Forecasting with measurement errors in dynamic models

Richard Harrison[*]

and

George Kapetanios[**]

and

Tony Yates[†]

Working Paper no.

[*] Bank of England.
E-mail: Richard.Harrison@Bankofengland.co.uk

[**] Bank of England.
E-mail: George.Kapetanios@Bankofengland.co.uk

[†] Bank of England.
E-mail: Tony.Yates@Bankofengland.co.uk

Our work on this area was initiated by conversations with a number of people including Kosuke Aoki, Peter Andrews, Charlie Bean, Spencer Dale, Colin Ellis, Nigel Jenkinson, Steve Nickell, Kalin Nikolov, Simon Price, Meghan Quinn, Tom Sargent and Alasdair Scott. This paper represents the views and analysis of the authors and should not be thought to represent those of the Bank of England or Monetary Policy Committee members.

Copies of working papers may be obtained from Publications Group, Bank of England, Threadneedle Street, London, EC2R 8AH; telephone 020 7601 4030, fax 020 7601 3298, e-mail mapublications@bankofengland.co.uk

Working papers are also available at www.bankofengland.co.uk/wp/index.html

The Bank of England's working paper series is externally refereed.

©Bank of England 2002
ISSN 1368-5562

Contents

Abstract

This paper explores the effects of measurement error on dynamic forecasting models. The paper sets out to illustrate a trade off that confronts forecasters and policymakers when they use data that are measured with error. On the one hand, observations on recent data give valuable clues as to the shocks that are hitting the system and will be propagated into the variables to be forecast (and which ultimately will inform monetary policy). But on the other, those recent observations are likely to be those least well measured. Two broad classes of results are illustrated. The first relates to cases where it is imagined that the forecaster takes the coefficients in the data generating process as a given, and has to choose how much of the historical time series of data to use to form a forecast. It is shown that if recent data is sufficiently badly measured, relative to older data, that it can be optimal in this case not to use old data at all. The second class of results is more general. Here, it is shown that for a general class of linear autoregressive forecasting models, the optimal weight to place on a data observation of some age, relative to the weight in the true data generating process, will depend on the measurement error in that data. The gains to be had in forecasting are illustrated using a model of UK business investment growth.

Key words: Forecasting, measurement error, signal extraction

JEL classification: C53

Summary

This paper explores a trade-off that confronts forecasters and monetary policymakers when they use data that are measured with error (as surely, in reality, they are forced to). On the one hand, observations on recent data give valuable clues as to the shocks that are hitting the system and will be propagated into the variables to be forecast (and which ultimately will inform monetary policy). But on the other, those recent observations are likely to be those least well measured. Older data may have been revised a number of times as more survey returns on the data or other kinds of corrobarative evidence were collected by the statistics agency. We begin by illustrating and proving how, faced with a choice between either using or not using most recent observations in forecasting, once measurement error is sufficiently large it can be optimal not to use it. We move on to consider a case when measurement error is larger, the more recent the data observation: this captures the idea that recent data, more likely to be a first release, will be more noisy than older data, which may have been revised and corrobrated with information that came later. We derive conditions under which a many-step-ahead forecast, (based on older data) will be a better forecast (in fact the optimal forecast) than a one-step-ahead forecast. The noisier are recent data, the more likely this is to be true. And the more persistent the data generating process, the more likely this is to be true (because old shocks contain more information for values of variables in the immediate future). By assuming a declining variance structure for the revision errors, we therefore create a tradeoff between more noisy but more recent data and less noisy older data. We generalise these results further, by allowing the forecaster to 'aim off' the coefficients in the true model, to improve forecasts still further. Finally, we derive the optimal forecasting model from the class of linear autoregressive models. We can view this as describing how the policymaker decides on the optimal weights to place on past data. This structure therefore allows the policymaker to include many lags of the data to construct the forecast and place different weights (coefficients) on different lags. This is clearly more general than the analysis in previous sections which constrained the forecaster to either use or not use at all some particular data. It is not surprising that the optimal weighting scheme differs from the weighting scheme that characterises the data generating process. The greater the signal about the future in a data point, the greater the weight in the optimal forecasting model. More recent and therefore more imprecisely measured data have a smaller weight. The greater the persistence in the data generating process, the greater the signal in older data for the future, and the more extra measurement error in recent data relative to old data makes it optimal to rely on that older data. We conclude with an application to UK

business investment growth, and illustrate the improvement in forecasting performance that can be got using our procedure, an improvement that turns out to be statistically significant.

1 Introduction

This paper explores a trade-off that confronts forecasters and monetary policymakers when they use data that are measured with error (as surely, in reality, they are forced to). On the one hand, observations on recent data give valuable clues as to the shocks that are hitting the system and will be propagated into the variables to be forecast (and which ultimately will inform monetary policy). But on the other, those recent observations, the ones likely to contain the most information about the future profile of the data, contain measurement error that will induce errors in the forecast. Moreover, the more recent a data observation is, the more measurement error it is likely to contain. The most recent data may be a first release. Older data may have been revised a number of times as more survey returns on the data or other kinds of corrobarative evidence were collected by the statistics agency.(1) The best forecast in this (very real) situation will balance the information about shocks contained in recent data against the contaminating noise of the measurement error. One option, of course, is to wait for data observations to improve; to wait for corroborative evidence to accumulate. However, this is not an available option in one prominent example where forecasting is important: monetary policymaking. Monetary policy takes time to have an effect. And the economy may be such that it is appropriate to respond to shocks sooner rather than later to avoid larger fluctuations in (say) inflation and/or output which are themselves undesirable. So there is an advantage to using the signal contained in the most recent data for forecasting and policy.

This is of course a very familiar problem, and there is a large literature that attempts to examine aspects of forecasting and monetary policymaking in 'real time', when data are likely to be of poorer quality than at some future date. We do not attempt to do any justice to the depth or diversity of this literature here, but to put our work in context it is worth mentioning a few important strands of research.(2) Real-time data sets that enable economists to study the properties of different vintages of data relevant to policymaking have been compiled by **?** for the US, and by **?**, **?** and **?** for the UK. Others (for example **?**, **?**, though this literature is very large indeed) have studied whether the statistics agency behaves like a 'rational' forecaster by examining whether early releases of data predict later ones. Still others have studied the implications for monetary policy and inflation forecasts of having to use real-time measures of important indicators like the output gap (**?**and **?**).

