Forecasting in a changing world: from the great recession to the COVID-19 pandemic

Mariia Artemova[1]
Francisco Blasques[1]
Siem Jan Koopman[1]
Zhaokun Zhang[2]

[1] Vrije Universiteit Amsterdam
[2] Shanghai University

Forecasting in a changing world:
from the great recession to the COVID-19 pandemic.

Mariia Artemova[1,2], Francisco Blasques[1,2]*, Siem Jan Koopman[1,2,3],
Zhaokun Zhang[4]

[1] *Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands*

[2] *Tinbergen Institute, 1082 MS Amsterdam, The Netherlands*

[3] *CREATES, Aarhus University, 8210 Aarhus, Denmark*

[4] *School of Economics, Shanghai 200444, Shanghai University, China*

January 2021

Abstract

We develop a new targeted maximum likelihood estimation method that provides improved forecasting for misspecified linear autoregressive models. The method weighs data points in the observed sample and is useful in the presence of data generating processes featuring structural breaks, complex nonlinearities, or other time-varying properties which cannot be easily captured by model design. Additionally, the method reduces to classical maximum likelihood when the model is well specified, which results in weights which are set uniformly to one. We show how the optimal weights can be set by means of a cross-validation procedure. In a set of Monte Carlo experiments we reveal that the estimation method can significantly improve the forecasting accuracy of autoregressive models. In an empirical study concerned with forecasting the U.S. Industrial Production, we show that the forecast accuracy during the Great Recession can be significantly improved by giving greater weight to observations associated with past recessions. We further establish that the same empirical finding can be found for the 2008-2009 global financial crisis, for different macroeconomic time series, and for the COVID-19 recession in 2020.

Keywords: Autoregressive Models, Cross-Validation, Kullback-Leibler Divergence, Stationarity and Ergodicity, Macroeconomic Time Series.

JEL classification codes: C10, C22, C32, C51.

---

1 Introduction

For the vast majority of empirical studies in economics and finance of the last five decades, the linear autoregressive (AR) model has been used for the analysis and forecasting of time series observations. The flexibility of the AR model allows econometricians to describe relatively complex dynamics in a remarkably simple manner. From short-run temporal dependence with seasonal dynamics to long-run persistency generated by stochastic trends, the basic AR model is often able to provide a better in-sample fit and a more accurate out-of-sample forecast, when compared to more intricate and complex dynamic models. However, given the simplistic structures present in linear reduced-form AR models, and the complexity of data generating processes for large market economies, it is likely that autoregressions suffer from a form of model misspecification. A typical way in which model misspecification can manifest itself is parameter instability. We can find much evidence of parameter instability in economic and financial empirical studies and applications; see Stock and Watson (1996, 2007) in macroeconomic forecasting, and Wolff (1987), Schinasi and Swamy (1989) and Goyal and Welch (2003) in financial forecasting. As also discussed by Inoue et al. (2014), parameter instability is widely recognized as a crucial issue that significantly hampers the forecasting performance of econometric models; see, for example, Stock and Watson (1996), Clements and Hendry (1998), Goyal and Welch (2003), Koop and Potter (2004), Paye and Timmermann (2006), Giacomini and Rossi (2009), and Rossi (2013). Several methods have been proposed to improve forecasting performance in the presence of parameter instability. Rolling window estimation is among the most popular with important applications in finance, see Goyal and Welch (2003), in macroeconomics, see Swanson (1998), and in exchange rate forecasting, see Molodtsova and Papell (2009), among others. Using a window of observations for estimation can be regarded as a basic weighting scheme for the observations.

We propose a new targeted estimation procedure for the unknown parameters in the model. The procedure weighs observed data points differently for constructing a likelihood function that is used to estimate the parameters. The observation weights are chosen such that the forecasting performance of the misspecified linear AR model is optimized. Hence, the weights are designed to optimize a secondary criterion, which we call a *targeting function*. By providing a second forecasting target to the estimation procedure, we provide additional

structure to the estimation process. Furthermore, we are able to improve the properties of the MLE with respect to this secondary target. We refer to this procedure as *weighted maximum likelihood estimation* (WMLE).

In practice, our WMLE provides specific weights to individual observations, and obtains parameter estimates that are optimal for forecasting. For example, the WMLE can be used to give higher weights to more recent observations compared to observations far in the past. This may be desirable as various political, institutional, and technological developments change the economy, rendering past observations increasingly obsolete for the forecasting of economic variables. Similarly, the WMLE can be used to give higher weights to data from past periods of greater historical relevance. For example, Industrial Production forecasts during an economic recession can be improved by giving higher weights to past episodes of economic recession. In fact, at any given point in time, the WMLE weights can be adjusted to reflect the importance of past observations for improving forecast accuracy. It is also natural that forecasts for inflation during international oil crises can be improved by paying special attention to price dynamics during past oil crises. The forecasts for unemployment during a period of strong fiscal austerity can benefit from giving more prominence to observations that originate from past episodes of strong fiscal austerity. The WMLE estimator often implies a time-varying parameter AR model whose parameter estimates are optimal for forecasting observations in a given period of interest. We show in our empirical study that forecasts of the U.S. Industrial Production during the global 2008–2009 recession can be significantly improved by defining weights that are making past observations from recession periods more informative while those from the more remote past less informative.

The WMLE procedure can be applied generally for the analysis, modeling, and forecasting of economic and financial time series. The procedure is designed to provide optimal AR forecasts under model misspecification and it includes other weighted estimation methods (such as rolling window estimation techniques) as a special case. We show in our study how to obtain the optimal WMLE weights through cross-validation. This procedure automatically identifies the important features from past data and adjusts the weights to deliver optimal out-of-sample forecasting accuracy for the AR model. We will show that our method of weight selection by cross-validation ensures that the WMLE converges to the classical MLE when the AR model is correctly specified. In this special case, the WMLE converges to the

3

'true' time-invariant parameter vector, which is optimal for forecasting. At the same time, if the AR model is misspecified, then the cross-validation procedure will establish weights that are optimal according to the targeting criterion. In this case, the WMLE provides time-varying parameter estimates for the AR model which significantly improve the forecasting accuracy of the classical MLE method.

In contrast to earlier contributions in the literature as listed above, our method for finding optimal weights is shown to be valid under very general conditions for the data generation process which is not restricted to pre-specified breaks, trends, and related effects. We also deviate from a substantial part of the literature in basing our approach on the maximum likelihood estimation of AR models. Indeed, while Giraitis et al. (2012) have already proposed a cross-validation method for optimizing forecasts, their method is not related to maximum likelihood estimation of parameters and they do not use AR models to forecast. Instead, Giraitis et al. (2012) produce one-step-ahead predictions directly through a weighted sample average of past observations. In essence, our approach also defines such weights, but it does so implicitly through the structure of the AR model and its estimated parameters. By taking the AR model as a starting point, we take advantage of the ability of this model to describe relatively complex dynamic dependence patterns in the data and produce forecasts multiple steps ahead. Amongst other models, AR models are capable to capture short-run dynamics, seasonalities and stochastic trends in time series. Hence, we start at an appreciable level of forecasting accuracy. The WMLE is then used to further optimize the forecasting performance of the model. Additional empirical evidence reveals that the advantages of WMLE extend to the forecasting of a wide variety of time series.

Instead of introducing weights to the MLE method, to address possible misspecifications in a linear AR model, one could also concentrate on improving model specification. We fully recognize that many contributions in econometrics are concerned with efforts to improve the model design by, for example, introducing breaks, random coefficients, nonlinearities and non-parametric features. These attempts typically leads to a specification that is closer to the correct model. Indeed, testing for model specification and investing in model design that captures complex nonlinearities and time-varying parameters is an important part of time series analysis and forecasting. There are however cases where we believe that a targeted estimator, such as the one we develop here, can be advantageous: (*i*) when there is great

uncertainty about the correct model specification; (ii) when correct specification is difficult to achieve, even after incorporating nonlinear features into the model; (iii) when model complexity is undesirable from a perspective of implementation, computation, or interpretation. These issues are especially relevant in present days, when the world has been shaken by the COVID-19 pandemic, and time series models have to be revised without much evidence after such a major structural break. This situation illustrates the uncertainty referred to in case (i). The relevance of case (ii) becomes more prominent when we analyze complex data sets where it is highly challenging to obtain only a reasonable model specification, even after the introduction of nonlinear features in the model. Finally, case (iii) is particularly relevant when we analyze big data sets and we fit a model to thousands of different time series for producing forecasts, including confidence bounds, and impulse responses. In this setting, it is undesirable to treat complex nonlinear models by means of numerical estimation methods that can become computationally prohibitive or unreliable.

This paper is organized as follows. Section 2 introduces the basic concepts behind weighted maximum likelihood estimation and optimality in forecasting. Section 3 provides the details of our WMLE procedure including possible weight functions and their estimation by cross-validation. Section 4 investigates the theoretical aspects of WMLE; in particular, we show its asymptotic equivalence with MLE, provide conditions for optimal forecasting performance, and show the asymptotic validity of the Diebold-Mariano test for accuracy in forecasting. For a selection of empirical illustrations in Section 5, we show that the optimal weight functions can take intuitive forms that would be difficult to obtain from time-varying parameter models. Section 6 concludes and discusses possible directions for future research.

2   Basic Concepts

In order to easily grasp the basic concepts that underly the weighted maximum likelihood estimation (WMLE) procedure, we first consider a very simple example concerned with the estimation of a population mean.

Consider a sequence of random variables $X_1, \ldots, X_T$ that is assumed to come from an unknown data generating process (DGP) with length $T > 1$. Suppose that we decide to work with a statistical model where X_t is assumed to be an independent and identically

distributed (iid) sequence that is generated by a Gaussian distribution with unknown mean μ and unity variance, that is $X_t \sim N(\mu, 1)$, for $t = 1, \ldots, T$. In this case, the MLE of the unknown parameter μ, based on the sample $X_{s:T} := \{X_t\}_{t=s}^T$, for some $1 \le s < T$, is naturally given by

$$\hat{\mu}_{s:T} = \arg\max \sum_{t=s}^T (X_t - \mu)^2 \quad \Rightarrow \quad \hat{\mu}_{s:T} = \frac{1}{T-s+1} \sum_{t=s}^T X_t.$$

Since the statistical model assumes that the observations in the sample are iid, the n-step ahead forecast $\hat{X}_{T+n}$ conditional on the observed sample $X_s, \ldots, X_T$ is

$$\hat{X}_{T+n} = \mathbb{E}(X_{T+n}|X_s, \ldots, X_T) = \hat{\mu}_{s:T}.$$

If the assumptions imposed by the statistical model are correct, then the model is well-specified and the DGP consists of a sequence of iid $N(\mu_0, 1)$ random variables with some unknown μ_0. In this special case, it is optimal to use the complete available sample $X_1, \ldots, X_T$, and hence set $s = 1$. Furthermore, the MLE provides an unbiased forecast $\hat{X}_{T+n}$ and it is also the most accurate forecast among all unbiased estimators. We have that the bias is zero,

$$\text{Bias}(\hat{X}_{T+n}) = \mathbb{E}(\hat{X}_{T+n} - X_{T+n}) = \mathbb{E}\hat{\mu}_{s:T} - \mu_0 = \frac{1}{T-s+1} \sum_{t=s}^T \mu_0 - \mu_0 = 0,$$

and the variance attains the Cramer-Rao lower bound (CRLB),

$$\mathbb{V}\text{ar}(\hat{X}_{T+n}) = \mathbb{V}\text{ar}(\hat{\mu}_{s:T}) = \frac{1}{(T-s+1)^2} \sum_{t=s}^T \mathbb{V}\text{ar}(X_t) = \frac{1}{T-s+1} = \text{CRLB}.$$

As a result, the forecast mean squared error (MSE) is the smallest among all the unbiased forecasts and it becomes smaller by setting $s = 1$ and letting the sample size T increase, that is

$$\text{MSE}(\hat{X}_{T+n}) = \left(\text{Bias}(\hat{X}_{T+n})\right)^2 + \mathbb{V}\text{ar}(\hat{X}_{T+n}) = \frac{1}{T-s+1} \to 0 \quad \text{as} \quad T-s+1 \to \infty.$$

Unfortunately, this good state of affairs is not applicable to a model that is misspecified.

