Macroeconomic Forecasting and Evaluation with Supervised and Neural Network Reinforced Factor Models

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2020

vorgelegt
von

Simon Lineu Umbach, M.Sc.,

aus

Wittlich

Referent: Prof. Dr. Jörg Breitung
Korreferent: Prof. Dr. Robinson Kruse-Becher

Tag der Promotion: 18. Februar 2021

Für meine Eltern

Vorwort

Mein besonderer Dank gilt meinen beiden Betreuern Professor Dr. Jörg Breitung und Professor Dr. Robinson Kruse-Becher. Ihre Unterstützung mit zahlreichen Ideen und Anmerkungen zu meinen Forschungsprojekten hat entschieden zum Gelingen dieser Arbeit beitragen. Ich weiß sehr zu schätzen, dass ich mich an beide jederzeit wenden konnte. Herr Breitung hatte immer an ein offenes Ohr für meine Forschungsideen und hat mir überhaupt erst ermöglicht diesen nachzugehen. Herrn Kruse-Becher möchte ich noch einmal für ein intensives letztes Jahr danken, das in besonderer Erinnerung bleiben wird.

Mein Dank gilt auch meinen Arbeitskollegen, unter denen ich gute Freunde gefunden habe. Ein hohes Maß an Diskussions- und Hilfsbereitschaft untereinander hat für ein angenehmes und konstruktives Arbeitsklima gesorgt. Die Zeit mit Ihnen bleibt unvergesslich - sei es im Büro, beim "Lehrstuhlfußball" oder am Kiosk nach der Arbeit.

Nicht zuletzt möchte ich mich ganz besonders bei meinen Eltern und Geschwistern bedanken. Sie haben mir den Rückhalt gegeben, der mich (nicht nur) durch die letzten Jahre getragen hat. Ohne sie hätte ich die Arbeit nicht schreiben können.

Köln, am 21. Oktober 2020 Simon Lineu Umbach

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Forecasts of key macroeconomic variables are essential components for the decision making of central banks, fiscal authorities, households, and private sector businesses. For example, in assessing financial sustainability it is crucial to have reliable forecasts of the future path of economic activity. Due to their relevant support for policy making both in the private and public sector, statistical forecasting models and methods have found enormous interest in the academic literature. The availability of different forecasts of the same economic quantity has raised the need for formal statistical procedures to compare the competing forecasts' predictive accuracy. Forecast evaluation tests can provide evidence whether a superior predictive accuracy of a forecast series is merely good luck or truly indicative of a difference in population.

In many macroeconomic forecasting applications a large number of time series can be exploited to predict the variable of interest. As macroeconomic time series are typically sampled at quarterly or monthly frequency and only a few decades of observations are available, one has to cope with the so-called "course of dimensionality" (large number of parameters relative to the sample size), when estimating a statistical model with many predictor variables. This raises the issue of how to incorporate the information of a large amount of candidate predictors within a single forecasting model. Although variable selection procedures (statistical or motivated by economic theory) can be used to choose a small subset of predictors from a large set of potentially useful variables, the performance of these methods ultimately rests on the few variables that are chosen. An alternative to variable selection is to pool the information of all the candidate predictors within a few factors. These factors capture the common (linear) components of the individual series and discard their idiosyncratic variation.

In macroeconomics, factor models have proven to be successful both for forecasting purposes and for illustrating the state of the economy. They have been subject to extensive analysis in the academic literature and have found application in some well-known economic indicators. To name a few, the Conference

Board publishes Leading, Coincident, and Lagging Economic Indexes for the US that are constructed as factor estimates from selected macroeconomic time series. In Germany, the German Institute of Economic Research (DIW Berlin) provides a monthly economic indicator that is built upon a factor modeling approach. In the context of macroeconomic forecasting, factor models have become a conventional approach to overcome the "course of dimensionality". A popular approach is to form a forecasting model by using the latent factor estimates underlying the large set of candidate predictors as model inputs instead of relying on variable selection procedures.

This thesis considers macroeconomic forecasting in a data rich environment, and forecast evaluation tests. The statistical factor model provides the basic framework and is extended with the purpose of improving its forecasting capability. This thesis contributes to the literature by proposing and refining different adaptations of the factor model that aim at overcoming some of its hampering constraints.

First, it is analyzed how factor estimates can be tailored to forecasting applications by incorporating the forecasting target directly in the factor estimation process. For this purpose the Principal Covariate Regression (PCovR) technique of de Jong and Kiers (1992) is refined and it is analyzed under which circumstances gains in forecast accuracy can be achieved by integrating this form of supervision in the factor estimation.

Second, the statistical factor model is aligned with the variational autoencoder (VAE) framework of Kingma and Welling (2013) in the context of macroeconomic forecasting. It is studied whether factor models enriched by neural networks can provide superior forecasting power for macroeconomic time series. In contrast to the original factor model, the resulting neural network reinforced factor model is not subject to the linearity restriction anymore, and can capture nonlinear common dynamics in the set of candidate predictors as well. Furthermore it is proposed to incorporate the aforesaid supervision aspect within these models.

The extended factor models are applied to forecast key monthly macroeconomic variables such as industrial production, inflation, and employment. The findings suggest that their forecasting capability can be significantly improved by the analyzed and refined extensions.

As already mentioned in the beginning, only reliable forecasts are helpful for decision making. A comparison of two competing forecasts of the same economic quantity requires a formal statistical procedure to distinguish between a better predictive accuracy by coincidence and a fundamental advantage of one over the

other. To this end, one of the most popular statistics is the Diebold and Mariano (1995) test. This thesis contributes to the literature by showing how the power of the Diebold and Mariano (DM) test can be improved when the forecasts are rational, i.e., unbiased and efficient. In applied work, it is essential to uncover superior predictive ability and therefore, powerful and yet correctly-sized tests are needed.

Overall, this thesis comprises three self-contained essays on macroeconomic forecasting with factor models, and on forecast evaluation tests. The essays have been single author projects[1] and are summarized below.

Chapter 2 corresponds to the paper "Forecasting with Supervised Factor Models" (Umbach, 2020) and was published in *Empirical Economics*. This study analyzes in which forecasting settings it is promising to use factor estimates that are supervised with respect to the forecasting target via the PCovR technique. This contrasts the frequently used practice where the latent factors are estimated as the Principal Components of a large amount of candidate predictors, and then, in a separate step, are related to the forecasting target in a regression framework. The latter approach performs predictor space compression and estimation of the forecasting model in two separate steps. Hence, no information regarding the forecasting objective is taken into account during factor estimation. While the estimated factors capture the major common variation in the predictors, it is not guaranteed this information is most relevant for forecasting. In a simulation study and in a macroeconomic forecasting exercise, it is shown that supervised factors within the PCovR framework yield promising forecasting results when they are constructed from intermediate large predictor spaces. If the number of predictor variables is too large relative to the number of observations available for estimation, the supervised factors are vulnerable to overfitting problems.

Principal Covariate Regression requires the specification of a value for a supervision parameter that governs the trade-off between predictor space compression and target orientation of the estimated factors. In the literature, there has not been much guidance regarding this choice. Motivated by the application of PCovR in forecasting, the problem of choosing a value for the supervision parameter is solved by the development of an information criterion. The information criterion is shown to be an appropriate means to find a good balance between compression and target orientation of the estimated factor.

Furthermore, another favorable aspect of the supervised factor model is shown.

[1] I consider it important to emphasize the helpful support of my supervisors Prof. Dr. Jörg Breitung and Prof. Dr. Robinson Kruse-Becher.

It requires only a single factor to achieve its full forecasting power given the supervision parameter is adjusted appropriately. This is especially interesting if one seeks to express the state of the economy by a single index as it is the objective of the economic indicators mentioned in the beginning.

In Chapter 3, it is analyzed whether factor models enriched by elements from the machine learning literature, more precisely by neural networks, can achieve superior forecasting accuracy. First, it is shown how the statistical factor model and variational autoencoders from the machine learning literature are interrelated. VAEs provide a powerful framework for nonlinear dimensionality reduction. They estimate the distribution of the common latent factors underlying the data by combining the statistical factor model with a purely data-driven neural network approach. It is demonstrated that the resulting *deep factor model* can be interpreted as a flexible nonlinear extension of the standard factor model. The nonlinearity is achieved by the integration of a neural network that models the first two moments of the conditional distribution of the latent factors.

In their original formulation VAEs only provide a means for dimensionality reduction. To adapt the VAE framework to (macroeconomic) prediction tasks, an extension is proposed that relates to the supervision aspect of Chapter 2.

The flexible parametrization of the deep factor model comes at the cost of not having a closed form for the likelihood. Instead, the model parameters have to be estimated by variational inference. As a further consequence of their flexible parametrization variational autoencoders do not provide a fully identified latent model such that their major purpose in macroeconometrics should be seen in forecasting instead of revealing some interpretable latent economic dynamics. Indeed, the results of the empirical forecasting exercise suggest significant improvements of the deep factor model in the forecasting accuracy of four major US macroeconomic time series.

In Chapter 4, the Diebold and Mariano test under forecast rationality is examined. The DM test offers a framework for testing the null hypothesis of equal predictive accuracy of two competing forecast series. The test statistic is based on the forecast error loss differential, i.e., the difference between the two forecast error loss functions. Different loss measures can be exploited, but the most prominent measure is the mean squared error (MSE) loss differential.

This chapter contributes to the literature by deriving a simplified variant of the DM test statistic that is applicable under rational forecasts. The MSE loss differential is decomposed and adjusted by removing some components that are zero in expectation both under the null and under the alternative hypothesis.

Hence, these components only add noise to the test statistic. The resulting rationality adjusted DM test remains as simple to apply as the original approach and can improve the power of the testing procedure considerably.

When the forecast series stem from estimated statistical models, such as, for instance, from factor models presented in Chapters 2 and 3, the impact of parameter estimation uncertainty on the distribution of the (adjusted) test statistic generally has to be taken into account. Otherwise size distortions can occur, especially if the number of forecasts in relation to the number of observations used for parameter estimation of the forecasting models is relatively large. Ignoring these effects can result in misleading conclusions drawn from the test. To prevent size distortions, a simple-to-use adjustment of the estimation of the long-run variance of the test statistic is proposed. This adjustment accounts for the parameter estimation uncertainty. It holds under a fixed estimation scheme and shows good results for the rolling and recursive scheme as well.

Furthermore, the applicability of the rationality adjusted test statistic in a nested forecast comparison is discussed. Despite its nonstandard limiting distribution in the case of nested forecasts, it is argued that the adjusted DM test is still accurate for practical purposes if standard normal critical values are used. This is advantageous compared to the standard DM test, which is seriously undersized in nested forecast comparisons.

The small sample properties of the proposed test statistic are examined in extensive simulation studies that cover the cases of model-free forecasts, forecasts from estimated statistical models, and forecasts from nested models.

As indicated above, the adjusted DM test rests upon the assumption of rational forecasts. Under the MSE loss function a rational forecast is characterized by unbiasedness and efficiency. Although imposing forecast rationality seems to be an appealing assumption, there is some evidence that, e.g., analysts are not always (MSE-)rational in their forecasts. To take notice of the ongoing debate in the literature on forecast rationality, the effect of rationality violations on the adjusted DM test is discussed.

Chapter 2

Forecasting with Supervised Factor Models

2.1 Abstract

A conventional approach to forecast in a data-rich environment is to estimate factor-augmented predictive regressions with factors constructed by Principal Component Analysis. This study analyzes under which circumstances gains in forecast accuracy can be achieved by incorporating some form of supervision in the factor estimation process. Specifically, Principal Covariate Regression (PCovR) is considered. For the problem of choosing a value for the supervision parameter in PCovR an information criterion is proposed. The information criterion is shown to be an appropriate means to find a good balance between predictor space compression and target orientation of the estimated factors. A simulation study and an empirical application on a macroeconomic dataset show that supervised factors can improve the forecasting accuracy of factor models.

2.2 Introduction

In many forecasting applications in macroeconomics and finance a vast set of potential predictor variables can be exploited to forecast a variable or diffusion index of interest. A popular approach to cope with large predictor spaces is to use factor models that aim at finding a few latent factors underlying the high-dimensional predictor space. In a forecasting context, a frequently used practice is to first estimate the factors by means of Principal Component Analysis (PCA) and then relate them to a forecasting target in a regression framework.[1] This approach conducts predictor space compression and estimation of the forecast equation in two separate steps such that no information regarding the forecasting target is exploited in the factor estimation.

[1] See Breitung and Choi (2013) and Stock and Watson (2006) for reviews.

This study analyzes under which circumstances gains in forecast accuracy can be achieved by incorporating some form of supervision in the factor estimation process. Specifically, Principal Covariate Regression (PCovR) is considered which was introduced by de Jong and Kiers (1992). PCovR requires to choose a value for the supervision parameter that governs the trade-off between predictor space compression and target orientation of the estimated factors. The problem of determining an appropriate value for the supervision parameter is solved by deriving an information criterion which is shown to be an appropriate means to find a good balance between compression and target orientation. The information criterion yields better results than an alternative approach based on a stochastic extension of PCovR proposed by Vervloet et al. (2013). Additionally, exploiting the information criterion allows to obtain competitive forecasts with only a single supervised factor.

The simulation study shows that supervised factors are able to incorporate information of the regressor space that is relevant for forecasting but neglected or only captured in some minor principal components by the unsupervised factor model. The empirical application on a macroeconomic dataset corroborates the finding that supervised factors can provide more accurate forecast than their unsupervised counterparts. A complication of the supervised factor model, however, is its sensitivity to overfitting when the number of regressors is very large compared to the observations available. It is concluded that the supervised factor model has its strength mainly for medium-sized regressor spaces.

The remainder of this paper is organized as follows. Section 2.3 introduces PCovR and shows how it is related to Principal Components Regression (PCR) and Reduced Rank Regression (RRR). Furthermore, the information criterion for the supervision parameter is derived. The simulation study is given in section 2.4. Section 2.5 presents an empirical forecasting application for the key macroeconomic variables Industrial Production, and the Consumer Price Index. Finally, Section 2.6 offers some conclusions.

2.3 A Supervised Factor Model

A conventional approach to forecast an economic variable in a data-rich environment is to estimate common factors of the predictor space by means of PCA and include them in factor-augmented predictive regressions. This approach rests on the assumption that the n predictor variables x_t obey a factor structure of the

form

$$x_t = \Lambda f_t + e_t, \tag{2.1}$$

where f_t is an $r \times 1$ vector of common factors. The $n \times r$ matrix Λ contains the factor loadings and the error term e_t denotes the idiosyncratic components. The principal component estimator for Λ results from applying the least-squares principle on equation (2.1) subject to the identification restriction $\Lambda'\Lambda = I_r$. It is straightforward to show that the estimate $\widehat{\Lambda}$ consists of the first r eigenvectors belonging to the largest eigenvalues of the covariance matrix of x_t. The common factors are then estimated by $\widehat{f_t} = \widehat{\Lambda}' x_t$ (see, for instance, Breitung and Choi (2013)). Given such estimates, the factors can be exploited in the predictive regression:

$$y_{t+h} = \gamma f_t + u_{t+h}, \tag{2.2}$$

where h designates the forecasting horizon. The procedure described above treats predictor space compression and estimation of the forecasting equation separately such that the construction of the factors f_t does not take into account the relationship between the individual predictors x_{it}, $i = 1, \ldots, n$, and the target variable y_{t+h}. Instead, the factor estimates focus solely on compiling the major variation in x_t. However, this might not necessarily be the information that is most relevant for forecasting. For instance, it can be the case that relevant information hidden in some minor principal components, that are disregarded in the predictive regression (2.2), is lost.

Principal Covariate Regression allows to incorporate some form of supervision in the factor estimation process. It takes the forecasting target explicitly into account when estimating the factor subspace. Searching for a low-dimensional subspace of dimension $r \ll n$ spanned by $F = XA$, PCovR comprises the two stages of compressing the predictor space (2.1) and fitting the forecast equation (2.2) by one single criterion function:

$$\begin{aligned} Q_\theta(F, \Lambda, \gamma) = \theta \frac{(y - F\gamma')'(y - F\gamma')}{||y||^2} \\ + (1 - \theta)\, tr\left\{ \frac{(X - F\Lambda')'(X - F\Lambda')}{||X||^2} \right\}, \end{aligned} \tag{2.3}$$

where $F = [f_1, \ldots, f_{T-h}]'$, $X = [x_1, \ldots, x_{T-h}]'$, and $y = [y_{1+h}, \ldots, y_T]'$. For scaling purposes each residual sum is divided by the squared Frobenius norm of its corresponding regressand. Furthermore, all variables should be standardized to prevent scale effects. The supervision parameter $\theta \in [0, 1]$ specifies the orientation of the factor estimates, e.g. whether the focus is on summarizing the

common variation in X or on aligning the factors on the target variable y. In the original formulation of de Jong and Kiers (1992), PCovR is a purely data-based method that does not assign an explicit underlying statistical model. However, one can easily show that the least-squares criterion function (2.3) results from the gaussian likelihood when the two error components are independent and the variance ratio σ_u^2/σ_e^2 is assumend to be known. Ignoring the scaling factors, θ then equals $1/(1+\sigma_u^2/\sigma_e^2)$.

Minimization of the criterion function (2.3) subject to the identification restriction $T^{-1}F'F = I_r$ is equivalent to maximizing

$$tr\left\{\theta\frac{A'X'yy'XA}{||y||^2} + (1-\theta)\frac{A'X'XX'XA}{||X||^2}\right\} \tag{2.4}$$

subject to $T^{-1}A'X'XA = I_r$. By taking the first order condition of the Lagrangian, the constraint optimization describes a generalized eigenvalue problem. Accordingly, estimates for A can be obtained from solving the generalized eigenvalue problem

$$\left|\theta\frac{X'yy'X}{||y||^2} + (1-\theta)\frac{X'XX'X}{||X||^2} - \lambda X'X\right| = 0, \tag{2.5}$$

where the eigenvectors associated with the largest r eigenvalues are the estimator for A. Estimates for Λ and γ are then obtained from regression of X and y on $\widehat{F} = X\widehat{A}$.[2]

Principal Components Regression results as a special case of PCovR when setting $\theta = 0$ in the criterion function (2.3). Regarding the framework above, one obtains $\widehat{A} = \widehat{\Lambda}$. Whereas the unsupervised factors from PCR focus solely on the variation in X, PCovR finds an r-dimensional subspace of X spanned by F that accounts for a maximum amount of variation in both X and y. This immediate link between the supervised factors of PCovR and the target variable may be of advantage in forecasting with factor models.

2.3.1 Relationship to Reduced Rank Regression

As outlined above, PCR results from PCovR when choosing $\theta = 0$ in the criterion function (2.3). For the other extreme case of $\theta = 1$, PCovR corresponds with a regression of y on X. Then, the first factor in F is the common component of X that is maximally correlated with y. The remaining factors are the principal components of the residual part of X that is orthogonal to y.

[2] Heij et al. (2007) provide an alternative algorithm that is based on a singular value decomposition.

PCovR can be interpreted as a Reduced Rank Regression problem. To see this, let $Z_\theta = [(1-\theta)X, \theta y]$ and $B' = [\Lambda', \gamma']$ and consider the multivariate regression of Z_θ on X:

$$Z_\theta = X\Pi + V, \tag{2.6}$$

where $V = [\theta u, (1-\theta)E]$. The restriction that Π has reduced rank equal to r is expressed as $\Pi = AB'$, where A is $n \times r$ and B is $r \times (n+1)$. The loss function for equation (2.6) may be written as

$$\sigma(A, B) = tr\left\{(Z_\theta - XAB')'(Z_\theta - XAB')\right\}$$

subject to the rank and identification restriction $A'X'XA = I_r$. By the same argument as above, this describes a generalized eigenvalue problem. Accordingly, estimates for A can be obtained from solving the generalized eigenvalue problem

$$\left|X'Z_\theta Z_\theta' X - \lambda X'X\right| = 0, \tag{2.7}$$

where the eigenvectors associated with the largest r eigenvalues constitute $\widehat{A}$. Equation (2.7) is indeed equivalent to equation (2.5) as

$$\begin{aligned} X'Z_\theta Z_\theta' X &= X'\Big([(1-\theta)X, \theta y]\,[(1-\theta)X, \theta y]\Big)X \\ &= X'\Big(\theta yy' + (1-\theta)XX'\Big)X \\ &= \theta X'yy'X + (1-\theta)X'XX'X. \end{aligned}$$

Hence, PCovR can be interpreted as a special type of RRR where the predictor matrix X is also part of the multivariate response Z with a rank r restriction on the solution.

2.3.2 On the Choice of the Supervision Parameter

PCovR requires to specify the supervision parameter θ. In the literature, there is not much guidance regarding this choice. There are a few simulation studies from which one can take some heuristic advise for reasonable θ choices. These suggest to choose a rather small value for θ, i.e. below .5 or even below .1 (see Vervloet et al. (2013) for an overview). A more elaborate approach is based on a stochastic extension as suggested by Wilderjans et al. (2009) and analyzed in the context of PCovR by Vervloet et al. (2013). By assuming that the error terms in equations (2.1) and (2.2) follow normal distributions with zero means and variances $\sigma_e^2 I_n$ and σ_u^2, respectively, one can show that maximizing the likelihood $L\left(X, y | F, \Lambda, \gamma, \sigma_e^2, \sigma_u^2\right)$ is equivalent to minimizing the loss function (2.3) when

$\theta = ||y||^2 / \left(||y||^2 + ||X||^2 \, \sigma_u^2/\sigma_e^2\right)$. Exploiting this expression to obtain estimates for θ, however, requires knowledge on the amount of error in equations (2.1) and (2.2). Vervloet et al. (2015) estimate θ by setting $\hat{\sigma}_e^2$ equal to the percentage of unexplained variance from a PCA on X and $\hat{\sigma}_u^2$ equal to $1 - R^2$ of a regression of y on X. Another option is to adapt some cross-validation (CV) method to determine θ. However, depending on the specification of the CV procedure and the grid for $\theta \in [0, 1]$, this requires moderate computational efforts.

An Information Criterion for the Supervision Parameter

The maximum likelihood choice for θ has certain drawbacks. First, it depends crucially on having accurate estimates for σ_e^2 and σ_u^2. Second, it relies on the assumption that one is equally interested in recovering common factors in X and finding suitable factors to forecast y. While the latter is not a drawback in general, it might be of disadvantage when one is primarily interested in predicting y. Motivated by the application of PCovR in forecasting, an approximate information criterion for selecting θ can be derived. The criterion is based on the fit of the forecasting equation and a penalty for the 'pseudo-dimension' of the subspace spanned by the latent factors.

Without the stochastic extension, PCovR is a purely data-based method. The essential idea is to interpret the supervision parameter θ as a smoothing parameter within a nonparametric regression framework and exploit the improved Akaike information criterion for smoothing parameter selection of Hurvich et al. (1998). By simply rescaling the loss function (2.3) to

$$\begin{aligned} Q_\theta(F, \Lambda, \gamma) = & \frac{(y - F\gamma')'(y - F\gamma')}{||y||^2} \\ & + \underbrace{\frac{(1-\theta)}{\theta}}_{\tilde{\theta}} tr\left\{\frac{(X - F\Lambda')'(X - F\Lambda')}{||X||^2}\right\}, \end{aligned} \tag{2.8}$$

one can view $\tilde{\theta}$ as a parameter that governs the smoothness of the predictions $\hat{y}$ by forcing the factors F to recover the variance structure of the regressor variables simultaneously. The larger $\tilde{\theta}$ the more weight is put on the penalty for neglecting the fit of the regressor equation (2.1) which generally leads to smoother predictions for y.

To derive a smoother matrix $H_{\tilde{\theta}}$, define $f := F\gamma'$ and $\beta := A\gamma'$ and consider

the problem of estimating a single factor $f = X\beta$ based on

$$X = X\beta\lambda' + E,$$
$$y = X\beta + u.$$

By concentrating out λ, minimization of the PCovR loss function (2.8) can easily be shown to be equivalent to minimizing

$$\tilde{Q}_{\tilde{\theta}}(\beta) = (y - X\beta)'(y - X\beta) - \tilde{\theta}\frac{\beta' X' X X' X \beta}{\beta' X' X \beta}.$$

Note that the scaling constants $1/||X||^2$ and $1/||y||^2$ are omitted for readability. Let $\hat{\beta}_{\tilde{\theta}} = \arg\min \tilde{Q}_{\tilde{\theta}}(\beta)$. The first-order condition can be rewritten to

$$\hat{\beta}_{\tilde{\theta}} = \left[\left(1 + \tilde{\theta}\frac{\widehat{Q}_2}{\hat{\sigma}_f^2}\right) I_n - \frac{\tilde{\theta}}{\hat{\sigma}_f^2}\frac{1}{T}X'X\right]^{-1} \hat{\beta}_0, \tag{2.9}$$

where $\sigma_f^2 = T^{-1}\hat{\beta}_{\tilde{\theta}}' X' X \hat{\beta}_{\tilde{\theta}}$, $\widehat{Q}_2 = T^{-1}\hat{\beta}_{\tilde{\theta}}' X' X X' X \hat{\beta}_{\tilde{\theta}} / (\hat{\beta}_{\tilde{\theta}}' X' X \hat{\beta}_{\tilde{\theta}})$, and $\hat{\beta}_0 = (X'X)^{-1} X'y$. Equation (2.9) shows the shrinkage effect of the smoothing parameter $\tilde{\theta}$ on the least squares estimator $\hat{\beta}_0$. If $\tilde{\theta}$ equals zero, i. e. no penalty for neglecting the fit on the regressors, $\hat{\beta}_{\tilde{\theta}}$ results as the least squares estimator of an ordinary regression of y on X.

Using representation (2.9) the vector of fitted values results as

$$\hat{y} = X\hat{\beta}_{\tilde{\theta}} = H_{\tilde{\theta}} y,$$

where

$$H_{\tilde{\theta}} = X\left[\left(1 + \tilde{\theta}\frac{\widehat{Q}_2}{\hat{\sigma}_f^2}\right) I_n - \frac{\tilde{\theta}}{\hat{\sigma}_f^2}\frac{1}{T}X'X\right]^{-1} (X'X)^{-1} X'. \tag{2.10}$$

Note that $H_{\tilde{\theta}}|_{\tilde{\theta}=0} = H_0$ is the projection matrix onto the space spanned by the columns of X, whereas H_∞ yields a projection on the space spanned by the first principal component of X. Accordingly, the dimension of the subspace of fitted values is:

$$\begin{aligned} rk(H_0) = tr(H_0) = n \quad &\text{for} \quad \tilde{\theta} = 0, \\ rk(H_\infty) = tr(H_\infty) = 1 \quad &\text{for} \quad \tilde{\theta} \to \infty. \end{aligned} \tag{2.11}$$

While the first statement follows trivially from inspection of (2.10), a proof for result (2.11) is provided in Appendix A.1. For intermediate values of $\tilde{\theta}$ the interpretation of $\kappa_{\tilde{\theta}} = tr(H_{\tilde{\theta}})$ is less obvious. Following Breitung and Roling (2015) it is appealing to interpret $\kappa_{\tilde{\theta}}$ as the *pseudo-dimension* of the fitted values, where

$\kappa_{\tilde{\theta}}$ is a real number with $1 \leq \kappa_{\tilde{\theta}} \leq n$. Exploiting this *pseudo-dimension*, the framework of Hurvich et al. (1998) is adapted who provide a modified Akaike information criterion in the context of smoothing parameter selection. Interpreting $\tilde{\theta}$ as a parameter that governs the smoothness of the prediction $\hat{y}$, an approximate information criterion is considered:

$$\text{AIC}\left(\tilde{\theta}\right) = \log\left(\hat{\sigma}^2\right) + 2\frac{\kappa_{\tilde{\theta}} + 1}{T - \kappa_{\tilde{\theta}} - 2}, \tag{2.12}$$

where $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}(y_i - \hat{y}_i)^2$. A higher value for $\tilde{\theta}$ minimizes the first term in (2.12) by increasing the in-sample fit. The second term in (2.12) imposes a penalty depending on the 'pseudo-dimension' induced by $\tilde{\theta}$.[3]

The derivations above focused on choosing $\tilde{\theta}$ for a single factor f which will turn out to be totally sufficient. One advantage of choosing the supervision parameter $\theta = 1/(1 + \tilde{\theta})$ by means of the information criterion is that it makes the choice of the number of factors superfluous. Choosing a higher number of factors generally leads to a lower value of θ because the sample fit as well as the penalty term ceteris paribus increase due the enlarged factor number. The simulation study in Section 2.4 as well as the empirical application in Section 2.5 show that the supervised factor model can achieve competitive results with just one factor.

2.4 Simulation Study

The objective of the simulation study is two-fold. First, it is analyzed under which circumstances supervised factors have superior forecasting power. Second, the choice of the supervision parameter by the information criterion is evaluated and compared to cross-validation and the likelihood approach.

[3] Some computational hint: In practice, it can happen that for some small θ-values $H_{\tilde{\theta}}$ is hardly invertible causing the penalty $\kappa_{\tilde{\theta}} = tr(H_{\tilde{\theta}})$ to be not strictly monotonically increasing in $\tilde{\theta}$. If this happens to be the case, a practical approach to deal with this issue is to exploit that $\kappa_{\tilde{\theta}} = tr\left\{D^{-1}\right\}$, where D contains the eigenvalues of the matrix term in $H_{\tilde{\theta}}$ that is inverted (see Appendix A.1). Setting the smallest eigenvalue of D to one, helps to circumvent the problem. Following this route, one may interpret $\kappa_{\tilde{\theta}} - 1$ as the 'additional-dimension' of the fitted values induced by supervision of the factor estimate.

2.4.1 Data Generating Process

Corresponding to framework (2.1) - (2.2) the simulation study is based on the data generating process (DGP):

$$f_t = \phi f_{t-1} + \eta_t, \tag{2.13}$$

$$x_t = \Lambda f_t + e_t, \tag{2.14}$$

$$y_{t+h} = \gamma f_t + z_t \beta + u_{t+h}. \tag{2.15}$$

The r factors f_t follow an AR(1)-process with $\phi = 0.9$ and $\eta \sim N\left(0, 1-\phi^2\right)$. The orthogonal loadings matrix Λ is parametrized as $\Lambda = VD$ with V being equal to the first r normalized eigenvectors of $A'A$, where A is a random draw from a multivariate standard normal distribution. The entries in the diagonal matrix D are arranged in descending order and calculated by $\sqrt{d_i} = e^{-s\frac{i}{2r}}$ for $i = 1, \ldots, r$, where s denotes the decay constant. For scaling purposes, the diagonal of D is then normalized to length one. This results in the eigen-decomposition of $Var\left(X\right) = VDD'V' + \sigma_e^2 I$ with the fraction $d_i / \sum_{i \in r} d_i$ specifying the percentage of common variation in X that can be explained by its i^{th} principal component. Hence, by tuning the decay parameter s, one can govern the importance of common factors in the predictor space.

