Chapter 11

FORECASTING WITH TRENDING DATA

GRAHAM ELLIOTT
University of California

Contents

Handbook of Economic Forecasting, Volume 1
Edited by Graham Elliott, Clive W.J. Granger and Allan Timmermann

DOI: 10.1016/S1574-0706(05)01011-6

Abstract

This chapter examines the problems of dealing with trending type data when there is uncertainty over whether or not we really have unit roots in the data. This uncertainty is practical – for many macroeconomic and financial variables theory does not imply a unit root in the data however unit root tests fail to reject. This means that there may be a unit root or roots close to the unit circle. We first examine the differences between results using stationary predictors and nonstationary or near nonstationary predictors. Unconditionally, the contribution of parameter estimation error to expected loss is of the same order for stationary and nonstationary variables despite the faster convergence of the parameter estimates. However expected losses depend on true parameter values.

We then review univariate and multivariate forecasting in a framework where there is uncertainty over the trend. In univariate models we examine trade-offs between estimators in the short and long run. Estimation of parameters for most models dominates imposing a unit root. It is for these models that the effects of nuisance parameters in the models is clearest. For multivariate models we examine forecasting from cointegrating models as well as examine the effects of erroneously assuming cointegration. It is shown that inconclusive theoretical implications arise from the dependence of forecast performance on nuisance parameters. Depending on these nuisance parameters imposing cointegration can be more or less useful for different horizons. The problem of forecasting variables with trending regressors – for example, forecasting stock returns with the dividend–price ratio – is evaluated analytically. The literature on distortion in inference in such models is reviewed. Finally, forecast evaluation for these problems is discussed.

Keywords

unit root, cointegration, long run forecasts, local to unity

JEL classification: C13, C22, C32, C53

1. Introduction

In the seminal paper Granger (1966) showed that the majority of macroeconomic variables have a typical spectral shape dominated by a peak at low frequencies. From a time domain view this means that there is some relatively long run information in the current level of a variable, or alternately stated that there is some sort of 'trending' behavior in macroeconomic (and many financial) data that must be taken account of when modelling these variables.

The flip side of this finding is that there is exploitable information for forecasting, today's levels having a large amount of predictive power as to future levels of these variables. The difficulty that arises is being precise about what this trending behavior exactly is. By virtue of trends being slowly evolving by definition, in explaining the long run movements of the data there is simply not a lot of information in any dataset as to exactly how to specify this trend, nor is there a large amount of information available in any dataset for being able to distinguish between different models of the trend.

This chapter reviews the approaches to this problem in the econometric forecasting literature. In particular we examine attempts to evaluate the importance or lack thereof of particular assumptions on the nature of the trend. Intuitively we expect that the forecast horizon will be important. For longer horizons the long run behavior of the variable will become more important, which can be seen analytically. For the most part, the typical approach to the trending problem in practice has been to follow the Box and Jenkins (1970) approach of differencing the data, which amounts to the modelling of the apparent low frequency peak in the spectrum as being a zero frequency phenomenon. Thus the majority of the work has been in considering the imposition of unit roots at various parts of the model. We will follow this approach, examining the effects of such assumptions.

Since reasonable alternative specifications must be 'close' to models with unit roots, it follows directly to concern ourselves with models that are close on some metric to the unit root model. The relevant metric is the ability of tests to distinguish between the models of the trend – if tests can easily distinguish the models then there is no uncertainty over the form of the model and hence no trade-off to consider. However the set of models for this is extremely large, and for most of the models little analytic work has been done. To this end we concentrate on linear models with near unit roots. We exclude breaks, which are covered in Chapter 12 by Clements and Hendry in this Handbook. Also excluded are nonlinear persistent models, such as threshold models, smooth transition autoregressive models. Finally, more recently a literature has developed on fractional differencing, providing an alternative model to the near unit root model through the addition of a greater range of dynamic behavior. We do not consider these models either as the literature on forecasting with such models is still in early development.

Throughout, we are motivated by some general 'stylized' facts that accompany the professions experience with forecasting macroeconomic and financial variables. The first is the phenomenon of our inability in many cases to do better than the 'unit root

forecast', i.e. our inability to say much more in forecasting a future outcome than giving today's value. This most notoriously arises in foreign exchange rates [the seminal paper is Meese and Rogoff (1983)] where changes in the exchange rate have not been easily forecast except at quite distant horizons. In multivariate situations as well imposition of unit roots (or the imposition of near unit roots such as in the Litterman vector autoregressions (VARs)) tend to perform better than models estimated in levels. The second is that for many difficult to forecast variables, such as the exchange rate or stock returns, predictors that appear to be useful tend to display trending behavior and also seem to result in unstable forecasting rules. The third is that despite the promise that cointegration would result in much better forecasts, evidence is decidedly mixed and Monte Carlo evidence is ambiguous.

We first consider the differences and similarities of including nonstationary (or near nonstationary) covariates in the forecasting model. This is undertaken in the next section. Many of the issues are well known from the literature on estimation of these models, and the results for forecasting follow directly. Considering the average forecasting behavior over many replications of the data, which is relevant for understanding the output of Monte Carlo studies, we show that inclusion of trending data has a similar order effect in terms of estimation error as including stationary series, despite the faster rate of convergence of the coefficients. Unlike the stationary case, however, the effect depends on the true value of the coefficients rather than being uniform across the parameter space.

The third section focusses on the univariate forecasting problem. It is in this, the simplest of models, that the effects of the various nuisance parameters that arise can be most easily examined. It is also the easiest model in which to examine the effect of the forecast horizon. The section also discusses the ideas behind conditional versus unconditional (on past data) approaches and the issues that arise.

Given the general lack of discomfort the profession has with imposing unit roots, cointegration becomes an important concept for multivariate models. We analyze the features of taking cointegration into account when forecasting in section three. In particular we seek to explain the disparate findings in both Monte Carlo studies and with using real data. Different studies have suggested different roles for the knowledge of cointegration at different frequencies, results that can be explained by the nuisance parameters of the models chosen to a large extent.

We then return to the ideas that we are unsure of the trending behavior, examining 'near' cointegrating models where either the covariates do not have an exact unit root or the cointegrating vector itself is trending. These are both theoretically and empirically common issues when it comes to using cointegrating methods and modelling multivariate models.

In section five we examine the trending 'mismatch' models where trending variables are employed to forecast variables that do not have any obvious trending behavior. This encompasses many forecasting models used in practice.

In a very brief section six we review issues revolving around forecast evaluation. This has not been a very developed subject and hence the review is short. We also briefly review other attempts at modelling trending behavior.

2. Model specification and estimation

We first develop a number of general points regarding the problem of forecasting with nonstationary or near nonstationary variables and highlight the differences and similarities in forecasting when all of the variables are stationary and when they exhibit some form of trending behavior.

Define Z_t to be deterministic terms, W_t to be variables that display trending behavior and V_t to be variables that are clearly stationary. First consider a linear forecasting regression when the variable set is limited to $\{V_t\}$. Consider the linear forecasting regression

$$y_{t+1} = \beta V_t + u_{t+1},$$

where throughout β will refer to an unknown parameter vector in keeping with the context of the discussion and $\hat{\beta}$ refers to an estimate of this unknown parameter vector using data up to time T. The expected one step ahead forecast loss from estimating this model is given by

$$EL\big(y_{T+1} - \hat{\beta}' V_T\big) = EL\big(u_{T+1} - T^{-1/2}\big\{T^{1/2}\big(\hat{\beta} - \beta\big)' V_T\big\}\big).$$

The expected loss then depends on the loss function as well as the estimator. In the case of mean-square error (MSE) and ordinary least squares (OLS) estimates (denoted by subscript OLS), this can be asymptotically approximated to a second order term as

$$E\big(y_{T+1} - \hat{\beta}'_{\mathrm{OLS}} V_T\big)^2 \approx \sigma_u^2\big(1 + mT^{-1}\big),$$

where m is the dimension of V_t. The asymptotic approximation follows from mean of the term $T\sigma_u^{-2}(\hat{\beta}_{\mathrm{OLS}} - \beta)' V_T V_T (\hat{\beta}_{\mathrm{OLS}} - \beta)$ being fairly well approximated by the mean of a χ_m^2 random variable over repeated draws of $\{y_t, V_t\}_1^{T+1}$. (If the variables V_T are lagged dependent variables the above approximation is not the best available, it is well known that in such cases the OLS coefficients have an additional small bias which is ignored here.) The first point to notice is that the term involving the estimated coefficients disappears at rate T for the MSE loss function, or more generally adds a term that disappears at rate $T^{1/2}$ inside the loss function. The second point is that this is independent of β, and hence there are no issues in thinking about the differences in 'risk' of using OLS for various possible parameterizations of the models. Third, this result is not dependent on the variance covariance matrix of the regressors. When we include nonstationary or nearly nonstationary regressors, we will see that the last two of these results disappear, however the first – against often stated intuition – remains the same.

Before we can consider the addition of trending regressors to the forecasting model, we first must define what this means. As noted in the introduction, this chapter does not explicitly examine breaks in coefficients. For the purposes of most of the chapter, we will consider nonstationary models where there is a unit root in the autoregressive representation of the variable. Nearly nonstationary models will be ones where the largest root of the autoregressive process, denoted by ρ, is 'close' to one. To be clear, we require a definition of close.

A reasonable definition of what we would mean by 'close to one' is values for ρ that are difficult to distinguish from one. Consider a situation where ρ is sufficiently far from one that standard tests for a unit root would reject always, i.e. with probability one. In such cases, there we clearly have no uncertainty over whether or not the variable is trending or not – it isn't. Further, treating variables with such little persistence as being 'stationary' does not create any great errors. The situation where we would consider that there is uncertainty over whether or not the data is trending, i.e. whether or not we can easily reject a unit root in the data, is the range of values for ρ where tests have difficulty distinguishing between this value of ρ and one. Since a larger number of observations helps us pin down this parameter more precisely, the range over ρ for which we have uncertainly shrinks as the sample size grows.

Thus we can obtain the relevant range, as a function of the number of observations, through examining the local power functions of unit root tests. Local power is obtained by these tests for ρ shrinking towards one at rate T, i.e. for local alternatives of the form $\rho = 1 - \gamma/T$ for γ fixed. We will use these local to unity asymptotics to evaluate asymptotic properties of the methods below. This makes ρ dependent on T, however we will suppress this notation. It should be understood that any model we consider has a fixed value for ρ, which will be understood for any sample size using asymptotic results for the corresponding value for γ given T.

It still remains to ascertain the relevant values for γ and hence pairs (ρ, T). It is well known that our ability to distinguish unit roots from those less than one depends on a number of factors including the initialization of the process and the specification of the deterministic terms. From Stock (1994) the relevant ranges can be read from his Figure 2 (pp. 2774–2775) for various tests and configurations of the deterministic component when initial conditions are set to zero effectively, when a mean is included the range for γ over which there is uncertainty is from zero to about $\gamma = 20$. When a time trend is included uncertainty is greater, the relevant uncertain range is from zero to about $\gamma = 30$. Larger initial conditions extend the range over γ for which tests have difficulty distinguishing the root from one [see Müller and Elliott (2003)]. For these models approximating functions of sample averages with normal distributions is not appropriate and instead these processes will be better approximated through applications of the Functional Central Limit Theorem.