Consider, for example, the case where the mean has a trend component; that is, the true unknown DGP is $X_t \sim N(\beta t, 1)$. However, the statistical model still assumes that the data is iid with $X_t \sim N(\mu, 1)$. In this case, the MLE no longer provides the most accurate forecasts, and most importantly, it is no longer optimal to make use of the entire sample of data $X_1, \ldots, X_T$. This interesting situation occurs because the MLE attempts to approximate the distribution of the entire sample $X_s, \ldots, X_T$ by minimizing the empirical Kullback-Leibler divergence with respect to the true measure. To do so, the MLE becomes centered at the value $\beta T/2$ when we use the complete sample ($s = 1$). However, since $\mathbb{E}(X_{T+1}) = \beta(T+1)$, this is clearly not an interesting estimator when we are interested in forecasting accuracy rather than fitting the distribution of past observations. An MLE that takes into account only the last few sample values can greatly improve the forecast MSE. The optimal value of $1 \leq s < T$ finds a good compromise between the forecast bias (which decreases with s increasing) and the forecast variance (which increases together with s). The forecast bias is strictly decreasing in s since

$$\text{Bias}(\hat{X}_{T+n}) = \mathbb{E}\hat{\mu}_{s:T} - \mathbb{E}X_{T+n} = \frac{1}{T-s+1}\sum_{t=s}^{T} \beta t - \beta(T+n) = \beta n + \frac{\beta}{2}(T-s).$$

The forecast bias would be minimized by setting $s = T$ and hence making use of only the last observation X_T in the sample. However, this strategy is not optimal in terms of the MSE since the variance of the forecast produced by this MLE would be too large

$$\mathbb{V}\text{ar}(\hat{X}_{T+n}) = \mathbb{V}\text{ar}(\hat{\mu}_{s:T}) = \frac{1}{(T-s+1)^2}\sum_{t=s}^{T} \mathbb{V}\text{ar}(X_t) = \frac{1}{(T-s+1)}.$$

In order to minimize the MSE, it is instead better to make use of the information contained in past data values, even if this comes at the cost of a larger forecast bias. This trade-off is reflected in the MSE expression as given by

$$\text{MSE}(\hat{X}_{T+n}) = \left(\text{Bias}(\hat{X}_{T+n})\right)^2 + \mathbb{V}\text{ar}(\hat{X}_{T+n}) = \frac{1}{4}\beta^2(2n+T-s)^2 + (T-s+1)^{-1}.$$

Figure 1 below plots the 1-step-ahead forecast MSE as a function of s for the case $T = 100$ and $\beta = 0.02$. The figure shows that instead of using the entire sample $X_1, \ldots, X_{100}$, it is

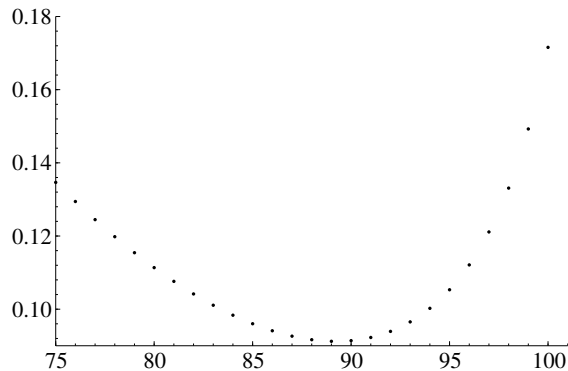optimal to use only the last 12 observations $X_{89}, \ldots, X_{100}$.

Figure 1: One-step-ahead forecast MSE as function of s for $T = 100$ and $\beta = 0.02$.

The estimator $\hat{\mu}_{s:T}$ proposed above is essentially an MLE that gives zero weight to the likelihood contributions from data before s, and unit weight to the contributions of the remained observations. In essence, it is a WMLE of the form

$$\hat{\mu}_T = \arg\max \sum_{t=1}^{T} \ell_t(X_t; \mu) \cdot w_t$$

where $\ell_t(X_t; \mu)$ denotes the log likelihood contribution of the X_t observation and w_t is the weight function

$$w_t = \begin{cases} 0, & \text{for } 1 \le t < s, \\ 1, & \text{for } s \le t \le T. \end{cases}$$

This weight function is however unnecessarily restrictive, and it can be easily improved upon, at least in terms of reducing the forecast MSE. Consider for example the case where $w_t = \rho^{T-t}$ for some $0 < \rho \le 1$. This weight function gives unit weight to the last observation X_T and decaying weights to past observations when $0 < \rho < 1$. The WMLE then takes the form

$$\hat{\mu}_T = \frac{\sum_{t=s}^{T} \rho^{T-t} X_t}{\sum_{t=s}^{T} \rho^{T-t}}.$$

In the special case of $\rho = 1$, we obtain the usual MLE. It is also easy to show that

$$\text{Bias}(\hat{X}_{T+1}) = \frac{\sum_{t=s}^{T} \rho^{T-t} \beta t}{\sum_{t=s}^{T} \rho^{T-t}} - \beta(T+1), \qquad \mathbb{V}\text{ar}(\hat{X}_{T+1}) = \frac{\sum_{t=s}^{T} \rho^{2(T-t)}}{\left(\sum_{t=s}^{T} \rho^{T-t}\right)^2},$$

8

from which it follows that

$$\text{MSE}(\hat{X}_{T+1}) = \frac{\beta^2 \left(\rho^{T+1}((\rho-1)T + \rho + s - s\rho - 2) + \rho^s\right)^2}{(\rho-1)^2 \left(\rho^s - \rho^{T+1}\right)^2} + \frac{(\rho-1)\left(\rho^{T+1} + \rho^s\right)}{(\rho+1)\left(\rho^{T+1} - \rho^s\right)}.$$

Figure 2 presents the forecast MSE of the WMLE as a function of the parameter ρ together with the optimal weight function. The first graph shows that the WMLE with $\rho \approx 0.89$ considerably outperforms the standard MLE with uniform weight ($\rho = 1$) while the second graph presents the optimal weight function for $\rho = 0.89$.
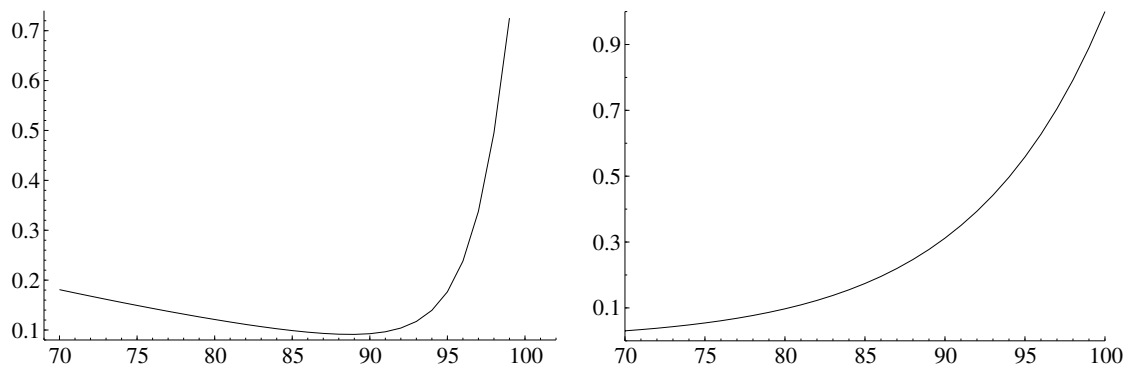

Figure 2: One-step-ahead forecast MSE as function of ρ for $T = 100$ and $\beta = 0.02$ (left), and optimal weight function $w_t = 0.89^{T-t}$ (right).

These results are obviously not exclusive to the linear trending mean considered above. On the contrary, any time-variation in the true mean that is not appropriately captured by the statistical model will result in a non-constant optimal weight function that improves the forecasting accuracy of the MLE. In fact, the WMLE opens the possibility for much more complex weight functions than the simple exponential decay presented above. These weights can optimally adapt to breaks in the DGP, cyclical fluctuations and others.

3   Weighted Maximum Likelihood Estimation

We consider the linear autoregressive (AR) model in our treatment of weighted maximum likelihood estimation. The AR(p) model assumes that the time series $\{X_t\}_{t\in\mathbb{Z}}$ is generated by the dynamic stochastic process

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \epsilon_t, \qquad t \in \mathbb{Z}, \tag{1}$$

where $p \in \mathbb{N}_+$ is the lag-order of the AR process, $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is a sequence of iid innovations with density $p_\epsilon(\boldsymbol{\lambda})$ indexed by the vector of parameters $\boldsymbol{\lambda}$. For simplicity, we collect all parameters in the vector $\boldsymbol{\theta} := (\boldsymbol{\alpha}, \boldsymbol{\lambda})$ where $\boldsymbol{\alpha} := (\alpha_0, \ldots, \alpha_p)$. The parameter vector $\boldsymbol{\theta}$ is easy to estimate via the exact maximum likelihood method or via regression (based on the conditional likelihood function). The autocorrelation function, the forecast function, including forecast confidence bounds, and the impulse response function for the AR(1) model rely on $\boldsymbol{\theta}$ and given an estimate of $\boldsymbol{\theta}$, they can be readily computed.

When a dynamic model is well specified, the classical MLE produces optimal forecasts in Kullback-Leibler divergence; see Blasques et al. (2015). In case of the Gaussian AR(p) model as given by (1) with $p_\epsilon(\boldsymbol{\lambda}) = N(0, \sigma^2)$ and $\boldsymbol{\lambda} = \sigma^2$, the forecast function also minimizes the mean squared error loss function. The main reason behind this result is that the in-sample fit and out-of-sample forecasting performance are closely connected when the model is well specified. However, in the case of a misspecified AR model, having a good in-sample fit over the entire data set on the one hand, and having a good out-of-sample forecasting accuracy on the other hand, become two separate problems. Our Proposition 3 in Section 4 shows that for every misspecified AR model, there exists a non-constant sequence of parameters that improves the approximation of the AR model to the true data-generating process, together with the forecasting accuracy. We further show that the WMLE can help in finding such an appropriate sequence of parameters, and that the WMLE outperforms the classical MLE.

3.1 The Estimator

Let $\mathbf{w} \in \mathbb{R}_+^T$ denotes a weight vector with elements $(w_1, \ldots, w_T)$ that defines the weight of the log likelihood contribution at time $t = 1, \ldots, T$. The weights w_t are non-negative and sum up to unity, that is $\sum_{t=1}^T w_t = 1$. For any given $\mathbf{w}$, the WMLE $\hat{\boldsymbol{\theta}}(\mathbf{w})$ for the vector $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}}(\mathbf{w}) := \arg\max_{\boldsymbol{\theta}\in\Theta} \frac{1}{T} \sum_{t=1}^T \ell_t(\boldsymbol{\theta}) \cdot w_t, \tag{2}$$

where $\ell_t(\boldsymbol{\theta})$ denotes the logarithm of the conditional density of X_t given $X_{t-1}, X_{t-2}, \ldots, X_1$,

$$\ell_t(\boldsymbol{\theta}) := \log p_\epsilon\big(e_t(\boldsymbol{\alpha}); \boldsymbol{\lambda}\big),$$

10

and $e_t(\boldsymbol{\alpha})$ denotes the residual term

$$e_t(\boldsymbol{\alpha}) := X_t - \alpha_0 - \alpha_1 X_{t-1} - \ldots - \alpha_p X_{t-p}, \qquad t = p+1, \ldots, T.$$

The fundamental difference between the WMLE and the classical MLE is the introduction of weights for the log likelihood of different observations. Observations with relatively large (small) weight will have a relatively larger (smaller) influence in the estimation of the parameter vector $\boldsymbol{\theta}$. As such, the WMLE parameter estimates will attempt to fit more accurately the dynamics of the time series at certain periods of interest. Since the weight vector $\mathbf{w}$ determines the parameter estimates, we consider the WMLE as a map $\hat{\boldsymbol{\theta}} : \Omega \times \mathbb{R}_+^T \to \Theta$, where Ω is the event space of the underlying probability space of interest. Similarly, given the vector $\mathbf{w}$, the random estimator maps elements of Ω to the parameter space Θ, that is $\hat{\boldsymbol{\theta}}(\mathbf{w}) : \Omega \to \Theta$.

In order to analyze the out-of-sample forecasting performance of the AR(p) under the WMLE estimates, it is important to define a class of weight vectors that use information only until a certain point in time. In particular, we let $\mathbf{w}_k$, $T' \leq k \leq T$ denote a weight vector whose elements are uniformly set to zero for all $t > k$

$$\mathbf{w}_k = (w_1, \ldots, w_k, 0, \ldots, 0)',$$

where all weight vectors have dimension $T \times 1$. The weight vector $\mathbf{w}_k$ defines a WMLE $\hat{\boldsymbol{\theta}}(\mathbf{w}_k)$ that only makes use of the shorter sample $X_1, \ldots, X_k$. Finally, we let $\mathcal{W}$ denote the space of all weight matrices $\mathbf{W}$ and each $\mathbf{w}_k$ corresponds to a row of $\mathbf{W}$. Throughout, we let the rows of $\mathbf{W}$ be normalized, for instance, to sum up to one.

In general, the WMLE estimates will differ from one time period to the next. For example, the parameter estimate $\hat{\boldsymbol{\theta}}(\mathbf{w}_k)$ obtained using the sub-sample $X_1, \ldots, X_k$, will typically be different from the parameter estimate $\hat{\boldsymbol{\theta}}(\mathbf{w}_{k+1})$ obtained using the sub-sample $X_1, \ldots, X_{k+1}$. For this reason, the WMLE can potentially be used to construct a sequence of parameter estimates

$$\hat{\boldsymbol{\theta}}(\mathbf{w}_k), \ \hat{\boldsymbol{\theta}}(\mathbf{w}_{k+1}), \ \hat{\boldsymbol{\theta}}(\mathbf{w}_{k+2}), \ \ldots$$

that describes the parameter instability in the AR(p).

Our general framework of WMLE with its different weight sequences encompasses a number of different well-known estimators. Next we highlight some of the relations between the WMLE and well-known estimators.

3.2   Special Classes of Weights

The famous recursive least squares filter originally proposed by Gauss in 1821 in his work *'Theoria combinationis observationum erroribus minimis obnoxiae'* and rediscovered by Plackett (1950), consists of a sequence of estimators obtained recursively over a window of increasing length. Recursive ML estimators can be obtained in the WMLE setting by defining a matrix of weights $\mathbf{W}$ with rows $\mathbf{w}_k$ having elements that satisfy $w_{k,1} = w_{k,2} = \cdots = w_{k,k} = 1$ for an increasing k.

The popular rolling window estimators with window size h used for improved forecasting performance in Swanson (1998), Goyal and Welch (2003), Molodtsova and Papell (2009) and Inoue et al. (2014), among others, are obtained as a special case of the WMLE by letting the matrix of weights $\mathbf{W}$ with rows $\mathbf{w}_k$ having elements that satisfy $w_{k,t} = 1$ if $k - t \leq l$ and $w_{k,t} = 0$ otherwise, for some window length l and every k.