The error terms in $e_t = [e_{1t}, \ldots, e_{nt}]'$ are noise processes with mean zero and variances $\sigma_{e_i}^2$. Let $\rho_{x_i f}^2 = Cov\left(x_i, f\right) Var\left(x_i\right)^{-1} Var\left(f\right)^{-1} Cov\left(x_i, f\right)'$ denote the squared correlation between regressor x_i and factors f, where $Cov\left(x_i, f\right)$ is a row vector containing all covariances between x_i and $f_1, \ldots, f_r$. The coefficient $\rho_{x_i f}^2$ indicates the amount of information the factors f carry on predictor x_i. Using $Var\left(x_i\right) = \lambda_i' \lambda_i + \sigma_{e_i}^2$, where λ_i denotes the i^{th} row of Λ, the desired squared correlation $\rho_{x_i f}^2$ is achieved by taking

$$\sigma_{e_i}^2 = \frac{Cov\left(x_i, f\right) Var\left(f\right)^{-1} Cov\left(x_i, f\right)'}{\rho_{x_i f}^2} - \lambda_i' \lambda_i.$$

The error process is weakly cross- and serially correlated as in Bai and Ng (2002):

$$e_{i,t} = \frac{1}{c_i} \left(\alpha e_{i,t-1} + \nu_{it} + \sum_{j \neq 0, j=-J}^{J} \delta \nu_{i-j,t} \right)$$

with scaling constant $c_i = \sigma_{e_i}^{-1} \sqrt{\left(1 + 2J\delta^2\right) / \left(1 - \alpha^2\right)}$ and ν_{it} being $i.i.d$ standard normal. α and δ are both set to 0.2. J is selected by $\lceil n/20 \rceil$ such that about 10% of the variables are cross-correlated.

y_{t+h} is generated from the common factors of the regressors plus an additional

term z_t consisting of p variables of the predictor space. Including z_t accounts for the potential scenario that the factors of the regressor space are not necessarily incorporating the predictor information in an optimal manner regarding their forecasting power. Obviously, if the components of z_t were few in numbers and known to the forecaster, it would be straightforward to include them in the estimated model. However, in the likely scenario that the exact composition of z_t is unknown and many variables have an individual effect on y_{t+h} beyond their common factor structure, one might still have to rely on a pure factor model. The simulations below are performed with and without z_t.

For the parameter vectors γ and β different specifications are considered. For scaling purposes γ and β are normalized such that $\gamma\gamma' = n^{-1}\sum d_i$ and $\beta'\beta = p/n$. The idiosyncratic error term u_t is *i.i.d.* noise with mean zero and variance σ_u^2, and independent of e_t. Analogously to above, the desired squared correlation $\rho^2_{yf|z}$ between y and f conditionally on z_t is realized by taking

$$\sigma_u^2 = \frac{Cov\,(y, f|z)\,Var\,(f)^{-1}\,Cov\,(y, f|z)'}{\rho^2_{yf|z}} - \gamma\gamma'.$$

For each parameter setting, $T = 100$ simulated observations are used to estimate the model and to compute a single forecast $\hat{y}_{T+1}$. The simulation experiments focus on one-step-ahead predictions. Instead of forecasting multi-steps ahead, different choices for ρ^2_{yf} are considered. Extending the forecasting horizon or reducing ρ^2_{yf} essentially has the same effect of a higher forecasting uncertainty whereby the latter provides a simple and neat control over the DGP. The forecast accuracy is measured by the mean squared forecast error (MSE) over 5000 simulation runs. For the purpose of having a benchmark, the MSE is divided by the error variance σ_u^2. If the data generating process (2.13) - (2.15) was perfectly estimated, the MSE would equal 1.

2.4.2 Results

Two simulation experiments are performed. In the first one, the data generating process follows a standard factor model and y_t depends on the factors only, i.e. $\beta = 0$. Table 2.1 reports the results for this factor-DGP. For the second simulation experiment reported in Table 2.2, the forecasting equation (2.15) is augmented by z_t which consists of p randomly chosen predictors of X. This factor-regression-DGP allows the predictor variables to have an individual effect on y_t beyond their common factor structure.

For both DGPs, different settings regarding the number of regressors n and

the squared correlation ρ^2_{yf}, i.e. the amount of information the factors f carry on y, are considered. Furthermore, two different specifications for γ are implemented. For the first specification, denoted by $\gamma^{(1)}$, the entries in $\gamma^{(1)}$ are set to $d_1, d_2, \ldots, d_r$. This ensures that the relative importance of the factors in predicting y is the same as in explaining the variation in X. For the second one, denoted by $\gamma^{(2)}$, all entries in $\gamma^{(2)}$ are set equal to one such that all factors have equal weight in predicting y. In both settings, γ is normalized such that $\gamma\gamma' = n^{-1}\sum d_i$.

Tables 2.1 - 2.2 show the MSEs of the factor models for different choices of the supervision parameter θ. For the unsupervised factor model the number of factors is chosen according to the IC_{p2} information criterion by Bai and Ng (2002).[4] For the supervised factor model with θ determined by the likelihood approach, θ_{ml}, the supervision parameter and the number of factors are specified according to the sequential procedure proposed in Vervloet et al. (2015). When the supervision parameter is chosen by the information criterion, θ_{aic}, or by cross-validation, θ_{cv}, only one supervised factor is estimated. Following this route results in a very parsimonious model structure.

Considering the simulation results of the factor-DGP in Table 2.1 it is not surprising that the supervised model cannot improve upon the unsupervised one in most cases as the latter fits the DGP exactly. However, there is still an interesting insight when considering the results for $n = 30$ and $\rho^2_{yf} = 0.8$. Even under the factor-DGP the supervised factor model can be slightly superior when the relative importance of the factors in predicting y is not the same as in explaining the variation in X which is shown in the columns for $\gamma^{(2)}$. If, for instance, a minor principal component that explains only a small proportion of the variation in the regressors has a relatively strong effect on the forecasting target, a single supervised factor can provide more accurate forecasts than multiple unsupervised factors.

The ability of the supervised factor model to focus on the information in X that is relevant for forecasting creates an advantage that becomes more pronounced for the factor-regression-DGP presented in Table 2.2. For $n = 30$ and $\rho^2_{yf} = 0.8$ the supervised factor model is clearly superior under both settings for γ. Supervision enables the factors to incorporate information in the individual regressors that is not captured by their common factor structure and, thereby, improves their forecasting power. However, if ρ^2_{yf} is low, i.e. the error in the

[4] Among the criteria developed by Bai and Ng (2002), the IC_{p2} was found to recover the true factor number most accurately in the simulations.

Table 2.1 Simulation with factor-DGP

	n	30				70			
DGP:	ρ^2_{yf}	0.8		0.4		0.8		0.4	
	γ	$\gamma^{(1)}$	$\gamma^{(2)}$	$\gamma^{(1)}$	$\gamma^{(2)}$	$\gamma^{(1)}$	$\gamma^{(2)}$	$\gamma^{(1)}$	$\gamma^{(2)}$
unsupervised		1.23	1.94	1.11	1.23	1.14	1.76	1.10	1.21
		(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
θ_{aic}		1.32	1.89	1.13	1.32	1.53	2.72	1.17	1.39
		(.28)	(.43)	(.17)	(.26)	(.11)	(.13)	(.06)	(.07)
θ_{cv}		1.33	1.90	1.16	1.34	1.57	2.57	1.18	1.41
		(.35)	(.49)	(.26)	(.34)	(.12)	(.20)	(.08)	(.12)
θ_{ml}		1.28	2.13	1.48	1.30	1.20	2.07	1.16	1.35
		(.03)	(.03)	(.03)	(.03)	(.01)	(.01)	(.01)	(.01)
θ	.00	2.17	3.60	1.21	1.45	2.08	3.46	1.19	1.42
	.05	1.89	3.41	1.19	1.43	1.82	3.27	1.18	1.41
	.10	1.66	3.14	1.17	1.41	1.66	3.01	1.18	1.40
	.20	1.39	2.48	1.14	1.36	1.67	2.57	1.36	1.48
	.30	1.33	2.03	1.15	1.31	1.97	2.71	1.83	2.08
	.40	1.34	1.90	1.19	1.32	2.30	3.07	2.31	2.72
	.50	1.37	1.89	1.24	1.38	2.60	3.34	2.68	3.05
	.60	1.41	1.91	1.30	1.43	2.86	3.55	2.94	3.24
	.90	1.51	1.97	1.42	1.51	3.41	3.92	3.39	3.50

Notes: The top line specifies the DGP settings. Furthermore it is $s = 5$, $r = 8$, $\rho^2_{xf} = 0.8$, $p = 0$. Reported are the MSEs of the unsupervised and the supervised factor model with θ chosen by the criterion (aic), 10-fold cross-validation (cv), and the likelihood approach (ml). The average θ-values are reported in parentheses. The lower part of the table provides MSEs when θ is fixed to a given value. Except for the unsupervised model and the supervised model with θ_{ml} only one factor is estimated.

forecasting equation (2.15) is large and the regressors are not particularly informative, this advantage vanishes. This might be noteworthy in the context of multi-step-ahead forecasts that typically suffer from a higher forecasting uncertainty. In this case, relying on unsupervised factors can provide better results. In the related case where the idiosyncratic errors in the regressor equation (2.14) are large, i.e. ρ^2_{xf} small, supervision improves upon the forecasting performance of factor models as it helps to identify the common variation in X that is relevant for forecasting.[5]

The downside of the supervised factor model becomes obvious when considering the results for $n = 70$ under both DGPs. When the number of regressors is very large compared to the observations available, the supervised factor model suffers from overfitting. Considering the lower part of both tables, one can see that there is no choice for θ that improves upon the unsupervised factor model

[5] Results for small values of ρ^2_{xf} are not reported for the sake of readability but are available upon request.

Table 2.2 Simulation with factor-regression-DGP

DGP:	n	30				70			
	ρ^2_{yf}	0.8		0.4		0.8		0.4	
	γ	$\gamma^{(1)}$	$\gamma^{(2)}$	$\gamma^{(1)}$	$\gamma^{(2)}$	$\gamma^{(1)}$	$\gamma^{(2)}$	$\gamma^{(1)}$	$\gamma^{(2)}$
unsupervised		1.68	2.39	1.19	1.30	1.91	2.51	1.22	1.33
		(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
θ_{aic}		1.46	1.93	1.25	1.40	2.50	3.71	1.42	1.664
		(.47)	(.52)	(.25)	(.31)	(.16)	(.15)	(.09)	(.09)
θ_{cv}		1.46	1.93	1.26	1.40	2.36	3.09	1.45	1.67
		(.53)	(.57)	(.33)	(.39)	(.23)	(.27)	(.11)	(.14)
θ_{ml}		1.85	2.69	1.27	1.43	2.09	3.03	1.32	1.50
		(.03)	(.03)	(.03)	(.03)	(.01)	(.01)	(.01)	(.01)
θ	.00	3.79	5.21	1.48	1.71	4.19	5.58	1.55	1.78
	.05	3.26	4.80	1.42	1.67	3.62	5.11	1.49	1.73
	.10	2.76	4.27	1.37	1.62	3.07	4.52	1.45	1.68
	.20	1.98	3.06	1.29	1.50	2.40	3.35	1.53	1.69
	.30	1.61	2.28	1.25	1.40	2.38	2.98	1.99	2.17
	.40	1.49	2.01	1.26	1.38	2.59	3.17	2.46	2.70
	.50	1.46	1.95	1.30	1.41	2.82	3.38	2.78	3.00
	.60	1.47	1.94	1.34	1.44	3.01	3.56	3.01	3.19
	.90	1.52	1.98	1.43	1.51	3.44	3.92	3.39	3.49

Notes: The top line specifies the DGP settings. Furthermore, it is $s = 5$, $r = 8$, $\rho^2_{xf} = 0.8$, $p = n/2$. Reported are the MSEs of the unsupervised and the supervised factor model with θ chosen by the criterion (aic), 10-fold cross-validation (cv), and the likelihood approach (ml). The average θ-values are reported in parentheses. The lower part of the table provides MSEs when θ is fixed to a given value. Except for the unsupervised model and the supervised model with θ_{ml} only one factor is estimated.

for $n = 70$.

An appealing feature of the supervised factor model with θ chosen by the information criterion or by cross-validation is that a single factor is sufficient for the model to be competitive under the circumstances outlined above. This parsimony can be of particular interest if one is interested in having a single factor estimate, e.g. to express the 'state of the economy' with respect to a specific economic figure (GDP growth, inflation, etc.) by a single index. Using a supervised factor that is aligned with the target variable might then be more reasonable than choosing the largest principal component of the dataset. Regarding the means for selecting the supervision parameter θ, both tables report an overall good performance of the information criterion. In almost all cases, it chooses θ close to the 'optimal' θ from the grid in the lower part of the tables. Determining θ by cross-validation yields similar results with θ_{cv} being slightly larger than θ_{aic} on average. For a comparison with the likelihood approach of selecting θ one has to mind that the latter chooses both θ and the number of

factors while for the information criterion a single factor is sufficient. It is striking that the likelihood choice of θ does not yield an improvement in forecasting accuracy compared to the unsupervised factor model in any case. Hence, under the circumstance that favor the use of supervised factors, it seems advisable to choose θ by the information criterion or by cross-validation.

2.5 Empirical Application

2.5.1 Data and Forecasting Model

For the empirical application, a pseudo real-time forecasting exercise of the coming h-months growth rate of the two key macroeconomic variables Industrial Production ('INDPRO') and the Consumer Price Index ('CPI') is conducted. The dataset over the period 1960-01 to 2015-12 is taken from the monthly macroeconomic database provided by the Federal Reserve Bank of St. Louis. It contains 135 time series and is updated in real-time. McCracken and Ng (2016) show that factors extracted from this dataset contain the same predictive information as those from the often used Stock & Watson datasets. All variables are transformed to obtain stationary series as described in McCracken and Ng (2016). In the dataset, some series contain missing values or are only available over a limited time span. In line with Stock and Watson (2002b), the expectation-maximization algorithm is employed to estimate a balanced panel.

As demonstrated in the simulation study, the supervised factor model takes full effect when the number of predictors is not too large compared to the number of observations. To circumvent the dimensionality problem, two different data subsets are considered. For the first one, the predictor set is simply replaced by its major principal components such that 90% of the original variation in the data is retained. Neglecting only 10% of the data variation should both yield a sufficient reduction and not limit the supervised factor model too much in its ability to exploit information from smaller principal components for forecasting.

The second subset choice is motivated by the study of Boivin and Ng (2006), who find that factors extracted from a smaller pre-screened dataset often yield satisfactory or even better results than those from the full dataset. The selection is governed by two principles: First, only series that are commonly known to have leading or at least coincident characteristics are chosen. The selection is motivated by the components of the leading and coincident indices of the Conference Board. Second, using very similar series twice is avoided to prevent oversampling from a particular group. For example, the series on 'Housing Starts:

Total New Privately Owned' is included but its sub-series that report the same information for different U.S. regions are excluded. The final sub-dataset that is used to forecast Industrial Production consists of 34 time series and comprises labor market indicators, new orders and speed of delivery indices, housing starts, interest rate spreads, consumer expectations, stock market series, as well as industrial production, income and sales indices (see Table A.1 in Appendix A.2). When forecasting the CPI, aggregate price indices are added and some of the real variables are changed for their nominal counterparts when available such that the dataset consists of 41 time series in total (see Table A.2 in Appendix A.2).[6]

Let y_t denote one of the two series of interest. When forecasting Industrial Production, the dependent variable is defined as average annualized monthly growth:

$$y^h_{t+h|t} = (1200/h) \ln (\mathrm{IP}_{t+h}/\mathrm{IP}_t).$$

The Consumer Price Index is defined similarly but treated as $I(1)$:

$$y^h_{t+h|t} = (1200/h) \ln (\mathrm{CPI}_{t+h}/\mathrm{CPI}_t) - 1200 \ \ln (\mathrm{CPI}_t/\mathrm{CPI}_{t-1}).$$

The general forecasting function takes the form

$$\hat{y}^h_{t+h|t} = \hat{\alpha}_h + \hat{f}_{h,t}\hat{\gamma}'_h + \sum_{j=1}^{p} \hat{\delta}_{h,j} y_{t-j+1}.$$

For the unsupervised model, the number of factors r used for forecasting and the number of auto-regressive lags p are chosen by BIC with $1 \leq r \leq r_{max}$ and $0 \leq p \leq 6$, where r_{max} is determined by the IC_{p2} information criterion of Bai and Ng (2002). The choice of p made for the unsupervised factor model is taken over by the supervised models to ensure comparability of the forecasting performances of the different factor estimates. For Industrial Production, lags are excluded, i.e. $p = 0$, which is in line with the results of Stock and Watson (2002b) who show amongst other results that the pure diffusion index model performs best for this series. Since the supervision parameter can compensate for the number of factors, the supervised factor model is restricted to one factor only. The first forecast is made in 1980-01. $f_{h,t}, \gamma_h$ and $\delta_{h,j}$ are estimated by using data from 1960-01 through 1980-01. For the following periods, estimation and prediction is performed recursively on an expanding window.

[6] Instead of (real) interest rate spreads, nominal interest rates are included as these contain information about expected future inflation.

Table 2.3 Out-of-sample forecasting performance CPI

	$h = 1$		$h = 3$		$h = 9$	
	subset based on principal components					
θ_0	1.0000	(.00)	1.0000	(.00)	1.0000	(.00)
θ_{aic}	**0.8968***	(.47)	**0.9086**	(.47)	0.9374	(.50)
θ_{cv}	0.9064*	(.44)	0.9113	(.50)	**0.9369**	(.54)
θ_{ml}	0.9692	(.01)	0.9827*	(.01)	0.9539*	(.01)
	subset based on leading indicators					
θ_0	1.0146	(.00)	0.9679	(.00)	0.9166	(.00)
θ_{aic}	0.8869**	(.59)	0.9137	(.63)	0.8958	(.64)
θ_{cv}	**0.8811****	(.57)	**0.9114**	(.63)	**0.8947**	(.66)
θ_{ml}	0.9562	(.02)	0.9234	(.02)	0.9113	(.02)

Notes: The columns show the relative MSE of h-months-ahead forecasts over the period 1980-01 to 2015-12. Values below one indicate improved forecasting accuracy. The lowest MSEs are indicated in bold. One (two) stars mean 0.10 (0.05) statistical significance for the Diebold-Mariano test (1995) with HAC standard errors. The average θ-values for different selection means are reported in parentheses.

2.5.2 Results

Tables 2.3 and 2.4 report the performance of the supervised factor model in forecasting CPI inflation and Industrial Production growth h months ahead. The forecasting accuracy of the unsupervised factor model applied to the full database serves as a benchmark and its MSE is normalized to one for each forecasting horizon. All results are reported relatively to the respective benchmark with a value below one indicating a lower MSE than the benchmark model.

Overall, the results indicate that forecasts can be improved by using supervised factors. Especially over short forecasting horizons, supervision has a positive effect on prediction accuracy. For longer forecasting horizons, however, gains from supervision vanish. This supports the observation from the simulation study that supervising the factors is more promising when the regressor space is more informative, which is naturally the case for shorter forecasting horizons. When forecasting CPI inflation, supervised factors improve forecasting accuracy by up to 11%. The improvements are present for both the principal components subset and the leading indicator subset. Regarding Industrial Production, the improvements are less consistent. While supervision improves forecasting accuracy remarkably on the leading indicators subset by up to 16%, gains are small for the principal components subset. For the latter it might be the case that the supervised factor model is limited in benefiting from its ability to incorporate relevant information hidden in small principal components as some of them are

Table 2.4 Out-of-sample forecasting performance INDPRO

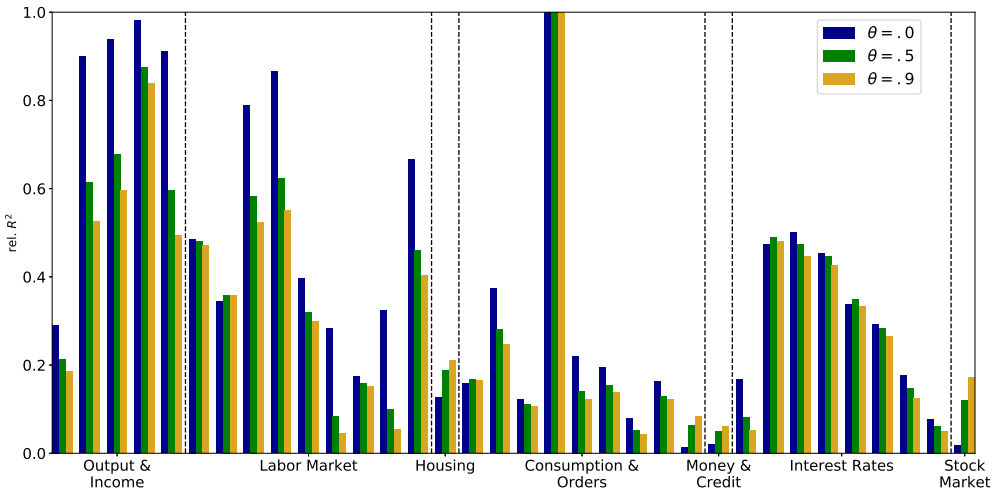
	$h = 1$		$h = 3$		$h = 9$	
	subset based on principal components					
θ_0	1.0000	(.00)	1.0000	(.00)	**1.0000**	(.00)
θ_{aic}	**0.9610**	(.35)	**0.9731**	(.40)	1.1265	(.47)
θ_{cv}	0.9638	(.38)	0.9910	(.42)	1.1320	(.48)
θ_{ml}	1.0040	(.01)	1.0086	(.01)	1.0411	(.01)
	subset based on leading indicators					
θ_0	1.0313	(.00)	1.0509	(.00)	1.1165	(.00)
θ_{aic}	**0.9225****	(.55)	0.8384**	(.60)	1.0376	(.67)
θ_{cv}	0.9229**	(.50)	**0.8356****	(.55)	1.0294	(.61)
θ_{ml}	1.0104	(.03)	1.0024	(.03)	**0.9502**	(.03)

Notes: The columns show the relative MSE of h-months-ahead forecasts over the period 1980-01 to 2015-12. Values below one indicate improved forecasting accuracy. The lowest MSEs are indicated in bold. One (two) stars mean 0.10 (0.05) statistical significance for the Diebold-Mariano test (1995) with HAC standard errors. The average θ-values for different selection means are reported in parentheses.

discarded by construction of the data subset. Finally, it might be noteworthy to recall that the supervised model requires only a single factor to achieve in most cases better or similar results as the model with multiple unsupervised factors.

Turning to the means for selecting the supervision parameter θ, it is apparent that the information criterion and 10-fold cross-validation choose almost the same θ value and, consequently, yield similar forecasting results. For a comparison with θ_{ml}, one has to recall that the likelihood approach chooses both θ and the number of factors while the information criterion estimates a single factor only. Hence, a one-to-one comparison of individual factor estimates would be pointless. However, when comparing the forecasting performances, Tables 2.3 and 2.4 show that the supervised factor model based on the information criterion or cross-validation yields better results than the one based on the likelihood approach in most cases. The use of more factors and a stronger focus on predictor space compression of the likelihood approach results in a small value for the supervision parameter such that forecasts do not differ a lot from the unsupervised factor model. While the supervision gains are rather small over short forecasting horizons, the conservative choice of θ_{ml} might be advantageous for less informative long-run forecasts to avoid overfitting.

Because the factors are only identified up to an orthogonal transformation, a detailed discussion of the individual factors is gratuitous. Nevertheless, the finding that forecasts can be improved by supervision suggests to briefly visualize the supervision effect. Figures 2.1 - 2.2 display the R^2 between the first factor



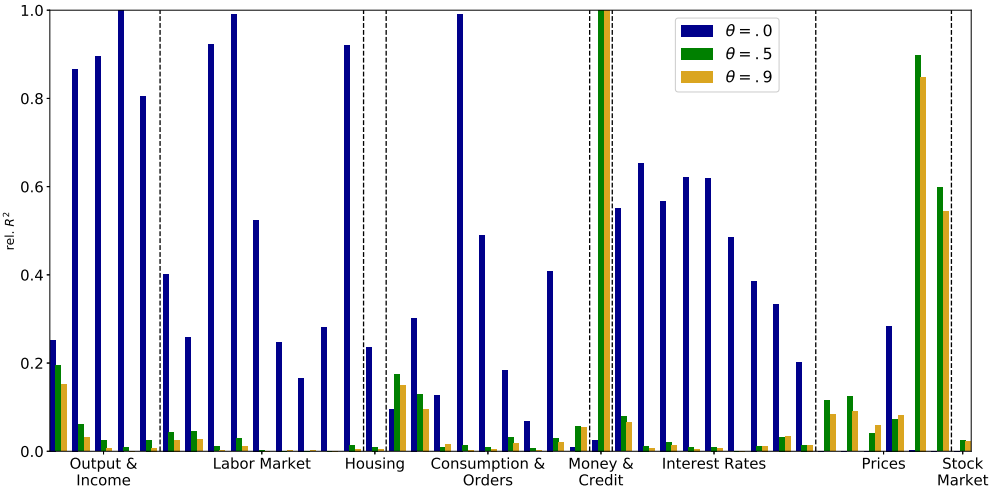
Notes: R^2 between the first factor and individual time series. The factor is supervised for 3-months-ahead predictions of Industrial Production growth.

Fig. 2.1 Effect of supervision with respect to Industrial Production.

and the individual regressor series for different degrees of supervision. Since the results look similar for all forecasting horizons, only the cases in which the factor is supervised for 3-months-ahead predictions are presented.[7] As by construction the correlation between factors and regressor series generally decreases with higher supervision, the bar charts are normalized such that for each degree of supervision the largest R^2 is set to one and the remaining bars are set to their value relative to the respective benchmark. As a consequence, figures 2.1 - 2.2 emphasize on differences in the relative importance of the individual regressor series for the first (supervised) factor.

Figure 2.1 shows that the loadings on those series that are commonly classified as coincident indicators mostly decrease with higher supervision. With Industrial Production as the target variable, the first factor puts less weight on variables from the 'Output & Income' group and on the 3$^{\text{rd}}$ to 5$^{\text{th}}$ series from the 'Labor Market' group. All of these series are known as having coincident characteristics. On the contrary, the higher the degree of supervision the more relative importance is attached to some leading indicators such as, for example, 'Housing Starts', the 'ISM: New Orders Index', (fourth bar in the 'Consumption & Orders' group), the 'Consumer Sentiment Index' (last bar in the 'Consumption & Orders' group), the 'Real M2 Money Stock', and the 'S&P: Industrials'

[7] Significant differences only show up when the factor is supervised for 9-months-ahead predictions of Industrial Production growth. In this case, supervision results in higher loadings on variables from the interest rates group.



Notes: R^2 between the first factor and individual time series. The factor is supervised for 3-months-ahead predictions of the Consumer Price Index.

Fig. 2.2 Effect of supervision with respect to the Consumer Price Index.

series.

For CPI inflation, figure 2.2 reports a much more pronounced difference between supervised and unsupervised factors than for Industrial Production. The supervised factors load almost exclusively on price series and the 'Real M2 Money Stock'. Some minor weight is put on the macro variables 'Income', 'Real Personal Consumption Expenditures', and 'Real Manufacturing and Trade Industries Sales' (first bars in respective groups) which are all usually listed as coincident indicators. These variables being relevant for forecasting might be reasonable against the backdrop of prices being lagging indicators. The major importance of the price variables in general reflects the persistence in the price series.

2.6 Conclusion

This study shows that supervision in factor estimation can improve the forecasting power of factor models. Supervised factors are particularly promising when relevant information in the regressor space is not captured by its major principal components. For the choice of the supervision parameter, the proposed information criterion finds an overall good balance between predictor space compression and target orientation of the estimated factors. A complication of the supervised factor model arises when the number of regressors is very large compared

to the observations available. In this case supervised factors are vulnerable to overfitting problems. It is concluded that incorporating supervision in factor estimation is especially promising for intermediate large regressor spaces.

In line with these findings, the empirical application compares the forecasting performance of the supervised factor model applied to a macroeconomic dataset consisting of 34-41 leading and coincident indicators to the unsupervised model using both the reduced and the full dataset comprising 135 time series. The results indicate that a single supervised factor can provide more accurate forecasts than the classical factor model.

Chapter 3

Macroeconomic Forecasting with Neural Network Reinforced Factor Models

3.1 Abstract

In many macroeconomic forecasting applications factor models are used to cope with large datasets. This study aligns variational autoencoders with macroeconomic factor modeling and proposes an extension to adapt this framework for forecasting exercises. Variational autoencoders are well suited for nonlinear dimensionality reduction. They estimate the distribution of the common latent variables by combining a statistical factor model with a purely data-driven neural network approach. It is demonstrated that the resulting deep factor model can be interpreted as a flexible nonlinear extension of the standard factor model. In the empirical part, it is analyzed whether factor models augmented by neural networks can achieve superior forecasting power. The results suggest significant improvements in the forecasting accuracy of four major US macroeconomic time series.

3.2 Introduction

Factor models have become popular in economics because they can cope with large data sets in an effective manner. They have served various purposes, for instance, in the construction of economic indicators or in forecasting real and nominal economic variables. The key idea behind factor models is that dependencies, i.e., covariances, among the variables in the data set are explained by a small number of latent factors which might then be used as a starting point for further analysis.

This study takes a look beyond the linear factor model. The objective is to analyze whether factor models enriched by elements from the machine learning

literature, more precisely by neural networks, can achieve superior forecasting power. While there have been some attempts to use neural networks in forecasting macroeconomic aggregates, e.g., Cook and Hall (2017), Nakamura (2005) and Tkacz (2001), this approach is different in the sense that the neural network is integrated into the factor model instead of serving as a substitute. For this purpose, variational autoencoder (VAE) can be exploited. VAEs have recently raised much attention in the machine learning literature. They seek to estimate the distribution of common latent variables underlying the data by combining a statistical factor model with purely data-driven neural networks. The explicit statistical model distinguishes VAEs from pure data-driven autoencoders that are a standard tool in the machine learning literature for dimensionality reduction[1] and have recently found applications in the finance literature (see, e.g., Gu et al. (2020) for an autoencoder asset pricing model).