Having determined what we mean by trending regressors, we can now turn to evaluating the similarities and difference with the stationary covariate models. We first split the trending and stationary covariates, as well as introduce the deterministics (as is familiar in the study of the asymptotic behavior of trending regressors when there are determin-

istic terms, these terms play a large role through altering the asymptotic behavior of the coefficients on the trending covariates). The model can be written

$$y_{t+1} = \beta_1' W_t + \beta_2' V_t + u_{1t},$$

where we recall that W_t are the trending covariates and V_t are the stationary covariates. In a linear regression the coefficients on variables with a unit root converge at the faster rate of T. [For the case of unit roots in a general regression framework, see Phillips and Durlauf (1986) and Sims, Stock and Watson (1990), the similar results for the local to unity case follow directly, see Elliott (1998).] We can write the loss from using OLS estimates of the linear model as

$$\begin{aligned} &L\big(y_{T+1} - \hat{\beta}_{1,\mathrm{OLS}}' W_T - \hat{\beta}_{2,\mathrm{OLS}}' V_T\big) \\ &\quad = L\big(u_{T+1} - T^{-1/2}\big[T\big(\hat{\beta}_{1,\mathrm{OLS}} - \beta_1\big)' T^{-1/2} W_T + T^{1/2}\big(\hat{\beta}_{2,\mathrm{OLS}} - \beta_2\big)' V_T\big]\big), \end{aligned}$$

where $T^{-1/2} W_T$ and V_T are $O_p(1)$. Notice that for the trending covariates we divide each of the trending regressors by the square root of T. But this is precisely the rate at which they diverge, and hence these too are $O_p(1)$ variables.

Now consider the three points above. First, standard intuition suggests that when we mix stationary and nonstationary (or nearly nonstationary) variables we can to some extent be less concerned with the parameter estimation on the nonstationary terms as they disappear at the faster rate of T as the sample size increases, hence they are an order of magnitude smaller than the coefficients on the stationary terms, at least asymptotically. However this is not true – the variables they multiply in the loss function grow at exactly this rate faster than the stationary covariates, so in the end they all end up making a contribution of the same order to the loss function. For MSE loss, this is that the terms disappear at rate T regardless of whether they are stationary or nonstationary (or deterministic, which was not shown here but follows by the same math).

Now consider the second and third points. The OLS coefficients $T(\hat{\beta}_{1,\mathrm{OLS}} - \beta_1)$ converge to nonstandard distributions which depend on the model through the local to unity parameter γ as well as other nuisance parameters of the model. The form depends on the specifics of the model, precise examples of this for various models will be given below. In the MSE loss case, terms such as $E[T(\hat{\beta}_{1,\mathrm{OLS}} - \beta_1)' W_T W_T' (\hat{\beta}_{1,\mathrm{OLS}} - \beta_1)]$ appear in the expected mean-square error.

Hence not only is the additional component to the expected loss when parameters are estimated now not well approximated by the number of parameters divided by T but it depends on γ through the expected value of the nonstandard term. Thus the OLS risk is now dependent on the true model, and one must think about what the true model is to evaluate what the OLS risk would be. This is in stark contrast to the stationary case. Finally, it also depends on the covariates themselves, since they also affect this nonstandard distribution and hence its expected value. The nature and dimension of any deterministic terms will additionally affect the risk through affecting this term. As is common in the nonstationary literature, whilst definitive statements can be made actual calculations will be special to the precise nature of the model and the properties of the

regressors. The upshot is that it is not true that we can ignore the effects of the trending regressors asymptotically when evaluating expected loss because of their fast rate of convergence, and that the precise effects will vary from specification to specification.

This understanding drives the approach of the following. First, we will ignore for the most part the existence and effect of 'obviously' stationary covariates in the models. The main exception is the inclusion of error correction terms, which are closely related to the nonstationary terms and become part of the story. Second, we will proceed with a number of 'canonical' models – since the results differ from specification to specification it is more informative to analyze a few standard models closely.

A final general point refers to loss functions. Numerical results for trade-offs and evaluation of the effects of different methods for dealing with the trends will obviously depend on the loss function chosen. The typical loss function chosen in this literature is that of mean-square error (MSE). If the h step ahead forecast error conditional on information available at time t is denoted $e_{t+h|t}$ this is simply $E[e_{t+h|t}^2]$. In the case of multivariate models, multivariate versions of MSE have been examined. In this case the h step ahead forecast error is a vector and the analog to univariate MSE is $E[e'_{t+h|t} K e_{t+h|t}]$ for some matrix of weights K. Notice that for each different choice of K we would have a different weighting of the forecast errors in each equation of the model and hence a different loss function, resulting in numerical evaluations of any choices over modelling to depend on K. Some authors have considered this a weakness of this loss function but clearly it is simply a feature of the reality that different loss functions necessarily lead to different outcomes precisely because they reflect different choices of what is important in the forecasting process. We will avoid this multivariate problem by simply choosing to evaluate a single equation from any multivariate problem.

There has also been some criticism of the use of the univariate MSE loss function in problems where there is a choice over whether or not the dependent variable is written in levels or differences. Consider an h step ahead forecast of y_t and assume that the forecast is conditional on information at time t. Now we can always write $y_{T+h} = y_T + \sum_{i=1}^{h} \Delta y_{t+i}$. So for any loss function, including the MSE, that is a function of the forecast errors only we have that

$$
\begin{aligned}
L(e_{t+h}) &= L(y_{t+h} - y_{t+h,t}) \\
&= L\left(y_t + \sum_{i=1}^{h} \Delta y_{t+i} - y_t + \sum_{i=1}^{h} \Delta y_{t+i,t} \right) \\
&= L\left(\sum_{i=1}^{h} (\Delta y_{t+i} - \Delta y_{t+i,t}) \right)
\end{aligned}
$$

and so the forecast error can be written equivalently in the level or the sum of differences. Thus there is no implication for the choice of the loss function when we consider

the two equivalent expressions of the forecast error.[1] We will refer to forecasting y_{T+h} and $y_{T+h} - y_T$ as being the same thing given that we will always assume that y_T is in the forecasters information set.

3. Univariate models

The simplest model in which to examine the issues, and hence the most examined model in the literature, is the univariate model. Even in this model results depend on a large variety of nuisance parameters. Consider the model

$$\begin{aligned} y_t &= \phi z_t + u_t, \quad t = 1, \ldots, T, \\ (1 - \rho L)u_t &= v_t, \quad t = 2, \ldots, T, \\ u_1 &= \xi, \end{aligned} \tag{1}$$

where z_t are strictly exogenous deterministic terms and ξ is the 'initial' condition. We will allow additional serial correlation through $v_t = c(L)\varepsilon_t$ where ε_t is a mean zero white noise term with variance σ_ε^2. The lag polynomial describing the dynamic behavior of y_t has been factored so that $\rho = 1 - \gamma/T$ corresponds to the largest root of the polynomial, and we assume that $c(L)$ is one summable.

Any result is going to depend on the specifics of the problem, i.e. results will depend on the exact model, in particular the nuisance parameters of the problem. In the literature on estimation and testing for unit roots it is well known that various nuisance parameters affect the asymptotic approximations to estimators and test statistics. There as here nuisance parameters such as the specification of the deterministic part of the model and the treatment of the initial condition affect results. The extent to which there are additional stationary dynamics in the model has a lesser effect. For the deterministic component we consider $z_t = 1$ and $z_t = (1, t)$ – the mean and time trend cases, respectively. For the initial condition we follow Müller and Elliott (2003) in modelling this term asymptotically as $\xi = \alpha\omega(2\gamma)^{-1/2}T^{1/2}$ where $\omega^2 = c(1)^2\sigma_\varepsilon^2$ and the rate $T^{1/2}$ results in this term being of the same order as the stochastic part of the model asymptotically. A choice of $\alpha = 1$ here corresponds to drawing the initial condition from its unconditional distribution.[2] Under these conditions we have

[1] Clements and Hendry (1993) and (1998, pp. 69–70) argue that the MSFE does not allow valid comparisons of forecast performance for predictions across models in levels or changes when $h > 1$. Note though that, conditional on time T dated information in both cases, they compare the levels loss of $E[y_{T+h} - y_T]^2$ with the difference loss of $E[y_{T+h} - y_{T+h-1}]^2$ which are two different objects, differing by the remaining $h - 1$ changes in y_t.

[2] It is common in Monte Carlo analysis to generate pseudo time series to be longer than the desired sample size and then drop early values in order to remove the effects of the initial condition. This, if sufficient observations are dropped, is the same as using the unconditional distribution. Notice though that α remains important – it is not possible to remove the effects of the initial condition for these models.

$$T^{-1/2}(u_{[Ts]}) \Rightarrow \omega M(s) = \begin{cases} \omega W(s) & \text{for } \gamma = 0, \\ \omega\alpha \mathrm{e}^{-\gamma s}(2\gamma)^{-1/2} + \omega \int_0^s \mathrm{e}^{-\gamma(s-\lambda)}\,\mathrm{d}W(\lambda) & \text{else,} \end{cases} \tag{2}$$

where $W(\cdot)$ is a standard univariate Brownian motion. Also note that for $\gamma > 0$,

$$\begin{aligned} E\big[M(s)\big]^2 &= \alpha^2 \mathrm{e}^{-2\gamma s}/(2\gamma) + \big(1 - \mathrm{e}^{-2\gamma s}\big)/(2\gamma) \\ &= (\alpha^2 - 1)\mathrm{e}^{-2\gamma s}/(2\gamma) + 1/(2\gamma), \end{aligned}$$

which will be used for approximating the MSE below.

If we knew that $\rho = 1$ then the variable has a unit root and forecasting would proceed using the model in first differences, following the Box and Jenkins (1970) approach. The idea that we know there is an exact unit root in a data series is not really relevant in practice. Theory rarely suggests a unit root in a data series, and even when we can obtain theoretical justification for a unit root it is typically a special case model [examples include the Hall (1978) model for consumption being a random walk, also results that suggest stock prices are random walks]. For most applications a potentially more reasonable approach both empirically and theoretically would be to consider models where $\rho \leqslant 1$ and there is uncertainty over its exact value. Thus there will be a trade-off between gains of imposing the unit root when it is close to being true and gains to estimation when we are away from this range of models.

A first step in considering how to forecast in this situation is to consider the cost of treating near unit root variables as though they have unit roots for the purposes of forecasting. To make any headway analytically we must simplify dramatically the models to show the effects. We first remove serial correlation.

In the case of the model in (1) and $c(L) = 1$,

$$\begin{aligned} y_{T+h} - y_T &= \varepsilon_{T+h} + \rho\varepsilon_{T+h-1} + \cdots + \rho^{h-1}\varepsilon_{T+1} + \big(\rho^h - 1\big)\big(y_T - \phi' z_T\big) \\ &\quad + \phi'(z_{T+h} - z_T) \\ &= \sum_{i=1}^{h} \rho^{h-i}\varepsilon_{T+i} + \big(\rho^h - 1\big)\big(y_T - \phi' z_T\big) + \phi'(z_{T+h} - z_T). \end{aligned}$$

Given that largest root ρ describes the stochastic trend in the data, it seems reasonable that the effects will depend on the forecast horizon. In the short run mistakes in estimating the trend will differ greatly from when we forecast further into the future. As this is the case, we will take these two sets of horizons separately.