The elements of the weight vectors $\mathbf{w}_k$ can be set a-priori or they can be estimated to provide optimal $\text{AR}(p)$ forecasts. For reasons of simplicity, computational efficiency, and, most importantly, statistical efficiency, it will often be beneficial to parameterize the weight function by a low-dimensional parameter vector $\boldsymbol{\rho}$. Instead of estimating all elements of any given weight vector $\mathbf{w}_k$ individually, we can estimate a small number of coefficients in $\boldsymbol{\rho}$ which determine the individual weights in a parametric fashion. For example, exponentially decaying weights are obtained from a scalar parameter ρ and with the specification $w_{k,t}(\rho) = \rho^{(k-t)}$. This set of weights normalizes the weight of the most recent observation $t = k$ to unity $w_{k,k} = 1$, and allows for data in the remote past to be given a lower weight during the estimation of $\boldsymbol{\theta}$. The exponential class of weights contains the MLE as a special case ($\rho = 1$) as well as WMLE based on a wide range of decay patterns ($0 \leq \rho < 1$).

Another interesting class of weights is obtained by letting the elements of $\mathbf{w}_k(\boldsymbol{\rho})$ depend on lagged values of X_t and/or other variables of interest Z_t. For example, in Section 5, we show that forecasts of Industrial Production during the Great Recession, can be significantly

improved by defining weights that make past recession periods more informative, and, at the same time, downweight more recent observations. In particular, we use the NBER recession indicator Z_t and define the weights of the vector $\mathbf{w}_k(\boldsymbol{\rho}, Z_t)$ as follows

$$w_{k,t}(\boldsymbol{\rho}, Z_t) = \rho_1^{(k-t)}(1 + (\rho_2 - 1) \cdot Z_t),$$

where $\boldsymbol{\rho} = (\rho_1, \rho_2)$ with $0 \leq \rho_1 \leq 1$, $\rho_2 \geq 1$. This specification illustrates the generality of our framework of weighing observations in the context of maximum likelihood estimation.

3.3   Estimation of Optimal Weights by Cross-Validation

The recursive and rolling window estimation techniques mentioned in Section 3.2, can provide important insight into parameter instability, be it in the form of breaks, trends, seasonality or random changes. However, those methods are not designed from the outset to optimize the forecasting performance of the AR model. Instead, any improvements in forecast accuracy are the result of an ad-hoc improvement in model specification. Indeed, rolling window estimation will provide improved forecasts if it turns out that the exclusive use of recent data is advantageous for forecasting. Similarly, time-varying parameter models, will deliver improved forecast accuracy over their static parameter counterparts, if the specific weights on past data introduced indirectly by the parameter updating equations happen to improve the forecasting ability of the model. In this section we propose a cross-validation method for finding the matrix of weights $\mathbf{W}$ that optimizes a secondary *targeting function* which measures the forecasting performance of the AR(p) model. There are various popular choices of targeting criterion functions that reflect forecasting accuracy. The mean absolute forecast error (MAFE), the mean squared forecast error (MSFE), the out-of-sample log likelihood, etc. If the MSFE is chosen then we say that the matrix of weights $\hat{\mathbf{W}}$ is the best WMLE in an n-step-ahead MSFE sense if

$$\hat{\mathbf{W}} = \arg\min_{\mathcal{W}} \frac{1}{T - T' - n} \sum_{k=T'}^{T-n} \left(\hat{X}_{k+n}\big(\hat{\boldsymbol{\theta}}(\mathbf{w}_k)\big) - X_{k+n}\right)^2 \tag{3}$$

where, as noted before, the WMLE $\hat{\boldsymbol{\theta}}$ maps events $\omega \in \Omega$ and weight vectors $\mathbf{w} \in \mathbb{R}_+^T$ to the parameter space, i.e. $\hat{\boldsymbol{\theta}} : \Omega \times \mathbb{R}_+^T \to \Theta$, and T' defines the sample point from which the

forecasting accuracy begins to be measured. A large T' gives us more data to estimate the parameter vector $\boldsymbol{\theta}$ by WMLE, but a small number of observations to evaluate the forecasting accuracy of the model and optimize the weights. On the contrary, a small T' increases the uncertainty in the estimation of $\boldsymbol{\theta}$ but gives us a larger sample to determine the optimal weights. When the matrix $\mathbf{W}$ is parameterized by a vector $\boldsymbol{\rho}$, then the optimal weights are defined as $\mathbf{W}(\hat{\boldsymbol{\rho}})$ where

$$\hat{\boldsymbol{\rho}} = \arg\min_{\boldsymbol{\rho}} \frac{1}{H} \sum_{k=T'}^{T-n} \left(\hat{X}_{k+n}\big(\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\boldsymbol{\rho}))\big) - X_{k+n} \right)^2, \tag{4}$$

where we denote $H := T - T' - n$. The optimization of the weights in either (3) or (4) relies on the relation between weight vectors $\mathbf{w}_k(\hat{\boldsymbol{\rho}})$ and the parameter estimates in $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}))$. The optimization would be trivially simple if, for a given data sample, the mapping $\hat{\boldsymbol{\theta}} : \mathbb{R}_+^T \to \Theta$ from weight vectors in $\mathbb{R}_+^T$ to point estimates in Θ were known analytically. In general, however, this map is unknown analytically. As a result, we need to carry out the optimization numerically.

The recursive algorithm designed to estimate the optimal weights is presented in the display Algorithm 1 below. We start with a uniform unity weight function $\mathbf{w}_t(\boldsymbol{\rho}) = 1 \; \forall \; t$ and optimize the likelihood to obtain the standard MLE. Then we optimize the weights using (4). Conditional on the new $\hat{\boldsymbol{\rho}}$, we can obtain a new WMLE. We iterate until convergence. In Algorithm 1, $Q_H(\boldsymbol{\rho})$ denotes the n-step-ahead forecasting performance criterion chosen to optimize the weights, of which the MSFE in (4) is a special case.

This simple steepest-ascent algorithm has revealed itself to be fast and stable in our Monte Carlo experiments reported in Appendix C and in our empirical application in Section 5. Since the algorithm is initiated at the classical ML estimates, the WMLE will feature non-uniform weights and differ from the MLE only when there is scope for improvement over the MLE.

4   Theoretical Foundations for the Weighted MLE

The cross-validation optimization described in Algorithm 1 is intuitively appealing as it leads to weights that improve the out-of-sample forecasting performance rather than the in-sample

ALGORITHM 1.

---

1. Set $\hat{\boldsymbol{\rho}}_1$ such that $w_{k,t}(\hat{\boldsymbol{\rho}}_1) = 1 \ \forall \ t \leq k$, and obtain the ML estimates $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}_1))$, $k = T', ..., T - n$.

2. Given the ML estimates $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}_1))$, $k = T', ..., T - n$, update the weights $\hat{\boldsymbol{\rho}}_2$ in the direction of steepest descent of the criterion function $Q_H(\boldsymbol{\rho})$.

3. For $j \geq 2$

   If $Q_H(\hat{\boldsymbol{\rho}}_j) < Q_H(\hat{\boldsymbol{\rho}}_{j-1})$:

   3.1. Given $\hat{\boldsymbol{\rho}}_j$ obtain $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}_j))$, $k = T', ..., T - n$.

   3.2. Given the ML estimates $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}_j))$, $k = T', ..., T - n$, update the weights $\hat{\boldsymbol{\rho}}_{j+1}$ in the direction of steepest descent of the criterion function $Q_H(\boldsymbol{\rho})$.

   3.3. Repeat step 3 with $j = j + 1$.

   If $Q_H(\hat{\boldsymbol{\rho}}_j) \geq Q_H(\hat{\boldsymbol{\rho}}_{j-1})$:

   3.4 Collect $(\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\rho}}_{j-1}), \hat{\boldsymbol{\rho}}_{j-1})$ and stop iterating.

---

fit of the AR(p) model. Below we provide theoretical foundations for this procedure.

First, we analyze the WMLE as a generalization of the classical MLE. In particular, we show that if the AR(p) model is well specified, then the WMLE is asymptotically equivalent to the MLE, and hence uncovers the true parameter vector and minimizes forecast errors. On the other hand, we also show that, if the model is misspecified, then there exist non-uniform weights that improve upon the MLE parameter estimates in terms of forecasting performance. Furthermore, we show that our algorithm for finding optimal weights delivers that WMLE outperforms the MLE under very general conditions. Second, we give conditions under which our cross-validation procedure delivers a WMLE that provide optimal forecasting performance. These results apply to a wide range of forecasting performance criteria. Third, we implement a Diebold-Mariano test that can be used to infer whether the improvements in forecasting accuracy from a change in weights are statistically significant or not and we analyze the validity of the asymptotic distribution of the statistic.

4.1 Asymptotic Properties of WMLE

The WMLE can be used to describe instability in the parameters of AR(p) models. For example, the WMLE class has recursive estimators as well as rolling window estimators as special cases. Proposition 1 shows that many DGPs can be written in the form of a time-varying parameter AR(p) model with Gaussian innovations. For simplicity, we restrict our attention to a DGP with stochastic contracting dynamics. This allows us to apply laws of large numbers and central limit theorems.

The contracting dynamics are not necessary for the theory that follows, but they are sufficient. As such, these results can be extended to a host of other settings that allow for heterogeneous dynamics, deterministic components, etc. Apart from the general contracting behavior, Theorem 1 still allows for a DGP with very general dynamics. Indeed, X_t can depend nonlinearly on its past, as well as on a potentially very large vector V_t of variables that may include not only innovations and random breaks, but also a wide range of exogenous strictly stationary and ergodic (SE) variables with complex dynamics and temporal dependence patterns.

The SE nature and bounded moments of the process $\{X_t\}_{t\in\mathbb{Z}}$ are obtained using Theorem 3.1 of Bougerol (1993) and Theorem A10.1 of Blasques et al. (2014). The SE nature of the time-varying AR(p) parameters is a direct consequence of Krengel's Theorem. Below, we let $\mathbb{C}^1(\mathcal{X} \times \mathcal{V})$ denote the space of real-valued continuously differentiable functions defined on the set $\mathcal{X} \times \mathcal{V}$. Similarly, $\mathbb{L}(\mathcal{X})$ denotes the space of Lipschitz continuous functions defined on $\mathcal{X}$, and $\mathbb{L}_V(\mathcal{X}) := \{\mathbb{L}_v(\mathcal{X}), v \in \mathcal{V}\}$ denotes the class of functions that are Lipschitz on $\mathcal{X}$ uniformly over $v \in V$.

PROPOSITION 1. *Let $\{X_t\}_{t\in\mathbb{Z}}$ generated according to*

$$X_t = \phi(X_{t-1}, V_t) \quad , \quad t \in \mathbb{Z} \tag{5}$$

where

(i) $\{V_t\}_{t\in\mathbb{Z}}$ is an SE n_V-variate stochastic sequence

(ii) $\phi \in \mathbb{C}^1(\mathcal{X} \times \mathcal{V})$ and $\phi \in \mathbb{L}_V(\mathcal{X})$;

(iii) $\mathbb{E}|\phi(x, V_t)|^4 < \infty$ for some $x \in \mathcal{X}$; and

16

(iv) $\mathbb{E}\sup_{x \in \mathcal{X}} |\phi'_x(x, Z_t)|^4 < 1.$

Then the following time-varying AR(p) representation holds

$$X_t = \alpha_{0,t} + \alpha_{1,t}X_{t-1} + ... + \alpha_{p,t}X_{t-p} + \epsilon_t \quad , \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \quad , \quad t \in \mathbb{Z} \qquad (6)$$

where $\{\alpha_{i,t}\}_{t \in \mathbb{Z}}$ is SE for every $i = 1, ..., p$, and furthermore $\{X_t\}_{t \in \mathbb{Z}}$ is also SE and has four bounded moments $\mathbb{E}|X_t|^4 < \infty$.

4.2   Conditions for Optimal WMLE Forecasting

A distinct feature of the WMLE estimator is the fact that it reduces to the MLE when the weights are unnecessary or undesirable. Proposition 2 shows precisely the weights converge in probability to unity when the model is well specified. Specifically, the cross-validation method that we propose for estimating the weights ensures that the weighted likelihood function converges in probability to the classical likelihood function as the size of the estimation sample $S := T' - p$ and cross-validation sample $H := T - T' - n$ diverge to infinity sequentially. The Monte Carlo evidence reported in Appendix C confirms that the weights remain close to unity, even in finite sample dimensions that are typical in empirical studies.

PROPOSITION 2. *Let $\mathcal{W}$ be compact and suppose the conditions of Proposition 1 hold with $V_t = \epsilon_t \ \forall \ t \in \mathbb{Z}$ and*

$$\phi(X_{t-1}, \epsilon_t) = \alpha_0 + \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \epsilon_t \ \forall \ t \in \mathbb{Z}.$$

Then the MSFE criterion in (3) ensures that $\hat{\mathbf{w}}_{k,t} \xrightarrow{p} 1 \ \forall \ k$ and $\forall \ t \leq k$, as $S \to \infty$ and $H \to \infty$ sequentially, for any given forecasting horizon $n \geq 1$ and lag order $p \geq 1$.

By application of Berge's Maximum Theorem, we obtain as a corollary that WMLE converges in probability to the MLE as the cross-validation sample H diverges to infinity. When both the cross-validation sample H and the estimation sample S diverge to infinity, WMLE converges to the true parameter $\boldsymbol{\theta}_0 \in \Theta$, just as MLE does. The Monte Carlo evidence reported in Appendix C reveals that WMLE performs well in finite samples.

COROLLARY 1. *Let the conditions of Proposition 2 hold. Then* $\|\hat{\boldsymbol{\theta}}(\mathbf{w}_k) - \hat{\boldsymbol{\theta}}(1)\| \overset{p}{\to} 0$ *as* $H \to \infty$ *and* $\hat{\boldsymbol{\theta}}(\mathbf{w}_k) \overset{p}{\to} \boldsymbol{\theta}_0$ *as* $S \to \infty$ *for* $k = T', ..., T$ *and any given* $n \geq 1$.

