A New Approach for Detecting Shifts in Forecast Accuracy*

Ching-Wai (Jeremy) Chiu† Bank of England

Simon Hayes‡ Bank of England

George Kapetanios§ Kings College London

Konstantinos Theodoridis¶ Cardiff University

November 9, 2018

Abstract

Forecasts play a critical role at inflation targeting central banks, such as the Bank of England. Breaks in the forecast performance of a model can potentially incur important policy costs. Commonly used statistical procedures, however, implicitly put a lot of weight on type I errors (or false positives), which result in a relatively low power of tests to identify forecast breakdowns in small samples. We develop a procedure which aims at capturing the policy cost of missing a break. We use data-based rules to find the test size that optimally trades off the costs associated with false positives with those that can result from a break going undetected for too long. In so doing, we also explicitly study forecast errors as a multivariate system. The covariance between forecast errors for different series, though often overlooked in the forecasting literature, not only enables us to consider testing in a multivariate setting but also increases the test power. As a result, we can tailor the choice of the critical values for each series not only to the in-sample properties of each series but also to how the series for forecast errors covary.

Key words: Forecast Breaks, Statistical Decision Making, Central Banking

JEL classification: C53, E47, E58

*We thank Michael McCracken (the Editor) and an anonymous referee for their helpful comments and suggestions. We also thank Barbara Rossi, Raffaella Giacomini, Martin Seneca and Rodrigo Guimaraes, as well as participants from Central Bank Forecasting Conference and the Bank of England for their comments. Any views expressed are solely those of the authors and so cannot be taken to represent those of the Bank of England or members of the Monetary Policy Committee, Financial Policy Committee or Prudential Regulation Authority Board.

†jeremy.chiu@bankofengland.co.uk

‡simon.hayes@bankofengland.co.uk

§george.kapetanios@kcl.ac.uk

¶theodoridisk1@cardiff.ac.uk

1 Introduction

Economic forecasts are essential inputs into many policy decisions. At the Bank of England, lags in the transmission process of monetary policy means that the Monetary Policy Committee (MPC) sets the policy stance on the basis of forecasts of inflation, output growth and unemployment typically over a two to three year period. Also, forecasts of household borrowing and debt are key inputs into the Bank's Financial Policy Committee's (FPC) assessment of the resilience of the financial system.

Economic forecasting is difficult to do accurately, however. The economy is complex and dynamic, and, in many dimensions, not well understood. Relationships between economic variables can exhibit unexpected shifts. Sometimes the presence of a break is obvious. For example, the global financial crisis in 2008 arguably presented a structural break in the economy, but it nevertheless takes time to gauge the quantitative effects and to understand exactly how the economy has changed. Often, however, breaks occur more subtly and only become apparent with the passage of time and the accumulation of forecast errors in one direction.

Given this challenging environment, the models and judgements that underlie economic forecasts cannot be set in stone. Some that are relatively accurate at one point in time may be inaccurate in other circumstances, and may need to be modified or replaced. It is important, therefore, that economic forecasters maintain a close eye on the accuracy of their forecasts, and much of the skill in forecasting lies in judging when a forecast breakdown has occurred and in taking appropriate action. To this end, we focus on a test for a change in the mean forecast error.

A number of approaches have been put forward in the macroeconometric literature to support this activity. Most involve some testing procedure which compares forecasts from two different subperiods for evidence of difference in performance. By nature, these testing procedures are related to tests of structural change. In particular, they are closely linked to fluctuation type tests, such as those developed by Brown, Durbin, and Evans (1975) and Ploberger, Kramer, and Kontrus (1989) (see, also, Kuan and Hornik (1995)), that aim to detect structural change without entertaining a particular alternative hypothesis and are therefore, in theory, at least, robust to a wide variety of deviations from stationarity.

One issue with such procedures is that the associated tests can have low power and, therefore, not pick up forecast breakdowns for some time after they occur. This issue clearly illustrates the asymmetry between the null hypothesis of no breakdown, whose probability of being rejected, while true, is supposed to be well controlled, and the alternative of forecast breakdown which can have a low probability of being selected, even if true. To address this asymmetry, we consider data based rules, based on statistical decision making, to select significance levels. This enables quicker detection of breakdowns, albeit at the cost of more false positives. This trade off is managed by explicitly considering loss functions that quantify the costs of Type I and Type II errors.

Therefore, in this paper we describe a set of new procedures that we have developed for identifying shifts in forecast accuracy. These procedures are designed to support a more regular, extensive and systematic monitoring of the forecast accuracy of the Bank's main forecast outputs. As such, they address the Bank of England Court's request for a better basis for evaluating the Bank's forecast

performance, and builds on the analysis of, and recommendations relating to, forecast performance produced by the Bank's Independent Evaluation Office (IEO) (Bank of England, 2015).

In its analysis of forecast errors, the IEO employed standard methods of statistical inference, while noting that such methods were not necessarily ideal given the properties of the data. In particular, the available time series for forecast errors is relatively short and exhibits considerable serial correlation. As a result, standard tests for structural breaks, derived without correcting for serial dependence and asymptotic approximations, tend to have problematic behaviour under the null hypothesis of no breaks; in other words, there are nominal size distortions. Further, even if the size distortions are corrected, such tests tend have low power to reject the null hypothesis of no breaks, as discussed, for example, in Groen et al. (2013).

The procedure we present aims at warning the analyst that a break might be occurring earlier than standard testing procedures would. It does so by taking fuller account of the empirical properties of the data — making appropriate allowances for small sample sizes and the observed serial dependence of forecast errors — and by putting the notion of loss at the centre of the forecast evaluation assessment, i.e. flagging shifts in forecast performance only to the extent that they are relevant to the forecaster's objectives. This explicitly recognises the fact that in many contexts some forecast errors are more important than others, depending on the use to which the forecasts are to be put. We believe this will enhance the quality of discussion surrounding developments in forecast accuracy and lead to sounder judgements on model design and use.

In practice, our exercise makes the case that standard forecast breakdown tests, such as those proposed by Giacomini and Rossi (2009), used to detect forecast failure be made more sensitive by being allowed to reject the null hypothesis of no shift in forecast accuracy more often. This essentially involves selecting less conservative significance levels for the tests than is normal practice. These significance levels are optimally chosen by explicitly trading off the losses associated with the two errors associated with any decision making: taking action when no action is needed and not taking action when action is needed. Of course, this approach is not confined to forecast breakdown tests, but applies more widely to statistical tests that detect or monitor structural change, such as Bai and Perron (1998) and Chu et al. (1996).

In order to appreciate the novelty and purpose of this approach it is important to present its decision theoretical background. A large literature on statistical decision problems exists, starting with Wald (1950) and Savage (1954), followed by DeGroot (1970) and Berger (1985). Our paper builds on these works and focuses on specifying a decision rule for making a choice in the presence of uncertainty about the true state of the world. Many approaches to this problem have been proposed. As discussed in Granger and Pesaran (2000a), Granger and Pesaran (2000b), Pesaran and Skouras (2002) and Granger and Machina (2006), one approach focused on forecasting involves selecting a course of action that minimises an expected loss function in the presence of some forecasting information. The forecasting information can be either a point forecast or a more general conditional forecast distribution. Our theoretical setup is consistent with the spirit of this strand of literature.

Statistical testing relates to this literature since testing has two outcomes and therefore provides a decision rule that depends on the choice of the significance level. While Granger and Machina

(2006) acknowledge the possibility of transposing their framework to a statistical testing context, they do not pursue this possibility.

Recent work by Tetenov (2012) transposes this machinery to the microeconomic context of deciding on the extent of using different treatments on a population based on a loss function and information on the efficacy of these treatments. In that work, the link between reaching decisions using standard decision rules and statistical tests is discussed but testing conventions such as using standard significance level are not relaxed.

In a follow-up paper, Tetenov (2016) continues to focus on the above microeconomic treatment decision framework and derives the optimal significance level of a test as a function of test power and costs of treatment adoption. While this moves closer to our framework of viewing the significance level of a test as a tuning parameter to be chosen by optimising a loss function, it is very specific to the treatment problem at hand. In our case, we initially state the general problem by leaving the losses incurred by each course of action available to the agent, unspecified. We then proceed to specialise our analysis to the forecast breakdown problem and discuss, in detail, ways in which this specialisation can be implemented.

The paper is structured as follows. Section 2 discusses the challenges of the current forecast breakdown procedures faced by policy makers. Section 3 gives an overview of our proposed solutions. Section 4 and 5 offer theoretical discussion and step-by-step procedures on our proposed testing procedures. Section 6 provides Monte Carlo evidence that our proposed algorithm performs better in most of the scenarios. Section 7 discusses the empirical results when we apply our procedures to the forecast produced by the Bank of England. Section 8 concludes.

2 Motivation

Many policy decisions rest on forecasts, so the accuracy of the latter will affect the appropriateness of the former. As mentioned above, the optimal stance of monetary policy depends on an accurate forecast for inflation, which, in turn, requires an assessment of the cost pressure outlook, so a forecast for wages for example. The forecast for wages may hinge in turn on the projected path for productivity. If the Great Recession caused a permanent fall in the level and/or growth rate of productivity that forecasters were slow to detect, a bias is likely to show up in the wage forecast which could then snowball into poor forecasts for other variables, such as inflation, and, ultimately, in suboptimal policy decisions.

Forecast breakdowns are, however, not easy to identify. Almost ten years after the start of the Great Recession, we only have about 30 post-recession forecast errors for a quarterly series. Identifying forecast breakdowns with samples this small is no easy feat, but we think the procedure we propose in this paper can help.

2.1 Four Challenges with the Textbook Procedure

The statistical test we consider in our analysis compares the bias in a base sample (which we also refer to as a pre-break sample) and in a subsequent post-break sample and we tailor our approach

so as to capture the small sample-size characteristics of our series. We proceed with the Giacomini and Rossi (2009) test but our procedure can also be applied to other statistics, such as the forecast errors' second moments (although working with first moments is more useful for our application). In particular, it relates more directly to the underlying economic dimension of the problem in that it makes it easier to identify what specific model or judgement may be the cause of the break.

Many papers in the literature assume forecast errors to be *identically and independently normally distributed*, and that a researcher would use a 5% as the confidence level, also known as the size of the test. Although this procedure is a well-trodden path and widely understood, it is not ideal for policy making in real time. There are four reasons for this.

1. **The sample sizes with which we work are unavoidably small**. Standard tests are shown to have good power against the null of no break in the mean forecast error for sample sizes of at least 100 observations (Giacomini and Rossi (2009)). This is natural, given they are derived under assumptions that hold asymptotically. Policy makers, typically dealing with quarterly series, cannot afford to wait that long so they would naturally put a premium on a testing strategy that would increase the power of the test on shorter samples.

2. **The forecast errors in our dataset are usually serially correlated**, which reduces the information content of each newly observed forecast error. As a result the actual size of the test in a "policy-relevant" sample can be significantly different from the nominal size computed asymptotically and the power of the test can be disappointingly low.