This study attempts to align VAEs with the statistical factor model in the context of macroeconomic forecasting. The object of interest is forecasting when a large predictor space is available. A precise identification of the latent factor space or the common component per se is not attempted, which can be important for macroeconomic modeling or index construction. For the purpose of forecasting, the literature tends to prefer the fairly simple approach of Stock and Watson (2002a), that augments the forecasting equation by nonparametric factor estimates obtained from principal components, over an explicit dynamic parametric state space formulation of the problem (Boivin and Ng, 2005). When dealing with large predictor spaces, the nonparametric framework seems to be more robust than the state space approach that can suffer from the high dimensionality.[2] Variational autoencoders may be interpreted as *deep factor models* that can capture nonlinear common dynamics in the predictor space. Due to their high flexibility, they can be exploited to model complex patterns in the data. However, they do not provide a fully identified latent model such that their major purpose should be seen in forecasting instead of macroeconomic modeling. Indeed, the empirical application shows that significant improvements over the factor model are possible.

Incorporating machine learning techniques to macroeconomic forecasting has recently raised much interest in the literature. See Coulombe et al. (2019) and the references therein. Coulombe et al. (2019) analyze which features of machine

[1] See, for example, Goodfellow et al. (2016).

[2] Apart from that Stock and Watson (2016) suggest from the literature that in many empirical applications the differences between parametric and nonparametric implementations are rather small.

learning techniques are most salient for forecast accuracy gains. Two of their major findings are, first, that the ability of these techniques to capture nonlinearities in a nonparametric way constitutes the primary benefits for macroeconomic forecasting. Second, the results indicate that the factor model still provides an accurate means for dimensionality reduction in a big data framework that is not outperformed by alternative regularization methods such as, for instance, the Lasso approach. The deep factor approach considered in this study takes up exactly these two findings. It retains a basic factor structure but incorporates nonlinearities to capture complex patterns in the data beyond linearity.

Section 3.3 relates the factor model to variational autoencoders, and exposes the similarities and differences inherent to both approaches. Furthermore, an adjustment is proposed that adapts the VAE framework for prediction tasks. Section 3.4 provides an empirical forecasting application on a large macroeconomic dataset and shows that significant improvements in the forecasting accuracy of four major US macroeconomic time series can be achieved. Section 3.5 concludes and provides some directions for further research.

3.3 The Model

This section illustrates the relationship between the linear factor model and variational autoencoders with a focus on the similarities and constraints inherent to both modeling frameworks. While both approaches differ with respect to their exact model specifications and require different estimation techniques, they are built on a common core such that one might view VAEs as an extension of factor models. The section starts by recapitulating both approaches with an emphasis on how VAEs enhance the factor model. For a review on the different models, the reader is referred to more comprehensive articles.[3] This summary focuses on their shared characteristics.

3.3.1 The Factor Model Revisited

The statistical factor model is perhaps the most common example of a latent variable model. It rests on the assumption that the n variables x_t obey a factor structure of the form

$$x_t = \Lambda f_t + e_t \quad \text{for} \quad t = 1, \ldots, T, \tag{3.1}$$

[3] See Breitung and Choi (2013) and Stock and Watson (2006) for reviews on the factor model and Kingma and Welling (2013) for an introduction to variational autoencoders.

where f_t is an $r \times 1$ vector of common factors. The $n \times r$ matrix Λ contains the factor loadings and the error term e_t represents the variance unique to each variable x_i for $i = 1, \ldots, n$.[4] The latent factors are conventionally defined as $f_t \overset{iid}{\sim} \mathcal{N}(0, I_r)$ and $\Lambda'\Lambda$ being diagonal to impose identification restrictions. In case of the strict factor model, the idiosyncratic component is $e_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_e^2 I_n\right)$.[5]

The strict factor model can be overly restrictive in economic applications. It sometimes seems unrealistic to assume that both components f_t and e_{it} are *i.i.d.* However, for ease of exposition of the relationship to variational autoencoders, these assumptions are retained.

Inference

A conventional approach to estimate the latent factors is by means of principal components. In order to expose the relationship between the factor model and variational autoencoders, a probabilistic estimation approach is taken. Instead of using the principal components estimator, estimation by maximum likelihood and via an EM algorithm is considered. Equation (3.1) implies a probability distribution for $x_t|f_t$ of the form

$$x_t|f_t \sim \mathcal{N}\left(\Lambda f_t, \sigma_e^2 I_n\right).$$

With the distribution over the latent factors defined by $f_t \overset{iid}{\sim} \mathcal{N}(0, I_r)$ the marginal distribution of x_t follows from

$$p(x_t) = \int p(x_t|f_t)\, p(f_t)\, \mathrm{d}f_t. \tag{3.2}$$

as

$$x_t \sim \mathcal{N}\left(0, \sigma_e^2 I_n + \Lambda\Lambda'\right).$$

The log-likelihood of observing $\{x_t\}_{t=1}^T$ is

$$\begin{aligned}
\mathcal{L}\left(\Lambda, \sigma_e^2; x\right) &= \sum_{t=1}^{T} \ln p(x_t; \Lambda, \sigma_e^2) \\
&= \sum_{t=1}^{T} \left(-\frac{1}{2} \ln\left(|\sigma_e^2 I_n + \Lambda\Lambda'|\right) - \frac{1}{2} \operatorname{tr}\left\{ \left(\sigma_e^2 I_n + \Lambda\Lambda'\right)^{-1} x_t x_t' \right\} \right) \\
&= -\frac{T}{2} \left(\ln\left(|\sigma_e^2 I_n + \Lambda\Lambda'|\right) + \operatorname{tr}\left\{ \left(\sigma_e^2 I_n + \Lambda\Lambda'\right)^{-1} S \right\} \right),
\end{aligned} \tag{3.3}$$

[4] For notational convenience it is assumed that $\mathbb{E}[x_t] = 0$. In practical applications the variables are typically standardized prior to the estimation process.

[5] The strict factor model is also known as probabilistic PCA in the machine learning literature. See, e.g., Tipping and Bishop (1999).

where S is the sample covariance of $\{x_t\}_t^T$, and terms independent of the model parameters are omitted. Tipping and Bishop (1999) show that the log-likelihood is maximized when the columns of Λ span the principal subspace of the data. Hence, the strict factor model affects a mapping from the latent space into the principal subspace of the data, irrespective if maximum likelihood or the principal components estimator is applied.

Alternatively, the model can by estimated by using an EM algorithm (Tipping and Bishop, 1999), which provides insights into the shared characteristics of the factor model and VAEs. In the EM approach, the latent factors f_t can be considered as 'missing data'. The EM algorithm iterates between computing the expectation of the 'complete-data' log-likelihood of the joint distribution $p(x_t, f_t)$ with respect to $p(f_t|x_t; \Lambda, \sigma_e^2)$ (E-Step) and maximizing this expression with respect to the model parameters (Λ, σ_e^2) (M-Step). The complete-data log-likelihood is given as

$$\begin{aligned}
\mathcal{L}_c\left(\Lambda, \sigma_e^2; x, f\right) &= \sum_{t=1}^{T} \ln\left(p\left(x_t, f_t; \Lambda, \sigma_e^2\right)\right) \\
&= \sum_{t=1}^{T} \ln\left(p\left(x_t|f_t; \Lambda, \sigma_e^2\right) p\left(f_t\right)\right) \\
&= \sum_{t=1}^{T} \left(-\frac{n}{2}\ln\left(\sigma_e^2\right) - \frac{1}{2\sigma_e^2}\left(x_t - \Lambda f_t\right)'\left(x_t - \Lambda f_t\right) - \frac{1}{2} f_t' f_t\right), \quad (3.4)
\end{aligned}$$

where terms independent of the model parameters are omitted again. In the E-step, $\mathcal{L}_c$ is evaluated at the expected value of f_t given the data x_t and at current parameter values (Λ, σ_e^2). The factors f_t are conditionally distributed as

$$f_t|x_t \sim \mathcal{N}\left(\left(\sigma_e^2 I_r + \Lambda'\Lambda\right)^{-1}\Lambda' x_t, \left(I_r + \frac{1}{\sigma_e^2}\Lambda'\Lambda\right)^{-1}\right). \quad (3.5)$$

Taking the expectation of $\mathcal{L}_c$ from equation (3.4) with respect to $p\left(f_t|x_t; \Lambda, \sigma_e^2\right)$ gives

$$\begin{aligned}
\mathbb{E}_{p(.)}\left[\mathcal{L}_c\right] &= \sum_{t=1}^{T}\left(-\frac{n}{2}\ln\left(\sigma_e^2\right) - \frac{1}{2\sigma_e^2}\mathbb{E}_{p(.)}\left[\left(x_t - \Lambda f_t\right)'\left(x_t - \Lambda f_t\right)\right] - \frac{1}{2}\mathbb{E}_{p(.)}\left[f_t' f_t\right]\right) \\
&= \sum_{t=1}^{T}\left(-\frac{n}{2}\ln\left(\sigma_e^2\right) - \frac{1}{2\sigma_e^2}\left(\operatorname{tr}\left\{x_t x_t'\right\} - 2x_t'\Lambda\ \mathbb{E}_{p(.)}\left[f_t\right]\right)\right. \\
&\quad \left. + \operatorname{tr}\left\{\Lambda'\Lambda\mathbb{E}_{p(.)}\left[f_t f_t'\right]\right\} - \frac{1}{2}\operatorname{tr}\left\{\mathbb{E}_{p(.)}\left[f_t f_t'\right]\right\}\right) \quad (3.6)
\end{aligned}$$

with

$$\mathbb{E}_{p(.)}\left[f_t\right] = \left(\sigma_e^2 I_r + \Lambda'\Lambda\right)^{-1}\Lambda' x_t,$$
$$\mathbb{E}_{p(.)}\left[f_t f_t'\right] = \left(I_r + \frac{1}{\sigma_e^2}\Lambda'\Lambda\right)^{-1} + \mathbb{E}_{p(.)}\left[f_t\right]\mathbb{E}_{p(.)}\left[f_t\right]'.$$

In the M-step, $\mathbb{E}_{p(.)}\left[\mathcal{L}_c\left(\Lambda, \sigma_e^2; x, f\right)\right]$ is maximized with respect to Λ and σ_e^2. Taking the first order conditions of equation (3.6) and solving for Λ and σ_e^2 yields the parameter update rules:

$$\widetilde{\Lambda} = x_t\ \mathbb{E}_{p(.)}\left[f_t\right]'\left(\mathbb{E}_{p(.)}\left[f_t f_t'\right]\right)^{-1},$$
$$\widetilde{\sigma}_e^2 = \frac{1}{nT}\sum_{t=1}^{T}\left(\mathrm{tr}\left\{x_t x_t'\right\} - 2x_t'\Lambda\ \mathbb{E}_{p(.)}\left[f_t\right] + \mathrm{tr}\left\{\Lambda'\Lambda\ \mathbb{E}_{p(.)}\left[f_t f_t'\right]\right\}\right).$$

To summarize, the EM algorithm iterates between the E-step and the M-step until convergence:

E-step:

$$\mathbb{E}_{p(.)}\left[\mathcal{L}_c\left(\Lambda, \sigma_e^2; x, f\right)\right] = \int \ln\left(p\left(x, f; \Lambda, \sigma_e^2\right)\right)\ p\left(f|x; \Lambda^{(\mathrm{old})}, \left(\sigma_e^2\right)^{(\mathrm{old})}\right)\mathrm{d}f.$$

M-step:

$$\left(\widetilde{\Lambda}, \widetilde{\sigma}_e^2\right) = \arg\max_{\Lambda, \sigma_e^2} \mathbb{E}_{p(.)}\left[\mathcal{L}_c\left(\Lambda, \sigma_e^2; x, f\right)\right],$$

where the superscript 'old' indicates the parameter estimates from the previous iteration. It can be shown that the likelihood increases with every iteration of the EM algorithm, unless there is a local maximum of the likelihood function (Dempster et al., 1977).

3.3.2 A Deep Factor Model

The factor model in equation (3.1) specifies a linear relationship between the latent factors f_t and observed x_t. Variational autoencoders as introduced by Kingma and Welling (2013) can be regarded as deep factor models that allow for nonlinear mappings from the latent space to the observation space. The nonlinearity is achieved by a neural network. VAEs are capable of modeling arbitrary data distributions by combining neural networks with a latent factor model approach.

To see the relationship to the strict factor model, recall that the probability

of x_t given the latent factors f_t is distributed as

$$x_t|f_t \sim \mathcal{N}\left(\mu_t^{(x|f)}, \Sigma_t^{(x|f)}\right), \tag{3.7}$$

where

$$\mu_t^{(x|f)} = \Lambda f_t, \quad \Sigma_t^{(x|f)} = \sigma_e^2 I_n. \tag{3.8}$$

In contrast, VAEs parmeterize the conditional distribution in (3.7) with a neural network. In the VAE model, one has for the conditional mean and variance

$$\left(\left(\mu_t^{(x|f)}\right)', \operatorname{vec}\left(\Sigma_t^{(x|f)}\right)'\right)' = g\left(f_t; \theta\right), \tag{3.9}$$

where g is a nonlinear function of f_t with parameter vector θ that *decodes* the latent factors f_t to observed x_t. In variational autoencoders, neural networks are used as *decoder* functions. There are many possible choices for the *decoder network* $g(f_t; \theta)$. By way of illustration, consider a relatively simple multilayered perceptron (MLP) network for $g(f_t; \theta)$ with one hidden layer denoted by h_t and the *Rectified Linear Unit (ReLU)* activation function. Furthermore, retain the assumptions on the conditional covariance matrix in (3.8). In this case, the conditional moments of $x_t|f_t$ are modeled as

$$\begin{aligned}
\mu_t^{(x|f)} &= W_2^{\text{dec}} h_t + b_2^{\text{dec}}, \\
\ln\left(\sigma_e^2\right) &= W_3^{\text{dec}} h_t + b_3^{\text{dec}}, \\
h_t &= \max\left(0, W_1^{\text{dec}} f_t + b_1^{\text{dec}}\right),
\end{aligned}$$

where $\theta = \left\{W_1^{\text{dec}}, W_2^{\text{dec}}, W_3^{\text{dec}}, b_1^{\text{dec}}, b_2^{\text{dec}}, b_3^{\text{dec}}\right\}$ are the weights and biases of the decoder MLP. The flexibility of this approach allows to model complex data structures beyond linearity, but raises identification issues of the latent factors which is discussed in Section 3.3.4.

Variational Inference

The flexible parametrization of the deep factor model comes at the cost of not having an analytically closed form for the likelihood anymore. In case of the strict factor model, the log-likelihood $\mathcal{L}\left(\Lambda, \sigma_e^2; x\right)$ can be derived analytically as summarized in equations (3.2) - (3.3). However, when a deep factor structure is imposed, it is not possible to specify the log-likelihood function

$$\mathcal{L}\left(\theta; x\right) = \sum_{t=1}^{T} \ln\left(p(x_t; \theta)\right) = \sum_{t=1}^{T} \ln\left(\int p(x_t|f_t; \theta)\ p(f_t)\, \mathrm{d}f_t\right) \tag{3.10}$$

in closed-form. This becomes clear when considering the conditional mean and variance of $p(x_t|f_t;\theta)$, which are modeled as nonlinear functions of f_t, as it has been shown with an example in the previous subsection. The integral in equation (3.10) remains intractable because the marginalization step which has been exploited for linear factor model is no longer applicable in this nonlinear setting. As a consequence no closed form for the likelihood exists.

The alternative route of estimating the model parameters θ via the EM algorithm is not feasible by the same argument. The EM algorithm requires the conditional $p(f_t|x_t;\theta)$, which is in case of the linear factor model given by equation (3.5). However, for the deep factor model, $p(f_t|x_t;\theta)$ is not analytically tractable:

$$p(f_t|x_t;\theta) = \frac{p(x_t, f_t;\theta)}{p(x_t;\theta)} = \frac{p(x_t, f_t;\theta)}{\int p(x_t|f_t;\theta)\ p(f_t)\,\mathrm{d}f_t}. \tag{3.11}$$

The problem still is that the denominator in (3.11) cannot be solved in closed form.

As neither the marginal likelihood contribution $p(x_t;\theta)$ nor the conditional $p(f_t|x_t;\theta)$ can be computed analytically, one has to rely on approximation techniques to estimate the model parameters. In case of VAEs, the problem can be solved by variational inference, which might be seen as an generalization of the EM algorithm outlined in Section 3.3.1. The unknown distribution $p(f_t|x_t;\theta)$ is approximated by choosing a distribution $q(f_t;\phi)$ from a predefined family of densities $\mathcal{D}$, e.g., the class of normal distributions, such that the Kullback-Leibler (KL) divergence between $p(f_t|x_t;\theta)$ and $q(f_t;\phi)$ is minimized. The vector ϕ denotes the *variational parameters*. In case of the variational autoencoder, it comprises the weights and biases of the *encoder network* which *encodes* the high-dimensional observation vector x_t into the moments of $q(f_t;\phi)$. The subsequent section will give concrete form to $q(f_t;\phi)$ and the encoding network.

The KL divergence between $q(f_t;\phi)$ and $p(f_t|x_t;\theta)$ is defined as

$$\begin{aligned}
\mathrm{KL}(q(f_t;\phi)||p(f_t|x_t;\theta)) &= \int q(f_t;\phi) \ln\left(\frac{q(f_t;\phi)}{p(f_t|x_t;\theta)}\right) \mathrm{d}f_t \\
&= \int q(f_t;\phi) \ln\left(\frac{q(f_t;\phi)}{p(f_t, x_t;\theta)}\right) \mathrm{d}f_t + \ln\left(p(x_t;\theta)\right) \\
&= \mathbb{E}_{q(f;\phi)}\left[\ln\left(q(f_t;\phi)\right)\right] - \mathbb{E}_{q(f;\phi)}\left[\ln\left(p_\theta(x_t, f_t;\theta)\right)\right] + \ln\left(p(x_t;\theta)\right).
\end{aligned}$$

Hence, the marginal loglikelihood of datapoint t can be decomposed to

$$\ln\left(p(x_t;\theta)\right) = \mathrm{ELBO}\left(x_t;\theta,\phi\right) + \mathrm{KL}\left(q(f_t;\phi)||p(f_t|x_t;\theta)\right), \tag{3.12}$$

where

$$\text{ELBO}\left(x_t;\theta,\phi\right) = \mathbb{E}_{q(f;\phi)}\left[\ln\left(\frac{p(x_t,f_t;\theta)}{q(f_t;\phi)}\right)\right]. \tag{3.13}$$

Due to the non-negativity of the KL divergence, the $\text{ELBO}\left(x_t;\theta,\phi\right)$ provides a lower bound to the likelihood function (called *Evidence Lower BOund*). Although there is no analytical expression for the log-likelihood in equation (3.12) since $p(f_t|x_t;\theta)$ in the KL term is unknown, one can optimize the ELBO in equation (3.13).

A closer look at the ELBO reveals an analogy to the EM algorithm from Section 3.3.1. One can rewrite the ELBO as

$$\text{ELBO}\left(x_t;\theta,\phi\right) = \mathbb{E}_{q(f;\phi)}\left[\ln p(x_t,f_t;\theta)\right] - \mathbb{E}_{q(f;\phi)}\left[\ln q(f_t;\phi)\right], \tag{3.14}$$

where the first summand is the expected *complete-data* log-likelihood of datapoint t similar to equation (3.6) from the strict factor model, and the second term is the entropy of $q(f_t;\phi)$. Now, optimization consists of iteratively maximizing the $\text{ELBO}\left(x_t;\theta,\phi\right)$ with respect to the variational parameters ϕ of $q(f_t;\phi)$ holding the parameters θ fixed ('E-step') and maximizing the $\text{ELBO}\left(x_t;\theta,\phi\right)$ with respect to θ under given $q(f_t;\phi)$ ('M-step'). Note that the EM algorithm for the strict factor model maximizes only the first term in equation (3.14), i.e., the expected *complete-data* log-likelihood. It exploits the fact that the ELBO equals the log-likelihood $\ln\left(p(x_t;\theta)\right)$ when one chooses $q(f_t;\phi) = p(f_t|x_t;\theta)$, which of course requires $p(f_t|x_t;\theta)$ to be known and tractable.[6]

Parameter Estimation in the Deep Factor Model

The latent factors are still defined as $f_t \overset{iid}{\sim} \mathcal{N}\left(0, I_r\right)$. Furthermore, let $p(x_t|f_t;\theta)$ be a multivariate Gaussian whose moments are computed from f_t with an MLP as defined by equation (3.9) and illustrated by the subsequent example in section 3.3.2. Since the posterior $p(f_t|x_t;\theta)$ cannot be computed analytically, it has to be approximated. To this end, it is assumed that the true posterior can be approximated by $q(f_t;\phi)$, where $q(f_t;\phi)$ is the probability density function of a normal distribution with mean vector $\mu_t^{(f|x)}$ diagonal covariance matrix $\Sigma_t^{(f|x)}$. More precisely, $q(f_t;\phi)$ is chosen as a probability density function from the class of normal distributions and can be described by

$$q(f_t;\phi) = \mathcal{N}\left(f_t;\mu_t^{(f|x)},\Sigma_t^{(f|x)}\right), \tag{3.15}$$

[6] For a more extensive and general description of variational inference, the review of Blei et al. (2017) is recommended.

where

$$\left(\left(\mu_t^{(f|x)}\right)', \text{vec}\left(\Sigma_t^{(f|x)}\right)'\right)' = v(x_t;\phi), \tag{3.16}$$

and where $v(x_t;\phi)$ is a neural network with weights and biases summarized in the variational parameter vector ϕ. $v(x_t;\phi)$ denotes the *encoding network* that *encodes* x_t into the distribution parameters of $q(f_t;\phi)$. Using the same network structure as in the example for the decoding network in Section 3.3.2, one obtains

$$h_t = \max\left(0, W_1^{\text{enc}} x_t + b_1^{\text{enc}}\right), \tag{3.17}$$

$$\mu_t^{(f|x)} = W_2^{\text{enc}} h_t + b_2^{\text{enc}}, \tag{3.18}$$

$$\ln\left(\sigma_t^2\right) = W_3^{\text{enc}} h_t + b_3^{\text{enc}}, \tag{3.19}$$

where σ_t^2 denotes the vector of the diagonal entries of $\Sigma_t^{(f|x)}$. The weights and biases $\phi = \{W_1^{\text{enc}}, W_2^{\text{enc}}, W_3^{\text{enc}}, b_1^{\text{enc}}, b_2^{\text{enc}}, b_3^{\text{enc}}\}$ of the encoding network are the variational parameters of the model.

The objective function is the ELBO which is obtained for *i.i.d.* data as the sum of individual-datapoint ELBOs:

$$\text{ELBO}(x;\theta,\phi) = \sum_{t=1}^{T} \text{ELBO}(x_t;\theta,\phi)$$

The individual ELBO contributions can be decomposed to

$$\begin{aligned}\text{ELBO}(x_t;\theta,\phi) &= \mathbb{E}_{q(f;\phi)}\left[\ln\left(\frac{p(x_t|f_t;\theta)\ p(f_t)}{q(f_t;\phi)}\right)\right] \\ &= \mathbb{E}_{q(f;\phi)}\left[\ln\left(p(x_t|f_t;\theta)\right)\right] - \text{KL}\left(q(f_t;\phi)||p(f_t)\right).\end{aligned} \tag{3.20}$$

The first term denotes the expected model fit. It describes how well the observations x_t can be reconstructed from the latent f_t by the decoder neural network $g(f_t;\theta)$. The second term gives the deviation of the distribution of the encoder output from the prior $p(f_t;\theta)$. The Kullback-Leibler divergence can be computed analytically because $q(f_t;\phi)$ and $p(f_t;\theta)$ are assumed to be Gaussian. One obtains

$$\begin{aligned}\text{KL}\left(q(f_t;\phi)||p(f_t)\right) &= \frac{1}{2}\left(\ln\left(\frac{|I_r|}{|\sigma_t^2 I_r|}\right) - r + \text{tr}\left\{I_r^{-1}\sigma_t^2 I_r\right\} + \left(\mu_t^{(f|x)}\right)' I_r^{-1}\mu_t^{(f|x)}\right) \\ &= \frac{1}{2}\left(-\sum_{j=1}^{r}\ln\left(\sigma_{jt}^2\right) - r + \sum_{j=1}^{r}\sigma_{jt}^2 + \sum_{j=1}^{r}\left(\mu_{jt}^{(f|x)}\right)^2\right) \\ &= \frac{1}{2}\sum_{j=1}^{r}\left(-\ln\left(\sigma_{jt}^2\right) - 1 + \sigma_{jt}^2 + \left(\mu_{jt}^{(f|x)}\right)^2\right),\end{aligned} \tag{3.21}$$

where the index j indicates the j^{th} element of the vectors $\mu_t^{(f|x)}$ and σ_t^2, resp. Recall that $\mu_t^{(f|x)}$ and σ_t^2 depend on the variational parameters ϕ as exemplified by the encoder network in equations (3.17) to (3.19).

Substituting equation (3.21) in (3.20) one obtains the individual ELBO contributions as

$$\begin{aligned}\text{ELBO}\left(x_t;\theta,\phi\right) = \mathbb{E}_{q(f;\phi)}\left[\ln\left(p(x_t|f_t;\theta)\right)\right] \\ + \frac{1}{2}\sum_{j=1}^{r}\left(\ln\left(\sigma_{jt}^2\right) + 1 - \sigma_{jt}^2 - \left(\mu_{jt}^{(f|x)}\right)^2\right).\end{aligned} \tag{3.22}$$

The ELBO is optimized by stochastic gradient descent with respect to the decoder parameters θ and the encoder/variational parameters ϕ. Before taking the gradients, the first term in equation (3.22) has to be approximated by a Monte Carlo estimator as the expectation cannot be computed analytically. It can be approximated by Monte Carlo sampling

$$\mathbb{E}_{q(f;\phi)}\left[\ln\left(p(x_t|f_t;\theta)\right)\right] \approx \frac{1}{L}\sum_{l=1}^{L}\ln\left(p(x_t|\tilde{f}_{lt};\theta)\right), \tag{3.23}$$

where $\tilde{f}_{lt}$ is sampled by $\tilde{f}_{lt} = \mu_t^{(f|x)} + \sigma_t^2 \odot \epsilon_l$ from $\epsilon_l \sim \mathcal{N}\left(0, I_r\right)$. $\odot$ denotes an element-wise multiplication. Note that this transformation (the so-called *reparametrization trick*) is necessary as it makes the Monte Carlo estimate of the expectation $\mathbb{E}_{q(f;\phi)}\left[\ln\left(p(x_t|f_t;\theta)\right)\right]$ differentiable with respect to ϕ. The problem is that the gradient of the expectation term has to be taken with respect to ϕ. However, the expectation is taken with respect to the distribution $q(f;\phi)$, which itself is a function of ϕ. In such cases, it holds in general that

$$\nabla_\phi \mathbb{E}_{q(f;\phi)}\left[\ln\left(p(x_t|f_t;\theta)\right)\right] \neq \mathbb{E}_{q(f;\phi)}\left[\nabla_\phi \ln\left(p(x_t|f_t;\theta)\right)\right].$$

Using the *reparametrization trick*, expectation and gradient operators become commutative:

$$\begin{aligned}\nabla_\phi \mathbb{E}_{q(f;\phi)}\left[\ln\left(p(x_t|f_t;\theta)\right)\right] &= \nabla_\phi \mathbb{E}_{p(\epsilon)}\left[\ln\left(p(x_t|\tilde{f}_t;\theta)\right)\right] \\ &= \mathbb{E}_{p(\epsilon)}\left[\nabla_\phi \ln\left(p(x_t|\tilde{f}_t;\theta)\right)\right]\end{aligned}$$

such that the Monte Carlo estimator in equation (3.23) is differentiable with respect to ϕ. The full proof is given by Kingma and Welling (2013).

Using the Monte Carlo estimator from equation (3.23) in (3.22) the objective

function for datapoint x_t becomes

$$\begin{aligned}\text{ELBO}\left(x_t;\theta,\phi\right) \approx \frac{1}{L}\sum_{l=1}^{L}\ln\left(p(x_t|\tilde{f}_{lt};\theta)\right) \\ + \frac{1}{2}\sum_{j=1}^{r}\left(\ln\left(\sigma_{jt}^2\right)+1-\sigma_{jt}^2-\left(\mu_{jt}^{(f|x)}\right)^2\right),\end{aligned}$$

which can be optimized by stochastic gradient descent with respect to the model parameters θ and ϕ.

3.3.3 A Supervised Deep Factor Model

The variational autoencoder focuses on compressing the common variation in $\{x_t\}_{t=1}^T$ within a few latent factors. If the overall objective is to exploit these factors in forecasting, it seems a natural extension to the model to include the forecasting target y_{t+h} in the factor estimation process as well, such that the factors are supervised with respect to the variable of interest. In order to extend the VAE framework, the likelihood is augmented to $p(x_t, y_{t+h})$ and analogously to above it is assumed that

$$\begin{aligned}x_t|f_t &\sim \mathcal{N}\left(\mu_t^{(x|f)}, \Sigma_t^{(x|f)}\right), \\ y_{t+h}|f_t &\sim \mathcal{N}\left(f_t'\beta, \sigma_u^2\right).\end{aligned}$$

For simplicity, a linear relationship between latent factors f_t and the forecasting target y_{t+h} and a constant error variance is assumed. Hence, the model allows for a nonlinear factor composition but retains the linearity in the forecasting relationship between f_t and y_{t+h}. It is straightforward to extent this to a nonlinear relationship by parametrizing the conditional mean and variance of $y_{t+h}|f_t$ with a neural network as well. Furthermore, conditional independence between the regressors x_t and the forecasting target y_{t+h} is assumed, such that their joint distribution conditional on f_t factorizes according to

$$p(x_t, y_{t+h}|f_t) = p(x_t|f_t)\ p(y_{t+h}|f_t).$$

This assumption is motivated by the idea that the predictable dynamics of y_{t+h} are accounted for by the latent factors underlying the predictor space. This idea can be seen as a general motivation for a factor modeling approach in forecasting exercises. Nevertheless, it remains to some extent a restrictive assumption, but it allows to simplify the estimation process considerably.

