Combined Forecasts and Forecast Breakdown Preselection

Inaugural-Dissertation
zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2008

vorgelegt von

Dirk Ulbricht

Referent: Prof. Dr. Gebhard Flaig
Korreferent: PD Dr. Klaus Abberger
Promotionsabschlussberatung: 04. Februar 2009

Acknowledgements

I thank Gebhard Flaig, Klaus Abberger, and Kai Carstensen for supervising my thesis. I am especially indebted to Gebhard Flaig who provided me with many helpful suggestions, important advice, and constant encouragement during the course of this work. I am grateful to Ifo for providing me with a stimulating research environment. I thank all my colleges at the Department of Business Cycle Analyses and Surveys for many helpful insights. I gratefully reckognize contributions of Johannes Mayr, Thomas Strobel, Klaus Wohlrabe, Georg Wamser, and Steffen Henzel. I am indebted to Ulrich Oberndorfer, Oliver Röhn, Ulrike Jäger, and Anja Rohwer for their editorial help. Participants at conferences in Brisbane, Nice, Vienna, and Kiel made very valuable comments to improve this work. All chapters profited from seminar presentations at Ifo. I am particularly indebted to my brother Jochen Ulbricht and Julia Kneissl for moral support.

Contents

List of Figures

List of Tables

Nomenclature

ADF	Augmented Dickey-Fuller test
AE	Average Earnings
ARCH	Autoregressive Conditional Heteroscedasticity
ARMA	Autoregressive Moving Average
BCI	Business Climate Indicator
BD	10-year UK Government benchmark bonds
BIC	Bayesian Information Criterion
CBI	Industrial Trends Survey of UK manufacturing
DEM	Exchange rate of the British Pound to the Deutschmark
DGP	Data Generating Process
ESI	Economic Sentiment Indicator
EW	Equal Weights scheme
FBP	Forcast Breakdown Preselection
FTSE	Financial Times Stock Exchange Index
GARCH	Generalized Autoregressive Conditional Heteroskedasticity

GC	GARCH Combination scheme
GP	GARCH Preselection scheme
iid	identically, independently distributed
INFL	UK inflation
INSTABLE	Instable period of the empirical experiment
INV	Inverse MSE weighting scheme
IP	Industrial Production
M0	Monetary aggregate 0
MA	Moving Average process
MC	Monte Carlo Study
Median	Median forecast
MS	Markov Switching model
ODD	Odds-Matrix combination scheme
ONS	Office of National Statistics
OPT	Optimal combination scheme
OVB	Omitted Variable Bias

RPI	Retail Price Index
RS	Retail Sales
RTDB	Real-Time Data Base
RW	Rank Weighting scheme
STABLE	Stable period of the empirical experiment
TB	Deflated UK Treasury Bill rate
TBUS	Deflated US Treasury Bill rate
TU	Theil's U
TVP	Time Varying Parameter model
U	Unemployment
USD	Exchange rate of the British Pound to the US Dollar

Chapter 1

Introduction

This dissertation takes a fresh look at combined forecasts in the presence of structural change. A new method is proposed that refines combined forecasts filtering out those models that are very likely to become unreliable over the forecast horizon.

Economic forecasting as much as forecasting in every other field relies entirely on the past to predict the future. The task of a forecaster is to find a link between today and future values of some variable. What makes the process of finding an adquate description of past economic relationships so difficult, is the constant evolution of the economy. In the 1920s, the birth of economic statistics in the US gave rise to large-scale statistical macroeconomic models. Researchers aimed at finding the 'true' data generating process (DGP) matching empirical data to theory based on statistical inference. However, these models failed to predict the stagflation of the 1970s, as it ran counter to the empirical finding thitherto of a negative correlation between unemployment and inflation. As a reaction, Lucas (1976) formulates his famous critique. It states that static models of macroeconomic quanitities will sooner or later fail in predicting macroeconomic aggregates as the underlying relationships differ depending on what macroeconomic policy regime is in place. To put it simply: as a reaction to policy changes people will adjust their behaviour – and thus the inner workings of the economy.

The failure of large models triggered the development of a wide range of new forecasting devices. In particular, small adaptive time-series models solely based on the target variable's own past as, e.g., the approaches developed by Holt (1960) came to widespread use in economic forecasting. However, the idea of finding a 'true' model was not discarded. As Chatfield (1996) points out, a forecaster would typically entertain a family of possible models selecting the best one according to its in-sample fit based on diagnostic tools such as the autocorrelation function. This has been widely criticized, as the same data is used to select the models, make inference, and compute predictions (Chatfield, 1996). The uncertainty related to the model selection process was largely ignored.

When it comes to the creation of a good predictive device it is crucial to know as much about change, as possible. This includes the knowledge about its form and timing. Then, in the best of all cases, change can be identified, anticipated and incorporated in the model. Elliott (2005) points out that the great diversity of the ways a model can be non-constant poses a huge challenge in implementing this task. Major differences concern the frequency and the form of change. A widespread notion of thinking about change is connected to the business cycle. Goldfeld and Quandt (1973) introduce the Markov-switching (MS) regression model to the economic literature. In its basic form it is designed such that the coefficient vector abruptly takes on different values in different time periods or regimes. Those regimes are mostly associated with times of economic expansion and recession. In contrast, Cooley and Prescott (1973) design the stochastically time-varying parameters (TVP) model to deal with slow but constant change. An economic issue featuring this kind of evolutionary process is the expectation formation, which is thought to be an on-going and smoothly developing process. Although the economic forecasting literature contains instances in which forecasts are improved by allowing for specific types of non-linearity, this has the effect of dramatically increasing the dimensionality of the models. Furthermore, as with every forecasting tool, non-linear models need to identify the pattern of change to incorporate in the model. Clements and Hendry (1998) concentrate on change that could not be anticipated given the historical data up to the point the forecast is made. Comparing several forms of structural change, they point out that such a structural break in form of a permanent shift in the parameter vector of a model, will have the most detrimental effect on forecasting accuracy.

Not all approaches dealing with change concentrate exclusively on the first moment. Frequently it is found that variances of economic variables evolve over time, showing a persistent pattern: large (small) shocks tend to be followed by large (small) shocks of either sign, a phenomenon termed volatility clustering. To capture this stylized fact Engle (1982) presents the autoregressive conditional heteroscedasticity (ARCH) approach, modelling the conditional variance such that it depends on elements in the information set in an autoregressive manner. It has been found already in the first empirical application of ARCH to UK inflation that a large number of parameters is required in the conditional variance function. The more parameters estimated, the more complicated inequalitiy restrictions need to be imposed to ensure a positive finite unconditional variance. Bollerslev (1986) proposes the generalized ARCH (GARCH) model that parsimonously paramterizes ARCH models in a manner analogous to the extension from AR to ARMA models in traditional times series. In the following, a rich strand of literature emerged modifying the basic set up to capture additional empirical features. Extensions include asymmetries in the volatility, mean effects of the conditional volatility, and the extension to multivariate ARCH models (see Bera and Higgins, 1993, for a review). Given huge jumps in the variance of economic time series that can not be reconciled with the ARCH framework (Dueker, 1997), researchers quickly realized the possibility of change in the structure of the ARCH models themselves. In their modelling they recurred to the approaches used in the modelling of shifts in the mean of economic time-series. Hamilton and Susmel (1994), e.g., model change in the conditional variance parameters in form of an MS process.

The identification and prediction of structural shift is complicated by the particularities of economic data. Revisions due to incoming information, methodological innovations and statistical reforms are rather the rule than the exception. Thereby, changes to the data are frequently quite substantial.

Furthermore, outliers, i.e., rare strong irregularities, are a common feature (Tsay 1988). Thus, an unusual observation in the last period does not necessarily mean that a structural shift has taken place. Finally, not only the coefficient estimates of a model, but also the predictors may change. A major reason for this is the change in the structure of the economy. The IT-revolution, e.g., profoundly transformed the aspect of office work and gave rise to a new and important sector of the economy. Relatively high labor costs in developed countries led to a large scale reduction of manufacturing industries. Thus, economic change may imply that certain economic variables cease to be useful while others become good predictors. All these features induce huge uncertainty to the detection and especially to the monitoring of structural change. As a consequence, the date of a sudden and substantial break can not be pinned on one concise date – not even ex-post, let alone ex-ante. Frequently, the candidate dates stretch over several years.

Given the huge challenges associated with the incorporation of change in the models as such, one strand of literature has focused on the robustness of existing forecasting tools to change. This work demonstrated that simple adaptive models like the naive forecast, i.e., simply using the last observation of the variable to be predicted as a forecast, frequently outperform more sophisticated alternatives. Furthermore, Clements and Hendry (2006) among others, show that simple adjustments help to make models less sensitive. They suggest to difference the data prior to estimation. This reduces the effect of a break to a 'blib', i.e., limiting its detrimental effect to one single forecast. Furthermore, they lend theoretical justification to intercept corrections (IC), a technique frequently used by the applied forecaster. The concept of IC makes use of last period's forecast errors to adapt models to change.

When it comes to forecasting, not every structural change markedly increases forecast error variances. If change is gradual it might well be that it only marginally affects forecast accuracy. Given the uncertainty generally associated with economic forecasts, the deterioration caused by the shift might be negligible. Furthermore, changes do not necessarily have to be detrimental, at all. This is the case if the relationships of the variables under consideration remain unchanged, despite of a structural change that simultaneously affects both of them. In view of this, Clements and Hendry (1998) introduce the concept of forecast failure, defined as a significant deterioration in forecast performance relative to the anticipated outcome based on an earlier fit. Giacomini and Rossi (2005) make this idea workable. Their forecast break down test compares in- and out-of-sample performance of predictive models to identify and predict forecast breakdowns. It brings together the opposing views of in- and out-of-sample accuracy measures, and gives an indicator of change that is free of any concrete definition of the form of the shift.

In the late 1970s the effort to find the econometric model or class that dominates alternative approaches in a wide range of settings led to the creation of forecast competitions. Their aim is to mimic a realistic forecast situation iteratively testing many candidate models in a horse race. The most prominent exponents are the so-called M-competions, whose outcome is summarized in Makridakis and Hibon (2000). There have been three rounds of them, so far. Each time – despite of the inclusion of more recently developed, highly sophisticated methods – the basic results have remained unchanged. One of the central results was that combinations of many models rather than single forecasting devices outperform their competitors on average. In a more recent analysis, Marcellino (2004) confirmes this finding, comparing 58 single methods in a dataset containing 500 European time series. He points out that more disaggregate analysis demonstrates that single non-linear models can outperform pooled forecasts for several series. However, they perform rather badly for other series so that on average their performance is not as good as that of combined forecasts.

This finding is calling into question the search for the 'true model'. However, what is the rationale for the combined forecasts to outperform the alternatives? Combined forecasts are frequently compared to diversification in asset management (Markowitz, 1952), where the money is spread over a portfolio of stocks rather than on one asset alone, thus minimizing the risk associated with the investment: it is highly unlikely that all stocks in the portfolio have negative returns at the same time. As a consequence, the return of the portfolio will be less volatile then the individual assets alone as huge outliers in both directions will compensate each other. Translated into terms of economic forecasting, combination is a way of adressing model uncertainty and structural change. When the true model either does not exist or can not be identified, it is reasonable to consider a set of likely models. Furthermore, structural change may affect different models in different ways. Positive and negative forecast errors potentially compensate each other so that pooling serves as a hedge against large forecast errors. Under benign conditions, combination can produce even better results than the best individual model alone (Bates and Granger, 1969).

The simplest and hard to beat way of combining a set of forecasts is taking the average or the median. Still, given that a researcher has information on the past performance of the individual models, it seems unreasonable to ignore this knowledge when designing the weights of the combined forecast. The optimal combination, exploiting all possibilities to off-set forecast errors, is based on the covariance matrix of the forecast errors. However, the estimation of the covariance matrix involves the estimation of a huge number of coefficients. Thus, the assumption of uncorrelatedness, i.e., disregarding the off-diagonal elements of the covariance matrix frequently leads to far better results (Timmermann, 2006). Then, the task reduces to weighting according to some kind of accuracy measure, where good models receive a higher weight than less accurate ones. Some controversy has emerged regarding the type of accuracy measure to be used. Proponents of in-sample measures such as the mean squared error of the estimation, point out that their approach best reflects the fit of the respective model to the data up to the forecast origin and thus optimally uses the available information. However, most of the research has concentrated on some form of aggregated past forecast error, such as the mean squared forecast error (MSE). This is due to the frequently found difference between in- and out-of-sample predictive accuracy (Goyal and Welch, 2007). Argueably, out-of-sample measures implicitly reflect the models capability of dealing with structural shifts.

Basing the weighting schemes on accuracy measures will in the presence of shifts in the economy ultimately lead to a shift in relative performances and thus the weights. Numerous approaches have been proposed to make the weighting schemes more adaptable. They include minimizing the impact of obselete information discounting past forecast errors, or discarding outdated information all together. Some researchers have modelled the changing weights in form of a TVP, an MS, or a combination of the two (Aiolfi and Timmermann, 2006). Granger and Ramanathan (1984) have used autoregressive conditional heteroscedasticity (ARCH) effects, an omnipresent feature of economic data, to predict forecast error variance and employ them to update the weighting scheme. Winkler and Clemen (1992) show that the possible gain of including a model in the combination scheme might be off-set or even lead to a loss in forecast accuracy, if the additional estimation uncertainty with respect to the weights is too big. For changing weights schemes this is particularly relevant. They either reduce the information set the estimation of the weight sequence is based on or induce additional uncertainty due to the need to identify the type of change. This leads to the central trade-off this dissertation will deal with: up-to-datedness versus precision.

The theory of combined forecasts demonstrates that large relative differences in the forecast error variance of two models tend to reduce the gains of combination. Thus, it is intuitively convincing to exclude those models from combining that, given their past performance, will perform poorly in the future, as well. This approach goes under the heading of trimming. In this line, Granger and Jeon (2004) advocate the use of models of 'similar quality'. Aiolfi and Timmermann (2006) propose the clustering of the models into groups according to the MSE, using only the best models for the combined forecast. The most radical approach is the so called predictive least squares (PLS, Stock and Watson, 1999), where the model that has produced the best forecast up to the forecast date is given a weight of one, and the other models zero. While most of these approaches have resulted in an improvement of the combined forecast, Timmermann (2006) shows that the selection of those models to be given zero weight has been rather ad-hoc and data dependent. As Aiolfi and Timmermann (2006) point out, for these approaches to work, it is necessary that the relative past performance is persistent, i.e., that the worst models remain the worst models over all periods. However, Stock and Watson (1996), analyzing a broad set of international macoreconomic forecasts demonstrate that the ranks of a huge set of different forecasts vary over two arbitrary sub-periods of the analysis. In view of the widespread occurance of structural shifts, the success of trimming will crucially depend on the time period under study.

This dissertation takes a fresh perspective on combined forecasts in the presence of structural change. It proposes a novel refinement technique, the Forecast Breakdown Preselection (FBP) that can be interpreted as a forward looking alternative to standard trimming schemes. Based on the dynamics of ARCH effects, the methodology of Giacomini and Rossi (2005) is employed to detect those models that are likely to face a forecast break down in the next forecast situation. This allows to directly address the trade-off between up-to-datedness and precision sorting out unreliable models. In contrast to standard trimming approaches it is free of any ad-hoc decision making. Rather than basing selection of candidate models on past and possibly out-dated information, it takes the decision employing a dynamic indicator for the period where it matters most: the forecast horizon. Furthermore, it lends additional flexibility to the weight sequence. Both, simple and adaptive versions of past performance weighting schemes need some time to react to drastic changes in the relative performance. In contrast, FBP can immediately set the weights of models that abruptly deteriorate to zero. Furthermore, as it merely filters a given sample of individual forecasts, the resulting subsample can be employed in any scheme of choice.

FBP is developed along the lines of an empirical experiment iteratively predicting UK inflation. This is accompanied by a short theoretical exposition of the workings of combined forecasts and the adaptive schemes most frequently found in the literature. A simulation analysis compares the performance of the latter and explores how much information is actually needed for those schemes to make sense. Existing studies Stock and Watson (2004), test combination schemes on series that have been found to feature one or more structural breaks without dating them. In contrast, this dissertation takes a look at the performance of combined forecasts at an identified break in the target series using the methodology of Bai and Perron (2003b). Thereby, the focus is laid on the individual forecasts. More specifically, periods of increased forecast error variances are analyzed using visualization tools, ARCH tests, and the methodology of Giacomini and Rossi (2005), and set into relation to the structural break analysis. Building on these empirical features, FBP is introduced and its functioning is shown in a second simulation study based on the empirical experiment. Then, FBP's usefulness for the prediction of UK inflation is explored in detail and compared to an alternative refinement technique. Furthermore, FBP is tested for the prediction of other economically relevant variables in the current data set.

This dissertation introduces the concept of forecast breakdowns to the context of combined forecasts to develop a novel refinement technique that is flexible, free of ad-hoc decision making, and forward-looking. The study is the first to apply the analysis of forecast breakdowns to a large number of models. This allows for new insights into the reliability of forecast devices around an identified break. Furthermore, the study is the first to consider such a wide range of different combination techniques to test the practicability of methods making combined forecast more adaptable. Finally, it extends traditional analyses of combined forecasts in that it considers several types of non-linearties, ARCH effects, breaks, and phases of high forecast uncertainty, in one study.

The following chapter gives a short overview of the basic theory of forecast combination. Chapter three describes the data and the empirical approach. Chapter four outlines the analysis of structural breaks and presents results for the series used. The fifth chapter relates ARCH effects to combined forecasts. The next chapter presents and compares adaptive combination schemes. Chapter seven describes the forecast breakdown analysis and introduces the FBP. Chapter eight presents a simulation analysis to demonstrate the usefulness of the approach. Chapter nine compares empirical results of the combination schemes with and without the application of FBP. The last chapter is a conclusion.

Chapter 2

Theory of combined forecasts

2.1 Combination of a pair of forecasts

This chapter presents some theoretical results on the optimal combination (OPT), the inverse MSE weighting scheme (INV), and the equal-weights-scheme (EW). These three are the most widely used combination approaches. OPT will not be used in the applied part of this dissertation as it provides rather weak results in empirical studies (Timmermann, 2006). However, it is very useful for illustrative purposes. In the following, it is demonstrated for the simple case of two forecasts that the interaction of relative performances of the models and their correlation play a crucial role for the success of the combination approach. For simplicity the analysis assumes the DGP to be time-invariant and known.

Past performance schemes like INV and OPT learn from forecast errors. Thus, they give better results when compared to EW the higher the differences of forecast accuracy of the constituent models. Moreover, past performance schemes tend to handle poor performing models more efficiently the worse they are. In the extreme they are attributed a weight close to zero. This self-selection raises some doubts on the necessity of the standard trimming approach that filters out some arbitrary percentiles of very inaccurate models.

However, disregarding information on past performances might sometimes be a reasonable approach. In applied work weight sequences need to be estimated. In particular, if the estimation is complicated by structural changes EW frequently outperforms more sophisticated methods.

The more negative the correlation of the errors and the smaller the difference of accuracy, the better pooling generally works. Under benign conditions, pooled forecasts can even beat the best individual model. However, even if the covariance of the errors does not allow for combination to beat all alternatives, it may still be a sensible alternative. If knowledge of the performance of individual models is limited due to short historical information or uncertainty related to shifts in the economy, pooling poses a valuable approach.

Bates and Granger (1969) transfering Markowitz' (1952) diversification argument to the field of forecasting demonstrate the theoratical merits of combination when the covariance of the individual forecast errors is known. They consider the case of two single competing point forecasts, $f_{1,t}$ and $f_{2,t}$, of some quantity y_t, derived h periods ago. As the single forecasts are assumed to be unconditionally unbiased, the forecast errors

$$e_{i,t} = y_t - f_{i,t} \tag{2.1}$$

are normally distributed

$$e_{i,t} \sim N\left(0, \sigma_i^2\right), \tag{2.2}$$

where $\sigma_i^2 = var\left(e_{i,t}\right)$, $\sigma_{1,2} = \rho\sigma_1\sigma_2$ denotes the covariance between $e_{1,t}$ and $e_{1,t}$, and ρ is their correlation. The linear combination

$$c_t = kf_{1t} + (1-k)f_{2t} \tag{2.3}$$

of the two forecasts, which is a weighted average, is then unbiased in the same sense, so that the forecast error of the combined forecast

$$e_{c,t} = ke_{1t} + (1-k)e_{2t} \tag{2.4}$$

has zero mean and variance

$$\sigma_c^2 = k^2\sigma_1^2 + (1-k)^2\sigma_2^2 + 2k(1-k)\sigma_{1,2}. \tag{2.5}$$

The optimal value k_{opt} can be derived by minimizing the error variance σ_c^2, differentiating Equation (2.5) with respect to k and solving the first order condition:

$$k_{opt} = \frac{\sigma_2^2 - \sigma_{1,2}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{1,2}} \tag{2.6}$$

$$1 - k_{opt} = \frac{\sigma_1^2 - \sigma_{1,2}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{1,2}}. \tag{2.7}$$

Thus, individual forecasts that yield lower forecast error variances (are more accurate) are assigned a higher weight in the combination process. In the extreme case of $\sigma_2^2 \to \infty$, $k_{opt} \to 1$. The resulting forecast error variance of the optimally combined forecast is

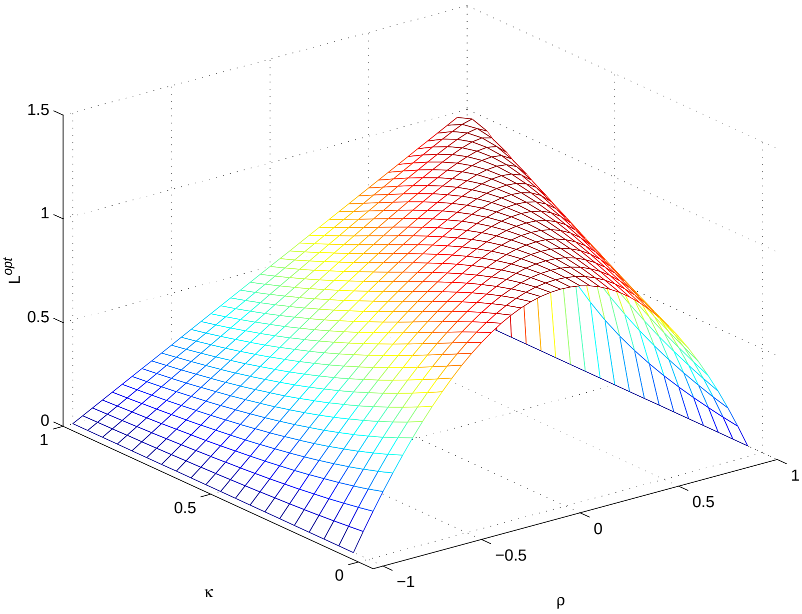
$$\sigma_{opt}^2 = \frac{\sigma_1^2\sigma_2^2\,(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}. \tag{2.8}$$

Let us define $\kappa = \sigma_2/\sigma_1$ with $\kappa < 1$, so that σ_2 is the best individual forecast. The loss $\mathrm{L}^{opt} = \frac{\sigma_{opt}^2}{\sigma_2^2}$ compares σ_{opt}^2 to the best individual forecast; For $\mathrm{L}^{opt} = 1$, the combined forecast is as good as the best model, for $\mathrm{L}^{opt} < 1$ it is better, and for $\mathrm{L}^{opt} > 1$ it is worse. Expressing L^{opt} in terms of $\rho_{1,2}$ and κ gives

$$\mathrm{L}^{opt} = \frac{1-\rho^2}{1+\kappa^2 - 2\rho\kappa}. \tag{2.9}$$

To highlight the dependence of the combined forecasts on the covariance, Figure 2.1 plots L^{opt} for different values of κ and ρ. If the two constituent forecasts are equally accurate ($\kappa = 1$), it follows that the combined forecast will deteriorate, the higher ρ. If the individual forecast variances differ ($\kappa \neq 1$), the combined forecast will deteriorate with an increasing ρ relative to f_2 as long as $\rho < \kappa$ and improve upon f_2 with a higher ρ if $\rho > \kappa$. Only if $\kappa = \rho$, the forecast are equally accurate ($\mathrm{L}^{opt} = 1$). Thus, the combined forecast is always at least as good as the best individual forecast, and can lead to a

Figure 2.1: L^{opt} as a function of κ and ρ

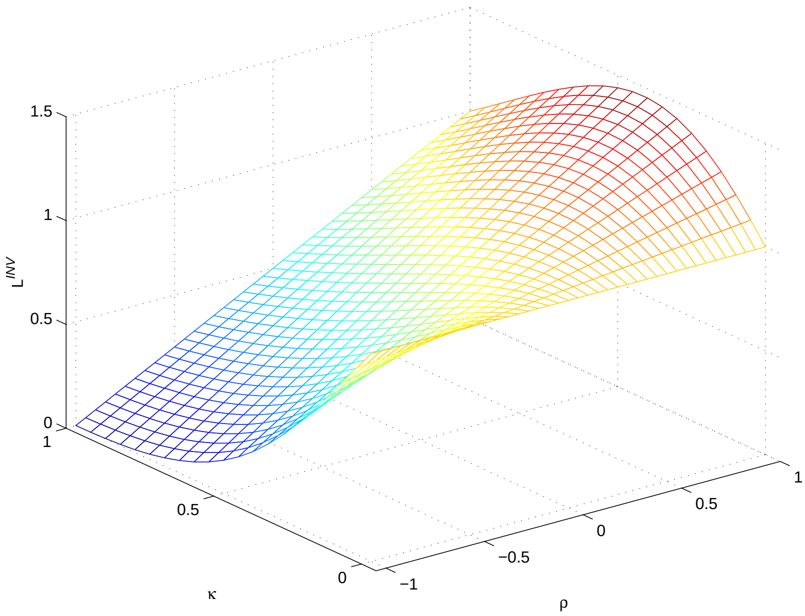


considerable improvement over f_2.

In finite samples, however, especially when the sample size is small relative to the number of candidate forecasts, the estimation of the off-diagonal elements of the variance-covariance matrix has proven to be an especially demanding task. Furthermore, the estimated weights can become very instable, considerably deteriorating the forecast performance of the pooled forecasts (Winkler and Clemen, 1992). A pragmatic solution proposed already by Bates and Granger (1969) and sucessfully applied to a wide range of forecasts by Stock and Watson (2004) is to concentrate on the variances of the forecast errors. Assuming the mean of the individual forecast errors to be zero, the combination scheme thus weights the individual forecasts inversely to their relative mean squared forecast error (MSE) giving INV. In the case of combining two single forecasts k_{opt} reduces to k_{inv} giving:

$$k_{inv} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{2.10}$$

Figure 2.2: L ^{inv} as a function of κ and ρ



$$1 - k_{inv} = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \tag{2.11}$$

The resulting forecast error variance is then

$$\sigma_{inv}^2 = \frac{\sigma_1^2 \sigma_2^2 \left(\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2\right)}{\left(\sigma_1^2 + \sigma_2^2\right)^2}. \tag{2.12}$$

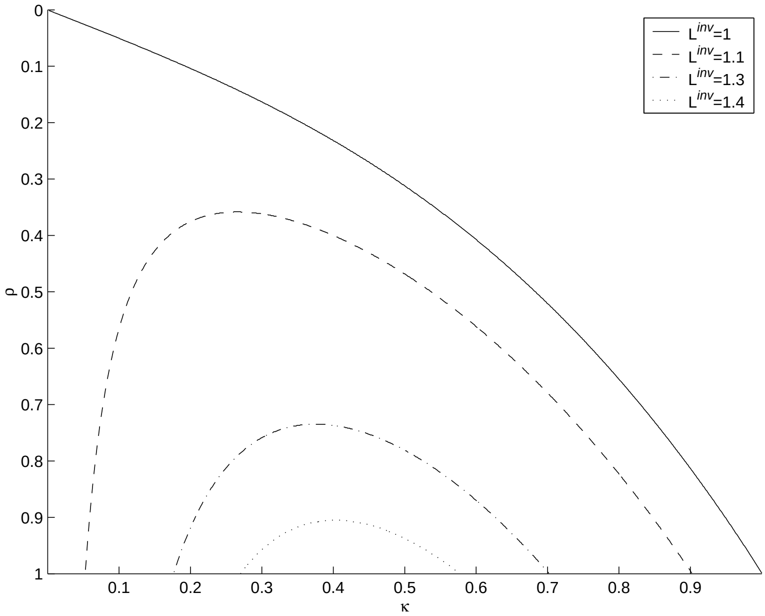
The loss over the best individual model is $\mathrm{L}^{inv} = \frac{\sigma_{inv}^2}{\sigma_2^2}$; using κ and ρ will then give

$$\mathrm{L}^{inv} = \frac{1 + \kappa^2 + 2\rho\kappa}{(1 + \kappa^2)^2}. \tag{2.13}$$

Figure 2.2 gives L^{inv} for different values of κ and ρ. As in the case of the OPT, if $\kappa = 1$ the ratio is an increasing function of ρ, and it is smaller than one, except if $\rho = 1$ where it takes the value of unity.

However, ignoring the correlation of the forecast errors comes at a cost. When ρ is high and κ is low, the combined forecast is no longer necessarily better or at least as good as the best individual model. Figure 2.2 is divided into

Figure 2.3: Cross section of L^{inv}



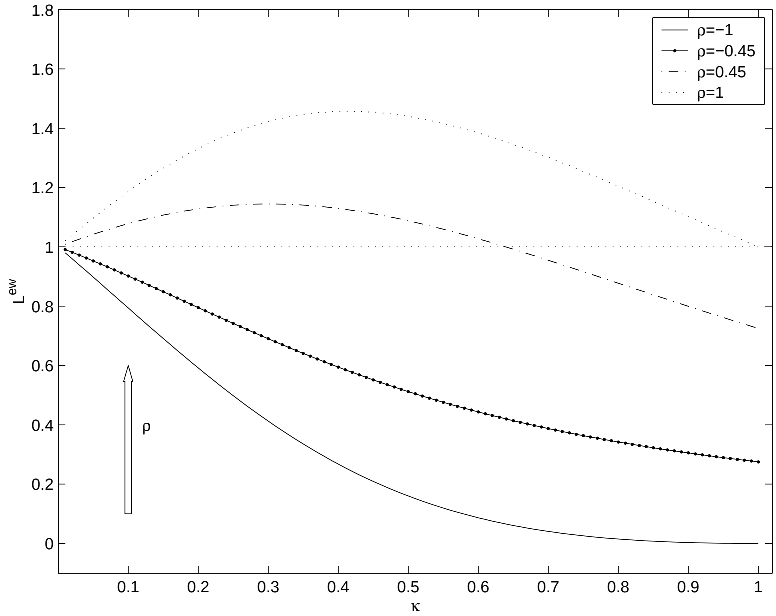
two zones, confined by the condition $\rho = \frac{\kappa+\kappa^3}{2}$. When $\rho < \frac{\kappa+\kappa^3}{2}$, $\mathrm{L}^{inv} < 1$ holds, and the higher κ and the lower ρ the better INV works. I.e., in this zone, the combination improves, the smaller the differences in the variances and the more negative the correlation of the individual models.