3. **The test size is set to 5 percent almost by default**. It is such standard practice to use a 5 percent critical value in statistical testing that this choice is rarely questioned. However, given that this figure should reflect the share of false positives that is acceptable to the decision maker, there is no particular reason why it should be the same for all enquiries.

4. **The standard test procedure has no explicit consideration of the power of the test**. The power of the test is dependent (amongst other factors) on the size of the break in forecast performance that has occurred, which is series-specific.

3 Proposed Solutions

It should be clear by now that the problems we face are both conceptual and practical. On the conceptual side we want to overcome the *lexicographic structure* that underlies testing procedures (as Tetenov (2016) refers to it). In other words, we aim at relaxing the prominence that *Type I* errors have relative to *Type II* errors. It is not obvious that *false positives*(Type I errors) incur greater policy costs than *false negatives*(Type II errors) so we want our procedure to be more flexible in allowing us to trade off the two types of error in an optimal way. In this sense, our work relates to the recent literature on decision-making using statistical testing (in particular Tetenov (2012) and Tetenov (2016)), in that the optimal size for our tests (which will ultimately determine whether a forecasting model is updated or not) is obtained by trading off the two error types according to a loss function.

3.1 Two-state, two-action problem

In order to select optimally the size of our test, we need a criterion capturing how often a policy-maker is willing to tolerate a false alarm to catch a break promptly.

The specific problem that we face is the decision of whether or not to adjust a forecast model on the basis of an emerging indication of a shift in forecast accuracy. We can think of it as a 'two-state, two-action' decision problem, as is illustrated in Table 1, which shows the costs associated with the decision given the unobserved true state. The diagonal elements correspond to situations in which the correct decision is made, i.e. when the model is adjusted and there has been a break (bad state), and when the model is not adjusted and there has been no break (good state). In this general set-up, we denote the losses as L_{yb} and L_{ng} respectively, although these may be zero. The off-diagonal elements — adjusting the model when there has been no break and failing to adjust the model when a break has occurred — are assumed to entail losses of L_{yg} and L_{nb} respectively. These are the costs associated with running with an incorrectly specified model, one that will systematically produce forecasts exhibiting a bias relative to that of those made in-sample.

Table 1: A two-state, two-decision problem

	Bad state (a break)	Good state (no break)
Action taken ($d = 1$)	L_{yb}	L_{yg}
No action taken ($d = 0$)	L_{nb}	L_{ng}

The choice of loss function L is an arbitrary one, reflecting the decision-maker's preferences. L is defined as the norm of the difference between two vectors: (i) a vector of forecast biases resulting from a forecast breakdown in variable x; (ii) a vector of forecast biases when a variable x displays no forecast breaks.

Such a definition of L enables us to accommodate a general loss function, allowing us to capture the forecast bias of variables which may not necessarily be the same as the breaking variables. Throughout the paper, we use μ to denote the forecast bias of *a single variable*, for example, μ^x refers to the forecast bias of the *breaking variable* x. We use the upper case $\mathcal{M}$ to denote the vector of forecast biases of *a group of variables* such as the *headline variables of interest* to the researcher.[1] We also denote b^x as the magnitude of the forecast break (i.e. the change in forecast bias) for variable x across base and evaluation samples. Using a subscript '0' to denote the base sample, L is therefore defined as follows:

$$L(b^x) = ||\mathcal{M}(\mu_0^x + b^x) - \mathcal{M}(\mu_0^x)|| \tag{1}$$

This loss function is a function of the magnitude of the break b^x; and it is zero in the absence of a break; and it is positive when a break of either sign occurs.

We choose α, the optimal nominal size, on the basis of our expected losses. Specifically, we choose α so as to minimise the expected loss from the two-state, two-action problem. The probability of

[1] We will also use $\mathcal{M}^z$ to denote the forecast bias of variable z in the forecast bias vector. By definition, $\mathcal{M}^x = \mu^x$.

mistakenly changing the model is the probability of a Type I error, i.e. α itself, and the probability of mistakenly failing to change it is the probability of a Type II error, i.e. the probability of a break going unnoticed or $1 - \mathcal{P}$ (one minus the power of the test). $\mathcal{P}$ itself depends on α, as well as on the magnitude of the break under the alternative hypothesis (b) and the size of the sample (T). We will get into the details further below; for now it is important to note how this approach allows us to formalize the tradeoff between the size and the power of the test. Increasing α will increase the probability of taking action when in fact none would be required, but also increases the power ($\frac{\partial \mathcal{P}}{\partial \alpha} > 0$) thus diminishing the probability of *not* taking action in the 'bad' state. The optimal size will be the one that equates the marginal cost of increasing further to its marginal benefit.

3.2 A multivariate approach

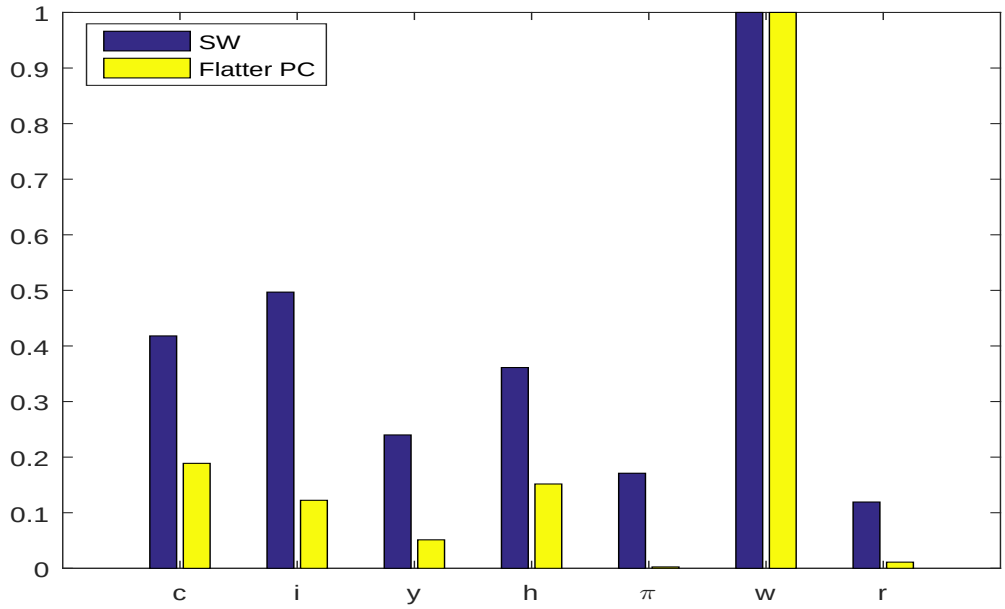
The above two-state, two-action decision problem is general enough to be applied to both univariate and multivariate settings. Much of the forecast breakdown literature tends to focus on one variable at a time. Yet, policymakers consider their variables of interest as a multivariate system. An excerpt from a recent *Inflation Report* publication by the Bank of England illustrates the importance of the multivariate dimension of the problem:

> CPI inflation had remained at -0.1% in October, as expected. The lower price of oil increased the likelihood that headline inflation rates would remain subdued in the near term. In addition, nominal wage growth had levelled off. Average hours worked had been lower than expected, however, which might have explained some of the flattening off in pay growth, with changes in the composition of employment an additional factor. To the extent that these were reflected in productivity as well as pay, their implications for inflation were likely to be small. A third potential factor behind weak pay growth was the low level of CPI inflation seen during the course of the year, which may have fed into pay negotiations.
>
> (Bank of England (2016), p. 3)

This quotation is taken from the section of the Inflation Report that discusses the evolution of variables of interest since the previous issue. What is interesting for our purposes is that the outturn for inflation is not explained in terms of some past value of inflation itself, as one would expect in a univariate setting. Rather, it is related to energy prices and labour market conditions. In turn the weak level of wages is understood as potentially stemming from inflation. This interdependence of variables is typical of the macroeconomic discussion and we will try to have our analysis reflect this.

The quotation above also clearly illustrates the sense in which some variables have a headline status while others play more of an instrumental role. Let us assume the end goal is to produce the most accurate forecast for inflation. An efficient allocation of scarce resources would imply that attention should be devoted to the variables whose forecast performance breakdown would mostly affect inflation. Our multivariate procedure will involve the estimation of the degree of

Figure 1: Bias simulation using Smets and Wouters (2007)



Note: Bias (in absolute value and relative to the bias in wages) in consumption, investment, output, hours, inflation, wage growth and the policy rate; Smets and Wouters (2007) calibration (blue) and 'flatter Phillips Curve' scenario (yellow).

co-variation in forecast errors for our variables of interest so as to be able to address this point. To get a sense of how that works, though, we begin considering a similar example in a controlled setting.

A DSGE model is the ideal environment in which the interaction of various macro variables can be studied. The model presented by Smets and Wouters (2007) is a popular benchmark for any policy-relevant DSGE, so we use it to conduct the following experiment. Suppose a judgment was made on labour market conditions that resulted in a one-percent bias in the forecast for wage growth.[2] Policy makers taking Smets and Wouters (2007) and applying this judgement on wages would produce biased forecasts for all other variables, as the blue bars in Figure 1 show. An incorrect assessment of the future evolution of wage growth would impact the assessment of cost pressures and so the forecast for inflation. In turn, this would affect the prediction for short-term rates which ultimately influences the intertemporal consumption and investment decisions. In the end, an unwarranted judgement on wages spills over to all the other variables.

The structure of the economy critically affects the degree to which this is the case though. If we repeat a similar experiment in a version of the same model in which the Phillips Curve is flatter the spillover of the incorrect judgement on wages onto other variables is much more muted because inflation responds a lot less to an incorrect forecast for marginal cost.[3]

So, if we think of inflation as our key variable of interest, forecast breaks in wages would matter much more in the former case than they would in the latter. The multivariate procedure we

[2]For concreteness, we assume a bias in the wage markup shock process, the shock more directly related to wage determination. Of course there is a degree of arbitrariness in the selection of the shock, yet this is enough to illustrate our point.

[3]Implemented by increasing the probability of firms not being able to re-optimize prices to .985.

employ acknowledges this interdependence and captures it by means of a multivariate model of the observed forecast errors. The observed covariance of the forecast errors will influence the selection of the critical values used to determine whether a suspect forecast break is worth attending to.

3.3 Data-driven calibrations

Since eliciting the preferences on power and size is not straightforward, we propose to have the loss function and most every other characteristic of our setup driven by the underlying data. A key feature of our data is the small sample size, which reduces the power of tests. Testing for breaks at the customary 5% significance value would, in practice, mean that forecast breaks would go unnoticed for a long time, causing delays in adopting the appropriate policy measures. So not only do we adopt a loss-function based approach but also we tailor our procedure to the specific characteristics of each of the series under consideration.

In particular, we move away from asymptotic results by simulating the distribution of our test statistics given the properties of the sample of data. So, for each of the series we consider, we simulate the distribution for the Giacomini and Rossi (2009) test under the null hypothesis of no break in the *mean* forecast error for each variable and forecast horizon.