A number of papers have examined these models analytically with reference to forecasting behavior. Magnus and Pesaran (1989) examine the model (1) where $z_t = 1$ with normal errors and $c(1) = 1$ and establish the exact unconditional distribution of the forecast error $y_{T+h} - y_T$ for various assumptions on the initial condition. Banerjee (2001) examines this same model for various initial values focussing on the impact of the nuisance parameters on MSE error using exact results. Some of the results given below are large sample analogs to these results. Clements and Hendry (2001) follow Sampson (1991) in examining the trade-off between models that impose the unit root

and those that do not for forecasting in both short and long horizons with the model in (1) when $z_t = (1, t)$ and $c(L) = 1$ where also their model without a unit root sets $\rho = 0$. In all but the very smallest sample sizes these models are very different in the sense described above – i.e. the models are easily distinguishable by tests – so their analytic results cover a different set of comparisons to the ones presented here. Stock (1996) examines forecasting with the models in (1) for long horizons, examining the trade-offs between imposing the unit root or not as well as characterizing the unconditional forecast errors. Kemp (1999) provides large sample analogs to the Magnus and Pesaran (1989) results for long forecast horizons.

3.1. Short horizons

Suppose that we are considering imposing a unit root when we know the root is relatively close to one. Taking the mean case $\phi = \mu$ and considering a one step ahead forecast, we have that imposing a unit root leads to the forecast y_T of y_{T+h} (where imposing the unit root in the mean model annihilates the constant term in the forecasting equation). Contrast this to the optimal forecast based on past observations, i.e. we would use as a forecast $\mu + \rho^h(y_T - \mu)$. These differ by $(\rho^h - 1)(y_T - \mu)$ and hence the difference between forecasts assuming a unit root versus using the correct model will be large if either the root is far from one or the current level of the variable is far from its mean.

One reason to conclude that the 'unit root' is hard to beat in an autoregression is that this term is likely to be small on average, so even knowing the true model is unlikely to yield economically significant gains in the forecast when the forecasting horizon is short. The main reason follows directly from the term $(\rho^h - 1)(y_T - \mu)$ – for a large effect we require that $(\rho^h - 1)$ is large but as the root ρ gets further from one the distribution of $(y_T - \mu)$ becomes more tightly distributed about zero.

We can obtain an idea of the size of these affects analytically. In the case where $z_t = 1$, the unconditional MSE loss for a h step ahead forecast where h is small relative to the sample size is given by

$$\begin{aligned} E[y_{T+h} - y_T]^2 &= E\big[\varepsilon_{T+h} + \rho\varepsilon_{T+h-1} + \cdots + \rho^{h-1}\varepsilon_{T+1} + \big(\rho^h - 1\big)(y_T - \mu)\big]^2 \\ &= E\big[\varepsilon_{T+1} + \rho\varepsilon_{T+h-1} + \cdots + \rho^{h-1}\varepsilon_{T+1}\big]^2 \\ &\quad + T^{-1}\big\{T^2\big(\rho^h - 1\big)^2\big\}E\big[T^{-1}(y_T - \mu)^2\big]. \end{aligned}$$

The first order term is due to the unpredictable future innovations. Focussing on the second order term, we can approximate the term inside the expectations by its limit and after then taking expectations this term can be approximated by

$$\sigma_\varepsilon^{-2}T^2\big(\rho^h - 1\big)^2E\big[T^{-1}(y_T - \mu)^2\big] \approx 0.5h^2\gamma\big(\alpha^2 - 1\big)\mathrm{e}^{-2\gamma} + \frac{h^2\gamma}{2}. \tag{3}$$

As γ increases, the term involving $\mathrm{e}^{-2\gamma}$ gets small fast and hence this term can be ignored. The first point to note then is that this leaves the result as basically linear in γ –

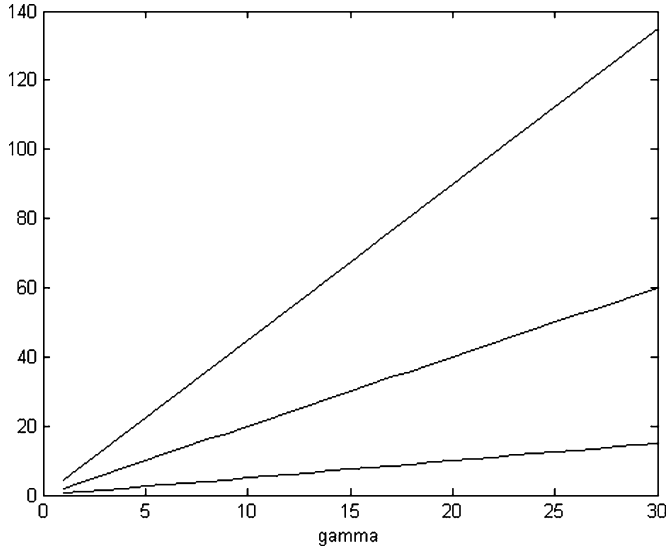

Figure 1. Evaluation of (3) for $h = 1, 2, 3$ in ascending order.

the loss as we expect is rising as the imposition of the unit root becomes less sensible and the result here shows that the effect is linear in the misspecification. The second point to note is that the slope of this linear effect is $h^2/2$, so is getting large faster and faster for any $\rho < 1$ the larger is the prediction horizon. This is also as we expect, if there is mean reversion then the further out we look the more likely it is that the variable has moved towards its mean and hence the larger the loss from giving a 'no change' forecast. The effect is increasing in h, i.e. given γ the marginal effect of a predicting an extra period ahead is $h\gamma$, which is larger the more mean reverting the data and larger the prediction horizon. The third point is that the effect of the initial condition is negligible in terms of the cost of imposing the unit root,[3] as it appears in the term multiplied by $e^{-2\gamma}$. Further, in the case where we use the unconditional distribution for the initial condition, i.e. $\alpha = 1$, these terms drop completely. For $\alpha \neq 1$ there will be some minor effects for very small γ.

The magnitude of the effects are pictured in Figure 1. This figure graphs the effect of this extra term as a function of the local to unity parameter for $h = 1, 2, 3$ and $\alpha = 1$. Steeper curves correspond to longer forecast horizons. Consider a forecasting problem where there are 100 observations available, and suppose that the true value for ρ was 0.9. This corresponds to $\gamma = 10$. Reading off the figure (or equivalently from the expression above) this corresponds to values of this additional term of 5, 20 and 45. Dividing these by the order of the term, i.e. 100, we have that the additional loss in MSE

[3] Banerjee (2001) shows this result using exact results for the distribution under normality.

as a percentage for the unpredictable component is of the order 5%, 10% and 15% of the size of the unpredictable component, respectively (since the size of the unpredictable component of the forecast error rises almost linearly in the forecast horizon when h is small).

When we include a time trend in the model, the model with the imposed unit root has a drift. An obvious estimator of the drift is the mean of the differenced series, denoted by $\hat{\tau}$. Hence the forecast MSE when a unit root is imposed is now

$$\begin{aligned}
&E[y_{T+1} - y_T - h\hat{\tau}]^2 \\
&\quad \cong E\big[\varepsilon_{T+h} + \rho\varepsilon_{T+h-1} + \cdots + \rho^{h-1}\varepsilon_{T+1} \\
&\qquad + T^{-1/2}\big\{T\big(\rho^h - 1\big) + h\big\}(y_T - \mu - \tau T) - hT^{-1/2}u_1\big]^2 \\
&\quad = E\big[\varepsilon_{T+h} + \rho\varepsilon_{T+h-1} + \cdots + \rho^{h-1}\varepsilon_{T+1}\big]^2 \\
&\qquad + T^{-1}E\big[\big\{T\big(\rho^h - 1\big) + h\big\}^2 T^{-1/2}(y_T - \mu - \tau T) - hT^{-1/2}u_1\big]^2.
\end{aligned}$$

Again, focussing on the second part of the term we have

$$\begin{aligned}
&\sigma_\varepsilon^{-2}E\big[\big\{T\big(\rho^h - 1\big) + h\big\}^2 T^{-1/2}(y_T - \mu - \tau T) - hT^{-1/2}u_1\big]^2 \\
&\quad \approx h^2\big[(1+\gamma)^2\big\{\big(\alpha^2 - 1\big)\mathrm{e}^{-2\gamma}/(2\gamma) + 1/(2\gamma)\big\} \\
&\qquad + \alpha^2/(2\gamma) - (1+\gamma)\mathrm{e}^{-\gamma}/\gamma\big]. \qquad (4)
\end{aligned}$$

Again the first term is essentially negligible, disappearing quickly as γ departs from zero, and equals zero as in the mean case when $\alpha = 1$. The last term, multiplied by $\mathrm{e}^{-\gamma}/\gamma$ also disappears fairly rapidly as γ gets larger. Focussing then on the last line of the previous expression, we can examine issues relevant to the imposition of a unit root on the forecast. First, as γ gets large the effect on the loss is larger than that for the constant only case. There are additional effects on the cost here, which is strictly positive for all horizons and initial values. The additional term arises due to the estimation of the slope of the time trend. As in the previous case, the longer the forecast horizon the larger the cost. The marginal effect of increasing the forecast horizon is also larger. Finally, unlike the model with only a constant, here the initial condition does have an effect, not only on the above effects but also on its own through the term $\alpha^2/2\gamma$. This term is decreasing the more distant the root is from one, however will have a nonnegligible effect for very roots close to one. The results are pictured in Figure 2 for $h = 1$, 2 and 3. These differential effects are shown by reporting in Figure 2 the expected loss term for both $\alpha = 1$ (solid lines) and for $\alpha = 0$ (accompanying dashed line).

The above results were for the model without any serial correlation. The presence of serial correlation alters the effects shown above, and in general these effects are complicated for short horizon forecasts. To see what happens, consider extending the model to allow the error terms to follow an MA(1), i.e. consider $c(L) = 1 + \psi L$. In the case where there is a constant only in the equation, we have that

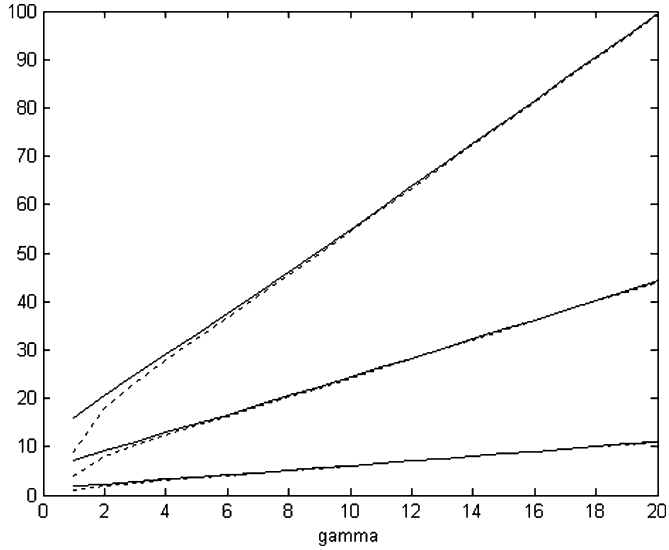

Figure 2. Evaluation of term in (4) for $h = 1, 2, 3$ in ascending order. Solid lines for $a = 1$ and dotted lines for $a = 0$.