(1) See (**?**, page 44) for a discussion of the reasons why data are revised in the UK.
(2) A helpful bibliography, can be found at http://phil.frb.org/econ/forecast/reabib.html.

Within this broad literature are papers that study the properties of forecast models in the presence of measurement error, and these are the closest intellectual antecedents of our own. One line of enquiry has been to study a problem of joint model estimation and signal extraction/forecasting. Optimal filters/forecasts are studied in a line of work that runs from, for example, **?** through to **?**. **?** present informal experiments that reveal the advantages for forecasting of using real-time data for model estimation. Another focus for study has been the idea of evaluating the properties of combinations of forecasts (see, for example, **?** and discussions in **?**). Observations on time series at dates leading up to time t are 'forecasts' of sorts of data at time t, so the problem of how best to make use of these data is a problem of combining forecasts.[(3)]

This paper puts to one side the problem of model estimation. We assume that the forecaster/policymaker knows the true model.[(4)] Taken at face value, this looks like a very unrealistic assumption. But it has two advantages. First, it enables us to isolate the forecasting problem, without any loss of generality. The second advantage is that it also emphasises an aspect of forecasting and policy that is realistic. The policymaker may have a noisy information source that is contaminated with measurement error, but also contains an important signal about shocks. The policymaker may also have an information source that is not contaminated by (at least that source of) measurement error – an economic prior – but that does not contain the same high frequency diagnosis of the state of the economy. The set up we use is just an extreme version of this. We assume that the policymaker's prior about the structure of the economy (the data generating proces) is correct.

We begin (in Section 2.1) by illustrating and proving how, faced with a choice between either using or not using most recent observations in forecasting, once measurement error is sufficiently large it can be optimal not to use it. Specifically, we assume that the policymaker has access to data for periods $t = 0, ..., T$ and makes forecasts for date $T + 1$ using data available at date $T + 1 - n$ and the true model. The policymaker's choice variable is n. Conventional forecasting analysis usually assumes that the forecaster/policymaker sets $n = 0$ which means using all available data. Here we relax that assumption. In the illustration we offer, of a mean-reverting dynamic process, the structure of which is known to policymakers, it is optimal to resort to using the mean as the forecast, rather than (in this simple example, any of) the noisy data. In effect this amounts to choosing $n = \infty$.

(3) This observation is made in **?**.

(4) We examine model estimation separately in a related paper, **?**.

We move on (in Section 2.2) to consider a case when measurement error is larger, the more recent the data observation: this captures the idea that recent data, more likely to be a first release, will be more noisy than older data, which may have been revised and corrobrated with information that came later. We derive conditions under which a many-step-ahead forecast, (based on older data) will be a better forecast (in fact the optimal forecast) than a one-step-ahead forecast. The noisier are recent data, the more likely this is to be true. And the more persistent the data generating process, the more likely this is to be true (because old shocks contain more information for values of variables in the immediate future).(5) By assuming a declining variance structure for the revision errors, we therefore create a tradeoff between more noisy but more recent data and less noisy older data. In general this means that there is some finite n that minimises the mean squared error of the forecasts.

In Section 3 we generalise these results a little further. We began by deriving optimal forecasts when forecasters took the model parameters as a given, and had to find the optimal point at which to end the data frontier for the forecast. In Section 3, the problem is characterised still as one in which the forecaster has to choose the optimal point (n) at which to end the data frontier for a forecast of a variable at time $T + 1$. But now we assume that the policymaker can also choose the parameters of the forecast model. Specifically we analyse the case in which the policymaker knows that the data is generated by a first order autoregressive process with parameter a. To generate the forecast, we allow the policymaker to use some other parameter $\tilde{a}$ that may differ from the true model parameter, a. We find that the optimal point at which to end the data frontier does not necessarily imply using all the most recent data, and that that optimal forecast does not imply a parameter equal to the 'true' one. (Except, of course, in the limiting case when there is no measurement error).

Section 4 generalises the results by deriving the optimal forecasting model from the class of linear autoregressive models. We can view this as describing how the policymaker decides on the optimal weights to place on past data. This structure therefore allows the policymaker to include many lags of the data to construct the forecast and place different weights (coefficients) on different lags. This is clearly more general than the analysis in previous sections which constrained the forecaster to either use or not use at all some particular data. It is not surprising that the optimal weighting scheme differs from the weighting scheme that characterises the data

(5) These points are illustrated too in **?** but we prove them here.

generating process. The greater the signal about the future in a data point, the greater the weight in the optimal forecasting model. More recent and therefore more imprecisely measured data have a smaller weight. The greater the persistence in the data generating process, the greater the signal in older data for the future, and the more extra measurement error in recent data relative to old data makes it optimal to rely on that older data.

In Section 5, we present an application of the results in Section 4 to a single equation forecasting model for investment spending in the UK, though we think the theoretical results have a very general implication for linear forecasting models. We use real time data on revisions to national accounts from **?** to estimate how the variance of measurement error declines as we move back in time from the data frontier at T to some $T - n$. We find, not surprisingly, that indeed the optimal forecasting model differs significantly from the weights put on data implied by the underlying estimated model, suggesting that the problem we study here may well be quantitatively important.

2 Optimal choice of the data frontier for forecasting

2.1 Age-invariant measurement error

We begin, as we described in the introduction, by illustrating how it may be optimal not to use recent (in fact in this example, any) data for forecasting, but instead to rely on the model, which we assume is known. In this section we use a very simple model, but we will relax some of our assumptions in later sections.