As a result, the critical values we use to evaluate the null hypothesis of no breaks are made to depend ultimately on the properties of the individual series; in particular on the likelihood of observing a break. The larger the likelihood of observing breaks, the larger the power the test has and hence renders smaller critical values.

4 Theoretical Considerations

As noted in Section 3.1, the loss function is constructed based on a two-state, two-action decision problem. In the standard setting the decision maker takes a decision between two actions (or rather action and inaction) each of which has a different payoff depending on the state of the world (good or bad, or in our case no forecast breakdown against forecast breakdown). The timing structure is one where the action is conditional on the state of the world which is observed via a noisy signal.

In terms of the loss function, we implicitly assume that $L_{yb}(\boldsymbol{b}) << L_{nb}(\boldsymbol{b})$ and that $L_{ng}(\mathbf{0}) << L_{yg}(\mathbf{0})$. Moreover, the loss depends on $\boldsymbol{b}$, the size of break which is not observable to policy makers. The decision is a function of an econometric testing outcome which aims to uncover the occurrence or not of the bad state. In particular, we specify that a test is carried out based on an observed sample of size T. The test statistic S_T leads to action being taken if it exceeds a particular threshold given by c_α which is a function of a tuning parameter α (the nominal size, or rejection probability under $\boldsymbol{b} = 0$). For example, a usual setting is one where a 5% two-sided test is used, based on an asymptotic normal approximation which implies that $\alpha = 0.05$ and $c_\alpha \sim 1.96$.

4.1 Finding the optimal size α

Define

$$\mathcal{P}_{T,\alpha}(\boldsymbol{b}) = \Pr(S_T > c_\alpha|\boldsymbol{b}) \tag{2}$$

The decision maker wishes to test the null hypothesis $\boldsymbol{b} = \mathbf{0}$ and minimise expected losses, under various possible values for $\boldsymbol{b}$, under the alternative, with respect to α. Usual econometric practice would fix this to say 0.05 or 0.1 but clearly this might not be optimal in the above setting. Expected loss conditional on $\boldsymbol{b}$, $E(L_\alpha|\boldsymbol{b})$, is given by

$$\begin{aligned} E(L_\alpha|\boldsymbol{b}) &= \mathcal{P}_{T,\alpha}(\mathbf{0})L_{yg}(\mathbf{0}) + (1-\mathcal{P}_{T,\alpha}(\mathbf{0}))L_{ng}(\mathbf{0}) + \mathcal{P}_{T,\alpha}(\boldsymbol{b})L_{yb}(\boldsymbol{b}) + (1-\mathcal{P}_{T,\alpha}(\boldsymbol{b}))L_{nb}(\boldsymbol{b}) \\ &= \mathcal{P}_{T,\alpha}(\mathbf{0})(L_{yg}(\mathbf{0}) - L_{ng}(\mathbf{0})) + L_{ng}(\mathbf{0}) + \mathcal{P}_{T,\alpha}(\boldsymbol{b})(L_{yb}(\boldsymbol{b}) - L_{nb}(\boldsymbol{b})) + L_{nb}(\boldsymbol{b}). \end{aligned}$$

Therefore, unconditional losses are given by

$$E\left(L_\alpha\right) = \mathcal{P}_{T,\alpha}(\mathbf{0})\left(L_{yg}(\mathbf{0}) - L_{ng}(\mathbf{0})\right) + L_{ng}(\mathbf{0}) + \int_{\boldsymbol{b}\in\boldsymbol{B}} \left(\mathcal{P}_{T,\alpha}(\boldsymbol{b})\left(L_{yb}(\boldsymbol{b}) - L_{nb}(\boldsymbol{b})\right) + L_{nb}(\boldsymbol{b})\right) dF(\boldsymbol{b}) \tag{3}$$

where $F(\boldsymbol{b})$ denotes some weighting scheme for various deviations from the null hypothesis and $\boldsymbol{B}$ denotes the set in which $\boldsymbol{b}$ takes values. In the following empirical analysis, we will approximate $F(\boldsymbol{b})$ with a normal distribution (see Section 5.3) such that large breaks occur with smaller probabilities. However, for simplicity and expositional purposes we now assume equal weights ($F(\boldsymbol{b}) = 1$) giving

$$E\left(L_\alpha\right) = \mathcal{P}_{T,\alpha}(\mathbf{0})\left(L_{yg}(\mathbf{0}) - L_{ng}(\mathbf{0})\right) + L_{ng}(\mathbf{0}) + \int_{\boldsymbol{b}\in\boldsymbol{B}} \left(\mathcal{P}_{T,\alpha}(\boldsymbol{b})\left(L_{yb}(\boldsymbol{b}) - L_{nb}(\boldsymbol{b})\right) + L_{nb}(\boldsymbol{b})\right) d\boldsymbol{b} \tag{4}$$

The problem amounts to minimising $E\left(L_\alpha\right)$ with respect to α or

$$\hat{\alpha} = \arg\min_{\alpha} E\left(L_\alpha\right) \tag{5}$$

Denoting $\mathcal{P}'_{T,\alpha}$ as the first derivative with respect to α, the first order condition is then

$$\mathcal{P}'_{T,\alpha}(\mathbf{0})\left(L_{ng}(\mathbf{0}) - L_{yg}(\mathbf{0})\right) = \int_{\boldsymbol{b}\in\boldsymbol{B}} \left(\mathcal{P}'_{T,\alpha}(\boldsymbol{b})\left(L_{yb}(\boldsymbol{b}) - L_{nb}(\boldsymbol{b})\right)\right) d\boldsymbol{b} \tag{6}$$

If S_T is well behaved under $\boldsymbol{b} = 0$, $\mathcal{P}_{T,\alpha}(\mathbf{0}) = \alpha$ and so (6) becomes

$$(L_{ng}(\mathbf{0}) - L_{yg}(\mathbf{0})) = \int_{\boldsymbol{b}\in\boldsymbol{B}} (\mathcal{P}'_{T,\alpha}(\boldsymbol{b})(L_{yb}(\boldsymbol{b}) - L_{nb}(\boldsymbol{b})))d\boldsymbol{b} \tag{7}$$

Of course, in general, $\mathcal{P}_{T,\alpha}(\mathbf{0})$ and $\mathcal{P}_{T,\alpha}(\boldsymbol{b})$ and therefore $\mathcal{P}'_{T,\alpha}(\mathbf{0})$ and $\mathcal{P}'_{T,\alpha}(\boldsymbol{b})$ are not known but are nonstochastic quantities that can be evaluated either analytically or, if that is too cumbersome as is likely the case for most realistic settings, by simulation. Noting that $\lim_{T\to\infty} \mathcal{P}_{T,\alpha}(\boldsymbol{b}) = 1$, which implies that $\lim_{T\to\infty} \mathcal{P}'_{T,\alpha}(\boldsymbol{b}) = 0$, it is easily seen that asymptotically the problem has a trivial solution given by $\alpha = 0$. In such cases, large breaks are uninteresting because most break tests will be able to detect them. However, in practice the power is less than unity, implying that breaks are small. We formalize this idea by using the local-to-zero argument. Specifically, one can define a sequence of local bad states indexed by $\boldsymbol{b}_T = \boldsymbol{b}/\sqrt{T}$, in which case $\lim_{T\to\infty} \mathcal{P}_{T,\alpha}(\boldsymbol{b}_T) = \mathcal{P}_\alpha(\boldsymbol{b}) \neq 1$, and then the problem has a non-trivial solution even asymptotically.

This procedure is a form of cross-validation whereby an objective function is optimised with respect to a tuning parameter. It is important, therefore, to note that the idea can be extended to allow for further tuning parameters. For example, the test statistic S_T can be modified to allow for a data dependent data window that only considers recent data. This might be useful in the context of forecasting. Then the optimisation would be undertaken with respect to both α and the window, or more generally, bandwidth.

4.2 An analytical example

A very simple derivation that illustrates the above considerations can be constructed as follows. Let the data generating process be:

$$y_i \sim iidN(\frac{\mu}{\sqrt{T}}, 1) \tag{8}$$

The form is chosen to ensure that the power of the test does not converge to 1 asymptotically, as discussed in the previous subsection. We wish to test $E(y_t) = 0$. It follows that

$$S_T = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_t \sim N(\mu, 1) \tag{9}$$

In the context of the previous general discussion, $\boldsymbol{b} = \mu$. Therefore

$$\begin{aligned} \mathcal{P}_{T,\alpha}(\mu) &= \Pr\left(S_T > c_\alpha | \mu\right) \\ c_\alpha &= \Phi^{-1}(1-\alpha) \end{aligned}$$

where $\Phi^{-1}(\cdot)$ is the inverse function of the cumulative standard normal distribution. Then,

$$\begin{aligned}\mathcal{P}_{T,\alpha}(\mu) &= \Pr(S_T > c_\alpha|\mu) \\ &= \Pr(S_T - \mu > c_\alpha - \mu|\mu) \\ &= 1 - \Phi(c_\alpha - \mu) \\ &= 1 - \Phi(\Phi^{-1}(1-\alpha) - \mu)\end{aligned}$$

Of course

$$\begin{aligned}\mathcal{P}_{T,\alpha}(\mathbf{0}) &= \Pr(S_T > c_\alpha|\mathbf{0}) \\ &= 1 - \Phi(c_\alpha) \\ &= 1 - \Phi(\Phi^{-1}(1-\alpha)) \\ &= \alpha\end{aligned}$$

This gives closed form expressions that can be plugged in (6) together with user defined loss functions, $L_{yb}(\mu)$ and $L_{nb}(\mu)$. Note that we do not place any restrictions on these loss functions.

5 Our procedure step by step

In this section, we provide a detailed discussion of our new approach to detecting shifts in forecast accuracy.

5.1 The loss function

Recall that the purpose of this paper is to design a test to detect any changes in the mean forecast errors across two samples subject to the set of challenges discussed earlier. As set out in Section 3, we are interested in not only detecting breaks in a univariate but also in a multivariate setting. The set-up of the two-state, two-action decision problem addresses our need to balance the cost of making mistakes for policymakers, i.e. working under the assumption there is a break when indeed there is none (*Type I error*) or missing out a break that is actually driving a change in the forecast error bias (*Type II error*).

Losses would, in principle, depend on the policymaker's preference ordering. Since it is hard to elicit those preferences, we adopt a simple statistical criterion. We assume that the loss incurred in not capturing a break would be the same as that incurred if a perceived (but unrealized) break caused the analyst to update the forecasting model unduly, in other words,$L_{yg}(0) = L_{nb}(b) = L(b) > 0$. If, instead, the break was correctly captured, i.e. either no correction was applied and there was no break or a correction was applied and there was a break, then the loss would be zero ($L_{ng}(0) = L_{yb}(b) = 0$). This choice is made for communication purposes. For instance, it is much easier to explain that an x percentage points bias in nominal wages has caused an average inflation

forecast error of y percentage points than talking about mean squared forecast errors. In this case, if there is no wage bias then there should be no inflation bias. Furthermore, focussing on the mean of forecast errors allows us to identify those periods that contribute most to the 'break' and help us to identify not only the 'source' but also to assess the plausibility of the statistical inference.