Figure 2.3 gives the contour lines of the part of Figure 2.2 where INV is worse or equally good than f_2 ($\rho \geq \frac{\kappa+\kappa^3}{2}$ and $\mathrm{L}^{inv} \geq 1$). On the outer line, where $\rho = \frac{\kappa+\kappa^3}{2}$, and the y-axis, where $\kappa \to 0$, INV is (approximately) as accurate as f_2.[1] Every combination of ρ and κ that is to the right of y- axis and below the line where $\rho = \frac{\kappa+\kappa^3}{2}$, INV performs worse than f_2. The most detrimental combinations of ρ and κ to INV are enclosed by the inner line, where $\mathrm{L}^{inv} = 1.4$, and the x-axis. The maximum of 1.46 is reached when $\rho = 1$ and $\kappa \approx 0.41$.

For $\kappa \neq 1$ the first derivative of $\mathrm{L}^{inv}, \frac{\delta \mathrm{L}^{inv}}{\delta \kappa}$, is positive if $\rho > \frac{\kappa^3+\kappa}{(1-3\kappa^2)}$, and negative if $\rho < \frac{\kappa^3+\kappa}{(1-3\kappa^2)}$. This is a counterintuitive result. Suppose the researcher wants to combine a given forecast with another one and has two alternatives

[1] $\kappa = 0$ is not defined as this would imply f_2 being infinitely more accurate than f_1.

Figure 2.4: L^{inv} as a function of κ and some values of ρ



at his disposal. Then, the combination with the less accurate of the two candidates may lead to better results then the combination with the better one.

Figure 2.4 highlights this point, plotting $\mathrm{L}^{inv}(\kappa)$ for different values of ρ, where ρ is more positive the higher the line. $\kappa = 1$ on the right indicates that the models are equally accurate, and the smaller κ the higher the differences. For $\rho = -1$ and $\rho = -0.45$ moving from the right to the left the loss is monotoneously increasing. For the positive correlations, that is, for $\rho = 1$ and $\rho = 0.45$ firstly this holds, as well. However, the higher the differences, e.g., when $\kappa < 0.2$, a further increase of the differences leads to a lower loss.

This weakens the rationale behind the standard trimming approach as the correlation of forecast errors is mostly positive (Clemen and Winkler, 1986) so that the counterintuitive result applies in practical work. Trimming prevents the inclusion of very inaccurate models in the combination process. However, the worse those models the more efficiently past performance schemes like INV tend to deal with them. Thus in view of the difficulties in measuring past forecast accuracy that will be outlined in chapter 6 dropping some arbitrary percentiles could do more harm than good.

Disregarding all information of the covariance matrix of the forecast errors taking the average, thus attributing all models the same weight, yields EW. The forecast error variance of INV given in Equation (2.12) simplifies to:

$$\sigma_{ew}^2 = \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\sigma_1\sigma_2\rho_{1,2}. \tag{2.14}$$

The loss L^{ew} of σ_{ew}^2 over f_2 is then given as

$$\mathrm{L}^{ew} = \frac{1+\kappa^2+2\kappa\rho}{4\kappa^2}. \tag{2.15}$$

In contrast to OPT and INV, L^{ew} is always a negative function of κ, irrespective of the value of ρ; the first derivative of Equation (2.15) with respect to κ is positive if $\rho > \frac{\kappa}{\kappa^2-2\kappa}$, which always holds, as $0 < \kappa \leq 1$. Again simplification leads to a loss in forecast accuracy. Figure B.1 in the appendix plotting L^{ew} as a function of ρ and κ demonstrates that the extent of the relative loss can be very much higher than for INV, especially if $\kappa \to 0$. Still, if condition $\rho < \frac{3\kappa^2-1}{2\kappa}$ holds EW outperforms even the best individual model.

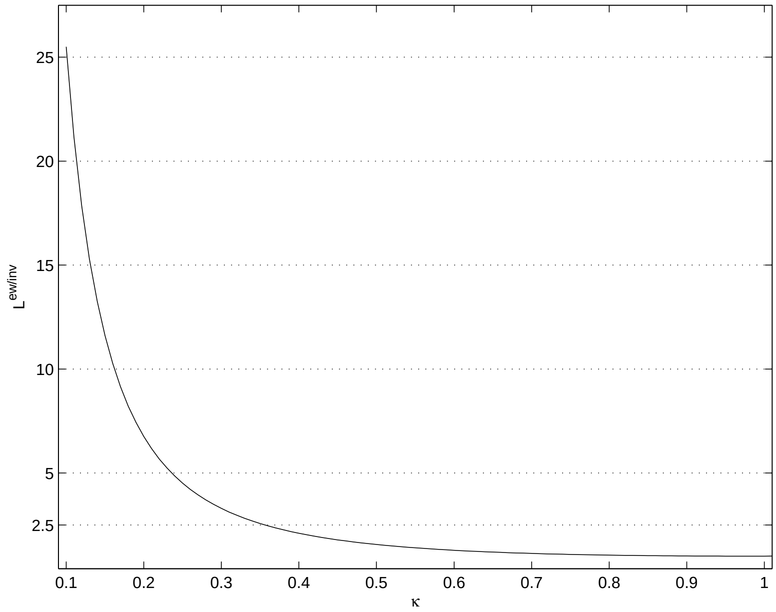
As both, INV and EW, disregard the off-diagonal elements of the covariance matrix of the forecast errors, the relative performance of the two approaches depends on κ alone. To evaluate the relative loss as a result of disregarding the diagonal elements consider the ratio:

$$\mathrm{L}^{ew/inv} = \frac{\sigma_{EW}^2}{\sigma_{INV}^2} = \frac{(1+\kappa^2)^2}{4\kappa^2} \tag{2.16}$$

Figure 2.5 plots $\mathrm{L}^{ew/inv}$ for some values of κ.[2] $\mathrm{L}^{ew/inv}$ is a monotenously decreasing function of κ, with the minimum being one. While the relative loss is relatively little up to $\kappa = 0.3$, where $\mathrm{L}^{ew/inv} = 1,8125$, it steeply increases with $\kappa \to 0$. Thus, using EW instead of INV is particularly detrimental, if the relative differences of forecast accuracy of the different models is very high. INV will attribute smaller weights, while EW can not adapt.

[2]For ease of presentation, smaller values of κ that give very high values of $\mathrm{L}^{ew/inv}$ are not plotted.

Figure 2.5: $\mathrm{L}^{ew/inv}$ as a function of κ



2.2 Average forecast given instable processes

Up to now the analysis was confined to constant forecast error covariances. However, as laid out in the introduction, the structure of the errors will vary over time as the data generating process is changing, i.e., the economy evolves. While the merits of combination in this setting are confirmed empirically in the seminal article of Bates and Granger (1969), theoretical proofs have emerged rather recently. With an eye to the specificaton of the simulation experiment in chapter 8 this section aims at highlighting the central issues involved: the need for the modeling of the misspecification of the individual models, the type of structural change, and the importance of the relative size of the break.

In the following, we will concentrate on two of the more eminent analyzes. Complexity is increasing sharply when structural change is considered. In order not to aggravate the situation any further, most of the research excluded weight optimization issues concentrating on EW.

2.2.1 Shift in the DGP

Aiolfi and Timmermann (2006) present a model that highlights potential gains in the case of a switch in the DGP between two states F_1 and F_2, which, e.g., could be interpreted in the context of the business cycle as phases of expansion and recession. Misspecification of the individual models is given in that they only recur to the information on one of the phases:

$$y_t = S_t F_{1t} + (1 - S_t) F_{2t} + \epsilon_{yt} \tag{2.17}$$

$$f_{1t} = F_{1t} + \epsilon_{1t}, \tag{2.18}$$

$$f_{2t} = F_{2t} + \epsilon_{2t}, \tag{2.19}$$

where all variables are assumed normal with factors $F_{it} \sim N(\mu_i, \sigma^2_{F_i})$, $i = 1, 2$, and the disturbance terms $\epsilon_{it} \sim N(0, \sigma^2_{\epsilon_i})$ and $\epsilon_{yt} \sim N(0, \sigma^2_{\epsilon_y})$. Furthermore, while the disturbance terms are mutually uncorrelated and uncorrelated with the factors, the two factors themselves may be correlated. The state transition probabilities are constant $P(S_t = 1) = p$ and $P(S_t = 0) = 1 - p$. The forecast errors using only the information of one of the two factors are $e_{it} = y_t - f_{it}$ with variance $\sigma^2_{e_i}$.

The authors show that the population MSE of the equal-weighted combined forecast will be lower than the population MSE of the best model if:

$$\frac{1}{3}(\frac{p}{1-p})^2 \frac{1+\psi_2}{1+\psi_1} < \frac{\sigma^2_{e_2}}{\sigma^2_{e_1}} < 3(\frac{p}{1-p})^2 \frac{1+\psi_2}{1+\psi_1},$$

where $\psi_i = \sigma^2_{\epsilon_i}/\sigma^2_{F_i}$ are the noise-to-signal ratios for forecasts one and two. Imposing equal probabilities for the two states, i.e., $p = 0.5$ and equal noise-to-signal ratios, i.e., $\psi_1 = \psi_2$ simplifies the expression to:

$$\frac{1}{3} < \frac{\sigma^2_{e_2}}{\sigma^2_{e_1}} < 3,$$

which illustrates that hedging against breaks using combined forecasts makes sense for a wide range of relative factor variances.

2.2.2 Extraneous breaks

Hendry and Clements (2004) use a very similar set-up to demonstrate potential gains of combined forecasts in the presence of abrupt changes in the mean of the DGP analogous to the change analyzed by the structural break literature presented in chapter 4. They consider a wide range of designs for the shift whereby it is assumed to occur after the estimation period so that parameter estimates are not affected. The individual models are misspecified as each of them only uses one of the variables that form the target variable resulting in an omitted variable bias (OVB) . In the following, we will consider the situation of a break in the intercept of a single variable in the data generating process (extraneous break).

As in Aiolfi and Timmermann (2004) the target variable y_t is given as the sum of two constituent processes:

$$y_t = \beta_1' x_{1,t-1} + \beta_2' x_{2,t-1} + \epsilon_{yt}, \tag{2.20}$$

with $\epsilon_{yt} \sim N(0, \sigma^2_{\epsilon_y})$ and independent of $x_{1,t-1}$ and $x_{2,t-1}$. Now, two investigators ignoring the true DGP fit separate models using only part of the information:

$$y_t = a' w_t + \epsilon_{1t} \tag{2.21}$$

and

$$y_t = b' z_t + \epsilon_{2tt} \tag{2.22}$$

where $x_{1,t-1}$ and $x_{2,t-1}$ are replaced with w_t and z_t to simplify presentation. Moreover,

$$\begin{pmatrix} w_t \\ z_t \end{pmatrix} = N_n \left(\begin{pmatrix} \phi_{z,t} \\ \phi_{w,t} \end{pmatrix}, \begin{pmatrix} \sigma^2_w & \sigma^2_{wz} \\ \sigma^2_{wz} & \sigma^2_z \end{pmatrix} \right),$$

where $\phi_{z,t} = \phi_{w,t} = 0$ for $t = 1, ..., T$. The estimation errors ϵ_{it} reflect the ommitted variable bias.[3] In $T+1$ the z-process shifts $\phi_{z,T+1} = \mu_z$. Now, the two investigators predict y_{T+1} using the data up to T. The analysis thereby abstracts of sampling variability in the coefficients a and b. The 1-horizon

[3] For a detailed exposition see the original paper.

forecast using Equation (2.21) is $\hat{y}_{T+1} = \hat{a}_0 + \hat{a}_1' w_{T+1}$ giving forecast error $\hat{u}_{T+1} = y_{T+1} - \hat{y}_{T+1}$. Analogously, the other model's forecast is $\tilde{y}_{T+1} = \hat{b}_0 + \hat{b}_1' z_{T+1}$ resulting in forecast error $\tilde{\nu}_{T+1} = y_{T+1} - \tilde{y}_{T+1}$. Making some weak assumptions to reduce the notational burden, the authors show that combined forecast outperforms both individual forecasts if

$$\frac{1}{3}\left(1 + \frac{\mu_z^2}{\sigma_z^2}\right) < \rho^2 < 3\left(1 + \frac{\mu_z^2}{\sigma_z^2}\right).$$

Thus, combining forecasts can offset the instability in the individual forecasts and in effect serves as an intercept correction. Moreover, the result depends on the size of the break relative to the variances of the underlying processes z_t and w_t, and their correlation.

Chapter 3

Empirical set-up

3.1 The models and the experiment

How does pooling of forecasts perform in practice? To answer this question the major point of reference in the following chapters will be an empirical experiment. It demonstrates what an applied forecaster trying to predict UK inflation on a monthly basis could have done in the past. Table 3.1 illustrates the approach. Due to data restrictions the points in time forecast situations are simulated, the forecast origins, stretch from December 1984 to June 1999. As the actual values of inflation are known by now, we can evaluate the performance of the models used. In real-time forecasting, only the data up

Table 3.1: Pseudo-out-of-sample experiment

Estimate	Forecast ahead			
	$k = 1$	$k = 3$	$k = 6$	$k = 12$
1948:6 $\longrightarrow$ 1984:12	1985:1	1985:3	1985:6	1985:12
1948:6 $\longrightarrow$ 1985:1	1985:2	1985:4	1985:7	1986:1
1948:6 $\longrightarrow$ 1985:2	1985:3	1985:5	1985:8	1986:2
1948:6 $\longrightarrow$ 1985:3	1985:4	1985:6	1985:9	1986:3
...				
1948:6 $\longrightarrow$ 1999:6	1999:7	1999:9	1999:12	2000:6

to the forecast origin is available. The precision of the estimated coefficients is lower when compared to estimations using the complete information set ex-post. Thus, it is common practice to implement so-called pseudo-out-of-sample forecasts, restricting estimation at each iteration to the information set that would have been available in the past. As the results could differ according to the distance in time the forecasts bridge, the analysis considers four different forecast horizons typically found in applied work: 1-month, 3-month, 6-month, and 12-month horizons.

Suppose the forecaster fits a large number of single OLS models using lagged values of exogenous variables and inflation as regressors x_{it} giving

$$INFL_t = \alpha + \sum_{i=1}^{n} \sum_{j=h+1}^{p} \beta_{ij} L^j x_{it} + u_t, \tag{3.1}$$

where u_t is assumed to be independently identically distributed (iid), and α and β_{ij} are the constant respectively coefficients of lag j of regressor i, with $i = 1, ..., n$. Moreover, $t = 1, ..., T$ is the time index and p denotes the number of lags considered. L^j is the lag operator that shifts a variable j periods back in time, e.g., $L^1 x_{it} = x_{it-1}$, $L^2 x_{it} = x_{it-2}$, ..., $L^p x_{it} = x_{it-p}$. Following common practice (Stock and Watson, 2004), the number of lags considered in the regression is selected to minimize the Bayesian information criterion (BIC):

$$\text{BIC} = S + l \ln(T),$$

where l is the number of estimated parameters. Here, l is the number of regressors times the number of lags plus the constant. It minimizes the sum of squared residuals, S. The second term penalizes overfitted models as it increases BIC the higher l. The maximum number of possible lags is set to eight.

Table 3.2: Example of model-building

model	A	B	INFL
1	x		x
2		x	x
3	x	x	x
4			x

Note: an "x" indicates that the variable is considered in the respective model.

The models are formed, using all possible combinations of the regressors plus inflation which is always included. Table 3.2 illustrates this for the case of two regressors A and B. There are three possible combinations including inflation and at least one additional regressor, model one to three, plus the autoregressive model number four. In the experiment, where the highest number of exogenous regressors per model is 16, the number of possible combinations of the variables is $2^{16} + 1$. The number of coefficients to be estimated, $i \times j + 1$, rises quickly, the more lags are considered, limiting the degrees of freedom used in each regression. Thus, the maximum of variables per model is set to three reducing the total number of models to be estimated each iteration and horizon to 121.

A common method of dealing with structural changes is to use only a fixed number of past observations of the data τ. This would, e.g., result in using observations $T - \tau, T - \tau + 1, ..., T$ at forecast origin $t = T$ and $T - \tau + 1, T - \tau + 2, ..., T + 1$ in $t = T + 1$, shifting the window one period forward. The rationale is to facilitate the adaptation of the model parameters to the new (post-change) situation giving more recent observations more weight and discarding obsolete information. However, this comes at the cost of artificially increasing estimation uncertainty. As empirical evidence on the success of the rolling window approach is mixed, the analysis employs all observations up to forecast origin t: the models use observations $1, ..., T$ at forecast origin $t = T$ and $1, ..., T + 1$ in $t = T + 1$.

3.2 The series

There is an ample set of candidate exogenous variables a forecaster can choose of. However, there are some restrictions, a realistic simulation of the forecasting situation has to take into account. Some of the data changes over time, as statistical agencies frequently revise their publications due to new incoming information, canonical changes, or due to the correction of errors made. Thus, a forecaster in $t = T$ may have a different version, or vintage, of the information set spanning observations $1, ..., T$ at his disposal than a forecaster in $t = T + 1$ though the same time period $1, ..., T$ is considered. This might have considerable influence on the models used. Croushore (2006) points out that revisions change the data input, the estimated coefficients, and the model itself (e.g., the number of lags). Usually, statistical agencies only publish the most recent vintage, so that obsolete information sets get lost. However, there are some collections of real-time data, real-time data bases (RTDB), that contain all vintages.

The real-time data considered here cover fundamental variables contained in Egginton, Pick, and Vahey (2002) comprising real industrial production (IP), the total claimant count as a measure of unemployment (U), monetary aggregate (M0), retail sales volume (RS), and average earnings (AE). The variables have been collected out of the printed publications Economic Trends and Financial Statistics of the Office of National Statistics (ONS, formally the Central Statistical Office).

As an example of real-time data, consider Table 3.3 presenting the vintages of IP between December 1984 and June 1985 (in columns) covering the period between October 1984 and April 1985 (in rows). The upper left field contains the value of IP for October 1984 as it has been available in December 1984. In January 1985 the value for October 1984 is revised and the new information for November 1984 becomes available. In February 1985 the two values already published are revised and another observation is added, and so forth. The first measurements of the data are always published with a time lag of two month to the period they refer to. Furthermore, the data are

Table 3.3: Example of real-time data, IP

	Vintage						
Obs.	1984:12	1985:1	1985:2	1985:3	1985:4	1985:5	1985:6
1984:10	0.79	1.28	0.10	0.10	-0.19	-0.29	-0.39
1984:11		-0.19	0.39	0.49	0.19	0.29	0.29
1984:12			0.10	0.48	0.68	0.48	0.48
1985:1				1.35	1.54	0.87	0.48
1985:2					-0.19	0.38	0.38
1985:3						1.90	2.20
1985:4							0.56

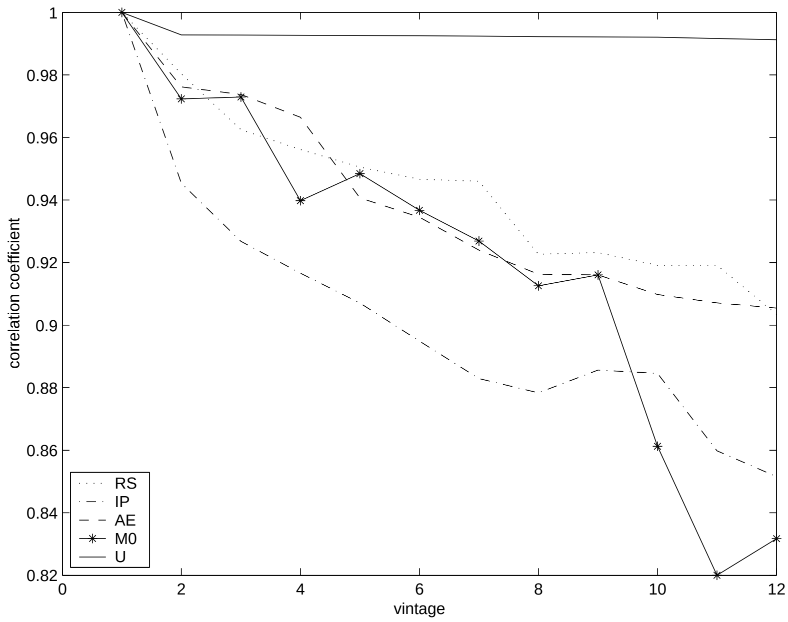
Table 3.4: Extent of revision

Variable	Mean abs. rev.	Mean abs. chg.	Ratio	Min rev.	Max rev.
RS	0.50	1.22	0.41	-1.80	3.15
IP	0.62	1.25	0.49	-2.59	1.63
AE	0.31	0.66	0.47	-1.24	2.40
M0	0.27	0.46	0.58	-1.86	1.05
U	0.25	0.55	0.45	-2.21	0.90

revised considerably, frequently even changing sign.

Table 3.4 gives an impression of the extent of the data revisions considering the difference between the first measurements and the last vintage in June 1999. The information on each variable is arranged in rows. The second column shows the mean of the absolute revisions. The third column shows the mean of the absolute changes of the respective variables over time. As the data are transformed to month-on-month changes, the units of the figures are percentage points. In order to make the revisions interpretable, the fourth column presents an indicator of the relevance of the revision. It sets the mean absolute revisions in relation to the mean absolute changes from one observation to the next measured in the last vintage in June 1999, giving the ratio of the two numbers. The bigger the ratio, the more important

Figure 3.1: Correlation coefficients of the first with the 11 subsequent vintages



are revisions. The last two columns present the minimum and the maximum revision. The ratios range from 0.41 for retail sales to 0.58 for money, meaning that changes due to revisions are about two to three fifth as big as the variation of the data over time. Furthermore, the revisions may even be considerably higher. The maximum revision of average earnings, e.g., is nearly eight times higher than the mean absolute revision.

Figure 3.1 shows that the differences between the vintages materialize quickly. It plots the correlation coefficients of the first vintage with the eleven subsequent vintages. The correlations between the vintages are decreasing sharply for retail sales, average earnings and especially for industrial production and monetary aggregate, the bigger the distance in time. The correlation coefficients of the latter two drop to well below 0.9 within the first year after the first publication. However, the correlation coefficients of unemployement stays close to one. Table 3.5 on the facing page gives the corresponding results for the total revisions. In case of money the correlation coefficient of the first to the last vintage is merely 0.68, whereas the correlation coefficient

Table 3.5: Correlation coefficients of the first and the last vintage

Variable	Corr Coef
RS	0.84
IP	0.69
AE	0.78
M0	0.68
U	0.97

of unemployment is still close to one.

Additional to the fundamental variables, the data set contains several variables that are not subject to revisions, including the target variable, UK inflation rate (INFL), measured as the 12-month percentage change of the Retail Price Index (RPI).[1] A set of financial indicators comprise short-term interest rates, measured as the three-month UK Treasury Bill rate (TB) deflated with RPI, the monthly average of the Financial Times Stock Exchange Index (FTSE), and the yields of the 10-year UK Government benchmark bonds (BD). External indicators cover the end of month values of the exchange rate of the British Pound to the US Dollar (USD), the exchange rate of the British Pound to the Deutschmark (DEM), the Treasury Bill rate of the United States of America (TBUS) deflated with the US inflation rate, and the price of one barrel Brent oil (OIL).[2] Survey data are given in form of the inflation expectations for the next three months of the Industrial Trends Survey of UK manufacturing collected by the Confederation of British Industry (CBI), the Business Climate Indicator (BCI), and the Economic Sentiment Indicator (ESI). The latter two are published by th Directorate General for Economic and Financial Affairs (DG ECFIN). Table 3.6 presents the four different groups of exogenous variables used.

[1] For the time period covered, the RPI was the main measure of prices in the UK and the RPI-inflation rate was the target rate when the Bank of England (BoE) adopted inflation targeting in October 1992. In 2003 the Harmonized Consumer Price Index (HCPI) replaced the RPI-inflation rate as the target of the BoE.

[2] US inflation is measured as the 12-month percentage changes of the US consumer price index (CPI).

Table 3.6: The exogenous variables

Fundamental variables	Financial indicators	External position	Survey data
IP ind. prod.	TB treasury bill	USD exch. rate	BCI business climate
U unemployment	BD bonds	DEM exch. rate	ESI econonmic sent.
AE avg. earnings	FTSE stock market	OIL price	CBI infl. expectations.
RS retail sales			
M0 money			

One particular form of change is given by non-stationarity in form of a unit root. To control for this, all fundamental variables, FTSE, BD, USD, DEM, and OIL have been transformed to month-on-month percentage changes. All variables are tested for a unit root using the standard augmented Dickey-Fuller (ADF) procedure (Dickey and Fuller (1979)). The ADF test is carried out by estimating:

$$\Delta y_t = \alpha y_{t-1} + \sum_{j=1}^{p} \beta_j L^j \Delta y_{t-j} + \nu_t, \tag{3.2}$$

where the null hypothesis is $H_0 : \alpha = 0$ and the alternative is $H_1 : \alpha < 0$ are evaluated using the simple t-ratio for α:

$$t_\alpha = \frac{\hat{\alpha}}{se(\hat{\alpha})}, \tag{3.3}$$

where $\hat{\alpha}$ and $se(\hat{\alpha})$ are the estimates of α and its standard error. The critical values are non-standard so that the values tabulated in Mackinnon (1996) are used. The lag length is chosen using BIC allowing for a maximum lag-length of eight. The results presented in the Appendix in Table B.1 demonstrate that the transformed variables employed are stationary, giving p-values below 0.05.

Table A.1 gives some more details on the variables, the transformations and the sources the data have been taken from. Table A.2 gives the descriptives.

Chapter 4

Analysis of structural breaks

4.1 Theory

In order to evaluate and analyze the performance of individual models and combination schemes in the presence of shifts, the latter need to be identified beforehand. As pointed out in the introduction, there are many ways change can happen. The shift most widely tested for is a sudden and sizeable break. In particular, Stock and Watson (1996), analyzing a huge number of macroeconomic variables, find structural breaks to be present in the majority of cases. Furthermore, Clements and Hendry (1994) identify abrupt shifts of the parameter vector as the class of structural change most detrimental in the context of forecasting.

Firstly, this section presents the basic notational framework and the literature testing for a single break. Then the framework of Bai and Perron (1998) is outlined which will be applied to the variables of the empirical experiment in the following subsection. It is less restrictive with respect to the number of structural breaks. The procedure can be organized into the estimation of the breakdates, the tests to determine if there has been a structural break at all, and the estimation of the number of structural breaks.

Consider a structural change model with m breaks resulting in $m+1$ regimes of the form:

$$\begin{aligned}
y_t &= x_t'\beta_1 + u_t && t = 1, 2, ..., T_1 \\
y_t &= x_t'\beta_2 + u_t && t = T_1 + 1, ..., T_2 \\
&\vdots && \vdots \\
y_t &= x_t'\beta_{m+1} + u_t && t = T_m + 1, ..., T,
\end{aligned} \tag{4.1}$$

The time index is denoted by t; y_t is the dependent variable and $x_t(q \times 1)$ is a vector of independent variables with the corresponding vector of coefficients with $\beta_i \neq \beta_{i+1}(1 \leq i \leq m)$. Both, the vector of coefficients and the break points $(T_1, ..., T_m)$ are explicitly treated as unknown. In the following $T_0 = 0$ and $T_{m+1} = T$. Expressing the equation system in (4.1) in matrix form yields:

$$Y = \overline{X}\beta + U$$

where $Y = (y_1, ..., y_T)'$ and $\overline{X}$ is the matrix diagonally partitioning X at the m-partition $(T_1, ..., T_m)$, $\overline{X} = diag(X_1, ..., X_{m+1})$, with $X_i = (x_{T_{i-1}+1}, .., x_{T_i})'$; the coefficient vector is given as $\beta = (\beta_1', \beta_2', ..., \beta_{m+1}')$ and the disturbance being $U = (u_1, ..., u_T)'$. In order to ensure that each break date is asymptotically distinct and bounded from the limits of the sample usually some restrictions on the possible values of the break dates are imposed. For some arbitrary small positive number ε the following set is defined:

$$\Lambda_\varepsilon = \{(\lambda_1, ..., \lambda_m); |\lambda_{i+1} - \lambda_i| \geq \varepsilon, \lambda_1 \geq \varepsilon, \lambda_m \leq 1 - \varepsilon\},$$

where $\lambda_i = T_i/T$ and $0 < \lambda_1 < ... < \lambda_m < 1$.

The first authors analyzing structural breaks considered only the presence of a *single break*, i.e., m=1. Chow (1960), tested the null-hypothesis H_0 : $\beta_1 = \beta_2$ against the alternative $H_0 : \beta_1 \neq \beta_2$. The Chow-test is then a simple F-test of the form:

$$F_T(T_1) = \frac{\bar{S}_T - S_T(T_1)/(q+1)}{S_T(T_1)/(T - 2q - 2)} \tag{4.2}$$

where $\bar{S}_T$ is the sum of squared residuals under the null hypothesis and

$S_T(T_1)$ is the sum of squared residuals under the alternative hypothesis, which depends on the break point T_1. The test statistic follows the F-distribution with q and $T-2q$ degrees of freedom. However, the use of the Chow statistics is limited in two ways. The test is only applicable if one break is present, and the breakpoint has to be known in advance. According to Hansen (2001), the researcher has two options: She may test for an arbitrary break date or for a break date based on some known feature of the data. The first solution is prone to errors, as the acception of the null hypothesis could always mean that a true break date has been missed by chance. The second one is likely to indicate a structural change when there is in fact none, as the potential break date is correlated with the data. Additionally, both approaches are highly arbitrary so that different researchers can easily obtain different results.