Under incorrect model specification, recursive or rolling-window estimators, can often improve upon full-sample estimators by allowing for time-varying parameters that better capture the dynamics of the data at any given period of time. Similarly, the WMLE will be able to improve the forecasting performance of the AR(p) by allowing for time-varying parameters that can improve the out-of-sample performance of the model. The existence of such a sequence of parameters is another simple, albeit important and general, consequence of Proposition 1. Below we let $\text{MSFE}_n(\boldsymbol{\theta})$ denote the n-step ahead mean squared error achieved by the AR(p) model under some parameter vector $\boldsymbol{\theta} \in \Theta$,

$$\text{MSFE}_n(\boldsymbol{\theta}) = \mathbb{E}_t\Big(X_{t+n} - \hat{X}_{t+n}(\boldsymbol{\theta})\Big).$$

COROLLARY 2. *Let the conditions of Proposition 1 hold, and suppose that*

$$\phi(X_{t-1}, V_t) \neq \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \epsilon_t \quad , \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \ ,$$

for every $\boldsymbol{\theta} \in \Theta$ *and some* $t \in \mathbb{Z}$. *Then there exists a non-constant sequence* $\{\boldsymbol{\theta}_t\}_{t \in \mathbb{Z}}$ *of points in* Θ *such that* $\text{MSFE}_n(\boldsymbol{\theta}_t) < \text{MSFE}_n(\boldsymbol{\theta})$ *for any given* $\boldsymbol{\theta} \in \Theta$ *and* $n \geq 1$.

Corollary 2 highlights that time-varying parameters can improve the forecasting of the AR(p) when the model is a simplistic representation of the data. Corollary 1 reveals that the WMLE will only deliver time-varying parameters in large samples, when the model is well specified. Proposition 3 below focuses on the properties of the WMLE algorithm proposed in the previous section. First, it highlights that the algorithm is designed to ensure that the WMLE outperforms (or is at least as good as) the MLE in terms of the forecasting accuracy of the AR(p) model in the cross-validation sample. Furthermore, Proposition 3 shows that under appropriate regularity conditions, the WMLE algorithm will actually uncover the weights that optimize the forecasting performance of the AR(p) model in the cross-validation sample. We let $\hat{\boldsymbol{\rho}}_j$ denote the j-th iteration weights and we let $Q_H(\boldsymbol{\rho})$ denote the mean

squared error in the cross-validation sample obtained under $\boldsymbol{\rho}$. We obtain

$$Q_H(\boldsymbol{\rho}) := \frac{1}{H} \sum_{k=T'}^{T-n} \left(\hat{X}_{k+n}\big(\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\boldsymbol{\rho}))\big) - X_{k+n} \right)^2.$$

The Monte Carlo evidence reported in Appendix C and the empirical evidence for U.S. Industrial Production in Section 5 reveal that the WMLE is indeed capable of significantly improving the forecasting performance of the $AR(p)$ model.

PROPOSITION 3. *For any given realized sample* $\{x_t\}_{t=1}^{T}$, *Algorithm 1 ensures that*

$$Q_H(\boldsymbol{\rho}_{j+1}) \le Q_H(\boldsymbol{\rho}_j), \qquad \forall\, j \ge 1,$$

and hence the WMLE outperforms the MLE under the Q_H criterion.
If furthermore it holds that

$$\sup_T \sup_{\boldsymbol{\rho}} \left| \partial \hat{\boldsymbol{\theta}}(\boldsymbol{\rho})/\partial \boldsymbol{\rho} \right| < 1, \quad and \quad \sup_T \sup_{\boldsymbol{\theta}} \left| \partial \hat{\boldsymbol{\rho}}(\boldsymbol{\theta})/\partial \boldsymbol{\theta} \right| < 1.$$

Then $\boldsymbol{\rho}_j \to \boldsymbol{\rho}^*$ *and* $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ *as* $j \to \infty$, *for any given* $n \ge 1$.

The two main conditions of Proposition 3 ensure the contraction of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\rho}}$ as maps $\hat{\boldsymbol{\theta}} : \boldsymbol{\rho} \mapsto \boldsymbol{\theta}$ and $\hat{\boldsymbol{\rho}} : \boldsymbol{\theta} \mapsto \boldsymbol{\rho}$. Since these maps are not known analytically, the contracting behavior can only be verified numerically. This can be achieved by optimizing the derivatives stated above, and ensuring that there maximum is less than one.

4.3   Test for Forecast Precision Improvement

The result established in Proposition 3 is important, but it ensures only that the WMLE improves the finite sample MSFE. In other words, the algorithm discussed in Section 3.3 delivers weights that optimize the forecasting performance within the cross-validation sample. However, due to sampling error it is impossible to ensure that the true forecasting performance has improved from the MLE to the WMLE with some weight matrix $\mathbf{W}$. We will use a simple Diebold-Mariano (DM) test statistic to assess whether the improvement in forecasting performance is statistically significant or not. Proposition 4 highlights the validity of the

asymptotic distribution derived by Diebold and Mariano (1995) under the assumptions of Proposition 1 when the null hypothesis compares the MLE against an alternative WMLE.

Below, we let MSFE($\mathbf{W}$) denote the MSFE achieved by the AR(p) model under the WMLE with weight matrix $\mathbf{W}$, and let $\mathbf{W}^*$ denote the best possible WMLE weight matrix for the AR(p) model

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} MSFE(\mathbf{W}).$$

Furthermore, we let MSFE(1) denote the MSFE achieved by the MLE. Under correct model specification, we naturally have that MSFE(1)=MSFE($\mathbf{W}^*$). Proposition 4 states a DM test with a null hypothesis of correct specification (both MLE and WMLE provide equal forecasting accuracy)

$$\mathrm{H}_0 : \mathrm{MSFE}(\mathbf{W}^*) = \mathrm{MSFE}(1)$$

against an alternative of incorrect specification (WMLE provides improved forecast accuracy)

$$\mathrm{H}_1 : \mathrm{MSFE}(\mathbf{W}) < \mathrm{MSFE}(1).$$

Diebold (2015) discusses the question whether the DM assumptions hold in practice and argues that it is an empirical issue for which there exist tests that one may wish to employ. Below we let $\bar{d}_k(\mathbf{w}_k)$ and $\mathrm{Ste}\big(d_k(\mathbf{w}_k)\big)$ denote the sample average and standard error of the MSFE difference between WMLE and MLE, at period k,

$$\bar{d}_k(\mathbf{w}_k) := u_k(1)^2 - u_k(\mathbf{w}_k)^2 \quad \text{where} \quad u_k(\mathbf{w}_k) := \big(\hat{X}_{k+n}\big(\hat{\boldsymbol{\theta}}(\mathbf{w}_k)\big) - X_{k+n}\big)^2.$$

PROPOSITION 4. *Let the conditions of Proposition 1 hold. Then*

$$\bar{d}_k(\mathbf{w}_k) \,/\, \mathrm{Ste}\big(d_k(\mathbf{w}_k)\big) \xrightarrow{d} N(0,1), \quad as \quad T \to \infty,$$

for any given pair (T', n) *under the null hypothesis* $\mathrm{H}_0 : \mathrm{MSFE}(\mathbf{W}^*) = \mathrm{MSFE}(1)$, *and*

$$\bar{d}_k(\mathbf{w}_k) \,/\, \mathrm{Ste}\big(d_k(\mathbf{w}_k)\big) \to \infty, \quad as \quad T \to \infty,$$

under the alternative hypothesis $\mathrm{H}_1 : \mathrm{MSFE}(\mathbf{W}) < \mathrm{MSFE}(1)$.

We regard the DM test as the natural tool for comparing the forecasts produced under two WMLE estimates. In the more recent contributions of Giacomini and White (2006) and Diebold (2015), the bulk of the literature has moved from *testing forecasts* to *testing models*, evaluated at their pseudo-true parameters; see, for example, West (1996), and Clark and McCracken (2001). Clearly, here we are not interested in testing different models, as all the forecasts come from the same AR(p) model. Instead, we are interested in testing the forecasts obtained under the different parameter estimates produced by WMLE and MLE.

4.4 Monte Carlo Evidence

We verify in a set of simulation experiments whether the small-sample properties match our theoretical findings. We investigate the finite sample performance of WMLE in the context of the basic AR(1) model and by means of our Monte Carlo simulation experiments. We consider four data generating processes (DGP) for the time series y_t, these are four departures of the AR(1) model : (1) the basic AR(1) model with no departures, (2) with a time-varying AR coefficient, (3) with a structural break for the AR coefficient, and (4) a regime-switching AR coefficient. For each case, we assess whether the optimal weight function does improve the one-step-ahead forecasting accuracy of the MLE. The forecasts are based on rolling-window estimates. The WMLE method is only applied to the standard AR(1) model. Two different weight functions are adopted: (a) the exponential weight function with $w_t = \rho_1^{k-t}$ with $\rho \in [0, 1]$, and (b) the binary weight function with decay $w_t = \rho_1^{(k-t)}(1 + (\rho_2 - 1) \cdot Z_t)$ where Z_t is a predetermined indicator for the recession period. The exponential weight method (a) is used in experiments (1), (2) and (3), while the binary weight method (b) is used in experiment (4). More details for the Monte Carlo experiments and their results are presented in Appendix C. The overall conclusions are as follows. In case the AR model is well-specified as in DGP (1), we do not find any significant improvements in forecasting accuracy. In effect, WMLE and MLE results are overall the same. However, in all three other cases where the AR model is not correctly specified for the DGP, the WMLE succeeds in producing higher maximized likelihood function values using non-uniform weight functions. In particular, the increases in forecast precisions of WMLE compared to MLE are high and they confirm the asymptotic findings in this Section.

5 Empirical Illustrations

We study the empirical performance of our WMLE procedure for several economic relevant applications. In Section 5.1, we focus on forecasting the yearly growth rate of monthly U.S. Industrial Production Index (IPI) during the global recession of 2008. We find that the WMLE can deliver a significantly better forecasting performance than the standard MLE method. This result is achieved by increasing the weights associated with observations coming from past recession periods. Moreover, we find that the improved out-of-sample forecasting accuracy delivered by WMLE is not driven by a single, or just a few, observations. Instead, the improved forecasting performance is experienced throughout the entire validation sample. In Sections 5.2 and 5.3 we show that the improved forecasting performance delivered by WMLE is not restricted to forecasting the IPI during a recession period. In particular, Section 5.2 shows that the WMLE outperforms the MLE also during expansion periods. Section 5.2 also illustrates that it is often important to allow for more complex weight functions that allow for both ρ_1 and ρ_2 to be different from unity. In Section 5.3, we show that the WMLE outperforms the MLE in other data sets as well. In effect, it reveals that the WMLE method can, in some cases, improve the forecasting performance of the model in a remarkable fashion. Finally, in Section 5.4, we focus on forecasting during the COVID-19 recession. We find that even when simple exponential weights are used the WMLE method can improve substantially the forecast accuracy especially in the presence of a structural break.

5.1 Forecasting Industrial Production During Great Recession

We study whether the WMLE method improves the forecast accuracy for the growth in the U.S. Industrial Production Index (IPI), when compared to the MLE method. The IPI growth rate is a core indicator of the U.S. economy. The time series of U.S. IPI is monthly and our sample covers the period from January 1950 until December 2009. We evaluate measures of forecasting accuracy in terms of root mean squared error (RMSE) and mean absolute error (MAE). We compare these for MLE and WMLE by means of the Diebold-Mariano (DM) test for relative predictive accuracy.

We consider two window lengths of 25 years (period 1975–1999, contains 46 recession

months) and 50 years (period 1950–1999, 85 recession months) for the estimation of the parameters. We have 8 years (period 2000–2007, 8 recession months) for the cross-validation of the forecasting accuracy. Finally, we take 1.5 years (period 2008–2009 June, contains 18 recession months) for the out-of-sample forecasting period. We report the results for the forecast horizons of one month ($n = 1$) and one quarter ($n = 4$) ahead. The weight functions are flexible and allow for a combination of binary and exponentially decaying weights. The binary weights can be used to emphasize the weight of past recessions in the estimation of parameters. The intuition behind this weighting scheme is that IPI observations coming from past recession periods may be more informative about the dynamics of the IPI during the global recession of 2008–2009. The exponential weights can be used to weigh down observations from the remote past. The intuition that underlies the exponential decay is straightforward: as the economy changes, observations far in the past are likely to be less informative about the current dynamics of IPI than the more recent observations. Specifically, the weight function for $\mathbf{w}_k$ takes the form

$$w_{k,t} = \rho_1^{(k-t)}(1 + (\rho_2 - 1) \cdot Z_t), \qquad 0 \leq \rho_1 \leq 1, \quad \rho_2 \geq 1,$$

where the indicator variable $Z_t \in \{0, 1\}$ is the NBER recession indicator. Setting $\rho_1 = \rho_2 = 1$ delivers standard MLE (unweighted). Naturally, when $\rho_1 < 1$, past observations receive less weight than recent ones. Similarly, when $\rho_2 > 1$, observations from recession periods receive more weights for estimation. Since the coefficients ρ_1 and ρ_2 are estimated using our cross-validation method, we will infer from the data how strong these effects are.

For simplicity, we first set $\rho_1 = 1$ and apply the WMLE method with a binary weight function to an AR(1) model and an AR(p) model of which the number of lags p is selected by means of the Akaike's information criterion. We refer to the second model as the AR(AIC) model. We focus on out-of-sample forecasting of the monthly U.S. IPI growth rate during the NBER global recession period which spans from January 2008 to June 2009. We emphasize that the sample for which the forecasting accuracy was evaluated was not used for estimating the AR(1) parameters or the likelihood weights. Hence, the MLE estimator can potentially outperform the WMLE estimator in this separate validation sample. Any improvement in forecasting accuracy should only be expected if the WMLE method presents a substantial

estimation advantage over the MLE.