We now lay down the generic notation for our application of the theoretical discussion in Section 4. We denote z as the headline (or reference) variable the forecast accuracy of which policy makers are interested in studying, and x as an 'instrumental variable', where policy makers are not directly interested in its forecast accuracy but in how its display in forecast breakdowns affect z. For example, we think of x as one of the series for import prices which is relevant inasmuch as it helps produce an accurate inflation forecast — and z as one of the reference variables — for example, inflation. Of course, in the univariate setting, the variables z and x coincide.

Following (4) and (5), we proceed to consider a grid of possible breaks for variable x at forecast horizon h. (In this section, we omit the superscript h in order to simplify our notation. But it is understood that our analysis is specific to the forecast horizon.) For each of them, we compute $L^{z,x}(\cdot)$, i.e. the change in the bias in forecast error for variable z that would result from a break in x. Table 2 presents our two-state, two-decision problem with the simplified loss structure.

Table 2: A two-state, two-decision problem with simplified loss structure

	Bad state (a break)	Good state (no break)
Action taken ($d = 1$)	0	$L^{z,x}(b^x)$
No action taken ($d = 0$)	$L^{z,x}(b^x)$	0

Note: L denotes the loss function. The superscript x refers to the variable with a forecast break (denoted as b^x), whereas z refers to the reference or headline variable(s) of interest to policy makers.

5.1.1 Modelling $L^{z,x}(\cdot)$ using a vector auto-regressive (VAR) model

Recall in equation (1) the loss function is defined as the difference between the *mean* of the forecast error distribution under a break and that under no break. Generically, we model the mean of the our forecast errors jointly as a vector autoregressive process:

$$\overline{\varepsilon}_t = A_0 + B_0 \overline{\varepsilon}_{t-1} + \overline{u}_t \tag{10}$$

where $\overline{\varepsilon}_t$ is a vector including forecast errors for our variables of interest over our base period $t = 1, ..., T_0$, and $\overline{u}_t \sim N(0, \Sigma_0^u)$. Note that we add the subscript '0' to denote the estimation using the base sample.[4]

We include only one lag in our VAR specification owing to sample length considerations. Furthermore, due to small sample issues the estimation of the VAR model is carried out subject to the constraint that the mean of the vector of the forecast implied by the multivariate model (equation

[4]Such notation corresponds to that in Section 4.2 if we let $T_0 = (1-p)T$, $T_1 = (1-p)T + 1$ and $T_2 = T$.

11) coincides with the univariate estimates.

$$\mathcal{M}_0 = (I - B_0)^{-1} A_0 \tag{11}$$

where $\mathcal{M}_0 = [\mu_0^1, ..., \mu_0^Q]'$ denotes the vector of forecast biases of all of the $Q = 13$ variables in the base sample (recall our notation in Section 3.1 that $\mathcal{M}$ denotes the bias of a group of variables whereas μ denotes the bias of a variable). In the following, we use the subscript Y to indicate the variables of interest to policy makers.

5.1.2 Inversion

Denote X as the set of all variables we are studying. In particular, consider a variable $x \in X$ and a vector $Y = X \backslash x$ which includes all variables but x, and define $D_0 \equiv (I - B_0)^{-1}$. After dropping the subscript '0', the constraint (11) can be partitioned as:

$$\begin{bmatrix} \mu^x \\ \mathcal{M}^Y \end{bmatrix} = \begin{bmatrix} D_{xx} & D_{xY} \\ D_{Yx} & D_{YY} \end{bmatrix} \begin{bmatrix} A_x \\ A_Y \end{bmatrix}$$

And we can express:

$$A_x = D_{xx}^{-1} \left(\mu^x - D_{xY} A_Y \right) \tag{12}$$

$$\mathcal{M}^Y = D_{Yx} A_x + D_{YY} A_Y \tag{13}$$

so that we can use (12) to substitute for A_x in (13) to obtain:

$$\mathcal{M}^Y = D_{Yx} \left(D_{xx}^{-1} \left(\mu^x - D_{xY} A_Y \right) \right) + D_{YY} A_Y \tag{14}$$

Equation (14) is key to understand how, under our inversion scheme, a break in a variable spills over to others. In particular, if we define as $d\mathcal{M}^z$ the *change in forecast bias* for some generic variables $z \in Y$, it can be seen that:

$$\begin{aligned} d\mathcal{M}^z &= e_z' d\mathcal{M}^Y \\ &= e_z' D_{Yx} D_{xx}^{-1} d\mu^x \\ &= e_z' D_{Yx} D_{xx}^{-1} b^x \end{aligned}$$

where e_z is a vector that selects z out of Y, and $d\mu_x = b^x$, the change in the forecast error bias for x from the base sample to be used in our simulations.[5] This linear mapping gives us an idea about how much an accurate forecast for variable x matters for that of variable z by capturing

[5] When carrying out hypothesis testing, we compute the test statistic using the break found between the evaluation and base samples. In other words, $d\mu_x = b^x \equiv \mu_1^x - \mu_0^x$, where μ_0^x and μ_1^x respectively denote the bias in the base and evaluation samples.

how forecast errors for a certain variable x co-vary with forecast errors for z. This idea parallels that in our DSGE example in Section 3. In that case we use the model's controlled environment to study the effects of a bias in the forecast for one variable onto another. Here we follow the same principle but let the sample covariance of forecast errors determine how a change in forecast bias in the forecast for one variable will impact another. The underlying idea, however, is the same: inaccurate forecasts for one variable can cause deteriorations in the forecast performance for other variables. If variable z is the variable of interest and variable x is instrumental to the forecast of z, a small value of $e_z' D_{Yx} D_{xx}^{-1}$ suggests that a change in forecast error bias for x will not have a dramatic effect on that of the variable of interest.

Following (1), we define the loss function L for our univariate and multivariate tests as follows:

$$L^{z,x}\left(b^x\right) \equiv \begin{cases} \|\mathcal{M}(\mu_0^x + b^x) - \mathcal{M}(\mu_0^x)\| & \text{when } z = x \\ \|\mathcal{M}^z(\mu_0^x + b^x) - \mathcal{M}^z(\mu_0^x)\| = \left\|e_z' D_{Yx} D_{xx}^{-1} b^x\right\| & \text{when } z \neq x \end{cases} \tag{15}$$

The outstanding question is then whether this effect is significant or not. We rely on information from the sample and conduct simulations to characterize the distribution under the null hypothesis of no breaks. Moreover, in the case where $z \neq x$, while at the moment we focus on one variable at a time, it is straightforward to apply our proposed approach to linear combinations of bias changes (by replacing e_z' with any arbitrary vector in equation (15)) or other functions or moments of the data.

5.2 Finding the optimal size

We next compute the optimal size and average optimal sizes over the grid of breaks we consider. We first express the expected loss conditional on a break as:

$$E(L_\alpha|b^x) = \begin{cases} \alpha L^{z,x}\left(b^x\right) + \left(1 - \mathcal{P}_{T,\alpha}\left(b^x\right)\right) L^{z,x}\left(b^x\right) & \text{when } z = x \\ \alpha L^{z,x}\left(b^x\right) + \left(1 - \mathcal{P}_{T,\alpha}\left(b^x, L^{z,x}\left(b^x\right)\right)\right) L^{z,x}\left(b^x\right) & \text{when } z \neq x \end{cases} \tag{16}$$

where b^x is the size of the break in x and $\mathcal{P}_{T,\alpha}(\cdot)$ is the test power as defined in Section 4. In the case where $z \neq x$, the power function $\mathcal{P}_{T,\alpha}(\cdot)$ depends on the *dynamic covariance* between variables in addition to the magnitude of the break.

We then compute the optimal size $\alpha^{*z,x}$ as:

$$\alpha^{*z,x} = \sum_{b^x \in \boldsymbol{B}} \pi\left(b^x\right) \left\{\arg\min_\alpha E(L_\alpha|b^x)\right\} \tag{17}$$

where b^x, the size of the break in x, is collected in the set $\boldsymbol{B}$, and $\pi(\cdot)$ is its associated probability. Details are discussed in Section 5.3.

The minimization problem has a very intuitive interpretation under simple regularity conditions on the power function.[6] For a given break magnitude, one would want to increase the nominal size to the point such that further decrease in the probability of a *Type II* error will not be able to offset the increase in *Type I* error.[7]

5.3 Estimating reference breaks in the data

For a given series x under the *base sample*, forecast errors at a certain forecast horizon are denoted as ε_t^x. As mentioned above, these are considered primitives in our analysis and we model them as a simple time-series autoregressive process:

$$\varepsilon_t^x = \rho_w(0) + \sum_{k=1}^{P} \rho_w(k)\varepsilon_{t-k}^x + u_t$$

where the number of lags P is selected optimally using the Bayesian Information Criterion, and the intercept $\rho_w(0)$ and the autoregressive coefficients $\rho_w(k)$ are estimated on a 24-quarter rolling window (hence the subscript w on the autoregressive polynomial) for $w = 1, ..., W$. We then define ν_w^x, the rolling-window bias for series x:

$$\nu_w^x = \frac{\rho_w(0)}{1 - \sum_{k=1}^{p} \rho_w(k)}$$

We then compute the reference break $\tilde{b}^x$ for variable x as the difference (in absolute values) between the largest of the biases estimated within the windows and the bias estimated over the entire base sample (denoted as μ^x). In other words,

$$\tilde{b}^x \equiv max\left\{|\nu_w^x|\right\}_{w=1}^{W} - |\mu^x|$$

We approximate the distribution of breaks in the base sample as follows. We first define a grid of non-zero breaks $\boldsymbol{B} = \{.25\tilde{b}^x,\ .5\tilde{b}^x,\ \tilde{b}^x,\ 1.5\tilde{b}^x,\ 2\tilde{b}^x\}$ based on the reference breaks we computed before. Then we compute the associated probabilities as $\pi(b^x) = \phi(\frac{b^x}{\sigma^x})$ where $b^x \in \boldsymbol{B}$, σ^x is the standard deviation of $|\nu_w^x|_{w=1}^W$ and ϕ is the probability density function (pdf) for Normality.

[6] We use the grid method to solve the minimisation problem in order to ease the computational burden in our already computationally intensive exercise.

[7] To clarify, our 'loss function' is different from that used by Giacomini and Rossi (2009). In our case the end-goal of having a loss function is to pin down the optimal size for the test at hand, while in their setup the loss is a function of parameter estimates (so ultimately of the sample) which measures the quality of the forecast. In our case, we focus on the forecast-error bias as our measure of forecast quality so the *surprise loss* (as defined by Giacomini and Rossi (2009), p. 672) is simply the *standardised* difference in the forecast-error bias in the base and evaluation samples, and we do not refer to this quantity as *loss* to avoid possible confusion.

5.4 The algorithm

Our aim is to test if the contribution of one variable to another one is significant. We proceed in steps as detailed in the Algorithm below.