Quandt (1958) proposes a first solution to this challenge. The idea behind the test is to compute the $F_T(T_1)$-statistics for all potential break points in a given interval focusing on the one that most likely rejects the null hypothesis, the supremum, giving:

$$\sup F_T = \sup_{T_1 \epsilon [\varepsilon T, (1-\varepsilon)T]} F_T(T_1) \tag{4.3}$$

If Quandt's statistic exceeds certain bounderies, the null hypothesis of no structural break is rejected. However, the Chow-statistic follows the F-distribution only if the break date is known. If one parameter is identified only under the alternative hypothesis of a break, testing for the significance becomes a non-standard problem (Hansen, 2001). Andrews (1993) makes the Chow-test operational for the case of an unknown breakpoint. He derives the limiting distribution of $\sup F_T$ showing that the limiting distribution of the statistic depends on the trimming parameter ε presenting simulated critical values. Hansen (2000) presents a heteroskedastic fixed-regressor bootstrap procedure that delivers the correct asymptotic distribution for the $\sup F_T$-statistic in the presence of general non-stationarities in the regressors, including mean and variance breaks and unit roots.[1]

[1]Note: Andrews (1993) used a Wald-like test. However, for normal linear regression models the two tests are equivalent.

Bai and Perron (1998) present a framework for the analysis of *multiple structural breaks*. The estimation of the break points builds on the simple OLS approach. For each m-partition $(T_1, ..., T_m)$, denoted $\{T_j\}$ the respective least-squares estimate of β_j is computed. It minimizes $\sum_{i=1}^{m+1} \sum_{t=T_{i-1}}^{T_i} [y_t - x_t'\beta]^2$, where $T_0 = 0$ and $T_{m+1} = T$. The disturbance term u_t is assumed to have mean zero, but its variance $\sigma_{u_t}^2$ may be heteroskedastic. However, the changes in variance are only permitted to take place at the same dates the conditional mean of y_t changes. Let $\hat{\beta}(\{T_j\})$ denote the resulting estimate. Employing this in the objective function yields the sum of squared residuals denoted as $S_T(T_1, ..., T_m)$. The estimated break dates are such that

$$(\hat{T}_1, ..., \hat{T}_m) = \arg \min_{(T_1,...,T_m)} S_T(T_1, ..., T_m),$$

where the minimization is taken over all partitions $(T_1, ..., T_m)$, and $T_i - T_{i-1} \geq [\varepsilon T]$, i.e., the break point estimators $(\hat{T}_1, ..., \hat{T}_m)$ are global minimizers of the objective function and the estimated regression coefficients at the estimated m-partition $\left\{\hat{T}_j\right\}$, i.e., $\hat{\beta} = \hat{\beta}(\left\{\hat{T}_j\right\})$. The break points have a discrete number so that they can be estimated by a grid search. However, the standard grid search procedure requires least squares operations of order $O(T^m)$. Bai and Perron (1998) reduce this number to operations of order $O(T^2)$ making use of an algorithm based on the principle of dynamic programming considered by Fisher (1958).

Building on the break estimates, the authors extend Andrews (1993)'s framework of testing for a single structural break to the case of multiple structural breaks. It is a subsample procedure utilizing a $\sup F$-type statistic, testing the null hypothesis of structural stability against the alternative hypothesis that there is a *known* number of breaks n, i.e., $m = 0$ versus $m = n$ breaks:

$$F_T(\lambda_1, ..., \lambda_n; q) = \frac{1}{T}\left(\frac{T-(n+1)q}{nq}\right)\hat{\beta}'R'(R\hat{V}(\hat{\beta})R')^{-1}R\hat{\beta}.$$

Here, R is such that $(R\beta)' = (\beta_1' - \beta_2', ..., \beta_n' - \beta_{n+1}')$ and $\hat{V}(\hat{\beta})$ is a heteroscedasticity and serial correlation robust estimate of the variance covariance matrix of $\hat{\beta}$, and q are the degrees of freedom.[2] The sup$F$-type test statistic is then defined as

$$\sup F_T(n,q) = \sup_{(\lambda_1,...,\lambda_n)\epsilon\Lambda_E} F_T(\lambda_1,...,\lambda_n,q) = F_T(\hat{\lambda}_1,...,\hat{\lambda}_n;q)$$

where the break point estimates $(\hat{\lambda}_1,...,\hat{\lambda}_n)$ minimize the global sum of squared residuals. Bai and Perron (2003a) present variations of the test for the case of different assumptions made with respect to the distribution of the regressors and the errors across segments.

The authors relax the restriction that the number of breaks has to be known presenting the double maximum tests. However, they require the specification of an upper bound M for the number of possible breaks m. They are defined for some fixed weights $\{a_1,...,a_M\}$ as

$$\begin{aligned} & \mathrm{D}\max F_T(M,q,a_1,...,a_M) \\ = \; & \max\nolimits_{1\leq m\leq M} a_m \sup\nolimits_{(\lambda_1,...,\lambda_m)\in\Lambda_\varepsilon} F_T(\lambda_1,...,\lambda_m,q) \\ = \; & \max\nolimits_{1\leq m\leq M} a_m F_T(\hat{\lambda}_1,...,\hat{\lambda}_m;q) \end{aligned}$$

The weights $\{a_1,...,a_M\}$ are set according to the priors on the likelihood of single change points. Bai and Perron (1998) do not give any theoretical guidelines on how to set the weights, however they propose two particular versions. The first sets all weigths equal to unity, giving the statistic labeled

$$\mathrm{UDmax} F_T(M,q) = \max_{1\leq m\leq M} \sup_{(\lambda_1,...,\lambda_m)\in\Lambda_\varepsilon} F(\lambda_1,...,\lambda_m,q)$$

Bai and Perron (1998) point out that for a fixed m, $F(\lambda_1,...,\lambda_m,q)$ is the sum of m dependent chi-square random variables where all the elements of the sum are divided by m. This implies that – holding q constant – the critical values for the individual tests $\sup_{(\lambda_1,...,\lambda_m)\in\Lambda_\varepsilon} F(\lambda_1,...,\lambda_m,q)$ and in turn the marginal p-values decrease with a higher number of possible breaks m. Hence, the test might have low power for large m.

[2]For a discussion of the alterntives of $\hat{V}(\hat{\beta})$ see Bai and Perron (2003a).

In order to alleviate this problem the authors propose an alternative weighting scheme, where the weights are constructed such that the marginal p-values only depend on q and the significance level α. If $c(q,\alpha,m)$ denotes the asymptotical critical value of the test $\sup_{(\lambda_1,...,\lambda_m)\in\Lambda_\epsilon} F(\lambda_1,...,\lambda_m,q)$ for significance level α, the weights are defined as $a_1 = 1$ for $m = 1$ and $a_m = c(q,\alpha,1)/c(q,\alpha,m)$ for $m > 1$. The resulting test statistic is denoted

$$\text{WDmax}F_T(M,q) = \max_{1\leq m\leq M} \frac{c(q,\alpha,1)}{c(q,\alpha,m)} \times \sup_{(\lambda_1,...,\lambda_m)\in\Lambda_\epsilon} F(\lambda_1,...,\lambda_m,q)$$

Finally, they propose an iterative procedure to find the exact number of breaks. It tests the null hypothesis of l structural changes against the alternative that one additional break exists, i.e., $m = l+1$. The starting point are the estimates of the break points $(\hat{T}_1,...,\hat{T}_l)$ of the model with l breaks that are obtained by the global minimization of the sum of squared residuals. Such a model has $l+1$ segments. Each of these segments containing the observations $\hat{T}_{i-1}$ to $\hat{T}_i (i = 1,...,l+1)$ is in turn tested for the presence of a break, i.e., the null hypothesis of no structural break against one structural break is tested for $l+1$ times. If the minimum of the resulting $l+1$ sums of squared residuals is sufficiently smaller than the sum of squared residuals of the model with l breaks, the null hypothesis is rejected. Thus, the test statistic is defined as

$$\sup F_T(l+1|l) =$$
$$\left\{ S_T(\hat{T}_1,...,\hat{T}_l) - \min_{1\leq i\leq l+1} \inf_{\tau\in\Lambda_{i,\eta}} S_T(\hat{T}_1,...,\hat{T}_{i-1},\tau,\hat{T}_i,...,\hat{T}_l) \right\} / \hat{\sigma}^2$$

where $\Lambda_{i,\eta} = \left\{\tau; \hat{T}_{i-1} + (\hat{T}_i - \hat{T}_{i-1})\eta \leq \tau \leq \hat{T}_i - (\hat{T}_i - \hat{T}_{i-1})\eta\right\}$; for $i = l+1$, $S_T(\hat{T}_1,...,\hat{T}_l,\tau)$ and for $i = 1$, $S_T(\hat{T}_1,...,\hat{T}_{i-1},\tau,\hat{T}_i,...,\hat{T}_l)$ is understood as $S_T(\tau,\hat{T}_1,...,\hat{T}_l)$. It is the sum of squared residuals resulting form the least squares estimation from each m-partition $(T_1,...,T_m)$. $\hat{\sigma}^2$ is a consistent estimate of σ^2 under the null hypothesis.

4.2 Empirical application

This section analyzes the target variable inflation as well as the explanatory variables for structural breaks. Which structure should be tested for? Following the approach of Hansen (2001) and Bai and Perron (2003a), the analysis presented here focuses on the breaks in the structure of a linear AR(p)-model. It has a proven track record as a forecasting model and is an often hard to beat benchmark in forecasting competitions (Stock and Watson, 2006). Furthermore, the empirical experiment is based on ARX models employing a variety of different explanatory variables. The only structure they all have in common is the autoregressive component. In order to test for shifts in the level and persistence of the variables employed an autoregressive model AR(p) of the form

$$y_t = \alpha + \sum_{i=1}^{p} \beta_i L^i y_t + u_t \tag{4.4}$$

is fitted to each of them. Thereby, t, with $t = 1, ..., T$, again indicates the period, y_t is the variable under analysis, L^i is the lag-operator, α is a constant, β_i is a $p \times 1$ vector of regression coefficients and u_t are iid with mean zero and variance σ_u^2. Here and in the following break analyses, the November 2007 vintage of the variables with the sample sizes given in table A.2 are employed. The number of lags considered in the regression, p, is selected making use of BIC. Setting the maximum of possible lags to eight, the lag length for the AR(p) model of inflation given in Equation (4.4) selected is six.

As a starting point for the analysis of structural breaks, Bai and Perron (1998) propose to employ WDmax$F_T(M, q)$ and the UDmax$F_T(M, q)$ to test for the presence of breaks in general. If the null hypothesis of no structural break can be rejected, the number of breaks is determined using the sequential $\sup F_T(l + 1|l)$-statistic. Consider first the results for the inflation rate given in Table 4.1. It presents the results for UDmax, WDmax, and $\sup F_T(l + 1|l)$, for $l = 2, 3, 4$.[3] The $\sup F_T(1)$, the UDmax and the WDmax

[3] Following the recommendations of Bai and Perron (2003a)ϵ is set to 0.15 and $M = 5$.

Table 4.1: Tests statistics for breaks in the inflation rate

$\sup F_T(1)$	UDmax	WDmax	$\sup F_T(2\|1)$	$\sup F_T(3\|2)$	$\sup F_T(4\|3)$
33.10***	33.1***	45.47***	40.41***	41.63***	16.63
(22.62)	(22.80)	(28.87)	(22.62)	(24.64)	(26.54)

The asymptotic critical values at the 5 % significance level are given in parenthesis; *, **, and *** indicate significance at the 10, 5, and 1 % level.

Table 4.2: Breakdates and confidence intervals of the inflation rate

Estimators	$\hat{T}_1$	$\hat{T}_2$	$\hat{T}_3$
Break dates	1961:9	1980:4	1991:2
95 % C.I.	(1960:10;	(3/31/1977;	(1990:3;
	1963:6)	6/30/1981)	1992:5)

reject the null hypothesis at the 5 percent level, indicating that at least one structural break is present. The $\sup F_T(l+1|l)$ rejects up to $l = 3$ that is the estimated number of breaks is $m = 3$.

The break point estimates are presented in Table 4.2. The dates of the 95 percent confidence intervals are given in parenthesis. As the procedure allows for different variances across segments, the intervals are not symmetric. The last estimate, the one at February 1991, will be in the centre of the analysis of the following chapters as it lies in the period of the empirical experiment. Its confidence band is tight when contrasted with the results of comparable studies (see, e.g., Jouini and Boutahar, 2003) spanning 26 months. A plot of the inflation rate and the breakpoint estimates and 95 percent confidence intervals is given in Figure 4.1 on the next page.

Causality of the breakdates goes beyond the framework used and the purpose of this dissertation. However, the timing of the break of interest in the early 1990s suggests that it can be associated with three major events in the economic history of the UK: the financial turbulance initiated by George Sorros ending British membership of the European Exchange Rate Mechanism (ERM) in September 1992, the Bank of England adopting an inflation

Figure 4.1: UK inflation, breaks and 95 percent confidence bands

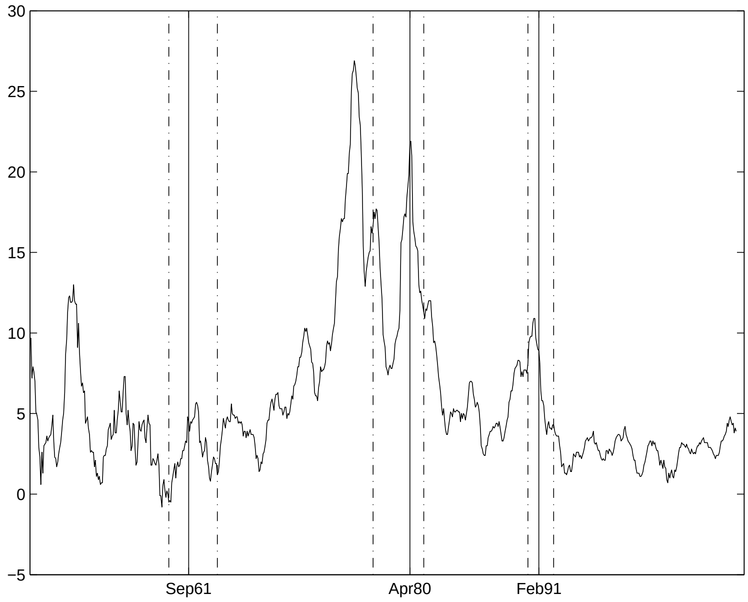


Table 4.3: Estimated std. dev., mean, and sum of AR coefficients

Segment	I	II	III	IV
	(1948:6–1961:8)	(1961:9–1980:3)	(1980:4–1991:1)	(1991:2–2007:10)
$\hat{\sigma}_{infl}$	2.98	6.23	3.97	1.15
$\hat{\mu}_{infl}$	3.95	8.8	7.38	2.88
$\sum_{i=1}^{p} \hat{\beta}_i$	0.95	0.99	0.96	0.90

targeting framework in October 1992, and the oil crises triggered by the Gulf War in 1990/91.

Table 4.3 contrasts the periods between the breaks giving some statistical measures: the sum of the estimated autocorrelation coefficients as a measure of persistence, the estimated mean, $\hat{\mu}_{infl}$, and standard deviation of inflation, $\hat{\sigma}_{infl}$. The estimated break point in February 1991 marks the transition to a markedly less volatile period with a moderation in the inflationary development and a less persistent effect of shocks to the inflation rate. The estimated standard deviation rises from the first segment having a value of

2.98 to 6.23 in the second period, and falls to a value of 1.15 in the last segment. The estimated mean and the persistence follow a very similar pattern. The fall in all of the three indicators from the third to the last segment is very pronounced. The estimated standard deviation drops from 3.97 to 1.15, the estimated mean from 7.38 to 2.88 and the estimated persistence from 0.96 to 0.90.

Hendry and Clements (2004) introduce the concept of co-breaking. They point out that structural breaks per se need not cause forecasting models to fail. If a break in the target variable is accompanied by breaks in the regressors, potentially compensating its impact, forecast accuracy may only be affected to a minimal degree, or not at all. Table 4.4 on the facing page analyzes, whether breaks in the regressors have occured.

In contrast to other studies testifying the omnipresence of structural breaks in macroeconomic variables (Stock and Watson, 1996) for most of the variables in the current analysis, evidence for the presence of structural breaks is limited. The $\sup F_T(1)$, the UDmax and the WDmax test reject the null of no structural break only for average earnings, the interest rate series, and the survey data. For these series, the three tests reject at least at the five percent significance level. In case of the exchange rate with the Deutschmark (DEM) there is weak evidence of a structural break with the WDmax test rejecting at the five percent level, the $\sup F_T(1)$ and the UDmax not rejecting at the ten percent level. Looking at the number of breaks, only in case of the economic sentiment indicator (ESI) there is significant evidence for more than one structural break with the $\sup F_T(2|1)$ rejecting the null hypothesis at the one percent level. Evidence for more than two breaks is weak.

Table 4.4: Test results for the presence and number of breaks

	sup $F_T(1)$	UDmax	WDmax	sup $F_T(2\|1)$	sup $F_T(3\|2)$
RS	3.44	5.34	7.7		
	(11.47)	(11.70)	(12.81)		
IP	1.56	4.75	8.40		
	(11.47)	(11.7)	(12.81)		
AE	59.61***	59.61***	59.61***	20.12*	
	(21.87)	(22.04)	(28.76)	(21.87)	
MO	4.91	4.91	6.60		
	(8.58)	(8.88)	(9.91)		
U	4.67	12.03	16.37		
	(17.60)	(16.37)	(17.92)		
BD	13.49**	13.49**	13.49***	5.65	
	(11.47)	(11.47)	(17.01)	(11.47)	
FTSE	10.91	10.91	10.91		
	(13.98)	(14.23)	(15.59)		
USD	6.76	6.76	8.93		
	(8.58)	(8.88)	(9.91)		
DEM	4.63	5.98	13.19**		
	(8.58)	(8.88)	(11.67)		
TB	35.56***	35.56***	37.51***	13.33*	
	(13.98)	(14.23)	(19.86)	(13.98)	
CBI	36.27***	36.27***	40.63***	21.45*	
	(21.87)	(22.04)	(28.67)	(21.87)	
OIL	3.58	5.39	9.39		
	(8.58)	(8.88)	(13.83)		
RIUS	41.02***	38.59***	44.02***	13.55	
	(16.19)	(16.37)	(21.95)	(14.26)	
ESI	63.81***	63.81***	63.81***	29.46***	16.19*
	(16.19)	(16.37)	(21.95)	(16.19)	(18.11)
BCI	19.71**	19.71**	20.24**	9.36	
	(18.23)	(18.42)	(19.96)	(18.23)	

The asymptotic critical values at the 5 % significance level are given in parenthesis; *, **, and *** indicate significance at the 10, 5, and 1 % level respectively.

Figure 4.2: Estimated break points and 95 percent confidence intervals

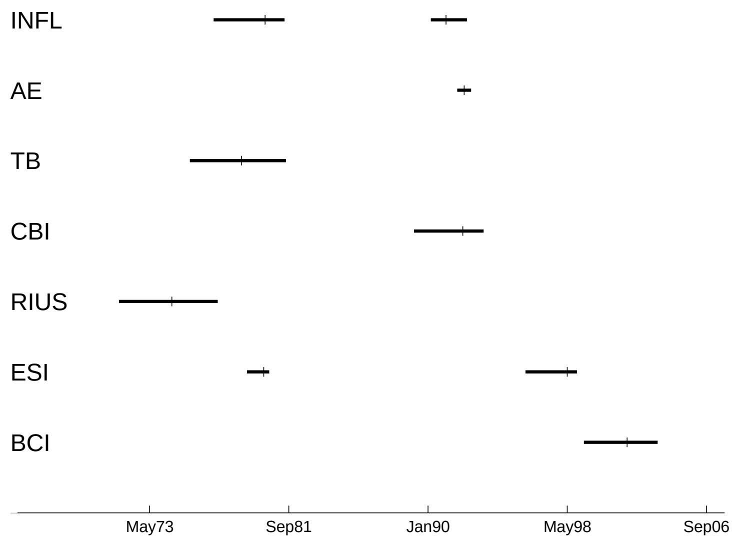


Figure 4.2 presents the break estimates of all variables along a timeline. In order to optimize presentation, the period considered in the figure starts in January 1972. The 95 percent confidence intervals are thereby given as horizontal lines. The break estimates are marked by a small vertical line crossing these lines. The estimated breaks do not all lie within certain periods. In particular, only the point estimates of the breakdates of average earnings and CBI are relatively close to the structural break of the inflation rate that is at the center of the analysis in February 1991. They are located in March and February 1992, respectively. The confidence band for average earnings is very thight, spanning eleven months, starting October 1992 and ending in August 1993, while the break estimate of CBI is very imprecise with the confidence band spanning more than four years starting in March 1989 and ending in May 1993. In face of these result co-breaking can be ruled out for most of the models. Thus, the mitigating effects of pooling derived by Hendry and Clements (2004) and laid out in section 2.2.2 might be useful.

Chapter 5

ARCH effects and combination

5.1 A first visual impression

Weighting schemes which are based on the accuracy of the constituent models learn from errors made. To be successful they require relative past performance to be constant over time. If this is not the case they could attribute large (small) weights to low (high) performing models leading to increased forecast errors of the combined forecast. Before analysing the ranks being the standard indicator of relative performances of the models in the next chapter, here, the focus is laid on the squared forecast errors. Although they represent a cruder measure, they allow to highlight some important features and relate performances to the break estimate.

Figure 5.1 on the next page gives a visual impression. It plots the squared forecast errors (y-axes) of the individual models (z-axes) over time (x-axes) for the 1-month horizon. The dates give the forecast origins. The higher the squared errors, the less accurate the forecasts.

The performance is correlated over time and models, giving the figure a wave-like pattern. While there are phases, where most of the models perform relatively well, there are at least two phases when most of the models have considerably higher squared forecast errors than over the other iterations. These phases are around February 1987 and February 1992. However, there

Figure 5.1: Forecast errors of the individual models over time, h=1

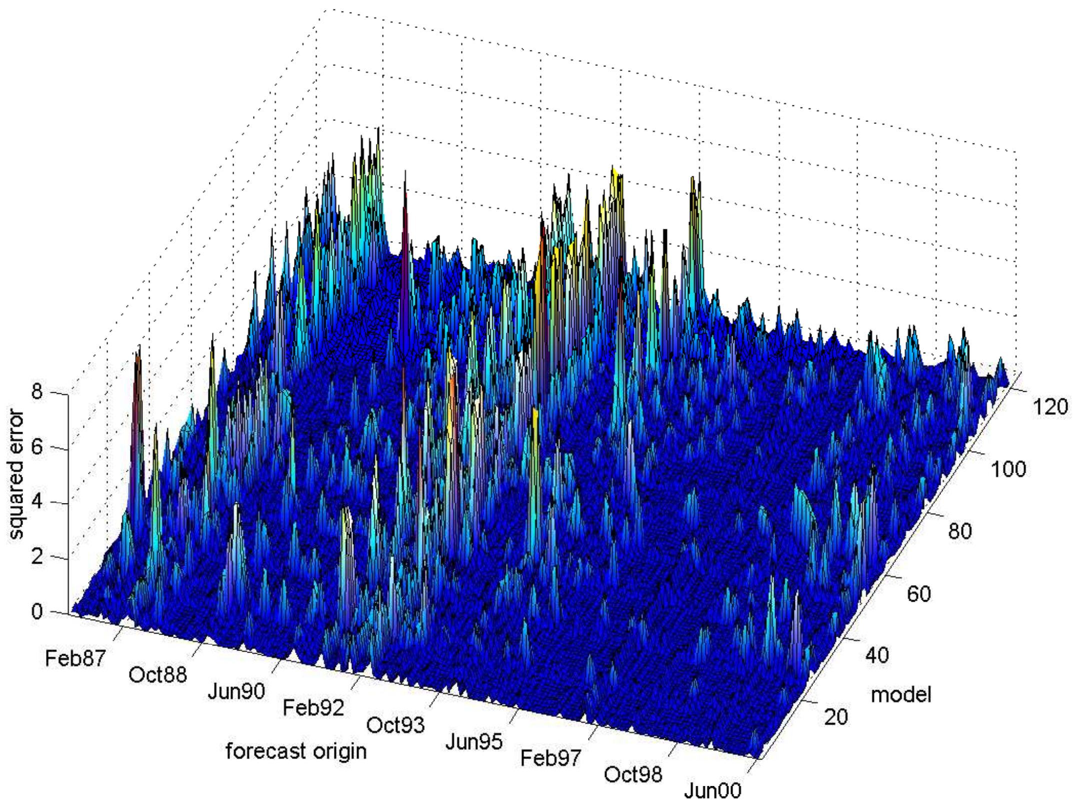
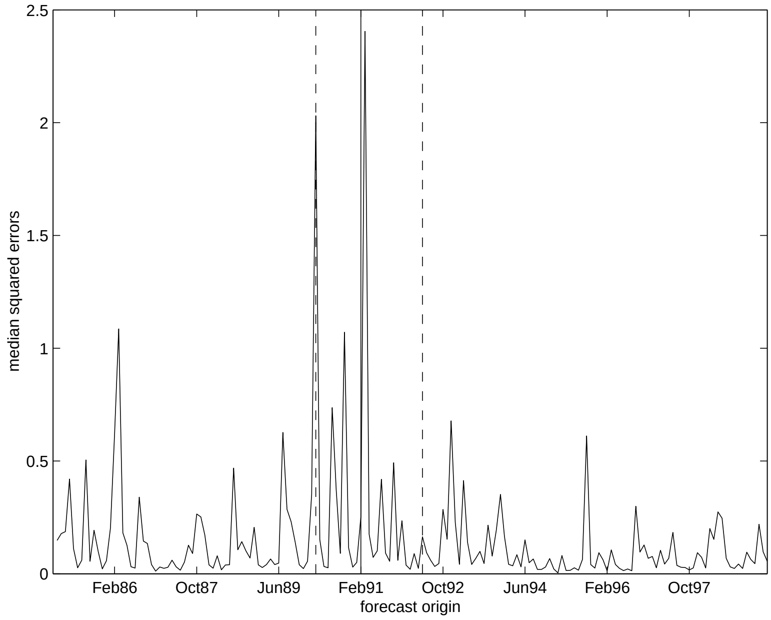


Figure 5.2: Median squared forecast errors over time, $h = 1$



are phases, e.g., around October 1988 and October 1993 when only part of the models are affected.[1]

The figure is dominated by extreme squared errors. To get a more precise picture, Figure 5.2 gives the median squared forecast error at each iteration. The break estimate is indicated by the solid vertical line and the 95 percent confidence bands by the dashed lines. The median squared forecast errors are mostly below 0.5. Within the confidence band of the break estimates the median is highest. Shortly after the break the median reaches its maximum of 2.3. Noteably, the third highest median emerges considerably earlier than the breakpoint estimate, in April 1990, which is the forecast origin after the lower bound of the 95 percent interval. Moreover, in February 1986 it doubles to well above one.

These results show that the estimated break can be associated with a phase of high forecast uncertainty that coincidentally affects most models. However, there are other sources of increased forecast errors as well. The uncertain

[1] The corresponding plots for higher forecast horizons are not presented separately as main features are very similar.

phase at the beginning of the period considered can neither be associated with a break date of the target variable nor the regressors (see Figure 4.2). Strikingly the second highest median squared forecast error precedes the break estimate by more than a year. Still, it is within the 95 percent confidence interval. Possibly, there are two breaks located too close to each other for the approach of Bai and Perron (2003a) to identify them separately.[2]

If structural change only affected the extent of the forecast errors while the relative performance were constant over time its negative impact on combined forecasts should not be very high. However, there are several phases when only part of the models feature high squared forecast errors. Thus, in these periods past performance weighting schemes most likely attribute the respective models a weight that does not correspond to their relative performance.

Central to the following analysis are the individual performances of each model. They show a persistent pattern that can be exploited to improve combined forecasts in the face of changing relative performances: large (small) shocks tend to be followed by large (small) shocks of either sign, a phenomenon termed volatility clustering. The next section outlines, how this pattern can be modelled and directly employed in the context of combined forecasts. In chapter 7 it will be used in an indirect way to facilitate the prediction of forecast breakdowns.

5.2 (G)ARCH effects

Apart of the results provided here, volatility clustering is frequently found in applied work (Harvey, Leybourne, and Newbold, 2001). Moreover, not only forecast errors, but economic time series themselves rarely exhibit constant variances. To model this feature Engle (1982) introduces the concept of autoregressive conditional heteroscedasticity (ARCH). Leaning on Bera and

[2] The technique requires a minimum number of observations between two candidate break dates.

Higgins (1993) it can be described in terms of a dynamic linear regression model:

$$y_t = x_t'\beta + \varepsilon_t \qquad t = 1, ..., T \tag{5.1}$$

where x_t represents a $k \times 1$ vector of explanatory variables which may include lagged values of y_t; β is a $k \times 1$ vector of regression parameters.[3] The ARCH model characterizes the distribution of the stochastic error conditional on the realized values of the set of variables $\Psi_{t-1} = \{x_{t-1}, x_{t-2}, ...\}$:

$$\varepsilon_t | \Psi_{t-1} \sim N(0, h_t), \tag{5.2}$$

where

$$h_t = \alpha_o + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_q \varepsilon_{t-q}^2, \tag{5.3}$$

with $\alpha_0 > 0$ and $\alpha_i \geq 0$, $i = 1, ..., q$, to ensure that the conditional variance is positive.[4] In the regression model, a large shock is represented by a large deviation of y_t from its conditional mean $x_t'\beta$ or equivalently, a large positive or negative value of ε_t. In the ARCH regression model, the standard deviation of the current error ε_t is conditional on the realized values of the lagged errors ε_{t-i}, $i = 1, ..., q$. Thus, it captures the volatility clustering as h_t is an increasing function of the magnitude of the lagged errors, irrespective of their signs. The order of the lag q determines the length of time for which a shock persists in conditioning the variance of subsequent errors. The more lagged values in Equation (5.3), i.e., the larger q the longer the different episodes of volatility will tend to be.

However, the simple ARCH model has a major detriment that limits its practical use. In the first empirical applications of ARCH to the volatility of UK inflation Engle (1982) found that a large lag q was involved in the conditional variance function. This would require the estimation of a large

[3]The following exposition concentrates on the mean and the variance. For a survey extending to higher moments and extensions of the ARCH framework see Bera and Higgins (1993).