$$y_{T+h} - y_T = \big(\varepsilon_{T+h} + (\rho+\psi)\varepsilon_{T+h-1} + \cdots + \rho^{h-2}(\rho+\psi)\varepsilon_{T+1}\big) + \big[\big(\rho^h - 1\big)(y_T - \mu) + \rho^{h-1}\psi\varepsilon_T\big],$$

where the first bracketed term is the unpredictable component and the second term in square brackets is the optimal prediction model. The need to estimate the coefficient on ε_T is not affected to the first order by the uncertainty over the value for ρ, hence this adds a term approximately equal to σ_ε^2/T to the MSE. In addition to this effect there are two other effects here – the first being that the variance of the unpredictable part changes and the second being that the unconditional variance of the term $(\rho^h - 1)(y_T - \mu)$ changes. Through the usual calculations and noting that now $T^{-1/2}y_{[T.]} \Rightarrow (1+\psi)^2\sigma_\varepsilon^2 M(\cdot)$ we have the expression for the MSE

$$E[y_{T+h} - y_T]^2 \simeq \sigma_\varepsilon^2\Bigg(1 + (h-1)(1+\psi)^2 + T^{-1}\bigg[(1+\psi)^2\bigg\{0.5h^2\gamma\big(\alpha^2 - 1\big)\mathrm{e}^{-2\gamma} + \frac{h^2\gamma}{2}\bigg\} + 1\bigg]\Bigg).$$

A few points can be made using this expression. First, when $h = 1$ there is an additional wedge in the size of the effect of not knowing the root relative to the variance of the unpredictable error. This wedge is $(1+\psi)^2$ and comes through the difference between the variance of ε_t and the long run variance of $(1 - \rho L)y_t$, which are no longer the same in the model with serial correlation. We can see how various values for ψ will then change the cost of imposing the unit root. For $\psi < 0$ the MA component reduces

the variation in the level of y_T, and imposing the root is less costly in this situation. Mathematically this comes through $(1+\psi)^2 < 1$. Positive MA terms exacerbate the cost. As h gets larger the differential scaling effect becomes relatively smaller, and the trade-off becomes similar to the results given earlier with the replacement of the variance of the shocks with the long run variance.

The costs of imposing coefficients that are near zero to zero needs to be compared to the problems of estimating these coefficients. It is clear that for ρ very close to one that imposition of a unit root will improve forecasts, but what 'very close' means here is an empirical question, depending on the properties of the estimators themselves. There is no obvious optimal estimator for ρ in these models. The typical asymptotic optimality result when $|\rho| < 1$ for the OLS estimator for ρ, denoted $\hat{\rho}_{\mathrm{OLS}}$, arises from a comparison of its pointwise asymptotic normal distribution compared to lower bounds for other consistent asymptotic normal estimators for ρ. Given that for the sample sizes and likely values for ρ we are considering here the OLS estimator has a distribution that is not even remotely close to being normal, comparisons between estimators based on this asymptotic approximation are not going to be relevant. Because of this, many potential estimators can be suggested and have been suggested in the literature. Throughout the results here we will write $\hat{\rho}$ (and similarly for nuisance parameters) as a generic estimator.

In the case where a constant is included the forecast requires estimates for both μ and ρ. The forecast is $y_{T+h|T} = (\hat{\rho}^h - 1)(y_T - \hat{\mu})$ resulting in forecast errors equal to

$$y_{T+h} - y_{T+h|T} = \sum_{i=1}^{h} \rho^{h-i}\varepsilon_{T+i} + (\hat{\mu} - \mu)\big(\hat{\rho}^h - 1\big) + \big(\rho^h - \hat{\rho}^h\big)(y_T - \mu).$$

The term due to the estimation error can be written as

$$\begin{aligned}&(\hat{\mu} - \mu)\big(\hat{\rho}^h - 1\big) + \big(\rho^h - \hat{\rho}^h\big)(y_T - \mu)\\ &\quad = T^{-1/2}\big\{T^{-1/2}(\hat{\mu} - \mu)T\big(\hat{\rho}^h - 1\big) + T\big(\rho^h - \hat{\rho}^h\big)T^{-1/2}(y_T - \mu)\big\},\end{aligned}$$

where $T^{-1/2}(\hat{\mu} - \mu)$, $T(\hat{\rho}^h - 1)$ and $T(\rho^h - \hat{\rho}^h)$ are all $O_p(1)$ for reasonable estimators of the mean and autoregressive term. Hence, as with imposing a unit root, the additional term in the MSE will be disappearing at rate T. The precise distributions of these terms depend on the estimators employed. They are quite involved, being nonlinear functions of a Brownian motion. As such the expected value of the square of this is difficult to evaluate analytically and whilst we can write down what this expression looks like no results have yet been presented for making these results useful apart from determining the nuisance parameters that remain important asymptotically.

A very large number of different methods for estimating $\hat{\rho}^h$ and $\hat{\mu}$ have been suggested (and in the more general case estimators for the coefficients in more general dynamic models). The most commonly employed estimator is the OLS estimator, where we note that the regression of y_t on its lag and a constant results in the constant term in this regression being an estimator for $(1-\rho)\mu$. Instead of OLS, Prais and

Winsten (1954) and Cochrane and Orcutt (1949) estimators have been used. Andrews (1993), Andrews and Chen (1994), Roy and Fuller (2001) and Stock (1991) have suggested median unbiased estimators. Many researchers have considered using unit root pretests [cf. Diebold and Kilian (2000)]. We can consider any pretest as simply an estimator, $\hat{\rho}_{\mathrm{PT}}$ which is the OLS estimator for samples where the pretest rejects and equal to one otherwise. Sanchez (2002) has suggested a shrinkage estimator which can be written as a nonlinear function of the OLS estimator. In addition to this set of regressors researchers making forecasts for multiple steps ahead can choose between estimating $\hat{\rho}$ and taking the hth power or directly estimating $\hat{\rho}^h$.

In terms of the coefficients on the deterministic terms, there are also a range of estimators one could employ. From results such as in Elliott, Rothenberg and Stock (1996) for the model with y_1 normal with mean zero and variance equal to the innovation variance we have that the maximum likelihood estimators (MLE) for μ given ρ is

$$\hat{\mu} = \frac{y_1 + (1-\rho)\sum_{t=2}^{T}(1-\rho L)y_t}{1 + (T-1)(1-\rho)^2}. \tag{5}$$

Canjels and Watson (1997) examined the properties of a number of feasible GLS estimators for this model. Ng and Vogelsang (2002) suggest using this type of GLS detrending and show gains over OLS. In combination with unit root pretests they are also able to show gains from using GLS detrending for forecasting in this setting.

As noted, for any of the combinations of estimators of ρ and μ taking expectations of the asymptotic approximation is not really feasible. Instead, the typical approach in the literature has been to examine this in Monte Carlo. Monte Carlo evidence tends to suggest that GLS estimates for the deterministic components results in better forecasts that OLS, and that estimators such as the Prais–Winsten, median unbiased estimators, and pretesting have the advantage over OLS estimation of ρ. However general conclusions over which estimator is best rely on how one trades off the different performances of the methods for different values for ρ.

To see the issues, we construct Monte Carlo results for a number of the leading methods suggested. For $T = 100$ and various choices for $\gamma = T(\rho - 1)$ in an AR(1) model with standard normal errors and the initial condition drawn so $\alpha = 1$ we estimated the one step ahead forecast MSE and averaged over 40,000 replications. Reported in Figure 3 is the average of the estimated part of the term that disappears at rate T. For stationary variables we expect this to be equal to the number of parameters estimated, i.e. 2. The methods included were imposing a unit root (the upward sloping solid line), OLS estimation for both the root and mean (relatively flat dotted line), unit root pretesting using the Dickey and Fuller (1979) method with nominal size 5% (the humped dashed line) and the Sanchez shrinkage method (dots and dashes). As shown theoretically above, the imposition of a unit root, whilst sensible if very close to a unit root, has a MSE that increases linearly in the local to unity parameter and hence can accompany relatively large losses. The OLS estimation technique, whilst loss depends on the local to unity parameter, does so only a little for roots quite close to one. The trade-off between imposing the root at one and estimating using OLS has the imposition of the root

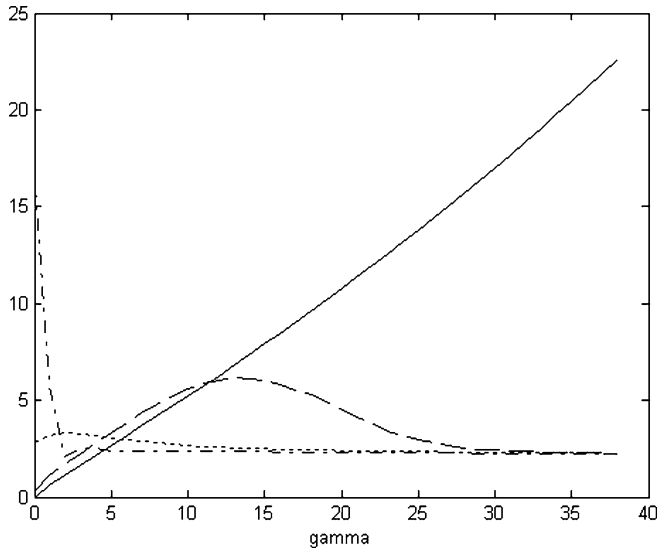

Figure 3. Relative effects of various estimated models in the mean case. The approaches are to impose a unit root (solid line), OLS (short dashes), DF pretest (long dashes) and Sanchez shrinkage (short and long dashes).

better only for $\gamma < 6$, i.e. for one hundred observations this is for roots of 0.94 or above. The pretest method works well at the 'ends', i.e. the low probability of rejecting a unit root at small values for γ means that it does well for such small values, imposing the truth or near to it, whilst because power eventually gets large it does as well as the OLS estimator for roots far from one. However the cost is at intermediate values – here the increase in average MSE is large as the power of the test is low. The Sanchez method does not do well for roots close to one, however does well away from one. Each method then embodies a different trade-off.

Apart from a rescaling of the y-axis, the results for h set to values greater than one but still small relative to the sample size result in almost identical pictures to that in Figure 3. For any moderate value for h the trade-offs occurs at the same local alternative.

Notice that any choice over which of the method to use in practice requires a weighting over the possible models, since no method uniformly dominates any other over the relevant parameter range. The commonly used 'differences' model of imposing the unit root cannot be beaten at $\gamma = 0$. Any pretest method to try and obtain the best of both worlds cannot possibly outperform the models it chooses between regardless of power if it controls size when $\gamma = 0$ as it will not choose this model with probability one and hence be inferior to imposing the unit root.