The estimation strategy remains the same as outlined in the previous sec-

tions. The unknown distribution $p(f_t|x_t, y_{t+h}; \theta, \beta)$ has to approximated by choosing a function $q(f_t; \phi)$ from the class of normal distributions with mean vector $\mu_t^{(f|x)}$ and diagonal covariance matrix $\Sigma_t^{(f|x)}$, where the conditional moments are parametrized by the encoding network $v(x_t; \phi)$ as shown in equation (3.16) and the subsequent example. Again $q(f_t; \phi)$ is chosen such that the KL divergence between $p(f_t|x_t, y_{t+h}; \theta, \beta)$ and $q(f_t; \phi)$ is minimized:

$$\begin{aligned}\mathrm{KL}\left(q(f_t;\phi)||p(f_t|x_t, y_{t+h};\theta,\beta)\right) &= \int q(f_t;\phi) \ln\left(\frac{q(f_t;\phi)}{p(f_t|x_t, y_{t+h};\theta,\beta)}\right) \mathrm{d}f_t \\ &= \mathbb{E}\left[\ln\left(q(f_t;\phi)\right)\right] - \mathbb{E}_{q(f;\phi)}\left[\ln\left(p_\theta(x_t, y_{t+h} f_t;\theta,\beta)\right)\right] + \ln\left(p(x_t, y_{t+h};\theta,\beta)\right).\end{aligned}$$

Hence, the marginal loglikelihood of datapoint pair (x_t, y_{t+h}) can be decomposed to

$$\ln\left(p(x_t, y_{t+h};\theta)\right) = \mathrm{ELBO}\left(x_t, y_{t+h};\theta,\phi,\beta\right) + \mathrm{KL}\left(q(f_t;\phi)||p(f_t|x_t, y_{t+h};\theta,\beta)\right),$$

with the ELBO objective being

$$\begin{aligned}\mathrm{ELBO}\left(x_t, y_{t+h};\theta,\phi,\beta\right) &= \mathbb{E}_{q(f;\phi)}\left[\ln\left(\frac{p(x_t, y_{t+h}, f_t;\theta,\beta)}{q(f_t;\phi)}\right)\right] \\ &= \mathbb{E}_{q(f;\phi)}\left[\ln\left(\frac{p(x_t|f_t;\theta)p(y_{t+h}|f_t;\beta)p(f_t)}{q(f_t;\phi)}\right)\right] \\ &= \mathbb{E}_{q(f;\phi)}\left[\ln\left(p(x_t|f_t;\theta)\right)\right] + \mathbb{E}_{q(f;\phi)}\left[\ln\left(p(y_{t+h}|f_t;\beta)\right)\right] \\ &\quad - \mathrm{KL}\left(q(f_t;\phi)||p(f_t)\right).\end{aligned} \tag{3.24}$$

As a further extension, it can be beneficial to control the degree to which the latent factors f_t focus on the reconstruction of x_t or on the prediction of y_{t+h}. For this purpose a *supervision* parameter α is introduced. The posterior distribution of the latent factors f_t given x_t and y_{t+h} in this supervised deep factor model can be formalized as

$$p\left(f_t|x_t, y_{t+h}\right) \propto \left(p\left(x_t|f_t\right)\right)^{(1-\alpha)} \left(p\left(y_{t+h}|f_t\right)\right)^{\alpha} p\left(f_t\right),$$

where α is a parameter from 0 to 1 that governs the degree of supervision. Since $p\left(x_t|f_t\right)$ and $p\left(y_{t+h}|f_t\right)$ are Gaussian distributions with covariances $\Sigma_t^{(x|f)}$ and σ_u^2, respectively., $\left(p\left(x_t|f_t\right)\right)^{(1-\alpha)}$ and $\left(p\left(y_{t+h}|f_t\right)\right)^{\alpha}$ are also Gaussian with covariances $\frac{1}{1-\alpha}\Sigma_t^{(x|f)}$ and $\frac{1}{\alpha}\sigma_u^2$, respectively.[7] The *supervised* ELBO follows as

[7] Strictly speaking $\left(p\left(x_t|f_t\right)\right)^{(1-\alpha)}$ and $\left(p\left(y_{t+h}|f_t\right)\right)^{\alpha}$ are proportional to a Gaussian up to the normalization factor $1/\int_x \left(p\left(x_t|f_t\right)\right)^{(1-\alpha)} \mathrm{d}x$ and $1/\int_y \left(p\left(y_{t+h}|f_t\right)\right)^{\alpha} \mathrm{d}y$, respectively.

$$\begin{aligned}\text{ELBO}^{(\alpha)}\left(x_t, y_{t+h}; \theta, \phi\right) &= \mathbb{E}_{q(f;\phi)}\left[\ln\left(\frac{\left(p(x_t|f_t;\theta)\right)^{(1-\alpha)}\left(p(y_{t+h}|f_t;\beta)\right)^{\alpha} p(f_t)}{q(f_t;\phi)}\right)\right] \\ &= (1-\alpha)\,\mathbb{E}_{q(f;\phi)}\left[\ln p(x_t|f_t;\theta)\right] + \alpha\mathbb{E}_{q(f;\phi)}\left[\ln p(y_{t+h}|f_t;\beta)\right] \\ &\quad - \text{KL}\left(q(f_t;\phi)||p(f_t)\right).\end{aligned} \tag{3.25}$$

For $\alpha = 0$, the original variational autoencoder is restored. Note that the supervised deep factor model exhibits some similarites to Principal Covariate Regression (PCovR) of de Jong and Kiers (1992) who estimate the factors f_t by a loss function that is a weighted sum of the squared reduction error of f_t on x_t and a prediction error of f_t on y_t.

3.3.4 Flexibility, Identification, and Robustness

While both the factor model and variational autoencoders share some common characteristics, they follow a distinct approach to modeling the data distribution. This holds especially for the requirements regarding parameter identification, which makes both models different in their scope of application.

Identification

As pointed out, the flexible VAE framework allows to describe complex data structures. However, this comes at the cost of unidentifiability of the true model parameters. Identification in latent variable models is by nature a difficult task. Even for the strict factor model, which is subject to restrictive assumptions, it is

$$\begin{aligned}(f_t, x_t)' &\sim \mathcal{N}\left(\begin{bmatrix}0\\ \mu_x\end{bmatrix}, \begin{bmatrix}I_r & \Lambda' \\ \Lambda & \sigma I_n + \Lambda\Lambda'\end{bmatrix}\right), \\ f_t|x_t &\sim \mathcal{N}\left(\left(\sigma_e^2 I_r + \Lambda'\Lambda\right)^{-1}\Lambda' x_t, \left(I_r + \frac{1}{\sigma_e^2}\Lambda'\Lambda\right)^{-1}\right),\end{aligned}$$

such that the model only is identified up to an orthogonal transformation, e.g., $\Lambda Q'Q\Lambda' = \Lambda\Lambda'$, where Q is a 'rotation matrix' with $Q'Q = I$. The ML-estimators reviewed in Section 3.3.1 require these assumptions to identify the factor space. Weakening the assumptions by allowing for cross-sectional correlation immediately raises severe identification problems since the factor model with an unrestricted error covariance matrix Σ_e involves $n\left(n+1\right)/2 + rn$ parameters whereas the covariance matrix of the data entails only $n\left(n+1\right)/2$ parameters.[8] Never-

[8] Note that under certain conditions it is still possible to consistently estimate the factor space by Principal Components even if the covariance parameters of the idiosyncratic components

theless, identification of the posterior factor space is at least possible to some extent such that the factor model offers some room for interpretation of the factors and their loadings.

The VAE framework does not allow to derive an explicitly identified form for $p(f_t, x_t)$ and $p(f_t|x_t)$ as in the case of the factor model, where both distributions are identified up an orthogonal transformation as explained above. The VAE model assumes that the observed data x_t stem from an underlying joint distribution $p_{\theta^*}(x_t, f_t) = p_{\theta^*}(x_t|f_t)p_{\theta^*}(f_t)$, where θ^* denotes the true but unknown model parameters. The model then gives rise to the observed distribution of the data by

$$p_\theta(x_t) = \int p_\theta(x_t, f_t)\,\mathrm{d}f = \int p_\theta(x_t|f_t)p_\theta(f_t)\,\mathrm{d}f_t.$$

The VAE approach estimates a full generative model $p_\theta(x_t, f_t) = p_\theta(x_t|f_t)p_\theta(f_t)$ and an inference model $q_\phi(f_t|x_t)$ that approximates the unknown $p_\theta(f_t|x_t)$. However, there are no guarantees whether these distributions actually identify the true data generating process. All that can be inferred is that the VAE optimizes the parameters θ towards the approximate maximum marginal likelihood objective (ELBO) of the data such that after optimization

$$p_\theta(x_t) \approx p_{\theta^*}(x_t).$$

However, it is impossible to learn the true joint distribution over both observed and latent variables or the true posterior distribution of the latent variables. This would only be possible for a fully identified model.[9]

The lack of explicit identification of the factor space affects that there is no (economic) interpretation of the composition of the latent variables. Obviously, this restricts VAEs in their scope of econometric applications. However, if the interest lies primarily in the forecasting power of latent factor models, this aspect is of minor importance and the deep factor model can be considered a reasonable extension to the factor model.

Flexibility and Robustness

Both the factor model and VAEs aim at learning a low-dimensional representation of the observed data x_t. The stochastic factor model can be regarded as a restricted VAE with a linear decoder function that generates x_t by a linear

are left unidentified. For a review see, e.g., Breitung and Choi (2013).

[9] Khemakhem et al. (2020) adress the issue of identification of the true joint distribution over observed and latent variables by implementing a factorized prior distribution over the latent variables that is conditioned on an additionally observed variable.

transformation of the latent f_t. This restriction to linear dependencies among the variables in x_t makes the factor model neat and robust to overfitting.

The flexible VAE framework allows to describe more complex patterns in the data. By modeling the first two moments of $p(x_t|f_t)$ with a neural network, VAEs can capture a rich class of data distributions $p(x_t)$. The deep factor model finds a low-dimensional manifold spanned by f_t that summarizes the (nonlinear) common characteristics of the observation space x_t. The parametrization via a neural network has the benefit that one does not have to manually engineer the most appropriate (nonlinear) function $x_t = g^*(f_t)$ connecting the latent space with the observation space. Instead, the weights and biases summarized in the parameter vector θ of the neural network are used to learn g from a broad class of functions such that $g(f_t;\theta)$ is driven to match $g^*(f_t)$ during the training process.

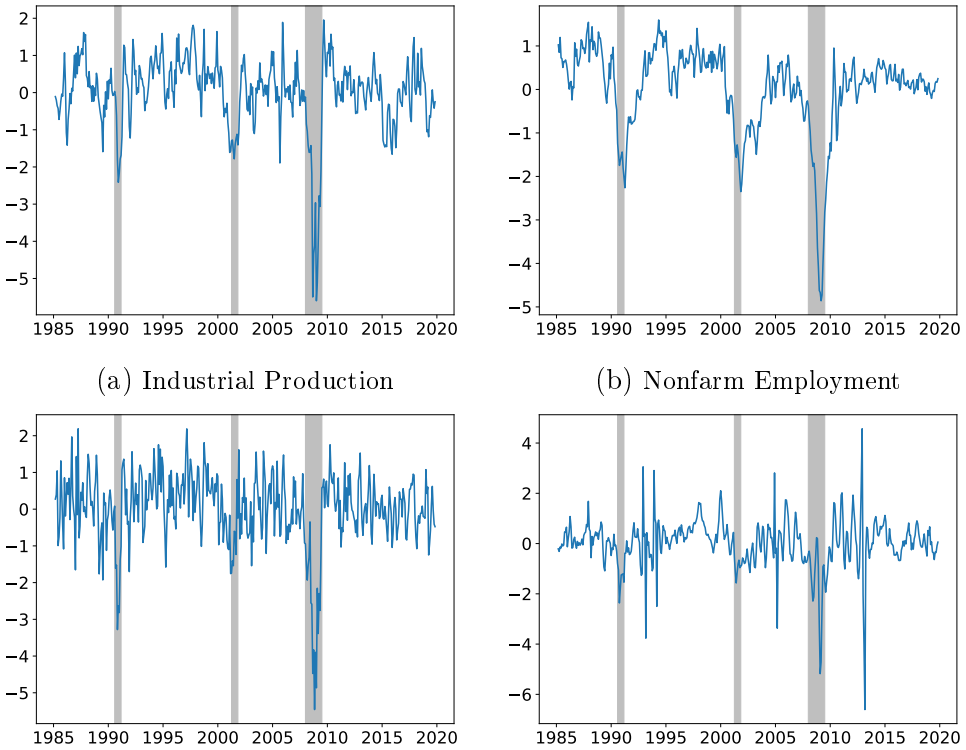
The ability of the deep factor model to learn complex low-dimensional representations of the data x_t points out the risk of overfitting. When the capacity of the decoder network $g(f_t;\theta)$ is allowed to become too great, the model can fail to learn informative latent factors. Theoretically, one could imagine that each variable in x_t could be recovered almost exactly by a single latent factor f_t with a very powerful nonlinear decoder $g(f_t;\theta)$.[10]

3.4 Empirical Application

3.4.1 Data and Forecasting Models

For the empirical application, an out-of-sample forecasting exercise of the coming h-months growth rate of the macroeconomic variables Industrial Production, Nonfarm Employment, Real Manufacturing and Trade Industries Sales and Real Personal Income ex Transfer Receipts over the period 1985-01 to 2019-12 is conducted. These four variables constitute the economic Coincident Index maintained by the Conference Boards that can be used to describe the current state of the economy. Figure 3.1 visualizes the series over the forecasting period. The dataset is taken from the monthly macroeconomic database provided by the Federal Reserve Bank of St. Louis. It contains 128 time series and is updated in real-time. All variables are transformed to obtain stationary series as described in McCracken and Ng (2016). In the dataset, some series contain missing values or are only available over a limited time span. In line with Stock and Watson (2002b), the EM algorithm is employed to estimate a balanced panel. The forecasting time span encompasses the so-called period of *Great Moderation* be-

[10] This problem is known as *posterior collapse* in the literature.



(a) Industrial Production

(b) Nonfarm Employment

(c) Real Manufacturing and Trade Industries Sales

(d) Real Personal Income ex Transfer Receipts

Notes: The figure shows the average annualized monthly growth of the forecast targets over the period 1985-01 to 2019-12. For visualization purposes the series are standardized and averaged over $h = 3$ months as described in equation (3.26). The shaded areas indicate recession periods.

Fig. 3.1 Forecasting targets.

ginning in the mid 80ies, where macroeconomic fluctuations began to dampen in comparison to the decades before, and the distortions in succession of the financial crisis in 2007. The first forecast is made in 1984-12 such that the performance comparison includes overall $T = 420 - h$ forecasts.

Let y_t denote the forecasting target. The dependent variable is defined as average annualized monthly growth, i.e., in case of Industrial Production

$$y_{t+h|t} = (1200/h) \ln (\mathrm{IP}_{t+h}/\mathrm{IP}_t). \tag{3.26}$$

As a benchmark model the general forecasting function of the Stock and Watson (2002b) approach is taken

$$\hat{y}_{t+h|t} = \hat{c}_h + \hat{f}_t' \hat{\beta}_h$$

with the factors $\hat{f}_t$ estimated by Principal Components. The number of factors

is specified according the IC_{p2} information criterion of Bai and Ng (2002) with a maximum number of factors set to 8. To ensure comparability of the forecasting performances, the choice of IC_{p2} information criterion is also used for the deep factor model. Note that both lags of $\hat{f}_t$ and autoregressive lags of the forecasting target are excluded. This is in line with the findings of Stock and Watson (2002b) who show amongst other results that the pure diffusion index model exhibits a comparable or even better forecasting performance than a forecasting model that is augmented by the respective lags. Furthermore, the exclusion of lags allows for a tailored comparison of the factor model and its deep counterpart as no other variables beyond the factors have an impact on the forecasting power.

For the deep factor model the forecasts are obtained by

$$\hat{y}_{t+h|t} = \hat{c}_h + \hat{f}'_{h,t}\hat{\beta}_h \quad \text{with} \quad \hat{f}_{h,t} = \mu_t^{(f|x)} = v(x_t; \hat{\phi}), \tag{3.27}$$

where $v(x_t; \hat{\phi})$ denotes the encoding neural network as specified in equations (3.15) - (3.16). Note the explicit dependence of $\hat{f}_{h,t}$ on h which should indicate that in case of the supervised deep factor model the parameters ϕ of the encoding network and thus the factor estimates are influenced by the forecasting target.

To obtain the first forecast, the model parameters and the number of factors are estimated using data from 1975-01 through 1984-12. For the following periods, estimation and prediction is performed recursively on an rolling window of 10 years.

Although the analogy of the supervised deep factor model to PCovR is mentioned in Section 3.3.3, PCovR is not included in the forecast comparison as it performs poorly on the given dataset.[11]

3.4.2 Implementation

The deep factor model requires the choice of several hyperparameters regarding the neural network topologies and the stochastic gradient descent optimizer to calculate the model parameters. These hyperparameters are either chosen by crossvalidation or according to recommendations from the literature.

Neural Network Topologies

The architectures of both the encoding $v(x_t; \phi)$ and decoding $g(f_t; \theta)$ networks are guided by the following considerations: First, rather small models are pre-

[11] The reason may be that PCovR suffers from overfitting when the predictor space in relation to the number of observations available is too large (see, e.g., Heij et al., 2007, and Umbach, 2020).

ferred, especially since the number of observations in macroeconomic forecasting is limited. Using a rolling estimation window of 10 years, only 120 observations are available for parameter estimation at each time step. If the capacity of the networks becomes too great, the model might fail to learn informative latent factors as argued in Section 3.3.4. Second, prefer *depth* over *width*. Theoretically, a neural network with just a single hidden layer and an appropriate nonlinearity can realize arbitrary mappings as long as the number of hidden units (the width) is large enough (Hornik et al., 1989). Nevertheless, more hidden layers are preferred over additional units in a single hidden layer since the number of paths through the network grows exponentially with the number of hidden layers whereas the number of parameters only grows linearly. Thus the flexibility becomes much higher without having to overly increase the number of parameters.

Taking these considerations into account six different network topologies are evaluated: The encoder and decoder networks are allowed to have up to 3 hidden layers. Furthermore, the number of hidden units per layer is set to 8 or 32 such that the minimum number of hidden units is equal to the maximum number of latent factors that can be chosen by BIC. The maximum number of hidden units equals 1/4 of the total number of predictor variables in the dataset. Out of these six specifications the final model used for forecasting is selected by a cross-validation procedure that is explained in the following subsection. Furthermore, note that only symmetric variational autoencoders are considered, i.e., both the encoding and the decoding network have the same architecture.

The supervised deep factor model requires the choice of the supervision parameter α as in equation (3.25). To make the choice scale independent of the number of predictor variables x_t, the original *unsupervised* part of the ELBO is divided by the squared Frobenius norm of the regressors, and the additional *supervising* term of the ELBO is divided by the squared Frobenius norm of the target variable, respectively, in the estimation process. Given these normalizations, α is set to 0.5 which constitutes equal weighting of both objectives of approximating the predictor variables and fitting the target variable. Note that the optimal choice of α requires further investigation. Setting $\alpha = 0.5$ is a rather heuristic approach. Using some cross-validation procedure for α may further enhance the forecasting accuracy of the deep factor model, but is associated with additional computational effort.

Training, Validation, and Testing

Generally, when training neural networks, the estimation sample is split into a training and a validation set. The former is used to estimate the model parameters subject to a specific set of hyperparameter values, for instance, the network architecture. The latter is used for tuning the hyperparameters. As the estimation sample is rather small in this forecasting exercise, the model architecture is determined by cross-validation instead of splitting the estimation sample in a training and validation set once. Using 5-fold cross-valdiation, the estimation sample of 120 observations is randomly divided into five equally sized sets. One set is used as the validation sample whereas the remaining four constitute the training sample. This split ratio of 80:20 between training and validation set is a standard choice for fitting neural networks. To select the final forecasting model, the fit of the different model architectures is calculated on each of the five validation samples and the model that performed best on average is chosen to produce the out-of-sample forecast.

Optimization Algorithm

The model parameters are determined by minibatch stochastic gradient descent (SGD). Unlike standard gradient descent that uses the entire training sample at each iteration of the optimization routine, minibatch SGD evaluates the gradient at a small random subsample of the data at each iteration. This approximation reduces the computational costs and lowers the risk for the optimizer to get stuck in a local minimum. In this study a batch size of 24 observations is employed.[12]

For the optimizer, the AdamW algorithm of Loshchilov and Hutter (2019) is used which is a refined version of the Adaptive Moment estimation algorithm (Adam) introduced by Kingma and Ba (2014). The Adam algorithm computes adaptive learning rates for individual parameters using estimates of first and second moments of the gradient.[13]

Regarding the hyperparameters for the optimization process, recommendations from the literature are used. Both a default learning rate of 0.001 as proposed by Kingma and Ba (2014), and the optimizer's default weight decay coefficient of 0.01 are chosen. Furthermore, a dropout probability of 10% is ap-

[12] Experiments indicate that the results do not critically depend on the batch size in this forecasting application.

[13] Loshchilov and Hutter (2019) demonstrate that L2 regularization of the model parameters is not effective in Adam, and propose an adaptive moment algorithm with decoupled weight decay (AdamW) which they state to generalize substantially better than the original Adam algorithm.

plied to the hidden nodes. Dropout evokes effective regularization by randomly dropping out nodes of the network during the training process. Srivastava et al. (2014) state that the choice of the dropout probability should be coupled with the choice of the number of hidden units. A higher dropout ratio is appropriate for a larger number of hidden nodes. As the neural networks employed in this study are rather compact, a small dropout probability of 10% is used.

The Monte Carlo estimator for the expectations operator in equation (3.23) requires the choice of L, i.e., the number of samples drawn per datapoint to approximate the expectations term. Kingma and Welling (2013) found in their experiments that L can be set to 1 as long as the minibatch size is large enough (e.g., 100). As in the given forecasting exercise a batch only consists of 24 observations, L is set to 5 to meet a *simulated* batch of more than 100 observations.[14]

Finally, note that the stochastic nature of the optimizer and its random initialization of the neural network parameters can cause the optimizer to settle at different optima. To enhance the stability of the results, multiple random seeds are used to initialize estimation. For the final forecast, the mean value of the resulting predictions is taken. More precisely, the model is estimated 10 times per cross-validation fold.[15] As there are 5 folds, the mean value from 5×10 estimates of the selected model is taken to produce the final forecast.

3.4.3 Results

Table 3.1 reports the performance of the deep factor model in forecasting Industrial Production, Nonfarm Employment, Real Manufacturing and Trade Industries Sales and Real Personal Income ex Transfer Receipts h months ahead. The forecasting accuracy of the linear factor model serves as a benchmark and its MSFE is normalized to one for each forecast horizon. The results for the deep factor models are reported relatively to the respective benchmark with a value below one indicating a lower MSFE than the benchmark principal components forecast. The results in Table 3.1 show the forecasting performance of the deep factor model with the factors nonlinearly determined as described in Section 3.4.2. For the relationship between the factors and the forecasting target linearity is retained as stated in equation (3.27). The forecasting accuracy of the deep factor model is reported for two different degrees of supervision: $\alpha = 0$ (no

[14] The choice of L does not seem to be critical for the model performance. However, computation time increases considerably the larger L.

[15] The restriction to 10 repetitions is in line with Gu et al. (2020). Moreover, recall that the forecasting performance is measured over a sample of 420-h observations, which provides a further hedge against *lucky* results.

Table 3.1 Out-of-sample forecasting performance 1985-2019

	$h = 1$	$h = 2$	$h = 3$	$h = 6$
	Industrial Production			
deep factors, $\alpha = .0$	0.9281**	0.97712	1.0109	1.0499
deep factors, $\alpha = .5$	0.9018***	0.9312**	0.9426	0.9542
	Nonfarm Employment			
deep factors, $\alpha = .0$	1.0596	1.0991*	1.0908	1.1321*
deep factors, $\alpha = .5$	0.9708	0.8810***	0.8742***	0.8546**
	Real Manufacturing and Trade Industries Sales			
deep factors, $\alpha = .0$	0.9193**	0.9717*	0.9865	0.9997
deep factors, $\alpha = .5$	0.8760***	0.9352***	0.9755	0.9575
	Real Personal Income ex Transfer Receipts			
deep factors, $\alpha = .0$	0.9773*	0.9729	0.9863	0.9496
deep factors, $\alpha = .5$	0.9723*	0.9492**	0.9607*	0.8989**

Notes: The columns report the relative MSFE of the deep factor model with different degrees of supervision. Pseudo out-of-sample forecasts are made h-months-ahead over the period 1985-01 to 2019-12. Values below 1.00 indicate improved forecast accuracy relative to the linear factor model with the benchmark MSFE being normalized to 1.00 for each forecasting target and time horizon. One (two, three) stars indicate .10 (.05, .01) statistical significance for the Diebold and Mariano (1995) test with HAC standard errors using the Bartlett kernel and a bandwidth of $\lfloor T^{(1/3)} \rfloor = 7$.

supervision) and $\alpha = 0.5$ (intermediate supervision).

Overall, the results indicate that the forecast accuracy of factor models can be improved by using deep factors. This hold particularly true when the factors are supervised ($\alpha = 0.5$). For the case of unsupervised deep factors, $\alpha = 0$, the results are mixed, but the deep factor model benefits considerably from the supervision extension proposed in Section 3.3.3. The empirical results in Table 3.1 indicate that this extension is necessary to adapt the VAE framework for forecasting purposes. Consequently the following analysis focuses on the supervised deep factor model.

The supervised deep factors improve upon the linear factor model by up to 10-14% depending on the target series. For all forecasting targets and horizon, the supervised deep factor model yields more precise forecasts than the linear factor model. In case of Industrial Production, significant accuracy gains are achieved up to a forecasting horizon of 2 months. Beyond this time span, no improvements at a significance level below 0.10 can be observed. Taking into consideration that Industrial Production is a volatile and noisy time series, it might be not surprising that improvements are only found in the short-run.

Forecast beyond one quarter are generally not very informative for this series. A similar pattern can be observed for Real Manufacturing and Trade Industries Sales. Significant improvements in forecasting accuracy are observed only over short forecasting horizons.

Interestingly, for Nonfarm Employment and Real Personal Income ex Transfer Receipts, the gains from using the supervised deep factor model are mainly present for forecasts beyond one month ahead. Especially over a forecast horizon of six months ahead, the deep factor model clearly outperfoms the linear factor model. Both variables react with a greater delay to changes in the economy than Industrial Production and Manufacturing and Trade Industries Sales such that forecasts beyond one quarter are potentially still informative, which is better exploited by the deep factor model than by its linear counterpart.

Table 3.2 provides some additional summary statistics for the comparison of the linear factor model and the supervised deep factor model. It reports the coefficient on the supervised deep factor forecast from the forecast combining regression,

$$y_{t+h} = \gamma \hat{y}^{(d)}_{t+h|t} + (1-\gamma)\, \hat{y}^{(l)}_{t+h|t} + u_{t+h}, \tag{3.28}$$

where $\hat{y}^{(d)}_{t+h|t}$ is the supervised deep factor forecast and $\hat{y}^{(l)}_{t+h|t}$ is the benchmark forecast from the linear factor model. Note that γ also is the coefficient of the encompassing regression which follows immediately from equation (3.28) as

$$\hat{e}^{(l)}_{t+h|t} = \gamma \left(\hat{e}^{(l)}_{t+h|t} - \hat{e}^{(d)}_{t+h|t} \right) + u_{t+h},$$

with $\hat{e}^{(l)}_{t+h|t}$ and $\hat{e}^{(f)}_{t+h|t}$ denoting the forecast errors of the linear factor models and the deep factor model, respectively. Additionally, heteroscedastic and autocorrelation robust standard errors are reported. Consider, for example, the entries for the 3-months-ahead forecast of Industrial Production: γ is estimated to be 0.83 with a standard error of 0.29. Thus, the hypothesis that the weight on the supervised deep factor forecast is 0 ($\gamma = 0$) is rejected at the 5% level, but the hypothesis that the deep factor forecast receives unit weight cannot be rejected.

It is apparent that in some cases the weight γ exceeds unity. As mentioned by Claeskens et al. (2016) this phenomenon can happen in scenarios of a high positive forecast error correlation together with a relatively high variation in forecast reliability. Indeed, the last column of Table 3.2 shows a high empirical cross-correlation coefficients between both forecast error series. Against the backdrop that the target series are very noisy, and that both forecasting models use the same information and have a similar structure, it is not surprising that

Table 3.2 Summary statistics

Statistic	forecast horizon	combination weight γ	std. error	cross-correlation
Industrial Production	$h = 1$	1.15	0.20	0.96
	$h = 2$	0.97	0.20	0.96
	$h = 3$	0.83	0.29	0.96
	$h = 6$	0.85	0.39	0.97
Nonfarm Employment	$h = 1$	0.61	0.14	0.93
	$h = 2$	0.90	0.13	0.92
	$h = 3$	0.99	0.17	0.94
	$h = 6$	1.24	0.30	0.95
Real Manufacturing and Trade Industrie Sales	$h = 1$	1.56	0.25	0.97
	$h = 2$	1.17	0.22	0.98
	$h = 3$	0.70	0.27	0.97
	$h = 6$	0.81	0.35	0.97
Real Personal Income ex Transfer Receipts	$h = 1$	0.95	0.35	0.98
	$h = 2$	1.20	0.35	0.98
	$h = 3$	0.96	0.36	0.98
	$h = 6$	1.33	0.34	0.97

Notes: γ denotes the coefficient on the supervised deep factor forecast from the forecast combining regression (3.28). Column 4 gives the corresponding HAC standard errors using the Bartlett kernel and a bandwidth of $\lfloor T^{(1/3)} \rfloor = 7$. The last column reports the cross-correlation between the forecast error series of the linear factor model and the supervised deep factor model.

the forecasting error series are highly correlated.

Sometimes improvements in forecasting accuracy can be attributed to a good performance during a particular period of time only. Table 3.3 reports the forecasting performance of the supervised deep factor model in relation to the linear factor model over two subsamples. The first periods covers the "Great Moderation"' (1985-1 to 2006-12). The second period ranges from 2007-01 to 2019-12 and includes the financial crisis. In both subsamples the relative MSFE is below 1 for almost all forecasting horizons and forecasting targets which indicates a superior forecasting accuracy of the deep factor model. For the first period the improvement is significant at the 5% level in 11 of 16 cases, for the second in 8 of 16 cases. Overall there is no clear indication that the superiority of deep factor model differs between both periods.