The results are presented in Table 1 where we take the recession period in the year 2000 as our cross-validation period to determine the WMLE estimates that are used in the global recession of 2008. We report the RMSE and MAE ratios for the WMLE against MLE methods such that a ratio value smaller than unity indicates WMLE improvements in forecast accuracy. We also report the p-values of the DM test (in parentheses) and the cross-validation estimates of the WMLE parameter ρ_2.

Table 1: AR Model Forecasting: Binary Weights for Recessions

	RMSE				MAE			
	$n=1$		$n=4$		$n=1$		$n=4$	
	Ratio (DM)	ρ_2	Ratio (DM)	ρ_2	Ratio (DM)	ρ_2	Ratio (DM)	ρ_2
AR(1)								
25 years	1.0908 (0.3446)	35.5	0.8297* (0.0679)	27.7	0.8310 (0.1564)	31.5	0.7919* (0.0500)	27.1
50 years	0.9102** (0.0262)	15.7	0.7430** (0.0241)	12.3	0.8007*** (0.0011)	13.8	0.5975*** (0.0000)	11.1
AR(AIC)								
25 years	1.4909 (0.1460)	137.5	0.9189 (0.2706)	11.0	1.4039 (0.1785)	228.5	0.9450 (0.3918)	8.8
50 years	0.9094* (0.0667)	15.9	0.7782* (0.0523)	9.7	0.8427** (0.0153)	25.7	0.6955*** (0.0021)	11.5

Table 1: WMLE forecasting results for monthly U.S. IPI growth rate for the AR(1) model and the AR(p) model where p is imposed by the Akaike's information criterion, this is AR(AIC). Rolling window samples have length 25 and 50 years, starting from periods 1975-1999 and 1950-1999, respectively. The column labeled Ratio shows the RMSE and MAE using the WMLE method of the AR model relative to the RMSE and MAE, respectively, using the MLE method. Cases where the forecasting improvement is statistically significant at the 90%, 95%, and 99% confidence levels are indicated by *, **, and ***, respectively. Entries in parentheses show the p-value of DM test. The column labelled "ρ_2" contains its WMLE parameter estimates for the binary weighting $(1 - \rho_2) \cdot Z_t$ where $Z_t \in \{0, 1\}$ is the NBER recession indicator. The cross-validation sample is 2000–2007. The out-of-sample forecasts are computed for 2008–2009 June (18 months).

The results from Table 1 reveal that WMLE strongly outperforms MLE in the majority of cases. The improvements of up to 31% in the forecasting accuracy of the AR(AIC) are quite remarkable. The p-values of the Diebold-Mariano statistics reveal that forecast accuracy

improvements relative to MLE are statistically significant at standard confidence levels. Most importantly, the parameter estimates of ρ_2 indicate that the improved performance is due to giving more weight to past recession periods. We want to highlight that the improved forecasting performance is not only achieved for the simple AR(1) model, which is likely to be a misspecified model, but also for the AR(AIC) model, which in all cases has $p > 1$ in the AR(p) model and is therefore more flexible as it includes more lags.

The cross-validation estimates of ρ_2 vary over a wide range of values larger than one. The more reasonable, smaller estimates of ρ_2 lead to better results than the more extreme ones (say those larger than 30). For example, consider the AR(AIC) model and the rolling window estimation sample of 50 years, for the RMSE of the quarterly forecasting horizon ($n = 4$) we have the estimate of 9.7 for ρ_2. The estimate of 9.7 implies that observations coming from a recession period receive almost 9 times more weight than observations coming from an expansion period. Hence an observation in the recession year of 2000 is given approximately the same weight as an observation in the 10 expansion years before 2000. Most importantly, this weighting scheme leads to significant improvements in the RMSE as highlighted by the ratio of 0.778. In contrast, the ρ_2 estimate of 137.5 (obtained for the monthly forecast horizon ($n = 1$) and the rolling window of 25 years) implies essentially that only the observations in recession periods have an impact on the estimation; the data coming from expansion periods are ignored. Such extreme estimates of ρ_2 lead to poor results in out-of-sample forecasting evaluations when compared to the standard MLE method.

To illustrate that the results reported in Table 1 are not driven by a single or by only a few observations, we present in Figure 3 the accumulated RMSE statistics over the 18 months of our forecasting window of 2008 January to 2009 June, for both the MLE and WMLE methods. Figure 3 shows these results for both the AR(1) and AR(AIC) model. They clearly show that the performance of the WMLE method is preferred, in terms of RMSE, for all time periods in the forecasting window. It is a strong empirical result that the WMLE achieves better forecasting accuracy by outperforming the MLE over the entire forecasting evaluation sample. Finally, the results for the weight function that combines the exponential decaying weight (ρ_1) and the binary weight for recession periods (ρ_2) are reported in Appendix C. From these results we can conclude that allowing for $0 < \rho_1 \leq 1$ only leads to marginal improvements when compared to the results in Table 1.
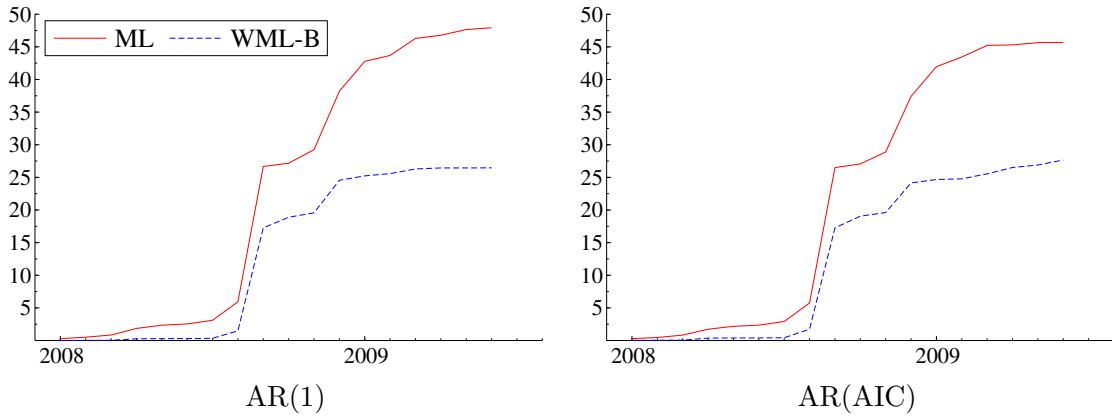
Figure 3: The accumulated root mean squared forecasting error from the MLE and WMLE methods during the forecasting window of 2008–2009 June. These results are presented for both the AR(1) and AR(AIC) models with a 50-year rolling window. The forecasts are computed for a forecasting horizon of $n = 4$ periods.

5.2 Forecasting Industrial Production During 2008-Expansion

Table 2: AR Model Forecasting: for Expansion Period,
with Exponential Decay (ρ_1) and Binary (ρ_2) Weights

	$n=1$			$n=4$		
	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2
	(DM)			(DM)		
	0.9708	0.977	53.6	0.9413*	1.000	2.2
	(0.2748)			(0.0986)		
AR(AIC)	0.9126*	0.994	79.2	0.8709	0.998	8.8
	(0.0983)			(0.1609)		

Table 2: WMLE forecasting results for monthly U.S. IPI growth rate. The ratios are based on the RMSE statistics and a rolling window size of 25 years. The forecasts are computed for the expansion period 2001–2007. For further explanations, we refer to Table 1.

Given the encouraging results as reported above, we have concluded that the WMLE method can lead to improved forecasting performances for IPI growth during recession periods. We next investigate whether these improvements are not specific to the global financial recession period only. Does the WMLE method also provide significant improvements in forecasting accuracy during expansion periods? For this purpose, we replace the recession indicator Z_t by the expansion indicator $(1 - Z_t)$ and take the NBER expansion period from January 2001 to December 2007 as our forecasting evaluation sample. The results are re-

ported in Table 2 that only focuses on the RMSE ratios and a rolling window size of 25 years; the full results, including the MAE ratios and the rolling window size of 50 years, are provided in Appendix C. We learn from Table 2 that the WMLE method delivers higher out-of-sample forecasting precisions in the evaluation sample when compared to those from the MLE method. In two of these cases, the DM p-values show improvements that are marginally statistically significant at the 90% confidence level.

In particular, Table 2 presents overall improvements in forecasting accuracy of up to 13%. Although the magnitude of these improvements is not as impressive as in the recession case, we still find statistically significant improvements in terms of the DM test. The improvements are especially relevant when related to the forecasts of the AR(AIC) model which is a highly competitive benchmark in many forecasting studies. In any case, improvements in forecasting accuracy are expected to be smaller in the expansion case rather than in the recession case. The main reason is that the number of observations originating from economic expansion periods outnumbers by far the number of observations originating from recession periods. As a result, the total weight that expansion periods have in the likelihood criterion is much larger. It leads to ML parameter estimates that are already adequate for expansion periods. As a result, there is a smaller margin for obtaining an improvement for the WMLE method.

5.3 Forecasting Other Economic Variables During 2008-Recession

To show that the WMLE method can also be effective for other relevant economic time series in delivering notable improvements in forecasting accuracy, we present two additional illustrations. We consider the monthly economic time series of the U.S. Unemployment Rate and U.S. Total Non-Farm Payrolls and focus again on forecasting these variables during the global recession of 2008. In Table 3 we report the RMSE and MAE ratios obtained from the same out-of-sample forecasting study as done for U.S. Industrial Production. The tables with full estimation results for the two time series, including the estimates for ρ_1 and ρ_2, are presented in Appendix C.

In the upper part of Table 3, the RMSE and MAE ratios are presented for the monthly time series of the U.S. Unemployment Rate. These results show that the WMLE is capable of delivering significant reductions in the out-of-sample forecasts. In particular, we report

RMSE reductions of up to 34% for the monthly forecasts from the AR(1) model, and MAE reductions of up to 23% for the forecasts from the AR(AIC) model, both at a quarterly horizon. The lower part of Table 3 is for the forecasting of the monthly time series of the U.S. Non-Farm Payrolls during the global recession. For the AR(1) model, remarkable reductions of up to 48% in the out-of-sample forecasting RMSE are obtained, and reductions of more than 50% in the MAE, at both monthly and quarterly forecasting horizons. Even more convincingly, for the AR(AIC) we find reductions of more than 25% in the RMSE and over 40% in the MAE. Most of these reductions are statistically significant at any reasonable confidence level.

Table 3: AR Model Forecasting: Other Economic Time Series

	RMSE Ratio		MAE Ratio	
	$n=1$	$n=4$	$n=1$	$n=4$
U.S. Monthly Unemployment Rate				
AR(1)	0.6814***	0.7746*	0.7162**	0.7421*
	(0.0023)	(0.0788)	(0.0266)	(0.0586)
AR(AIC)	0.8710	0.8480	0.8235*	0.7780
	(0.1136)	(0.2922)	(0.0915)	(0.1567)
U.S. Monthly Total Non-farm Payrolls				
AR(1)	0.5212***	0.5455**	0.4401***	0.4582*
	(0.0000)	(0.0155)	(0.0000)	(0.0003)
AR(AIC)	0.6700***	0.6369	0.6087***	0.5659
	(0.0060)	(0.1715)	(0.0009)	(0.1538)

Table 3: WMLE forecasting results for two other economic time series. The forecasts are computed over the sample period 2008 January - 2009 June. We refer to Table 1 for further explanations.

5.4 Forecasting During the COVID-19 Recession

We also study the forecasting performance of the WMLE method during the COVID-19 recession. The COVID-19 global recession is one of the deepest recessions in the world history and it is characterized by a huge decline in economic activity in many sectors as well

as increased uncertainty. Therefore, in our application, we consider two time series: the U.S. Economic Policy Uncertainty Index (EPUI) proposed by Baker et al. (2016) and the first differences of the U.S. Weekly Economic Index (WEI) constructed by Lewis et al. (2020). The former is based on the U.S. newspapers and it characterizes policy uncertainty, while the latter is constructed using ten indicators of the U.S. real economic activity which cover consumer behavior, labor market, and production.

For the daily time series of the U.S. EPUI, we take the estimation and cross-validation samples spanning from 1 February 2018 until 1 February 2020, and for the U.S. Weekly Economic Index (WEI) from 5 January 2008 until 1 February 2020 with the rolling windows' length of 550 days and 533 weeks, respectively. Then for the cross-validation, we have 180 days (around half a year) for the EPUI and 96 weeks (around 2 years) for the WEI. Starting from 1 February 2020, right before the recession began, we then use 121 days (3 months) and 33 weeks, respectively, for the out-of-sample forecasts. To compare the forecasting performance we consider one-step-ahead forecasts ($n = 1$) and exponentially decaying weights for the WMLE.

In Table 4 we report the RMSE and MAE ratios for the WMLE against the MLE method. We find that for the U.S. EPUI series the WMLE method substantially outperforms the standard MLE method and the improvements are of up to 16%. Moreover, this result is significant at a 1% level. These improvements most likely occur due to the presence of a structural break in the series as during the coronavirus pandemic the policy uncertainty increased substantially since the middle of March 2020, shortly after the recession started. Therefore, the data far in the past became less relevant for forecasting and receive much lower weights since the end of March 2020. Furthermore, in Figure 4 the accumulated RMSEs for MLE and WMLE are similar in February but, after the break, the accumulated RMSE based on the MLE is much higher, which again supports the idea that after the break the AR model is misspecified and then the WMLE method is preferable for forecasting. This is in line with the findings of the Monte Carlo experiment B.3 with a structural break.

For the U.S. WEI, the WMLE method slightly outperforms the standard MLE based on the RMSE but the difference in the forecast accuracies is not significant (Table 4). In Figure 4 we also observe that at one moment the WMLE outperforms MLE but overall their performance is comparable. This result could be explained by the fact that the series

29

exhibited a large downturn at the end of March and after that quickly started bouncing back. The WMLE method reacts quicker to the downturn but then the difference in forecast accuracies disappears. This is in line with the theoretical findings discussed in Section 4 that in some cases the WMLE method simplifies to a standard MLE. Overall we can conclude that the use of the WMLE method appears to be highly beneficial for accurate economic forecasting.