When a time trend is included the trade-off between the measures remains similar to that of the mean case qualitatively however the numbers differ. The results for the same experiment as in the mean case with $\alpha = 0$ are given in Figure 4 for the root imposed to one using the forecasting model $y_{T|T+1} = y_T + \hat{\tau}$, the model estimated by OLS and also

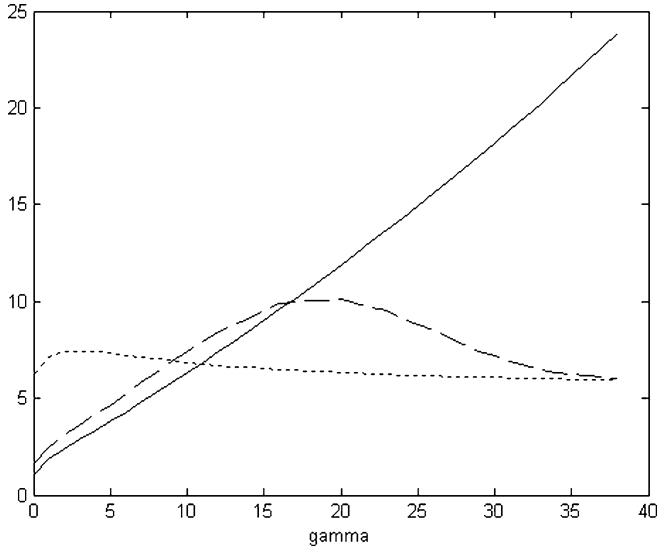

Figure 4. Relative effects of the imposed unit root (solid upward sloping line), OLS (short light dashes) and DF pretest (heavy dashes).

a hybrid approach using Dickey and Fuller t statistic pretesting with nominal size equal to 5%. As in the mean case, the use of OLS to estimate the forecasting model results in a relatively flat curve – the costs as a function of γ are varying but not much. Imposing the unit root on the forecasting model still requires that the drift term be estimated, so loss is not exactly zero at $\gamma = 0$ as in the mean case where no parameters are estimated. The value for γ for which estimation by OLS results in a lower MSE is larger than in the mean case. Here imposition of the root to zero performs better when $\gamma < 11$, so for $T = 100$ this is values for ρ of 0.9 or larger. The use of a pretest is also qualitatively similar to the mean case, however as might be expected the points where pretesting outperforms running the model in differences does differ. Here the value for which this is better is a value for γ of over 17 or so. The results presented here are close to their asymptotic counterparts, so these implications based on γ should extend relatively well to other sample sizes. Diebold and Kilian (2000) examine the trade-offs for this model in Monte Carlos for a number of choices of T and ρ. They note that for larger T the root needs to be closer to one for pretesting to dominate estimation of the model by OLS (their L model), which accords with the result here that this cutoff value is roughly a constant local alternative γ in h not too large. The value of pretesting – i.e. the models for which it helps – shrinks as T gets large. They also notice the 'ridge' where for near alternatives estimation dominates pretesting, however dismiss this as a small sample phenomenon. However asymptotically this region remains, there will be an interval for γ and hence ρ for which this is true for all sample sizes.

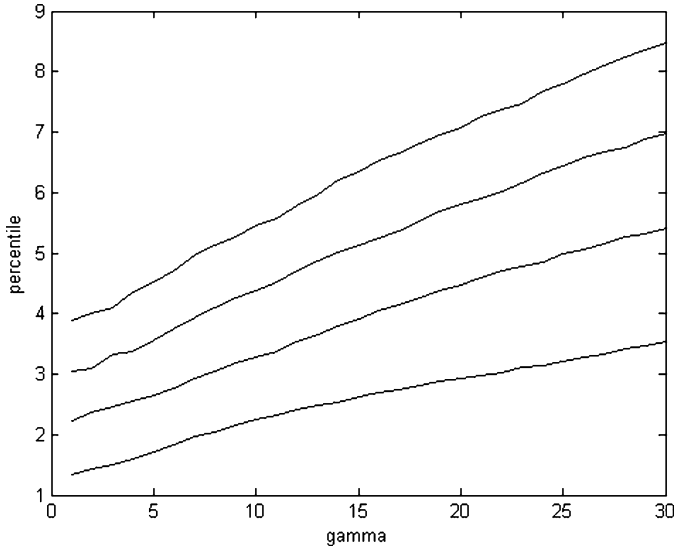

Figure 5. Percentiles of difference between OLS and Random Walk forecasts with $z_t = 1$, $h = 1$. Percentiles are for 20, 10, 5 and 2.5% in ascending order.

The 'value' of forecasts based on a unit root also is heightened by the corollary to the small size of the loss, namely that forecasts based on known parameters and forecasts based on imposing the unit root are highly correlated and hence their mistakes look very similar. We can evaluate the average size of the difference in the forecasts of the OLS and unit root models. In the case of no serial correlation the difference in h step ahead forecasts for the model with a mean is given by $(\hat{\rho}^h - 1)(y_T - \hat{\mu})$. Unconditionally this is symmetric around zero – whilst the first term pulls the estimated forecast towards the estimated mean the estimate of the mean ensures asymptotically that for every time this results in an underforecast when y_T is above its estimated mean there will be an equivalent situation where y_T is below its estimated mean. We can examine the percentiles of the limit result to evaluate the likely size of the differences between the forecasts for any (σ, T) pair. The term can be evaluated using a Monte Carlo experiment, the results for $h = 1$ and $h = 4$ are given in Figures 5 and 6, respectively, as a function of γ. To read the figures, note that the chance that the difference in forecasts scaled by multiplying by σ and dividing by $\sqrt{T}$ is between given percentiles is equal to the values given on the figure. Thus the difference between OLS and random walk one step ahead forecasts based on 100 observations when $\rho = 0.9$ has a 20% chance of being more than $2.4/\sqrt{100}$ or about one quarter of a standard deviation of the residual. Thus there is a sixty percent chance that the two forecasts differ by less than a quarter of a standard deviation of the shock in either direction. The effects are of course larger when $h = 4$, since there are more periods for which the two forecasts have time to diverge. However

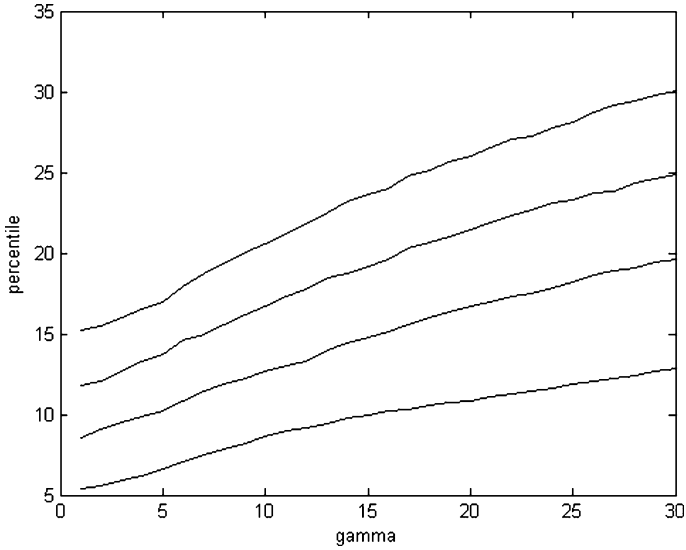

Figure 6. Percentiles of difference between OLS and Random Walk forecasts with $z_t = 1$, $h = 4$. Percentiles are for 20, 10, 5 and 2.5% in ascending order.

the difference is roughly h times as large, thus is of the same order of magnitude as the variance of the unpredictable component for a h step ahead forecast.

The above results present comparisons based on unconditional expected loss, as is typical in this literature. Such unconditional results are relevant for describing the outcomes of the typical Monte Carlo results in the literature, and may be relevant in describing a best procedure over many datasets, however may be less reasonable for those trying to choose a particular forecast model for a particular forecasting situation. For example, it is known that regardless of ρ the confidence interval for the forecast error in the unconditional case is in the case of normal innovations itself exactly normal [Magnus and Pesaran (1989)]. However this result arises from the normality of $y_T - \phi' z_T$ and the fact that the forecast error is an even function of the data. Alternatively put, the final observation $y_T - \phi' z_T$ is normally distributed, and this is weighted by values for the forecast model that are symmetrically distributed around zero so for every negative value there is a positive value. Hence overall we obtain a wide normal distribution. Phillips (1979) suggested conditioning on the observed y_T presented a method for constructing confidence intervals that condition on this final value of the data for the stationary case. Even in the simplest stationary case these confidence intervals are quite skewed and very different from the unconditional intervals. No results are available for the models considered here.

In practice we typically do not know $y_T - \phi' z_T$ since we do not know ϕ. For the best estimates for ϕ we have that $T^{-1/2}(y_T - \hat{\phi}' z_T)$ converges to a random variable and hence we cannot even consistently estimate this distance. But the sample is not completely

uninformative of this distance, even though we have seen that the deviation of y_T from its mean impacts the cost of imposing a unit root. By extension it also matters in terms of evaluating which estimation procedure might be the one that minimizes loss conditional on the information in the sample regarding this distance. From a classical perspective, the literature has not attempted to use this information to construct a better forecast method. The Bayesian methods discussed in Chapter 1 by Geweke and Whiteman in this Handbook consider general versions of these models.

3.2. Long run forecasts

The issue of unit roots and cointegration has increasing relevance the further ahead we look in our forecasting problem. Intuitively we expect that 'getting the trend correct' will be more important the longer the forecast horizon. The problem of using lagged levels to predict changes at short horizons can be seen as one of an unbalanced regression – trying to predict a stationary change with a near nonstationary variable. At longer horizons this is not the case. One way to see mathematically that this is true is to consider the forecast h steps ahead in its telescoped form, i.e. through writing $y_{T+h} - y_T = \sum_{i=1}^{h} \Delta y_{T+i}$. For variables with behavior close to or equal to those of a unit root process, their change is close to a stationary variable. Hence if we let h get large, then the change we are going to forecast acts similarly to a partial sum of stationary variables, i.e. like an $I(1)$ process, and hence variables such as the current level of the variable that themselves resemble $I(1)$ processes may well explain their movement and hence be useful in forecasting for long horizons.

As earlier, in the case of an AR(1) model

$$y_{T+h} - y_T = \sum_{i=1}^{h} \rho^{h-i} \varepsilon_{T+i} + \left(\rho^h - 1\right)\left(y_T - \phi' z_T\right).$$

Before we saw that if we let h be fixed and let the sample size get large then the second term is overwhelmed by the first, effectively $(\rho^h - 1)$ becomes small as $(y_T - \mu)$ gets large, the overall effect being that the second term gets small whilst the unforecastable component is constant in size. It was this effect that picked up the intuition that getting the trend correct for short run forecasting is not so important. To approximate results for long run forecasting, consider allowing h get large as the sample size gets large, or more precisely let $h = [T\lambda]$ so the forecast horizon gets large at the same rate as the sample size. The parameter λ is fixed and is the ratio of the forecast horizon to the sample size. This approach to long run forecasting has been examined in a more general setup by Stock (1996) and Phillips (1998). Kemp (1999) and Turner (2004) examine the special univariate case discussed here.

For such a thought experiment, the first term $\sum_{i=1}^{h} \rho^{h-i} \varepsilon_{T+i} = \sum_{i=1}^{[T\lambda]} \rho^{[T\lambda]-i} \varepsilon_{T+i}$ is a partial sum and hence gets large as the sample size gets large. Further, since we have $\rho^h = (1 + \gamma/T)^{[T\lambda]} \approx \mathrm{e}^{\gamma\lambda}$ then $(\rho^h - 1)$ no longer becomes small and both terms have the same order asymptotically. More formally we have for $\rho = 1 - \gamma/T$ that in

the case of a mean included in the model

$$T^{-1/2}(y_{T+h} - y_T) = T^{-1/2}\sum_{i=1}^{h}\rho^{h-i}\varepsilon_{T+i} + \big(\rho^h - 1\big)T^{-1/2}(y_T - \mu)$$
$$\Rightarrow \sigma_\varepsilon^2\big\{W_2(\lambda) + \big(\mathrm{e}^{-\gamma\lambda} - 1\big)M(1)\big\},$$

where $W_2(\cdot)$ and $M(\cdot)$ are independent realizations of Ornstein Uhlenbeck processes where $M(\cdot)$ is defined in (2). It should be noted however that they are really independent (nonoverlapping) parts of the same process, and this expression could have been written in that form. There is no 'initial condition' effect in the first term because it necessarily starts from zero.