Assume that the true model is

$$y_t^* = a y_{t-1}^* + e_t \quad \textbf{(1)}$$

where $|a| < 1$ and y_t^* denotes the true series. Data is measured with error, and the relationship between the true and observed series is given by

$$y_t = y_t^* + v_t \quad \textbf{(2)}$$

For this section, we make the following assumptions about the processes for e and v:

$$e_t \sim i.i.d.(0, \sigma_e^2)$$

and

$$v_t \sim i.i.d.(0, \sigma_v^2)$$

which encompasses the assumption that the measured data are unbiased estimates of the true data.[6]

Importantly, at this stage we are assuming that the variance of the distribution of the measurement error does not depend on how recently the data observation was released, or that the measurement error is, as the title of this section dubs it 'age-invariant'. We will relax this assumption later in the paper. We assume that we have a sample from period $t = 1$ to period $t = T$ and we wish to forecast some future realisation y^*_{T+1}. The standard forecast, (when there is no measurement error) for y^*_{T+1} is denoted by $\hat{y}^{(0)}_{T+1}$ and given by $\hat{y}^{(0)}_{T+1} = ay_T$: this is the forecast that simply projects the most recent observation of y_t using the true model coefficient a. We investigate the mean square properties of this forecast compared with the general forecast $\hat{y}^{(n)}_{T+1} = a^{n+1}y_{T-n}$, a class of forecasts that project using data that are older than the most recent outturn.

We begin by finding an expression for the forecast error, and then computing the mean squared error for different forecasts amongst the general class described above. The (true) forecast error (which of course we never observe) is given by $\hat{u}^{(n)}_{T+1} = y^*_{T+1} - \hat{y}^{(n)}_{T+1}$. We know that from (**(1)**) we can write:

$$y^*_{T+1} = a^{n+1}y^*_{T-n} + \sum_{i=0}^{n} a^i e_{T+1-i}$$

and from (**(2)**) we have:

$$\begin{aligned}\hat{y}^{(n)}_{T+1} &= a^{n+1}y_{T-n} \\ &= a^{n+1}y^*_{T-n} + a^{n+1}v_{T-n}\end{aligned}$$

So:

$$\begin{aligned}\hat{u}^{(n)}_{T+1} &= a^{n+1}y^*_{T-n} + \sum_{i=0}^{n} a^i e_{T+1-i} - a^{n+1}y^*_{T-n} - a^{n+1}v_{T-n} \\ &= \sum_{i=0}^{n} a^i e_{T+1-i} - a^{n+1}v_{T-n}\end{aligned}$$

(6) The analysis in **?** focuses on the first moment properties of revisions and finds some evidence of bias. But we abstract from that issue here.

Therefore, the mean square error is simply given by:[7]

$$MSE(n) = a^{2(n+1)}\sigma_v^2 + \left(1 + \frac{a^2 - a^{2(n+1)}}{1 - a^2}\right)\sigma_e^2$$

The next step is to explore the condition that the mean squared error from a forecast using the most recent data is less than the mean squared error that uses some other more restricted information set, or $MSE(0) < MSE(n)$ for some $n > 0$. This will tell us whether there are circumstances under which it is worth forecasting without using the latest data. Doing this gives us:

$$MSE(0) < MSE(n) \Rightarrow a^2\sigma_v^2 + \sigma_e^2 < a^{2(n+1)}\sigma_v^2 + \left(1 + \frac{a^2 - a^{2(n+1)}}{1 - a^2}\right)\sigma_e^2$$

which can be written as:

$$\left(a^2 - a^{2(n+1)}\right)\sigma_v^2 < \frac{a^2 - a^{2(n+1)}}{1 - a^2}\sigma_e^2$$

which reduces to

$$\sigma_v^2 < \frac{\sigma_e^2}{1 - a^2} \qquad \textbf{(3)}$$

So if $\sigma_v^2 > \frac{\sigma_e^2}{1-a^2}$ it is better in terms of MSE *not* to use the most recent data. The intuition is simply that if the variance of the measurement error σ_v^2 is very large relative to the shocks that hit the data generating process, (σ_e^2), then it is not worth using the data to forecast: the more so the smaller is the parameter that propagates those shocks (a). In fact it follows that if $\sigma_v^2 > \frac{\sigma_e^2}{1-a^2}$ then $MSE(n-1) > MSE(n)$ for all n and therefore we are better off using the unconditional mean of the model to forecast the true series than any other data.

There are two alternative ways to describe this result. The first uses the signal-to-noise ratio, defined as $\sigma^2 = \sigma_e^2/\sigma_v^2$. Then, if $\sigma^2 < (1 - a^2)$, it is optimal not to use the most recent data. Under this interpretation, the lower the signal-to-noise ratio, the more likely it is that using the mean of the process will provide the best forecast. The critical value of the signal-to-noise ratio is a decreasing function of the persistence of the process – as the persistence of the process increases, the strength of past signals in later data increases. The second interpretation involves the observation that $\sigma_e^2/\left(1 - \alpha^2\right)$ corresponds to the unconditional variance of the true data, y^*. If we denote the variance of the data as $\sigma_{y^*}^2$, then it is optimal not to use recent data if condition is $\sigma_v^2 > \sigma_{y^*}^2$. This interpretation shows that it is only worth using recent data if the unconditional variance of the data measurement errors is smaller than the unconditional variance of the true data, which is an intuitive result. The practical relevance of this result may be questionable as it seems likely that there are few data series in practice that are so badly measured. But the underlying

(7) Notice that this expression requires that the revision errors in (**(2)**) are uncorrelated with future shocks to the model (**(1)**). This seems like a reasonable assumption.

assumptions in this section are somewhat over-simplified and the analysis of the subsequent sections applies more generally.

The above analysis concentrated on a simple AR(1) model. However, the intuition is clear and is valid for general AR models and more general dynamic models. The increasing cost of the cumulative sum of the structural errors (e_t) is balanced against the falling cost of smaller measurement errors ($a^i v_t$) when older data are used. In the case of this simple model there is no solution where some old data are useful. Either the most recent data should be used or no data at all, once – and this is crucial – the mean of the series is known. This is because older data are as well measured as more recent data.