To see how the relative forecasting accuracy evolves over time, Figure 3.2 plots the Giacomini and Rossi (2010) fluctuation test statistic, that is obtained as the standardized difference between the MSFE of the linear factor model and the MSFE of the supervised deep factor model calculated over a 10-years-rolling window. While the accuracy improvements of the deep factor model are signif-

Table 3.3 Subset results

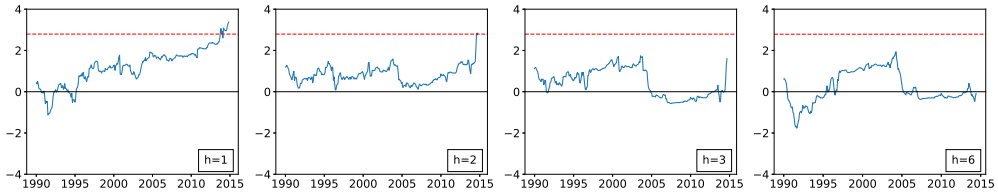
Sample		1985 – 2006		2007 – 2019	
Statistic	forecast horizon	rel. MSFE	DM statistic	rel. MSFE	DM statistic
Industrial Production	$h = 1$	0.96	1.14	0.85	2.53
	$h = 2$	0.94	1.69	0.92	1.57
	$h = 3$	0.91	1.84	0.97	0.29
	$h = 6$	0.91	1.36	0.98	0.24
Nonfarm Employment	$h = 1$	1.01	-0.16	0.90	1.67
	$h = 2$	0.91	2.05	0.84	1.97
	$h = 3$	0.89	1.91	0.85	1.87
	$h = 6$	0.93	1.08	0.77	2.14
Real Manufacturing and Trade Industrie Sales	$h = 1$	0.88	2.96	0.88	1.94
	$h = 2$	0.92	2.72	0.96	0.92
	$h = 3$	0.93	1.50	1.03	-0.69
	$h = 6$	0.92	1.53	0.99	0.17
Real Personal Income ex Transfer Receipts	$h = 1$	0.97	1.20	0.98	0.47
	$h = 2$	0.94	2.12	0.96	1.05
	$h = 3$	0.95	1.32	0.97	0.69
	$h = 6$	0.98	0.45	0.82	3.11

Notes: "rel. MSFE" denotes the relative root mean squared forecast error of the deep factor model with supervision parameter $\alpha = 0.5$ against the linear factor model. Values below 1.00 indicate a higher forecast accuracy of the deep factor model. Forecast from the deep factor model are obtained according to equation (3.27). "DM statistic" refers to the results of the Diebold and Mariano (1995) test with HAC standard errors using the Bartlett kernel and a bandwidth of $\lfloor T^{(1/3)} \rfloor$, where T denotes the sample size.

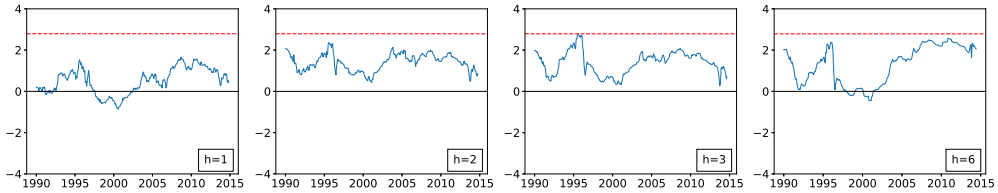
icant over the entire forecasting period in most cases as reported in Table 3.1, Figure 3.2 shows that significance is not guaranteed over rolling subperiods according to the Giacomini and Rossi (2010) test. Nevertheless Figure 3.2 indicates that the superior performance of the deep factor model cannot be attributed to a single time period or event. Although the relative forecasting performance fluctuates over time, the deep factor model seems to be the preferred forecasting tool in most periods.

3.5 Conclusion

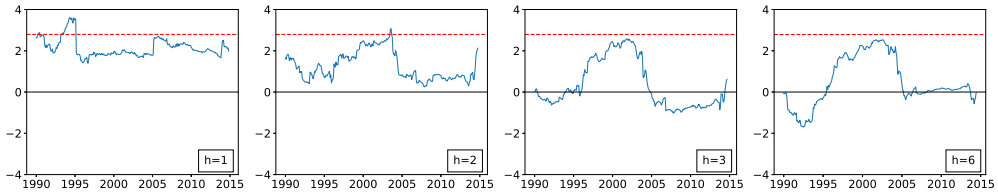
This study shows that variational autoencoders, which can be understood as a nonlinear extension of the factor model (*deep factor models*), have the potential to provide improved forecasts in a data-rich environment. Both approaches aim at estimating a low-dimensional representation of the observed data and share a common core, that is outlined in this study. They differ, however, with respect to the estimation approach and the ability to identify the latent factor space.



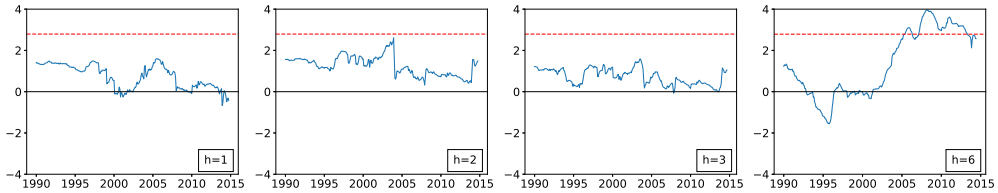
(a) Industrial Production



(b) Nonfarm Employment



(c) Real Manufacturing and Trade Industries Sales



(d) Real Personal Income ex Transfer Receipts

Notes: The figure shows the fluctuation test statistic and the 5% critical value (red dashed line) of the Giacomini and Rossi (2010) fluctuation test. The test statistic is obtained as the standardized difference between the MSFE of the linear factor model and the MSFE of the supervised deep factor model calculated over a 10-years-rolling window. The test statistic is calculated with HAC standard errors using the Bartlett kernel and a bandwidth of $\lfloor T^{(1/3)} \rfloor = 4$. Values above 0 indicate a better forecast accuracy of the supervised deep factor model.

Fig. 3.2 Fluctuation test statistic.

An adjustment is proposed that adapts the VAE framework for the forecasting exercise. The empirical application on four major US macroeconomic time series reveals the potential of variational autoencoders to significantly improve the forecasting accuracy over the famous factor model approach of Stock and Watson (2002b). The architecture of the neural networks applied within the VAE model is rather elementary in this study. The impact of exploiting more sophisticated network structures, that allow, e.g., for the inclusion of time lags of the predictor

variables, remains subject to further investigation. Moreover, the choice of the supervision parameter can be refined to enhance the forecasting capability of the deep factor model.

Natural extensions are the inclusion of autoregressive lags of the forecasting target and a dynamic factor structure within the variational autoencoder framework. From an empirical perspective, the former is, for instance, expected to be relevant for inflation forecasting. Another interesting objective for further research is how the deep factor model can be reframed to provide some useful information about the latent factor space itself, which concerns questions regarding the interpretability and the identification of the factor estimates.

Chapter 4

Improving the Diebold & Mariano Test under Forecast Rationality

4.1 Abstract

One of the most popular statistics to compare the predictive accuracy of two competing forecasts is the Diebold and Mariano (DM) test. In this study, it is suggested to decompose the mean squared error loss differential such that a simplified and more powerful variant of the test statistic can be derived under the assumption of rational forecasts. When comparing forecasts from estimated models, the estimation error uncertainty generally has to be taken into account. To prevent size distortions that can occur when the number of forecasts in relation to the number of estimation sample observations is relatively large, a simple-to-use adjustment to account for parameter estimation uncertainty is proposed. This adjustment remains valid under a fixed estimation scheme and shows good results for the rolling and recursive scheme as well. Furthermore, the applicability of the adjusted test statistic in a nested forecast comparison is discussed. Despite the nonstandard limiting distribution in the case of nested forecasts, simulation evidence suggests that the use of standard normal critical values yields actual sizes close to nominal size in finite samples.

4.2 Introduction

Forecasting plays a critical role both in economic research and policy-making. Being able to compare the accuracy of competing forecasts of the same outcomes by a formal statistical procedure is important to discriminate between good an bad forecasts. The evaluation of competing forecasts has become an extensive field in the econometric literature. One of the most popular statistics to compare the predictive accuracy of two forecast is the Diebold and Mariano (1995, henceforth DM) test that considers the null hypothesis of zero mean in a series defined as the loss differential, i.e., the difference between the two forecast er-

rors' loss functions. In empirical applications the most prominent measure for the accuracy difference of two forecast is the mean squared error (MSE) loss differential. In this study the MSE loss differential is decomposed and it is suggested to exploit a *rationality adjusted* loss differential in the framework of the DM test. Under the assumption of rational forecasts a simplified variant of the DM test for the null hypothesis of equal MSE is derived and it is shown that the power can be considerably improved by this adjustment.

The key assumption for the power improvement of the adjusted DM test statistic is forecast rationality. Imposing forecast rationality seems to be an appropriate assumption. At first sight it is not reasonable why a forecaster whose objective is to minimize the MSE, should stick to a nonrational (i.e., biased and/or inefficient) forecast, when there is room for presumably easy improvement. Nevertheless, there is some evidence that analysts are not always rational in their forecasts. The effect of rationality violations on the adjusted DM test is briefly discussed.

The original DM test only exploits the forecast errors, and makes assumptions directly on the forecast error loss differential. When the forecasts stem from estimated statistical models, nested or non-nested, the impact of the parameter estimation uncertainty on the distribution of the DM test statistic generally has to be taken into account. West (1996) derives how parameter estimation error may affect the limiting distribution of the DM test statistic. Ignoring the effect of parameter estimation uncertainty can result in non-negligible size distortions when the number of forecasts in relation to the number of observations used for parameter estimation is not sufficiently small. For the rationality adjusted DM test the effect of parameter estimation error is analyzed and a simple-to-implement adjustment is proposed to approximately incorporate the effect of parameter estimation uncertainty on the rational DM test statistic. This adjustment holds under a fixed forecasting scheme and shows decent results for the rolling and recursive scheme as well.

Additionally, the applicability of the rationality adjusted DM test in a nested forecast comparison is discussed. McCracken (2007) derives the non-Gaussian limit distribution of the DM test statistic in a nested forecast environment. Clark and West (2007) argue that after an adjustment of the MSE loss differential standard normal critical values are still reasonably accurate for practical purposes. This adjustment accounts for the additional noise of the nesting model from estimating parameters whose population values are zero under the null. The rational DM test benefits from its adjustment on the MSE loss differential in a

similar fashion as the approach of Clark and West. Simulation results indicate that using standard normal critical values yields nearly accurate results in the context of nested models.

Furthermore one should note that this study compares predictive ability at the population level, i.e., the accuracy of forecasts at unknown population values of the forecasting model parameters (West, 1996). In contrast, Giacomini and White (2006) consider an environment with asymptotically non-vanishing estimation uncertainty. They suggest tests of finite-sample predictive ability that are designed to assess the accuracy of a forecasting method in a (finite) sample of the size at hand.

The remainder of this paper is structured as follows. Section 4.3 introduces the rationality adjusted DM test statistic both in a model-free framework and in an environment where the forecast are built on either nested or non-nested estimated statistical models. Furthermore, the adjustment for incorporating the effect of parameter estimation uncertainty is derived. The section concludes with a discussion of the rationality assumption and the effect of rationality violations on the adjusted DM test statistic. Section 4.4 provides Monte Carlo evidence of the small sample properties of the rationality adjusted DM test for survey and model predictions. Additionally the case of nested forecasts is examined. Section 4.5 concludes and provides some directions for further research.

4.3 The Diebold & Mariano Test under Rational Forecasts

Diebold and Mariano (1995) base their test for comparing predictive accuracy on forecast errors only, and make assumptions directly on the forecast error loss differential. The potential effect of parameter estimation errors of underlying statistical models that may have generated the forecasts is not taken into account. As Diebold (2015) reaffirms, "the DM test was intended for comparing forecasts [...]. The DM test was not intended for comparing models". Hence, in its original form, the DM test is designed for a model-free environment, e.g., for comparing forecasts that stem from survey data or expert judgment.

West (1996) extends the DM test to account for parameter estimation error. Using the Diebold-Mariano-West framework, the original question can be recast as whether the forecast error loss differential can be used to learn something about the accuracy of the competing forecasts when the true values of the model

parameters were known.[1] Despite its original purpose and the extension of West to include the effect of parameter estimation error, the DM test in its original form has commonly been used for comparing models in pseudo-out-of-sample forecasting exercises.

This section introduces the proposed adjustment of the DM test first in the original framework to compare forecasts that are not based on (estimated) statistical models. The effect of parameter estimation error on the adjusted test statistic is analyzed subsequently.

4.3.1 Comparing Rational Forecasts in a Model-Free Environment

The objective is to compare the accuracy of forecasts of some univariate time-series $\{Y_t\}$ that is generated by a stationary and ergodic stochastic process. The two competing h-step ahead forecasts of Y_{t+h} based on information up to time period t are denoted by $\hat{Y}_{i,t+h|t}$ for $i = 1, 2$. The corresponding forecast errors are obtained as $e_{i,t+h|t} = Y_{t+h} - \hat{Y}_{i,t+h|t}$ for $i = 1, 2$.

As recapitulated by Patton and Timmermann (2007), a rational forecast $\hat{Y}^*_{t+h|t}$ is defined as the value of $\hat{Y}_{t+h|t}$ that, conditional on the information set at time t, minimizes the expected loss

$$\hat{Y}^*_{t+h|t} \equiv \underset{\hat{Y}_{t+h|t}}{\arg\min}\, \mathbb{E}\left[L\left(Y_{t+h}, \hat{Y}_{t+h|t}\right)\right],$$

Provided that the competing forecasts are covariance stationary, it follows for the rational forecast errors under the mean squared error (MSE) loss function (Diebold and Lopez, 1996):

$$(i) \quad \mathbb{E}(e^*_{1,t+h|t}|\mathcal{I}_{1,t|t}) = 0 \;\text{ and }\; (ii) \quad \mathbb{E}(e^*_{2,t+h|t}|\mathcal{I}_{2,t|t}) = 0, \tag{4.1}$$

where $\mathcal{I}_{1,t}$ and $\mathcal{I}_{2,t}$ denote the relevant information sets associated with time period t. These conditions can be checked by using the Mincer and Zarnowitz (1969) approach for example. In many empirical applications, it is natural to assume that forecasts are rational. If some forecasts violate the rationality condition, this implies that the forecasting method does not fully incorporate the available information and should be revised. It is therefore appealing to impose the rationality condition in order to improve the power of the Diebold-Mariano test.

[1] Note the difference to the Giacomini and White (2006) test for conditional predictive ability which compares the accuracy of forecasts given that these are constructed using estimated parameters.

The MSE loss differential of the Diebold-Mariano test can be decomposed as

$$\begin{aligned} d_t &= e_{1,t+h|t}^2 - e_{2,t+h|t}^2 \\ &= \left(Y_{t+h} - \hat{Y}_{1,t+h|t}\right)^2 - \left(Y_{t+h} - \hat{Y}_{2,t+h|t}\right)^2 \\ &= \hat{Y}_{2,t+h|t} e_{1,t+h|t} - \hat{Y}_{1,t+h|t} e_{2,t+h|t} - \hat{Y}_{1,t+h|t} e_{1,t+h|t} + \hat{Y}_{2,t+h|t} e_{2,t+h|t}. \end{aligned} \tag{4.2}$$

Under rational forecasts (4.1) the latter two terms are zero in expectation no matter whether the null hypothesis is true or not. Since these two terms just add noise to the test statistic, the power of the test can be improved by using the reduced expression

$$d_{r,t} = \hat{Y}_{2,t+h|t} e_{1,t+h|t} - \hat{Y}_{1,t+h|t} e_{2,t+h|t} = \left(e_{1,t+h|t} - e_{2,t+h|t}\right) Y_{t+h}. \tag{4.3}$$

Analogously to the original formulation of Diebold and Mariano (1995), a test statistic for the null hypothesis of equal forecast accuracy, $\mathbb{E}\left[d_{r,t}\right] = 0$, is constructed as

$$t_r = \frac{1}{\hat{\omega}\sqrt{T}} \sum_{t=1}^{T} d_{r,t}, \tag{4.4}$$

which is asymptotically $\mathcal{N}(0,1)$ under the null. $\hat{\omega}^2$ denotes a consistent long-run variance estimator for

$$\omega^2 = \lim_{T\to\infty} \mathbb{E}\left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} d_{r,t}\right)^2.$$

4.3.2 The Rational Diebold & Mariano Test under Parameter Uncertainty

The analysis so far considered forecasts that are not based on (estimated) statistical models. As it has been shown by West (1996), parameter estimation error must in principle be taken into account when the forecasts are generated by estimated models. Although the impact of parameter estimation errors vanishes asymptotically, neglecting them can lead to size distortions in finite samples as the asymptotic variance ω^2 is generally underestimated in this case. To analyze the behavior of the proposed rational DM test under parameter estimation uncertainty, the following framework based on Breitung and Knüppel (2020) is used.

Let $\{y_{t+h}\}_{t=1}^{P}$ denote the sample of P observations of a random variable y_t to be predicted, where h denotes the forecasting horizon. y_t is assumed to be generated by a stationary and ergodic stochastic process $\{Y_t\}$. Let $\hat{y}_{1,t+h|t}$ and $\hat{y}_{1,t+h|t}$

denote the corresponding out-of-sample forecasts based on the relevant information sets $\mathcal{I}_{1,t}$ and $\mathcal{I}_{2,t}$ of two competing models. The forecasts are realizations of the forecast generating processes $\left\{Y_{i,t+h|t}^{(\theta_i)}\right\}$, where θ_i is the parameter vector of forecasting model $i = 1, 2$. The parameters are estimated under the recursive estimation scheme such that the parameter vectors are updated at each forecasting origin $t = 1, \ldots, P$ using the sample $\{-R+1, -R+2, \ldots, t\}$.[2] Hence, $\hat{\theta}_{i,t}$ indicates an estimate based on $t+R$ observations. The parameter estimates of the competing models and the pre-evaluation sample size R are assumed to be unknown. For the analysis, only the observed actual values $\{y_{t+h}\}_{t=1}^{P}$ and their h-step ahead forecasts $\left\{\hat{y}_{i,t+h|t}\right\}_{t=1}^{P}$, for $i = 1, 2$, are considered. For the competing forecast functions $Y_{i,t+h|t}^{(\hat{\theta}_{i,t})}$, the following assumption is made:[3]

Assumption 1. *(i) The parameters are estimated consistently with*

$$a) \quad \hat{\theta}_0 - \theta = O_p\left(R^{-1/2}\right),$$

$$b) \quad \sup_{t\in\{1,\ldots,P\}} ||\hat{\theta}_t - \hat{\theta}_0|| = O_p\left(\frac{\sqrt{t}}{R}\right) \quad \textit{for} \quad t = 1, 2, \ldots, P,$$

where $\hat{\theta}_0$ denotes the estimator based on time periods $\{-R+1, -R+2, \ldots, 0\}$. (ii) Let $D_{t+h}(\theta) = \partial Y_{t+h|t}^{(\theta)}/\partial\theta$ and $\bar{D}_{t+h}(\theta) = P^{-1}\sum_{t=1}^{P} D_{t+h}(\theta)$. For all $\theta_i^* \in [\theta_i - \epsilon, \theta_i + \epsilon]$ with $\epsilon > 0$ it holds that

$$\frac{1}{P}\sum_{t=1}^{P} ||D_{t+h}(\theta^*) - \bar{D}_{h+h}(\theta^*)|| \xrightarrow{p} \bar{D}^2 \quad \textit{with } 0 \leq \bar{D} \leq \infty$$

$$\mathbb{E}||D_{t+h}(\theta^*)Y_{t+h}||^{2+\delta} < \infty \quad \textit{for some } \delta > 0 \textit{ and all } t.$$

Part (i) $a)$ supposes the usual parametric convergence rate of the estimation error in the pre-evaluation sample $\{-R+1, \ldots, 0\}$, whereas part (i) $b)$ limits the variation of the estimators in recursive estimation procedure within the evaluation sample. Part (ii) guarantees the existences of a central limit theorem.

As before the adjusted DM test considers the reduced loss differential derived from equation (4.2)

$$d_{r,t}^{(\hat{\theta})} = \hat{Y}_{2,t+h|t}\hat{e}_{1,t+h|t} - \hat{Y}_{1,t+h|t}\hat{e}_{2,t+h|t}, \tag{4.5}$$

[2] The analysis carries over to a rolling window estimation scheme if one assumes that the window size R gets large relative to the size of the evaluation period.

[3] Note that for notational convenience the superscript indicating the dependence on parameter estimates is henceforth omitted. $\hat{Y}_{i,t+h|t}$ refers to forecast function i based on estimated parameters $\hat{\theta}_i$, and $Y_{i,t+h|t}$ denotes the (hypothetical) forecast function based on the true parameter vector θ_i. The same notational distinction applies to the forecasts error $\hat{e}_{i,t+h|t}$ and $e_{i,t+h|t}$, respectively.

and the test statistic based on a sample of forecasts $\left\{\hat{y}_{1,t+h|t}, \hat{y}_{2,t+h|t}\right\}_{t=1}^{P}$ and observations $\{y_{t+h}\}_{t=1}^{P}$ is calculated by

$$t_r^{(\hat{\theta})} = \frac{1}{\hat{\omega}\sqrt{P}} \sum_{t=1}^{P} d_{r,t}^{(\hat{\theta})}. \tag{4.6}$$

For the adjusted DM test statistic under parameter estimation error, the following proposition holds:

Proposition 1. *Assume that Assumption 1 holds. If* $R \to \infty$, $P \to \infty$, $P/R \to 0$, *then*

$$t_r^{(\hat{\theta})} = \sqrt{P}\frac{\bar{d}_r}{\hat{\omega}} + O_p\left(\frac{P}{R}\right) \quad \xrightarrow{d} \quad \mathcal{N}(0,1),$$

where $\bar{d}_r = \frac{1}{P}\sum_{t=1}^{P}\left(e_{1,t+h|t} - e_{2,t+h|t}\right)Y_{t+h}$ *and* $\hat{\omega}^2$ *denotes a consistent long-run variance estimator for* $\omega^2 = \lim_{P\to\infty}\mathbb{E}\left(\frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}\right)^2$.

The proof is relegated to Appendix B.1.

In the analysis, it has been assumed that only the observed values $\{y_{1+h}, \ldots, y_{P+h}\}$ and the competing forecasts $\{\hat{y}_{i,t+h}, \ldots, \hat{y}_{i,P+h}\}$ are available. It has been shown that the parameter estimation error can be ignored when $R \to \infty$, $P \to \infty$ and $P/R \to 0$. Unlike in the case of the original Diebold-Mariano test, the estimation error is asymptotically irrelevant only for $\lim_{P,R\to\infty}\frac{P}{R} = 0$. For the DM test this requirement can be relaxed under certain conditions as shown by West (1996). The most prominent example is when the measure of forecast accuracy has also been used for estimating the model parameters. This arises, for instance, when a quadratic loss function (MSE) is used to evaluate the accuracy of forecasts that stem from two non-nested models estimated by ordinary least squares.

Following West (1996) it is possible to derive the limiting distribution for the case that P/R converges to some constant. This, however, requires to know the estimation samples and complete information about how the forecasting models have been estimated. As these information are frequently not available or simply neglected for reasons of practicality when dealing with complex models, the effect of parameter estimation error is often not taken into account in empirical forecast comparisons. Unfortunately, for typical sample sizes the additional term due to the parameter estimation error has a non-negligible effect on the adjusted DM test statistic and can induce size distortion if the ratio P/R is not sufficiently small.

An (Approximate) Size Correction Strategy

In line with the argument above regarding the implementation of the adjusted DM test in practice, two approximate size corrections are proposed to account for the impact of parameter estimation uncertainty. The objective is to provide a simple-to-use size correction in the spirit of West (1996), that can be applied even when only the observed values of the forecasting target and the competing forecasts are available, i.e., there is no information on the parameter estimates of the competing models and the pre-evaluation sample. The proposed size adjustments are derived under fairly restrictive assumptions but display a decent performance under various settings in the simulation experiments.

Assume the forecast function of model i to be $\hat{Y}_{i,t+h|t} = X'_{i,t}\hat{\beta}_i$, where the vector $X_{i,t}$ contains the predictor variables. To analyze the asymptotic distribution of the adjusted DM test statistic consider the following decomposition

$$\begin{aligned}
\frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}^{(\hat{\theta})} &= \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\left(\hat{Y}_{2,t+h|t}\hat{e}_{1,t+h|t} - \hat{Y}_{1,t+h|t}\hat{e}_{2,t+h|t}\right) \\
&= \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\left(X'_{2,t}\hat{\beta}_2\left(Y_{t+h} - X'_{1,t}\hat{\beta}_1\right) - X'_{1,t}\hat{\beta}_1\left(Y_{t+h} - X'_{2,t}\hat{\beta}_2\right)\right) \\
&= \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\Big(e_{1,t+h|t} + X'_{1,t}\left(\beta_1 - \hat{\beta}_1\right) \\
&\quad - e_{2,t+h|t} - X'_{2,t}\left(\beta_2 - \hat{\beta}_2\right)\Big) Y_{t+1}.
\end{aligned} \tag{4.7}$$

The case of a fixed estimation scheme for the parameter estimates is studied. This allows us to apply a simple approach to account for the estimation uncertainty as will be shown below. When the model parameters are estimated by least squares over a fixed estimation window using the sample $\{-R+1, -R+2, \ldots, 1\}$ one obtains

$$\left(\hat{\beta}_i - \beta_i\right) = \left(\sum_{s=-R+1}^{1-h} X_{i,s}X'_{i,s}\right)^{-1} \sum_{s=-R+1}^{1-h} X_{i,s}e_{i,s+h} \quad \text{for} \quad i = 1, 2,$$

where $e_{i,s+h}$ denotes the residual from the estimation sample of model i.

Consequently, equation (4.7) can be rewritten as

$$\begin{aligned}\frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}^{(\hat{\theta})} &= \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\left(e_{1,t+h|t} - e_{2,t+h|t}\right) Y_{t+h} \\ &- \sqrt{\frac{P}{R}}\left(\frac{1}{P}\sum_{t=1}^{P} X_{1,t}' Y_{t+h}\right)\left(\frac{1}{R}\sum_{s=-R+1}^{1-h} X_{1,s} X_{1,s}'\right)^{-1}\frac{1}{\sqrt{R}}\sum_{s=-R+1}^{1-h} X_{1,s} e_{1,s+h} \\ &+ \sqrt{\frac{P}{R}}\left(\frac{1}{P}\sum_{t=1}^{P} X_{2,t}' Y_{t+h}\right)\left(\frac{1}{R}\sum_{s=-R+1}^{1-h} X_{2,s} X_{2,s}'\right)^{-1}\frac{1}{\sqrt{R}}\sum_{s=-R+1}^{1-h} X_{2,s} e_{2,s+h}.\end{aligned}$$

The first term on the right-hand side represents uncertainty about the adjusted loss differential $d_{r,t}^{(\theta)}$ when the true parameter vectors β_1 and β_2 were known. The latter terms represent uncertainty about the parameter estimates.

So far only algebraic manipulations have been done. In order to derive the asymptotic distribution of the adjusted loss differential under parameter estimation uncertainty, the assumptions in West (1996) are required. Specifically, one has to assume that the sequences $\{(e_{i,s+h}, X_{i,s})'\}$ and $\{(Y_{t+h}, X_{i,t})'\}$ are covariance stationary, mixing, and have bounded fourth moments. With these assumptions in hand one obtains for $P \to \infty$, $R \to \infty$ such that $\lim_{P,R\to\infty} \frac{P}{R} = \pi$

$$\begin{aligned}\frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}^{(\hat{\theta})} &= \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\left(e_{1,t+h|t} - e_{2,t+h|t}\right) Y_{t+h} \\ &\quad - \mathbb{E}\left[X_1' Y_{t+h}\right]\left(\mathbb{E}\left[X_1 X_1'\right]\right)^{-1}\sqrt{\frac{\pi}{R}}\sum_{s=-R+1}^{1-h} X_{1,s} e_{1,s+h} \\ &\quad + \mathbb{E}\left[X_2' Y_{t+h}\right]\left(\mathbb{E}\left[X_2 X_2'\right]\right)^{-1}\sqrt{\frac{\pi}{R}}\sum_{s=-R+1}^{1-h} X_{2,s} e_{2,s+h} + o_p(1) \\ &= \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\left(e_{1,t+h|t} - e_{2,t+h|t}\right) Y_{t+h} \\ &\quad - \beta_1'\sqrt{\frac{\pi}{R}}\sum_{s=-R+1}^{1-h} X_{1,s} e_{1,s+h} + \beta_2'\sqrt{\frac{\pi}{R}}\sum_{s=-R+1}^{1-h} X_{2,s} e_{2,s+h} + o_p(1) \\ &= \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\left(e_{1,t+h|t} - e_{2,t+h|t}\right) Y_{t+h} \\ &\quad + \sqrt{\frac{\pi}{R}}\sum_{s=-R+1}^{1-h}\left(Y_{2,s+h} e_{2,s+h} - Y_{1,s+h} e_{1,s+h}\right) + o_p(1).\end{aligned}$$

Hence,

$$\frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}^{(\hat{\theta})} = \frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}^{(\theta)} + \underbrace{\sqrt{\frac{\pi}{R}}\sum_{s=-R+1}^{1-h}\xi_s + o_p(1)}_{\text{impact of parameter estimation error}},$$

where $\xi_s = Y_{2,s+h}e_{2,s+h} - Y_{1,s+h}e_{1,s+h}$. The second summand accounts for the parameter estimation error and vanishes only for $\lim_{P,R\to\infty}\frac{P}{R} = 0$ as it has already been stated in Proposition 1. When P/R converges to some constant, the impact of parameter estimation uncertainty has to be taken into account when specifying the limiting distribution of $d_{r,t}^{(\hat{\theta})}$.

Letting both P and R tend to infinity, both statistics $P^{-1/2}\sum_{t=1}^{P} d_{r,t}^{(\theta)}$ and $R^{-1/2}\sum_{s=-R+1}^{1-h}\xi_s$ are asymptotically normal and independent with asymptotic variances S_{ff} and S_{hh}, respectively. The asymptotic normality of the sample mean of $\xi_s = Y_{2,s+h}e_{2,s+h} - Y_{1,s+h}e_{1,s+h}$ follows from the assumptions made above. Recall that the forecast function of model i has been assumed to be $Y_{i,t|t+h} = X_{i,t}'\beta_i$. Together with the assumption that the sequence $\{(e_{i,s+h}, X_{i,s})'\}$ is covariance stationary, mixing, and has bounded fourth moments the asymptotic normality of $R^{-1/2}\sum_{s=-R+1}^{1-h}\xi_s$ can be deduced by using a central limit theorem under weak dependence. The asymptotic independence of the samples means of $d_{r,t}^{(\theta)}$ and ξ_s intuitively follows from their mixing properties and the fact that $P^{-1/2}\sum_{t=1}^{P} d_{r,t}^{(\theta)}$ and $R^{-1/2}\sum_{s=-R+1}^{1-h}\xi_s$ are built from non-overlapping samples due to the fixed estimation scheme employed.