Algorithm 1 *Our procedure in detecting shifts in forecast accuracy.*

1. *Given estimates for* $(\hat{A}_0, \hat{B}_0, \hat{\Sigma}_0^u)$ *from equation 10 we simulate N samples by drawing random disturbances via a wild bootstrap.*
2. *We simulate the null distribution by computing* $\mathcal{M}^n(b^x = 0), n = 1, ..., N$*, under the null hypothesis of no break in* x*.*
3. *We simulate N samples using this same model but under the alternative of breaks* b^x *given by the following grid of non-zero breaks for each variable* x*:* $b^x \in \boldsymbol{B} = \{.25\tilde{b}^x,\ .5\tilde{b}^x,\ \tilde{b}^x,\ 1.5\tilde{b}^x,\ 2\tilde{b}^x\}$*, where* $\tilde{b}^x$ *is the reference break estimated in the base sample in Section 5.3.*
4. *We then compute* $\mathcal{M}^n(b^x), n = 1, ..., N$*, and hence the loss function* $L^{z,x}(b^x)$ *as specified in* (15)*.*
5. *We also compute the power function* $\mathcal{P}_{T,\alpha}(b^x, L^{z,x}(b^x))$*, using the size-adjusted critical values simulated from step 2.*
6. *And finally we compute the optimal size* $\alpha^{*z,x}$ *following 17.*

So, for any variable pair (z, x), we have the optimal size ($\alpha^{*z,x}$, averaged across break sizes) that indicates when we should reject the null hypothesis that a break in the forecast error for variable x has no statistically significant effect on the bias in the forecast for variable z at the same time horizon.

6 Monte Carlo Simulations

In order to demonstrate the usefulness of our new approach to detecting shifts in forecast accuracy, we conduct Monte Carlo simulations on the simple analytical example set out in Section 4.2. In this section, we consider two possible scenarios: when the actual break size μ is (i) known and (ii) unknown to the researcher. Our results show that, under most circumstances, our proposed algorithm 1 does better than assuming a fixed level of significance.

6.1 When Type II loss is known to the researcher

We assume $L_{nb}(\mu) = \mu > 0$, $L_{yb}(\mu) = 0$, $L_{ng} = 0$ and $L_{yg} = c > 0$. In other words, Type II loss is known and is equal to the actual size of break μ. The loss matrix is summarised in Table 3.

In this situation equation 7 can be simplified as

$$\mathcal{P}'_{T,\alpha}(\mu) = \frac{c}{\mu} \tag{18}$$

Table 3: Monte Carlo simulation: loss matrix

	Bad state $\mu > 0$	Good state $\mu = 0$
Null hypothesis is rejected	0	$L_{yg}(0) = c$
Null hypothesis is not rejected	$L_{nb}(\mu) = \mu$	0

Note: L denotes the loss function.

The optimal level of α which is a function of c and μ, denoted as $\alpha(c, \mu)$, can be solved analytically or numerically.

Algorithm 2 describes in detail the Monte Carlo simulations we conduct.

Algorithm 2 *Monte Carlo Simulations when Type II loss is known*

1. *Define the grids* $\mu \in [0.5 : 0.01 : 5]$, $c \in [0.5 : 0.01 : 5]$, *with the step size of 0.01 each. Solve for the optimal* $\alpha(c, \mu)$ *according to equation 18 for each pair of* (c, μ).
2. *Simulate data with the true data-generating process with positive breaks and conduct hypothesis testing*
   (a) *Given* (c, μ) *and assume* $T = 100$, *simulate data according to 8.*[8]
   (b) *Construct the test statistic* S_T *following (9).*
   (c) *Conduct hypothesis testing for* $H_0 : \mu = 0$ *using the textbook procedure of setting* $\alpha = 0.05$. *If the hypothesis is not rejected, the loss is recorded as* μ*; otherwise, the loss is recorded as* 0.
   (d) *Repeat the above three steps for 5000 times and compute the average loss given* (c, μ). *This corresponds to Type II loss and denote it as* $\boldsymbol{L}^{c,\mu}_{TypeII,\alpha=0.05}$.
3. *Repeat step 2 with simulated data with the true data-generating process with zero breaks, i.e.* $\mu = 0$. *Note that when the null hypothesis is not rejected, the loss is recorded as* 0*; otherwise,* c. *The average loss computed corresponds to Type I loss and denote it as* $\boldsymbol{L}^{c,\mu}_{TypeI,\alpha=0.05}$.
4. *Compute the average loss* $\boldsymbol{L}^{c,\mu}_{\alpha=0.05} = \boldsymbol{L}^{c,\mu}_{TypeI,\alpha=0.05} + \boldsymbol{L}^{c,\mu}_{TypeII,\alpha=0.05}$.
5. *Repeat steps (2), (3) and (4) with the optimal critical values* $\alpha(c, \mu)$ *derived in step (1). Compute* $\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$.

Figure 2 displays the optimal alpha $\alpha(c, \mu)$ in step 1. We observe that (i) as c increases the optimal alpha decreases monotonically. This is intuitive because a smaller $\alpha(c, \mu)$ will be chosen when the cost of Type I error increases, holding all other factors constant; (ii) Given c, as μ increases, the optimal $\alpha(c, \mu)$ decreases monotonically only after a certain break size. Since the power increases with μ, a trade-off between Type I and Type II errors exists such that a larger $\alpha(c, \mu)$ is optimal until the break size μ reaches beyond a certain threshold.

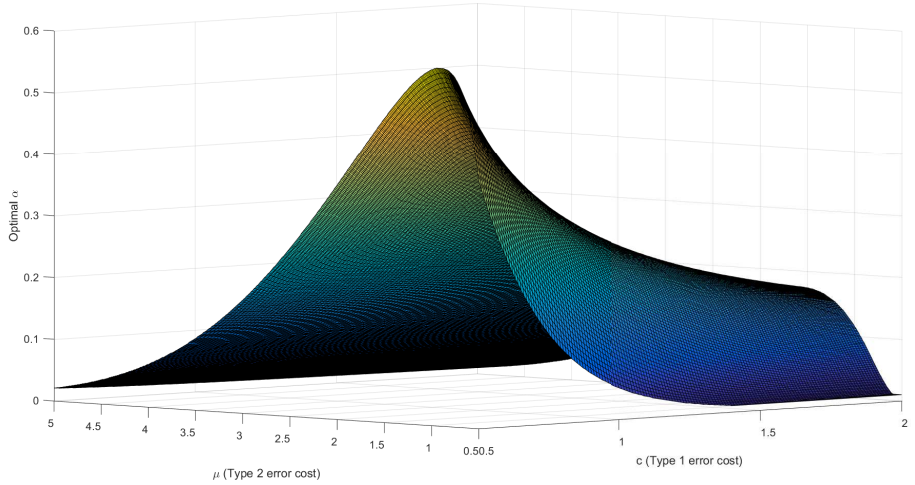
Figure 2 compares the average loss when a researcher uses the optimal chosen $\alpha(c, \mu)$ ($\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$) and the usual textbook procedure $\alpha = 0.05$ ($\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$) in our simulations. Results of selected pairs

[8] Owing to our calibration in our analytical example in Section 4.2, T does not matter in our simulations. We present supporting evidence in the Appendix.

of (c, μ) are reported in Table 4. It is found that the average loss when we use the textbook procedure is uniformly bigger than the loss using the optimal picked $\alpha(c, \mu)$. We also observe that the difference in average loss between the two tests decreases as μ increases. This is because $\alpha(c, \mu)$ tends to be a very small number (close to the textbook size 0.05) when the break is large in size and it is generally easy for either test to detect the break. In addition, the difference in average loss between the two tests decreases as c, the Type I error cost, becomes larger.

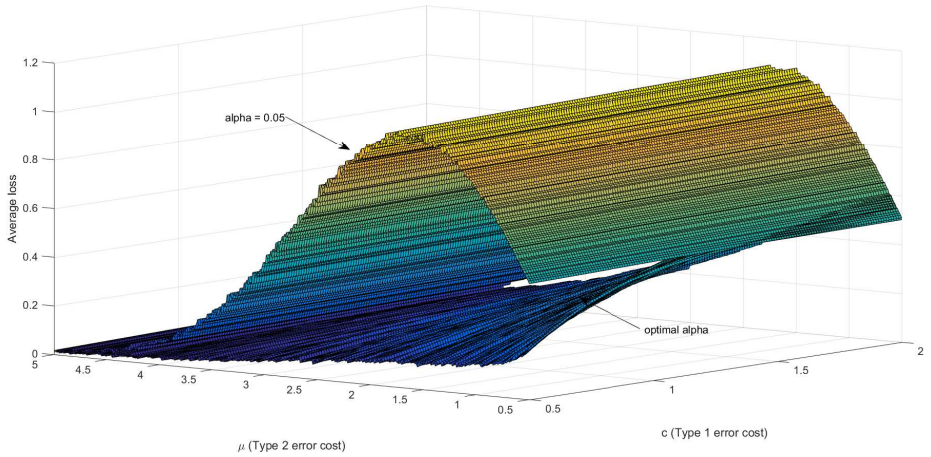
In short, when a researcher knows Type II error loss our proposed algorithm works uniformly better .

Figure 2: Monte Carlo simulations when Type II loss is known: optimal chosen $\alpha(c, \mu)$



Note: $\alpha(c, \mu)$ is computed using step 1 described in Algorithm 2. Refer to main text for details.

Figure 3: Average loss for Monte Carlo simulations when Type II loss is known, T=100



Note: Average losses with simulations using the optimally chosen sizes under known Type II loss ($\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$) and using the usual textbook procedure of $\alpha = 0.05$ ($\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$). Computation follows the Algorithm 2. Refer to main text for details.

6.2 When Type II loss is unknown to the researcher

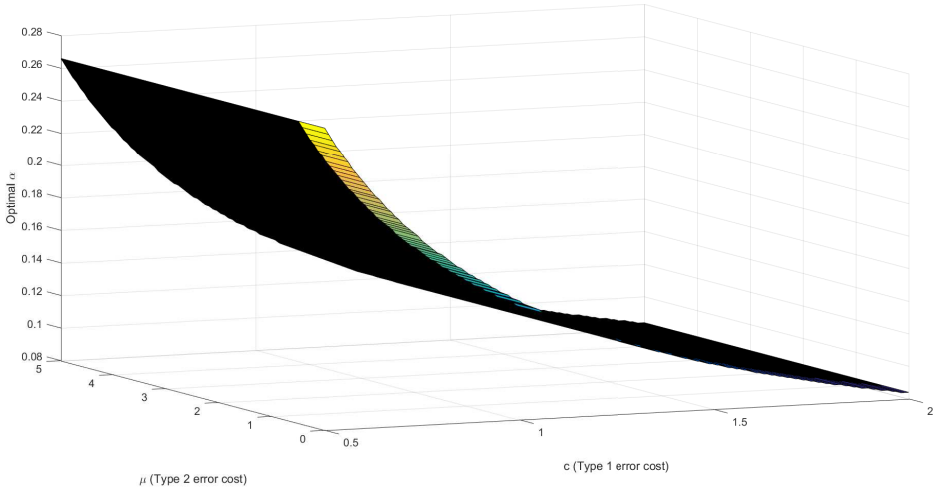
More realistically, a researcher does not know the loss when the hypothesis fails to reject the wrong null hypothesis, implying that $L_{nb}(\mu)$ is unknown. To perform our simulations we again

follow equation 7 to solve for $\alpha(c, \mu)$. We note that $\boldsymbol{b} = \mu$ and $\boldsymbol{B} = [0.5 : 0.01 : 5]$, and assume that μ follows a uniform distribution for simplicity. The modified simulations are described in Algorithm 3.