2.2 Age-dependent measurement error

We now investigate a slightly more complex case where the variance of the data measurement error v_t is assumed to tail off over time. This assumption reflects the observation that, in practice, we observe that statistics agencies revise data often many times after the first release. If we assume that successive estimates of a particular data point are subject to less uncertainty (since they are based on more information), then it seems reasonable to assume that the variance of the revision error embodied in the estimate of a particular data point diminishes over time.

The specific assumption we make here is that:

$$Var(v_{T-i}) = \begin{cases} b^i \sigma_v^2, & i = 0, 1, \ldots, N \\ 0, & i = N+1, \ldots \end{cases}$$

for a parameter $0 < b < 1$. We therefore assume that after a finite number of periods $N + 1$, there are no further revisions to the data. But for the first $N + 1$ periods, the variance of the revision error declines geometrically over time at a constant rate measured by b. This is a fairly specific assumption which we make here for simplicity and tractability (again the analysis of later sections is more general). Indeed, we know that data are revised for reasons other than new information specific to that series (for example re-basing and methodology changes) so the specification of revision error variance may be more complicated than we have assumed here. But the purpose of the assumption is to be more realistic than the homoskedastic case considered in Section 2.1.

Under our assumptions, the MSE as a function of n is given by

$$MSE(n) = a^{2n+2}b^n\sigma_v^2 + \sum_{i=0}^{n} a^{2i}\sigma_e^2, \;\; n = 0, 1, \ldots, N$$

$$\begin{aligned} MSE(n) &= \sum_{i=0}^{n} a^{2i}\sigma_e^2 \\ &= \sum_{i=0}^{N} a^{2i}\sigma_e^2 + \sum_{i=N+1}^{n} a^{2i}\sigma_e^2, \;\; n = N+1, \ldots \end{aligned}$$

We want to examine when $MSE(n) > MSE(N+1), n = 0, 1, \ldots, N$. It is clear that $MSE(n) > MSE(N+1), n = N+2, \ldots,$. So, for $n = 0, 1, \ldots, N$

$$MSE(n) > MSE(N+1) \Rightarrow a^{2n+2}b^n\sigma_v^2 + \sum_{i=0}^{n} a^{2i}\sigma_e^2 > \sum_{i=0}^{n} a^{2i}\sigma_e^2 + \sum_{i=n+1}^{N+1} a^{2i}\sigma_e^2$$

or

$$a^{2n+2}b^n\sigma_v^2 > \sum_{i=n+1}^{N+1} a^{2i}\sigma_e^2$$

or

$$a^{2n+2}b^n\sigma_v^2 > a^{2n+2}\frac{1-a^{2(N-n+1)}}{1-a^2}\sigma_e^2$$

or, in terms of the signal-noise ratio, σ:

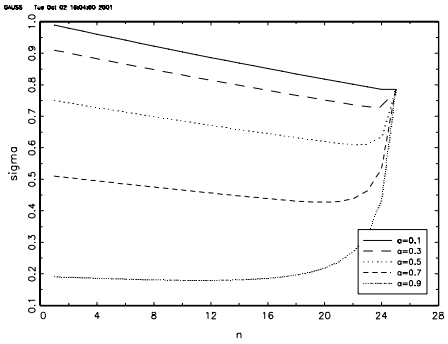
$$\frac{b^n(1-a^2)}{1-a^{2(N-n+1)}} > \sigma^2$$

So if $\sigma^2 < \frac{b^n(1-a^2)}{1-a^{2(N-n+1)}}$ for all n then the best forecast for y_{t+1} is $\hat{y}_{t+1}^{(N+1)}$. To clarify the range of relevant values for σ we graph the quantity $\frac{b^n(1-a^2)}{1-a^{2(N-n+1)}}$ over n for $N = 24$, $b = 0.99, 0.95, 0.9, 0.5$ and $a = 0.1, 0.3, 0.5, 0.7, 0.9$ in Figures 1a-1d. If each period corresponds to one quarter, then our assumption $N = 24$ corresponds to the situation in which data are unrevised after six years. While this is naturally an approximation (since rebasing and methodological changes can imply changes to official figures over the entire length of the data series) it seems a plausible one.

Clearly, the more persistent the process is (the larger the a) the lower σ^2 has to be for $\hat{y}_{t+1}^{(N+1)}$ to be the best forecast. Also, the more slowly the revision error dies out (the larger the b), the lower σ^2 has to be for $\hat{y}_{t+1}^{(N+1)}$ to be the best forecast. Note that some of the curves in the figures are not monotonic. This indicates that although $\hat{y}_{t+1}^{(N+1)}$ is a better forecast than $\hat{y}_{t+1}^{(0)}$, there exists some $N+1 > n > 0$ such that $\hat{y}_{t+1}^{(n)}$ is better than $\hat{y}_{t+1}^{(N+1)}$.

Diagram 1: Critical values for the signal:noise ratio

1a: b=0.99



1b: b=0.95

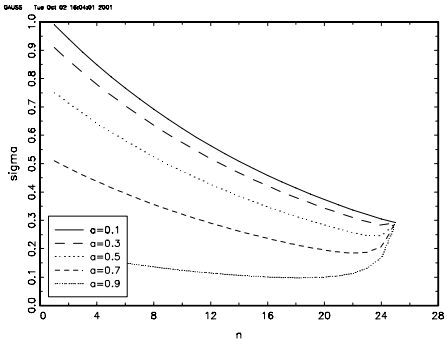
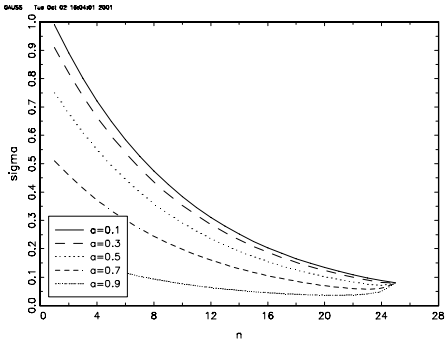
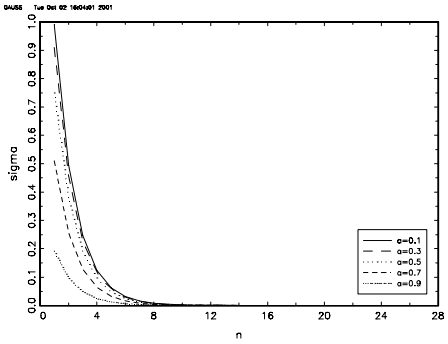