Since a linear combination of normal random variables is normally distributed it holds that under the null of equal forecast accuracy and for $\lim_{P,R\to\infty}\frac{P}{R} = \pi$

$$\frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}^{(\hat{\theta})} \xrightarrow{d} \mathcal{N}\left(0, \omega^2\right),$$

where

$$\omega^2 = S_{ff} + \pi S_{hh}$$

with the variance components $S_{ff} = \lim_{P\to\infty}\mathrm{Var}\left(P^{-1/2}\sum_{t=1}^{P} d_{r,t}^{(\theta)}\right)$ and $S_{hh} = \lim_{R\to\infty}\mathrm{Var}\left(R^{-1/2}\sum_{s=-R+1}^{1-h}\xi_s\right)$.

S_{ff} represents the asymptotic variance of the adjusted loss differential given the true parameter values were known. S_{hh} captures the additional uncertainty due to parameter estimation error of the competing models. Ignoring the latter can results in non-negligible size distortions when the ratio π is not sufficiently small. To get an appropriate estimate of S_{hh}, however, requires knowledge about

the estimation error of the competing models. In particular for the framework above, the residuals $\{\hat{e}_{i,s+h}\}_{s=-R+1}^{1-h}$ and predictions $\{\hat{y}_{i,s+h}\}_{s=-R+1}^{1-h}$ from the estimation sample are required to construct an estimate for S_{hh}. When these are not available or simply for the sake of convenience, it is proposed to replace the estimation sample values by their forecast sample counterparts $\{\hat{e}_{i,t+h}, \hat{y}_{i,t+h}\}_{t=1}^{P}$. These values are available to the researcher anyhow, when conducting the test. When the parameter estimates are stable over time, this strategy yields asymptotically valid results.

Summarizing the derivations above: In order to account for the effect of parameter estimation error on the adjusted DM test statistic $t_r^{(\hat{\theta})}$ from equation (4.6) in the case of $\lim_{P,R\to\infty} P/R = \pi$, it is proposed to approximately estimate the asymptotic variance ω^2 by

$$\hat{\omega}^2 = \hat{S}_{ff} + \pi \hat{S}_{hh}, \tag{4.8}$$

where $\hat{S}_{ff}$ and $\hat{S}_{hh}$ denote consistent long-run variance estimators of $d_{r,t}^{(\theta)}$ and ξ_s. When the pre-evaluation sample $\{-R+1, \ldots, 1-h\}$ is unknown it is suggested to exploit values from the evaluation sample for calculating $\hat{S}_{hh}$. A typical estimate for the long-run variance is then the weighted autocovariance estimate, for instance, in the case of S_{hh}

$$\hat{S}_{hh} = \hat{\gamma}_0 + 2 \sum_{j=1}^{P-1} k\left(\frac{j}{M}\right) \hat{\gamma}_j,$$

where $\hat{\gamma}_j = P^{-1} \sum_{t=1}^{P-1} (\hat{\xi}_t - \bar{\xi})(\hat{\xi}_{t+j} - \bar{\xi})$, and k being an appropriate kernel function (e.g. the Bartlett kernel). M denotes the bandwidth parameter.

The adjustment on the long-run variance estimate $\hat{\omega}^2$ for incorporating the effect of parameter estimation uncertainty will simplify even further if one is willing to assume that $Y_{i,s+h} \approx Y_{s+h}$ for $i = 1, 2$. In this case ξ_s can be approximated by

$$\xi_s = e_{2,s+h} Y_{2,s+h} - e_{1,s+h} Y_{1,s+h} \approx \left(e_{1,s+h|t} - e_{2,s+h|t}\right) Y_{s+h}.$$

Using this approximation and only exploiting values from the evaluation sample as proposed above, an approximate estimate of the long-run variance ω^2 results from simply pre-multipliying $\hat{S}_{ff}$ by a constant factor $(1 + \pi)$. Hence, an alternative approximate estimate for ω^2 that takes into account the parameter estimation uncertainty is given by

$$\hat{\omega}^2 = (1 + \pi)\, \hat{S}_{ff}, \tag{4.9}$$

where $\hat{S}_{ff}$ still denotes a long-run variance estimator of $d_{r,t}^{(\theta)}$.

The approximate long-run variance estimates for ω^2 in equations (4.8) and (4.9) that account for the impact of parameter estimation error on the test statistic have been derived under the assumption of a fixed estimation scheme. Under a recursive or a rolling estimation scheme the appropriate adjustment à la West (1996) becomes more demanding, and the estimation of the long-run variance ω^2 cannot be simply approximated by using values from the evaluation sample only. However, it is suggested to employ the estimates from equation (4.8) or (4.9) under these estimation schemes as well if a thorough adjustment in the spirit of West (1996) is either not possible or not intended. The major contribution to the uncertainty due to parameter estimation errors typically stems from $\hat{\theta}_0 - \theta$, i.e., from the pre-evaluation sample part that is not overlapping the evaluation sample. The difference from applying the recursive scheme instead of the fixed scheme, $\hat{\theta}_t - \hat{\theta}_0$, is typically small if R is reasonably large (see Assumption 1). Consequently, the estimates in equations (4.8) and (4.9) may serve as a means to obtain approximate estimates of the long-run variance ω^2 under the recursive estimation scheme as well. Indeed, the simulations in Section 4.4.3 show that the adjusted DM test is reasonably sized under the recursive and rolling estimation scheme when ω^2 is estimated according to equation (4.8) or (4.9).

4.3.3 Comparing Nested Forecasts

It is well known that in particular situations the DM test statistic has a non-standard limiting distribution. The typical discussion of the problem considers forecasts from linear models that are estimated by least squares:

$$\text{Model 1:} \quad Y_{1,t+h} = X_{1,t}'\beta_1 + e_{1,t+h}, \tag{4.10}$$

$$\text{Model 2:} \quad Y_{2,t+h} = X_{1,t}'\delta + Z_t'\gamma + e_{2,t+h} \equiv X_{2,t}'\beta_2 + e_{2,t+h}, \tag{4.11}$$

where $X_{2,t}' = (X_{1,t}', Z_t')'$ and $\beta_2 = (\delta', \gamma')'$. Furthermore, it is assumed that $\mathbb{E}\left[e_{1,t+h}X_{1,t}\right] = 0$ and $\mathbb{E}\left[e_{2,t+h}X_{2,t}\right] = 0$.

Since the first forecast results as a special case of the second forecast when $\gamma = 0$, the forecast $\hat{Y}_{1,t+h|t} = X_{1,t}'\hat{\beta}_1$ is said to be nested within $\hat{Y}_{2,t+h|t} = X_{2,t}'\hat{\beta}_2$. The problem with testing the null hypothesis of equal population-level predictive ability is that the forecasts errors are asymptotically identical under the null causing the asymptotic variance ω^2 of the loss differential to degenerate. However, the results in West (1996) are only applicable if ω^2 is positive, which

is a crucial condition for asymptotic normality of the test statistic. Otherwise

$$\frac{1}{\sqrt{P}} \sum_{t=1}^{P} d_t \xrightarrow{p} 0$$

if ω^2 is zero. McCracken (2007) develops a different set of asymptotics that allows for testing equal one-step-ahead population-level predictive ability between two nested models. He shows that when the number of observations used to generate initial estimates of the models R and the number of forecasts observations P increase at the same rate, the limiting distribution of the DMW test is non-standard and can be represented as functionals of Brownian motions. The asymptotic null distributions under the different estimation schemes depend on the number of excess parameters of the larger model and the ration $\pi = P/R$. The appropriate critical values for various comparison settings are simulated and provided by McCracken. Clark and McCracken (2005) extend the analysis to allow for a comparison of direct multistep-ahead forecasts and show how to construct asymptotic critical values from Monte Carlo simulations of the asymptotic distribution. As an alternative procedure they propose a bootstrap algorithm such that the critical values can be obtained as percentiles of the bootstrapped test statistic. In subsequent work, Clark and McCracken (2012) develop a fixed regressor bootstrap that may be easier to implement from a practitioner's point of view.

A different approach to compare nested forecasts is proposed by Clark and West (2007). They argue that under the null the larger model exhibits additional noise from estimating parameters whose population values are zero. This affects the MSE of the parsimonious model to be smaller than that of the larger model. By a simple adjustment of the MSE loss differential they discard the additional noise from parameter estimation and find that the use of standard normal critical values yields size close to, but a little less than, nominal size. In particular, Clark and West (2007) rewrite the MSE loss differential as

$$\begin{aligned} \frac{1}{P} \sum_{t=1}^{P} d_t &= \frac{1}{P} \sum_{t=1}^{P} \left(\hat{e}^2_{1,t+h|t} - \hat{e}^2_{2,t+h|t} \right) \\ &= -2 \frac{1}{P} \sum_{t=1}^{P} \left(\hat{Y}_{1,t+h|t} - \hat{Y}_{2,t+h|t} \right) \hat{e}_{1,t+h|t} \\ &\quad - \frac{1}{P} \sum_{t=1}^{P} \left(\hat{Y}_{1,t+h|t} - \hat{Y}_{2,t+h|t} \right)^2, \end{aligned} \tag{4.12}$$

and argue that it is reasonable to expect that $P^{-1}\sum_{t=1}^{P}(\hat{Y}_{1,t+h|t}-\hat{Y}_{2,t+h|t})\hat{e}_{1,t+h|t}$ is approximately zero under the null hypothesis of equal forecast accuracy. Since $-P^{-1}\sum_{t=1}^{P}(\hat{Y}_{1,t+h|t}-\hat{Y}_{2,t+h|t})^2<0$, it can be expected that the sample MSE from the parsimonious model is less than that of the alternative model. To put it differently the latter term in equation (4.12) induces noise from parameter estimation and causes a bias in the test statistic. Consequently, Clark and West suggest to discard this term to properly center the statistic such that its expectation will be zero under the null. Clark and West find that the use of conventional standard errors yields an asymptotic approximate normal test statistic that is accurate for practical purposes. Their approach considers the reduced loss differential

$$d_{cw,t}=\left(\hat{Y}_{2,t+h|t}-\hat{Y}_{1,t+h|t}\right)\hat{e}_{1,t+h|t}. \tag{4.13}$$

Note the similarity to Harvey et al. (1998) who propose to test $\mathbb{E}[e_{1,t+h|t}(e_{1,t+h|t}-e_{2,t+h|t})]$ as an implication of encompassing. Clark and West prefer the interpretation of executing a comparison of MSEs after adjusting for the upward bias in the MSE of the larger model. This interpretation requires to distinguish between tests of $\mathbb{E}[e_{1,t+h|t}(e_{1,t+h|t}-e_{2,t+h|t})]$ in nested and non-nested forecasts comparisons.

In the context of nested forecasts the rationality adjusted DM test benefits from the same advantage as the approach of Clark and West (2007). The simulations in Section 4.4.4 show that it remains properly centered under null even when the nesting model suffers from a potentially larger parameter estimation error. Similar to Clark and West, the rationality adjustment affects a centering of the DM test statistic such that is has approximate mean zero under the null.

Indeed, both tests are closely related when the forecast comparison is nested. Recall that the rationality adjusted DM test is based on the sample

$$\frac{1}{P}\sum_{t=1}^{P}d_{r,t}=\frac{1}{P}\sum_{t=1}^{P}\left(\hat{Y}_{2,t+h|t}\hat{e}_{1,t+h|t}-\hat{Y}_{1,t+h|t}\hat{e}_{2,t+h|t}\right), \tag{4.14}$$

whereas Clark and West (2007) consider

$$\frac{1}{P}\sum_{t=1}^{P}d_{cw,t}=\frac{1}{P}\sum_{t=1}^{P}\left(\hat{Y}_{2,t+h|t}-\hat{Y}_{1,t+h|t}\right)\hat{e}_{1,t+h|t} \tag{4.15}$$

to test the null hypothesis of equal MSE. Imposing the assumption of rational forecasts, i.e. $\mathbb{E}[e_{1,t+h|t}Y_{1,t+h|t}]=0$, the expectation of equation (4.15) reduces to $\mathbb{E}\left[Y_{2,t+h|t}e_{1,t+h|t}\right]$. Thus, under rational forecasts the difference between the expectation of the Clark and West and the rational DM statistic is given by

$\mathbb{E}[Y_{1,t+h|t}e_{2,t+h|t}]$. When this term is zero both tests are equal in expectation and test the same hypothesis. This is in general not the case, but holds for the nested scenario of equations (4.10) and (4.11). Under the null $e_{2,t+h|t}$ is uncorrelated with $X_{2,t}$. Since in the nested case $X_{1,t}$ is a strict subset of the elements of $X_{2,t}$, it follows that $\mathbb{E}[X_{1,t}e_{2,t+h|t}] = 0$ and, hence $\mathbb{E}[Y_{1,t+h|t}e_{2,t+h|t}] = 0$.

As argued above, only in a nested forecast comparison both approaches test the same hypothesis. The adjustment of Clark and West essentially results in an encompassing test whereas the rationality adjusted DM test is based on the null hypothesis of equal forecasting accuracy. Clark and West distinguish with their approach between tests of $\mathbb{E}[e_{1,t+h|t}(e_{1,t+h|t} - e_{2,t+h|t})]$ for nested and non-nested models. Only when the models are nested their approach can be interpreted as a test of equal MSE after adjusting for the additional noise from parameter estimation of the larger model. In contrast, the rationality adjusted DM test remains (approximately) applicable for both nested and non-nested forecasts.

When the models are nested, a higher power can be achieved by testing for encompassing instead of equal predictive accuracy as the simulations in section 4.4.4 indicate. However, whether the forecasts stem from nested models may not always be as distinct as in equations (4.10) and (4.11). A simple example is when the researcher has only excess to the forecasts and does not know the underlying model structures. In this case, it is not clear whether a test for nested models or the usual DM test statistic is appropriate. When the forecasts stem from non-nested models both approaches have different implications as the former tests for encompassing whereas the latter considers equal predictive accuracy under the null.

Even when the forecasting models are known, the distinction between nested and non-nested can be ambiguous. As a simple example assume that the benchmark forecast $\hat{y}_{1,t+h|t}$ involves a few variables only (e.g. an ARMA model), whereas the alternative model employs a large data set with hundreds of variables, say by using diffusion indices as predictors. For illustration, assume that $\hat{Y}_{1,t+h|t}$ is obtained from a simple dynamic regression model $Y_{1,t+h|t} = \alpha_1 Y_t + \beta_1 X_t + e_{1,t+h}$. The second prediction is generated by a factor model with $Y_{2,t+h|t} = \alpha_2 Y_t + \beta_2 f_t + e_{2,t+h}$, where f_t is a single factor $f_t = \gamma' Z_t$ and the vector X_t is included in Z_t. Is this a nested or a non-nested forecast comparison? Strictly speaking, the small model results as a special case from setting the loadings of all other variables except X_t in the factor model equal to zero. Of course this will never happen in practice and, therefore, the difference between the forecasts $Y_{1,t+h|t}$ and $Y_{2,t+h|t}$ does not tend to zero unless the coefficients β_1

and β_2 are zero such that the standard DM test would usually be applied. The crucial point is that it does not really matter whether the two models are nested or not but whether the forecasts are asymptotically identical. One possibility is that one model results as a special case of some more general model but this is neither sufficient nor necessary. Whether the forecasts are asymptotically identical depends on the fact, first, whether there exist a region in the parameter space where both forecasts are identical and, second, whether the actual parameters are an element of this region.

A conclusion from the discussion is that it cannot always be decided whether the forecast comparison is nested or non-nested. When this distinction is ambiguous, the decision whether to use the framework of Clark and West (2007) or the usual DM approach to test for equal MSE matters. Due to its encompassing type the Clark and West test can be employed to test the null of equal MSE only in an environment where the forecasts are clearly nested. In contrast, the rationality adjusted DM test accounts for the additional noise of the larger model in a similarly fashion as the Clark and West approach but still remains a test of equal predictive ability that is both valid in non-nested and approximately applicable in nested forecast comparisons.

4.3.4 A Brief Discussion of Forecast Rationality

The considerable power improvements of the rationality adjusted DM test (shown in the subsequent simulation studies) are achieved by discarding the last two terms of the decomposed MSE loss differential in equation (4.2). Under rational forecasts, these terms are zero in expectation and just add noise to the test statistic. Imposing rationality seems to be an appealing assumption. At first sight it is not reasonable why a forecaster whose objective is to minimize the MSE should stick to a nonrational, inefficient forecast, when there is room for presumably easy improvement. Nevertheless, there is some evidence that analysts are not always rational in their forecasts. The literature does not allow to make a clear statement whether (survey) forecasts generally are rational or not. It presents mixed results depending on the subsamples under consideration and the testing strategies employed. Recent studies in this field include, e.g., Croushore (2010), Jonsson and Österholm (2012), Patton and Timmermann (2012) and Romer and Romer (2000). Related studies such as, for instance, Croushore (2012) highlight that while forecasts may appear rational over the whole sample they may not be rational during certain periods of time. Following this, Rossi and Sekhposyan (2016) suggest a test for forecast rationality in the presence of instabilities.

Another discussion in the literature on forecast rationality centers around the question whether biases in forecasts should be dedicated to irrationality or to an asymmetric loss function employed in the forecast generating process (see, e.g., Elliott et al. (2008)). If the latter is the main driving force, the forecaster's objective is not minimizing the MSE. Consequently, it is questionable if the MSE loss differential is the appropriate measure for the comparison of competing forecast. Nevertheless, it remains the most prominent measurement in empirical applications and is commonly used to evaluate the DM test statistic.

4.4 Monte Carlo Evidence

To compare the small sample properties of the adjusted DM test in different forecasting scenarios, a number of Monte Carlo experiments are conducted. The performance of the adjusted DM test is analyzed both in a framework where forecasts are taken as (statistical) model-free (Section 4.4.2), and in a setting where parameter estimation error uncertainty from estimated statistical models is present (4.4.3). Furthermore, the behavior of the adjusted DM test in a nested forecast comparison is analyzed (4.4.4).

4.4.1 Long-run Variance Estimation

Before proceeding to the Monte Carlo results, a note on the long-run variance estimators employed in the simulation experiments seems to be appropriate. Computation of the (adjusted) DM test statistic requires an estimate for the long-run variance of the (adjusted) loss differential. A typical estimate for the long-run variance is the weighted autocovariances estimate

$$\hat{\omega}^2 = \hat{\gamma}_0 + 2 \sum_{j=1}^{T-1} k\left(\frac{j}{M}\right) \hat{\gamma}_j, \tag{4.16}$$

where $\hat{\gamma}_j = T^{-1} \sum_{t=1}^{T-1} (\hat{d}_t - \bar{d})(\hat{d}_{t+j} - \bar{d})$, and the kernel function k being determined according to the Bartlett or the uniform kernel. M is the bandwidth parameter that is often set to $M = T^{1/3}$ or $M = h - 1$, resp. The former is based on the finding of Newey and West (1994) that the optimal bandwidth, in minimal MSE sense, is proportional to $M = T^{1/3}$. The latter is motivated by the argument that optimal h-step-ahead forecasts are at most $(h-1)$-dependent (Diebold and Mariano, 1995).

It is well-known that estimation of the long-run variance can be a challenging task that becomes more difficult as the forecast horizon and, thus, the order

of auto-correlation increases, and the forecast sample size declines.[4] Instead of following the usual route in estimating ω^2 via equation (4.16), the fixed-m estimator of Coroneo and Iacone (2020) is used in some simulation settings. Coroneo and Iacone (2020) report better small sample size properties of the fixed-m estimator in case of autocorrelated forecast errors - a finding that is supported by Harvey et al. (2017) who evaluate the finite size and power of different approaches, and also corroborated by the simulation experiments in this study. Furthermore, this approach is guaranteed to provide positive long-run variance estimates even in small samples.

The fixed-m estimator of Coroneo & Iacone (2020) is based on a weighted periodogram estimate. Let

$$I(\lambda_j) = \left| \frac{1}{\sqrt{2\pi T}} \sum_{t=1}^{T} d_t e^{-i\lambda_j t} \right|^2$$

denote the periodogram of d_t and $\lambda_j = 2\pi j/T$ for $j = 0, \pm 1, \ldots, \pm \lfloor T/2 \rfloor$ being the Fourier frequencies.[5] Coroneo and Iacone (2020) construct an estimator of the long-run variance by using the Daniell kernel:

$$\hat{\omega}^2 = 2\pi \frac{1}{m} \sum_{j=1}^{m} I(\lambda_j), \tag{4.17}$$

where m is a function of the bandwidth M. As suggested by Coroneo and Iacone (2020) m is set to $\lfloor T^{(1/3)} \rfloor$ which they find to yield the best size-power combination in their simulation results. Note that when using the estimate (4.17) the resulting test statistic is asymptotically t-distributed with $2m$ degrees of freedom.

4.4.2 Survey Forecasts

In this section the small sample properties of the adjusted DM test are analyzed in a model-free environment. The size and power properties of the adjusted DM test are evaluated under a variety of specifications of forecast error contemporaneous and serial correlation. The adjusted DM test statistic in equation (4.4) requires both forecast errors and forecasts themselves which is why the simulation setup has to deviate from the typical experimental design of a model-free environment where forecast errors only are simulated (see, e.g., Harvey et al.,

[4] See, e.g., the extensive summary article of Clark and McCracken (2013). For a review on long-run variance estimation with emphasis on the spectral perspective see Müller (2014).

[5] $\lfloor \cdot \rfloor$ refers to the integer value of a number.

1997).

Consider the following data generating process, in which the forecasting target is simulated according to

$$y_{t+h} = \mu_{c,t} + \mu_{s,t} + u_{t+h}, \tag{4.18}$$

where $\mu_{c,t}$ and $\mu_{s,t}$ represent the impact of exogenous explanatory variables, and $u_t = \theta(L)\nu_t$ with $\nu_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\nu^2)$ and a finite lag order that will be specified below. The component $\mu_{c,t}$ can be interpreted as capturing the (market) information that is commonly available to and exploited by both of the competing forecasters, whereas $\mu_{s,t}$ summarizes all influences on y_{t+h} that are known by either forecaster 1 or forecaster 2.

Let $\mu_{c,t}$ follow an AR(1) process with autoregressive parameter $|\gamma| < 1$

$$\mu_{c,t} = \gamma\mu_{c,t-1} + \eta_t. \tag{4.19}$$

with $\eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2)$, and let $\mu_{s,t}$ be generated from a sum of two independent MA(q) processes, $\mu_{s,t} = \alpha\mu_{s_1,t} + (1-\alpha)\mu_{s_2,t}$, with

$$\mu_{s_1,t} = \theta(L)\epsilon_{1,t}, \quad \text{and} \quad \mu_{s_2,t} = \theta(L)\epsilon_{2,t},$$

where are $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are *iid*-normal with variances $\sigma_{\epsilon_1}^2$ and $\sigma_{\epsilon_2}^2$, respectively. The lag order q will be chosen corresponding to the forecasting horizon h with $q = h - 1$. Note that the same MA(q)-structure for u_t, $\epsilon_{1,t}$, and $\epsilon_{2,t}$ is imposed for simplicity. The choice of the lag order q will give a neat control over the degree of serial correlation of the forecast errors. The competing forecasts use the commonly known information $\mu_{c,t}$ and the individual information represented by $\mu_{s_1,t}$ and $\mu_{s_2,t}$. They are obtained as

$$y_{1,t+h} = \mu_{c,t} + \alpha\mu_{s_1,t}, \quad \text{and} \quad y_{2,t+h} = \mu_{c,t} + (1-\alpha)\mu_{s_2,t}$$

with associated forecast errors

$$e_{1,t+h} = (1-\alpha)\mu_{s_2,t} + u_{t+h}, \quad \text{and} \quad e_{2,t+h} = \alpha\mu_{s_1,t} + u_{t+h}.$$

Hence, the forecast errors consist of a common component u_{t+h} and individual parts that arise from not using the information in $\mu_{s_1,t}$ or $\mu_{s_2,t}$, resp. The choice of the common error variance, σ_u^2, and the individual error variances, $(\sigma_{\epsilon_1}^2, \sigma_{\epsilon_2}^2)$, governs the degree of cross-correlation between the forecast errors, which is given

Table 4.1 Empirical size, survey forecasts

	low serial correlation				high serial correlation			
T	25	50	100	200	25	50	100	200
A. moderate cross-correlation								
1-step-ahead								
MSE-t	4.15	4.44	4.63	4.68	4.15	4.44	4.63	4.68
MSE-t_r	4.36	4.45	4.68	4.54	4.36	4.45	4.68	4.54
3-steps-ahead								
MSE-t	4.63	4.67	5.05	4.78	3.66	4.47	4.56	4.73
MSE-t_r	4.33	4.73	4.78	4.86	3.68	4.36	4.77	4.81
6-steps-ahead								
MSE-t	4.66	4.67	4.93	5.07	4.02	4.24	4.14	4.94
MSE-t_r	4.56	4.46	4.80	4.85	3.75	4.40	4.11	4.66
B. high cross-correlation								
1-step-ahead								
MSE-t	4.21	4.77	4.72	4.92	4.21	4.77	4.72	4.92
MSE-t_r	4.49	4.55	4.80	4.86	4.49	4.55	4.80	4.86
3-steps-ahead								
MSE-t	4.31	4.61	4.74	4.59	3.91	4.17	4.98	4.74
MSE-t_r	3.70	4.41	4.64	4.94	3.93	4.15	4.80	4.93
6-steps-ahead								
MSE-t	4.38	4.42	4.77	4.98	4.31	4.41	4.42	4.79
MSE-t_r	4.69	4.73	4.55	4.81	3.94	4.51	4.56	4.64

Notes: Reported are the empirical rejection rates at a nominal size of 5% from 10,000 Monte Carlo simulations. MSE-t denotes the DM test, and MSE-t_r refers to the proposed version under forecast rationality. T denotes the number of forecast error observations.

by

$$Corr\left(e_{1,t+h}, e_{2,t+h}\right) = \frac{\sigma_\nu^2}{\sqrt{\left((1-\alpha)^2\sigma_{\epsilon_2}^2 + \sigma_\nu^2\right)\left(\alpha^2\sigma_{\epsilon_1}^2 + \sigma_\nu^2\right)}}.$$

Depending on the specifications of the MA(q)-processes for $\mu_{1,t}$, $\mu_{2,t}$, and u_t, one obtains different degrees of forecast error autocorrelation. The k^{th}-order autocorrelation of forecast error $e_{i,t+h}$ is given by

$$\rho_k = \frac{\theta_k + \sum_{j=1}^{q-k}\theta_j\theta_{j+k}}{1 + \sum_{j=1}^{q}\theta_j^2}$$

for $k \leq q$ and $\rho_k = 0$ for $k > q$.

Empirical Size

The empirical size properties are evaluated under different degrees of contemporaneous and serial forecast error correlation. As in Harvey et al. (1997) contemporaneous correlations of 0.5 and 0.9 are considered. To mimic h-step

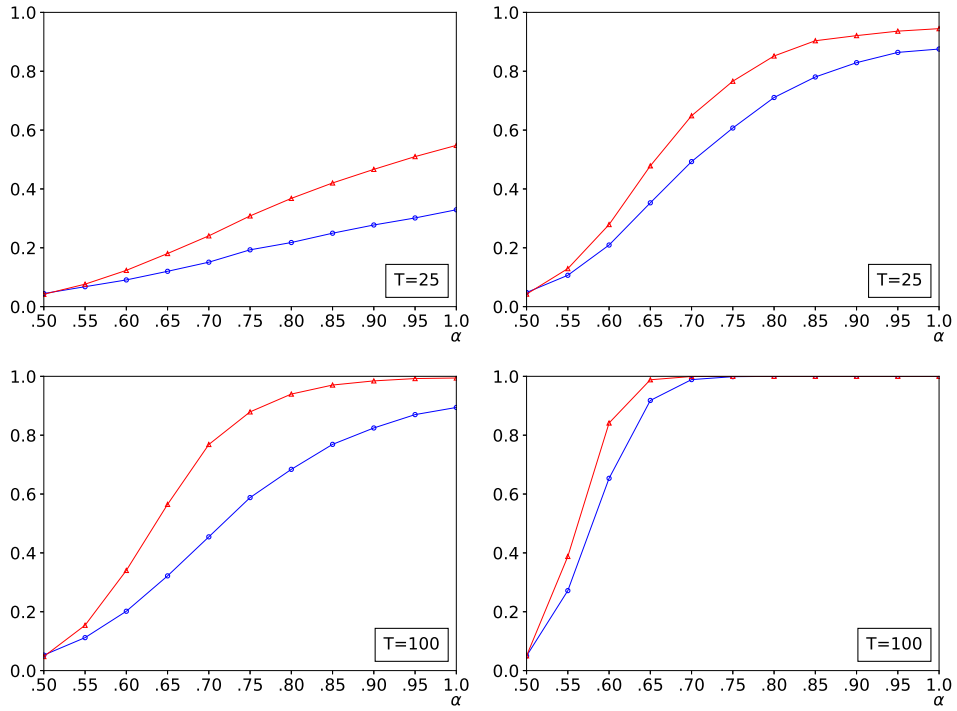
ahead forecasts, $\mu_{s_1,t}$, $\mu_{s_2,t}$, and u_t are generated by either a white noise process for $h = 1$ or an MA(q) process for $h > 1$ with $q = h-1$. The latter configuration represents a scenario of multistep-ahead forecasts that typically exhibit serial error correlation (at least) up to a lag order of $h - 1$.

More specifically, the following parametrization is used: σ_ν^2, determining the common error variance, is set to one. To achieve a degree of forecast error cross-correlation of 0.9 (0.5), $\sigma_{\epsilon_1}^2$ and $\sigma_{\epsilon_2}^2$ are set to $0.\overline{44}$ (4). Regarding the serially correlation of the forecast errors, two different autocorrelation profiles are examined: A high serial correlation scenario where $\theta_k = 0.8$ for $k \leq q$ and $\theta_k = 0$ for $k > q$, and a low serial correlation scenario where $\theta_k = 0.2$ for $k \leq q$ and $\theta_k = 0$ for $k > q$. Note that by setting θ_k equal to a constant for all $k \leq q$ the autocorrelation at each lag increases with the forecasting horizon h, which is considered to be a realistic property. The autocorrelation profiles are visualized in Figure B.1 in Appendix B.2. Furthermore, the autoregressive parameter γ in equation (4.19) is set to 0.9 and the error variance σ_η^2 is set to 0.1. To ensure equal forecast accuracy under the null α equals 0.5.