[4]Literature has demonstrated that these restrictions can be relaxed (Bera and Higgins, 1993).

number of coefficients subject to inequality restrictions. Bollerslev (1986) adressed this issue, introducing the generalized ARCH (GARCH). He extends the conditional variance Equation (5.3) adding p autoregressive elements:

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + ... + \alpha_q \varepsilon_{t-q}^2 + \beta_1 h_{t-1} + ... + \beta_p h_{t-p}, \tag{5.4}$$

where the inequality restrictions $\alpha_0 > 0$, $\alpha_i \geq 0$, and $\beta_j \geq 0$, with $i = 1, ..., q$ and $j = 1, ..., p$, are imposed to ensure that the conditional variance is strictly positive. Given the orders of p and q the GARCH process is usually denoted as GARCH(p,q). It can be demonstrated (Bera and Higgins, 1993) that GARCH is a parsimonious representation of a higher order ARCH process, similar to the generalization of an moving average (MA) process, to an autoregressive moving average (ARMA) process.

A major contribution of Engle (1982) was to show that changes in the volatility need not necessarily be the result of an exogenous structural shift, but may be predictable and characterized by a specific form of non-linear dependence. Under certain conditions the GARCH framework implies that the unconditional variance is constant over time and can be derived analytically.

With the aim of laying some foundations for the understanding of the simulation analysis in chapter 8, consider the example of a GARCH(1,1). It is the specification which is most widely found to adequately describe economic and financial data. Using the law of iterated expectations yields

$$\begin{aligned} E(\varepsilon_1^2) &= E[E(\varepsilon_t^2 \mid \Psi_{t-1})] \\ &= E(h_t) \\ &= \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2) + \beta_1 E(h_{t-1}) \\ &= \alpha_0 + (\alpha_1 + \beta_1) E(\varepsilon_{t-1}^2). \end{aligned} \tag{5.5}$$

If we assume that the process began infinitely far in the past with a finite initial variance the sequence of variances converge to a constant value giving

$$\sigma_\varepsilon^2 = E(\varepsilon_t^2) = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}, \tag{5.6}$$

Table 5.1: Percentage of models having a significant ARCH effect

Horizon	%
1	28.93
3	31.41
6	23.14
12	15.70

where the necessary and sufficient condition for the existence of the variance is $\alpha_1 + \beta_1 < 1$. It can be shown that if this holds the process is weakly stationary.

5.3 Testing for ARCH effects

Are there any ARCH effects in the forecast errors of the empirical experiment and how many models are affected? Engle (1982) proposed a test for the presence of ARCH effects in the residuals of a regression. Following his procedure, in a first step the levels of the forecast errors are filtered by a simple autoregressive model to control for any mean effects. The lag-length of this auxiliary regression is selected using BIC. The resulting squared residuals e_t^2 are then regressed on their own q lagged values:

$$e_t^2 = \beta_0 + \sum_{s=1}^{q} \beta_s e_{t-s}^2 + \nu_t \tag{5.7}$$

The null hypothesis of the ARCH LM test is that there are ARCH effects up to the $q-th$ lag. The test statistic is the product of the number of observations T and the R^2 of the estimate of Equation (5.7) which asymptotically follows an $\chi^2(q)$-distribution. Table 5.1 gives the percentage of the number of models that feature significant ARCH effects to the total number of models for the four forecast horizons analyzed. For all tests, q was set to 20. Even for this high value of q, which tends to weaken the power of rejecting the null as the chance of including irrelevant lags rises, there is ample evidence of

ARCH effects. The share of affected models ranges from 15.7 percent in the case of the 12-month horizon to 31.41 percent in the case of the one month horizon.

5.4 The GARCH combination scheme

Shortly after the development of ARCH its usefulness in the context of combined forecasts was recognized. Engle, Granger, and Kraft (1984) fit a stylized monetarist and a mark-up model to US inflation and treat the forecast errors as a bivariate ARCH-process. In an iterative forecasting experiment, the weights of the model are up-dated each step using the predicted conditional heteroscedasticity. Analogous to the INV method, the weights of the respective model are higher, the lower the predicted conditional variance. The bivariate ARCH-process is specified as:

$$\varepsilon_t|\psi_{t-1} \sim N(0, H(\varepsilon_{t-1})), \tag{5.8}$$

where $H(\varepsilon_{t-1}) \equiv H_t \equiv [H_{ijt}]$ is a 2×2 positive definite symmetric matrix. The specification is chosen such that each element of H_t is a quadratic form in ε_{t-1}:

$$\begin{bmatrix} H_{11t} \\ H_{21t} \\ H_{22t} \end{bmatrix} = \begin{bmatrix} a_{01} \\ a_{02} \\ a_{03} \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t-1}^2 \\ \varepsilon_{1t-1}\varepsilon_{2t-1} \\ \varepsilon_{2t-1}^2 \end{bmatrix} \tag{5.9}$$

The weighted average of the two forecasts f_{1t} and f_{2t} is given as $f_t = \lambda_t f_{1t} + (1-\lambda_t) f_{2t}$ yielding the combined forecast error $e_t = \lambda_t \varepsilon_{1t} + (1-\lambda_t)\varepsilon_{2t}$. Analogous to the formation of the optimal weights in Bates and Granger (1969), Equation (2.10) in chapter 2, the authors form the weights at time t, as

$$\lambda_t = (H_{22t} - H_{21t})/(H_{11t} + H_{22t} - 2H_{21t}). \tag{5.10}$$

Thus, the weights change over time, as the variance and covariances of the errors evolve.

Though the approach is intuitively appealing, the results have been mixed at best. Diebold and Pauly (1987) point out that the prediction of conditional variances will lead to an extremely noisy weight sequence, which is seen to be very harmful to combined forecasts. In particular, he attributes the weak results to the misspecification of the off-diagonal elements of the covariance matrix H_t.

Thus, the empirical and simulation analysis of this dissertation will disregard the covariance between the forecast error sequences of the individual models along the lines in Timmermann (2006). The scheme will be labeled GARCH combination (GC). The weights at forecast origin t for models $i = 1, ..., n$ for horizon h are computed as

$$k_{it}^{GC} = \hat{h}_{i,t+h|t}^{-1} / \sum_{i=1}^{n} \hat{h}_{i,t+h|t}^{-1} \tag{5.11}$$

where $\hat{h}_{i,t+h|t}^{-1}$ is the predicted conditional standard deviation based on a GARCH(1,1) fitted to the forecast errors available up to t.

Chapter 6

Adaptive combination schemes

The previous chapter demonstrated that the squared forecast error variances vary over time. Furthermore, it showed that the changes can be modelled in form of ARCH effects which can directly be employed to up-date weight sequences. However, due to the mixed results of GC, most of the combination schemes employed in applied work concentrate solely on the relative performance of the models to be pooled. Thus, these schemes crucially depend on the relative accuracy to be constant over time – irrespective of changes in the level of squared forecast errors. Now, this section takes a more precise look at the relative performances considering the ranking of the models over time. It demonstrates that the ranks of the individual models actually alter substantially over time. Consequently, the following subsections present approaches to improve past performance schemes in face of this challenge.

6.1 Changing relative performance

To illustrate the non-constancy of relative forecast performances, consider the example of a forecaster faced with the task of predicting inflation using the best of the individual models described in chapter 3. Table 6.1 displays the most accurate models for the 1-month, 3-month, 6-month and 12-month horizon forecasts according to MSE. The result is presented for three different (sub)periods: for the whole period of the experiment, for the period preceeding the break estimate, and the period following the break estimate. It presents the variables used in the models additionally to inflation, which is always included. The MSE itself is given below the variable abbreviations.

Table 6.1: Best models over different (sub-)samples

	h=1	h=3	h=6	h=12
All periods	DEM, TB	CBI	CBI	BD, CBI
MSE	0.12	0.65	1.61	3.94
Before the break	TB, CBI	BD, CBI	BD, CBI	AE, CS
MSE	0.17	0.93	2.16	4.24
After the break	DEM, TB	USD, TB	MO, CBI	MO, BS
MSE	0.07	0.28	0.69	1.89

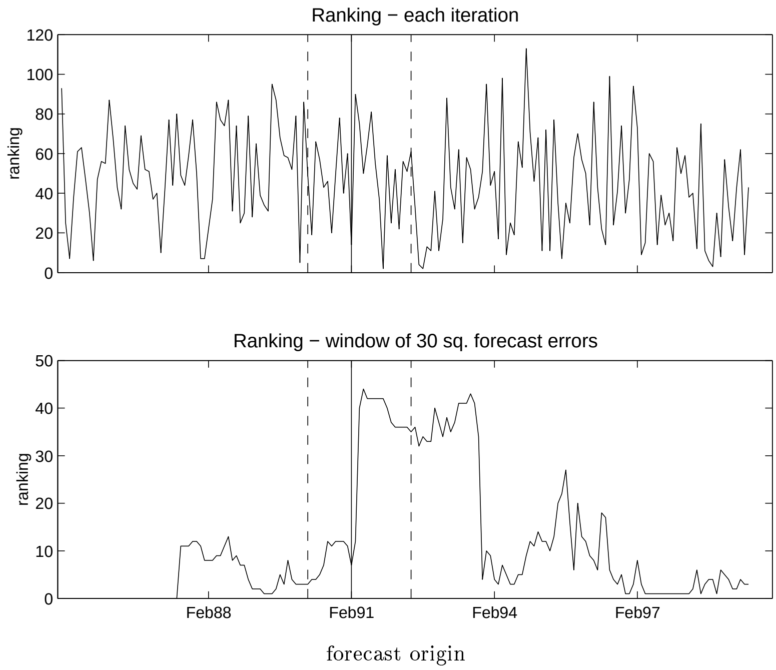
Consider the whole sample first. There is no model that outperforms the alternatives over all horizons. The model containing CBI is the best model, both for the 3-month and for the 6-month horizon forecasts. For the 1-month horizon forecasts the model containing DEM and TB is dominant while for 12-month horizon forecasts the model including bonds and CBI give the best model.

More importantly in the current context, there is no model that is the most accurate for one particular horizon and all sub-periods. While for the 1-month horizon the model containing DEM and TB excels for the complete sample and the post-break period, the model containing TB and CBI dominates the pre-break period. For the other three forecast horizons the result is still more striking as there is no model that dominates the alternatives in two of the three periods.

The forecaster most probably does not only search for the best model for some period, but rather for the one that is best for every point in time. The first subplot of Figure 6.1 illustrates the difference. It shows the ranks based on the squared errors of the best model over all periods for the 1-month horizon forecasts (DEM, TB). The dates refer to the forecast origins. The break and the 95 percent interval are displayed as solid and dashed lines.

The best position the model has taken is two. However, this occurs only twice, in November 1991 and in September 1992. In 46 out of 174 iterations, the model has a higher squared error than the median squared error in the respective iteration. What is more, the ranks seem to oscillate around 50. The worst position, were the model ranks 118th, corresponds to forecast origin October 1994.

Figure 6.1: Ranking and sq. forecast error best model, $h = 1$



Even if we control for some of the noise that results of the usual forecast uncertainty the picture remains unchanged. The second subplot gives the ranks that are formed on the basis of the last 30 squared forecast errors representing a smoothed version of the upper subplot. It demonstrates that there is at least one phase where the model ranked only 30th or worse. This period between April 1991 and November 1993 begins shortly after the break estimate.

Summarizing, the measurement of forecast accuracy is highly dependent on the time period under analysis. Furthermore, the break seems to have an effect on the ranking of the models.

6.2 Inverse MSE weights and weight stability

6.2.1 Methods of discounting outdated information

Up to now we have assumed the weight vector k to be fixed but unknown. The only exception was GC which uses conditional variance predictions to up-date weight sequences. In applied work k has to be estimated from the sample observations. A fixed k implies that changes in the recursively estimated weight vector $\hat{k}_t$ are interpreted as a result of the convergence in probability of $\hat{k}_t$ to k (Diebold and Pauly, 1987). However, in view of the results of the preceding analysis, the true covariance of the forecast errors and hence k is unlikely to be constant over time rendering a fixed weighting scheme suboptimal. The multiforecast analog to Equation (2.10) on page 26 for the computation of INV weights for individual model $i = 1, ..., n$ at forecast origin t is

$$k_{it}^{INV} = A_{it}^{-1} / \sum_{i=1}^{n} A_{it}^{-1}, \tag{6.1}$$

where, $A_{it} = \sum_{s=1}^{t-h}(Y_{s+h} - \hat{Y}_{i,s+h|s})^2$ is the sum of squared forecast errors of the last $t-h$ forecast origins. It slowly adapts to changes in the relative forecast performance of the individual models. The more past forecast errors enter A_{it} and the smaller the changes in the relative sizes of the forecast errors, the smaller the influence of new information on the weights.

Bates and Granger (1969) propose several simple methods of making the weights more responsive to recent information. Among others, they suggest to use a moving window of ν observations rather than the whole history of forecast errors to calculate the weights. Thus, $A_{it} = \sum_{s=t-v}^{t-h}(Y_{s+h} - \hat{Y}_{i,s+h|s})^2$. Furthermore, they suggest to discount past forecast errors, such that $A_{it} =$

$\sum_{s=1}^{t-h} \delta(Y_{s+h} - \hat{Y}_{i,s+h|s})^2$ and $\delta = \lambda^s$ a discount factor, such that for $\lambda > 1$ more weight is given to the more recent forecast errors than to distant ones.

Diebold and Pauly (1987) analyzing the combination of forecasts under structural change make some additional proposals with respect to the use of discount factors. Though their analyis focuses on the combination of a limited number of models allowing the weights to be estimated by OLS, their suggestions are easily transferable to the analysis of larger model spaces. Their suggestions are presented in Table 6.2; here t denotes past forecast origins.

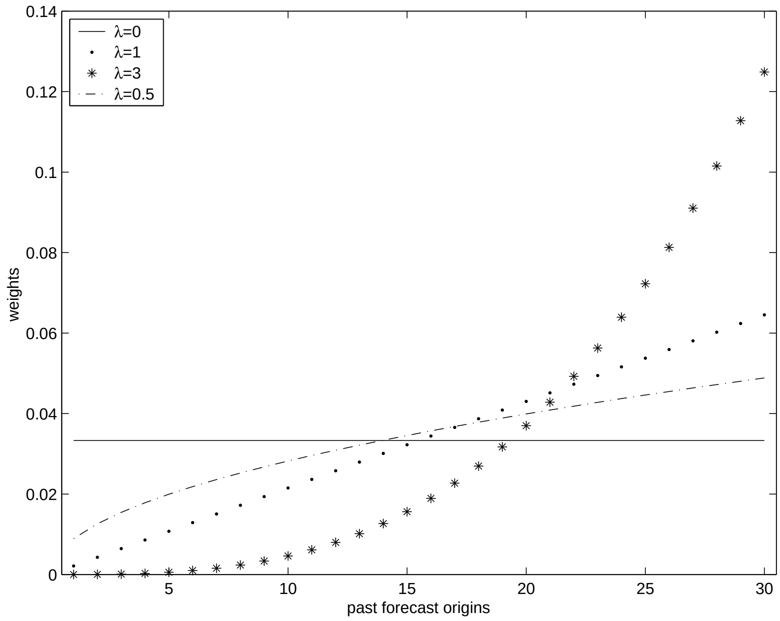
Table 6.2: Alternative schemes for discounting past information

(i)	$\delta = 1$, for all t	(equal weight)
(ii)	$\delta = t$, for all t	(linear)
(iii)	$\delta = \lambda^{T-t}$, $0 < \lambda \leq 1$, or $\delta = \lambda^t, \lambda \geq 1$.	(geometric)
(iv)	$\delta = t^\lambda$, $\lambda \geq 0$	(t-lambda)
(v)	$\delta = (t^\lambda - 1)/\lambda$ if $0 < \lambda \leq 1$; $\ln t$ if $\lambda = 0$	(Box-Cox)

Note: t denotes past forecast origins.

The equal weighting scheme (i) and the linear weighting scheme (ii) are special cases of the rest, where (i) does not discount past forecast errors, and (ii) gives more weight to recent forecast errors at a constant rate. The geometric scheme (iii) is desirable, if the researcher is interested in letting the weight on past information decline very rapidly. However, as $\frac{\mathrm{d}^2\delta}{\mathrm{d}t^2} = (\ln \lambda)^2 \lambda^{T-1} > 0$ for $0 < \lambda \leq 1$ and $\frac{\mathrm{d}^2\delta}{\mathrm{d}t^2} = -(\ln \lambda)\lambda^{t-2} > 0$ for $\lambda \geq 1$ it is limited to increasing rates of growth of the weights. In contrast, the Box-Cox weighting scheme (v) is restricted to decreasing rates of growth, as $\frac{\mathrm{d}^2\delta}{\mathrm{d}t^2} = -t^{\lambda-2} < 0$ if $0 < \lambda \leq 1$ and $\frac{\mathrm{d}^2\delta}{\mathrm{d}t^2} = -t^{-2} < 0$ if $\lambda = 0$. Furthermore, it is bounded by the linear and the log-linear schemes. The t-lambda scheme (iv) is the most flexible of the five as it allows for the weights to grow at an increasing and decreasing rate for more recent information. As an illustrative example the discount function of the t-lambda scheme is graphed in Figure 6.2 for values of λ of 0, 0.5, 1, and 3, where the number of errors of past forecast origins is set to 30, and 30 denotes the error of the most recent forecast origin. As $\frac{\mathrm{d}\delta}{\mathrm{d}t} = \lambda t^{(\lambda-1)} > 0$ and $\frac{\mathrm{d}^2\delta}{\mathrm{d}t^2} = \lambda(\lambda-1)t^{(\lambda-2)} > 0$, the

Figure 6.2: Discount functions for the t-lambda scheme



weights grow at an increasing rate if $\lambda > 1$ and at an decreasing rate if $\lambda < 1$. The border between the in- and decreasing part of the scheme, i.e., where $\lambda = 1$ corresponding to scheme (ii), is represented by the dotted line. The higher λ, the more recent is the information that is assigned a smaller weight compared to the situation when the observations are not discounted, represented by the horizontal line corresponding to $\lambda = 0$.

6.2.2 Weight stability

However, making weight estimates more responsive to changes in the covariance of the forecast errors comes at a cost. The higher the discounting and the smaller the window, the higher the influence single – more recent – observations have. Noise in form of the usual forecast uncertainty will more likely bias weight estimates. Thus, there is a trade-off between adaptivity and volatility of the weight sequence.

Even, if the the covariance is constant over time, but the history of forecast errors is relatively short, e.g., 20–50 observations, the weights calculated on the basis of the average of the past forecast errors are very sensitive to single observations and fluctuate around their optima. Kang (1986) investigates this issue in a simulation and an empirical experiment. In his MC study he analyzes the weights and the performance of combined forecasts of simple regression models, where some of the single forecasting models only use part of the underlying DGP. In both, the simulation and the empirical part of his paper, the weights display large standard deviations with wide ranges. The best combination scheme turns out to be EW – the one whose weight volatility is zero by definition. INV does not compare very favourably.

Winkler and Clemen (1992) corroborate this finding in a MC analysis of the sampling distribution of weight sequences, both for the combination of two forecasts and for the case of multiple forecasts. They demonstrate that INV is particularly sensitive. In the two-forecast-combination case the weights are very instable if, both, the ratio of the forecast error variances, κ, and the correlation of the forecasts, ρ, are close to one.

Their analysis of the combination of multiple forecasts is based on the sequential combination of forecasts: In a first step, two forecasts are combined. In turn, the combined forecast is treated as a single forecast which is then combined with the next forecast. Thus, the changes in the weights that are due to additional forecasts can be analyzed separately. Furthermore, the results of the two-forecast-case can be applied to the sequential analogs of κ, κ_{ji}, where $\kappa_{i,j} = \frac{\sigma_i^2}{\sigma_j^2}$ and $\sigma_i^2 < \sigma_j^2$, and ρ, ρ_{ji}, being the ratio of the forecast error variances and the correlation of two combined forecasts of forecasts i and j.

The first section of the multiple analysis considers exchangeable forecasts where the forecast errors have equal standard deviations, $\sigma_i = \sigma$ and equal correlations $\rho_{ij} = \rho$ for all i and j, where $i \neq j$. The optimal combination would assign equal weights, and as Winkler and Clemen note, this would lead the researcher to intuitively expect the weights to be relatively insensitive. However, the exchangeable-forecasts scenerio leads to high values of κ_{ji} and ρ_{ji} and hence very instable weights. The second scenario regards nonexchangeable forecasts that are correlated with ρ_{ij} taking values between 0.5 and 0.95 and the standard deviations ranging from one to 1.9. They point out that this scenario is often found in practice as forecasters have access to similar information and use more or less the same techniques. In this situation, estimated weights are highly susceptible to fluctuations in the error processes, as well.

6.3 Robust approaches

6.3.1 The Odds-Matrix Approach

The issue of weight instability and changing covariances of forecast errors imply the need for schemes that give robust weight sequences even when based on a relatively short information set. Gupta and Wilton (1988) present an approach that is supposed to compare favourably to the alternatives in this setting. The odds-matrix method (ODD) replaces the variance-covariance matrix by an odds-matrix, O.

Let π_{ij} represent the probability that the ith model will outperform the jth model. Outperformance is implied if the respective model has a lower absolute forecast error. Then, each element $o_{ij} = \frac{\pi_{ij}}{\pi_{ji}}$ represents the odds that forecast i will outperform forecast j. If the weights k_i are the 'true' weights, then the probability of model i outperforming model j should be given by $k_i/(k_i+k_j)$ whereas the odds should be given by k_i/k_j. Good models will have larger odds than bad models. Table 6.3 gives an example of the odds-matrix for the three model case. For all elements $o_{ij} = 1/o_{ji} > 0$ and the diagonal elements are ones. It can be shown that $Ok = vk$, so that the underlying weight vector k is given by the solution to $(O - vI)k = 0$. As O has unit rank and its trace is equal to v, there is only one eigenvalue which is different from zero. If the odds matrix is perfectly consistent, it is always possible to solve for k. However, if the the entries into the matrix are not consistent the matrix is no longer transitive. In this case, the solution of the problem $Ok = \tau_{max}k$ will be an eigenvector of positive values corresponding

Table 6.3: Example of the odds-matrix for three models

	model 1	model 2	model 3
model 1	$o_{11} = \frac{k_1}{k_1} = 1$	$o_{12} = \frac{k_1}{k_2}$	$o_{13} = \frac{k_1}{k_3}$
model 2	$o_{21} = \frac{k_2}{k_1}$	$o_{22} = \frac{k_2}{k_2} = 1$	$o_{23} = \frac{k_2}{k_3}$
model 3	$o_{31} = \frac{k_3}{k_1}$	$o_{32} = \frac{k_3}{k_2}$	$o_{33} = \frac{k_3}{k_3} = 1$

to τ_{max}. When normalized, this vector will be unique.

The remaining task is to specify O which consists in specifying $\binom{n}{2}$ pairwise probabilities, π_{ij}, where n again is the number of indiviual models. This can be easily implemented if sufficiently large data sets are available. Let a_{ij} be the number of times, model i outperformed model j, then $\pi_{ij} = a_{ij}/(a_{ij}+a_{ji})$.

What are the detriments and benefits of the odds matrix approach when compared to the variance covariance method? There is no analytical form of the expression at hand and there is no theoretical analysis available proving its usefulness in the presence of structural breaks. Furthermore, it does not consider the extent of the relative differences in the models performance, i.e., it is ordinal. However, over time, major performance differences should be reflected in the odds. Its set-up makes it less sensitive to outliers, and will thus generate more stable weights compared to the variance-covariance combination schemes, such as INV or OPT. As Gupta and Wilton (1987) point out, it can be expected that the method performes relatively good if only limited data are available, as the resulting weights are insensitive to small changes in the odds-ratios.

In an MC study Gupta and Wilton (1987) demonstrate the good relative performance with respect to EW, INV and OPT based on OLS estimation of the weights. In particular, they confirm findings that the ignorance of the covariance structure does not pose a major drawback neither to ODD nor

INV. The results show that ODD does not suffer from not accounting for the differences in the magnitude of the forecast performances. However, the rate at which the weights are stabilising is not analyzed. This is due to the design of their study which does not model a pseudo out-of-sample analysis. Instead of updating the weights on the basis of the information available up to each forecast origin, they apply a uniform weighting scheme over all iterations which is formed using the whole sample.

6.3.2 Rank-based approaches

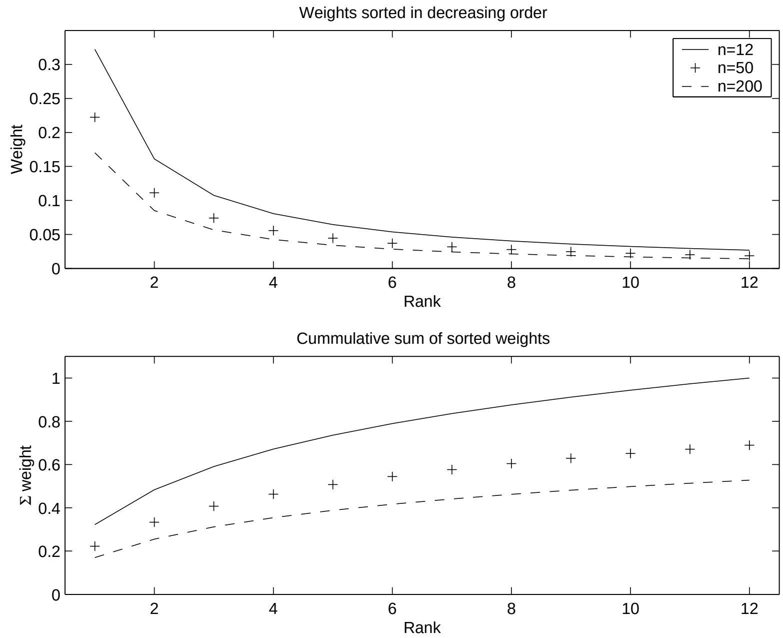
Rank-based combination schemes (RW) are very similar to ODD as they only consider the relative performance of the constituent models. However, while ODD considers the relative performance each iteration, RW merely take the average accuracy over all past iterations into account. Equal to ODD, RW disregards the extent of the differences in forecast performance. Again, this means a loss of information, however, large outliers are less likely to derail the scheme even if only little information is available to estimate the weights. The computationally easiest representative of this class of combination techniques is the simple median of the n individual forecast. In the following, it will be denoted Median.

Timmermann (2006) proposes a rank-based scheme that is very similar to the computation of INV. The weight at forecast origin t for the forecast at $t+h$ is computed as the ratio of the inverse rank of the respective model i at t, $IR_{it} = 1/R_{it}$ to the sum of the inverse ranks of all n models. The weight estimate of the individual model $\hat{k}_{it}^{RW}$ is thus given as:

$$\hat{k}_{it}^{RW} = IR_{it} / \sum_{i=1}^{n} IR_{it}. \tag{6.2}$$

The weights are a function of the number of models that are considered in the combination, n. This implicitly restricts the weights that can be attributed to the individual models. If, e.g., the researcher combines two alternative models, one of the models will always be given a weight of $k_1 =$

Figure 6.3: Weights of the RW method



$1/(1/1 + 1/2) = 2/3$, and the other a weight of $1/3$. No matter how big the differences in forecast accuracy are RW will always lead to these weights in the two model case. Moreover, the scheme can be characterized as favouring the very best models. Figure 6.3 illustrates the dependence of the twelve highest weights that can be attributed to single models as a function of the total number of models n. The upper plot demonstrates that the bigger n, the lower the weight attributed to the twelve best models. However, while receiving a weight of over 30 percent when $n = 12$, the best model still gets a weight of about 16 percent when $n = 200$. The lower graph gives the cummulative sum of weights of the twelve best models. Even for a huge number of models, when $n = 200$, the twelve best models are attributed more than 50 percent of the total sum of weights.

6.4 Comparison to EW

As pointed out before, if covariances of forecast errors of individual models evolve over time weight sequences need to be up-to-dated. However, up-to-dateness comes at the cost of higher weight uncertainty, the shorter the period employed for estimation. Furthermore, the use of past forecast performance implies a cost. In contrast, EW does not imply weight uncertainty and no prior information on the models is needed as the weights are attributed fixed values by definition. Thus, the issue is not how much information is needed for the weights to converge to their optima, but rather how much information does it take for past performance schemes to beat EW.

As there is no analytical way to answer this question, let us consider a small simulation experiment. The set-up thereby leans on Gupta and Wilton (1988). Three series, each containing 31 observations are created using pseudo-random draws of a normal distribution to simulate a sequence of hypothetical forecast errors of three models, at time $t = 1, ..., 31$. As a simulation of the total space of possible combinations of bilateral correlation coefficients, ρ_{ij}, and relative variance differences, κ_{ij}, analogously to the analysis in chapter 2 is computationally infeasable, the experiment concentrates on nine variance covariance patterns, presented in Table 6.4. Only positive values of ρ_{ij} are considered, as they argueably represent the most likely case found in reality (Winkler and Clemen, 1992). All constituent models in the first row of Table 6.4, schemes I, II and III, are equally accurate and the forecast error variances are set to one. In the second row, the variance difference is increasing to a 'moderate' level, giving $\kappa_{1,2} = 0.\bar{6}$, $\kappa_{1,3} = 0.5$, and $\kappa_{2,3} = 0.75$. In the last row, the difference in the variances are 'huge', with $\kappa_{1,2} = 0.4$, $\kappa_{1,3} = 0.25$, and $\kappa_{2,3} = 0.625$. The correlation between the forecast errors is increasing from the left to the right column. For the schemes in the left column, the bilateral correlation between the forecast errors, $\rho_{i,j}$, is zero. The schemes in the second column feature a 'medium' positive correlation of $\rho_{i,j} = 0.5$ and the forecast errors of the models in the third column feature a 'high' positive correlation of $\rho = 0.9$. For ease of analysis, the bilateral correlation coefficients are the same for all $\rho_{i,j}$, $i \neq j$, in each column.

