2	26	0.91	0.06	0.40	1.40	1.00
2005Q3	29	1.89	0.04	1.40	2.20	2.00	2014Q3	25	1.14	0.05	0.80	1.80	1.10
2005Q4	29	1.89	0.05	1.50	2.30	2.20	2014Q4	24	1.28	0.10	0.50	2.00	1.30
2006Q1	25	1.67	0.08	1.10	2.30	3.00	2015Q1	24	1.27	0.06	0.80	1.70	1.80
2006Q2	27	1.64	0.07	1.10	2.20	3.40	2015Q2	25	1.00	0.10	0.40	1.60	2.00
2006Q3	29	1.92	0.04	1.60	2.30	3.30	2015Q3	27	1.15	0.04	0.80	1.50	1.90
2006Q4	29	2.16	0.05	1.70	2.60	3.70	2015Q4	24	1.62	0.12	0.65	2.30	2.00
2007Q1	24	2.04	0.04	1.70	2.40	3.60	2016Q1	23	1.72	0.06	1.30	2.20	1.70
2007Q2	26	1.99	0.07	1.40	2.60	3.10	2016Q2	25	1.67	0.09	1.20	2.70	1.60
2007Q3	27	2.00	0.07	1.40	2.50	3.00	2016Q3	23	1.68	0.04	1.30	1.93	1.80
2007Q4	27	2.12	0.06	1.60	2.60	2.40	2016Q4	23	1.58	0.05	1.10	2.00	1.80
2008Q1	24	2.28	0.04	1.70	2.60	2.10	2017Q1	23	1.31	0.08	0.80	1.90	1.70
2008Q2	30	2.10	0.12	0.90	2.70	1.20							

'Quarter' the given period, 'Obs.' the number of available observations after filtering, 'Mean' the mean value of the individual forecasts after filtering, 'Var.' the variance of these forecasts, 'Min.' and 'Max.' the minimum and maximum forecast after filtering, respectively, 'Outc.' the actual outcome of GDP growth rate in the period.

Table 2: Summary statistics of filtered ECB SPF data on GDP growth two-year-ahead

Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.	Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.
2000Q3	22	2.63	0.07	2.20	3.20	3.90	2009Q1	20	2.13	0.06	1.60	2.50	-5.50
2000Q4	24	2.55	0.11	1.70	3.00	3.40	2009Q2	24	2.12	0.04	1.70	2.40	-5.40
2001Q1	22	2.69	0.04	2.30	3.00	3.20	2009Q3	24	2.05	0.03	1.70	2.40	-4.60
2001Q2	24	2.73	0.12	1.90	3.40	2.30	2009Q4	23	1.81	0.09	1.40	2.50	-2.40
2001Q3	22	2.87	0.05	2.50	3.30	1.90	2010Q1	21	1.64	0.06	1.10	2.10	1.00
2001Q4	25	2.92	0.04	2.50	3.30	1.20	2010Q2	24	1.13	0.30	-0.80	2.00	2.30
2002Q1	20	2.90	0.09	2.10	3.50	0.50	2010Q3	24	0.81	0.21	-0.80	1.50	2.40
2002Q2	26	2.90	0.05	2.50	3.30	0.90	2010Q4	24	0.94	0.40	-0.90	2.25	2.40
2002Q3	26	2.75	0.13	1.20	3.20	1.20	2011Q1	22	1.19	0.20	0.30	2.30	2.80
2002Q4	24	2.85	0.09	2.40	3.90	1.20	2011Q2	24	1.60	0.13	0.70	2.20	1.80
2003Q1	21	2.53	0.10	1.80	2.90	0.80	2011Q3	26	1.52	0.10	0.90	2.00	1.40
2003Q2	19	2.39	0.20	1.70	3.40	0.40	2011Q4	23	1.60	0.12	0.80	2.50	0.50
2003Q3	23	2.63	0.19	1.80	3.50	0.50	2012Q1	23	1.39	0.14	0.50	2.00	-0.50
2003Q4	23	2.57	0.07	2.10	3.10	1.10	2012Q2	23	1.61	0.14	1.00	2.30	-0.80
2004Q1	21	2.58	0.17	2.00	3.60	1.90	2012Q3	25	1.79	0.11	1.20	2.40	-1.00
2004Q2	23	2.49	0.16	1.60	3.50	2.40	2012Q4	23	1.74	0.06	1.30	2.20	-1.10
2004Q3	26	2.36	0.09	1.80	3.20	2.20	2013Q1	20	1.80	0.04	1.40	2.20	-1.20
2004Q4	21	2.37	0.16	1.60	3.40	1.80	2013Q2	25	1.42	0.18	0.50	2.20	-0.40
2005Q1	20	2.15	0.14	1.30	3.00	1.40	2013Q3	23	1.06	0.16	0.20	1.80	0.10
2005Q2	27	2.18	0.08	1.60	2.60	1.50	2013Q4	22	1.21	0.12	0.10	1.70	0.70
2005Q3	22	2.29	0.03	2.00	2.60	2.00	2014Q1	16	1.21	0.16	0.30	1.87	1.30
2005Q4	27	2.18	0.07	1.50	2.50	2.20	2014Q2	20	1.15	0.15	0.00	1.80	1.00
2006Q1	23	2.29	0.06	1.90	3.00	3.00	2014Q3	20	1.27	0.10	0.40	1.90	1.10
2006Q2	26	2.16	0.04	1.90	2.60	3.40	2014Q4	19	1.18	0.07	0.50	1.60	1.30
2006Q3	26	2.13	0.04	1.80	2.60	3.30	2015Q1	16	1.17	0.06	0.70	1.60	1.80
2006Q4	26	2.05	0.10	1.00	2.80	3.70	2015Q2	20	1.39	0.07	0.60	1.70	2.00
2007Q1	20	1.96	0.06	1.50	2.50	3.60	2015Q3	20	1.47	0.03	1.10	1.80	1.90
2007Q2	24	1.99	0.07	1.40	2.60	3.10	2015Q4	20	1.55	0.03	1.20	1.80	2.00
2007Q3	27	1.87	0.08	1.40	2.50	3.00	2016Q1	19	1.59	0.04	1.10	1.90	1.70
2007Q4	26	1.78	0.10	1.10	2.40	2.40	2016Q2	20	1.48	0.09	0.90	2.10	1.60
2008Q1	20	1.86	0.06	1.10	2.20	2.10	2016Q3	21	1.47	0.05	1.10	1.80	1.80
2008Q2	23	1.97	0.09	1.30	2.50	1.20	2016Q4	22	1.67	0.03	1.30	1.90	1.80
2008Q3	26	2.09	0.06	1.7	2.5	0.10	2017Q1	17	1.71	0.01	1.50	1.82	1.70
2008Q4	25	2.16	0.06	1.8	2.6	-2.10							

'Quarter' the given period, 'Obs.' the number of available observations after filtering, 'Mean' the mean value of the individual forecasts after filtering, 'Var.' the variance of these forecasts, 'Min.' and 'Max.' the minimum and maximum forecast after filtering, respectively, 'Outc.' the actual outcome of GDP growth rate in the period.

Table 3: Summary statistics of filtered ECB SPF data on unemployment one-year-ahead

Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.	Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.
1999Nov	22	10.54	0.09	9.90	11.00	9.54	2008Aug	26	6.74	0.04	6.30	7.20	7.38
2000Feb	24	10.29	0.09	9.60	11.00	8.88	2008Nov	26	7.11	0.06	6.70	7.50	8.03
2000May	21	9.90	0.10	9.00	10.80	9.27	2009Feb	25	7.18	0.11	6.70	7.90	9.86
2000Aug	21	9.66	0.07	9.30	10.30	9.04	2009May	24	7.33	0.16	6.30	7.80	9.01
2000Dec*	22	8.97	0.06	8.50	9.30	8.57	2009Aug	26	8.06	0.15	7.40	8.80	9.55
2001Feb	24	8.78	0.05	8.30	9.20	8.37	2009Nov	26	9.20	0.37	8.20	10.50	10.07
2001May	21	8.55	0.05	8.20	9.00	8.42	2010Feb	26	10.45	0.55	9.20	12.00	10.17
2001Aug	27	8.49	0.07	8.10	9.30	8.39	2010May	25	11.00	0.28	10.10	12.10	10.20
2001Nov	26	8.37	0.06	7.80	9.00	8.45	2010Aug	25	10.72	0.17	10.20	11.60	10.30
2002Feb	23	8.35	0.04	8.00	8.80	8.72	2010Nov	27	10.72	0.10	10.06	11.50	10.14
2002May	21	8.21	0.07	7.80	8.70	8.50	2011Feb	24	10.49	0.07	9.90	10.90	10.22
2002Aug	20	8.56	0.13	7.80	9.30	8.56	2011May	25	10.18	0.12	9.50	10.70	10.01
2002Nov	24	8.66	0.11	8.00	9.50	8.86	2011Aug	27	10.02	0.06	9.50	10.50	10.00
2003Feb	24	8.39	0.06	7.80	8.90	9.07	2011Nov	25	9.92	0.09	9.20	10.40	10.60
2003May	20	8.16	0.25	6.40	8.90	8.99	2012Feb	25	9.62	0.05	9.10	10.10	11.50
2003Aug	25	8.35	0.08	7.50	8.80	9.04	2012May	23	9.56	0.03	9.10	10.00	10.93
2003Nov	25	8.46	0.08	7.50	9.00	9.08	2012Aug	27	10.04	0.13	9.30	11.00	11.29
2004Feb	22	8.87	0.05	8.40	9.30	9.23	2012Nov	25	10.69	0.15	10.00	11.80	11.79
2004May	21	8.89	0.11	8.03	9.50	9.24	2013Feb	21	11.11	0.03	10.70	11.40	12.02
2004Aug	28	8.83	0.04	8.50	9.20	9.28	2013May	20	11.40	0.13	10.90	12.30	12.05
2004Nov	21	8.73	0.05	8.40	9.50	9.26	2013Aug	22	11.69	0.22	10.60	12.40	12.05
2005Feb	27	8.63	0.02	8.30	8.90	9.03	2013Nov	22	12.14	0.25	11.00	13.00	11.90
2005May	26	8.69	0.03	8.20	8.90	9.18	2014Feb	22	12.36	0.07	11.90	12.90	11.48
2005Aug	28	8.75	0.03	8.30	9.20	9.16	2014May	17	12.42	0.14	11.50	13.10	11.89
2005Nov	25	8.72	0.03	8.10	9.00	8.95	2014Aug	23	11.98	0.12	11.20	12.90	11.65
2006Feb	25	8.73	0.04	8.20	9.30	8.25	2014Nov	23	11.90	0.05	11.30	12.40	11.50
2006May	23	8.68	0.03	8.40	9.00	8.76	2015Feb	22	11.64	0.04	11.20	12.00	10.68
2006Aug	24	8.43	0.02	8.00	8.63	8.48	2015May	19	11.39	0.08	11.10	12.40	11.18
2006Nov	27	7.99	0.02	7.60	8.30	8.05	2015Aug	20	11.32	0.03	11.10	11.80	11.06
2007Feb	26	7.92	0.02	7.50	8.30	7.48	2015Nov	22	11.13	0.05	10.70	11.61	10.47
2007May	21	7.69	0.04	7.40	8.20	7.75	2016Feb	21	10.80	0.05	10.40	11.30	9.92
2007Aug	24	7.56	0.03	7.20	7.90	7.53	2016May	21	10.62	0.05	10.15	10.90	10.33
2007Nov	25	7.38	0.02	7.10	7.70	7.31	2016Aug	22	10.47	0.03	10.00	11.00	10.14
2008Feb	26	6.99	0.03	6.70	7.30	7.58	2016Nov	22	10.10	0.04	9.60	10.50	9.74
2008May	21	6.75	0.06	6.30	7.20	7.30	2017Feb	18	9.87	0.07	9.20	10.50	9.54

'Quarter' the given period, 'Obs.' the number of available observations after filtering, 'Mean' the mean value of the individual forecasts after filtering, 'Var.' the variance of these forecasts, 'Min.' and 'Max.' the minimum and maximum forecast after filtering, respectively, 'Outc.' the actual outcome of unemployment rate in the period.

* For this period, forecasts are provided for December instead of for November.