Table 4.1 shows some empirical size results of the DM test and the adjusted DM test for different degrees of forecast error serial- and cross-correlation. The complete simulation results can be found in Appendix B.2 and are qualitatively similar. Both the DM test and its rationality adjusted variant behave very similarly and are well sized. Even in case of a small number of available observations both tests meet the nominal size of 5%. This may be attributed to the weighted periodogramm long-run variance estimator employed. Many studies report (over)size problems of the DM test in small samples when serial correlation in the forecast errors is present (see e.g. Harvey et al., 2017). These size problems are usually attributed to the difficulty of appropriately estimating the long-run variance in small sample via a weighted autocovariance estimator as described in equation (4.16). Using the weighted periodogramm estimator of Coroneo and Iacone (2020) the empirical size seem to be stable with a tendency to be rather conservative than oversized.

Empirical Power

Figures 4.1 and 4.2 show the empirical power functions for different parameter constellations of the data generating process. A complete summary for all settings regarding the degrees of forecast error serial- and cross-correlations can found in the Appendix B.2. The power of the DM test and the adjusted DM test is plotted against the parameter α governing the distance from the null



(a) high forecast error cross-correlation (b) low forecast error cross-correlation

Notes: Results from 10,000 Monte Carlo simulations of one-step-ahead forecasts. Nominal size = 5%. T refers the number of observations. The blue line denotes the DM test. The red line denotes the adjusted DM test.

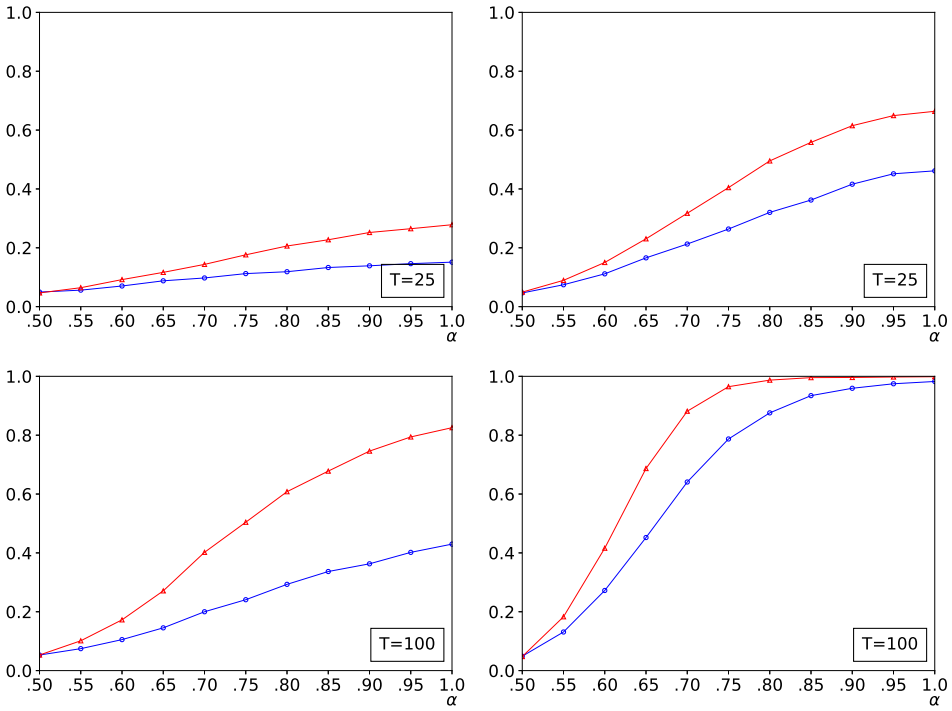
Fig. 4.1 Empirical power functions for the case of one-step-ahead forecasts.

hypothesis.

When α is increased to a range between 0.5 and 1, $y_{1,t+h}$ presents the more precise forecast. The adjusted DM test improves the power under all different settings regarding forecast error cross- and serial-correlation. Especially in the case of high forecast error cross-correlation - a setting where differences in forecast accuracy can be hard to detect - the adjusted DM test shows remarkable power improvements. Its rejection frequency is up to twice as high as under the standard DM test. The superiority of the adjusted DM test shows up both for one-step-ahead and autocorrelated multi-step-ahead forecasts.

4.4.3 Model Predictions

To evaluate the finite sample properties of the adjusted DM test when forecasts stem from estimated statistical models, the Monte Carlo design of Clark and Mccracken (2014) is employed. The data generating process is loosely based on



(a) high forecast error cross-correlation (b) low forecast error cross-correlation

Notes: Results from 10,000 Monte Carlo simulations of six-step-ahead forecasts. Nominal size = 5%. T refers the number of observations. The blue line —○— denotes the DM test. The red line —△— denotes the adjusted DM test.

Fig. 4.2 Empirical power functions for the case of six-step-ahead forecasts with high serial correlation.

the empirical relationship between GDP growth and interest rate spreads, and takes the following form:

$$y_{t+1} = 0.3y_t + b_1 x_{1,t} + b_2 x_{2,t} + u_{t+1}, \tag{4.20}$$

$$x_{i,t} = a x_{i,t-1} + \epsilon_{i,t} \quad \text{for} \quad i = 1, 2. \tag{4.21}$$

The error terms are uncorrelated and $iid-$normally distributed with $\text{Var}\,(u_t) = 10$ and $\text{Var}\,(\epsilon_{i,t}) = 0.1$ for $i = 1, 2$.[6] The competing forecasting models are

$$y_{1,t+1} = \alpha_1 y_t + \alpha_2 x_{1,t} + \alpha_3 x_{1,t-1} + e_{1,t+1}, \tag{4.22}$$

$$y_{2,t+1} = \beta_1 y_t + \beta_2 x_{2,t} + \beta_3 x_{2,t-1} + e_{2,t+1}. \tag{4.23}$$

[6] The initial observations necessitated by the autoregressive structure of the DGP are generated by draws form the unconditional distribution implied by the DGP.

Clark and Mccracken (2014) parametrize equations (4.20) and (4.21) with $a = 0.9$ and consider different values for b_1 and b_2 with $b_1 = b_2$ under the null of equal forecast accuracy.

In this simulation study different values for the persistence parameter a are examined due to the following considerations. First, one should note that y_{t+1} can be represented by an augmented ARMA(2, 1) process. Depending on whether $x_{1,t}$ or $x_{2,t}$ is exploited one can rewrite equation (4.20) as

$$y_{t+1} = (0.3 + a)\, y_t - 0.3a y_{t-1} + b_1 x_{1,t} - a b_1 x_{1,t-1} + u_t + b_2 \epsilon_{2,t} - a u_{t-1} \quad (4.24)$$

or

$$y_{t+1} = (0.3 + a)\, y_t - 0.3a y_{t-1} + b_2 x_{2,t} - a b_2 x_{2,t-1} + u_t + b_1 \epsilon_{1,t} - a u_{t-1}. \quad (4.25)$$

Comparing equations (4.24) and (4.25) with the forecast functions (4.22) and (4.23) one can see that the forecasting models are misspecified when $a \neq 0$ as they do not incorporate the appropriate lag structures. Hence, as a by-product of this simulation experiment the performance of the tests under (dynamic) model misspecification can be considered. Furthermore it is worth mentioning that the persistence of the data increases with a. Neglecting the exogenous regressors the ARMA(2,1) representation displays a degree of persistence, measured by the sum of autoregressive roots, equal to $0.3 + a - 0.3a$. Even without the persistent exogenous variables, the parametrization of Clark and McCracken with $a = 0.9$ results in a high degree of persistence of 0.93. Busetti and Marcucci (2013) consider in their comprehensive Monte Carlo study a data generating process that is similar to the one above. They observe that the presence of persistent regressors can lead to power losses. By varying a in this simulation study the performance of the tests can be compared under different degrees of persistence. However, these aspects are rather a by-product of the simulation study and point the direction to further assessments. The focus remains on the performance of the adjusted DM test when forecasts are exposed to parameter estimation uncertainty.

Before proceeding to the empirical size and power results a note on the long-run variance estimator employed is necessary. The fixed-m approach of Coroneo and Iacone (2020) described in Section 4.4.1 is based on the asymptotic framework of conditional predictive ability à la Giacomini and White (2006). In particular this assumes the presence of asymptotically nonvanishing estimation uncertainty, and consequently is only applicable for fixed and rolling estimation

schemes. The main point, however, is that this perspective considers a philosophically different null hypothesis than the Diebold-Mariano-West approach as already mentioned in Section 4.3. As the objective of these simulation experiments is to analyze the effect of estimation error uncertainty, and to evaluate the strategies proposed in Section 4.3.2, the Newey and West (1994) estimator based on equation (4.16) is used for estimating the long-run variance of the rationality adjusted DM test. For the standard DM test, it suffices to use the sample variance as an estimator. Since the forecasts are one-step-ahead the DM loss differential does not exhibit serial correlation under the null.

Empirical Size

For the empirical size a range of sample sizes (R, P) is considered, where R and P refer to the number of in-sample observations and one-step-ahead forecasts, respectively. Furthermore the performance is analyzed for both a rolling and a recursive estimation scheme and for different values for the persistence parameter a in equation (4.21). The parameters b_1 and b_2 are set to -1 under the null of equal forecast accuracy.

Varying the parameter a induces different variances of the regressors $x_{1,t}$ and $x_{2,t}$, and thus different rejection rates for the tests. The effect of varying a therefore needs to be analyzed while holding constant the variances of $x_{i,t}$. In the original framework of Clark and Mccracken (2014) it is $a = 0.9$ and $\text{Var}(\epsilon_i) = \sigma^2_{\epsilon_i} = 0.1$ such that $\text{Var}(x_i) = 0.5263$. Thus, $\sigma^2_{\epsilon_i}$ has to be selected such that $\sigma^2_{\epsilon_i}/(1 - a^2) = 0.5263$.

Table 4.2 displays the empirical rejection rates of the DM test (MSE-t) and its rationality adjusted companion (MSE-t_r). Furthermore, the results for the adjusted DM test with the long-run variance estimated according to equations 4.8 and 4.9 are presented by MSE-t_{r,adj_1} and MSE-t_{r,adj_2}, respectively. These are expected to approximately account for the effect of parameter estimation errors on the rationality adjusted test statistic.

Starting with the baseline scenario where $a = 0$ such that the forecasting models are correctly specified, two main observations are apparent. First, the adjusted DM test is oversized and this effect increases with π. The size problems are due to the impact of parameter estimation uncertainty, and are expected to increases as the ratio between forecasts and in-sample observations becomes larger. Second, both size adjustment strategies clearly alleviate the size problems as can be seen in the lines for MSE-t_{r,adj_1} and MSE-t_{r,adj_2}.

Note that the test statistics MSE-t_{r,adj_1} and MSE-t_{r,adj_2} are still solely based

Table 4.2 Empirical size, model predictions

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
$a = 0$								
MSE-t	6.11	5.30	4.88	5.10	5.63	5.68	5.36	4.87
MSE-t_r	9.50	10.61	11.31	14.35	9.24	10.44	10.85	12.03
MSE-t_{r,adj_1}	7.63	6.77	6.33	6.45	7.28	6.84	6.22	4.99
MSE-t_{r,adj_2}	7.72	6.80	6.13	6.37	7.23	6.84	6.22	4.93
$a = 0.5$								
MSE-t	5.39	4.85	4.75	4.34	5.56	5.41	4.76	5.04
MSE-t_r	8.91	8.77	10.14	11.35	9.00	9.65	9.07	10.16
MSE-t_{r,adj_1}	7.17	6.22	6.40	5.78	7.51	7.02	5.73	5.25
MSE-t_{r,adj_2}	6.97	5.36	5.24	4.08	7.08	6.20	4.62	3.55
$a = 0.9$								
MSE-t	5.78	4.86	5.45	4.65	5.58	4.99	5.37	5.68
MSE-t_r	12.15	9.91	11.06	10.03	11.91	10.76	10.88	10.21
MSE-t_{r,adj_1}	10.58	7.90	8.04	5.90	10.66	8.75	8.00	6.77
MSE-t_{r,adj_2}	9.73	6.38	5.63	3.07	9.83	6.79	5.34	3.52

Notes: Reported are the empirical rejection rates at a nominal size of 5% from 10,000 Monte Carlo simulations. MSE-t denotes the DM test, and MSE-t_r refers to the proposed version under forecast rationality. MSE-t_{r,adj_1} and MSE-t_{r,adj_2} denote the adjusted DM tests with the long-run variance estimated according to equations 4.8 and 4.9, respectively. $\pi = P/R$ is the ratio of forecasts and in-sample observations with $R = 200$.

on out-of-sample forecasts and do not use information from the estimation sample. Consequently, these test statistics are only approximately valid as outlined in Section 4.3.2. Against this backdrop, Table 4.2 suggests an overall decent performance of MSE-t_{r,adj_1} and MSE-t_{r,adj_2}. The size problems for $\pi = 0.2$ in small samples may be partially explained by the well-known difficulties of long-run variance estimation via the Newey and West (1994) approach in small samples. For large π the size adjustments tend to yield conservative tests.

Turning to the cases with $a \neq 0$, especially to the case of $a = 0.9$, the size adjustment strategies still improve the size problematic but do not work as good as in the benchmark case. Furthermore it is interesting to note that for the adjusted DM test the size does not further deteriorate as π increases when $a = 0.9$

For the case of a smaller estimation sample of $R = 100$ the empirical size results can be found in Table B.5 in Appendix B.3. This table reports the same patterns as described above.

Table 4.3 Sized-adjusted empirical power, model predictions

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
$a = 0$								
MSE-t	13.09	20.51	27.11	36.06	15.49	20.54	28.16	42.24
MSE-t_r	21.34	33.17	43.09	52.69	22.51	34.39	45.19	63.67
MSE-t_{r,adj_1}	21.76	33.30	42.67	51.46	22.98	33.85	44.82	63.50
MSE-t_{r,adj_2}	21.34	33.17	43.09	52.69	22.51	34.39	45.19	63.67
$a = 0.5$								
MSE-t	14.68	21.30	27.43	39.07	15.06	21.36	30.19	42.04
MSE-t_r	20.95	31.63	37.76	51.54	21.15	29.92	43.66	60.36
MSE-t_{r,adj_1}	21.27	31.36	37.21	50.04	21.49	30.53	42.88	58.73
MSE-t_{r,adj_2}	20.95	31.63	37.76	51.54	21.15	29.92	43.66	60.36
$a = 0.9$								
MSE-t	13.28	20.73	23.37	36.36	14.66	21.55	24.87	35.45
MSE-t_r	13.45	21.04	25.70	41.30	12.87	21.65	29.39	42.72
MSE-t_{r,adj_1}	12.89	20.83	24.99	40.38	12.88	21.05	28.10	40.85
MSE-t_{r,adj_2}	13.45	21.04	25.70	41.30	12.87	21.65	29.39	42.72

Notes: See the notes to Table 4.2.

Empirical Power

Table 4.3 shows the size-adjusted empirical power of the different testing variants. The parameters b_1 and b_2 are set to -1 and 0, respectively. The rationality adjustment of the DM test clearly affects a power improvement. The improvements are especially strong for the cases of $a = 0$ and $a = 0.5$. When $a = 0.9$ the data generating process has a high persistence and the forecasting models are misspecified. In this case the power of both the DM test and the adjusted DM test decreases - an observation that is also made by Busetti and Marcucci (2013) for highly persistent data. While the power loss for the DM test is rather small, the adjusted DM test suffers more from the high persistence in the regressors. However, it still displays equal or larger power than the DM test in this simulation setting.

The same patterns can be observed in case of a smaller estimation sample size of $R = 100$ (see Table B.6 in Appendix B.3).

4.4.4 Nested Forecast Comparison

For the nested forecast comparison a data generating process used by Clark and West (2007) is adopted. The data generating process is motivated by the predictive content of factor indices of economic activity on output growth. It

takes the following form:

$$y_{t+1} = 2.247 + 0.261y_t + \gamma_1 z_t + \gamma_2 z_{t-1} + \gamma_3 z_{t-2} + \gamma_4 z_{t-3} + u_{t+1}, \tag{4.26}$$

$$z_t = 0.804z_{t-1} - 0.221z_{t-2} + 0.226z_{t-3} - 0.205z_{t-4} + \epsilon_t. \tag{4.27}$$

The nested competing forecasting models are

$$y_{1,t+1} = \alpha_0 + \alpha_1 y_t + e_{1,t+1}, \tag{4.28}$$

$$y_{2,t+1} = \beta_0 + 0.261y_t + \beta_1 z_t + \beta_2 z_{t-1} + \beta_3 z_{t-2} + \beta_4 z_{t-3} + e_{2,t+1}. \tag{4.29}$$

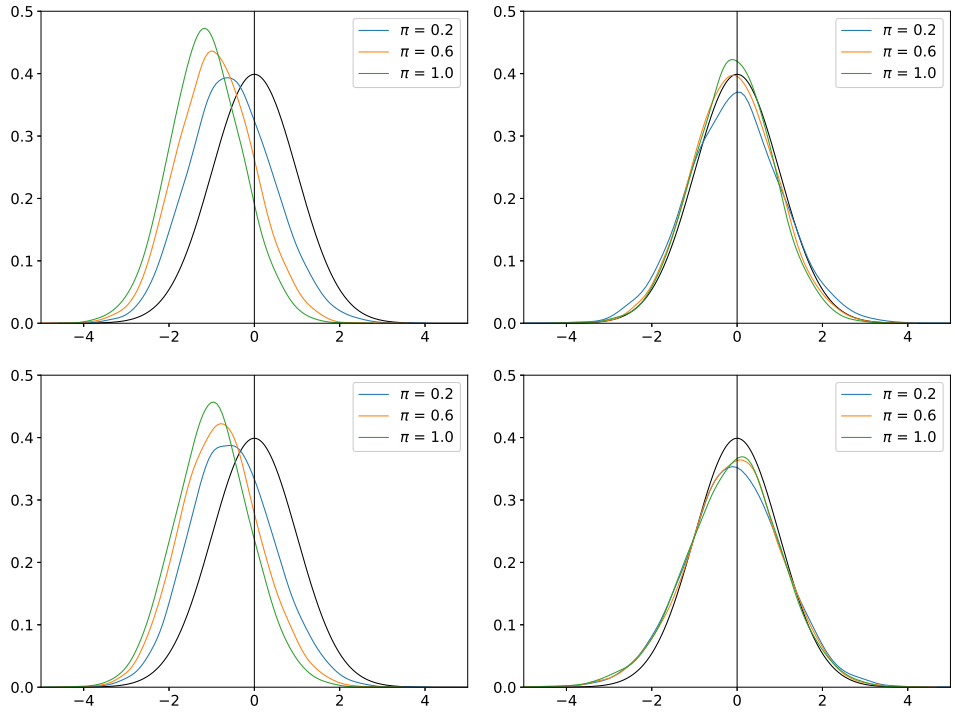
To simulate the size of the tests under the null of equal predictive ability, γ_i is set to 0 for $i = 1, \ldots, 4$. As in Clark and West (2007), equation (4.26) is parametrized by $\gamma = (3.363, -0.633, -0.377, -0.529)'$ in the power experiments.[7] Parameter estimates and forecasts of equations (4.28) and (4.29) are generated on a rolling as well as a recursive estimation scheme.

Empirical Size

As outlined in Section 4.3.3, the rationality adjusted DM test is expected to remain properly centered in a nested forecast comparison. As reported, for instance, by Clark and McCracken (2005) and Clark and West (2007) the DM test is seriously undersized when the competing models are nested. To some extent this effect can be attributed to a shift in the null distribution due to the lower parameter estimation uncertainty of the parsimonious model. This bias becomes more pronounced as the ratio between forecasts P and in-sample observations R gets larger. Figure 4.3 visualizes the shift of the null densities of the DM test statistic for different ratios of $\pi = P/R$. It presents smoothed density estimates of the DM and the adjusted DM test statistic under the null parametrization of the data generating process described above. The associated size results can be found in the first two lines of Table 4.4. Clearly, the DM is undersized as the right-sided test rejects at 1.6449 when a 5% level of significance is used. In contrast, the adjusted DM test does not suffer from the bias caused by the differing parameter estimation uncertainty of both models. As Figure 4.3 shows, the adjusted DM test remains centered under the null and shows only a marginal leftward shift.

Tables 4.4 and B.7 (Appendix B.4) report the empirical size results for different tests and a range of sample sizes (R, P). The first line shows the empirical

[7] The initial observations necessitated by the autoregressive structure of the DGP are generated by draws form the unconditional distribution implied by the DGP.



(a) DM test (b) adjusted DM test

Notes: Results from 10,000 Monte Carlo simulations according to the data generating process under the null specified in equations (4.26) - (4.29). The upper panel shows simulated densities under the rolling forecast scheme. The lower panel considers the case of recursive forecasts. The estimation sample size is $R = 200$. The number of forecasts P is varying.

Fig. 4.3 Null densities of simulated test.

rejection rates of the DM test that is clearly undersized due to the reasons outlined above. MSE-t_r refers to the rationality adjusted DM test and MSE-t_{cw} to the approach of Clark and West (2007) based on the loss differential in equation (4.13). MSE-t_{sim} denotes the DM test with simulated critical values as presented by McCracken (2007). The latter holds the size almost exactly as it is expected. Both the rationality adjusted DM test and the test of Clark and West are only approximately valid. Nevertheless, it is striking that both approaches alleviate the size distortions of the standard DM test to a great extent.

It is clear that the simulated critical values from the exact asymptotic distribution render the most exact test results. From a practical point of view one may take into account that generating these critical values is not trivial. McCracken (2007) reports the appropriate critical values for a broad variety of settings, which, however, still have to suit to the testing problem at hand.

Table 4.4 Empirical size, nested forecasts

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
MSE-t	1.82	0.69	0.31	0.09	1.85	0.91	0.59	0.17
MSE-t_r	6.19	4.58	4.21	3.08	7.31	5.33	5.45	5.93
MSE-t_{cw}	4.88	4.17	4.46	3.70	5.14	4.24	4.13	3.84
MSE-t_{sim}	5.66	4.44	5.40	4.82	5.68	4.83	4.56	5.16
MSE-t_{r,adj_1}	4.28	1.88	1.31	0.34	5.06	2.74	1.95	1.17
MSE-t_{r,adj_2}	4.82	2.12	1.55	0.39	5.69	3.24	2.10	1.39

Notes: Reported are the empirical rejection rates at a nominal size of 5% from 10,000 Monte Carlo simulations. MSE-t denotes the DM test, and MSE-t_r refers to the proposed version under forecast rationality. MSE-t_{r,adj_1} and MSE-t_{r,adj_2} denote the adjusted DM tests with the long-run variance estimated according to equations 4.8 and 4.9, respectively. MSE-t_{cw} refers to the encompassing test statistic of Clark and West (2007) and MSE-t_{sim} denotes the DM test under simulated critical values according to McCracken (2007). $\pi = P/R$ is the ratio of in-sample observations and forecasts with $R = 200$.

Furthermore it is interesting to note that the difficulties of the rational DM test to deal with parameter estimation uncertainty does not seem to be prevalent in this nested model comparison. An intuitive explanation might be that the bias towards the parsimonious model counteracts the size dispersion to some extent. However, when looking at the good *centering* of the rationality adjusted DM test in figure 4.3, it is questionable if this explanation is comprehensive.

Consequently, further accounting for parameter estimation uncertainty via equations 4.8 or 4.9 does not seem to be adequate and results in an undersized test as can be seen in the lower part of Tables 4.4 and B.7 (Appendix B.4).

Empirical Power

Tables 4.5 and B.8 (Appendix B.4) show the empirical power of the tests under consideration. The power figures are not size adjusted since the empirical size in the simulations has been close to the nominal size. The test of Clark and West (2007) clearly achieves the highest power, followed by the DM test based on the simulated critical values of McCracken (2007). The rationality adjusted DM test still yields a decent power improvement compared to the original DM test.

As discussed in Section 4.3.3 the approach of Clark and West can be used as a test for equal MSE only when the competing forecasts stem from nested models. In this case, the power results suggest that using the test statistic of Clark and West is an easy to implement and powerful approach. However, when it is not clear whether the forecast comparison is truly nested or when the distinction between nested and non-nested is ambiguous, the test of Clark and West does

Table 4.5 Empirical power, nested forecasts

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
MSE-t	48.30	74.69	88.93	98.82	49.38	75.19	90.32	98.61
MSE-t_r	64.68	89.01	96.93	99.90	66.05	88.74	97.28	99.77
MSE-t_{cw}	94.05	99.75	99.98	99.99	94.23	99.87	99.99	99.99
MSE-t_{sim}	70.97	93.51	99.08	99.98	71.12	93.05	98.90	99.96
MSE-t_{r,adj_1}	58.61	82.92	93.06	99.34	60.25	82.91	93.71	98.97
MSE-t_{r,adj_2}	58.07	81.22	91.46	98.69	59.61	80.84	92.22	98.46

Notes: See the notes to Table 4.4.

not allow for inference on equal MSE. The same argument holds for the DM test based on the simulated critical values. In this case, the rationality adjusted DM test can offer an alternative that guarantees a test of the null hypothesis of equal predictive accuracy and still improves the power compared to the standard DM test.

For completeness of the discussion one has to admit that potential size violations of the rationality adjusted DM test have been observed in the simulations for non-nested models. Including parameter estimation uncertainty via estimating the long-run variance according to equations 4.8 or 4.9 one gives up the power improvement to some extent, but still obtains a higher power than for the DM test both if the comparison is nested or non-nested.

4.5 Conclusion

This paper shows that the power of the Diebold and Mariano test can be improved when the competing forecasts are rational. By decomposing the MSE loss differential and exploiting forecast rationality, a simplified variant of the DM test statistic is derived. The rationality adjusted DM test is examined both in a model-free environment of survey forecasts, and under the influence of parameter estimation uncertainty, which arises when the competing forecasts stem from estimated (statistical) models. For the latter a simple-to-use adjustment of the long-run variance estimation of the test statistic is proposed. This adjustment accounts for the additional uncertainty from parameter estimation. Furthermore, it is shown that the rationality adjusted DM test has some appealing properties in nested forecast comparisons.

The proposed variant of the DM test statistic rests upon the assumption of forecast rationality. Whether forecasts are always rational is an debated issue

in the academic literature. The effect of potential rationality violations on the adjusted DM test is briefly discussed.

There are interesting objectives for further investigation. First, the application of the adjusted DM test in an empirical study is a natural and interesting complement to the study.

This paper examines the scenarios of survey- and model-based forecasts separately. From an empirical perspective, a mixed forecast comparison provides another relevant scenario, where one of the forecast series stems from an estimated statistical model whereas the other one is survey based. This scenario sometimes occurs when, for instance, the Survey of Professional forecasters is involved (see, e.g., the empirical part in Demetrescu et al. (2019)).

Moreover, the asymptotic behavior of the rationality adjusted DM test in the framework of asymptotically nonvanishing estimation uncertainty of Giacomini and White (2006) remains subject to further investigation.

Appendix A

Appendix for Chapter 2

A.1 Proof of Equation (2.11)

Lemma 1. *Let $H_{\tilde{\theta}}$ denote the projection of y on $\hat{y}$. It holds*

$$\lim_{\tilde{\theta}\to\infty} tr(H_{\tilde{\theta}}) = 1.$$

Proof. Let $\kappa_{\tilde{\theta}}$ denote the (pseudo-)dimension of the subspace of fitted values:

$$\kappa_{\tilde{\theta}} = tr\left\{H_{\tilde{\theta}}\right\} = tr\left\{ X \left[\underbrace{\left(1+\tilde{\theta}\frac{\hat{Q}_2}{\hat{\sigma}_f^2}\right) I_n - \frac{\tilde{\theta}}{\hat{\sigma}_f^2}\frac{1}{T}X'X}_{:=G} \right]^{-1} \left(X'X\right)^{-1} X' \right\}$$

$$\iff \quad \kappa_{\tilde{\theta}} = tr\left\{G^{-1}\right\} = tr\left\{(VDV')^{-1}\right\} = tr\left\{D^{-1}\right\},$$

where V and D contain the eigenvectors and eigenvalues, respectively, of the symmetric matrix G.
G can be rewritten as

$$G = I_n + \tilde{\theta}\left(\underbrace{\frac{\widehat{Q}_2}{\hat{\sigma}_f^2} I_n - \frac{1}{\hat{\sigma}_f^2}\Sigma}_{:=K}\right),$$

where $\Sigma = \frac{1}{T}X'X$. For the eigenvalues $d_j^{(G)}$ of G holds

$$d_j^{(G)}(\tilde{\theta}) = 1 + \tilde{\theta} d_j^{(K)}, \tag{A.1}$$

where $d_j^{(K)}$ denotes the j^{th} eigenvalue of K. Considering

$$K = \frac{\widehat{Q}_2}{\hat{\sigma}_f^2} I_n - \frac{1}{\hat{\sigma}_f^2}\Sigma = \frac{1}{\hat{\sigma}_f^2}\left(\widehat{Q}_2 I_n - \Sigma\right)$$

and recalling that $\hat{\sigma}_f^2$ and $\widehat{Q}_2$ are scalars, the eigenvalues $d_j^{(K)}$ of K are given by

$$d_1^{(K)} = \frac{1}{\hat{\sigma}_f^2}\left(-d_1^{(\Sigma)} + \widehat{Q}_2\right), \ldots, \; d_n^{(K)} = \frac{1}{\hat{\sigma}_f^2}\left(-d_n^{(\Sigma)} + \widehat{Q}_2\right),$$

where $d_1^{(\Sigma)}, ..., d_n^{(\Sigma)}$ denote the eigenvalues of Σ in decreasing order. For $\tilde{\theta} \to \infty$, $\widehat{Q}_2$ is equal to the largest eigenvalue of Σ (see end of section). Hence,

$$d_1^{(K)} = \frac{1}{\hat{\sigma}_f^2}\left(\underbrace{-d_1^{(\Sigma)} + \widehat{Q}_2}_{=0}\right), \; d_2^{(K)} = \frac{1}{\hat{\sigma}_f^2}\left(\underbrace{-d_2^{(\Sigma)} + \widehat{Q}_2}_{>0}\right), \; \ldots \; .$$

Exploiting relationship (A.1) it follows that

$$\lim_{\tilde{\theta}\to\infty} \kappa_{\tilde{\theta}} = \lim_{\tilde{\theta}\to\infty} tr\left\{D^{-1}\right\} = \lim_{\tilde{\theta}\to\infty} \sum_{j=1}^{n} \frac{1}{d_j^{(G)}} = \lim_{\tilde{\theta}\to\infty} \sum_{j=1}^{n} \frac{1}{1+\tilde{\theta} d_j^{(K)}} = 1.$$

To see why $\widehat{Q}_2$ is equal to the largest eigenvalue of Σ for $\tilde{\theta} \to \infty$ consider

$$\widehat{Q}_2 = \frac{1}{T}\frac{\beta_{\tilde{\theta}}' X'XX'X\beta_{\tilde{\theta}}}{\beta_{\tilde{\theta}}' X'X\beta_{\tilde{\theta}}} = \frac{\beta_{\tilde{\theta}}' \Lambda D \Lambda' \Lambda D \Lambda' \beta_{\tilde{\theta}}}{\beta_{\tilde{\theta}}' \Lambda D \Lambda' \beta_{\tilde{\theta}}},$$

where Λ and D consist of the eigenvectors and eigenvalues, respectively, of $\Sigma = \frac{1}{T}X'X$. For $\tilde{\theta} \to \infty$, one obtains the (unsupervised) principal components solution such that $\beta_{\tilde{\theta}} = \lambda_1$, where λ_1 denotes the first eigenvector of Σ belonging to its largest eigenvalue. Hence,

$$\lim_{\tilde{\theta}\to\infty} \widehat{Q}_2 = \frac{\lambda_1' \Lambda D \Lambda' \Lambda D \Lambda' \lambda_1}{\lambda_1' \Lambda D \Lambda' \lambda_1} = d_1^{(\Sigma)}.$$

□

A.2 Datasets

The dataset is taken from the monthly macroeconomic database provided by the Federal Reserve Bank of St. Louis. All variables are transformed to get stationary series as described in McCracken and Ng (2016).