Table 6.4: Variance-covariance patterns

increase in $\rho \longrightarrow$

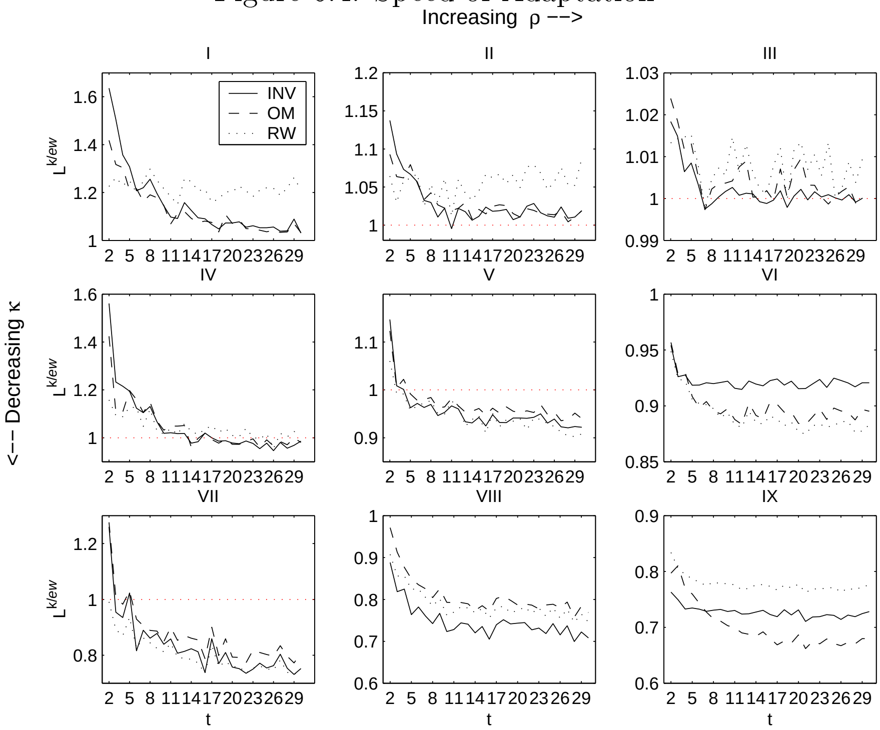
$\longleftarrow$ increase in $\min(\kappa_{i,j})$

	I			II			III	
1	0	0	1	0.5	0.5	1	0.9	0.9
0	1	0	0.5	1	0.5	0.9	1	0.9
0	0	1	0.5	0.5	1	0.9	0.9	1
	IV			V			VI	
1	0	0	1	0.6	0.7	1	1.1	1.3
0	1.5	0	0.6	1.5	0.9	1.1	1.5	1.6
0	0	2	0.7	0.9	2	1.3	1.6	2
	VII			VIII			IX	
1	0	0	1	0.8	1	1	1.4	1.8
0	2.5	0	0.8	2.5	1.6	1.4	2.5	2.8
0	0	4	1	1.6	4	1.8	2.8	4

Figure 6.4 plots the results given limited but growing information sets; in $t = 2$ only the forecast errors of the first period, $t = 1$, are used for the computation of the weights, in $t = 3$ the first two forecast errors have been used, and so forth. The lines give the loss of the three schemes compared to EW, $\mathrm{L}^k = \frac{\sigma_k^2}{\sigma_{EW}^2}$, where $k = \mathrm{INV}$, ODD, RW. Again, the higher L^k the worse the respective scheme is when contrasted with EW. $\mathrm{L}^k = 1$ indicates equal performance. 1000 replications are implemented for scenarios I to IX.

Given equal variances, every scheme disregarding the off-diagonal elements of the covariance matrix will tend to equal weights. Thus, EW clearly dominates the alternative schemes in scenario I, II, and III up to the 14th forecast origin and L^k is close to one hereafter. When the variance differences are not to big and the forecast errors are uncorrelated, for scenario IV, this holds, as well. The higher the potential gains of using past performance in the computation of the weights and the higher the correlation, the worse EW performs relative

Figure 6.4: Speed of Adaptation



to the alternatives. For scenario V, and VII, it only dominates intially. For scenario VI, VIII and IX it never dominates. When INV, ODD, and RW can make a difference in scenarios V to IX, there is no model that dominates for all situations. In scenario VIII the best model is INV, while ODD dominates in scenario IX, and RW excels in VI and V.

Adaptability of all schemes is high. The disadvantage of INV, ODD, and RW resulting of the need to learn the weights, evaporates very quickly. The biggest changes in L^k happen in the first 10 forecasts, though for some scenarios and some models that differs a little. The biggest additional gains over EW over time for all models can be realized in scenarios I, IV, and VII when the bilateral correlation coefficients are zero, and are smaller the higher the bilateral correlation coefficients are. RW schemes adapt faster than the alternatives, except when correlation and relative accuracy differences are high at the same time, i.e., in V, VIII, and in particular in IX.

When considering accuracy to the natural benchmark the trade-off between adaptability and weight estimation uncertainty is less of an issue. Even for relatively short informations sets, i.e., 15 observations, past performance schemes outperform EW for a wide range of κ-ρ-combinations. Furthermore, gains over EW through the inclusion of additional observations are small. With respect to the use of rolling windows this implies that sizes of about 15 or 20 observations should be sufficient. An extension to more observations limits the adaptability without corresponding gains through higher precision in the weight estimates.

Chapter 7

Forecast breakdowns and FBP

In the following a novel procedure for the robustification of combined forecasts in the presence of structural change, the Forecast Breakdown Preselection (FBP), is presented. It filters out those models, whose predictive performance will considerably deteriorate over the forecast horizon. This potentially reduces estimation error of the weights. And, it renders combinations more responsive to sudden changes, as bad performing models are immediately given zero weight – without having to wait for the weights to adjust. To achieve this, FBP recurs to ARCH effects in the forecast errors to apply the framework of Giacomini and Rossi (2005) of detecting and predicting forecast breakdowns to a wide range of models. The predicted failures are employed to identify reliable models, i.e., those models whose past and future performance is most likely comparable. The preselected models will then be used to compute combined forecasts along the lines presented in the preceding chapters.

7.1 Forecast break down

As mentioned before, computing weights based on relative past accuracy of the individual models becomes futile, if performances are changing drastically. In particular, a model that used to be a good predictive device and suddenly yields very inaccurate predictions, will be attributed an excessive weight. As the underlying DGP is unknown to the researcher, the 'usual' uncertainty associated with forecasting will sometimes result in high forecast errors. How bad need the predictions be to indicate that a single model is defective and how often has that to happen before we can say that it has not been a singular event?

A starting point in this respect is the definition of a forecast breakdown by Clements and Hendry (2005). They define a forecast breakdown as "a significant deterioration in forecast performance relative to the anticipated outcome, usually based on the historical performance of a model" (p. 2). Giacomini and Rossi (2005) present a framework to make this definition operable. They concretize a forecast failure as a significant difference between the in- and out-of-sample performance of a model in a repeated forecasting experiment. The authors consider a stochastic process $W \equiv \{W_t : \Omega \longrightarrow \mathbb{R}^{s+1}, s \epsilon \mathbb{N}, t = 1, ..., T\}$ which is defined on a complete probability space $(\Omega, \mathcal{F}, P)$. They partition the observed vector W_t into the variable of interest and a vector of predictors $W_t \equiv (Y_t, X_t')'$. The sample is divided into an in-sample window of size m and an out-of-sample window of size $n = T - m - h + 1$.[1] Let $f_t(\hat{\beta}_t)$ be the forecast of the iteration with the forecast origin t, with $\hat{\beta}_t$ being the $l \times 1$ parameter estimate. The starting point is the forecast produced by a direct forecasting method, i.e., the shortest lag to be considered in the estimation is h such that the most recent observation can be directly linked to the target period. Each iteration, the forecast correponds to a sequence of in-sample fitted values $\hat{y}_j(\hat{\beta}_t)$, where j varies over the in-sample window. Each iteration, the model is evaluated by a loss $L(\cdot)$, where each out-of-sample loss $L_{t+\tau}(\hat{\beta}_t) \equiv L(Y_{t+h}, f_t(\hat{\beta}_t))$ corresponds to an in-sample loss $L_j(\hat{\beta}_t) \equiv L(Y_j, \hat{y}_j(\hat{\beta}_t))$. Considering the OLS estimator of the linear model $Y_t = X_{t-h}'\beta + \varepsilon_t$, this will give $\hat{\beta}_t = \sum_{s=1}^{t-\tau}(X_s X_s')^{-1} \sum_{s=1}^{t-\tau} X_s Y_{s+\tau}$.[2] Thus the out-of-sample loss corresponding to the forecast at t is $L_{t+h}(\hat{\beta}_t) \equiv L(Y_{t+h}, X_t'\hat{\beta}_t)$ and the in-sample loss is

[1] h again denotes the forecast horizon.

[2] This exposition of the method of Giacomini and Rossi (2005) focuses on the recursive estimation scheme. The authors did consider rolling and fixed estimation schemes, as well.

$L_j(\hat{\beta}_t) \equiv L(Y_j, X'_{j+h}\hat{\beta}_t)$ with $j = h+1, ..., t$.

With the aim of quantifying a deterioration of the out-of-sample performance of a model relative to its in-sample analog, Giacomini and Rossi (2005) introduce the surprise loss, which is the difference of the out-of-sample losses with respect to the corresponding average in-sample losses $\bar{L}_t(\hat{\beta}_t)$

$$SL_{t+\tau}(\hat{\beta}_t) = L_{t+\tau}(\hat{\beta}_t) - \bar{L}_t(\hat{\beta}_t), \text{ for } t = m, ..., T-h. \tag{7.1}$$

Additionally to a forecast breakdown test that tests whether a model has suffered a forecast breakdown in the past, Giacomini and Rossi (2005) propose a framework to date past and to predict future forecast breakdowns. They propose linking forecast breakdowns to a set of economically meaningful variables. Let Z_t be the $r \times 1$ vector collecting those variables and let $\hat{\theta}_n$ be the OLS parameter estimate of equation:

$$SL_{t+\tau}(\hat{\beta}_t) = Z'_t\theta_n + \varepsilon_{t+\tau}, \tag{7.2}$$

over the out-of-sample period $t = m, ..., T-h$, where the regression always includes a constant. In order to test for the null hypothesis $H_0 : \theta_n = 0$ they propose a Wald test with the test statistic $W_{m,n,h} = n\hat{\theta}'_n\hat{\Omega}^{-1}_{m,n}\hat{\theta}_n$, where $\hat{\Omega}_{m,n}$ is an asymptotic variance estimator.[3] The null hypothesis is rejected, whenever $W_{m,n,h} > \chi^2_{r,1-\alpha}$, with $\chi^2_{r,1-\alpha}$ being the $(1-\alpha)-th$ quantile of a χ^2_r distribution. Furthermore, if fitted or predicted values of Equation (7.2) are significantly higher than zero, i.e., the lower $(1-\alpha)\%$ confidence band crosses the zero-line, a forecast breakdown is indicated. The one sided $(1-\alpha)\%$ confidence interval is easily computed as $\left(Z'_t\hat{\theta}_n - z_\alpha\left(Z'_t(\hat{\Omega}_{m,n}/n)\ Z_t\right)^{1/2}, +\infty\right)$, where z_α is the $(1-\alpha)-th$ quantile of a standard normal distribution.

Depending on the model to be analyzed, there is an ample set of potential explanatory variables to predict future surprise losses. Giacomini and Rossi (2005) forecasting US inflation propose financial indicators such as stock market volatility, business cycle leading indicators, and interest rates. In

[3]For a detailed describtion see Giacomini and Rossi (2005), proposition 7 and corollary 8.

another empirical application of the forecast breakdown analysis they relate forecast breakdowns of the traditional Phillips Curve model, linking changes in the inflation rate to past values of the unemployment gap, to changes in the parameters of a Taylor-type monetary policy reaction function.

In the case at hand, where a relatively large number of different models is used and the only common factor is the endogenous variable, a particularly attractive candidate for Z_t are lagged values of SL_t. As Giacomini and Rossi (2005) point out, this is feasible if ARCH effects induce persistence in the surprise losses. Thus, Equation (7.2) may be specified as:

$$SL_{t+h}(\hat{\beta}_t) = \sum_{i=0}^{p} \theta_i SL_{t-i}(\hat{\beta}_t) + \epsilon_{t+h}. \tag{7.3}$$

Equation (7.3) relates future values of SL_t to $p+1$ past observations of itself, enabeling the researcher to date past forecast breakdowns and to predict future ones for a wide range of heterogenously specified models.

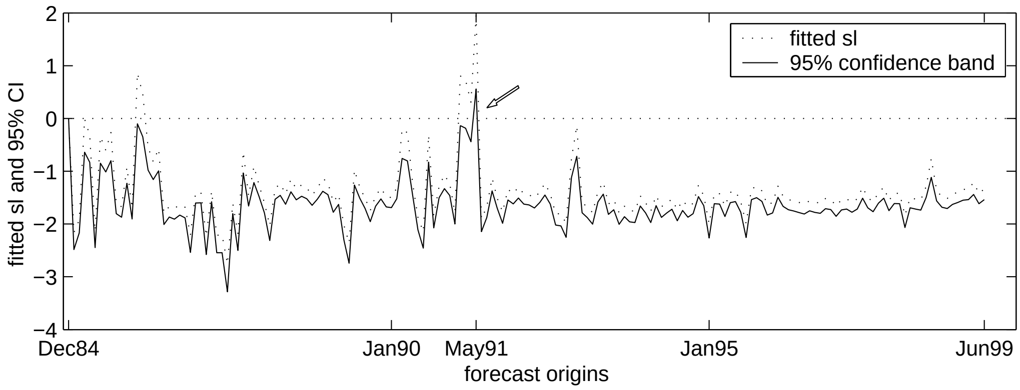
To illustrate the approach consider first the ex-post dating of past forecast breakdowns. The results are based on the whole information set that is on all in- and out-of-sample results from the first forecast origin December 1984 to the last forecast origin June 1999. Figure 7.1 on the facing page presents the results of an autoregressive model used for the three-month ahead forecasts of inflation.[4] The dotted line represents the fitted surprise losses, whereas the solid line represents the lower 95 percent confidence band. Whenever this lower bound crosses the zero line a forecast breakdown is indicated. This is the case in May 1991, which is indicated by an arrow.

7.2 Stable and instable periods

The presence of only one forecast breakdown for this model in 174 iterations raises doubts whether forecast breakdowns are sufficiently widespread to be

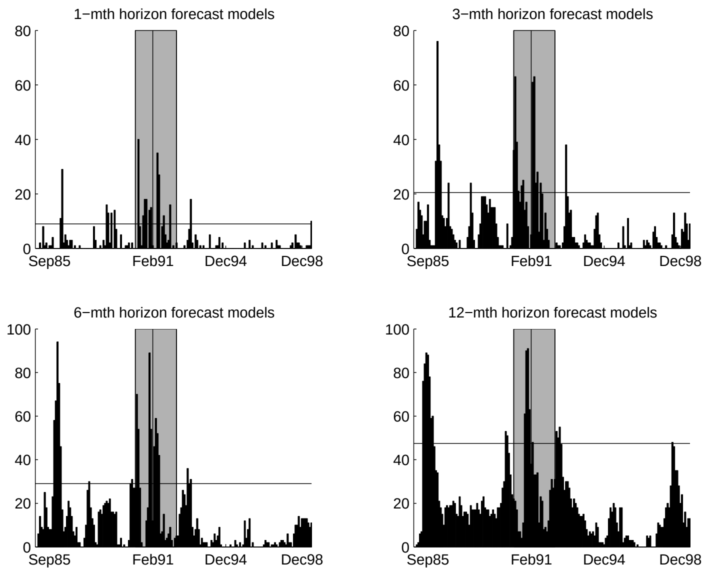
[4]For a detailed exposition of the estimation and forecasting approach chosen, see section 3.1 on page 35.

Figure 7.1: Example of dating forecast breakdowns



Note: the fitted SL and 95% confidence intervall result for a AR(p) used for the 3-month horizon forecast of inflation.

Figure 7.2: Number of forecast breakdowns per period



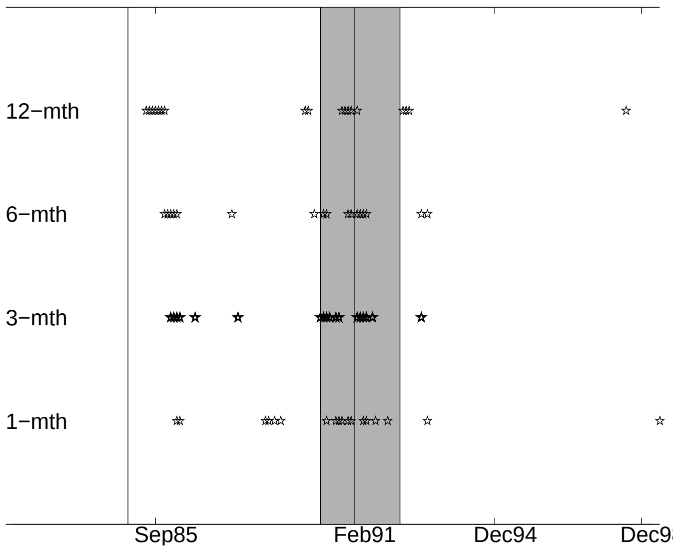
Note: the plots give the number of forecast breakdowns that have been dated ex-post for each iteration.

of any relevance in the current setting. Figure 7.2 gives the number of models that suffered a forecast breakdown for each iteration, for the 1-, 3-, 6- and 12-month forecast horizon. The horizontal lines indicate the upper ten percentiles of the number of forecast breakdowns per iteration. The shaded areas represent the 95 percent confidence bands of the break estimate of the inflation rate in February 1992. The break estimate itself is indicated by the vertical line within these bands.

The area covered by the black bars increases with the forecast horizons, indicating a rise in the overall number of forecast break downs. Given the total number of forecasts being 21054 per horizon, the overall numbers of failures, 504, 1431, 2142, and 3615, is considerable. The maximum number of breakdowns per period shows a similar pattern. At a single iteration at most 40, 76, 94, and 91 models fail over the period of analysis for the respective forecast horizons. In line with the analysis of the squared forecast errors in chapter 5 most of the peaks are located around the break estimate. Strikingly, the iterations associated with the highest number of forecast breakdowns for the 3- and 6-month horizon forecasts are not within or even close to the confidence interval of the break estimate of the inflation rate but rather around the end of 1985. Thus, the forecast breakdown analysis corroborates the existence of another period of uncertainty preceding the one associated with the break estimate of inflation.

Figure 7.3 on the facing page focuses on very instable periods where the stars indicate the location of the upper ten percentiles of the numbers of forecast breakdowns per iteration. From bottom to top the 1-, 3-, 6- and 12-month horizon forecasts are considered. Most of the iterations that feature high numbers of forecast breakdowns are clustered inside the 95 percent confidence bands of the structural break estimate of the inflation rate (indicated by the shaded area). However, while there are only few stars to the right of the confidence bands, there are clusters of stars at the beginning of 1986 until the end of 1988 for all four forecast horizons. This finding allows for a splitting of the sample that will be useful in the empirical analysis of the combined forecast in the following chapters: the approaches can be tested in

Figure 7.3: Phases of very high forecast uncertainty



Note: The stars indicate the location of the upper ten percentiles of the number of forecast breakdowns per iteration, dated ex-post.

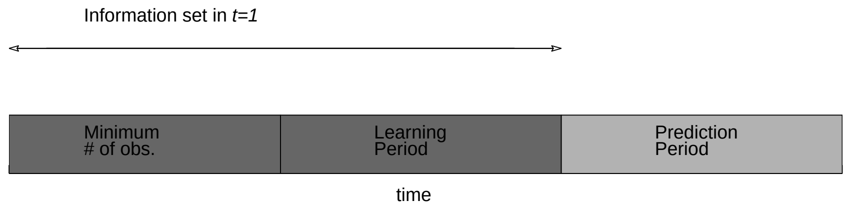
a setting of relative high and 'normal' forecast uncertainty. The period which is confined by the 95 percent confidence interval of the break estimate will henceforth be considered the INSTABLE period, the period between June 1992 to June 1999 will be labeled the STABLE period.

7.3 Forecast Breakdown Preselection

Given the ex-post evidence of forecast breakdowns being endemic, the prediction of forecast breakdowns ex-ante, i.e., when the information set is limited to the data available at each forecast origin t, appears very promising. If the predictions turn out to be reliable, the future forecast breakdowns can be employed for the preselection of models preceding the actual combination, resulting in the FBP filtered weights, defined as

$$\hat{k}_{sit}^{FBP} = \hat{k}_{sit} \cdot \mathrm{FBP}_{it} / (\sum_{i=1}^{n} \hat{k}_{sit} \cdot \mathrm{FBP}_{it}), \tag{7.4}$$

Figure 7.4: Timeline of FBP approach



where $\hat{k}_{sit}$ is the estimated weight of model i at the forecast origin t, and h is the forecast horizon. Weighting scheme s can be any of the approaches presented in the preceding chapters, i.e., OPT, EW, INV, ODD, RW, or GC. FBP$_{it}$ is a binary variable, being zero, if the lower 95 percent confidence interval of $SL_{it+h}(\hat{\beta}_{it})$, CI $^{SL,low}_{it+h} > 0$, and one else.

To illustrate the approach consider the timeline in Figure 7.4. The information set at the first forecast origin $t = 1$ is split into two sub-samples. To initialize the procedure the individual models need to be estimated based on a minimum number of observations. The second sub-sample is employed to evaluate the past performance of n individual models: in an iterative estimation and forecasting experiment, the in-sample squared errors and the squared forecast errors are collected, and the surprise losses are computed. Based on the surprise losses up to $t = 1$, Equation (7.3) on page 84 is estimated and the confidence bands are computed. The estimated equation is employed to predict the surprise loss of the next forecast step for each model i. If the lower 95 percent confidence band of the prediction crosses (does not cross) the zero line, FBP$_{it}$ is set to zero (one). In a last step, the adaptive combination scheme of choice is employed, and Equation (7.4) on the preceding page is used to sort out those models that are predicted to suffer a forecast breakdown. In the next period, $t = 2$, the forecast errors made in $t = 1$ are used to extend the learning period and the procedure is repeated. For $t = 3, ..., T$, the procedure is repeated in the same fashion. Table 7.1 presents a step-by-step description of the FBP method.

Table 7.1: Forecast Breakdown Preselection

1.	Estimation of the individual models
2.	In a learning period, surprise losses of the individual models are generated
3.	Estimation of $SL_{t+h}(\hat{\beta}_t) = \sum_{i=0}^{p} \theta_i SL_{t-i}(\hat{\beta}_t) + \epsilon_{t+h}$
4.	Prediction of the forecast breakdowns
5.	Preselection of the models
6.	Application of an adaptive combination scheme of choice

7.4 Relationship with the literature

The approaches presented in chapter 6 are designed to improve combined forecasts as they render weight sequences more adaptable to structural change. Thereby, they seek to trade off the use of past performance of the constituent models to estimate the weights against increased parameter estimation error. However, there are approaches to robustify combined forecasts that set in at a latter stage refining the 'raw' sequence of weights rescaling the weights before the actual prediction is made. The refinement techniques most widely used are shrinkage and trimming schemes. The shrinkage method, introduced by Diebold and Pauly (1987), adresses the situation when performance based weights tend to be very volatile and give huge (and possibly wrong) weights to some models. It reduces the chance of computing extreme weights shrinking them towards uninformative equal weights. To illustrate the basic concept consider the following simplified version proposed by Stock and Watson (2006)

$$k_{it}^{shr} = k_{it}\iota + (1-\iota)(1/n), \tag{7.5}$$

where $0 < \iota < 1$. The lower ι, the more the shrinkage weights k_{it}^{shr} tend to $1/n$, where n is the number of models, and the less the estimated performance differences represented by the weights k_{it} are taken into account.

The second refinement technique preselects the models to be employed prior to combination. Timmermann (2006) describes trimming as the process of dropping the worst α percent of the models based on their historical track record. This equates to setting the respective weights to zero and rescaling the restant weights such that they sum up to one. The idea is that models generating high forecast errors most probably contribute little to the improvement of the combined forecast, while the risk of harming the pooled forecast due to increased weight estimation error is relatively high.

So far, for both approaches the main parameters ι and α have been either estimated solely based on indicators of the past performance of the constituent models, or set according to some informal decision rule. However, while the approaches are intuitively appealing, empirical results have been mixed. Elliott (2004) has demonstrated that the performance is highly sensitive to the shrinkage parameter ι chosen.

As outlined in the theoretical analysis in chapter 2 on page 23 the efficiency of INV in dealing with very inaccurate models increases the worse their performance. Thus, gains through the exclusion of poor performing models may be limited. In view of the difficulties arising when measuring past forecast accuracy the success of trimming based on arbitrary decision rules appears even more unlikely.

FBP can be interpreted as a forward-looking version of the trimming approach. In contrast to the conventional trimming, it represents a formalized way of reducing the parameter estimation error that focuses on the performance over the time period where it matters, the forecast horizon. Furthermore, it concentrates on reliabilty rather than past forecast accuracy. This is of particular importance in periods of high forecast errors that put increased stress on the computation of the weights.

The adaptive weighting schemes presented in chapter 6 relied on the covariance of the forecast errors of the individual models to change relatively slowly over time. Though some of the more adaptive schemes might react faster than others, in case of a sudden and considerable change in the covariance of the forecast errors of the constituent models the combination schemes will need some time to learn. In contrast, FBP is more flexible, immediatly attributing a weight of zero, if the respective model is likely to be very inaccurate.

The GC scheme uses all of the information contained in the ARCH effects directly employing the conditional variance predictions for the estimation. It aims at fine tuning the weights, based on the knowledge of the likely evolution of the relative performances, and thus is a forward-looking procedure, as well. However, sucessful implementation requires to explicitly model and estimate the ARCH effects as precisely as possible. Even when huge financial market datasets are considered, predictive success of GARCH models is limited (Bera and Higgins, 1993). FBP is a crude and thus possibly less sensitive alternative, as it (implicitly) uses binary information on very high conditional variances only.

Chapter 8

Simulation analysis of FBP

The following Monte Carlo study (MC) demonstrates the usefulness of the FBP approach. The simulation assumes that the minimum conditions for the functioning of FBP are fulfilled: the availability of a huge number of forecasting tools and the presence of ARCH effects in the forecast errors. The preceding chapters showed that bilateral correlation coefficients of the forecast errors and relative forecast accuracy have a huge impact on the success of combination strategies. With the aim of testing robustness of the results to several relevant scenarios, the MC is designed to perfectly control for these features.

Moreover, the method is exposed to two forms of non-linear phenomena found to be present in the INSTABLE period in the previous chapters: structural breaks and phases of increased forecast error variances. Both phenomena frequently emerge in practical forecasting situations (Stock and Watson, 1996) and are possibly detrimental to FBP. In-sample and out-of sample performance measures employed by FBP are averages. Thus, the extreme values induced by a shock period or a break compared to 'normal' periods have a huge impact on measurement of the surprises losses. Table 8.1 summarizes the main features of the analysis.

Table 8.1: Central features of simulation analysis

1.	The presence of ARCH effects
2.	The availability of a large number of models/indicators
3.	Control of correlations between forecast errors
4.	Differently accurate models
5.	Uncertain periods
6.	Structural breaks in the DGP

8.1 The set-up

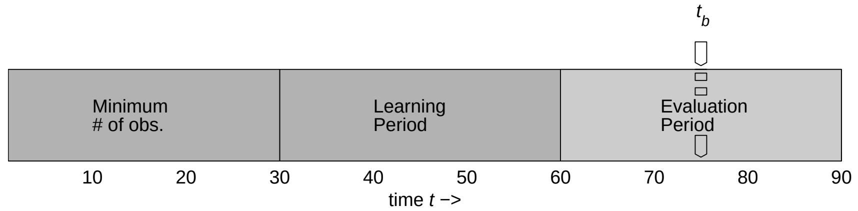
In contrast to the simulation study in section 6.4, the forecast errors can not be simulated directly as the forecast breakdown analysis requires indicators of in-sample fit. For the combination to make sense, individual models need to be specified, such that they represent only imperfect approximations to the DGP. Hendry and Clements (2004) model misspecification in form of an omitted variable bias while Aiolfi and Timmermann (2004) assume a switch in the DGP between two underlying factors, where each individual model uses only one factor.

In the two model case, it is comparatively easy to suit the covariance of the forecast errors to the needs of the analysis, as an analytical solution is mostly at hand. When a large set of models is considered, a straightforward way of controlling for the covariance is to model misspecification as a measurement error. This admits to directly influence the forecast errors. The explanatory variables have a core of useful information on the target variable blurred by noise – a description which is not unlikely to apply for many macroeconomic indicators, such as industrial production as a regressor for GDP, or monetary aggregates as predictors for inflation.

The target variable is defined as mean μ_t that is allowed to shift over time, plus a white-noise process ϵ_t, where $t = 1, ..., T$ is the time index, giving

$$y_t = \mu_t + \epsilon_t, \tag{8.1}$$

Figure 8.1: Set-up of the simulation analysis



with $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, where σ_ϵ^2 is the variance. The indicators x_{it}, where $i = 1, ...n$, are split into two parts: an informative content which is simply leading values of the target variable, and a noise part η_{it} giving:

$$x_{it} = y_{t+1} + \eta_{it} \tag{8.2}$$

The latter have mean zero and their variances $\sigma^2_{\eta_i,t}$ depend on past squared values of η_{it} in form of an ARCH(1):

$$\eta_{it} \sim N(0, \sigma^2_{\eta_i,t}) \tag{8.3}$$

$$\sigma^2_{\eta_i,t} = \gamma_0 + \gamma_1 \eta^2_{i,t-1} \tag{8.4}$$

The conditional variance of model i is given in Equation (8.4) where γ_0 is a constant and γ_1 is the coefficient of lagged values of η^2_{it}. This leads to an unconditional variance $\sigma^2_{\eta_i} = \frac{\gamma_0}{1-\gamma_1}$ and the $n \times n$ unconditional variance covariance matrix Σ_{η_i}, with correlation coefficients ρ_{ij}.[1] For the sake of perfect control of the forecast error variances the analysis abstracts of any dynamics in the mean equations, Equations (8.1) and (8.2).

Figure 8.1 illustrates the time-line of each recursion. Analogously to the empirical analysis, the forecasting experiment is implemented in a pseudo-out-of sample fashion that imitates the expanding information set typical to applied forecasting. In $t = 30$, n models $y_t = \hat{\beta}_{i0} + \hat{\beta}_{i1} x_{it-1}$ are estimated by OLS and one-step forecasts $y^F_{i,t+1}$ are computed. In $t = 31$ this procedure is repeated extending the information set of the OLS estimations to include the new observations of y_t and x_t in period $t = 31$, and so on. The evaluation

[1]The unconditional variance of a (G)ARCH process is introduced on page 60.