Algorithm 3 *Monte Carlo Simulations when Type II loss is unknown*

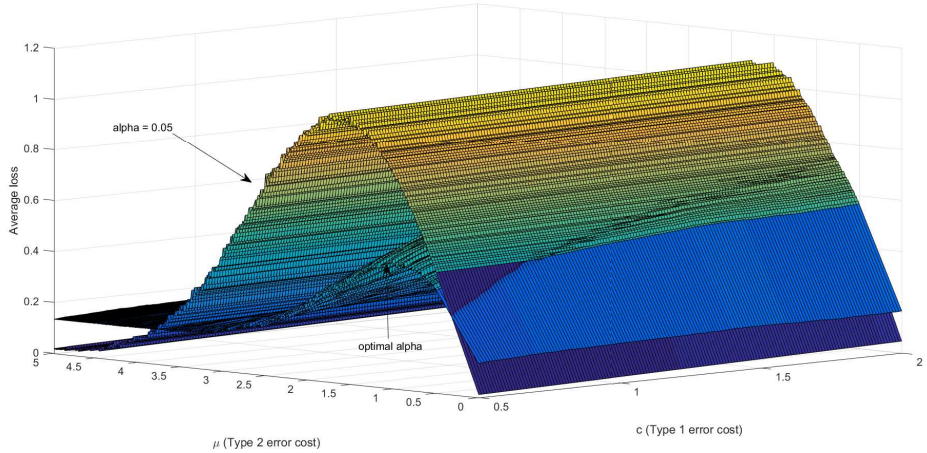
1. *As before, define grids* $\mu \in [0.5 : 0.01 : 5]$, $c \in [0.5 : 0.01 : 5]$, *with the step size of 0.01 each. Solve for the optimal* $\alpha(c, \mu)$ *according to equation 7 for each pair of* (c, μ).
2. *Repeat steps (2), (3), (4), (5) as stated in Algorithm (2).*

Figure 4: Monte Carlo simulations when Type II loss is unknown: optimal chosen $\alpha(c, \mu)$



Note: $\alpha(c, \mu)$ is computed using step 1 described in Algorithm 3. Refer to main text for details.

Figure 5: Average loss for Monte Carlo simulations when Type II loss is unknown, T=100



Note: Average loss is computed following the Algorithm 3. Refer to main text for details.

Figure 4 displays the optimally chosen sizes, which differ significantly from Figure 2 in that they no longer vary with the true μ as this is unobservable to the researcher. Otherwise, we still see that as c increases the optimal alpha decreases monotonically.

Average losses are presented in Figure 5. We find that the average loss under the optimally picked sizes are lower for most values of μ, with the exception of extremely small or large values. Table

4 compares the average losses under the situation with known and unknown Type II loss. The average loss under unknown Type II loss is larger relative to the loss under known Type II loss. Moreover, when $\mu = 0.5$ (very small break) or $\mu = 4, 5$ (very large break), losses with unknown Type II loss are larger than the loss using $\alpha = 0.05$. The intuition is as follows:

- when the actual break μ is very small (the power of the test tends to be small) but unknown, the researcher has to integrate out all the possible break sizes in $\boldsymbol{B}$. It implies that the optimally chosen $\alpha(c, \mu)$ is larger compared to the situation where μ is known, leading to over-rejection of the null hypothesis.
- similarly, when the actual break μ is very large, the textbook procedure of using $\alpha = 0.05$ is sufficiently good at detecting breaks because of high power. However, an ignorant researcher has to integrate out all the possible break sizes in $\boldsymbol{B}$, implying that a larger optimal $\alpha(c, \mu)$ is chosen compared to the situation where μ is known. Over-rejection of the null hypothesis is resulted.

In short, our proposed algorithm still works better when Type II loss is unknown, with the exception of extreme values of μ.

Table 4: Comparing the average losses when Type II loss is known and unknown to the researcher, T=100, for selected pairs of (c, μ)

		μ					
	c	0.5	1	2	3	4	5
$\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ (known Type II loss)	0.5	0.20	0.13	0.09	0.06	0.03	0.01
$\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ (unknown Type II loss)	0.5	0.40	0.48	0.32	0.16	0.13	0.13
$\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$	0.5	0.48	0.85	0.97	0.46	0.10	0.02
$\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ (known Type II loss)	1	0.44	0.33	0.19	0.11	0.05	0.02
$\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ (unknown Type II loss)	1	0.50	0.66	0.49	0.23	0.16	0.15
$\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$	1	0.49	0.86	0.98	0.47	0.11	0.03
$\boldsymbol{L}^{T,\mu,c}_{\alpha=\alpha(c,\mu)}$ (known Type II loss)	1.5	0.50	0.48	0.27	0.14	0.07	0.03
$\boldsymbol{L}^{T,\mu,c}_{\alpha=\alpha(c,\mu)}$ (unknown Type II loss)	1.5	0.55	0.76	0.61	0.28	0.17	0.17
$\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$	1.5	0.50	0.87	0.99	0.48	0.12	0.04
$\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ (known Type II loss)	2	0.50	0.60	0.34	0.17	0.08	0.04
$\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ (unknown Type II loss)	2	0.57	0.81	0.70	0.33	0.18	0.17
$\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$	2	0.51	0.88	1.00	0.49	0.13	0.05

Note: Average loss $\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ when Type II loss is known is computed using Algorithm (2), whereas average loss $\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ when Type II loss is unknown is computed using Algorithm (3). $\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$ is the average loss when the textbook procedure of $\alpha = 0.05$ is used. Refer to the main text for details.

7 Empirical Application

In this section we present the results of our exercise. We start off with our univariate analysis by presenting the optimal size of the test that our procedures results in. Then, given the optimal test size and the corresponding simulation-based critical values, we test whether Bank of England forecasts for each of the variables under consideration and at various key forecast horizons improved or deteriorated in the aftermath of the Great Recession. We will also discuss the implication for

the optimal test size under our multivariate analysis.

7.1 Data

Our analysis considers 13 quarterly macroeconomic time series, including the headline variables such as real GDP growth ('GDP'), CPI inflation ('CPI') and the unemployment rate ('Urate'). The other ten variables are nominal wages ('Wages'), hours, real investment ('RealInv'), real government spending ('RealG'), real exports ('RealX'), real imports ('RealIm'), real consumption ('RealC'), nominal house prices ('Houseprices'), as well as real GDP growth for the Euro Area ('GDP-EA') and the US ('GDP-US'). All series except the unemployment rate are expressed in annual growth, in line with Bank of England (2015). Our sample covers the period from 2000Q4 to 2016Q4 from the Bank of England database.[9]

We will focus on forecast errors at horizons of 1, 4, 8 and 12 quarters. Our implementation follows the fixed-scheme approach outlined by Giacomini and Rossi (2009) in that we fix the dates for our base and evaluation samples, primarily on considerations of sample sizes and on the occurrence of the Great Recession in the middle of our data sample. We fix the break date at 2011Q4 so that we have exactly 20 observations in the evaluation sample.

7.2 Univariate analysis

7.2.1 Optimal test size

Table 5 reports the test sizes we selected optimally following Algorithm 1, when we consider each variable in isolation (i.e. when $z = x$). Test sizes are differentiated across both variables and forecast horizons and tailored to the observed sample characteristics of each series. While there is no particular pattern across horizons or depending on whether 2008 and 2009 are included in the analysis, all the significance levels we end up selecting are larger than the customary 5 or 10 percent values.[10] Indeed only a few are below the 30 percent mark.[11] This suggests that given our setup and sample the power of the test is rather low at the standard 5 percent significance level. By selecting higher nominal size values our procedure increases the test's power. As a result, our procedure will be much more sensitive even to relatively small and recently occurred breaks. This is optimal, however, given the costs that such breaks incur.

[9] The start date is constrained by the earliest date for which all of the series are available for the vector autoregressive model in Section 5.1.1.

[10] Excluding the crisis period reduces the number of observation to the point of making it impractical to study 3-year ahead forecasts.

[11] Some optimal test sizes reach 50 percent. There are two possible explanations. Firstly, the fact that we use a grid in our minimisation problem, as explained in Section 5, to ease the computational burden comes with a trade-off. We anticipate that the use of a finer grid can help. Secondly, it reflects that, for some variables, the test has very low power for the alternative hypotheses considered and therefore has no ability to distinguish the null from the alternative. It implies that the choice is essentially a binary one between the two hypotheses and, with very high probability, the lower cost hypothesis is selected. The low power can be a result of our adoption of data-based rules. Recall in Section 5.3 we estimate, in the base sample, reference breaks $\tilde{b}^x$ and their associated probabilities as a function of the magnitude of the reference breaks standardised by the standard deviation of rolling window bias σ^x. If the likelihood of observing a 'reasonably large' break in the data is low, our algorithm will have low power to distinguish the null from the alternative hypotheses.

Table 5: Optimally selected test *size* (in percent) in our univariate analysis

	2000Q4-2011Q4				2000Q4-2011Q4 excl 2008-09		
	Forecast horizon (h)				Forecast horizon (h)		
	1	4	8	12	1	4	8
GDP	37.2	36.1	37.7	44.7	42.3	39.1	45.8
CPI	42.1	46.6	46.5	50.0	46.44	37.7	50.0
Urate	46.9	36.7	42.9	40.0	40.0	42.2	35.6
Wages	39.9	39.6	46.1	40.0	48.7	39.6	43.6
Hours	42.1	49.2	38.1	49.0	38.5	32.0	40.0
RealInv	39.8	45.3	38.8	48.6	49.9	44.3	44.2
RealG	35.8	46.2	39.9	48.0	37.2	47.0	49.9
RealX	42.6	37.0	39.5	36.9	48.9	46.0	50.0
RealIm	37.5	44.2	47.1	47.0	40.1	38.7	46.0
RealC	40.2	40.6	46.7	49.9	44.5	39.8	48.3
Houseprices	46.5	42.4	45.0	43.1	40.0	46.1	38.5
GDP-EA	44.3	44.9	45.1	44.8	46.4	43.8	40.0
GDP-US	42.7	27.6	46.2.0	40.0	39.8	42.0	46.3

Note: The left panel considers the base sample from 2000 to 2012. In the right panel, observations in 2008 and 2009 are excluded from the base sample to guard against the effects of the Great Recession years. In so doing, the sample becomes short to the point of making the evaluation of the 12-quarter ahead forecast impractical. The optimal size is computed based on the univariate analysis outlined in Section 5 and is summarised in Algorithm 1.