Diagram 1: (Continued)

1c: b=0.9



1d: b=0.5



3 Optimising over the choice of data frontier and the projection parameters

The analysis in the previous section constrained the policymaker/forecaster to use the true model when forecasting future outturns. The only choice variable was therefore the horizon n upon which to base the forecast $\hat{y}_{T+1}^{(n)} = a^{n+1} y_{T-n}$. This section generalises the problem of the policymaker/forecaster so that it is possible to construct a forecast that does not use the true model parameter. Specifically, we allow the policymaker/forecaster to use the forecast $\hat{y}_{t+1}^{(n)}(\tilde{a}) = \tilde{a}^{n+1} y_{t-n}$ where $\tilde{a}$ may differ from a. In this setting, there are two choice variables ($\tilde{a}$ and n) and so it might be more appropriate to use a different parameter value, $\tilde{a}$, and the most recent data as opposed to older data and the true model parameter, a.

This section therefore extends the setup and views the mean square error as a function of n and $\tilde{a}$ where the forecast is given by $\hat{y}_{t+1}^{(n)}(\tilde{a}) = \tilde{a}^{n+1} y_{t-n}$ and $\tilde{a}$, n are to be jointly determined given the structural parameters a, σ_v^2 and σ_e^2. We extend the analysis along these lines assuming that the revision error variance is given by $Var(v_{t-i}) = b^i \sigma_v^2, i = 0, 1, \ldots, N$ and $Var(v_{t-i}) = 0, i = N+1, \ldots$ as before.

Now the mean square error is a joint function of n and $\tilde{a}$ given by

$$MSE(n, \tilde{a}) = (a^{n+1} - \tilde{a}^{n+1})^2 \frac{\sigma_e^2}{1-a^2} + \frac{(1-a^{2(n+1)})\sigma_e^2}{1-a^2} + \tilde{a}^{2(n+1)} b^n \sigma_v^2, \;\; n = 0, 1, \ldots, N$$

$$MSE(n, \tilde{a}) = (a^{n+1} - \tilde{a}^{n+1})^2 \frac{\sigma_e^2}{1-a^2} + \frac{(1-a^{2(n+1)})\sigma_e^2}{1-a^2}, \;\; n = N+1, \ldots$$

and we wish to find the optimal values for n and $\tilde{a}$. To do so, we analyse a two-step minisimisation problem. First we will minimise the mean squared error with respect to the forecasting parameter $\tilde{a}$. This allows us to write down a set of mean-squared errors that use the optimal forecasting parameters as n changes. To find the best forecast simply requires choosing the n that gives the overall smallest mean-squared error.

We therefore begin by minimising $MSE(n, \tilde{a})$ with respect to $\tilde{a}$. The first order necessary conditions are:

$$\frac{\partial MSE(n, \tilde{a})}{\partial \tilde{a}} = \begin{cases} -2(a^{n+1} - \tilde{a}^{n+1})(n+1)\tilde{a}^n \frac{\sigma_e^2}{1-a^2} + 2(n+1)\tilde{a}^{2n+1} b^n \sigma_v^2 = 0 & n = 0, 1, \ldots, N \\ -2(a^{n+1} - \tilde{a}^{n+1})^2 (n+1)\tilde{a}^n \frac{\sigma_e^2}{1-a^2} = 0 & n = N+1, \ldots \end{cases}$$

Rearranging gives

$$-\tilde{a}^n \left[a^{n+1} - \left(1 - \frac{b^n(1-a^2)}{\sigma^2} \right) \tilde{a}^{n+1} \right] = 0, \;\; n = 1, \ldots, N \qquad \textbf{(4)}$$

$$-\tilde{a}^n \left(a^{n+1} - \tilde{a}^{n+1} \right) = 0, \;\; n = N+1, \ldots \qquad \textbf{(5)}$$

For (**(4)**), disregarding complex roots and under the convention of square roots being positive numbers, the solutions are $\tilde{a} = 0$ and $\tilde{a} = \sqrt[n+1]{\theta}$ where $\theta = \frac{\sigma^2 a^{n+1}}{\sigma^2 + b^n(1-a^2)}$. For (**(5)**), they are, intuitively, $\tilde{a} = 0$ and $\tilde{a} = a$. Note that for positive a, $\theta \geq 0$ making sure than the second solution of (**(4)**) is real. Just to verify that the solutions we have are proper minima we compute the second derivatives. These are given by

$$-n\tilde{a}^{n-1} \left[a^{n+1} - \left(1 - \frac{b^n(1-a^2)}{\sigma^2} \right) \tilde{a}^{n+1} \right] + \tilde{a}^n \left(1 - \frac{b^n(1-a^2)}{\sigma^2} \right) (n+1)\tilde{a}^n \;\; n = 0, 1, \ldots, N$$

$$-n\tilde{a}^{n-1} \left(a^{n+1} - \tilde{a}^{n+1} \right) + (n+1)\tilde{a}^{2n} \;\; n = N+1, \ldots$$

In both case the second derivatives are positive for the nonzero solution and zero for the zero solution. The non-zero solutions are therefore minima.

We can now incorporate the solutions of this minimisation into the expression for the mean squared error. We define $\hat{MSE}(n) = \min_{\tilde{a}} MSE(n, \tilde{a})$:

$$\hat{MSE}(n) = \left[(a^{n+1} - \theta)^2 + (1 - a^{2(n+1)})\right] \frac{\sigma_e^2}{1 - a^2} + \theta^2 b^n \sigma_v^2 \quad n = 0, 1, \ldots, N$$

$$\hat{MSE}(n) = \frac{(1 - a^{2(n+1)})\sigma_e^2}{1 - a^2} \quad n = N + 1, \ldots$$

which has to be minimised over n. Unfortunately standard methods do not apply as n takes only discrete values[8]. However, for given parameter values we can compute $\hat{MSE}(n)$ for a grid of n and get the minimum. What is clear is that it is not neccesary that the minimum is obtained at $n = 0$, thereby leading to the same conclusion as before.