Table 8.2: $\min(\kappa_i)$ and median κ_i

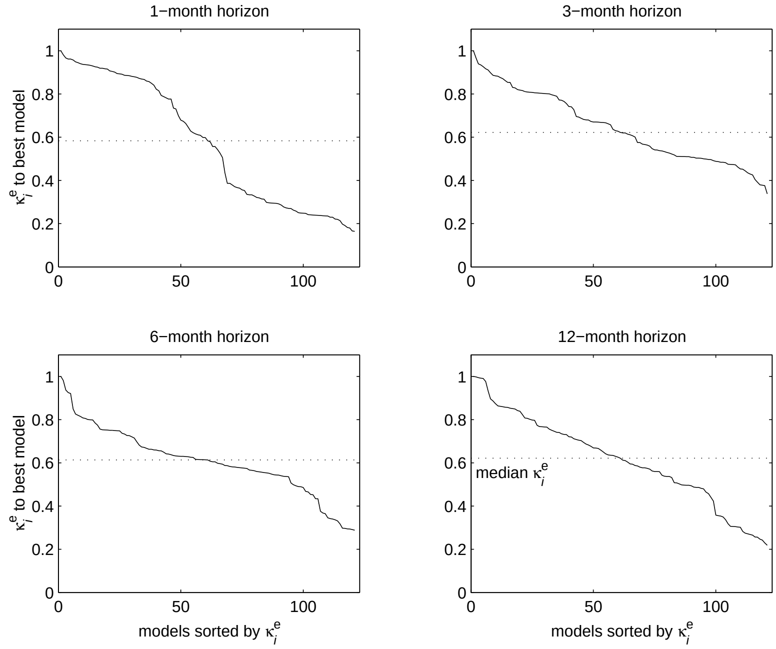
Horizon	$\min(\kappa_i)$	median κ_i
1	0.16	0.58
3	0.34	0.62
6	0.29	0.61
12	0.22	0.62

period (light gray) could be a shock period, where the variances of the forecast errors increase by a factor of Υ. Additionally, a break in the DGP, modeled as a change in the mean of the target variable y_t, could be added. It takes two values, $\mu_t = \alpha_1$ up to the break date $t_b = 75$, i.e., for $t = 1, ..., t_b - 1$, and $\mu_t = \alpha_2$ here after, i.e., $t = t_b, ..., T$.

8.2 Specification of the parameters

In the basic setting the parameters are mostly specified such that they match the empirical results and the robustness to changes is tested in alternative scenarios. Figure 8.2 on the facing page gives the ratio of the MSE of model i to the model having the smallest MSE, $\kappa_i^e = \frac{\min(MSE_i)}{MSE_i}$ for the different forecast horizons, whereby κ_i^e is sorted in decreasing order from the left to the right. The median κ_i^e for each period is given as a horizontal dotted line. The corresponding minimum and median κ_i^e are given in Table 8.2. κ_i^e is very similar for all horizons and displays a wide range of values. Most of the models have $\kappa_i^e > 0.5$, the lowest median being 0.58 in case of the 1-month horizon forecast. The worst performing models have values of κ_i^e between 0.16 in the case of the 1-month and 0.34 in case of the 3-month horizon forecast. Not giving any prior on the frequency of good or bad models, the unconditional variances $\sigma_{\eta_i}^2$ are pseudo-random draws of a uniform distribution $U(0.2, 1)$, leading to $0.2 \leq \kappa_{ij} \leq 1$. To minimize forecast errors arising from omitted variable bias the variance of the target variable y_t needs to be set considerably higher than $\max(\sigma_{\eta_i}^2)$, so that $\sigma_\epsilon^2 = 10$. This allows for a

Figure 8.2: Empirical κ_i^e to the best model

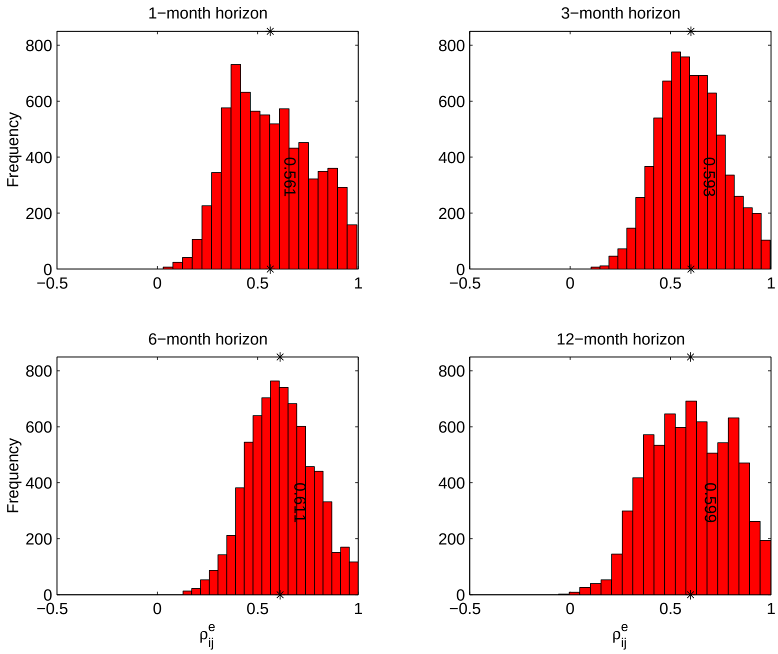


relatively tight control of the forecast error variances.

Figure 8.3 presents the histograms of the empirical bilateral correlation coefficients of the forecast errors of model i and model j, $\rho^e_{i,j}$, where $i \neq j$, for the different horizons. The distribution of $\rho^e_{i,j}$ appears to be symmetric and resembles a Gaussian distribution. The median, indicated by an asterisk, is between 0.56 (1-month horizon) and 0.61 (6-month horizon). The correlation coefficients are mostly positive, taking values between zero and one. Only for the 12-month horizon there are five negative correlation coefficients that are close to zero. Accordingly, the correlation coefficients $\rho_{i,j}$ of the MC are computed as pseudo-random draws from a standard normal distribution scaled such that they lie in the interval $(0;1)$ with mean 0.5. The ranges of both κ_{ij} and ρ_{ij} are in line with comparable studies, such as Gupta and Wilton (1987).

Given the results of Hendry and Clements (2004), the size of the break needs to be big relative to the in-sample error variances of the estimated models to have an influence on the forecast performance, so $\alpha_2 - \alpha_1$ is set to 10, giving

Figure 8.3: Empirical bilateral correlation coefficients ρ^e_{ij}



a ratio of $\frac{10}{1}$ to the forecast error variances of the worst models. For ease of computation, the analysis follows Giacomini and Rossi (2005) uniformly setting the ARCH coefficient of the models, γ_{i1}, to 0.6. In view of the results presented on the sizeable increase of the median squared forecast errors around the break estimates, the factor of the increase of the unconditional variances in the shock period, Υ, is set to 3. To robustify the analysis against the choice of the initial values the first 200 observations of η_{it} are dropped.

To test the sensitivity of the analysis to particular features of the set-up, six scenarios, A to F, are considered. The different scenarios are presented in Table 8.3. The number of observations assigned to the three basic periods, the minimum number of observations, the learning period, and the evaluation period, as well as the coefficients of the conditional variance are left unchanged. Scenario A is the baseline featuring neither a shock phase nor a break. Scenario B and C are set as Scenario A, Scenario B adding a period of uncertainty while Scenario C adds a break. Scenario D is Scenario A plus a shock phase and a break. Scenario E tests for the sensitivity to a mixture of positive and negative correlation coefficients. It is equal to D but the center of correlation coefficients ρ_{ij} is shifted to 0.25. Scenario F tests for the effect of lower relative MSE on D, differing only in that it has $\max \kappa_{ij} = 0.4$.

Table 8.3: List of scenarios

Scenario	Υ	$\alpha_2 - \alpha_1$	center ρ_{ij}	$\min \kappa_{ij}$
A	1	0	0.5	0.2
B	3	0	0.5	0.2
C	1	10	0.5	0.2
D	3	10	0.5	0.2
E	3	10	0.25	0.2
F	3	10	0.5	0.4

8.3 Results

The results are based on 1000 recursions for each scenario. To evaluate the gains of the application of FBP, combined forecasts are computed with and without applying the method. The schemes where FBP is not applied will be referred to as 'simple' in the following. The schemes comprise INV, ODD, RW, EW, Median, and GC. Adaptive versions of INV, ODD, and RW cover values of λ =0, 0.1, 0.5, 1, and 3 to discount more distant information or use a rolling window $\nu = 15$ or $\nu = 20$ in the way described in chapter 6. To minimize computational efforts, the analysis abstracts from estimation uncertainty with respect to the structure of the ARCH process. Assuming the structure to be known, a simple AR(1) model is fitted to the squared forecast errors to predict the conditional variances.

Table 8.4 presents the results for the baseline Scenario A. As the values of the MSE have no interpretation as such, the results are given as a ratio of a reference model, i.e., as Theil's U (TU). The benchmark model is EW without FBP. As it does not require any information on past performance

Table 8.4: Simulation results baseline scenario A

	Shock			
	simple		FBP	
Type (ν, λ)	TU	rk	TU	rk
INV (0,0)	0.61	9	0.57	9
INV (0,0.2)	0.60	8	0.56	8
INV (0,0.5)	0.59	7	0.56	7
INV (0,1)	0.58	6	0.55	6
INV (0,3)	0.54	4	0.52	4
INV (15,0)	1.55	15	1.48	15
INV (20,0)	2.26	16	1.94	16
ODD (0,0)	0.76	12	0.68	12
ODD (15,0)	0.65	10	0.61	10
ODD (20,0)	0.71	11	0.65	11
RW (0,0)	0.50	3	0.48	3
RW (15,0)	0.47	1	0.45	1
RW (20,0)	0.49	2	0.47	2
Median (0,0)	0.57	5	0.55	5
EW (0,0)	1.00	13	0.88	13
GC(0,0)	1.23	14	1.06	14

Note: TU is the Theil's U to the unfilt. EW;
rk gives the rank over the column.
ν = wind. size, λ = disc. factor.

of the constituent models, its MSE is the minimum result more elaborate models have to beat.

Consider column two and three displaying TU and ranks of the models for the shock period where FBP is not applied. As the relative differences of forecast accuracy are very high, schemes that consider past performance should improve upon EW. However, INV employing a window of 15 and 20 observations has the highest forecast errors giving TU of 1.55 and 2.26. Strikingly, GC, which correctly models the ARCH process of the forecast errors ranks merely 14th having TU of 1.23. All other schemes' TU are well below one, ranging from 0.47 for RW (15,0) to 0.76 for ODD(0,0). ODD which is one of the schemes explicitly developed to deal with changing forecast error variances ranks only 10th, 11th, and 12th, where the performance gets worse the more observations are used for the estimation of the weights. The most successful schemes are RW, ranking first when a rolling window of 15 observations is applied, ranking second and third, when the window size increases to 20 observations and when all observations are considered. The simplest scheme employing ranks, Median, ranks fifth. Considering INV schemes, where no rolling window is employed, the faster they discount, the better they are. Ordered according to the size of the discount factor λ, they take postion nine, eight, seven, six, and four. Columns four and five display the results for the different schemes, when FBP is applied prior to combining. All models improve, while the ranking is unaffected.

Table 8.5 gives the corresponding results for scenarios B to F. Considering the simple combinations first, most of the models have TU < 1. As in the baseline scenario, INV performs poorly when a rolling window is employed and improves, the higher λ, ODD schemes rank 10th at best, and RW outperforms the other models.

What effect does a break have? Consider the simple results of scenarios A to D first, as they differ only in the break and the shock period. The scenarios that do not feature a break, A and B, display considerably higher TUs than C and D for nearly all schemes. However, the relative performance of the schemes is almost unaffected by a break.

Table 8.5: Scenario B to F

	B				C				D				E				F			
	simple		FBP		simple		FBP		simple		FBP		simple		FBP		simple		FBP	
Type (ν, λ)	TU	rk	TU	rk	TU	rk	TU	rk	TU	rk	TU	rk	TU	rk	TU	rk	TU	rk	TU	rk
INV (0,0)	0.62	9	0.58	9	0.48	9	0.46	9	0.48	8	0.44	8	0.57	8	0.55	8	0.60	10	0.56	10
INV (0,0.2)	0.62	8	0.58	8	0.48	8	0.45	8	0.47	7	0.44	7	0.56	7	0.54	7	0.59	9	0.56	9
INV (0,0.5)	0.61	7	0.57	7	0.47	7	0.45	7	0.46	6	0.43	6	0.56	6	0.54	6	0.58	8	0.55	8
INV (0,1)	0.59	6	0.56	6	0.46	5	0.44	6	0.45	5	0.43	5	0.55	5	0.53	5	0.57	6	0.54	6
INV (0,3)	0.55	4	0.53	4	0.43	4	0.42	5	0.42	4	0.40	4	0.53	4	0.52	4	0.54	4	0.52	5
INV (15,0)	2.86	15	2.04	15	1.45	15	1.19	15	2.03	15	1.81	15	2.21	14	1.79	14	1.84	15	1.46	15
INV (20,0)	3.68	16	2.32	16	1.48	16	1.22	16	2.32	16	2.01	16	3.29	15	2.34	15	2.25	16	1.87	16
ODD (0,0)	0.80	12	0.70	13	0.63	13	0.57	13	0.68	12	0.58	12	0.75	12	0.68	12	0.78	13	0.68	13
ODD (15,0)	0.67	10	0.62	11	0.54	11	0.51	11	0.55	10	0.50	10	0.65	10	0.62	10	0.64	11	0.60	11
ODD (20,0)	0.74	11	0.66	12	0.58	12	0.54	12	0.61	11	0.54	11	0.69	11	0.65	11	0.71	12	0.64	12
RW (0,0)	0.51	3	0.50	3	0.40	3	0.38	3	0.40	3	0.38	3	0.48	1	0.48	1	0.53	3	0.51	4
RW (15,0)	0.49	1	0.47	1	0.39	2	0.38	2	0.39	1	0.37	1	0.50	3	0.50	2	0.51	1	0.49	2
RW (20,0)	0.50	2	0.48	2	0.39	1	0.38	1	0.39	2	0.37	2	0.50	2	0.50	3	0.52	2	0.50	3
Median (0,0)	0.58	5	0.55	5	0.52	10	0.49	10	0.51	9	0.47	9	0.61	9	0.58	9	0.58	7	0.55	7
EW (0,0)	1.00	14	0.86	14	1.00	14	0.85	14	1.00	14	0.83	14	1.00	13	0.89	13	1.00	14	0.85	14
GARCH (0,0)	0.84	13	0.62	10	0.46	6	0.40	4	0.88	13	0.68	13	34.07	16	33.88	16	0.56	5	0.36	1

Note: TU is the Theil's U (ratio of the MSEs) to the unfiltered EW-model; rk gives the rank over the column. ν = wind. size, where "0" indicates that all obs are used. λ = disc. factor; B is the scenario with a shock period, C features a break, D has shock period and break, E as D but center $\rho_{ij} = 0.25$, F as D but with $\max(\kappa_i = 0.4)$.

For INV schemes, one explanation is that a break mitigates weight estimation errors as it implies shrinkage towards EW. Box 1 gives an illustrative example. Scenario B corroborates this. Being less stable than A in that it features a shock period but not a break it shows the highest TU when comparing the results for individual schemes over the scenarios.

How do the results in D change, if ρ_{ij} shifts to 0.25 in E and if $\max \kappa_i = 0.4$ in F? In both scenarios most models' TU deteriorates except for three schemes in F: INV applying a window and GC. The increase of TU in E compared to D reveals that the stationary theory in section 2.1 is not applicable in a setting of shock periods and breaks. It predicts that INV always improves upon EW the higher relative performance differences are.

In view of GC correctly modelling the ARCH effects, its mostly poor performance is remarkable. In E, GC performance is worst showing an extreme value of TU of 34.07. However, for the other scenarios it can improve upon EW, and in F, GC even ranks 5th having TU of 0.56.

What effect does FBP have? All models improve by the application of FBP except for RW (0,0) in Scenario E, which has a gain of zero. However, the ranking changes only as GC improves in scenario B, C, and F.

To see in which scenarios FBP works best Table 8.6 shows the percentage gains due to the application of FBP the scheme for the six scenarios. The maximum gain is 36.97 for INV (20,0) in Scenario B, followed by GC in Scenario F and INV (20,0) in Scenario E, gaining 34.90 and 28.73 percentage points. Except for GC, which gains most in Scenario F, the highest gains for the models materialize if there is a shock period (Scenario B) or if there is a shock period and a break (Scenario D) and the correlation and relative performance structure is unchanged compared to the baseline. Comparing the scenarios D, E, and F, when a shift and a shock period is included, a shift of the center of ρ_{ij} to 0.25 (Scenario E) or smaller relative performance differences (Scenario F) lead to lower gains for most of the schemes. Thereby, the gains are markedly smaller in Scenario E than in Scenario F. Only INV schemes employing a rolling window of 15 and 20 give higher gains in Scenario E and F.

Table 8.6: Percentage gains using FBP

Type (ν, λ)	A	B	C	D	E	F
INV (0,0)	6.65	6.75	5.28	7.14	3.76	6.22
INV (0,0.2)	6.22	6.38	4.95	6.86	3.52	5.93
INV (0,0.5)	5.70	5.94	4.55	6.51	3.23	5.57
INV (0,1)	5.05	5.39	4.07	6.08	2.90	5.13
INV (0,3)	3.74	4.31	3.09	5.29	2.32	4.28
INV (15,0)	4.64	28.47	17.75	10.81	18.92	20.80
INV (20,0)	16.56	36.97	17.08	13.23	28.73	17.07
ODD (0,0)	12.50	13.16	10.42	14.44	9.11	12.68
ODD (15,0)	6.41	7.24	5.59	8.24	4.64	6.57
ODD (20,0)	9.53	10.40	7.68	11.25	6.81	9.43
RW (0,0)	4.77	3.29	4.90	4.67	0.00	3.57
RW (15,0)	3.98	3.48	4.52	5.01	1.09	3.49
RW (20,0)	4.12	3.46	4.74	4.85	0.20	4.25
Median (0,0)	4.81	6.35	4.90	8.58	4.37	5.99
EW (0,0)	13.89	13.91	15.06	17.27	10.80	15.33
GC	16.18	26.70	12.43	22.10	0.55	34.90

Note: The values given are gains expressed in percentage point changes of the columns labled simple and FBP in Table 8.5.

Summarizing the results, all models considerably gain by the application of FBP. The approach is particularly useful in periods of high forecast uncertainty, i.e., in the presence of shocks and breaks. However, the lower the differences of forecast errors and the more negative the bilateral correlation coefficients, the smaller the gains are.

Box 1: Example of shrinkage towards equal weights due to a break

This box illustrates, how a break that affects all models at the same time can shrink the weights of INV towards equal weights. Consider the simple two model example with weights computed according to INV, i.e., $k_{1t} = \frac{\Sigma sqe_{1t}}{\Sigma sqe_{1t}+\Sigma sqe_{2t}}$ and $k_{2t} = 1 - k_{1t}$ based on three squared forecast errors (sqe) in period $t = 1, 2, 3$. In period $t = 1$ and $t = 2$ sqe take values of 4 and 1 leading to weight $k_{1t} = 0.8$ for model 1, and $k_{2t} = 1 - k_1 = 0.2$ for model 2 in period $t = 2$. In $t = 3$ there is a sizeable break leading both sqe to increase by 40. The extreme sqe in $t = 3$ dominate the MSE of both models leading to weights that tend towards 0.5, giving $k_{1t} = 0.55$ and $k_{2t} = 0.45$.

Table: Two model example

t	1	2	3
sqe_{1t}	4	4	44
sqe_{1t}	1	1	41
k_{1t}	0.2	0.2	0.45
k_{2t}	0.8	0.8	0.55

Chapter 9

Empirical results

This chapter tests FBP using the experimental data giving the empirical counterpart to the MC of the last chapter. The data allow for a study of real life analogs of Scenario A and Scenario D. The analyses of structural breaks, the squared forecast errors, and the number of forecast breakdowns per iteration, identified the INSTABLE and the STABLE period. The former is characterized by relatively high forecast uncertainty whereas the latter period is marked by relatively accurate forecasts. The period preceding INSTABLE extending from the forecast origins December 1984 to February 1990 is employed as the learning period for the computation of the weights and FBP.

To test whether the filter works when other variables are forecasted, the results for the prediction of four additional targets are considered. Furthermore, FBP is compared to a valid alternative that applies a preselection strategy directly using ARCH effects.

9.1 Unfiltered combined forecasts

Table 9.1 investigates, whether there is a dominating combination scheme for both subperiods if FBP is not applied. The combination schemes are the same as in the simulation study. The first column gives the type of approach

Table 9.1: MSE and ranks; unfiltered

	INSTABLE								STABLE							
	$h=1$		$h=3$		$h=6$		$h=12$		$h=1$		$h=3$		$h=6$		$h=12$	
Type(τ,λ)	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk
INV (0,0)	0.30	5	1.44	10	3.67	5	7.92	5	0.07	7	0.26	11	0.75	6	2.36	12
INV (0,0.2)	0.30	3	1.44	11	3.67	4	7.92	3	0.07	8	0.26	12	0.75	7	2.34	11
INV (0,0.5)	0.30	2	1.44	9	3.66	2	7.92	4	0.07	6	0.26	10	0.75	8	2.33	10
INV (0,1)	0.30	4	1.44	6	3.66	3	7.94	6	0.07	5	0.26	8	0.75	9	2.31	9
INV (0,3)	0.30	6	1.43	5	3.72	8	8.01	7	0.07	4	0.26	9	0.76	11	2.31	8
INV (15,0)	0.30	8	1.44	8	3.73	9	8.33	12	0.07	9	0.27	13	0.78	12	2.30	7
INV (20,0)	0.30	7	1.44	7	3.79	10	8.11	9	0.07	10	0.26	6	0.76	10	2.40	13
ODD (0,0)	0.44	14	1.58	14	4.20	12	8.12	10	0.09	15	0.24	5	0.64	4	2.14	4
ODD (15,0)	0.52	15	1.73	15	4.22	13	8.02	8	0.08	11	0.20	1	0.45	1	1.58	1
ODD (20,0)	0.54	16	1.74	16	4.32	14	8.34	13	0.08	13	0.21	2	0.46	2	1.90	3
RW (0,0)	0.31	10	1.41	3	3.50	1	10.43	16	0.07	2	0.24	4	0.57	3	1.88	2
RW (15,0)	0.31	9	1.36	1	4.49	16	9.08	14	0.07	1	0.26	7	0.80	13	2.26	6
RW (20,0)	0.29	1	1.39	2	4.49	15	9.75	15	0.07	3	0.24	3	0.74	5	2.25	5
Median (0,0)	0.32	12	1.43	4	3.69	7	7.71	1	0.08	12	0.30	16	0.98	16	3.03	16
EW (0,0)	0.38	13	1.46	12	3.68	6	7.79	2	0.09	16	0.28	14	0.86	15	2.80	15
GC	0.31	11	1.46	13	3.79	11	8.30	11	0.08	14	0.29	15	0.84	14	2.41	14

Note: rk gives the rank over the column. ν = wind. size, where '0' indicates that all obs are used. λ = disc. factor.

used and specifies λ and ν in parenthesis. Columns two to nine display the MSE and the ranks across the column for the four different horizons for INSTABLE while the following columns give the corresponding results for STABLE.

Consider INSTABLE first. The best model for all forecast horizons is always a rank based method, whereby the corresponding ν differs for the RW for each horizon. Moreover, schemes that rank first for one horizon may have the last position for another one. For the one, three, six, and twelve month horizon forecasts the best models are RW(20,0), RW(15,0), RW(0,0), and Median, respectively. However, the first two are the worst models for six and twelve month horizon forecasts.

ODD combination schemes stick out as particularly inaccurate for all horizons. The application of a window leads to inferior results for most horizons except for the last one.

GC which exploits the ARCH effects found in-sample, is mediocre, mostly ranking 11th. However, the difference to the next better model is very small.

EW merely takes position 13 for the one-month horizon displaying a noteable difference to next better scheme. Still, the rank is getting better the longer the forecast horizon, even ranking second for the twelve month horizon.

INV schemes are in the middle-field, neither ranking worse than 12th nor better than second. INV give very similar MSE for all horizons. Only INV (0,3), INV (15,0), and INV (20,0) have markedly higher MSE for the two longest horizons. For all horizons except for the six month horizon, INV(0,0.5) is the best INV version whereas those that use a moving window show weak results.

Are there marked differences for STABLE? As could be expected given the analysis of the squared errors of the constituent models in section 5.1, the combined forecasts display a considerably lower MSE over all horizons when compared to the INSTABLE period.

The relative performances most noteably differ to the results of INSTABLE with respect to the improvement of ODD. While still ranking among the worst models for the 1-month horizon forecasts, ODD rank among the best models for the restant horizons, ODD(15,0) being first for all horizons. Median becomes even worse compared to INSTABLE, taking the last position for all horizons except for the shortest. Now, EW performs very poorly for all horizons, giving the worst results for the one month horizon. The GC based combination's relative performance deteriorates for all forecast horizons, as well. RW schemes give way to the improved ODD schemes. Nontheless, they gain for the longest and the shortest horizons. Again, INV schemes' MSE are very similar. Only INV(0,0), INV(0,0.2), and INV(20,0) give markedly higher MSE for the twelve month horizon forecasts.

As the results are very close it is indispensable to consider a test for statistical significant differences. In particular, none of the short horizon MSE of the INV schemes differ up to the second digit. The most widely used procedure to test for differences in forecast accuracy is proposed by Diebold and Mariano (1995). The h-steps ahead forecast errors of two competing models i and j at time t are given by $\{\hat{\epsilon}_t^i\}_{t=1}^T$ and $\{\hat{\epsilon}_t^j\}_{t=1}^T$. Accuracy is judged by the loss function $g(\{\hat{\epsilon}_t^i\}_{t=1}^T)$ and $g(\{\hat{\epsilon}_t^j\}_{t=1}^T)$. The sequence of loss differentials $\{\delta_t^i\}_{t=1}^T$ is defined as $\delta_t = g(\hat{\epsilon}_t^i) - g(\hat{\epsilon}_t^j)$. The test considers the null hypothesis $H_0 : E[\delta_t] = 0$ and is based on the observed sample mean

$$\bar{\delta} = \frac{1}{T^*} \sum_{t=T_0}^{T} d_t \tag{9.1}$$

with $T^* = T - T_0 + 1$. Here T_0 denotes the forecast origin. The sequence of forecast errors follows a MA$(h-1)$ process. If the autocorrelations of order h and higher are zero, the variance of the loss differential can be heteroscedastic and autocorrelation consistently (HAC) estimated as

$$\bar{V} = \frac{1}{T^*}(\hat{\gamma}_0 + 2\sum_{j=1}^{h-1} \hat{\gamma}_j) \tag{9.2}$$

where $\hat{\gamma}_j$ is the estimated $j - th$ autocovariance of the loss differential δ_t. Under the null hypothesis of equal forecast accuracy the DM test statistic can be computed as:

$$DM = \frac{\bar{\delta}}{\sqrt{\bar{V}}} \sim N(0,1) \tag{9.3}$$

To test if model i is not dominated by model j in terms of forecast accuracy a one-sided DM test has to be conducted. The modified null hypothesis is than given by $H_0 : E[\delta_t] \leq 0$. If the null is rejected model j dominates model i.

Table B.4 in the appendix presents the results of the cross comparision of the schemes for INSTABLE. Considering one model displayed in a column, positive (negative) values indicate that the model is better (worse) than the model in the respective row. However, significant difference are sparse. Only for the one month horizon there are two systematic patterns that can be identified: ODD are significantly worse than all other models at least at the five percent level for all forecast horizons considered and EW is significantly outperformed by all other models except for ODD.

Table B.2 gives the corresponding results for STABLE. For the six and twelve month horizon, the ODD schemes and in particular ODD(15,0) frequently outperform other approaches. For the shortest horizon, ODD(0,0) loses out to the rank based schemes and INV at the five percent level. EW, Median, and GC are significantly outperformed for the one, three and six month horizons by most alternative schemes. For the latter two this holds for the twelve month horizon, as well.

Comparing the simulation and the empirical analysis demonstrates that most of MC findings hold even for longer horizons though it merely considers one-step forecasts. The results confirm the dominance of the rank based method and the generally poor performance of EW, ODD, and GC. However, ODD for STABLE and EW for INSTABLE show very good results at longer forecast horizons. Given that it ranked 13th for the simulation analog of INSTABLE, Scenario D, GC performs considerably better than expected for the first horizon.

9.2 Filtered combined forecasts

Table 9.2 shows the MSE and ranks when FBP is applied. When compared to the unfiltered pendant the relative performance pattern remains mostly unaffected. Again, this is in line with the simulation that showed little changes in the ranking due to FBP.

For INSTABLE, GC emerges as the most accurate forecasting tool for the one month horizon forecast and as 7th for the twelve month horizon. Still, it is 12th and 13th for the three and the six month horizon forecasts. EW can improve from the 6th to the second, and from the second to the first position for the six month and the twelve month horizon. ODD gains for the two longest horizons. For STABLE the changes are mainly restricted to deterioration of RW(15,0) and RW(20,0) for some horizons.

The upper half of Figure 9.1 shows by how much the filtered and unfiltered versions of the schemes differ for INSTABLE. The red bars represent the percentage changes. A positive (negative) change that indicates an improvement (deterioration) of the filtered over the unfiltered alternative is represented by a red bar that extends to the right (left) of the zero point. The asterisks indicate the significance of the change according to the DM-test.

Most bars indicate an improvement of the filtered over the unfiltered alternative, except for eight schemes, whereby six of them are RW schemes for the two short horizons. While most improvements are significant at conventional levels, negative gains are never significant. Some of the improvements are considerable. In case of ODD schemes for the one and six month horizon forecast gains range from 13 to 20 percent. Looking at the different horizons, huge and highly significant gains emerge for the six month horizon. One and twelve month horizon forecasts are still markedly improved, while there are only two weakly significant gains for the three month horizon.