Table 4: AR Model Forecasting: COVID-19 recession

	RMSE Ratio $n=1$	MAE Ratio $n=1$
U.S. Daily Economic Policy Uncertainty Index		
AR(1)	0.8184***	0.8355***
	(0.0014)	(0.0062)
U.S. Weekly Economic Index		
AR(1)	0.9875	1.0138
	(0.6831)	(0.8780)

Table 4: WMLE forecasting results for two economic time series during the COVID-19 recession. The forecasts are computed over the sample period 1 February 2020–31 May 2020 and 1 February 2020–12 September 2020 for the EPUI and WEI, respectively. We refer to Table 1 for further explanations.
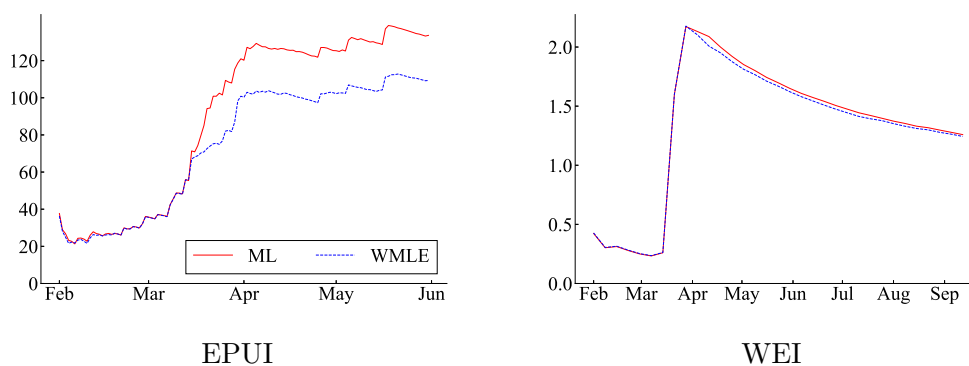

EPUI   WEI

Figure 4: The accumulated root mean squared forecasting error for the MLE and WMLE methods during the forecasting window of 1 February 2020–31 May 2020 for the EPUI and 1 February 2020–12 September 2020 for the WEI. The forecasts are computed for a forecasting horizon of $n = 1$ periods.

6 Conclusion

We have introduced a new weighted maximum likelihood estimation (WMLE) method that weighs the likelihood function contributions of individual observations differently for the purpose to deliver optimal forecasting accuracy for linear autoregressive models. We have shown how to estimate the optimal weights using a cross-validation method. We further have investigated the asymptotic properties of WMLE and we have considered four Monte Carlo experiments for studying the finite-sample properties. In empirical illustrations, we have analyzed the forecast accuracy of the WMLE method compared to standard MLE for several key economic indicators during the global recession of 2008–2009, the 2008 expansion, and the COVID-19 recession. The analyses have revealed that the WMLE can substantially improve forecast accuracy. In particular, forecast precision during recession periods can be significantly improved by increasing the weights of the most recent observations or by increasing the weights of observations corresponding to similar recession periods in the past. We have made the case that econometricians may need to give past recessions more attention for providing more accurate forecasts during periods of recessions and financial crises. In further work we can extend the method towards other linear dynamic models, including the autoregressive moving average model, and towards multivariate specifications such as the vector autoregressive model. Nonlinear extensions of the model classes can also be considered but may require an extended theoretical foundation.

References

Akcoglu, M. A. and U. Krengel (1979). A differentiation theorem in lp. *Mathematische Zeitschrift 169*, 31–40.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The quarterly journal of economics 131*(4), 1593–1636.

Blasques, F., S. J. Koopman, and A. Lucas (2014). Stationarity and ergodicity of univariate generalized autoregressive score processes. *Electron. J. Statist. 8*(1), 1088–1112.

Blasques, F., S. J. Koopman, and A. Lucas (2015). Information theoretic optimality of observation driven time series models for continuous responses. *Biometrika 102*(2), 325 – 343.

Bougerol, P. (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization 31*(4), 942–959.

Clark, T. E. and M. W. McCracken (2001, November). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics 105*(1), 85–110.

Clements, M. and D. F. Hendry (1998). *Forecasting Economic Time Series*. Number 9780521634809 in Cambridge Books. Cambridge University Press.

Diebold, F. 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.

Diebold, F. X. and R. S. Mariano (1995, July). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics 13*(3), 253–63.

Giacomini, R. and B. Rossi (2009). Detecting and Predicting Forecast Breakdowns. *Review of Economic Studies 76*(2), 669–705.

Giacomini, R. and H. White (2006, November). Tests of Conditional Predictive Ability. *Econometrica 74*(6), 1545–1578.

Giraitis, L., G. Kapetanios, and S. Price (2012). Adaptive Forecasting in the Presence of Recent and Ongoing Structural Change. Working Papers 691, Queen Mary University of London, School of Economics and Finance.

Goyal, A. and I. Welch (2003). Predicting the Equity Premium with Dividend Ratios. *Management Science 49*(5), 639–654.

Inoue, A., L. Jin, and B. Rossi (2014). Window Selection for Out-of-Sample Forecasting with Time-Varying Parameters. CEPR Discussion Papers 10168, C.E.P.R. Discussion Papers.

Koop, G. and S. Potter (2004). Forecasting in dynamic factor models using Bayesian model averaging. *Econometrics Journal 7*(2), 550–565.

Lewis, D. J., K. Mertens, J. H. Stock, et al. (2020). Monitoring real activity in real time: The weekly economic index. Technical report, Federal Reserve Bank of New York.

Molodtsova, T. and D. H. Papell (2009). Out-of-sample exchange rate predictability with Taylor rule fundamentals. *Journal of International Economics 77*(2), 167–180.

Paye, B. S. and A. Timmermann (2006). Instability of return prediction models. *Journal of Empirical Finance 13*(3), 274–315.

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika 37*(1/2), 149–157.

Rossi, B. (2013). *Advances in Forecasting under Instability*, Volume 2 of *Handbook of Economic Forecasting*, Chapter 0, pp. 1203–1324. Elsevier.

Schinasi, G. J. and P. A. V. B. Swamy (1989). The out-of-sample forecasting performance of exchange rate models when coefficients are allowed to change. *Journal of International Money and Finance 8*(3), 375–390.

Stock, J. H. and M. W. Watson (1996). Evidence on Structural Instability in Macroeconomic Time Series Relations. *Journal of Business & Economic Statistics 14*(1), 11–30.

Stock, J. H. and M. W. Watson (2007). Why Has U.S. Inflation Become Harder to Forecast? *Journal of Money, Credit and Banking 39*(s1), 3–33.

Swanson, N. R. (1998, May). Money and output viewed through a rolling window. *Journal of Monetary Economics 41*(3), 455–474.

West, K. D. (1996, September). Asymptotic Inference about Predictive Ability. *Econometrica 64*(5), 1067–84.

Wolff, C. C. P. (1987). Time-Varying Parameters and the Out-of-Sample Forecasting Performance of Structural Exchange Rate Models. *Journal of Business & Economic Statistics 5*(1), 87–97.

Technical Appendix

A   Proofs of Theorems and Propositions

Proof of Proposition 1. Theorem 3.1 in Bougerol (1993) implies that the sequence $\{X_t\}_{t\in\mathbb{N}}$ initialized at $X_1 = x$ and generated according to (5) for every $t \in \mathbb{N}$, converges exponentially almost surely (e.a.s) to an SE limit sequence $\{X_t\}_{t\in\mathbb{Z}}$, initialized in the infinite past, as long as $\{V_t\}_{t\in\mathbb{Z}}$ is an SE n_V-variate stochastic sequence, $\phi \in \mathbb{C}^1(\mathcal{X}, \mathcal{V})$, $\mathbb{E}\log^+ |\phi(x, V_t)| < \infty$ and $\mathbb{E}\log\sup_{x\in\mathcal{X}} |\phi'_x(x, Z_t)| < 0$. The first two conditions are directly given by *(i)* and *(ii)*. The two remaining conditions are implied by *(iii)* and *(iv)*. The four bounded moments of $\{X_t\}_{t\in\mathbb{Z}}$ are ensured by conditions *(i)-(iv)* (Blasques et al. (2014), Proposition SA.1).

The AR(p) representation follows trivially by re-writing the X_t as follows

$$
\begin{aligned}
X_t &= \phi(X_{t-1}, V_t) - \epsilon_t + \epsilon_t \\
&= \frac{\phi(X_{t-1}, V_t) - \epsilon_t}{\psi_0 + \Psi(L)X_t}(\psi_0 + \Psi(L)X_t) + \epsilon_t
\end{aligned}
$$

where $\Psi(L)$ denotes the lag polynomial $\Psi(L) = \psi_1 L + \cdots + \psi_p(L^p)$, and finally defining

$$
\alpha_{i,t} := \frac{\big(\phi(X_{t-1}, V_t) - \epsilon_t\big)}{\psi_0 + \Psi(L)X_t}\psi_i \ , \quad \text{for } i = 0, 1, ..., p.
$$

The SE nature of $\alpha_{i,t}$, $i = 0, ..., p$, follows directly from Krengel's Theorem (Akcoglu and Krengel (1979)) since every $\alpha_{i,t}$ is a measurable function of SE variables.   □

Proof of Proposition 2. Under the conditions of Proposition 1, we have that $\{X_t\}_{t\in\mathbb{Z}}$ is weakly stationary. Since furthermore, the Gaussian AR(p) model is well specified, it follows immediately that the Gaussian MLE converges to the true parameter as $S \to \infty$, i.e. $\hat{\boldsymbol{\theta}}(1) \xrightarrow{p} \boldsymbol{\theta}_0^*(1) = \boldsymbol{\theta}_0$ under the usual regularity conditions.

Application of a continuous mapping theorem as $S \to \infty$ implies that

$$
u_t(\hat{\boldsymbol{\theta}}(1))^2 \equiv \hat{X}_{k+n}\big(\hat{\boldsymbol{\theta}}(1)\big) - X_{k+n} \xrightarrow{p} u_t(\boldsymbol{\theta}_0^*(1))^2 \equiv \hat{X}_{k+n}\big(\boldsymbol{\theta}_0\big) - X_{k+n}.
$$

As a result, the limit as $S \to \infty$ of the n-step-ahead MSFE forecast criterion based on H

observed forecast errors under the true parameter $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0^*(\mathbf{1})$ is given by

$$Q_H(\mathbf{1}) := \frac{1}{H} \sum_{t=1}^{H} u_t(\boldsymbol{\theta}_0^*(\mathbf{1}))^2.$$

Since $u_t(\boldsymbol{\theta}_0) \; \forall \; t$ is a measurable function of $\{\epsilon_t\}_{t \in \mathbb{Z}}$, which are SE, by Krengel's theorem (Akcoglu and Krengel (1979)) the sequence $\{u_t(\boldsymbol{\theta}_0)\}_{t \in \mathbb{Z}}$ is also SE, and an application of the ergodic theorem then yields

$$Q_H(\mathbf{1}) \xrightarrow{p} \mathbb{E} u_t(\boldsymbol{\theta}_0^*(\mathbf{1}))^2 \quad \text{as} \quad H \to \infty.$$

Finally, note that Algorithm 1 is always initialized at a weight matrix $\mathbf{W}$ satisfying $\hat{\mathbf{w}}_{k,t} = 1 \; \forall \; k$ and $\forall \; t \leq k$. As a result, in the limit as $S \to \infty$ and $H \to \infty$, the probability that $\hat{\mathbf{w}}_{k,t} \neq 1$ for some (k,t) is given by

$$\mathbb{P}(\hat{\mathbf{w}}_{k,t} \neq 1) = \mathbb{P}\Big(\mathbb{E} u_t(\boldsymbol{\theta}_0^*(\mathbf{w}_k))^2 < \mathbb{E} u_t(\boldsymbol{\theta}_0^*(\mathbf{1}))^2\Big) = 0$$

for any weight vector $\mathbf{w}_k$ with some element $w_{k,t} \neq 1$ at some pair (t,k). $\qquad \square$

Proof of Proposition 3. The first claim follows trivially from the design of Algorithm 1. The second claim follows by noting that both the AR parameter vector $\hat{\boldsymbol{\theta}}$ and the weight parameter vector $\hat{\boldsymbol{\rho}}$ satisfy a recursive relation

$$\hat{\boldsymbol{\theta}}_{j+1} = \hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\theta}}_j)) \quad \text{and} \quad \hat{\boldsymbol{\rho}}_{j+1} = \hat{\boldsymbol{\rho}}(\hat{\boldsymbol{\theta}}(\hat{\boldsymbol{\rho}}_j)) \quad \forall \; j \geq 1.$$

It is well known that the uniform unit bound on the derivative ensures the stability of the recursion towards a unique global fixed point for any initialization. $\qquad \square$

Proof of Proposition 4. By Propositions 1 and 2, the MLE can only perform as well as the WMLE if the model is well specified. As a result, the condition that $\text{MSFE}(\mathbf{W}^*) = \text{MSFE}(1)$ ensures that the model is well specified under the null.