We can now easily consider the effect of wrongly imposing a unit root on this process in the forecasting model. The approximate scaled MSE for such an approach is given by

$$\begin{aligned} E\big[T^{-1}(y_{T+h} - y_T)^2\big] &\Rightarrow \sigma_\varepsilon^2 E\big\{W_2(\lambda) + \big(\mathrm{e}^{-\gamma\lambda} - 1\big)M(1)\big\}^2 \\ &= \frac{\sigma_\varepsilon^2}{2\gamma}\big\{\big(1 - \mathrm{e}^{-2\gamma\lambda}\big) + \big(\mathrm{e}^{-\gamma\lambda} - 1\big)^2\big(\big(\alpha^2 - 1\big)\mathrm{e}^{-2\gamma} + 1\big)\big\} \\ &= \frac{\sigma_\varepsilon^2}{2\gamma}\big\{2 - 2\mathrm{e}^{-\gamma\lambda} + \big(\alpha^2 - 1\big)\mathrm{e}^{-2\gamma}\big(\mathrm{e}^{-\gamma\lambda} - 1\big)^2\big\}. \end{aligned} \tag{6}$$

This expression can be evaluated to see the impact of different horizons and degrees of mean reversion and initial conditions. The effect of the initial condition follows directly from the equation. Since $\mathrm{e}^{-2\gamma}(\mathrm{e}^{-\gamma\lambda} - 1)^2 > 0$ then $\alpha < 1$ corresponds to a decrease the expected MSE and $\alpha > 1$ an increase. This is nothing more than the observation made for short run forecasting that if y_T is relatively close to μ then the forecast error from using the wrong value for ρ is less than if $(y_T - \mu)$ is large. The greater is α the greater the weight on initial values far from zero and hence the greater the likelihood that y_T is far from μ.

Noting that the term that arises through the term $W_2(\lambda)$ is due to the unpredictable part, here we evaluate the term in (6) relative to the size of the variance of the unforecastable component. Figure 7 examines, for $\gamma = 1, 5$ and 10 in ascending order this term for various λ along the horizontal axis. A value of 1 indicates that the additional loss from imposing the random walk is zero, the proportion above one is the additional percentage loss due to this approximation. For γ large enough the term asymptotes to 2 as $\lambda \to 1$ – this means that the approximation cost attains a maximum at a value equal to the unpredictable component. For a prediction horizon half the sample size (so $\lambda = 0.5$) the loss when $\gamma = 1$ from assuming a unit root in the construction of the forecast is roughly 25% of the size of the unpredictable component.

As in the small h case when a time trend is included we must estimate the coefficient on this term. Using again the MLE assuming a unit root, denoted $\hat{\tau}$, we have that

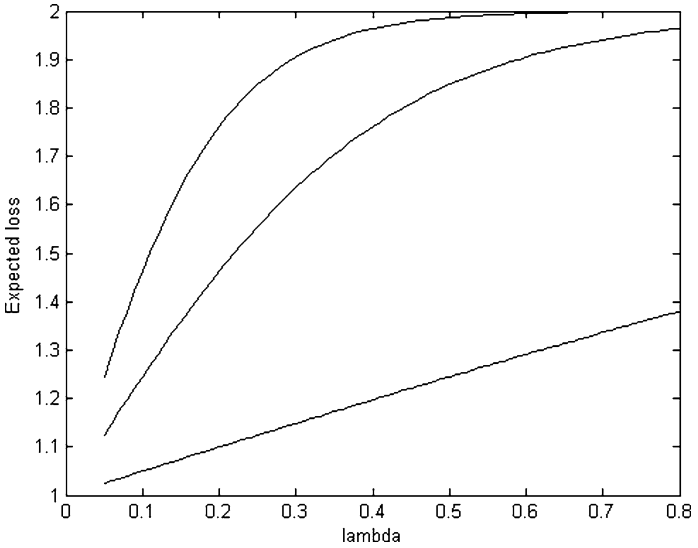

Figure 7. Ratio of MSE of unit root forecasting model to MSE of optimal forecast as a function of λ – mean case.

$$
\begin{aligned}
&T^{-1/2}(y_{T+h} - y_T - \hat{\tau}h) \\
&\quad = T^{-1/2}\sum_{i=1}^{h} \rho^{h-i}\varepsilon_{T+i} + \bigl(\rho^h - 1\bigr)T^{-1/2}\bigl(y_T - \phi' z_T\bigr) - T^{1/2}(\tau - \hat{\tau})(h/T) \\
&\quad \Rightarrow \sigma_\varepsilon^2\bigl\{W_2(\lambda) + \bigl(\mathrm{e}^{-\gamma\lambda} - 1\bigr)M(1) - \lambda\bigl(M(1) - M(0)\bigr)\bigr\}.
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
&E\bigl[T^{-1}(y_{T+h} - y_T)^2\bigr] \\
&\quad \Rightarrow \sigma_\varepsilon^2 E\bigl\{W_2(\lambda) + \bigl(\mathrm{e}^{-\gamma\lambda} - 1\bigr)M(1) - \lambda\bigl(M(1) - M(0)\bigr)\bigr\}^2 \\
&\quad = \sigma_\varepsilon^2 E\bigl\{W_2(\lambda) + \bigl(\mathrm{e}^{-\gamma\lambda} - 1 - \lambda\bigr)M(1) + \lambda M(0)\bigr\}^2 \\
&\quad = \frac{\sigma_\varepsilon^2}{2\gamma}\bigl\{\bigl(1 - \mathrm{e}^{-2\gamma\lambda}\bigr) + \bigl(\mathrm{e}^{-\gamma\lambda} - 1 - \lambda\bigr)^2\bigl(\bigl(\alpha^2 - 1\bigr)\mathrm{e}^{-2\gamma} + 1\bigr) + \lambda^2\alpha^2\bigr\} \\
&\quad = \frac{\sigma_\varepsilon^2}{2\gamma}\bigl\{1 + (1+\lambda)^2 + \lambda^2 a^2 - 2(1+\lambda)\mathrm{e}^{-\gamma\lambda} \\
&\qquad + (\alpha^2 - 1)\bigl(\bigl(1+\lambda\bigr)^2\mathrm{e}^{-2\gamma} + \mathrm{e}^{-2\gamma(1+\lambda)} - 2(1+\lambda)\mathrm{e}^{-\gamma(2+\lambda)}\bigr)\bigr\}. \qquad (7)
\end{aligned}
$$

Here as in the case of a few periods ahead the initial condition does have an effect. Indeed, for γ large enough this term is $1 + (1+\lambda)^2 + \lambda^2 a^2$ and so the level at which this tops out depends on the initial condition. Further, this limit exists only as γ gets large and differs for each λ. The effects are shown for $\gamma = 1, 5$ and 10 in Figure 8, where the

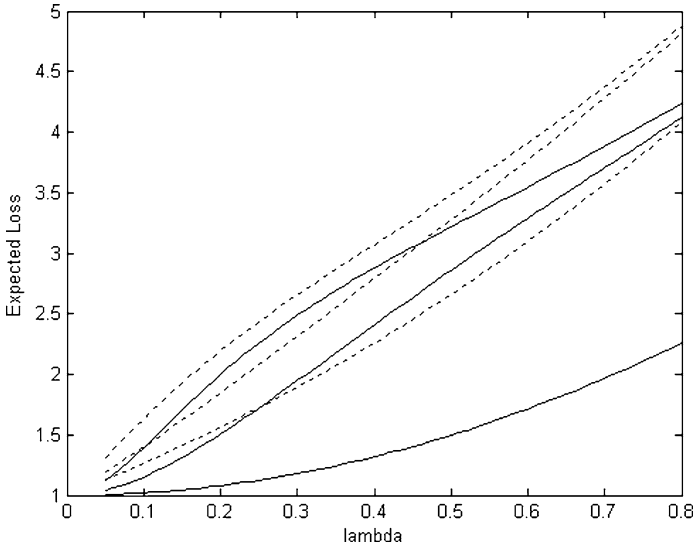

Figure 8. As per Figure 7 for Equation (7) where dashed lines are for $\alpha = 1$ and solid lines for $\alpha = 0$.

solid lines are for $\alpha = 0$ and the dashed lines for $\alpha = 1$. Curves that are higher are for larger γ. Here the effect of the unit root assumption, even though the trend coefficient is estimated and taken into account for the forecast, is much greater. The dependence of the asymptote on λ is shown to some extent through the upward sloping line for the larger values for γ. It is also noticeable that these asymptotes depend on the initial condition.

This trade-off must be matched with the effects of estimating the root and other nuisance parameters. To examine this, consider again the model without serial correlation. As before the forecast is given by

$$y_{T+h|T} = y_T + \big(\hat{\rho}^h - 1\big)\big(y_T - \hat{\phi}' z_T\big) + \hat{\phi}'(z_{T+h} - z_T).$$

In the case of a mean this yields a scaled forecast error

$$\begin{aligned} &T^{-1/2}(y_{T+h} - y_{T+h|T}) \\ &\quad = T^{-1/2}\varphi(\varepsilon_{T+h}, \ldots, \varepsilon_{T+1}) + \big(\rho^h - \hat{\rho}^h\big)T^{-1/2}(y_T - \mu) \\ &\qquad - \big(\hat{\rho}^h - 1\big)T^{-1/2}(\hat{\mu} - \mu) \\ &\quad \Rightarrow \sigma_\varepsilon^2\big(W_2(\lambda) + \big(\mathrm{e}^{\gamma\lambda} - \mathrm{e}^{\hat{\gamma}\lambda}\big)M(1) - \big(\mathrm{e}^{\hat{\gamma}\lambda} - 1\big)\varphi\big), \end{aligned}$$

where $W_2(\lambda)$ and $M(1)$ are as before, $\hat{\gamma}$ is the limit distribution for $T(\hat{\rho} - 1)$ which differs across estimators for $\hat{\rho}$ and φ is the limit distribution for $T^{-1/2}(\hat{\mu} - \mu)$ which also differs over estimators. The latter two objects are in general functions of $M(\cdot)$ and are hence correlated with each other. The precise form of this expression depends on the limit results for the estimators.

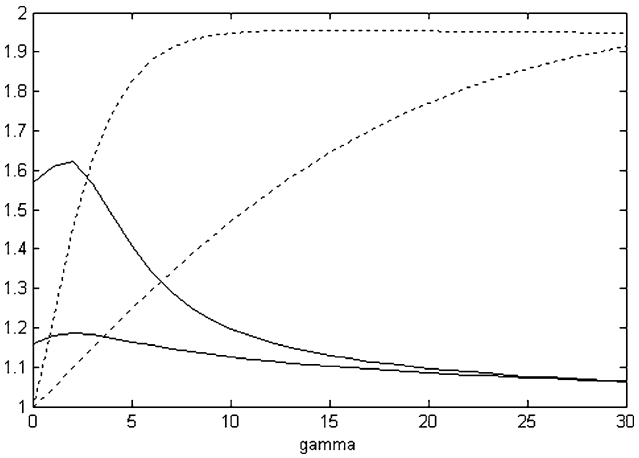

Figure 9. OLS versus imposed unit roots for the mean case at horizons $\lambda = 0.1$ and $\lambda = 0.5$. Dashed lines are the imposed unit root and solid lines for OLS.