4 A general approach to forecasting with dynamic models under data revisions

The analysis of previous sections has gradually increased the generality of the problem under consideration. To recap, we began by considering the choice of the the lag, n, of data that the policymaker/forecaster would use with the true model parameter in the face of (both age invariant and age dependent) measurement error. We then allowed the policymaker/forecaster to choose both the lag, n, and the parameter with which to project the lagged data, $\tilde{a}$.

In this section we propose a general method of forecasting in autoregressive models under a general known form of data revisions. The extension from the previous sections is that we optimise the forecasting model from within the linear class of models. Specifically, we allow the policymaker/forecaster to choose the optimal weights and lags on all past data. So, in the example of the previous sections the policymaker/forecaster knew that the data generating process was an AR(1) with parameter a. But even so, this section allows the forecaster/policymaker to choose from all linear AR models and it may be the case that the best forecasting model is (say) an AR(2) with coefficients b and c. The method described here can be easily extended to multivariate

(8) We can note that the minimum of any function is also the minimum of a positive monotonic transformation of that function. So we can take logs of the mean-squared-error expressions to show that, for real n, the minimum is obtained for

$$n = \frac{\log\left[\frac{(1-a^2)\theta \log(b)}{2\sigma^2 a \log(a)}\right]}{\log(a/b)}$$

models. In particular VAR models could easily be accomodated (including of course vector error correction models).

We assume a univariate $AR(p)$ model to illustrate the method. The true process is given by

$$y_t^* = \sum_{i=1}^{p} a_i y_{t-i}^* + e_t$$

This can be written as a VAR(1) model of the form

$$\mathbf{y}_t^* = \mathbf{A}\mathbf{y}_{t-1}^* + \epsilon_t$$

where $\mathbf{y}_t^* = (y_t^*, y_{t-1}^*, \ldots, y_{t-p+1}^*)'$, $\mathbf{A} = (\mathbf{a}, \mathbf{e}_1, \ldots, \mathbf{e}_{p-1})'$; $\mathbf{e}_i$ is a $p \times 1$ vector with an element of 1 at the i-th place and zeroes everywhere else; $\mathbf{a}$ is a $p \times 1$ vector of the autoregressive coefficients a_i; $\epsilon_t = (e_t, 0 \ldots, 0)'$. Now the observed data are given by

$$\mathbf{y}_t = \mathbf{y}_t^* + \mathbf{v}_t$$

where $\mathbf{v}_t = (v_t, v_{t-1}, \ldots, v_{t-p+1})$. At time T we wish to determine the optimal forecast for y_{T+1}^*. We assume that the revision error $\mathbf{v}_T$ has a variance matrix which is given by Σ_v^T. Our aim is to determine the optimal forecasting model of the form $\hat{y}_{T+1} = \tilde{\mathbf{A}}_1 \mathbf{y}_T$. in terms of mean square error, where $\tilde{\mathbf{A}}_1$ is a $1 \times p$ vector. Note that the restriction on the dimension of $\tilde{\mathbf{A}}_1$ to be the same as that of the order of the true process is not problematic because we can simply increase the order of the process by setting the higher order a's equal to zero. This means that the true data generating process might be an AR(1) even though we can write it as an AR(p) with the coefficients on lags $2, \ldots, p$ set equal to zero.

The forecast error for the forecast of the above form is given by

$$y_{T+1}^* - \hat{y}_{T+1} = \mathbf{A}_1 y_T^* + \epsilon_T - \tilde{\mathbf{A}}_1 \mathbf{y}_T^* + \tilde{\mathbf{A}}_1 \mathbf{v}_T = (\mathbf{A}_1 - \tilde{\mathbf{A}}_1)\mathbf{y}_T^* + \tilde{\mathbf{A}}_1 \mathbf{v}_T + \epsilon_{T+1}$$

where $\mathbf{A}_1$ is the first row of $\mathbf{A}$. The mean square error is given by

$$(\mathbf{A}_1 - \tilde{\mathbf{A}}_1)\grave{}(\mathbf{A}_1 - \tilde{\mathbf{A}}_1)' + \tilde{\mathbf{A}}_1 \Sigma_v^T \tilde{\mathbf{A}}_1' + \sigma_\epsilon^2$$

where $\Gamma = E(\mathbf{y}_T^* \mathbf{y}_T^{*'})$. The covariances of an AR(p) process are given by the first p elements of the first column of the matrix $\sigma_\epsilon^2[\mathbf{I}_{p^2} - \mathbf{A} \otimes \mathbf{A}]^{-1}$. We have assume that the error process is uncorrelated with the true process of the data. In the data revision literature this is referred to as the error-in-variables model. This assumption is not crucial to our analysis and could be relaxed as long as the covariances between the true process and the data revision errors could be estimated. We want to minimise the mean square error in terms of $\tilde{\mathbf{A}}_1$. We will use matrix optimisation calculus to solve this problem. We rewrite the expression for the mean square error

using only terms involving $\tilde{\mathbf{A}}_1$ since the rest of the terms will not affect the minimisation. We have that the mean square error is given by

$$\tilde{\mathbf{A}}_1\Gamma\tilde{\mathbf{A}}_1' + \tilde{\mathbf{A}}_1\Sigma_v^T\tilde{\mathbf{A}}_1' - \mathbf{A}_1\Gamma\tilde{\mathbf{A}}_1' - \tilde{\mathbf{A}}_1\Gamma\mathbf{A}_1' = \tilde{\mathbf{A}}_1(\Gamma + \Sigma_v^T)\tilde{\mathbf{A}}_1' - 2\tilde{\mathbf{A}}_1\Gamma\mathbf{A}_1'$$