Under the conditions of Proposition 1, the data $\{X_t\}_{t \in \mathbb{Z}}$ is SE with four bounded moments. Since the MLE and WMLE are constant over k, it follows immediately that both $u_k(1)^2$ and $u_k(1)^2$ are weakly stationary. The DM assumptions are thus satisfied under H_0

since $\text{MSFE}(\mathbf{W}^*) = \text{MSFE}(1)$ ensures also that $\mathbb{E}\bar{d}_k(\mathbf{w}_k) = 0$. □

B Simulation Experiment

We investigate the finite sample performance of the WMLE method in the context of AR(1) model using Monte Carlo simulation. We consider four different data generating processes (DGP) for the time series y_t. Among these four different AR(1) processes, the autoregressive coefficient could be time-invariant, time-varying, regime switching or subject to a structural break. We aim to investigate in which cases the optimal weight function can improve the forecasting accuracy of the MLE. We concentrate on one-step ahead forecasts based on rolling-window method with window length k. Although we consider four different data generating processes in the simulation experiment, our WMLE method is based on the ordinary AR(1) model as given by

$$y_t = \alpha + \beta y_{t-1} + \epsilon_t, \tag{7}$$

where intercept α and autoregressive coefficient β are treated as fixed and unknown, and where ϵ_t is an i.i.d. normally distributed variable with mean zero and unknown variance σ_ϵ^2.

We consider two kinds of weight functions which we apply in Monte Carlo experiment: the exponential weight function ($w_t = \rho_1^{k-t}$, where $\rho \in [0,1]$) and the binary weight function with decay ($w_t = \rho_1^{(k-t)}(1+(\rho_2-1)\cdot Z_t)$, where Z_t is the predetermined indicator for recession period). The exponential weight function is applied in the first three experiments; while the binary weight function with decay is applied in the last experiment.

B.1 Experiment 1: Time-Invariant AR(1) model

In the first experiment, the data are generated by an AR(1) model with time-invariant parameters and the WMLE parameter, ρ_1, is estimated based on Equation (7). The data generation process (DGP) is specified by:

$$\text{DGP:} \ y_t = \alpha + \beta y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{8}$$

where $\alpha = 0.13$, $\beta = 0.5$ and ϵ_ts are i.i.d distributed with variance $\sigma_\epsilon^2 = 0.5$. The exponential weight function is considered when applying WMLE algorithm and we use Equation (7) with the estimates, $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}))$, from the WMLE to calculate the forecasts.

In Experiment 1, we generate a time series by the AR(1) model. The rolling-window length, k, is selected to be 120 and forecasts are made for time period $t = 701\ldots760$. The generated data are considered as monthly data. For each simulated data set, we carry out WMLE and use the resulting estimates to calculate the one-step-ahead forecasts. Since the model is accurately specified we expect that the estimated WMLE parameter, ρ_1, is close to 1, which means the original AR(1) model can already provide accurate forecasts.
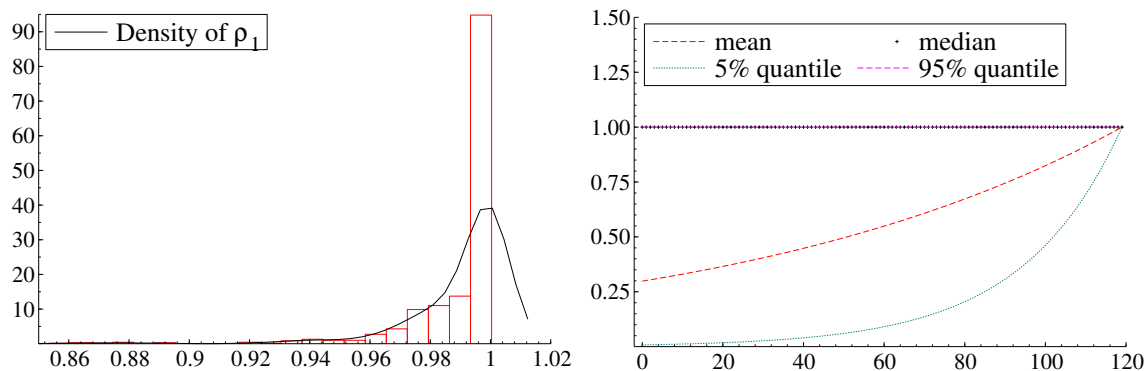

Figure 5: Simulation results for Experiment 1. The data are generated by an AR(1) model and the WMLE parameter, ρ_1, is estimated based on Equation (7). Exponential weight function is applied. The left panel presents the simulation density of WMLE parameter, ρ_1, over 1,000 simulations. The rights panel presents the average sample weights and its 90% confidence bound.

The left panel of Figure 5 shows the simulation density of WMLE parameter, ρ_1, in Experiment 1. The simulated parameters peak at 1. To be precise, the simulated parameter mean is 0.9953 and the median is 1. In the setup, we restrict our weight parameter to be between 0 and 1, so in this case we shall put more attention on the median of the simulation results rather than the mean. The feature of the simulated distribution indicates that the WMLE method can hardly improve forecasting accuracy. Such finding is consistent with Proposition 2 in Section 4 which shows the weights converge in probability to unity when the model is well specified. The right panel of Figure 5 shows the respective average sample weight and its 95% confidence bounds.

B.2 Experiment 2: Time-varying AR(1) model

In the second experiment, the time series are generated by an AR(1) model with time-varying coefficient, β, and the WMLE parameter, ρ_1, is estimated based on Equation (7). The data generation process (DGP) is specified by:

$$\text{DGP: } y_t = \alpha + \beta_t y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{9}$$

$$\beta_t = 0.5 + 0.5 \sin(2\pi t/B), \tag{10}$$

where $\alpha = 0.13$ and ϵ_ts are i.i.d distributed with variance $\sigma_\epsilon^2 = 0.5$. The coefficient, β is varying between 0 and 1 with respect to business cycle length $B = 72$. The exponential weight function is considered when applying WMLE algorithm and we use Equation (7) with the estimates, $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}))$, from the WMLE to calculate the forecasts.
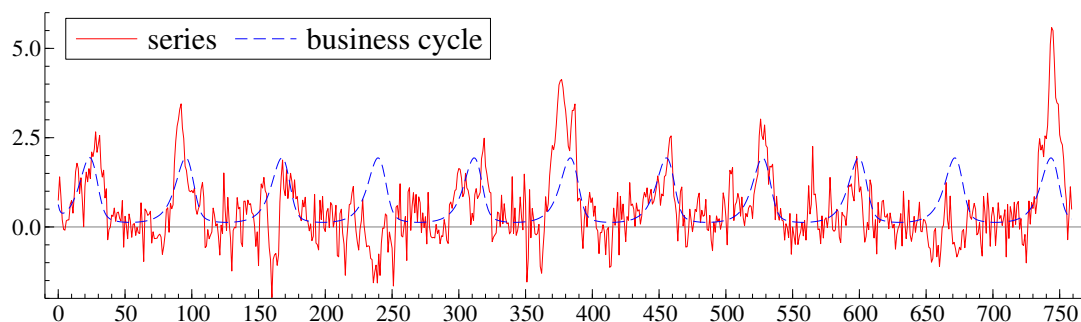

Figure 6: Illustration of DGP for Experiment 2.

In Experiment 2, we generate a time series of size $T = 760$. The generated data are considered as monthly data. We consider a DGP which includes a six-year business cycle, which is the average U.S. business cycle length. The rolling-window size is set to be 60 which is a bit shorter than the business cycle. An illustration of such DGP is shown in Figure 6. Forecasts are made for time period $t = 701 \ldots 760$. In this experiment, the time series imply instability of the coefficients in the model and the original AR(1) model is misspecified, thus the WMLE parameter ρ_1 is expected to be smaller than 1. This means recent observations are more relevant to the forecasts in the future.

The left panel in Figure 7 presents the simulated density of WMLE parameter, ρ_1. The
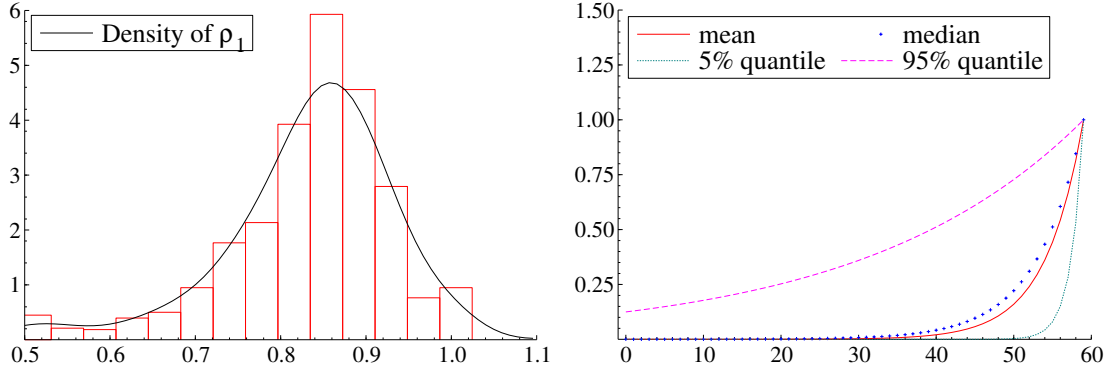
Figure 7: Simulation results of the WMLE method over 1,000 simulations. The data are generated by an AR(1) model with the time-varying coefficient, β, and the WMLE parameter, ρ_1, is estimated based on Equation (7). The exponential weight function is applied. The left panel presents the simulated density of the WMLE parameter, ρ_1. The right panel shows the average sample weights and its 90% confidence bound.

right panel shows the average sample weights and its 90% confidence bound. The right panel of Figure 7 shows that both the mean and the median of the simulated WMLE parameter, ρ_1, is smaller than 1. The general picture of Figure 7 is that when the time series contains certain time-varying components in it and the considered forecasting model is misspecified, the MLE method can improve forecasting accuracy by putting more weights to the recent observation. Such finding is also consistent with Proposition 3 in Section 4.

B.3   Experiment 3: AR(1) Model with a Structural Break

In Experiment 3, the data are generated by an AR(1) model with a structural break in the coefficient, β, and the WMLE parameter, ρ_1, is estimated based on Equation (7). The data generation process (DGP) is specified by:

$$y_t = \alpha + \beta_t y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{11}$$

$$\beta_t = 0.2 + 0.7 I_t, \tag{12}$$

where $\alpha = 0.13$ and ϵ_ts are i.i.d normally distributed with variance $\sigma_\epsilon^2 = 0.5$. The indicator I_t is set to $I_t = 0$ for $t < 420$ and $I_t = 1$ for $t \geq 420$. The exponential weight function is considered when applying WMLE algorithm and we use Equation (7) with the estimates,

$\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}))$, from the WMLE to calculate the forecasts.

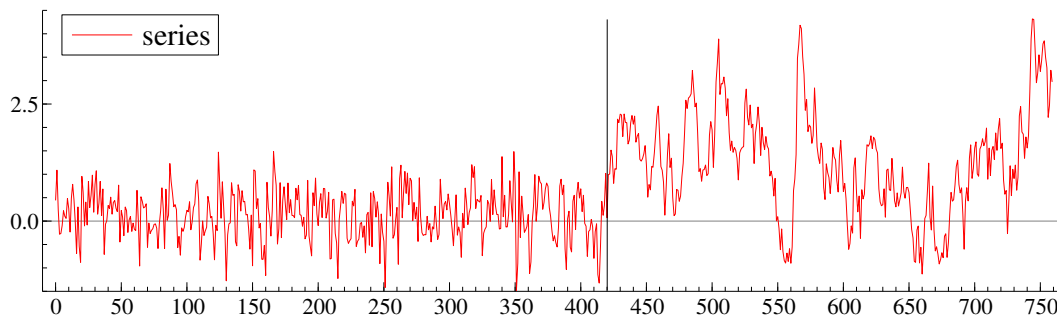

Figure 8: Illustration of DGP for Experiment 3.

In Experiment 3, we generate a time series with size $T = 760$. The rolling-window length, τ, is selected to be 120. The break-point is set to $t = 420$. We evaluate the forecasting performance in three different forecasting periods . The first forecasting period is the period before the break-point (BB) starting from $t = 360$. The second one is the period right after the break-point (JAB) starting from $t = 444$ and the last one is the period long after (LAB) the break-point starting from $t = 492$. For all cases, we evaluate the forecasting performances of 60 observations. The generated data are considered as monthly data. This means, the length of forecasting period is 5 years. Figure 8 presents a realization of the data from Experiment 3. When forecasts are calculated before the break point, the time series is generated by an ordinary AR(1) model and the weighted ML method converges to the classical ML method, while the simulated WMLE parameter is expected to be 1. When forecasts are calculated just after the break point, the data before the break point are less relevant for calculating future forecasts, thus the WMLE parameter, ρ_1, will be significantly smaller than 1. As the forecasting point is getting far away from the break point, the WMLE parameter ρ_1 moves generally back to 1 again.