As with the fixed horizon case, one can derive an analytic expression for the mean-square error as the mean of a complicated (i.e. nonlinear) function of Brownian motions [see Turner (2004) for the $\alpha = 0$ case] however these analytical results are difficult to evaluate. We can however evaluate this term for various initial conditions, degrees of mean reversion and forecast horizon length by Monte Carlo. Setting $T = 1000$ to approximate large sample results we report in Figure 9 the ratio of average squared loss of forecasts based on OLS estimates divided by the same object when the parameters of the model are known for various values for γ and $\lambda = 0.1$ and 0.5 with $\alpha = 0$ (solid lines, the curves closer to the x-axis are for $\lambda = 0.1$, in the case of $\alpha = 1$ the results are almost identical). Also plotted for comparison are the equivalent curves when the unit root is imposed (given by dashed lines). As for the fixed h case, for small enough γ it is better to impose the unit root. However estimation becomes a better approach on average for roots that accord with values for γ that are not very far from zero – values around $\gamma = 3$ or 4 for $\lambda = 0.5$ and 0.1, respectively. Combining this with the earlier results suggests that for values of $\gamma = 5$ or greater, which accords say with a root of 0.95 in a sample of 100 observations, that OLS should dominate the imposed unit root approach to forecasting. This is especially so for long horizon forecasting, as for large γ OLS strongly dominates imposing the root to one.

In the case of a trend this becomes $y_{T|T+h} = \hat{\rho}^h y_T + (1 - \hat{\rho}^h)\hat{\mu} + \hat{\tau}[T(1 - \hat{\rho}^h) + h]$ and the forecast error suitably scaled has the distribution

$$
\begin{aligned}
&T^{-1/2}(y_{T+h} - y_{T+h|T}) \\
&\quad = T^{-1/2}\varphi(\varepsilon_{T+h}, \ldots, \varepsilon_{T+1}) + \big(\rho^h - \hat{\rho}^h\big)T^{-1/2}\big(y_T - \phi' z_t\big) \\
&\qquad - \big(\hat{\rho}^h - 1\big)T^{-1/2}(\hat{\mu} - \mu) - T^{1/2}(\hat{\tau} - \tau)\big[\big(1 - \hat{\rho}^h\big) + \lambda\big] \\
&\quad \Rightarrow \sigma_\varepsilon^2\big(W_2(\lambda) + \big(\mathrm{e}^{\gamma\lambda} - \mathrm{e}^{\hat{\gamma}\lambda}\big)M(1) - \big(\mathrm{e}^{\hat{\gamma}\lambda} - 1\big)\varphi_1 + \big(1 + \lambda - \mathrm{e}^{\hat{\gamma}\lambda}\big)\varphi_2\big),
\end{aligned}
$$

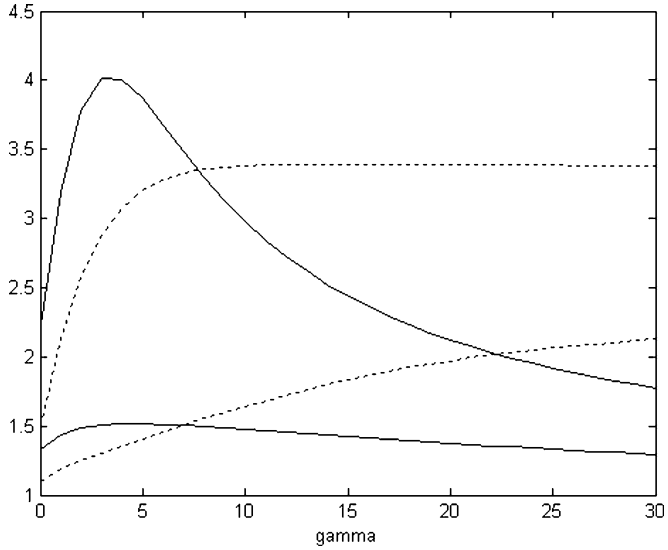

Figure 10. As per Figure 9 for the case of a mean and a trend.

where φ_1 is the limit distribution for $T^{-1/2}(\hat{\mu} - \mu)$ and φ_2 is the limit distribution for $T^{1/2}(\hat{\tau} - \tau)$. Again, the precise form of the limit result depends on the estimators.

The same Monte Carlo exercise as in Figure 9 is repeated for the case of a trend in Figure 10. Here we see that the costs of estimation when the root is very close to one is much greater, however as in the case with a mean only the trade-off is clearly strongly in favor of OLS estimation for larger roots. The point at which the curves cut – i.e. the point where OLS becomes better on average than imposing the root – is for a larger value for γ. This value is about $\gamma = 7$ for both horizons. Turner (2004) computes cutoff points for a wider array of λ.

There is little beyond Monte Carlo evidence on the issues of imposing the unit root (i.e. differencing always), estimating the root (i.e. levels always) and pretesting for a unit root (which will depend on the unit root test chosen). Diebold and Kilian (2000) provide Monte Carlo evidence using the Dickey and Fuller (1979) test as a pretest. Essentially, we have seen that the bias from estimating the root is larger the smaller the sample and the longer the horizon. This is precisely what is found in the Monte Carlo experiments. They also found little difference between imposing the unit root and pretesting for a unit root when the root is close to one, however pretesting dominates further from one. Hence they argue that pretesting always seems preferable to imposing the result. Stock (1996) more cautiously provides similar advice, suggesting pretests based on unit root tests of Elliott, Rothenberg and Stock (1996). All evidence was in terms of MSE unconditionally. Other researchers have run subsets of these Monte Carlo experiments [Clements and Hendry (1998), Campbell and Perron (1991)]. What is clear from the above calculations are two overall points. First, no method dominates every-

where, so the choice of what is best rests on the beliefs of what the model is likely to be. Second, the point at which estimation is preferred to imposition occurs for γ that are very close to zero in the sense that tests do not have great power of rejecting a unit root when estimating the root is the best practice.

Researchers have also applied the different models to data. Franses and Kleibergen (1996) examine the Nelson and Plosser (1982) data and find that imposing a unit root outperforms OLS estimation of the root in forecasting at both short and longer horizons (the longest horizons correspond to $\lambda = 0.1$). In practice, pretesting has appeared to 'work'. Stock and Watson (1999) examined many U.S. macroeconomic series and found that pretesting gave smaller out of sample MSE's on average.

4. Cointegration and short run forecasts

The above model can be extended to a vector of trending variables. Here the extreme cases of all unit roots and no unit roots are separated by the possibility that the variables may be cointegrated. The result of a series of variables being cointegrated means that there exist restrictions on the unrestricted VAR in levels of the variables, and so one would expect that imposing these restrictions will improve forecasts over not imposing them. The other implication that arises from the Granger Representation Theorem [Engle and Granger (1987)] is that the VAR in differences – which amounts to imposing too many restrictions on the model – is misspecified through the omission of the error correction term. It would seem that it would follow in a straightforward manner that the use of an error correction model will outperform both the levels and the differences models: the levels model being inferior because too many parameters are estimated and the differences model inferior because too few useful covariates are included. However the literature is divided on the usefulness of imposing cointegrating relationships on the forecasting model.

Christoffersen and Diebold (1998) examine a bivariate cointegrating model and show that the imposition of cointegration is useful at short horizons only. Engle and Yoo (1987) present a Monte Carlo for a similar model and find that a levels VAR does a little better at short horizons than the ECM model. Clements and Hendry (1995) provide general analytic results for forecast MSE in cointegrating models. An example of an empirical application using macroeconomic data is Hoffman and Rasche (1996) who find at short horizons that a VAR in differences outperforms a VECM or levels VAR for 5 of 6 series (inflation was the holdout). The latter two models were quite similar in forecast performance.

We will first investigate the 'classic' cointegrating model. By this we mean cointegrating models where it is clear that all the variables are $I(1)$ and that the cointegrating vectors are mean reverting enough that tests have probability one of detecting the correct cointegrating rank. There are a number of useful ways of writing down the cointegrating model so that the points we make are clear. The two most useful ones for our purposes

here are the error correction form (ECM) and triangular form. These are simply rotations of the same model and hence for any of one form there exists a representation in the second form. The VAR in levels can be written as

$$W_t = A(L)W_{t-1} + u_t, \tag{8}$$

where W_t is an $nx1$ vector of $I(1)$ random variables. When there exist r cointegrating vectors $\beta' W_t = c_t$ the error correction model can be written as

$$\Phi(L)\big[I(1-L) - \alpha\beta' L\big]W_t = u_t,$$

where α, β are nxr and we have factored stationary dynamics in $\Phi(L)$ so $\Phi(1)$ has roots outside the unit circle. Comparing these equations we have $(A(1) - I_n) = \Phi(1)\alpha\beta'$. In this form we can differentiate the effects of the serial correlation and the impact matrix α. Rewriting in the usual form with use of the BN decomposition we have

$$\Delta W_t = \Phi(1)\alpha c_{t-1} + B(L)\,\Delta W_{t-1} + u_t.$$

Let y_t be the first element of the vector W_t and consider the usefulness in prediction that arises from including the error correction term c_{t-1} in the forecast of y_{t+h}. First think of the one step ahead forecast, which we get from taking the first equation in this system without regard to the remaining ones. From the one step ahead forecasting problem then the value of the ECM term is simply how useful variation in c_{t-1} is in explaining Δy_t. The value for forecasting depends on the parameter in front of the term in the model, i.e. the (1, 1) element of $\Phi(1)\alpha$ and also the variation in the error correction term itself. In general the relevant parameter here can be seen to be a function of the entire set of parameters that define the stationary serial correlation properties of the model ($\Phi(1)$ which is the sum of all of the lags) and the impact parameters α. Hence even in the one step ahead problem the usefulness of the cointegrating vector term the effect will depend on almost the entire model, which provides a clue as to the inability of Monte Carlo analysis to provide hard and fast rules as to the importance of imposing the cointegration restrictions.