The lower half of Figure 9.1 displays the results for STABLE. For the three month and the six month horizon forecasts most of the models are significantly improved by the application of the filter. Overall, the extent of the gains for the six month horizon is comparable to INSTABLE. The predictions

Table 9.2: MSE and ranks; filtered

	INSTABLE								STABLE							
	$h = 1$		$h = 3$		$h = 6$		$h = 12$		$h = 1$		$h = 3$		$h = 6$		$h = 12$	
Type(τ, λ)	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk	MSE	Rk
INV (0,0)	0.29	2	1.43	7	3.42	4	7.69	2	0.07	6	0.26	10	0.71	7	2.26	10
INV (0,0.2)	0.29	3	1.43	9	3.43	5	7.70	3	0.07	8	0.26	11	0.71	8	2.24	8
INV (0,0.5)	0.29	4	1.43	8	3.45	6	7.73	4	0.07	7	0.26	9	0.71	9	2.23	7
INV (0,1)	0.29	5	1.43	6	3.47	7	7.78	5	0.07	5	0.25	7	0.72	10	2.22	6
INV (0,3)	0.30	6	1.43	5	3.56	11	7.88	10	0.07	4	0.25	8	0.72	11	2.22	5
INV (15,0)	0.30	9	1.44	11	3.59	13	8.23	13	0.07	9	0.26	13	0.72	12	2.25	9
INV (20,0)	0.30	8	1.44	10	3.64	14	7.94	11	0.07	10	0.25	6	0.71	6	2.34	13
ODD (0,0)	0.38	14	1.50	14	3.50	8	7.86	9	0.09	15	0.23	3	0.56	3	1.99	4
ODD (15,0)	0.42	15	1.58	15	3.50	9	7.86	8	0.08	11	0.20	1	0.43	1	1.59	1
ODD (20,0)	0.43	16	1.59	16	3.54	10	8.13	12	0.08	13	0.21	2	0.45	2	1.89	2
RW (0,0)	0.32	12	1.42	3	3.39	1	9.58	15	0.07	2	0.24	5	0.57	4	1.97	3
RW (15,0)	0.30	10	1.36	1	4.46	16	9.03	14	0.07	1	0.26	12	0.74	13	2.26	11
RW (20,0)	0.30	7	1.39	2	4.44	15	9.66	16	0.07	3	0.24	4	0.68	5	2.27	12
Median (0,0)	0.31	11	1.42	4	3.41	3	7.78	6	0.08	12	0.29	16	0.95	16	2.85	16
EW (0,0)	0.35	13	1.44	12	3.40	2	7.50	1	0.09	16	0.27	14	0.81	15	2.60	15
GC	0.29	1	1.46	13	3.59	12	7.79	7	0.08	14	0.28	15	0.81	14	2.46	14

Note: rk gives the rank over the column. ν = wind. size, where '0' indicates that all obs are used. λ = disc. factor.

Figure 9.1: Gain INSTABLE

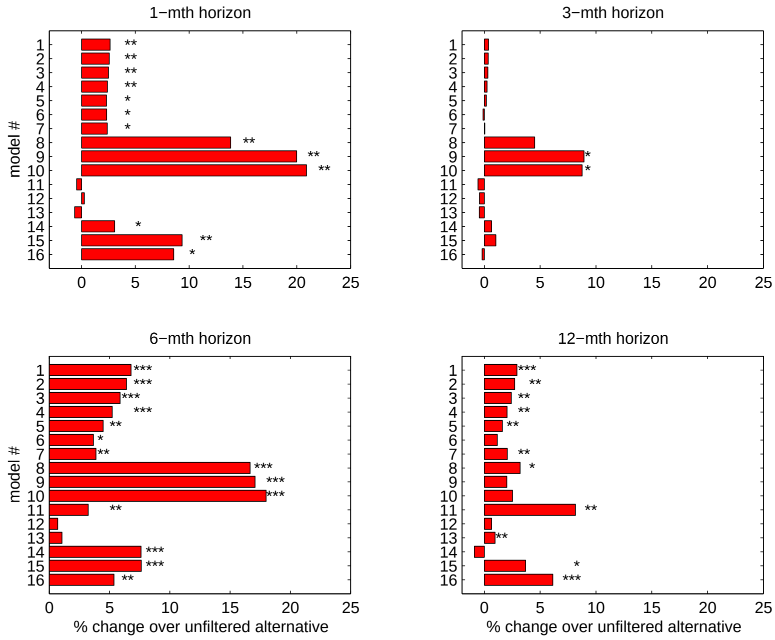
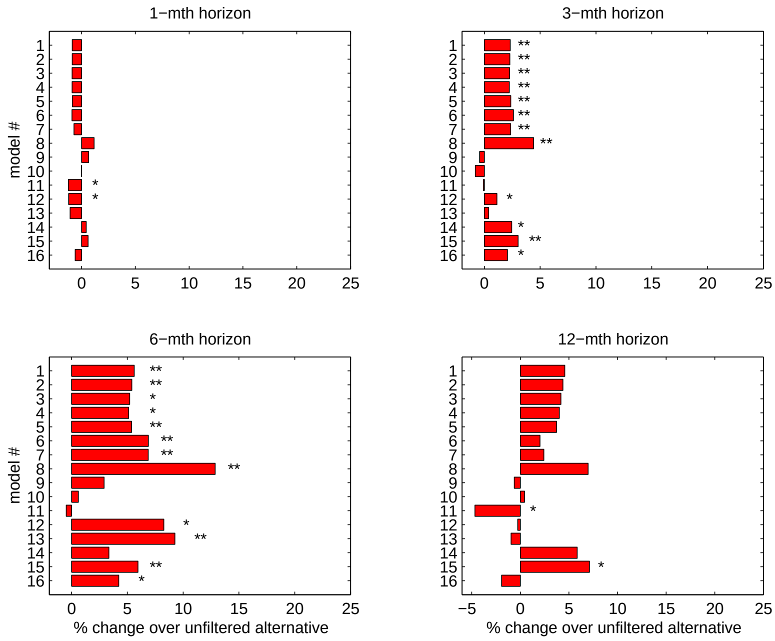


Figure 9.2: Gain STABLE



Note: 1=INV(0,0), 2=INV(0,0.2), 3=INV(0,0.5), 4=INV(0,1), 5=INV(0,3), 6=INV(15,0), 7=INV (20,0), 8=ODD (0,0), 9=ODD (15,0), 10=ODD(20,0), 11=RW(0,0), 12=RW(15,0), 13=RW(20,0), 14=Median (0,0), 15=EW (0,0), 16=GC; '*', '**','***' indicate significance at the ten, five, and one percent level.

for the one month horizon forecasts are virtually not affected, at all. Except for the weakly significant deterioration of model RW(0,0) and RW(15,0). Though there are high gains for the twelve month horizon only EW improves to the ten percent level. For RW(0,0) there is a weakly significant deterioration.

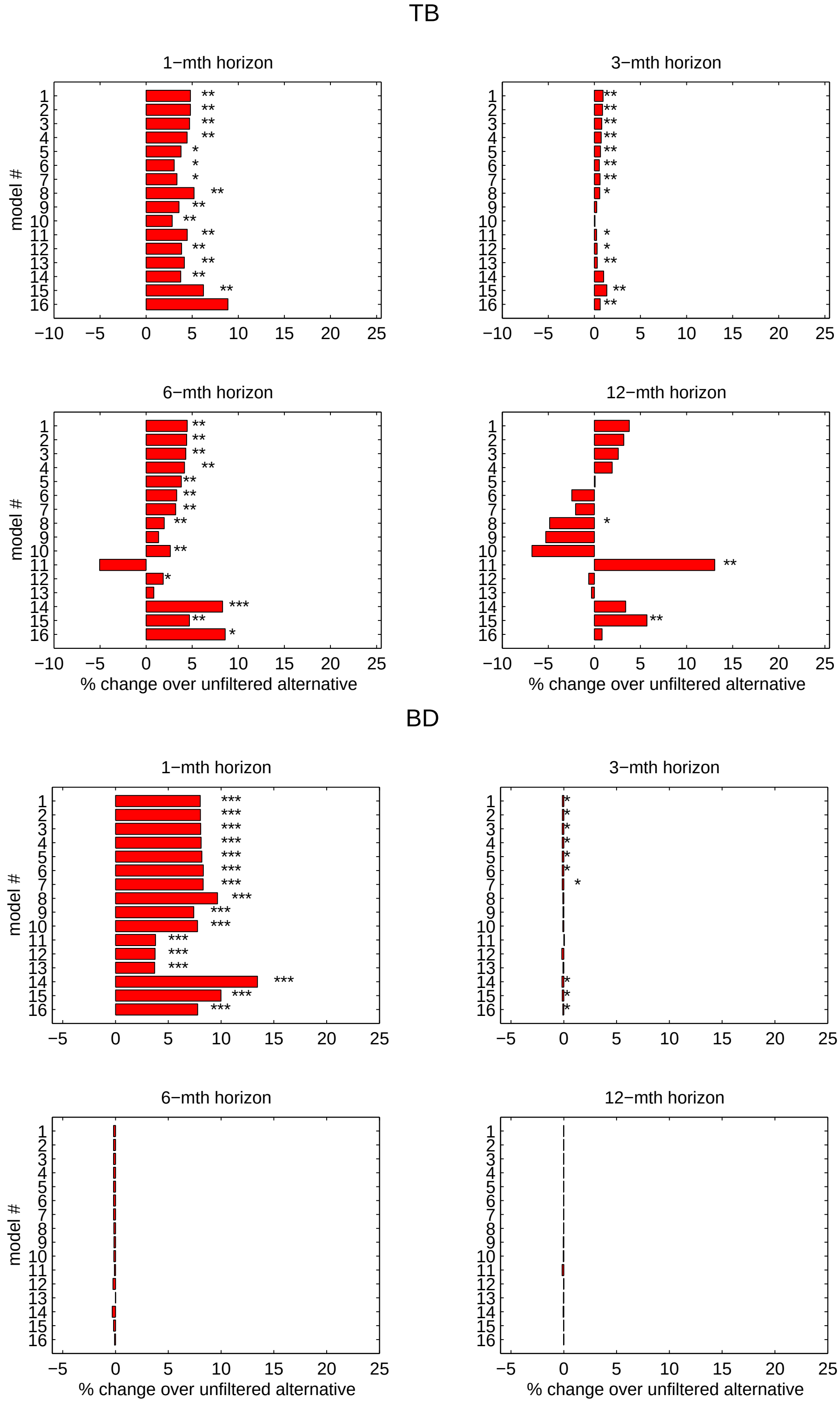
9.3 Other target variables

With the aim of testing FBP to the prediction of other variables than UK inflation, Figure 9.3 and 9.4 present the results for four other variables that are frequently forecasted in applied work: two of them, TB and BD, being financial and the other two, U and IP, being economic indicators. As the latter are real-time variables there are several options to define their 'true' values. Assuming the quality of the data to improve over time the results presented in the following reflect the deviation of the forecasts from the last vintage. As it is rather unusual to predict month-on-month changes for horizons up to one year, here, all of them are transformed to year-on-year percentage changes. In contrast to the analysis of inflation, the results are not split into sub-samples. The first 40 iterations are used as a learning period.

Where there are a significant differences at conventional levels, the bars indicate a positive gain. However, the size and the frequency differ across variables and forecast horizons. For the financial variables gains tend to be higher than for the economic variables.

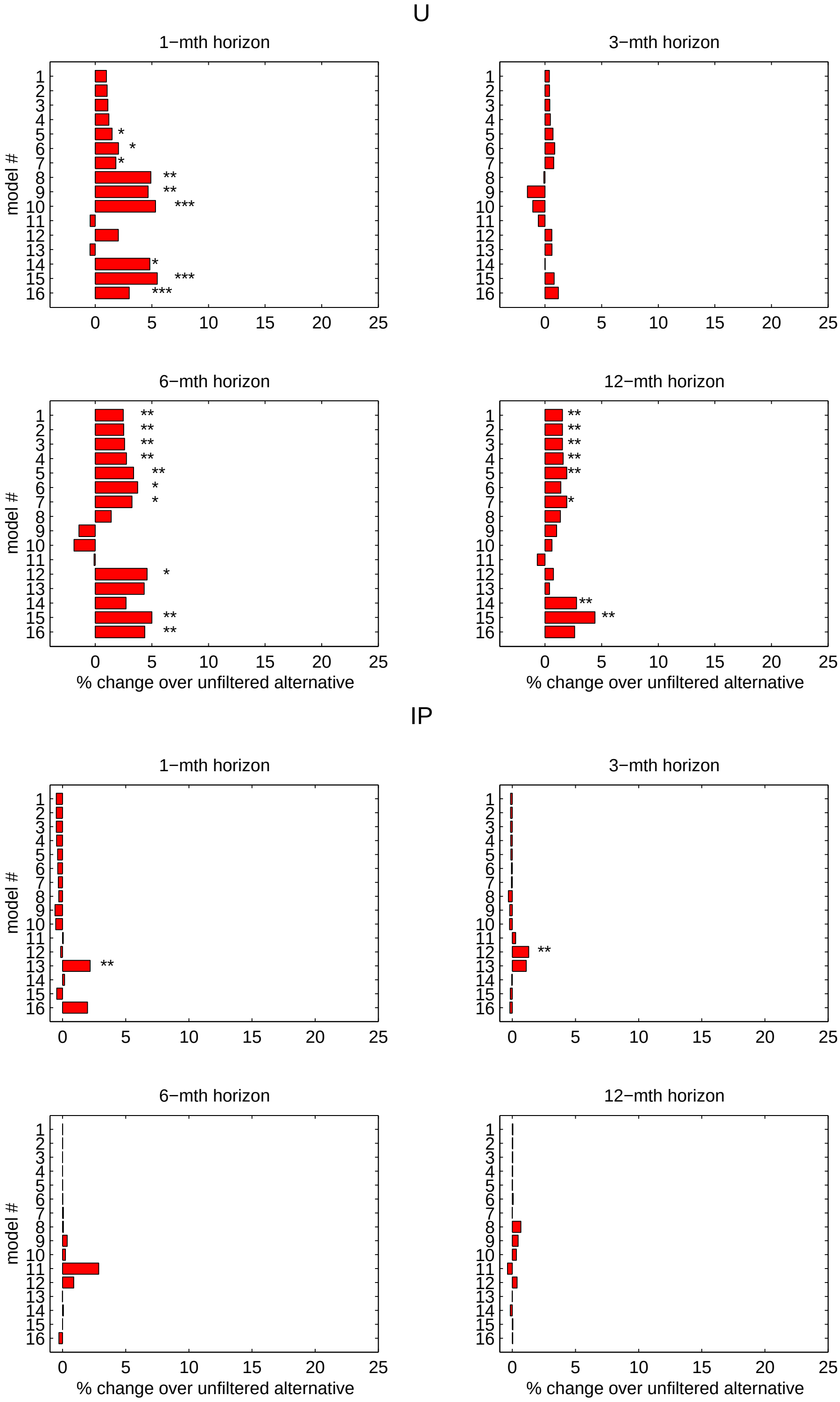
TB-forecasts can be improved at conventional levels up to the 6-month horizon. The gains for the 1-month and the 6-month horizon are mostly around five percentage points. For the 3-month horizon forecasts they are about two percentage points. For the 12-month horizon only RW(0,0) and EW can be improved significantly, yet the gains are relatively high. Furthermore, ODD(0,0) has a weakly significant loss of about five percentage points to the unfiltered alternative. Significant gains for BD are limited to the first two horizons. For the 1-month horizon the improvements are around five percent.

Figure 9.3: Gains for TB and BD



Note: 1=INV(0,0), 2=INV(0,0.2), 3=INV(0,0.5), 4=INV(0,1), 5=INV(0,3), 6=INV(15,0), 7=INV (20,0), 8=ODD (0,0), 9=ODD (15,0), 10=ODD(20,0), 11=RW(0,0), 12=RW(15,0), 13=RW(20,0), 14=Median (0,0), 15=EW (0,0), 16=GC; '*', '**','***' indicate significance at the ten, five, and one % level.

Figure 9.4: Gains for U and IP



Note: 1=INV(0,0), 2=INV(0,0.2), 3=INV(0,0.5), 4=INV(0,1), 5=INV(0,3), 6=INV(15,0), 7=INV (20,0), 8=ODD (0,0), 9=ODD (15,0), 10=ODD(20,0), 11=RW(0,0), 12=RW(15,0), 13=RW(20,0), 14=Median (0,0), 15=EW (0,0), 16=GC; '*', '**','***' indicate significance at the ten, five, and one % level.

For the 3-month horizons there are more gains than losses. However, all of them are very small and only weakly siginificant. Prediction errors of U can be significantly ameliorated at all horizons except for the 3-month horizon. Concerning IP the forecasts remain virtually unaffected. There are only two small significant gains: RW(20,0) for the one-month horizon and RW(15,0) for the 3-month horizon.

9.4 Comparison with a rule-based alternative

How does FBP compare to alternative approaches? Standard trimming is a poor contestant as it captures a different aspect of the covariance information: the past performance of the models. Given the compatibility of refinement techniques, it might be reasonable to combine FBP with standard trimming, rather than implementing either of them alone. Therefore, in the following FBP will be contrasted with an ad-hoc preselection method that is comparable in that it uses GARCH effects, as well. Though the direct use of these effects via conditional variance predictions in GC gave poor results, an alternative refinement scheme directly based on GARCH predictions might be promising. Similar to FBP, a forward looking version of trimming is employed. The models, whose conditional standard deviation of forecast errors is predicted to be highest are filtered out prior to combination. This gives the GARCH preselection (GP), defined as:

$$k_{sit}^{GP} = k_{sit} \cdot \mathrm{GP}_{it} / (\sum_{i=1}^{n} k_{sit} \cdot \mathrm{GP}_{it}), \tag{9.4}$$

where k_{sit} is the weight of model i at forecast origin t using weighting scheme s. GP_{it} is a binary variable, being zero, if $\hat{\sigma}_{it+h} > \mathrm{P}_{100-\alpha}^{\sigma_{t+\tau}}$, and one else. $\mathrm{P}_{100-\alpha}^{\sigma_{t+\tau}}$ is the $(100-\alpha)$th percentile of the predicted conditional standard deviation $\hat{\sigma}_{t+h}$. Following Granger and Jeon (2004), two alternative trimming factors ($\alpha_1 = 10$ and $\alpha_2 = 5$) are applied, meaning that the models predicted to have the ten and five percent highest conditional variances are excluded. The conditional variances are thereby forecasted using estimates

Table 9.3: Percentage gain of FBP over GP; INSTABLE

	INSTABLE								STABLE							
	$h=1$		$h=3$		$h=6$		$h=12$		$h=1$		$h=3$		$h=6$		$h=12$	
Type(τ,λ)	α_1	α_2	α_1	α_2	α_1	α_2	α_1	α_2	α_1	α_2	α_1	α_2	α_1	α_2	α_1	α_2
INV (0,0)	3.27	6.86	-0.48	-0.35	5.23***	5.3***	5.54*	4.72**	-3.09**	-1.16	1.3	2.2*	4.31**	3.34*	-0.85	-1.79
INV (0,0.2)	1.82	6.08	-1.47	-0.12	7.18***	7.51***	2.83*	2.97**	-5.77**	-1.4	0.17	0.99*	5.19**	5.36*	6.82	6.73
INV (0,0.5)	-4.93	-0.61	-1.68	-0.49	7.15***	7.78***	-0.94	0.66**	-5.41*	-3.08	-0.95	-1.4*	3.49**	3.5*	4.12	4.69
INV (0,1)	-0.92	-0.72	-1.47	-1	2.35***	1.99***	-3.05	0.68**	-0.64*	-0.71	1.56	1.69**	2.62**	4.54*	-0.51	-1.41
INV (0,3)	-0.45	-0.05	-0.86	-0.42	1.59**	1.27**	-2.14	0.72*	-0.78*	-0.91	0.79	1.64**	3.84**	5.39*	-1.14	-1.33
INV (15,0)	-0.76	-0.57	-0.52	-0.5	2.32	2.97*	8.82	8.54	-0.87**	-0.9	0.52	0.21**	-0.56**	-0.4*	-4.16	-4.44
INV (20,0)	8.49	15.62	0.75	4.26	11.48*	15.89*	3.65	3.22**	-3.87**	-2.98	3.86	1.5**	1.73**	2.03*	1.78	1.09
ODD (0,0)	8.09	15.18**	0.84	4.46	10.52***	14.63***	2.71*	2.49*	-3.24***	-2.03	3.1	1.06	3.7*	3.4*	0.67	0.38
ODD (15,0)	4.91*	9.92*	-0.99	1.03	11.35***	15.11***	4.27**	3.63**	-6.44	-1.88	-0.71**	0.86	7.65	9.99	8.89	7.26
ODD (20,0)	-0.77*	1.03**	-1.26	-0.52	4.65***	4.24***	0.26*	1.61*	-2.23	-1.06	1.22***	1.73*	4.55	4.79	2.83	1.9
RW (0,0)	-0.84*	0.88	-1.35*	-0.52	4.71***	4.14***	0.12**	1.15**	-2.5	-1.31	1.08**	1.87	5.28	5.23	2.24	1.47*
RW (15,0)	-0.73	1.09	-1.22**	-0.46	5.04	4.81	0.83	1.68*	-2.24	-1.14	0.88	1.66	4.17	4.26	3.95	3.01
RW (20,0)	-0.67*	1.15*	-1.17**	-0.43*	5.02*	5.23*	1.9	2.34	-2.36	-1.23	0.63	1.37	3.87	4.09*	4.44	3.64
Median (0,0)	-0.64	1.2	-1.19*	-0.43	5.28*	5.72**	2.48	2.81	-2.42**	-1.26	0.44	1.19	3.82*	4.14	4.73	3.98
EW (0,0)	-0.63	1.22*	-1.22	-0.43	5.51**	6.12***	2.89**	3.16**	-2.45***	-1.28	0.32	1.08	3.85**	4.26*	5.01*	4.29*
GC	-0.62	1.24*	-1.24	-0.43	5.69*	6.41*	3.16***	3.39***	-2.48**	-1.28	0.25	1.01*	3.89*	4.39	5.22	4.54

Note: rk gives the rank over the column. ν = wind. size, where '0' indicates that all obs are used. λ = disc. factor. '*', '**','***' indicate significance at the ten, five, and one % level. The trimming factors are $\alpha_1 = 10$ and $\alpha_2 = 5$.

of GARCH(1,1) processes fitted to the forecast errors up to the respective forecast origin.

The left part of Table 9.3 presents the percentage gains of FBP over GP for INSTABLE. The stars indicate the significance levels. The dominance of the FBP filter at longer horizons is striking, the maximum gain being 15.1 percentage points for ODD (15,0) over the five percent GP for the six month horizon. Though there are some – weakly significant – better results for GP for $\alpha = 10$ for the one and six month horizon, there are only two models, RW(15,0) and RW(20,0), $h = 3$ that gain more through GP at conventional significance levels. Especially for the six month horizon forecasts the FBP filter outperforms the GP filter for most of the models, indistinctly, whether the five percent or the ten worst models are excluded by GP.

The right part of Table 9.3 presents the analogous results for the STABLE period. The results are mixed. FBP still performs better for many schemes for the three and six month horizon. However, GP performs significantly better for the shortest horizon and $\alpha = 10$ for half of the schemes. There are no significantly different results for the twelve month horizon.

However, one of the major advantages of FBP is the objectivity with which the preselection of adaquate models is implemented. A comparison with a rule-based trimming scheme, is highly dependent on α. It will be a matter of data-mining to obtain results that favour the one or the other.

Chapter 10

Conclusion

Structural change is an omnipresent feature of economic data and models. As predictive tools require models to be reliable it represents a major source of forecast errors. Empirical and theoretical research – mostly analyzing structural breaks – have shown that the combination of many forecasts is a useful tool to address this issue. However, if past performance measures are employed to estimate adaptive weight sequences the researcher faces a trade-off between up-to-datedness and precision. This dissertation takes a closer look at the performance of the individual models to improve the latter proposing a novel refinement technique, the Forecast Breakdown Preselection (FBP). Building on the methodology of Giacomini and Rossi (2005) it makes use of forecast breakdown predictions to filter out unreliable models prior to combination. This reduces the uncertainty related to the estimation of the weight sequence, as it eliminates those models whose track record has become out-dated. In contrast to existing alternatives it offers an objective decision rule for the preselection of canditate models and – through the employment of ARCH dynamics – is a strictly foreward looking scheme. It is highly flexible in that it is compatible with most combination approaches and it directly reacts to deteriorations without requiring a learning period.

FBP as such does not yield forecasts and needs a combination scheme which uses the output of reliable models. Thus, the range of combination techniques most frequently found in the literature have been presented. As forecast error variances evolve over time, both simple and adaptive versions are considered. Adaptivity implies concentration on the more recent performance, discarding or downweighing possibly obsolete information. However, the smaller the number of errors the computation is based on the higher is the volatility of the weight estimates. As a consequence, forecast errors increase. Thus, a MC investigates how many observations are needed for past performance based methods to beat the natural alternative, the average forecast. It is demonstrated that the latter can be outperformed on the basis of a relatively short set of past forecast errors; the cost of processing information pays off when at least 15 observations of forecast errors are available. This holds for error covariance patterns typically found in applied work.

FBP itself is tested for in another MC which reflects the main features of the empirical analysis of the individual models: ARCH effects, phases of uncertainty, and structural breaks. The effect of the latter two are analyzed separately and in combination. Several robustness-checks investigate FBP's performance for a set of different relative forecast error variances and biltaral correlation patterns of the individual models. It demonstrates that the method leads to considerable improvements in forecast accuracy for all of the combination schemes employed. FBP gives the best results if there is a break or if there is a break and a phase of uncertainty. The gains are smaller the more bilateral correlation coefficients of the forecast errors tend to minus one and the smaller the relative performance differences of the individual models. Across the different scenarios and with or without FBP the relative performances of the combination schemes remain comparatively unchanged.

The relative accuracy as such merits a closer look, as this is the first study to consider such an ample set of combination schemes in the presence of these specific forms of non-linearities, in particular the ARCH effects. Rank weighted methods outperform the alternatives followed by INV schemes. The application of a rolling window is beneficial to RW, while it is detrimental to INV. However, the latter improves the faster past errors are discounted. Strikingly, EW which has been found to perform very good in the literature loses out against most of the alternatives. The same is true for GC, which correctly models the ARCH effects due to the set-up chosen, mostly ranked very poor. Furthermore, ODD, which is designed to give stable weight sequences in face of changing forecast error variances, gives mediocre results.

The empirical analysis largely corroborates the MC, though it considers multistep forecasts, as well. The experiment predicts UK inflation in a pseudo-out-of-sample fashion using real-time data. Applying the methodology of Bai and Perron (2003a) the variables in the data set are tested for structural breaks. Around an identified break of UK inflation in February 1991 forecast errors and the number of forecast breakdowns of individual models rise sharply. This allows for the identification of a period of increased forecast uncertainty (INSTABLE) and a 'normal' period (STABLE) period. Furthermore, the evolution of the forecast errors of many models can be explained by ARCH effects. For both, the more demanding INSTABLE and the STABLE period, the application of FBP leads to significant improvements in forecast accuracy over the unfiltered combination schemes. The percentage gain is mostly positive reaching up to 20 percent. There are a few situations when FBP leads to a loss, none of them being statistically significant. The results with respect to the relative performance of the single combination techniques found in the MC is largely confirmed. The extension of the analysis on four other economically meaningful variables in the data set confirms the usefulness of FBP. With the aim of conducting a comparison with a valid contestant, FBP is contrasted with a forward-looking version of the standard trimming approach. In contrast to FBP, it directly recurs to the ARCH effects. When the differences in accuracy were significant, FBP mostly outperformed the alternative. However, in the STABLE period for the shortest forecast horizon, the alternative is significantly more accurate.

Summarizing the results, FBP turns out to be a valuable refinement technique. This is confirmed by the simluation as well as the empirical analyses. In case of the latter, FBP provided considerable gains for many different combination schemes while there has been no significant deterioration – although several target variables and forecast horizons have been considered. Moreover, it is particularly useful in phases of high forecast uncertainty marked by structural breaks, increased forecast errors, and high numbers of forecast breakdowns.

Despite its success in the current real-time experiment the functioning of FBP still has to be tested for different data sets and time periods. However, the extension of the framework to cover other frequencies and contexts appears particularly promising: ARCH effects are even more widespread in weekly and daily financial market data than in economic data.

This dissertation proves the usefulness of ARCH-based approaches in the context of combined forecasts, which have largely been neglected in the literature. As the comparative scheme – directly using the ARCH effects – demonstrates there are still some possible alternatives that merit closer examination. E.g., alternative schemes could be based on exponentially weighted moving averages (EWMA) to predict conditional variances. However, objective decision rules need to be developed for such methods to be of practical use.

The estimation of many models paired with the framework of Giacomini and Rossi (2005) offers new research opportunities that go beyond the optimization of combination schemes. The number of models to be predicted to suffer a forecast breakdown could be employed as an indicator of rising forecast uncertainty. Moreover, it might prove to be a useful indicator for the business cycle analysis, as INSTABLE coincides with the beginning of an economic downturn.

Bibliography

AIOLFI, M., AND A. TIMMERMANN (2004): "Structural Breaks and the Performance of Forecast Combinations," mimeo, Bocconi University.

——— (2006): "Persistence in Forecasting Performance and Conditional Combination Strategies," *Journal of Econometrics*, 135(1-2), 31–53.

ANDREWS, D. (1993): "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 61(4), 821–856.

BAI, J., AND P. PERRON (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66(1), 47–78.

——— (2003a): "Computation and Analysis of Multiple Structural Change Models," *Journal of Applied Econometrics*, 18(1), 1–22.

——— (2003b): "Critical Values for Multiple Structural Change Tests," *The Econometrics Journal*, 6(1), 72–78.

BATES, J., AND C. W. J. GRANGER (1969): "The Combination of Forecasts," *Operations Research Quarterly*, 20(4), 451–468.

BERA, A., AND M. HIGGINS (1993): "ARCH Models: Proporties, Estimation And Testing," *Journal of Economic Surveys*, 7(4), 305–366.

BOLLERSLEV, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31(3), 307–327.

CHATFIELD, C. (1996): "Model Uncertainty and Forecast Accuracy," *Journal of Forecasting*, 15(7), 495–508.

Chow, G. (1960): "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28(3), 591–605.

Clemen, R., and R. Winkler (1986): "Combining Economic Forecasts," *Journal of Business and Economic Statistics*, 4(1), 39–46.

Clements, M., and D. Hendry (1994): *Towards a Theory of Economic Forecasting,* chap. 2, pp. 9–52, Non-stationary Time Series Analysis and Cointegration. Clements, M.P. and Hendry, D., Oxford University Press.