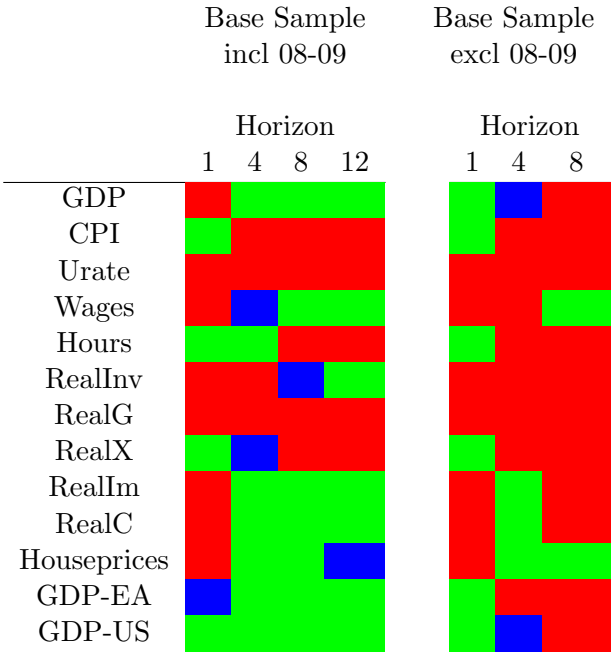
7.2.2 Application to Bank of England Forecast Errors

Having worked out the optimal test size, we turn to study whether forecasts for our main variables of interest have displayed significant improvements or deteriorations over our sample. Table 6 summarizes the main results for our univariate analysis.[12] It reports the thirteen variables under consideration in the rows and the forecast horizons in the columns. The left-hand panel compares the 2012-2016 period to the 2000-2011 period while the right-hand panel reports result that exclude the Great Recession period (2008-2009) from our base sample. We use a simple colour-coding scheme to report our procedure's results. A blue cell corresponds to the situation in which, based on the test size selected with our loss-function scheme and the simulated critical values reported in Table 5, we cannot reject the null hypothesis of no change in forecast performance. In other words, it corresponds to a situation in which the forecast error biases in the evaluation period are not statistically different from those observed in the evaluation period. The other cells correspond to situations in which our procedures flags a statistically significant change in the forecasting for a certain variable/horizon pair. When a significant change is observed we then check if the forecast performance has deteriorated (which we mark with a red cell) or improved (which we represent with a green cell).

Starting with the left-hand panel, we observe a predominance of green at horizons of one year and beyond. That suggests that forecasts have generally become more accurate over the last 5 years. That is in part, however, due to the inclusion of the Great Recession period in our baseline

[12]Table B.1 in the Appendix documents the forecast bias of GDP and CPI (both of the series are available in the public domain) in the base and evaluation samples.

Table 6: Hypothesis testing: univariate analysis applied to Bank of England forecast errors



Note: The base sample is from 2000Q4 to 2011Q4. In the left panel observations between 2008 and 2009 are included whereas in the right panel these observations are excluded. The evaluation sample is from 2012Q1-2016Q4. The null hypothesis is that there is no change in forecast bias across the base and evaluation samples. A blue cell corresponds to non-rejection of the null hypothesis using the simulated critical values reported in Table 5. A red cell denotes a statistically significant deterioration of forecast performance in the evaluation period relative to the base, and a green cell denotes a statistically significant improvement.

sample. The right-hand panel shows that forecast accuracy has generally deteriorated from the pre-crisis period. This is hardly surprising though. The post Great Recession economy presents more challenges to a forecaster than the Great Moderation world we lived in at the turn of the century. For instance, the exceptionally slow recovery of labour productivity in the aftermath of the recession — the *UK productivity puzzle* — poses a set of new questions which can only be answered with time. And uncertainties surrounding trend growth rates can easily result in biased forecasts for many of the real variables.

Another interesting point can be made observing the left-hand panel of Table 6: the tendency to observe forecast improvements at horizons of 4 quarters and above, does not extend to nowcasts (i.e. $h = 1$). With the notable exception of the nowcast for inflation, biases worsened across a host of variables. And, for the most part, the relative performance of nowcasts is unaffected (in terms of the rejection of the null of no change in the forecast bias as well as in terms of improvement or deterioration) by the inclusion of the crisis period in the baseline sample.

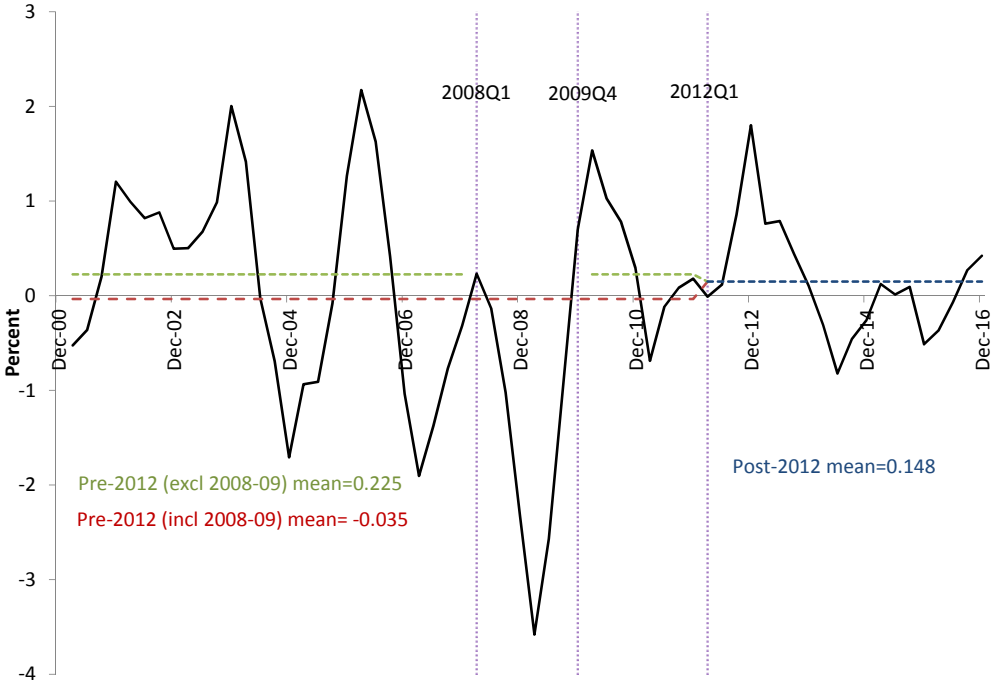
A notable exception is GDP. The GDP nowcast bias displays a significant deterioration when we compare the 2012-2016 period to 2000-2011 but it does display a significant improvement when the crisis period is removed. This is, at first, puzzling since both common wisdom and much of this analysis suggests the Great Recession was not easy to handle for forecasters. Yet, it can be better understood by studying the time series of forecast errors for GDP nowcasts, which we present in Figure 6. Recent errors have displayed a positive bias on average (of 0.148%), while the average error for the full base period was close to zero (-0.035%). However, the latter is the result of a positive pre-crisis bias (0.225%) being offset by a large negative bias during the crisis period. From this perspective, the nowcast errors for GDP over the most recent period are not particularly troublesome at all.

We hope that this illustration shows how this framework can be used in practice. We see a null-hypothesis rejection simply as the first step in what we could consider a *triage process*, i.e. it should be the starting point of some more in-depth analysis.

In this sense, it is also useful to reiterate the benefits of using the first moment of forecast errors for our analysis. The main advantage lies in the possibility of relating the forecast analysis to the underlying economics. For instance, when it comes to unemployment, forecasts have ex-post turned out to over-predict it, which relates to the puzzling weak growth. The problems behind the apparent deterioration in the accuracy of the forecasts for government spending have a more subtle explanation, instead, for which there is no substitute for an analyst's expertise. The Bank of England traditionally takes forecasts for nominal government spending directly from government publications, yet a forecast for *real* government spending requires a projection for the government spending deflator, which is where the problem appears to have emerged.

To sum up, these very different examples show how our test procedures on forecast errors can be used as a trigger for further analyses. We cannot emphasise enough that it is important to investigate forecast breaks *case by case* by considering a range of metrics, including the size of forecast bias in sub-samples, the optimal test size and the hypothesis testing results, rather than basing a decision on a single metric.

Figure 6: Nowcast errors for GDP



Note: This figure plots the nowcast ($h = 1$) errors for GDP and compares the mean in various subsamples. 'Pre-2012' represents the base sample (2000Q4 - 2011Q4) whereas 'Post-2012' represents the evaluation sample (2012Q1 - 2016Q4). The dotted green line denotes the mean for the base period excluding the financial crisis period, defined between 2008Q1 and 2009Q4, whereas the dotted red line includes the period. The dotted purple line denotes the mean for the evaluation sample.

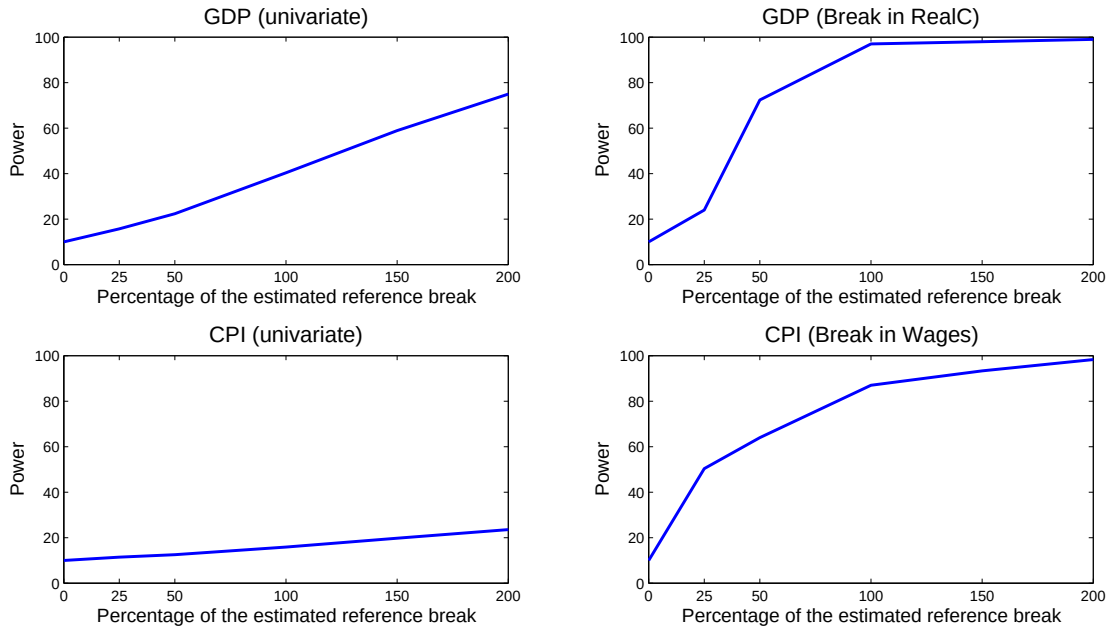
7.3 Multivariate analysis

Recall our discussion in Section 5 that the set-up of univariate and multivariate analyses seeks to answer two different questions. In the multivariate analysis, we consider variables jointly rather than in isolation. In other words, it involves testing not just whether there has been a break in a particular series, but whether that break is material for another variable of interest. This necessarily alters the consideration of the choice of the optimal size of the test as it changes the loss characteristics of each variable. We find that such multivariate analysis leads to a dramatic fall in the optimal test sizes, which is the direct result of the increase in test power.