Figure 9 presents the density of WMLE parameter, ρ_1, in Experiment 3. The WMLE method can hardly improve forecasting accuracy for the forecasting period before the breaking point because the forecasting model is well specified just as the finding of Experiment 1. For the forecasting period right after the structural break the simulated mean of the WMLE parameter is 0.93485, which indicates more weight should be put on recent observation in order to provide better forecasts. Finally, for the period long after structural break,
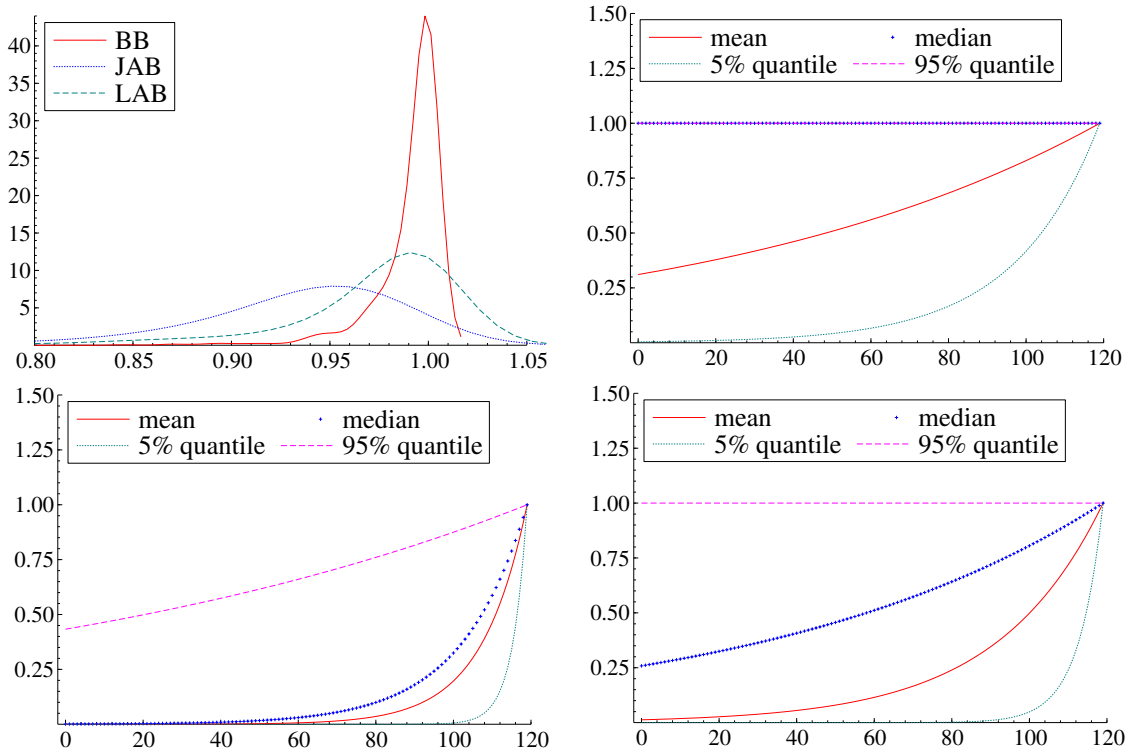
Figure 9: Simulation densities of WMLE parameter, ρ_1, over 1,000 simulations. The data are generated by an AR(1) model with a structural break in the coefficient, β, and the WMLE parameter, ρ_1, is estimated based on Equation (7). Exponential weight function is applied. The upper-left panel presents the simulated densities for three forecasting periods: the period before the breaking point (BB), the period just after the breaking point (JAB) and the period long after the breaking point (LAB). The upper-right panel presents the average sample weights and its 90% bound for forecasting period before the breaking point. The bottom-left panel presents the average sample weights and its 90% bound for forecasting period right after the breaking point. The bottom-right panel presents the average sample weights and its 90% bound for forecasting period long after the breaking point.

fewer observations before the breaking point are included in the estimation window and the WMLE parameter tends to peak at 1 again.

B.4   Experiment 4: regime-switching AR(1) model

In Experiment 4, the data are generated by an AR(1) model with a two-state regime-switching coefficient, β, and the WMLE parameters are estimated based on Equation (7). The data generation process (DGP) is specified by:

$$y_t = \alpha + \beta_t y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \tag{13}$$

$$\beta_t = 0.9 - 0.7 I_t, \tag{14}$$

where $\alpha = 0.13$ and ϵ_ts are i.i.d normally distributed with variance $\sigma_\epsilon^2 = 0.5$. I_t is a predetermined recession indicator. A value of 1 indicates a recessionary period; while a value of 0 indicates an expansionary period. Model is estimated using the binary weight function with decay.
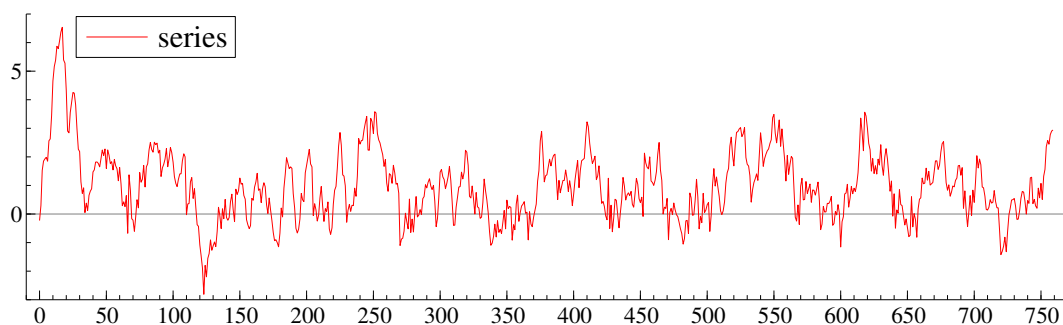
Figure 10: Illustration of DGP for Experiment 4.

In Experiment 4, we generate a time series with size $T = 720$. The rolling-window length, τ, is selected to be 120. Forecasts are made for time period $t = 709 \ldots 720$. The generated data are considered as monthly data. The time series we generated follows a 6-year business cycle, where a 61-month expansion is followed by an 11-month recession. Notice that the forecasting period we considered is the last recession period in the simulated series. An illustration of the DGP is shown in Figure 10.

The time series is estimated for (7) using WMLE algorithm with binary weight function and we use the estimates, $\hat{\boldsymbol{\theta}}(\mathbf{w}_k(\hat{\boldsymbol{\rho}}))$, from the WMLE to calculate the forecasts. Simulated densities of the weight parameters using binary weight function with decay are presented in Figure 11. The simulated binary parameter ρ_2 tends to have value larger than 1, which indicates that when calculating forecasts during recessions, more attention should be paid to the past recessions as discussed in the previous section. Moreover, parameter ρ_1 has peak near one. To be precise, the simulated parameter mean is 0.9781 and the median is 1. This result indicates that by using binary weight function we do not observe strong evidence of decaying weights in the past observations.
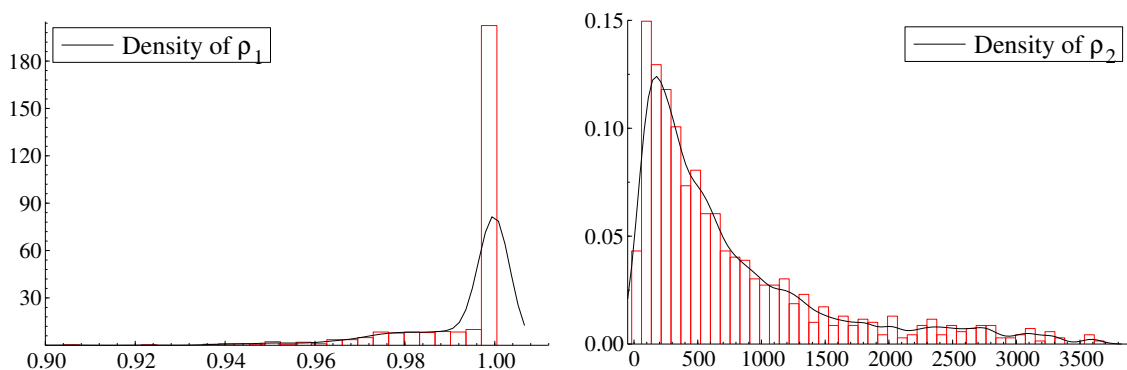
Figure 11: Simulation results of the WMLE method using binary weight function, over 1,000 simulations. The data are generated by an AR(1) model with a two-state regime switching coefficient, β. The WMLE parameters are estimated based on Equation (7). The upper-left panel presents the simulated density of parameter ρ_1. The upper-right panel presents the simulated density of parameter ρ_2. The bottom panel presents the median sample weights and its 90% bound for the selected forecasting period.

C Further Empirical Findings

We provide further (rolling-window) forecasting results for our empirical illustrations on the basis of different weight functions within the WMLE procedure. Tables 5 and 6 present the forecasting results for the monthly time series of U.S. IPI growth rate and obtained from the AR(1) model and the WMLE procedure with binary and decaying weights. Table 6 presents the results for an expansion period. Tables 7 and 8 present the forecasting results for the monthly time series of U.S. Unemployment Rate and U.S. Total Non-farm Payrolls, respectively.

Table 5: AR Model Forecasting Results: Binary Weight with Decaying

	RMSE						MAE					
	$n=1$			$n=4$			$n=1$			$n=4$		
Window	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2
AR(1)												
25-year	1.0908	1.000	35.5	0.8297*	1.000	27.7	0.8310	1.000	31.5	0.7919*	1.000	27.1
	(0.3446)			(0.0679)			(0.1564)			(0.0500)		
50-year	0.9108**	1.000	15.7	0.7423**	1.000	12.3	0.8007***	1.000	13.8	0.5975***	1.000	11.1
	(0.0268)			(0.0244)			(0.0011)			(0.0000)		
AR(AIC)												
25-year	1.4691	0.999	119.0	0.9189	1.000	11.1	1.4131	1.000	250.0	0.9450	1.000	8.8
	(0.1481)			(0.2707)			(0.1754)			(0.3918)		
50-year	0.9293	0.998	15.9	0.7782*	1.000	9.7	0.8427**	1.000	25.7	0.6955***	1.000	11.5
	(0.1292)			(0.0523)			(0.0153)			(0.0021)		

Table 5: Rolling-window forecasting results for monthly U.S. IPI growth rate. Entries Ratio show the root mean-squared-forecast-error (RMSE) or mean-absolute-error (MAE) using the WMLE method for the AR model relative to RMSE or MAE using normal ML method. Columns labelled ρ_1 and ρ_2 show the optimal weight parameter for the combined weight function. The forecasts were computed over the sample period 2008Jan–2009June.

Table 6: AR Model Forecasting Results: Binary Weight with Decaying, Expansion Period

| | RMSE | | | | | | MAE | | | | | |
| | n=1 | | | n=4 | | | n=1 | | | n=4 | | |
window length	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2
AR(1)												
25-year	0.9708	0.977	53.6	0.9413*	1.000	2.2	0.9652	0.967	32.5	1.0000	1.000	1.0
	(0.2748)			(0.0986)			(0.2397)			(0.8536)		
50-year	0.9605	0.977	129.3	1.0000	1.000	1.0	0.9530	0.967	32.2	1.0000	1.000	1.0
	(0.2456)			()			(0.1943)			()		
AR(AIC)												
25-year	0.9126*	0.994	79.2	0.8709	0.998	8.8	0.9430	0.990	71.9	0.9154	0.993	12.6
	(0.0983)			(0.1609)			(0.1452)			(0.1632)		
50-year	0.9727	0.990	47.3	1.0155	0.999	1.0	0.9487	0.989	71.8	1.0686	0.994	1.0
	(0.2604)			(0.8852)			(0.1332)			(0.9716)		

Table 6: Rolling-window forecasting results for monthly U.S. IPI growth rate. Entries Ratio show the root mean-squared-forecast-error (RMSE) or mean-absolute-error(MAE) using WMLE method of the AR model relative to RMSE or MAE using normal ML method. Columns labelled ρ_1 and ρ_2 show the optimal weight parameter for the combined weight function. The forecasts were computed over the sample period 2001Jan–2007Dec.

Table 7: AR Model Forecasting Results: U.S. Unemployment Rate

| | RMSE | | | | | | MAE | | | | | |
| | $n=1$ | | | $n=4$ | | | $n=1$ | | | $n=4$ | | |
Window	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2
AR(1)												
25-year	0.6626***	1.000	72.3	0.7449*	1.000	79.8	0.7095**	0.999	202.9	0.7596*	1.000	300.7
	(0.0009)			(0.0541)			(0.0153)			(0.0642)		
50-year	0.6814***	0.996	40.8	0.7746*	1.000	15.3	0.7162**	0.994	482.2	0.7421*	1.000	38.8
	(0.0023)			(0.0788)			(0.0266)			(0.0586)		
AR(AIC)												
25-year	0.8694	1.000	7.2	1.0291	0.946	1.0	0.9159	1.000	6.8	0.8398	0.952	1.0
	(0.1010)			(0.5258)			(0.2351)			(0.3481)		
50-year	0.8710	1.000	13.0	0.8480	0.998	14.3	0.8235*	1.000	39.7	0.7780	1.000	19.5
	(0.1136)			(0.2922)			(0.0915)			(0.1567)		

Table 7: Rolling-window forecasting results for monthly U.S. Unemployment. Entries Ratio show the root mean-squared-forecast-error(RMSE) or MAE using WMLE method of the AR model relative to RMSE or MAE using normal ML method. Columns labelled ρ_1 and ρ_2 show the optimal weight parameter for the combined weight function. The forecasts were computed over the sample period 2008Jan–2009June.

Table 8: AR Model Forecasting Results: U.S. Total Non-farm Payrolls

| | RMSE | | | | | | MAE | | | | | |
| | n=1 | | | n=4 | | | n=1 | | | n=4 | | |
Window	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2	Ratio	ρ_1	ρ_2
AR(1)												
25-year	0.5860***	1.000	66.5	0.7019	1.000	56.0	0.5027***	1.000	83.3	0.5943*	1.000	58.1
	(0.0010)			(0.1672)			(0.0002)			(0.0883)		
50-year	0.5212***	1.000	18.4	0.5455**	1.000	18.0	0.4401***	1.000	33.4	0.4582***	1.000	25.3
	(0.0000)			(0.0155)			(0.0000)			(0.0003)		
AR(AIC)												
25-year	0.7688**	0.990	7.4	0.8236	1.000	32.1	0.6393***	0.992	7.5	0.7253	0.997	16.4
	(0.0348)			(0.2893)			(0.0059)			(0.1954)		
50-year	0.6700***	1.000	8.7	0.6369	0.990	39.0	0.6087***	1.000	6.4	0.5659	0.989	23.2
	(0.0060)			(0.1715)			(0.0009)			(0.1538)		

Table 8: Rolling-window forecasting results for monthly U.S. Total Non-farm Payrolls. Entries Ratio show the root mean-squared-forecast-error (RMSE) or MAE using WMLE method of the AR model relative to RMSE or MAE using normal ML method. Columns labelled ρ_1 and ρ_2 show the optimal weight parameter for the combined weight function. The forecasts were computed over the sample period 2008Jan-2009June.