——— (1998): *Forecasting Economic Time Series.* Cambridge University Press.

——— (2005): "Evaluating a Model by Forecast Performance," *Oxford Bulletin of Economics & Statistics*, 67(1), 931–956.

——— (2006): *Forecasting With Breaks,* chap. 12, pp. 605–658, Handbook of Economic Forecasting. G. Elliott, C. Granger, and A. Timmermann, Elsevier.

Cooley, T., and E. Prescott (1973): "An Adaptive Regression Model," *International Economic Review*, 14(2), 364–371.

Croushore, D. (2006): *Forecasting With Real-Time Macroeconomic Data,* chap. 17, pp. 961–982, Handbook of Economic Forecasting. G. Elliott, C. Granger, and A. Timmermann, Elsevier.

Dickey, D., and W. Fuller (1979): "Distribution of the Estimators for Autoregressive Time Series With a Unit Root," *Journal of the American Statistical Association*, 74(366), 427–431.

Diebold, F., and P. Pauly (1987): "Structural Change and the Combination of Forecasts," *Journal of Forecasting*, 6(1), 21–40.

Diebold, F. X., and R. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Studies*, 13(3), 253–263.

Dueker, M. (1997): "Markov Switching in GARCH Processes and Mean-Reverting Stock-Market Volatility," *Journal of Business and Economic Statistics*, 15(1), 26–34.

Egginton, D., A. Pick, and S. Vahey (2002): "Keep it Real!: A Real-Time UK Macro Data Set," *Economics Letters*, 77(1), 15–20.

Elliott, G. (2004): "Forecast Combination With Many Forecasts," mimeo, Department of Economics, University of California, San Diego.

——— (2005): "Forecasting When There is a Single Break," mimeo, Department of Economics, University of California, San Diego.

Engle, R. (1982): "Autoregressive Conditional Heteroscedasticity With Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50(4), 987–1007.

Engle, R., C. Granger, and D. Kraft (1984): "Combining Competeting Forecasts of Inflation Using a Bivariate ARCH Model," *Journal of Economic Dynamics and Control*, 8(2), 151–165.

Fisher, W. (1958): "On Grouping for Maximum Homogeneity," *Journal of the American Statistical Association*, 53(284), 789–798.

Giacomini, R., and B. Rossi (2005): "Detecting and Predicting Forecast Breakdowns," University of California, Los Angeles, and Duke University Working Paper.

Goldfeld, S., and R. Quandt (1973): "A Markov Model for Switching Regressions," *Journal of Econometrics*, 1(1), 3–15.

Goyal, A., and I. Welch (2007): "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, 57(4), 1648–1649.

Granger, C., and Y. Jeon (2004): "Thick Modeling," *Economic Modelling*, 21(2), 323–343.

GRANGER, C. W. J., AND R. RAMANATHAN (1984): "Improved Methods for Combining Forecasts," *Journal of Forecasting*, 3(2), 197–204.

GUPTA, S., AND P. WILTON (1987): "Combination of Forecasts: An Extension," *Management Science*, 33(3), 356–372.

——— (1988): "Combination of Economic Forecasts: An Odds-Matrix Approach," *Journal of Business & Economic Statistics*, 6(3), 373–379.

HAMILTON, J., AND R. SUSMEL (1994): "Autoregressive Conditional Heteroskedasticity and Changes in Regime," *Journal of Econometrics*, 64(1-2), 307–333.

HANSEN, B. (2000): "Testing for Structural Change in Conditional Models," *Journal of Econometrics*, 97(1), 93–115.

——— (2001): "The New Econometrics of Structural Change: Dating Breaks in US Labor Productivity," *Journal of Economic Perspectives*, 15(4), 117–128.

HARVEY, D., S. LEYBOURNE, AND P. NEWBOLD (2001): "Analysis of a Panel of UK Macroeconomic Forecasts," *Econometrics Journal*, 4(1), 37–55.

HENDRY, D., AND M. CLEMENTS (2004): "Pooling of Forecasts," *Econometrics Journal*, 7(1), 1–31.

HOLT, C. (1960): *Planning Production, Inventories, and Work Force*. Prentice Hall.

JOUINI, J., AND M. BOUTAHAR (2003): "Structural Breaks in the US Inflation Process: A Further Investigation," *Applied Economics Letters*, 10(15), 985–988.

KANG, H. (1986): "Unstable Weights in the Combination of Forecasts," *Management Science*, 32(6), 683–695.

Lucas, R. (1976): "Econometric Policy Evaluation: A Critique," *Carnegie-Rochester Conference Series on Public Policy*, 1(2), 19–46.

Mackinnon, J. (1996): "Numerical Distribution Functions for Unit Root and Cointegration Tests," *Journal of Applied Econometrics*, 11(6), 601–618.

Makridakis, S., and M. Hibon (2000): "The M3-Competition: Results, Conclusions and Implications," *International Journal of Forecasting*, 16(4), 451–476.

Marcellino, M. (2004): "Forecast Pooling for European Macroeconomic Variables," *Oxford Bulletin of Economics & Statistics*, 66(1), 91–112.

Markowitz, H. (1952): "Portfolio Selection," *The Journal of Finance*, 7(1), 77–91.

Quandt, R. (1958): "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *Journal of the American Statistical Association*, 53(2), 873–880.

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

——— (1999): *A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series,* chap. 1, pp. 1–44, Cointegration, Causality and Forecasting: A Festschrift, In Honour of Clive WJ. Granger. Engle R, White H.

——— (2004): "Combination Forecasts of Output Growth in a Seven-country Data Set," *Journal of Forecasting*, 23(6), 405–430.

——— (2006): *Forecasting with Many Predictors,* chap. 10, pp. 515–554, Handbook of Economic Forecasting. G. Elliott, C. Granger, and A. Timmermann, Elsevier.

Timmermann, A. (2006): *Forecast Combinations,* chap. 4, pp. 135–196, Handbook of Economic Forecasting. G. Elliott, C. Granger, and A. Timmermann, Elsevier.

Tsay, R. (1988): "Outliers, Level Shifts, and Variance Changes in Time Series," *Journal of Forecasting*, 7(1), 1–20.

Winkler, R., and R. Clemen (1992): "Sensitivity of Weights in Combining Forecasts," *Operations Research*, 40(3), 609–614.

Appendix A

The variables employed

Table A.1: Sources of the variables and transformations

Variable	Acronym	Source	Description	transformation
Retail Sales	RS	Egginton	Volume of retail sales per week, all retailers (index)	mom % chg.
Industrial Production	IP	Egginton	Output for all production industries (index)	mom % chg.
Average Earnings	AE	Egginton	Average earnings in the whole economy (index)	mom % chg.
Monetary aggregate 0;	M0	Egginton	Average amount outstanding in the mth.	mom % chg.
Unemployment	U	Egginton	Total claimant count	mom % chg.
Long-term governm. bonds	BD	EcoWin	UK 10 Year Government Benchmarks, Yield, end of mth.	mom % chg.
Fin, Times Stock Exch. Ind.	FTSE	EcoWin	FTSE All Share	mom % chg.
Exch. rate, £/$	USD	EcoWin	Spot rate, end of period	mom % chg.
Exch. rate, £/DM	DM	EcoWin	Spot rate, end of period	mom % chg.
T-bill rate UK	TB	EcoWin	UK 3-mth. treasury bill, yield, end of mth., defl.with RPI	levels
Industrial trends	CBI	EcoWin	United Kingdom, average domestic prices, next three mths.collected by the the Confederation of British Industry	levels
Oil price	OIL	EcoWin	Brent, end of mth., USD	mom % chg.
T-bill rate US	TBUS	EcoWin	US 3-mth. treasury bill, yield, end of mth., defl.with US CPI	levels
Economic sent. indicator	ESI	EcoWin	Directorate General for Economic and Financial Affairs	levels
Business climate indicator	BCI	EcoWin	Directorate General for Economic and Financial Affairs	levels
Inflation	INFL	EcoWin	Retail Prices, index	yoy % chg.

Note: mom % chg. are mth. on mth. percentage changes; yoy % chg. are year on year percentage changes. Eggington refers to Egginton, Pick, and Vahey (2002); the data has been downloaded from the authors website, http://www.econ.cam.ac.uk/dae/keepitreal/ . January the first 2007. EcoWin is a commercial data provider. For details see www.EcoWin.com

Table A.2: Descriptives of the variables

	RS	IP	AE	M0	U	BD	FTSE	USD
Mean	0.27	0.09	0.62	0.41	-0.02	0.07	0.91	0.18
Median	0.28	0.19	0.54	0.42	-0.27	0.16	1.12	0.12
Maximum	10.65	6.76	4.24	3.70	6.26	11.06	52.62	13.31
Minimum	-9.78	-5.66	-1.70	-2.87	-4.51	-11.41	-26.60	-12.03
Std. Dev.	1.48	1.14	0.78	0.60	1.71	3.36	5.84	3.18
Skewness	0.51	0.05	0.81	0.08	0.97	-0.13	1.12	0.33
Kurtosis	20.28	9.08	6.31	12.66	4.51	3.76	17.57	4.56
# of Obs.	263	266	263	245	263	521	458	318

	DM	TB	CBI	OIL	TBUS	ESI	BCI	INFL
Mean	-0.03	1.44	23.75	0.75	1.45	-10.27	-7.92	6.24
Median	-0.23	2.84	22.00	0.00	1.66	-10.30	-5.05	4.60
Maximum	12.95	7.69	78.00	60.22	6.83	12.00	28.00	26.90
Minimum	-10.41	-16.67	-30.00	-30.54	-7.31	-32.00	-56.00	-0.80
Std. Dev.	3.26	4.51	23.93	9.57	2.12	9.54	16.12	5.04
Skewness	0.30	-1.78	0.29	1.49	-0.55	0.18	-0.43	1.70
Kurtosis	4.06	6.50	2.51	11.46	4.21	2.25	3.29	5.94
# of Obs.	318	438	305	306	522	318	294	625

Note: the # of obs. referes to the sample size available in November 2007.

Appendix B

Tests and figures

Figure B.1: L^{ew} as a function of ρ and κ

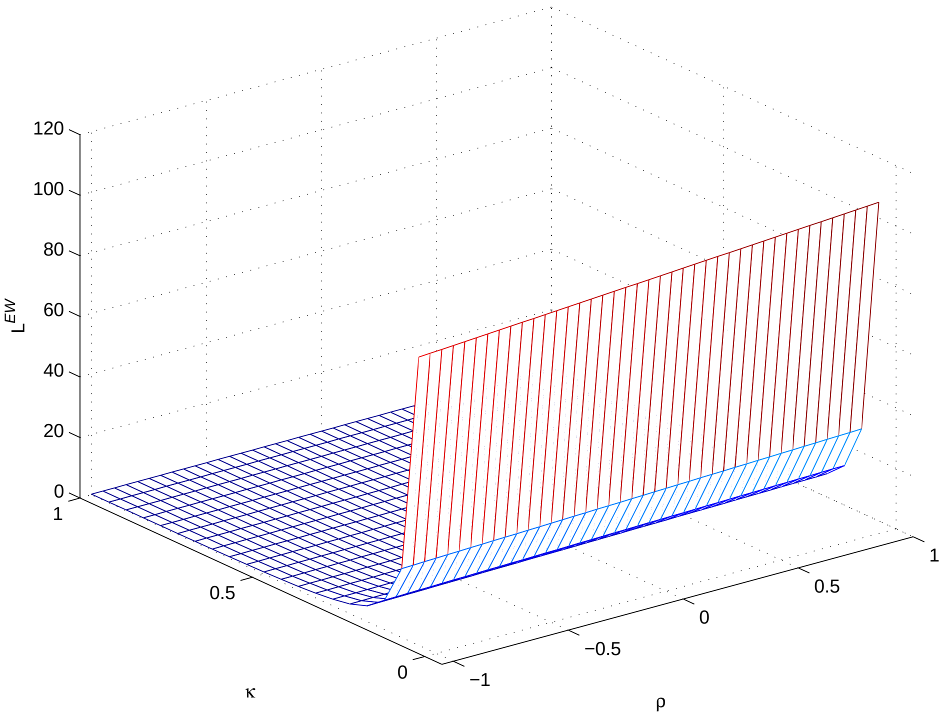


Table B.1: ADF-test results

Variable	# of obs	p	t_α	p-value
RS	433	1	-21.26	0.00
IP	470	0	-26.21	0.00
AE	526	4	-6.63	0.00
M0	441	0	-20.19	0.00
U	432	2	-4.67	0.00
BD	602	0	-19.95	0.00
FTSE	538	1	-17.34	0.00
USD	399	0	-18.68	0.00
DM	399	0	-19.14	0.00
TB	518	1	-2.86	0.05
CBI	381	3	-2.85	0.05
OIL	387	0	-19.73	0.00
TBUS	601	2	-3.47	0.01
CS	399	0	-3.35	0.01
BS	372	3	-3.58	0.01
INFL	692	6	-3.61	0.01

Results of the ADF test, where the lag length p is selected minimizing BIC.

1-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	0.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	-0.05	-0.25	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	-0.31	-0.47	-0.62	–	–	–	–	–	–	–	–	–	–	–	–	–
5	-0.34	-0.41	-0.42	-0.34	–	–	–	–	–	–	–	–	–	–	–	–
6	0.22	0.21	0.23	0.31	0.49	–	–	–	–	–	–	–	–	–	–	–
7	0.38	0.38	0.41	0.5	0.72	0.41	–	–	–	–	–	–	–	–	–	–
8	2.52***	2.51***	2.50***	2.49***	2.45***	2.29***	2.27**	–	–	–	–	–	–	–	–	–
9	0.69	0.68	0.68	0.69	0.7	0.63	0.62	-1.47*	–	–	–	–	–	–	–	–
10	0.87	0.86	0.86	0.87	0.88	0.81	0.79	-1.15	0.83	–	–	–	–	–	–	–
11	-0.8	-0.8	-0.8	-0.79	-0.8	-0.97	-1.05	-2.16**	-0.97	-1.13	–	–	–	–	–	–
12	-0.78	-0.79	-0.79	-0.79	-0.8	-0.98	-1.04	-1.98**	-0.99	-1.14	-0.33	–	–	–	–	–
13	-0.28	-0.28	-0.28	-0.27	-0.26	-0.4	-0.47	-1.73**	-0.72	-0.88	0.49	1.14	–	–	–	–
14	1.95**	1.94**	1.95**	1.98**	2.03**	1.86**	1.74**	-1.41*	0.02	-0.19	2.00**	1.94**	1.53*	–	–	–
15	3.17***	3.16***	3.15***	3.13***	3.07***	2.85***	2.81***	1.19	1.49*	1.24*	2.50***	2.28**	2.05**	1.92**	–	–
16	2.13**	2.13**	2.15**	2.17**	2.23**	2.21**	2.12**	-0.3	0.72	0.53	2.69***	2.61***	2.29**	1.21	-0.6584	–

3-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	0.04	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	-0.28	-0.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	-0.46	-0.59	-0.65	–	–	–	–	–	–	–	–	–	–	–	–	–
5	-0.15	-0.17	-0.1	0.11932	–	–	–	–	–	–	–	–	–	–	–	–
6	0.86	0.9	1.01	1.22	1.54*	–	–	–	–	–	–	–	–	–	–	–
7	-0.27	-0.28	-0.27	-0.21	-0.35	-1.97**	–	–	–	–	–	–	–	–	–	–
8	-1.32	-1.30*	-1.27*	-1.23	-1.23	-1.65**	-0.93	–	–	–	–	–	–	–	–	–
9	-1.20	-1.21	-1.21	-1.21	-1.26*	-1.45*	-1.31*	-1.01	–	–	–	–	–	–	–	–
10	-1.08	-1.08	-1.08	-1.09	-1.14	-1.34*	-1.18	-0.83	1.22	–	–	–	–	–	–	–
11	-0.67	-0.67	-0.67	-0.65768	-0.69	-0.93	-0.66	-0.19	1.64*	1.3*	–	–	–	–	–	–
12	-0.12	-0.13	-0.11	-0.07963	-0.11	-0.6	0.02	0.67	1.47*	1.32*	0.76	–	–	–	–	–
13	-0.87	-0.88	-0.88	-0.88581	-0.98	-1.40*	-1.08	-0.23	1.02	0.81	-0.05	-2.06**	–	–	–	–
14	2.44***	2.44***	2.43***	2.37***	2.12**	1.50*	1.8**	2.02**	1.58*	1.51*	1.26*	1.16	1.57*	–	–	–
15	2.15**	2.11**	2.05**	1.93**	1.52*	0.69	1.16	1.85**	1.36*	1.26*	0.95	0.68	1.19	-1.65**	–	–
16	1.8**	1.81**	1.85**	1.91**	1.86**	1.20*	1.72**	2.13**	1.67**	1.58*	1.39*	1.18	1.8**	-0.43	1	–

Note: 1=INV(0,0), 2=INV(0,0.2), 3=INV(0,0.5), 4=INV(0,1), 5=INV(0,3), 6=INV(15,0), 7=INV (20,0), 8=ODD (0,0), 9=ODD (15,0), 10=ODD(20,0), 11=RW(0,0), 12=RW(15,0), 13=RW(20,0), 14=Median (0,0), 15=EW (0,0), 16=GC; '*', '**','***' indicate significance at the ten, five, and one percent level.

6-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	0.29	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	0.19	0.11	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	0.12	0.7	0.5	–	–	–	–	–	–	–	–	–	–	–	–	–
5	0.36	0.36	0.4	0.5	–	–	–	–	–	–	–	–	–	–	–	–
6	0.3	0.3	0.3	0.32	0.25	–	–	–	–	–	–	–	–	–	–	–
7	0.8	0.7	0.7	0.7	-0.9	-1	–	–	–	–	–	–	–	–	–	–
8	-2.25**	-2.24	-2.24	-2.23***	-2.17**	-1.53*	-1.32*	–	–	–	–	–	–	–	–	–
9	-1.66**	-1.68**	-1.7**	-1.73**	-1.86**	-2.2**	-1.92**	-1.31*	–	–	–	–	–	–	–	–
10	-1.66**	-1.68**	-1.7**	-1.74**	-1.88**	-2.7***	-1.95**	-1.29*	0.57	–	–	–	–	–	–	–
11	-1.6*	-1.7**	-1.9**	-1.11	-1.2	-1.27*	-1.18	-0.53	2.31***	1.9**	–	–	–	–	–	–
12	0.4	0.39	0.4	0.41	0.38	0.48	0.67	1.25*	2.15**	2.23***	1.35*	–	–	–	–	–
13	-0.5	-0.6	-0.6	-0.6	-0.14	-0.32	-0.15	0.64	2.2**	2.8***	1.8**	-0.91	–	–	–	–
14	2.42***	2.41***	2.38***	2.32***	2.11**	1.64**	1.76**	2.4***	1.98**	2**	1.57*	1.4*	1.11	–	–	–
15	2.3**	2.23**	2.12**	1.95**	1.45*	0.8	0.95	2.37***	1.8**	1.81**	1.32*	0.34	0.57	-2.2**	–	–
16	1.44*	1.48*	1.53*	1.64**	1.74**	1.1	1.41*	2.66***	2.35***	2.37***	1.77**	0.32	0.72	-1.12	-0.18	–

12-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	-1.02	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	-0.86	-0.71	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	-0.7	-0.56	-0.42	–	–	–	–	–	–	–	–	–	–	–	–	–
5	-0.42	-0.3	-0.2	-0.1	–	–	–	–	–	–	–	–	–	–	–	–
6	-0.21	-0.15	-0.1	-0.06	-0.04	–	–	–	–	–	–	–	–	–	–	–
7	0.13	0.22	0.32	0.41	0.69	0.99	–	–	–	–	–	–	–	–	–	–
8	-2.04**	-1.96**	-1.8**	-1.55**	-1.02	-0.5	-0.9	–	–	–	–	–	–	–	–	–
9	-1.75**	-1.76**	-1.78**	-1.8**	-1.81**	-1.58*	-1.87**	-1.53*	–	–	–	–	–	–	–	–
10	-1.18	-1.17	-1.16	-1.17	-1.17	-1	-1.3*	-0.74	2.88***	–	–	–	–	–	–	–
11	-0.92	-0.91	-0.9	-0.9	-0.89	-0.76	-1.01	-0.57	1.52*	-0.05	–	–	–	–	–	–
12	-0.19	-0.16	-0.13	-0.11	-0.11	-0.16	-0.42	0.2	1.16	0.65	0.55	–	–	–	–	–
13	-0.18	-0.15	-0.13	-0.12	-0.12	-0.15	-0.38	0.17	1.1	0.6	0.53	-0.08	–	–	–	–
14	2.24**	2.18**	2.1**	1.99**	1.76**	1.36*	1.24*	2.44***	2.07**	1.77**	1.54*	0.97	0.89	–	–	–
15	2.2**	2.12**	2**	1.86**	1.52*	1.05	0.92	2.5***	2.01**	1.63*	1.4*	0.73	0.68	-1.89**	–	–
16	0.25	0.39	0.52	0.63	0.91	0.64	0.07	1.18	1.74**	1.23*	0.93	0.36	0.32	-1.5*	-1.15	–

Note: 1=INV(0,0), 2=INV(0,0.2), 3=INV(0,0.5), 4=INV(0,1), 5=INV(0,3), 6=INV(15,0), 7=INV (20,0), 8=ODD (0,0), 9=ODD (15,0), 10=ODD(20,0), 11=RW(0,0), 12=RW(15,0), 13=RW(20,0), 14=Median (0,0), 15=EW (0,0), 16=GC; '*', '**','***' indicate significance at the ten, five, and one percent level.

1-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	-0.15	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	-0.07	-0.02	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	0	0.03	0.05	–	–	–	–	–	–	–	–	–	–	–	–	–
5	0.22	0.24	0.27	0.33	–	–	–	–	–	–	–	–	–	–	–	–
6	0.35	0.37	0.39	0.44	0.53	–	–	–	–	–	–	–	–	–	–	–
7	0.5	0.54	0.59	0.73	1.68	0.83	–	–	–	–	–	–	–	–	–	–
8	2.76***	2.77***	2.78***	2.79***	2.83***	2.8***	2.83***	–	–	–	–	–	–	–	–	–
9	2.58***	2.58***	2.59***	2.6***	2.65***	2.66***	2.64***	2.07**	–	–	–	–	–	–	–	–
10	2.66***	2.66***	2.67***	2.68***	2.74***	2.76***	2.73***	2.22**	1.66**	–	–	–	–	–	–	–
11	0.76	0.76	0.75	0.72	0.55	0.39	0.45	-2.34***	-2.33***	-2.4***	–	–	–	–	–	–
12	0.19	0.2	0.2	0.21	0.17	0.07	0.06	-2.34***	-2.53***	-2.67***	-0.21	–	–	–	–	–
13	-0.57	-0.57	-0.57	-0.56	-0.57	-0.6	-0.7	-2.44***	-2.4***	-2.48***	-1.21	-0.42	–	–	–	–
14	0.97	0.98	1.01	1.05	1.09	0.98	0.96	-2.87***	-2.68***	-2.8***	0.11	0.43	0.96	–	–	–
15	2.57***	2.58***	2.6***	2.63***	2.69***	2.61***	2.69***	-2.95***	-2.46***	-2.55***	1.88**	1.89**	2.16**	2.57***	–	–
16	0.43	0.43	0.44	0.44	0.39	0.32	0.32	-2.07**	-2.27**	-2.36***	0.02	0.26	0.58	-0.04	-1.51*	–

3-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	0.09	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	-0.07	-0.19	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	-0.21	-0.3	-0.38	–	–	–	–	–	–	–	–	–	–	–	–	–
5	-0.15	-0.18	-0.17	-0.06	–	–	–	–	–	–	–	–	–	–	–	–
6	-0.04	-0.06	-0.01	0.17	0.51	–	–	–	–	–	–	–	–	–	–	–
7	-0.05	-0.06	-0.03	0.12	0.45	-0.04	–	–	–	–	–	–	–	–	–	–
8	0.98	0.95	0.93	0.9	0.86	0.86	0.82	–	–	–	–	–	–	–	–	–
9	1.18	1.16	1.14	1.13	1.1	1.11	1.07	1.36*	–	–	–	–	–	–	–	–
10	1.13	1.12	1.1	1.08	1.06	1.07	1.03	1.23*	0.29	–	–	–	–	–	–	–
11	-0.25	-0.25	-0.24	-0.21	-0.2	-0.23	-0.21	-1.02	-1.34*	-1.29*	–	–	–	–	–	–
12	-0.62	-0.62	-0.61	-0.59	-0.62	-0.66	-0.62	-1.17	-1.45*	-1.41*	-0.57	–	–	–	–	–
13	-0.52	-0.53	-0.52	-0.51	-0.56	-0.61	-0.57	-0.99	-1.25*	-1.21	-0.19	0.46	–	–	–	–
14	-0.13	-0.13	-0.1	-0.04	-0.02	-0.07	-0.06	-1.17	-1.29*	-1.22	0.18	0.5	0.36	–	–	–
15	0.71	0.63	0.59	0.56	0.47	0.43	0.39	-0.99	-1.2	-1.14	0.39	0.73	0.64	0.85	–	–
16	0.3	0.31	0.36	0.47	0.63	0.47	0.61	-0.57	-0.89	-0.87	0.34	0.74	0.75	0.27	0.02	–

Note: 1=INV(0,0), 2=INV(0,0.2), 3=INV(0,0.5), 4=INV(0,1), 5=INV(0,3), 6=INV(15,0), 7=INV (20,0), 8=ODD (0,0), 9=ODD (15,0), 10=ODD(20,0), 11=RW(0,0), 12=RW(15,0), 13=RW(20,0), 14=Median (0,0), 15=EW (0,0), 16=GC; '*', '**','***' indicate significance at the ten, five, and one percent level.

6-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	-0.07	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	-0.07	-0.07	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	-0.03	-0.02	0.01	–	–	–	–	–	–	–	–	–	–	–	–	–
5	0.2	0.24	0.31	0.45	–	–	–	–	–	–	–	–	–	–	–	–
6	0.18	0.21	0.26	0.35	0.07	–	–	–	–	–	–	–	–	–	–	–
7	0.36	0.4	0.48	0.64	1.08	2.89***	–	–	–	–	–	–	–	–	–	–
8	1.41*	1.31*	1.2	1.05	0.75	0.69	0.6	–	–	–	–	–	–	–	–	–
9	1.12	1.06	0.98	0.89	0.66	0.62	0.54	0.15	–	–	–	–	–	–	–	–
10	1.45*	1.37*	1.27	1.13	0.85	0.79	0.71	1.31	1.11	–	–	–	–	–	–	–
11	-0.28	-0.27	-0.25	-0.23	-0.27	-0.26	-0.32	-1.55*	-1.72**	-1.74**	–	–	–	–	–	–
12	1.16	1.21	1.3*	1.44*	1.69**	1.85**	1.73**	0.27	0.23	0.15	0.79	–	–	–	–	–
13	1	1.03	1.1	1.19	1.36*	1.46*	1.37*	0.25	0.21	0.14	0.72	-0.01	–	–	–	–
14	0.16	0.22	0.35	0.39	-0.2	-0.17	-0.44	-1.05	-0.87	-1.13	0.27	-1.34*	-1.13	–	–	–
15	0.09	0.13	0.22	0.34	-0.31	-0.25	-0.55	-1.02	-0.87	-1.11	0.25	-1.38*	-1.16	-0.1	–	–
16	0.45	0.51	0.62	0.83	0.84	0.66	0.05	-0.64	-0.58	-0.76	0.36	-1.44*	-1.18	0.52	0.77	–

12-month horizon forecasts

model #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
2	-0.07	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
3	-0.02	0.01	–	–	–	–	–	–	–	–	–	–	–	–	–	–
4	0.02	0.04	0.07	–	–	–	–	–	–	–	–	–	–	–	–	–
5	0.08	0.09	0.11	0.12	–	–	–	–	–	–	–	–	–	–	–	–
6	0.28	0.31	0.35	0.43	1.1	–	–	–	–	–	–	–	–	–	–	–
7	0.16	0.18	0.21	0.28	0.93	-0.74	–	–	–	–	–	–	–	–	–	–
8	0.29	0.26	0.21	0.15	0.06	-0.1	0.01	–	–	–	–	–	–	–	–	–
9	0.23	0.2	0.15	0.09	0	-0.17	-0.06	-0.26	–	–	–	–	–	–	–	–
10	1.06	0.87	0.66	0.45	0.21	0.01	0.16	0.58	5.53	–	–	–	–	–	–	–
11	1.17	1.12	1.04	0.93	0.73	0.59	0.71	1.54	1.28	1.12	–	–	–	–	–	–
12	1.04	1.15	1.37	1.9	3.44	1.69	3.32	0.55	0.74	0.52	-0.42	–	–	–	–	–
13	4.98***	5.17***	4.6***	3.12	1.52	1.02	1.48	1.99	3.27	2.77	-0.29	0.66	–	–	–	–
14	-0.2	-0.22	-0.27	-0.44	-1.63	-1.46	-2.5	-0.24	-0.22	-0.45	-0.85	-3.74	-1.96	–	–	–
15	-0.14	-0.15	-0.19	-0.35	-1.06	-1.13	-1.65	-0.2	-0.17	-0.42	-0.85	-4.22	-2.08	0.59	–	–
16	0.53	0.63	0.83	1.31	0.51	-0.03	0.32	0.13	0.25	-0.04	-0.77	-1.34	-1.98	1.16	1.23	–

Note: 1=INV(0,0), 2=INV(0,0.2), 3=INV(0,0.5), 4=INV(0,1), 5=INV(0,3), 6=INV(15,0), 7=INV (20,0), 8=ODD (0,0), 9=ODD (15,0), 10=ODD(20,0), 11=RW(0,0), 12=RW(15,0), 13=RW(20,0), 14=Median (0,0), 15=EW (0,0), 16=GC; '*', '**','***' indicate significance at the ten, five, and one percent level.

Eidesstattliche Versicherung

Ich versichere hiermit eidesstattlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie mir gegebene Anregungen sind als solche kenntlich gemacht. Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, 22. September 2008

Dirk Ulbricht

Curriculum Vitae

04/2006 – 09/2008	Ph.D. Student at the University of Munich, Junior Researcher at the Ifo Institute for Economic Research
10/2000 – 02/2006	Diploma in Economics, Free University of Berlin and University of Passau (Vordiplom)
07/1997	Abitur, Mollgymnasium, Mannheim
30.05.1978	Born in Mannheim, Germany