We differentiate with respect to $\tilde{\mathbf{A}}_1$ and set to zero to get

$$(\Gamma + \Sigma_v^T)\tilde{\mathbf{A}}_1' - \Gamma\mathbf{A}_1' = 0$$

giving

$$\tilde{\mathbf{A}}_1^{opt\ '} = (\Gamma + \Sigma_v^T)^{-1}\Gamma\mathbf{A}_1'$$

The second derivative is given by $(\Gamma + \Sigma_v^T)^{-1}$ and by the positive definiteness of this matrix the second order condition for minimisation of the mean square error is satisfied. This result is of some interest because it may be viewed as analogous to similar results in other literatures. Note first the similarity between this result and the standard signal extraction result which says that the optimal filter for distinguishing between signal and noise is equal to the autocovariance of the signal (Γ in our case) divided by the sum of the signal and noise autocovariances. Note that if $p = 1$ then $\tilde{\mathbf{A}}_1^2 \leq \mathbf{A}_1^2$. By the positive-definiteness of Γ and Σ_v^T one might conjecture that this result would extend to the multivariate case where $\tilde{\mathbf{A}}_1\tilde{\mathbf{A}}_1' \leq \mathbf{A}_1\mathbf{A}_1'$. Unfortunately, this is not the case. Although it is likely that this result will hold it is by no means certain. Another interesting corollary of the above result is that the method applies equally to measurement error. The only assumption we have made is that there exist an error in the measurement of the true data whose covariance is given by Σ_v^T. This clearly covers cases of data measurement error.

The above analysis concentrated on one-step ahead forecasts. The general problem of n-step ahead forecasting can be dealt with similarly by minimising the sum of the 1 to n-step ahead forecast errors with respect to a suitably defined set of coefficients $\tilde{A}$ just as we did above. We analyse this case in what follows: We want to minimise the variance of the forecast errors of the 1-step to n-step ahead forecasts. As we need to minimise a scalar function we choose to minimise the trace of the forecast error variance-covariance matrix of the 1 to n step forecasts. We assume for simplicity that $p > n$. If this is not case it can always be made the case by increasing p. Using the previous notation we know that

$$y_{T+n}^* = A^n y_T^* + A^{n-1}\epsilon_T + \ldots + \epsilon_{T+n}$$

So

$$y_{T+n,n}^* = (y_{T+1}^*, \ldots, y_{T+n}^*) = A^{(n)} y_T^* + A^{(n-1)}\epsilon_T + \ldots + \epsilon_{T+n,n}$$

where $A^{(n)}$ denote the first n rows of A^n and $\epsilon_{T+n,n}$ is a vector of the first n of the vector ϵ_{T+n}. So the forecast error is given by

$$y^*_{T+n,n} - \hat{y}_{T+n,n} = A^{(n)} y^*_T + A^{(n-1)} \epsilon_T + \ldots + \epsilon_{T+n,n} - \tilde{A} y^*_T - \tilde{A} v_T$$

The part of the variance of the forecast error, depending on $\tilde{A}$, which is relevant for the minimisation problem, is given as before by

$$\tilde{A}(\Gamma + \Sigma^T_v)\tilde{A}' - 2\tilde{A}\Gamma A^{(n)\prime}$$

Differentiating and noting that the derivative of the trace of the above matrix is the trace of the derivative gives

$$tr((\Gamma + \Sigma^T_v)\tilde{A}' - \Gamma A^{(n)\prime}) = 0$$

If the matrix is equal to zero then the trace is equal to zero and so

$$(\Gamma + \Sigma^T_v)\tilde{A}' - \Gamma A^{(n)\prime} = 0$$

or

$$\tilde{A}^{opt\ '} = (\Gamma + \Sigma^T_v)^{-1}\Gamma A^{(n)\prime}$$

Clearly the method we suggest is optimal in terms of mean square forecasting error conditional on being restricted to use p periods of past data, where $p = T$ is a possibility. It is therefore equivalent to using the Kalman filter on a state space model[(9)] once $p = T$. Nevertheless, the method we suggest may have advantages over the Kalman filter in many cases. Firstly, the method we suggest is transparent and easy to interpret structurally. For example, one can say something about the coefficients entering the regression and how they change when revisions occur. It is also possible to carry out inference on the new coefficients. We can obtain the standard errors of the modified coefficients from the standard errors of the original coefficients. So in forecasting one can say something about the importance (weight) of given variables and the statistical significance of those weights. From a practical point of view where a large model with many equations is being used for forecasting, and one which must bear the weight of economic story-telling, one may want to fix the coefficients for a few periods and not reestimate the whole model. Our method has some advantages over the Kalman Filter in uses of this sort, since it just uses the same coefficients rather than applying a full Kalman filter every period. Finally, the method we have is rather nonparametric as far as variances for the revision error are concerned. We have a $Tx1$ vector of errors at time T. In the most general case, these errros can have any TxT covariance matrix that represents all possibilities for how the variance of measurement error varies by

(9) For more details on the state space representation of the case we consider saee **?**.

vintage, over time, (and, in a multivariate setting, across variables). In other words, our procedure allows for time variation in the covariances, heteroscedasticity and serial correlation. The state space cannot easily attain that sort of generality. In fact a standard state space imposes rather strict forms of covariance to the errors that are unappealing in the context we are envisaging. These can be relaxed with great difficulty only and by experienced state space modellers. Some of these advantages may be come clearer with an empirical illustration.