Table A.1 Sub-dataset to forecast industrial production

	tcode	fred	description
Group 1: Output and income			
1	5	W875RX1	Real Personal Income ex Transfer Receipts
2	5	INDPRO	IP Index
3	5	IPMANSICS	IP: Manufacturing (SIC)
4	1	NAPMPI	ISM Manufacturing: Production Index
5	2	CUMFNS	Capacity Utilization: Manufacturing
Group 2: Labour market			
6	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed
7	5	CLAIMSx	Initial Claims
8	5	PAYEMS	All Employees: Total nonfarm
9	5	MANEMP	All Employees: Manufacturing
10	5	SRVPRD	All Employees: Service-Providing Industries
11	1	CES0600000007	Avg Weekly Hours: Goods-Producing
12	2	AWOTMAN	Avg Weekly Overtime Hours: Manufacturing
13	1	AWHMAN	Avg Weekly Hours: Manufacturing
14	1	NAPMEI	ISM Manufacturing: Employment Index
Group 3: Housing			
15	4	HOUST	Housing Starts: Total New Privately Owned
Group 4: Consumption, orders, and inventories			
16	5	DPCERA3M086SBEA	Real Personal Consumption Expenditures
17	5	CMRMTSPLx	Real Manu. and Trade Industries Sales
18	5	RETAILx	Retail and Food Services Sales
19	1	NAPMNOI	ISM: New Orders Index
20	1	NAPMSDI	ISM: Supplier Deliveries Index
21	5	AMDMNOx	New Orders for Durable Goods
22	5	ANDENOx	New Orders for Nondefense Capital Goods
23	5	AMDMUOx	Unfilled Orders for Durable Goods
24	2	UMCSENTx	Consumer Sentiment Index
Group 5: Money and credit			
25	5	M2REAL	Real M2 Money Stock
Group 6: Interest and exchange rate			
26	1	COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS
27	1	TB3SMFFM	3-Month Treasury C Minus FEDFUNDS
28	1	TB6SMFFM	6-Month Treasury C Minus FEDFUNDS
29	1	T1YFFM	1-Year Treasury C Minus FEDFUNDS
30	1	T5YFFM	5-Year Treasury C Minus FEDFUNDS
31	1	T10YFFM	10-Year Treasury C Minus FEDFUNDS
32	1	AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS
33	1	BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS
Group 7: Stock market			
34	5	S&P: indust	S&P's Common Stock Price Index: Industrials

Notes: 'tcode' denotes the following data transformation for a series x: (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $log(x_t)$; (5) $\Delta log(x_t)$; (6) $\Delta^2 log(x_t)$; (7) $\Delta(x_t/x_{t-1} - 1)$. 'fred' gives mnemonics in FRED.

Table A.2 Sub-dataset to forecast the Consumer Price Index

	tcode	fred	description
Group 1: Output and income			
1	5	W875RX1	Real Personal Income ex Transfer Receipts
2	5	INDPRO	IP Index
3	5	IPMANSICS	IP: Manufacturing (SIC)
4	1	NAPMPI	ISM Manufacturing: Production Index
5	2	CUMFNS	Capacity Utilization: Manufacturing
Group 2: Labour market			
6	2	HWIURATIO	Ratio of Help Wanted/No. Unemployed
7	5	CLAIMSx	Initial Claims
8	5	PAYEMS	All Employees: Total nonfarm
9	5	MANEMP	All Employees: Manufacturing
10	5	SRVPRD	All Employees: Service-Providing Industries
11	1	CES0600000007	Avg Weekly Hours: Goods-Producing
12	2	AWOTMAN	Avg Weekly Overtime Hours: Manufacturing
13	1	AWHMAN	Avg Weekly Hours: Manufacturing
14	1	NAPMEI	ISM Manufacturing: Employment Index
Group 3: Housing			
15	4	HOUST	Housing Starts: Total New Privately Owned
Group 4: Consumption, orders, and inventories			
16	5	DPCERA3M086SBEA	Real Personal Consumption Expenditures
17	5	CMRMTSPLx	Real Manu. and Trade Industries Sales
18	5	RETAILx	Retail and Food Services Sales
19	1	NAPMNOI	ISM: New Orders Index
20	1	NAPMSDI	ISM: Supplier Deliveries Index
21	5	AMDMNOx	New Orders for Durable Goods
22	5	ANDENOx	New Orders for Nondefense Capital Goods
23	5	AMDMUOx	Unfilled Orders for Durable Goods
24	2	UMCSENTx	Consumer Sentiment Index
Group 5: Money and credit			
25	5	M2REAL	Real M2 Money Stock
Group 6: Interest and exchange rate			
26	2	FEDFUNDS	Effective Federal Funds Rate
27	2	CP3Mx	3-Month AA Financial Commercial Paper Rate
28	2	TB3MS	3-Month Treasury Bill:
29	2	TB6MS	6-Month Treasury Bill:
30	2	GS1	1-Year Treasury Rate
31	2	GS5	5-Year Treasury Rate
32	2	GS10	10-Year Treasury Rate
33	2	AAA	Moody's Seasoned Aaa Corporate Bond Yield
34	2	BAA	Moody's Seasoned Baa Corporate Bond Yield
Group 7: Stock market			
35	5	S&P: indust	S&P's Common Stock Price Index: Industrials
Group 8: Prices			
36	6	WPSFD49207	PPI: Finished Goods
37	6	WPSFD49502	PPI: Finished Consumer Goods
38	6	OILPRICEx	Crude Oil, spliced WTI and Cushing
39	1	NAPMPRI	ISM Manufacturing: Prices Index
40	6	CPIAUCSL	CPI : All Items
41	6	PCEPI	Personal Cons. Expend.: Chain Index

Notes: 'tcode' denotes the following data transformation for a series x: (1) no transformation; (2) Δx_t; (3) $\Delta^2 x_t$; (4) $log(x_t)$; (5) $\Delta log(x_t)$; (6) $\Delta^2 log(x_t)$; (7) $\Delta(x_t/x_{t-1} - 1)$. 'fred' gives mnemonics in FRED.

Appendix B

Appendix for Chapter 4

B.1 Proof of Proposition 1

Proof. Consider the rational loss differential (4.5) under parameter estimation uncertainty:

$$\begin{aligned} d_{r,t}^{(\hat{\theta})} &= \hat{Y}_{2,t+h|t}\hat{e}_{1,t+h|t} - \hat{Y}_{1,t+h|t}\hat{e}_{2,t+h|t} \\ &= \hat{Y}_{2,t+h|t}\left(Y_{t+h} - \hat{Y}_{1,t+h|t}\right) - \hat{Y}_{1,t+h|t}\left(Y_{t+h} - \hat{Y}_{2,t+h|t}\right) \\ &= \left(\hat{Y}_{2,t+h|t} - \hat{Y}_{1,t+h|t}\right) Y_{t+h}. \end{aligned}$$

Applying a mean-value expansion of the form

$$\hat{Y}_{i,t+h|t} = Y_{i,t+h|t} + D_{t+h}(\bar{\theta}_{i,t})(\hat{\theta}_{i,t} - \theta_i),$$

where $\bar{\theta}_{i,t}$ denotes a value between $\hat{\theta}_{i,t}$ and θ_i, yields

$$\begin{aligned} d_{r,t}^{(\hat{\theta})} &= \Big(Y_{2,t+h|t} + D_{t+h}(\bar{\theta}_{2,t})(\hat{\theta}_{2,t} - \theta_2) \\ &\quad - \left(Y_{1,t+h|t} + D_{t+h}(\bar{\theta}_{1,t})(\hat{\theta}_{1,t} - \theta_1)\right)\Big) Y_{t+h} \\ &= \left(Y_{2,t+h|t} - Y_{1,t+h|t}\right) Y_{t+h} \\ &\quad + \left(D_{t+h}(\bar{\theta}_{2,t})(\hat{\theta}_{2,t} - \theta_2) - D_{t+h}(\bar{\theta}_{1,t})(\hat{\theta}_{1,t} - \theta_1)\right) Y_{t+h} \\ &= \left(Y_{2,t+h|t} - Y_{1,t+h|t}\right) Y_{t+h} \\ &\quad + Y_{t+h}D_{t+h}(\bar{\theta}_{2,t})O_p\left(\sqrt{P}/R\right) - Y_{t+h}D_{t+h}(\bar{\theta}_{1,t})O_p\left(\sqrt{P}/R\right) \\ &= \left(Y_{2,t+h|t} - Y_{1,t+h|t}\right) Y_{t+h} + O_p\left(\sqrt{P}/R\right). \end{aligned}$$

It follows that

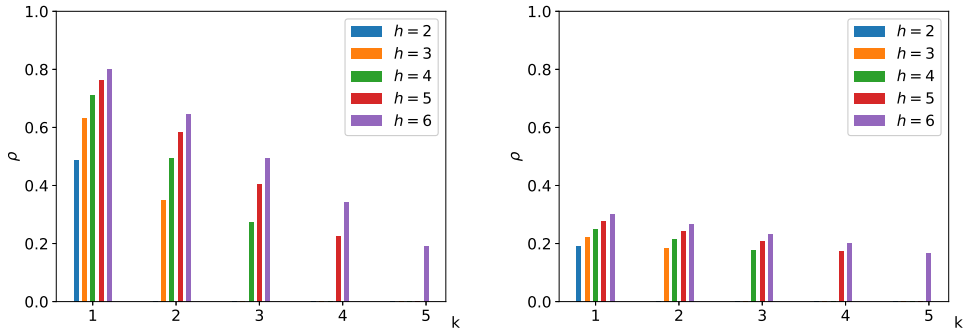
$$\frac{1}{\sqrt{P}}\sum_{t=1}^{P} d_{r,t}^{(\hat{\theta})} = \frac{1}{\sqrt{P}}\sum_{t=1}^{P}\left[e_{1,t+h|t}Y_{2,t+h|t} - e_{2,t+h|t}Y_{1,t+h|t}\right] + O_p\left(P/R\right)$$

and

$$\begin{aligned}\widehat{\Gamma}_j &= \frac{1}{P}\sum_{t=1+j}^{P} d_{r,t}^{(\hat{\theta})} d_{r,t-j}^{(\hat{\theta})} \\ &= \frac{1}{P}\sum_{t=1}^{P} \left(e_{1,t+h|t}Y_{2,t+h|t} - e_{2,t+h|t}Y_{1,t+h|t}\right)\left(e_{1,t+h|t}Y_{2,t+h|t} - e_{2,t+h|t}Y_{1,t+h|t}\right) \\ &\quad + O_p\left(1/R\right) \\ &\xrightarrow{p} \Gamma_j\end{aligned}$$

It follows that $\widehat{\omega}^2 \xrightarrow{p} \omega^2$ and Proposition 1 follows immediately. □

B.2 Simulation Results Survey Forecasts



a) high autocorrelation setup b) low autocorrelation setup

Notes: Forecast error autocorrelations profiles for h-step-ahead forecasts generated according the data generating process described in Section 4.4.2.

Fig. B.1 Forecast error autocorrelations profiles.

Table B.1 Empirical size, high forecast error cross-correlation

	low error serial correlation				high error serial correlation			
T	25	50	100	200	25	50	100	200
1-step-ahead								
MSE-t	4.21	4.77	4.72	4.92	4.21	4.77	4.72	4.92
MSE-t_r	4.49	4.55	4.80	4.86	4.49	4.55	4.80	4.86
2-steps-ahead								
MSE-t	4.45	4.72	5.09	5.11	4.25	4.64	4.43	5.33
MSE-t_r	4.27	4.27	4.45	5.11	3.61	4.36	4.33	5.09
3-steps-ahead								
MSE-t	4.31	4.61	4.74	4.59	3.91	4.17	4.98	4.74
MSE-t_r	3.70	4.41	4.64	4.94	3.93	4.15	4.80	4.93
4-steps-ahead								
MSE-t	4.39	4.35	4.34	4.52	3.75	4.26	4.70	4.95
MSE-t_r	4.04	4.44	4.34	4.81	3.60	4.25	4.78	4.88
5-steps-ahead								
MSE-t	4.68	4.89	4.94	4.63	3.41	4.17	4.64	4.88
MSE-t_r	4.34	4.32	4.56	4.94	3.45	4.21	4.63	4.53
6-steps-ahead								
MSE-t	4.38	4.42	4.77	4.98	4.31	4.41	4.42	4.79
MSE-t_r	4.69	4.73	4.55	4.81	3.94	4.51	4.56	4.64

Notes: Reported are the empirical rejection rates at a nominal size of 5% from 10,000 Monte Carlo simulations. MSE-t denotes the DM test, and MSE-t_r refers to the proposed version under forecast rationality. T denotes the number of forecast error observations.

Table B.2 Empirical size, moderate forecast error cross-correlation

	low error serial correlation				high error serial correlation			
T	25	50	100	200	25	50	100	200
1-step-ahead								
MSE-t	4.15	4.44	4.63	4.68	4.15	4.44	4.63	4.68
MSE-t_r	4.36	4.45	4.68	4.54	4.36	4.45	4.68	4.54
2-steps-ahead								
MSE-t	4.26	4.44	4.79	5.07	4.06	4.57	4.58	4.96
MSE-t_r	4.49	4.53	4.82	5.08	3.89	4.39	4.44	5.13
3-steps-ahead								
MSE-t	4.63	4.67	5.05	4.78	3.66	4.47	4.56	4.73
MSE-t_r	4.33	4.73	4.78	4.86	3.68	4.36	4.77	4.81
4-steps-ahead								
MSE-t	4.50	4.46	4.74	5.20	3.93	4.40	4.55	5.05
MSE-t_r	4.03	3.19	4.53	4.80	4.02	4.43	4.58	5.01
5-steps-ahead								
MSE-t	4.08	4.73	4.84	4.96	4.01	4.02	4.31	4.89
MSE-t_r	4.17	4.38	4.56	4.59	4.32	4.16	4.25	5.05
6-steps-ahead								
MSE-t	4.66	4.67	4.93	5.07	4.02	4.24	4.14	4.94
MSE-t_r	4.56	4.46	4.80	4.85	3.75	4.40	4.11	4.66

Notes: See the notes to Table B.1.

Table B.3 Empirical power, high forecast error cross-correlation

	low error serial correlation				high error serial correlation			
T	25	50	100	200	25	50	100	200
1-step-ahead								
MSE-t	19.11	33.58	58.06	85.69	19.11	33.58	58.06	85.69
MSE-t_r	30.29	59.31	88.02	99.29	30.29	59.31	88.02	99.29
2-steps-ahead								
MSE-t	18.39	32.47	56.17	83.58	14.94	26.99	44.57	71.78
MSE-t_r	28.24	55.19	84.51	98.60	24.64	49.28	78.34	97.15
3-steps-ahead								
MSE-t	17.80	31.34	53.62	80.61	12.61	21.59	36.25	59.64
MSE-t_r	27.46	52.27	81.19	97.59	21.52	41.54	69.68	93.30
4-steps-ahead								
MSE-t	17.70	29.46	50.40	77.13	11.71	19.19	30.65	50.15
MSE-t_r	26.61	49.44	77.81	96.44	18.61	36.16	61.36	88.38
5-steps-ahead								
MSE-t	17.14	28.65	48.41	74.23	11.38	16.59	27.50	44.99
MSE-t_r	26.11	47.17	74.98	94.94	18.02	31.89	55.30	82.81
6-steps-ahead								
MSE-t	17.18	27.73	45.16	70.51	10.15	15.99	24.09	38.72
MSE-t_r	26.76	46.56	71.18	93.91	16.59	29.64	50.47	77.56

Notes: Reported are the empirical rejection rates at a nominal size of 5% from 10,000 Monte Carlo simulations. For the power analysis, the parameter α of the DGP from Section 4.4.2 is set to 0.75. MSE-t denotes the DM test, and MSE-t_r refers to the proposed version under forecast rationality. T denotes the number of forecast error observations.

Table B.4 Empirical power, moderate forecast error cross-correlation

	low error serial correlation				high error serial correlation			
T	25	50	100	200	25	50	100	200
1-step-ahead								
MSE-t	61.00	92.53	99.92	99.99	61.00	92.53	99.92	99.99
MSE-t_r	76.59	98.90	99.99	99.99	76.59	98.90	99.99	99.99
2-steps-ahead								
MSE-t	58.48	92.23	99.84	99.99	46.63	83.24	98.98	99.99
MSE-t_r	73.91	98.22	99.98	99.99	64.39	96.18	99.99	99.99
3-steps-ahead								
MSE-t	56.99	90.62	99.81	99.99	37.93	71.63	95.88	99.92
MSE-t_r	71.65	97.68	99.99	99.99	55.52	90.97	99.90	99.99
4-steps-ahead								
MSE-t	54.91	88.64	99.64	99.99	32.27	61.22	90.45	99.53
MSE-t_r	70.19	96.84	99.98	99.99	49.13	85.13	99.26	99.99
5-steps-ahead								
MSE-t	53.14	86.70	99.29	99.99	28.58	54.39	84.16	98.69
MSE-t_r	69.24	96.05	99.93	99.99	45.07	78.60	97.85	99.98
6-steps-ahead								
MSE-t	54.33	85.16	98.87	99.99	26.64	49.34	78.54	97.50
MSE-t_r	69.31	95.26	99.95	99.99	41.23	73.07	96.18	99.96

Notes: See the notes to Table B.3.

B.3 Simulation Results Model Forecasts

Table B.5 Empirical size, model predictions, small sample

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
$a = 0$								
MSE-t	6.50	5.56	4.59	4.45	5.86	5.58	5.10	5.55
MSE-t_r	10.22	10.29	10.69	13.43	9.87	10.62	10.50	12.54
MSE-t_{r,adj_1}	7.92	6.66	5.43	5.10	7.62	6.65	5.58	5.04
MSE-t_{r,adj_2}	8.21	7.01	5.61	5.17	7.88	6.98	5.83	4.96
$a = 0.5$								
MSE-t	6.01	5.58	5.10	4.06	6.21	5.17	5.41	4.63
MSE-t_r	10.22	9.89	10.42	10.35	10.65	9.50	10.31	10.31
MSE-t_{r,adj_1}	8.03	6.62	5.87	4.36	8.51	6.52	6.17	4.53
MSE-t_{r,adj_2}	8.03	6.36	5.19	3.63	8.44	6.15	5.30	3.30
$a = 0.9$								
MSE-t	6.12	5.70	4.79	4.42	6.12	5.75	5.09	5.30
MSE-t_r	12.88	12.19	10.87	10.35	13.17	11.75	10.88	11.03
MSE-t_{r,adj_1}	11.00	8.85	7.07	5.23	11.19	8.96	7.51	6.37
MSE-t_{r,adj_2}	10.86	7.73	5.53	3.55	10.81	7.77	5.77	3.74

Notes: Reported are the empirical rejection rates at a nominal size of 5% from 10,000 Monte Carlo simulations. MSE-t denotes the DM test, MSE-t_r refers to the proposed version under forecast rationality. MSE-t_{r,adj_1} and MSE-t_{r,adj_2} denote the adjusted DM tests with the long-run variance estimated according to eq. 4.8 and eq. 4.9, resp. $\pi = P/R$ with $R = 100$.

Table B.6 Size-adjusted empirical power, model predictions, small sample

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
$a = 0$								
MSE-t	9.38	12.35	15.98	19.81	10.32	12.90	17.07	21.65
MSE-t_r	11.92	17.10	21.17	25.24	13.79	18.46	22.32	30.80
MSE-t_{r,adj_1}	12.00	17.16	20.85	24.34	13.86	18.38	22.88	29.92
MSE-t_{r,adj_2}	11.92	17.10	21.17	25.24	13.79	18.46	22.32	30.80
$a = 0.5$								
MSE-t	9.53	12.49	14.47	20.81	10.04	14.19	15.75	23.36
MSE-t_r	11.57	15.54	18.91	25.22	11.71	16.96	20.25	30.52
MSE-t_{r,adj_1}	11.75	15.19	18.56	25.57	11.68	17.16	20.10	29.94
MSE-t_{r,adj_2}	11.57	15.54	18.91	25.22	11.71	16.96	20.25	30.52
$a = 0.9$								
MSE-t	9.23	11.46	14.45	17.78	9.70	12.69	16.19	20.56
MSE-t_r	8.42	11.18	14.10	17.91	8.93	11.70	15.57	21.19
MSE-t_{r,adj_1}	8.42	11.03	13.53	17.61	8.92	11.33	14.66	20.57
MSE-t_{r,adj_2}	8.42	11.18	14.10	17.91	8.93	11.70	15.57	21.19

Notes: See the notes to Table B.5.

B.4 Simulation Results Nested Forecasts

Table B.7 Empirical size, nested forecasts, small sample

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
MSE-t	2.13	0.94	0.32	0.09	2.16	1.00	0.51	0.22
MSE-t_r	8.27	5.81	4.89	3.66	8.57	6.66	6.27	6.34
MSE-t_{cw}	6.12	4.73	4.55	4.65	5.42	5.00	4.45	4.08
MSE-t_{sim}	7.06	5.15	5.43	4.85	6.10	5.57	4.76	5.25
MSE-t_{r,adj_1}	5.52	2.47	1.33	0.44	5.75	3.16	2.25	1.35
MSE-t_{r,adj_2}	6.50	3.28	1.87	0.74	6.88	3.93	2.69	1.68

Notes: Reported are the empirical rejection rates at a nominal size of 5% from 10,000 Monte Carlo simulations. MSE-t denotes the DM test, MSE-t_r refers to the proposed version under forecast rationality. MSE-t_{r,adj_1} and MSE-t_{r,adj_2} denote the adjusted DM tests with the long-run variance estimated according to equations 4.8 and 4.9, resp. MSE-t_{cw} refers to the encompassing test statistic of Clark and West (2007) and MSE-t_{sim} denotes the DM test under simulated critical values according to McCracken (2007). $\pi = P/R$ with $R = 100$.

Table B.8 Empirical power, nested forecasts, small sample

	rolling scheme				recursive scheme			
π	0.2	0.4	0.6	1.0	0.2	0.4	0.6	1.0
MSE-t	27.78	41.89	55.11	75.53	28.21	44.00	58.34	79.85
MSE-t_r	42.88	62.34	75.90	92.60	42.29	62.88	76.34	92.56
MSE-t_{cw}	70.41	92.88	98.36	99.89	70.72	93.03	98.37	99.98
MSE-t_{sim}	49.49	73.35	89.67	98.69	50.05	73.81	87.03	98.27
MSE-t_{r,adj_1}	35.27	49.63	60.31	77.41	34.83	51.35	61.63	80.26
MSE-t_{r,adj_2}	36.34	49.39	58.04	74.02	35.72	50.33	58.82	76.25

Notes: See the notes to Table B.7.

Bibliography

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Boivin, J. and Ng, S. (2005). Understanding and Comparing Factor-Based Forecasts. *International Journal of Central Banking*, 1(3).

Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, 132(1):169–194.

Breitung, J. and Choi, I. (2013). Factor models. In *Handbook of Research Methods and Applications in Empirical Macroeconomics*, chapter 11, pages 249–265. Edward Elgar Publishing.

Breitung, J. and Knüppel, M. (2020). How far can we forecast? Statistical tests of the predictive content. *Journal of Applied Econometrics*, forthcoming.

Breitung, J. and Roling, C. (2015). Forecasting inflation rates using daily data: A nonparametric midas approach. *Journal of Forecasting*, 34(7):588–603.

Busetti, F. and Marcucci, J. (2013). Comparing forecast accuracy: A monte carlo investigation. *International Journal of Forecasting*, 29(1):13 – 27.

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

Clark, T. and McCracken, M. (2013). Advances in forecast evaluation. In *Handbook of economic forecasting*, volume 2, chapter 20, pages 1107–1201. Elsevier.

Clark, T. E. and McCracken, M. W. (2005). Evaluating direct multistep forecasts. *Econometric Reviews*, 24(4):369–404.

Clark, T. E. and McCracken, M. W. (2012). Reality checks and comparisons of nested predictive models. *Journal of Business & Economic Statistics*, 30(1):53–66.

Clark, T. E. and Mccracken, M. W. (2014). Tests of equal forecast accuracy for overlapping models. *Journal of Applied Econometrics*, 29(3):415–430.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291 – 311. 50th Anniversary Econometric Institute.

Cook, T. R. and Hall, A. S. (2017). Macroeconomic indicator forecasting with deep neural networks. Technical report, Federal Reserve Bank of Kansas City.

Coroneo, L. and Iacone, F. (2020). Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. *Journal of Applied Econometrics*, 35(4):391–409.

Coulombe, P. G., Leroux, M., Stevanovic, D., and Surprenant, S. (2019). How is machine learning useful for macroeconomic forecasting? Cirano working papers, CIRANO.

Croushore, D. (2010). An evaluation of inflation forecasts from surveys using real-time data. *The B.E. Journal of Macroeconomics*, 10(1):1–32.

Croushore, D. (2012). Forecast bias in two dimensions. Working Papers 12-9, Federal Reserve Bank of Philadelphia.

de Jong, S. and Kiers, H. A. L. (1992). Principal covariates regression. part i. theory. *Chemometrics and Intelligent Laboratory Systems*, 14:155–164.

Demetrescu, M., Hanck, C., and Kruse, R. (2019). Technical report, Robust Inference under Time-Varying Volatility: A Real-Time Evaluation of Professional Forecasters.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1):1–38.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

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):1–1.

Diebold, F. X. and Lopez, J. A. (1996). Forecast evaluation and combination. In Maddala, G. and Rao, C., editors, *Handbook of Statistics*, volume 14, chapter 8, pages 241–268. Elsevier.

Elliott, G., Komunjer, I., and Timmermann, A. (2008). Biases in macroeconomic forecasts: Irrationality or asymmetric loss? *Journal of the European Economic Association*, 6(1):122–157.

Giacomini, R. and Rossi, B. (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics*, 25(4):595–620.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gu, S., Kelly, B., and Xiu, D. (2020). Autoencoder asset pricing models. *Journal of Econometrics*.

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

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics*, 16(2):254–259.

Harvey, D. I., Leybourne, S. J., and Whitehouse, E. J. (2017). Forecast evaluation tests and negative long-run variance estimates in small samples. *International Journal of Forecasting*, 33(4):833 – 847.

Heij, C., Groenen, P. J., and van Dijk, D. (2007). Forecast comparison of principal component regression and principal covariate regression. *Computational Statistics & Data Analysis*, 51(7):3612 – 3625.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.*, 2(5):359 – 366.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):271–293.

Jonsson, T. and Österholm, P. (2012). The properties of survey-based inflation expectations in sweden. *Empirical Economics*, 42:79–94.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ica: A unifying framework. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217, Online. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. cite arxiv:1412.6980.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. cite arxiv:1312.6114.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

McCracken, M. W. (2007). Asymptotics for out of sample tests of granger causality. *Journal of Econometrics*, 140(2):719 – 752.

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

Mincer, J. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pages 3–46. National Bureau of Economic Research, Inc.

Müller, U. K. (2014). Hac corrections for strongly autocorrelated time series. *Journal of Business & Economic Statistics*, 32(3):311–322.

Nakamura, E. (2005). Inflation forecasting using a neural network. *Economics Letters*, 86(3):373–378.

Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653.

Patton, A. J. and Timmermann, A. (2007). Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics*, 140:884–918.

Patton, A. J. and Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, 30(1):1–17.

Romer, C. D. and Romer, D. H. (2000). Federal reserve information and the behavior of interest rates. *American Economic Review*, 90(3):429–457.

Rossi, B. and Sekhposyan, T. (2016). Forecast rationality tests in the presence of instabilities, with applications to federal reserve and survey forecasts. *Journal of Applied Econometrics*, 31(3):507–532.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Stock, J. and Watson, M. (2016). Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In Taylor, J. B. and Uhlig, H., editors, *Handbook of Macroeconomics*, volume 2 of *Handbook of Macroeconomics*, chapter 0, pages 415–525. Elsevier.

Stock, J. H. and Watson, M. (2006). Forecasting with many predictors. In Elliott, G., Granger, C., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, chapter 10, pages 515–554. Elsevier, 1 edition.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Tkacz, G. (2001). Neural network forecasting of Canadian GDP growth. *International Journal of Forecasting*, 17(1):57–69.

Umbach, S. L. (2020). Forecasting with supervised factor models. *Empirical Economics*, 58(1):169–190.

Vervloet, M., Deun, K. V., den Noortgate, W. V., and Ceulemans, E. (2013). On the selection of the weighting parameter value in principal covariates regression. *Chemometrics and Intelligent Laboratory Systems*, 123:36 – 43.

Vervloet, M., Kiers, H., den Noortgate, W. V., and Ceulemans, E. (2015). Pcovr: An r package for principal covariates regression. *Journal of Statistical Software*, 65(1):1–14.

West, K. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–84.

Wilderjans, T., Ceulemans, E., and Mechelen, I. V. (2009). Simultaneous analysis of coupled data blocks differing in size: A comparison of two weighting schemes. *Computational Statistics & Data Analysis*, 53(4):1086 – 1098.