Figure 7 displays the power function of GDP (the first row) and inflation (the second row) in the univariate analysis (the first column) and in the multivariate case (the second column, as a function of consumption and wages respectively). Figure 7 illustrates that the power increases dramatically when we model cross-variable relationships — more specifically, when we model how a break in forecast accuracy of an instrumental variable x impacts the forecast accuracy of a headline variable z. In other words, it is easier for the test to detect changes in the forecast accuracy of GDP (CPI) conditional on breaks in consumption (wages), when compared to using GDP (CPI) only.

What has caused this improvement? Note that the univariate analysis neglects any information about cross-equation restrictions (Hansen and Sargent, 1980) which can carry economic implications between variables. Since consumption makes up 60% of the GDP, the two variables are unsurprisingly highly correlated. While such correlation is neglected in the univariate case, it is

Figure 7: Simulated Power Functions



Note: This figure displays the power functions for GDP (the first row) and CPI (the second row) in terms of the percentage of the estimated reference breaks based on the base sample (2000Q4-2011Q4). The first column shows the power functions in our univariate analyses, where we consider breaks in forecast performance (changes in forecast biases across the base and evaluation samples) in GDP and CPI *per se*. The second column reports the power functions where we consider forecast breaks in real consumption (the first row) and in nominal wages (the second row). The loss function is described in equation 15.

fully exploited in the multivariate one. The same argument goes for CPI inflation and nominal wages. Table 7 presents the optimally selected size for GDP and CPI *conditional on the forecast break* of each of the variables in the 13-variable VAR model. We observe that the optimal size drops dramatically when comparing them to the univariate analysis in Table 5: many of the sizes are now in the range of 5 and 30 percent. It reflects the substantial increase in the test power in our multivariate setting.

8 Conclusion

Detecting forecast breakdowns is a difficult task. Data spans are limited and past structural change further reduces informative data availability. Further, tests for structural change are well known to have low power and therefore struggle to pick up recent breaks. The low power property seems endemic and not associated with specific tests. Therefore, a different approach is needed to overcome this hurdle. This paper provides one based on decision theory. Given that low power is one of two forms of error associated with testing it is reasonable to consider trading off low power with higher Type I errors. Standard practice which fixes Type I errors and accepts any

Table 7: Optimally selected test size (in percent) for GDP and CPI *conditional on the forecast break of each of the variables* in our multivariate VAR analysis

	GDP			CPI		
	Forecast horizon (h)			Forecast horizon (h)		
	1	4	8	1	4	8
Forecast Performance Break (i.e. Δ forecast bias) in						
GDP	9.70	8.53	11.4	19.4	22.6	11.4
CPI	23.8	9.23	6.96	9.75	10.0	9.58
Urate	9.70	26.6	8.28	29.3	16.8	8.28
Wages	19.9	19.6	41.2	20.0	10.0	7.97
Hours	19.8	50.0	20.0	27.0	20.0	20.0
RealInv	16.5	9.72	7.55	8.29	28.9	33.9
RealG	25.8	9.74	17.1	8.97	9.74	39.9
RealX	28.1	5.00	5.00	18.0	8.59	5.00
RealIm	24.0	38.9	5.00	43.9	18.9	5.00
RealC	17.9	24.6	5.00	9.04	7.82	8.40
Houseprices	20.0	5.00	5.00	18.4	9.47	18.2
GDP-EA	9.16	9.40	9.39	9.85	15.0	48.8
GDP-US	14.5	5.00	16.3	5.00	5.00	39.7

Note: The base sample is from 2000Q4 to 2011Q4. The optimal size is computed based on the multivariate analysis outlined in Section 5 and is summarised in Algorithm 1.

resulting Type II ones seems ill equipped to address the problem. Therefore, we approach the problem of setting Type I errors based on loss functions. This leads to a strategy for trading off the two errors. The use of decision theoretic considerations to modify statistical testing seems novel, certainly within our macroeconometric context.

The paper discusses extensively the problem, sets out the detailed theoretical setting and implements it using Bank of England forecasts. We find that, in a number of cases, resulting choices for Type I error, based on our chosen loss functions, exceed considerably standard values. Our new approach for detecting shifts in forecast accuracy opens up a new avenue for policy makers to study forecast breaks when the sample size is small. As a matter of fact, our procedure serves as a starting point for a *case-by-case* investigation into the cause of forecast breaks. In our discussion, we demonstrate that, rather than relying on one single metric, we use a series of metrics for investigation.

References

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*(1), 47–78.

Bank of England (2015). *Evaluating forecast performance*. Indepndent Evaluation Office. Bank of England, November 2015.

Bank of England (2016). *Inflation Report*. Bank of England, February 2016.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer.

Brown, R. L., J. Durbin, and J. M. Evans (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological) 37*(2), 149–192.

Chu, C.-S. J., M. Stinchcombe, and H. White (1996). Monitoring structural change. *Econometrica 64*(5), 1045–1065.

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill Series in Probability and Statistics. McGraw Hill.

Giacomini, R. and B. Rossi (2009). Detecting and predicting forecast breakdowns. *Review of Economic Studies 76*(2), 669–705.

Granger, C. W. J. and M. J. Machina (2006). Forecasting and decision theory. In C. W. G. G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*. Elsevier.

Granger, C. W. J. and M. H. Pesaran (2000a). A decision-theoretic approach to forecast evaluation. In W. S. Chan, W. K. Li, and H. Tong (Eds.), *Statistics and Finance: An Interface*. London: Imperial College Press.

Granger, C. W. J. and M. H. Pesaran (2000b). Economic and statistical measures of forecast accuracy. *Journal of Forecasting 19*, 537–560.

Groen, J. J. J., G. Kapetanios, and S. Price (2013). Multivariate Methods For Monitoring Structural Change. *Journal of Applied Econometrics 28*, 250–274.

Hansen, L. P. and T. J. Sargent (1980, May). Formulating and estimating dynamic linear rational expectations models. *Journal of Economic Dynamics and Control 2*(1), 7–46.

Kuan, C.-M. and K. Hornik (1995). The generalized fluctuation test: a unifying view. *Econometric Reviews 14*, 135–61.

Pesaran, M. H. and S. Skouras (2002). Decision-based methods for forecast evaluation. In M. P. Clements and D. F. Hendry (Eds.), *A Companion to Economic Forecasting*. Blackwell.

Ploberger, W., W. Kramer, and K. Kontrus (1989). A new test for structural stability in the linear regression model. *Journal of Econometrics 40*(2), 307 – 318.

Savage, L. J. (1954). *The Foundations of Statistics*. Dover Publications Inc.

Smets, F. and R. Wouters (2007, June). Shocks and frictions in us business cycles: A bayesian dsge approach. *American Economic Review 97*(3), 586–606.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics 166*, 157–165.

Tetenov, A. (2016). An economic theory of statistical testing. Working paper, University of Bristol.

Wald, A. (1950). *Statistical Decision Functions*. John Wiley and Sons.

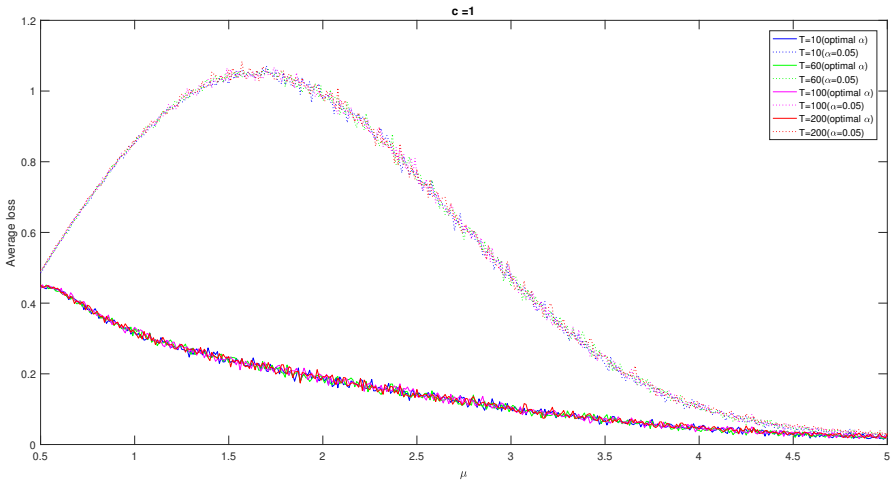
Appendix

A Monte Carlo Simulations

In Section 6, we assume $T = 100$ in our simulations. We perform evidence that our simulations are unaffected by the choice of T because of our simulation setup in section 4.2. For brevity, we only present evidence when Type II loss is known, but results extend to the situation where Type II loss is not known.

To that end, we show Figure A.1, a plot of the average losses $\boldsymbol{L}^{c,\mu}_{\alpha=\alpha(c,\mu)}$ and $\boldsymbol{L}^{c,\mu}_{\alpha=0.05}$ with $c = 1$ with the number of observations $T \in 10, 60, 100, 200$. There is little difference between the four set of simulations. Such result is held across all values of c.

Figure A.1: Average loss for Monte Carlo simulations when Type II loss is known, different values of T



Note: Average loss is computed following the Algorithm 2. Refer to main text for details.

B Forecast bias in the data

Table B.1: Forecast bias for each variable in the data

	Base sample bias 2000Q4-11Q4				Base sample bias 2000Q4-11Q4 excl 08-09			Evaluation sample bias 2012Q1-16Q4			
	1	4	8	12	1	4	8	1	4	8	12
GDP	-0.035	-0.793**	-1.275**	-1.447**	0.225*	-0.146	-0.330**	0.148	-0.137	-0.551**	-0.752**
CPI	-0.398**	0.030	0.276*	0.097	-0.509**	-0.034	-0.157	-0.091**	-0.540**	-0.046	-0.032

Note: This table documents the forecast bias for each variable in each subsample period. The unit is in percent. The left panel considers the base sample from 2000 to 2012 including observations from the financial crisis period between 2008-2009; the central panel considers the base sample excluding observations from the financial crisis period; and the right panel considers the evaluation sample between 2012 and 2016. '*' denotes that the bias is statistically significant from zero at 10%, whereas '**' denotes significance at 5%. The p-values which are used to conduct inferences are simulated to adjust for potential small sample bias in the data. Refer to the main text for details.