5 Empirical illustration

We apply the general method of optimising a forecast model to an investment forecasting equation similar to those used in a number of macroeconometric models for the UK economy. A stylised fact from the data revision analysis in the UK is that some of the investment series are among the most heavily revised series in the national accounts, see, for example **?**. This is one motivation for considering investment equations. The general equation we consider is given by

$$\Delta i_t = a_0 + \sum_{i=1}^{p} a_i \Delta i_{t-i} + a_{p+1} \Delta g_{t-1} + e_t$$

where i_t is the (log of) business investment and g_t is the (log of) GDP at market prices. Many equations of this general form include an error correction term. These terms however, usually include variables such as the capital stock and the cost of capital. For our analysis, these variables suffer from two key problems: they are difficult to measure in the first place; and no satisfactory analysis of the properties of the revisions in these series exists. Further, there is significant evidence to indicate that error correction terms may not be very helpful in a forecasting context. Evidence presented by **?** demonstrates that the forecasting performance of VAR models may be better than that of error correction models over the short forecasting horizons which concern us. Only over long horizons are error correction models shown to have an advantage. **?** cast doubt on the notion that error correction models are better forecasting tools even at long horizons, at least with respect to the standard root mean square forecasting error criterion. They also argue that although unit roots are estimated consistently, modelling nonstationary series in (log) levels is likely to produce forecasts which are suboptimal in finite samples relative to a procedure that imposes unit roots, such as differencing, a phenomenon exacerbated by small sample estimation bias.

We use real time data from 1975Q1-1995Q2. We use the revision data available to provide estimates of the revision error variances. We assume that revisions do not occur in general after 24

revision rounds. More details on the estimation of the data revision variances are given in **?**. We want to investigate the out-of-sample performance of the above equation. We consider eight variants of it. Four variants do not include GDP in the equation and four do. The four variants reflect the number of lags of investment considered which varies from 1 to 4. We assume that the revision error becomes smaller and eventually disappears after 24 rounds of revisions.

We compare the forecasting performance of the optimal and standard parameter estimates. The out-of-sample forecast evaluation exercise is carried out as follows. Starting at 1985Q2 the model is estimated over the period 1975Q1-1979Q1 and investment at 1985Q3 is forecast. The reason for not using the period 1979Q2-1985Q1 data for estimating the coefficients is to ensure (within the assumptions of the experiment) that the original parameter estimate reflects the true parameter rather than be contaminated by revision errors in the data. We continue producing investment forecasts until 1995Q2. So the whole forecast evaluation period is 1985Q1-1995Q2 (10 years). The reason we stop at 1995Q2 is because we need to use the most recently available data for the evaluation period as proxies for the true data (uncontaminated by noise).

We look at the RMSE ratios of the forecasts coming from optimal and standard parameter estimates and we also look at the Diebold-Mariano tests (see **?**) looking at the null hypothesis that the two forecast are equally good in terms of RMSE[(10)] . Results are also considered for the two five year subperiods within the whole evaluation period. Results are presented in Table A.

Table A: MSE ratios and Diebold-Mariano tests

	Whole period		First subperiod		Second subperiod	
Model	MSE Ratio	D-M Test	MSE Ratio	D-M Test	MSE Ratio	D-M Test
AR(1)	0.8690	2.4557*	0.9185	0.8346	0.8353	2.7867*
AR(2)	0.7837	3.6246*	0.7710	2.1530*	0.7945	3.1363*
AR(3)	0.7809	3.3514*	0.7678	1.9543	0.7919	2.9845*
AR(4)	0.8466	2.2081*	0.9233	0.5510	0.8003	2.8204*
ARDL(1)	0.8497	2.7007*	0.8909	1.1836	0.8158	2.7332*
ARDL(2)	0.8275	3.5666*	0.8631	1.8871	0.7881	3.3172*
ARDL(3)	0.8595	3.0335*	0.9138	1.2267	0.8033	3.3275*
ARDL(4)	0.9320	0.9047	1.1211	-1.3071	0.7818	3.0568*

* denotes significance at the 1% level

Clearly the forecasts using the optimal coefficients outperform the standard forecasts for all

(10) Positive test statistics indicate superiority of the forecasts based on the optimal forecasting coefficients and vice versa.

models for the whole period and the second subperiod. In all but one model they outperform the standard forecast in the first subperiod as well. Even in that model, this result is not statistically significant according to the Diebold-Mariano statistic. On the other the Diebold Mariano statistics indicate statistically significant superiority (at the 1% significance level) in the whole period and the second subperiod.

6 Summary and conclusion

A brief summary of our analysis can be given as follows:

- Section 2.1 assumed a data generating process $y_t^* = ay_{t-1}^* + e_t$ and a measurement equation $y_t = y_t^* + v_t$ with homoskedastic errors, v_t. We showed that the unconditional mean of the process forms a better forecast of the true data than any forecast $\hat{y}_{T+1}^{(n)} = a^{n+1}y_{T-n}$ when the variance of the measurement errors is greater than the variance of the true data.
- Section 2.2 generalised the analysis to the case in which the variance of v_t increases with t reflecting the fact that more recent data are measured less reliably. In that case we found that it can be optimal to forecast using $\hat{y}_{T+1}^{(n)} = a^{n+1}y_{T-n}$ for some $n \geq 1$: it is better not to use the most recent data in constructing the forecast.
- Section 3 generalised the problem to allow the forecaster to chose the parameter of the forecasting model as well as the forecasting lag. In that case we found that it can be optimal to forecast using $\hat{y}_{T+1}^{(n)} = \tilde{a}^{n+1}y_{T-n}$ for some $n \geq 1$: it is better to use less recent data and a parameter $\tilde{a}$ that differs from the true model parameter a.
- Section 4 generalised the problem further to allow the forecaster to choose a forecasting model from a general linear specification. Specifically, we allowed the forecaster to choose the coefficients on a general AR process that minimise the mean-squared error of the forecast. We found that it can be optimal to forecast with an AR model of higher order than the true data generating process.
- Section 5 applied the analysis of Section 4 to an equation estimated on UK investment data. We found that the forecasts using the optimal forecast model outperformed forecasts from the standard model when compared using the Diebold-Mariano test.

So in this paper we have explored the effects of data revision on forecasting models. We have shown that in the presence of data revisions it is possible that forecasting with older data may provide superior forecasts in terms of mean square error compared to forecasts which use the most recent data. This conclusion is not affected even if we allow for adjustments in the parameters of the dynamic model to optimise the forecast in terms of mean square error. Finally, we have provided a general method of determining the optimal forecasting model in the presence of data measurement and revision errors with known covariance structure.