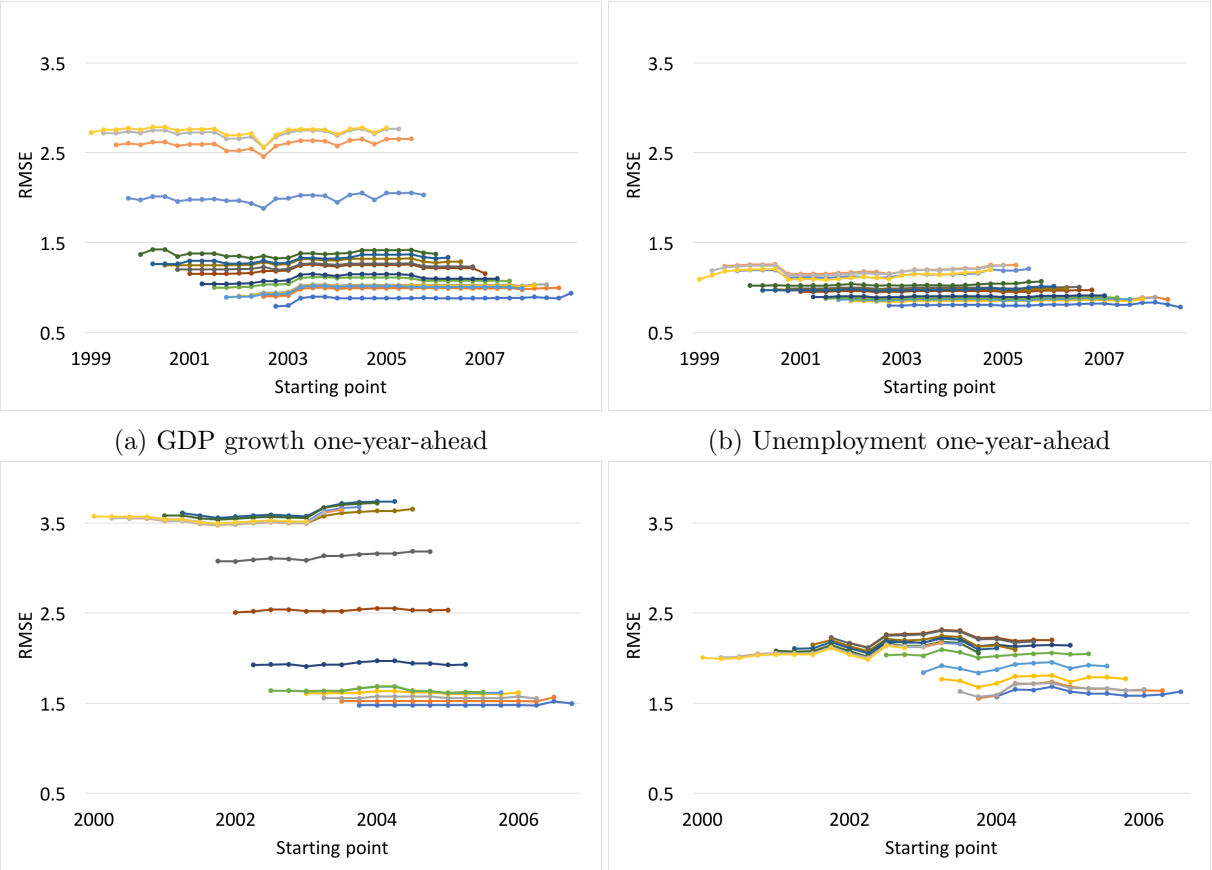
Table 4: Summary statistics of filtered ECB SPF data on unemployment two-year-ahead

Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.	Quarter	Obs.	Mean	Var.	Min.	Max.	Outc.
2000Nov	22	10.14	0.15	9.50	11.00	8.65	2009Feb	23	6.83	0.06	6.50	7.50	9.01
2001Feb	22	9.90	0.17	9.20	11.00	8.42	2009May	18	6.66	0.10	6.10	7.40	9.55
2001May	20	9.47	0.25	8.00	10.70	8.39	2009Aug	23	6.59	0.09	5.80	7.40	9.86
2001Aug	21	9.20	0.11	8.30	9.60	8.37	2009Nov	23	6.95	0.14	6.40	7.70	10.07
2001Dec*	22	8.57	0.13	7.90	9.50	8.47	2010Feb	21	6.96	0.22	6.00	8.00	10.20
2002Feb	23	8.31	0.18	7.40	9.30	8.5	2010May	19	7.39	0.30	6.00	8.40	10.30
2002May	19	8.21	0.26	7.30	9.30	8.56	2010Aug	23	8.04	0.40	7.10	9.60	10.17
2002Aug	24	8.15	0.23	7.40	9.50	8.72	2010Nov	24	9.28	0.89	7.10	11.10	10.14
2002Nov	24	8.09	0.31	7.40	10.00	8.86	2011Feb	23	10.37	1.18	8.50	12.40	10.01
2003Feb	22	8.05	0.12	7.50	8.90	8.99	2011May	21	10.62	0.51	9.70	12.00	10.00
2003May	20	7.90	0.13	7.40	8.70	9.04	2011Aug	22	10.37	0.42	9.40	11.80	10.22
2003Aug	20	8.21	0.30	7.00	9.20	9.07	2011Nov	25	10.49	0.51	9.30	13.00	10.60
2003Nov	22	8.31	0.13	7.70	9.20	9.08	2012Feb	22	9.96	0.27	8.90	11.20	10.93
2004Feb	23	8.04	0.11	7.20	8.50	9.24	2012May	21	9.80	0.20	9.00	10.70	11.29
2004May	21	7.87	0.13	7.20	8.60	9.28	2012Aug	23	9.57	0.27	8.50	10.21	11.50
2004Aug	23	7.95	0.14	7.00	8.50	9.23	2012Nov	24	9.51	0.26	8.50	10.40	11.79
2004Nov	23	8.07	0.13	7.00	8.60	9.26	2013Feb	22	9.22	0.13	8.50	10.00	12.05
2005Feb	18	8.45	0.13	7.50	9.00	9.18	2013May	19	9.23	0.10	8.30	9.70	12.05
2005May	20	8.58	0.17	7.80	9.40	9.16	2013Aug	25	9.74	0.46	8.40	12.10	12.02
2005Aug	25	8.46	0.09	7.80	9.00	9.03	2013Nov	23	10.41	0.46	9.00	12.00	11.90
2005Nov	19	8.40	0.06	8.00	8.90	8.95	2014Feb	20	10.62	0.11	10.00	11.20	11.89
2006Feb	25	8.30	0.07	7.80	8.80	8.76	2014May	17	10.81	0.34	9.80	11.70	11.65
2006May	21	8.34	0.10	7.50	8.80	8.48	2014Aug	19	11.23	0.47	9.80	12.40	11.48
2006Aug	25	8.42	0.12	7.30	8.80	8.25	2014Nov	20	11.48	0.57	9.50	12.60	11.50
2006Nov	24	8.41	0.10	7.40	8.80	8.05	2015Feb	19	11.79	0.29	10.50	12.90	11.18
2007Feb	23	8.41	0.10	7.50	9.00	7.75	2015May	16	11.98	0.36	10.40	13.00	11.06
2007May	20	8.41	0.10	7.40	8.80	7.53	2015Aug	21	11.47	0.28	10.10	12.30	10.68
2007Aug	23	8.17	0.08	7.40	8.50	7.48	2015Nov	20	11.44	0.16	10.60	12.00	10.47
2007Nov	26	7.72	0.06	7.10	8.10	7.31	2016Feb	20	11.16	0.13	10.30	11.80	10.33
2008Feb	24	7.65	0.08	7.00	8.30	7.3	2016May	17	10.93	0.24	9.90	12.20	10.14
2008May	19	7.50	0.08	7.10	8.20	7.38	2016Aug	18	10.89	0.17	9.60	11.50	9.92
2008Aug	22	7.39	0.03	7.10	7.80	7.58	2016Nov	21	10.66	0.15	9.80	11.30	9.74
2008Nov	24	7.21	0.06	6.80	7.80	8.03	2017Feb	21	10.33	0.21	9.50	11.30	9.54

'Quarter' the given period, 'Obs.' the number of available observations after filtering, 'Mean' the mean value of the individual forecasts after filtering, 'Var.' the variance of these forecasts, 'Min.' and 'Max.' the minimum and maximum forecast after filtering, respectively, 'Outc.' the actual outcome of unemployment rate in the period.

* For this period, forecasts are provided for December instead of for November.

B Pseudo-out-of-sample RMSE



(a) GDP growth one-year-ahead

(b) Unemployment one-year-ahead

(c) GDP growth two-year-ahead

(d) Unemployment two-year-ahead

Forecasts are computed pseudo-out-of-sample for $\tilde{v}_2 = 16$ periods using estimation windows based on $\underline{v} = 8$, and evaluated following Eq. 2. The line with earliest starting points corresponds to the earliest set of $\tilde{v}_2 = 16$ pseudo-out-of-sample periods and the line with the latest starting points corresponds to the latest set of $\tilde{v}_2 = 16$ pseudo-out-of-sample periods.

Figure 1: Pseudo-out-of-sample RMSE for different starting points using a rolling window