When we consider forecasting more steps ahead, another critical feature will be the serial correlation in the error correction term c_t. If it were white noise then clearly it would only be able to predict the one step ahead change in y_t, and would be uninformative for forecasting $y_{t+h} - y_{t+h-1}$ for $h > 1$. Since the multiple step ahead forecast $y_{t+h} - y_t$ is simply the sum of the changes $y_{t+i} - y_{t+i-1}$ from $i = 1$ to h then it will have proportionally less and less impact on the forecast as the horizon grows. When this term is serially correlated however it will be able to explain the future changes, and hence will affect the trade-off between using this term and ignoring it. In order to establish properties of the error correction term, the triangular form of the model is useful. Normalize the cointegrating vector so that the cointegrating vector $\beta' = (I_r, -\theta')$ and define the matrix

$$K = \begin{pmatrix} I_r & -\theta' \\ 0 & I_{n-r} \end{pmatrix}.$$

Note that $Kz_t = (\beta' W_t, W'_{2t})$ where W_{2t} is the last $n - r$ elements of W_t and

$$K\alpha\beta' W_{t-1} = \begin{pmatrix} \beta'\alpha \\ \alpha_2 \end{pmatrix} \beta' W_{t-1}.$$

Premultiply the model by K (so that the leading term in the polynomial is the identity matrix as per convention) and we obtain

$$K\Phi(L)K^{-1}K\big[I(1-L) - \alpha\beta' L\big]W_t = Ku_t,$$

which can be rewritten

$$K\Phi(L)K^{-1}B(L)\begin{pmatrix} \beta' W_t \\ \Delta W_{2t} \end{pmatrix} = Ku_t, \tag{9}$$

where

$$B(L) = I + \begin{pmatrix} \alpha_1 - \theta\alpha_2 - I_r & 0 \\ \alpha_2 & 0 \end{pmatrix} L.$$

This form is useful as it allows us to think about the dynamics of the cointegrating vector c_t, which as we have stated will affect the usefulness of the cointegrating vector in forecasting future values of y. The dynamics of the error correction term are driven by the value of $\alpha_1 - \theta\alpha_2 - I_r$ and the roots of $\Phi(L)$ and will be influenced by a great many parameters in the model. This provides another reason for why Monte Carlo studies have proved to be inconclusive.

In order to show the various effects, it will be necessary to simplify the models considerably. We will examine a model without 'additional' serial correlation, i.e. one for which $\Phi(L) = I$. We also will let both y_t and $W_{2t} = x_t$ be univariate. This model is still rich enough for many different effects to be shown, and has been employed to examine the usefulness of cointegration in forecasting by a number of authors. The precise form of the model in its error correction form is

$$\begin{pmatrix} \Delta y_t \\ \Delta x_t \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \begin{pmatrix} 1 & -\theta \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}. \tag{10}$$

This model under various parameterizations has been examined by Engle and Yoo (1987), Clements and Hendry (1995) and Christoffersen and Diebold (1998). In triangular form the model is

$$\begin{pmatrix} c_t \\ \Delta x_t \end{pmatrix} = \begin{pmatrix} \alpha_1 - \theta\alpha_2 + 1 & 0 \\ \alpha_2 & 0 \end{pmatrix} \begin{pmatrix} c_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} - \theta u_{2t} \\ u_{2t} \end{pmatrix}.$$

The coefficient on the error correction term in the model for y_t is simply α_1, and the serial correlation properties for the error correction term is given by $\rho_c = \alpha_1 - \theta\alpha_2 + 1 = 1 + \beta'\alpha$. A restriction of course is that this term has roots outside the unit circle, and so this restricts possible values for β and α. Further, the variance of c_t also depends on the innovations to this variable which involve the entire variance covariance matrix of u_t as well as the cointegrating parameter. It should be clear that in thinking about the effect of

various parameters on the value of including the cointegrating vector in the forecasting model controlled experiments will be difficult – changing a parameter involves a host of changes on the features of the model.

In considering h step ahead forecasts, we can recursively solve (10) to obtain

$$\begin{pmatrix} y_{T+h} - y_T \\ x_{T+h} - x_T \end{pmatrix} = \left(\sum_{i=1}^{h} \rho_c^{i-1} \right) \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} (1 \quad -\theta) \begin{pmatrix} y_T \\ x_T \end{pmatrix} + \begin{pmatrix} \tilde{u}_{1T+h} \\ \tilde{u}_{2t+h} \end{pmatrix}, \tag{11}$$

where $\tilde{u}_{1T+h}$ and $\tilde{u}_{t+h}$ are unpredictable components. The result shows that the usefulness of the cointegrating vector for the h step ahead forecast depends on both the impact parameter α_1 as well as the serial correlation in the cointegrating vector ρ_c which is a function of the cointegrating vector as well as the impact parameter in both the equations. The larger the impact parameter, all else held equal, the greater the usefulness of the cointegrating vector term in constructing the forecast. The larger the root ρ_c also the larger the impact of this term.

These results give some insight as to the usefulness of the error correction term, and show that different Monte Carlo specifications may well give conflicting results simply through examining models with differing impact parameters and serial correlation properties of the error correction term. Consider the differences between the results[4] of Engle and Yoo (1987) and Christoffersen and Diebold (1998). Both papers are making the point that the error correction term is only relevant for shorter horizons, a point to which we will return. However Engle and Yoo (1987) claim that the error correction term is quite useful at moderate horizons, whereas Christoffersen and Diebold (1998) suggest that it is only at very short horizons that the term is useful. In the former model, the impact parameter is $\alpha_y = -0.4$ and $\rho_c = 0.4$. The impact parameter is of moderate size and so is the serial correlation, and so we would expect some reasonable usefulness of the term for moderate horizons. In Christoffersen and Diebold (1998), these coefficients are $\alpha_y = -1$ and $\rho_c = 0$. The large impact parameter ensures that the error correction term is very useful at very short horizons. However employing an error correction term that is not serially correlated also ensures that it will not be useful at moderate horizons. The differences really come down to the features of the model rather than providing a general notion for all error correction terms.

This analysis abstracted from estimation error. When the parameters of the model have to be estimated then the relative value of the error correction term is diminished on average through the usual effects of estimation error. The extra wrinkle over a standard analysis of this estimation error in stationary regression is that one must estimate the cointegrating vector (one must also estimate the impact parameters 'conditional' on

[4] Both these authors use the sum of squared forecast error for both equations in their comparisons. In the case of Engle and Yoo (1987) the error correction term is also useful in forecasting in the x equation, whereas it is not for the Christoffersen and Diebold (1998) experiment. This further exacerbates the magnitudes of the differences.

the cointegrating parameter estimate, however this effect is much lower order for standard cointegrating parameter estimators). We will not examine this carefully, however a few comments can be made. First, Clements and Hendry (1995) examine the Engle and Yoo (1987) model and show that using MLE's of the cointegrating vector outperforms the OLS estimator used in the former study. Indeed, at shorter horizons Engle and Yoo (1987) found that the unrestricted VAR outperformed the ECM even though the restrictions were valid.

It is clear that given sufficient observations, the consistency of the parameter estimates in the levels VAR means that asymptotically the cointegration feature of the model will still be apparent, which is to say that in the overidentified model is asymptotically equivalent to the true error correction model. In smaller samples there is the effect of some additional estimation error, and also the problem that the added variables are trending and hence have nonstandard distributions that are not centered on zero. This is the multivariate analog of the usual bias in univariate models on the lagged level term and disappears at the same rate, i.e. at rate T. Abidir, Kaddour and Tzavaliz (1999) examine this problem. In comparing the estimation error between the levels model and the error correction model many of the trade-offs are the same. However the estimation of the cointegrating vector can be important. Stock (1987) shows that the OLS estimator of the cointegrating vector has a large bias that also disappears at rate T. Whether or not this term will on average be large depends on a nuisance parameter of the error correction model, namely the zero frequency correlation between the shocks to the error correction term and the shocks to Δx_t. When this correlation is zero, OLS is the efficient estimator of the cointegrating vector and the bias is zero (in this case the OLS estimator is asymptotically mixed normal centered on the true cointegrating vector). However in the more likely case that this is nonzero, then OLS is asymptotically inefficient and other methods[5] are required to obtain this asymptotic mixed normality centered on the true vector. In part, this explains the results of Engle and Yoo (1987). The value for this spectral correlation in their study was -0.89, quite close to the bound of one and hence OLS is likely to provide very biased estimates of the cointegrating vector. It is in just such situations that efficient cointegrating vector estimation methods are likely to be useful, Clements and Hendry (1995) show in a Monte Carlo that indeed for this model specification there are noticeable gains.

The VAR in differences can be seen to omit regressors – the error correction terms – and hence suffers from not picking up the extra possible explanatory power of the regressors. Notice that as usual here the omitted variable bias that comes along with failing to include useful regressors is the forecasters friend – this omitted variable bias is picking up at least part of the omitted effect.

The usefulness of the cointegrating relationship fades as the horizon gets large. Indeed, eventually it has an arbitrarily small contribution compared to the unexplained

[5] There are many such methods. Johansen (1991) provided an estimator that was asymptotically efficient. Many other asymptotically equivalent methods are now available, see Watson (1994) for a review.

part of y_{T+h}. This is true of any stationary covariate in forecasting the level of an $I(1)$ series. Recalling that $y_{T+h} - y_t = \sum_{i=1}^{h}(y_{t+i} - y_{t+i-1})$ then as h gets large this sum of changes in y is getting large. Eventually the short memory nature of the stationary covariate is unable to predict the future period by period changes and hence becomes a very small proportion of the difference. Both Engle and Yoo (1987) and Christoffersen and Diebold (1998) make this point. This seems to be at odds with the idea that cointegration is a 'long run' concept, and hence should have something to say far in the future.

The answer is that the error correction model does impose something on the long run behavior of the variables, that they do not depart too far from their cointegrating relation. This is pointed out in Engle and Yoo (1987), as h gets large $\beta' W_{T+h,t}$ is bounded. Note that this is the forecast of c_{T+h}, which as is implicit in the triangular relation above bounded as ρ_c is between minus one and one. This feature of the error correction model may well be important in practice even when one is looking at horizons that are large enough so that the error correction term itself has little impact on the MSE of either of the individual variables. Suppose the forecaster is forecasting both variables in the model, and is called upon to justify a story behind why the forecasts are as they are. If they are forecasting variables that are cointegrated, then it is more reasonable that a sensible story can be told if the variables are not diverging from their long run relationship by too much.

5. Near cointegrating models

In any realistic problem we certainly do not know the location of unit roots in the model, and typically arrive at the model either through assumption or pretesting to determine the number of unit roots or 'rank', where the rank refers to the rank of $A(1) - I_n$ in Equation (8) and is equal to the number of variables minus the number of distinct unit roots. In the cases where this rank is not obvious, then we are uncertain as to the exact correct model for the trending behavior of the variables and can take this into account.

For many interesting examples, a feature of cointegrating models is the strong serial correlation in the cointegrating vector, i.e. we are unclear as to whether or not the variables are indeed cointegrated. Consider the forecasting of exchange rates. The real exchange rate can be written as a function of the nominal exchange rate less a price differential between the countries. This relationship is typically treated as a cointegrating vector, however there is a large literature checking whether there is a unit root in the real exchange rate despite the lack of support for such a proposition from any reasonable theory. Hence in a cointegrating model of nominal exchange rates and price differentials this real exchange rate term may or may not appear depending on whether we think it has a unit root (and hence cannot appear, there is no cointegration) or is simply highly persistent.

Alternatively, we are often fairly sure that certain 'great ratios' in the parlance of Watson (1994) are stationary however we are unsure if the underlying variables them-

selves have unit roots. For example, the consumption income ratio is certainly bounded and does not wander around too much, however we are uncertain if there really is a unit root in income and consumption. In forecasting interest rates we are sure that the interest rate differential is stationary (although it is typically persistent), however the unit root model for an interest rate seems unlikely to be true but yet tests for the root being one often fail to reject.

Both of these possible models represent different deviations from the cointegrated model. The first suggests more unit roots in the model, the competitor model being closer to having differences everywhere. For example, in the bivariate model with one potential cointegrating vector, the nearest model to a highly persistent cointegrating vector would be a model with both variables in differences. The second suggests fewer unit roots in the model. In the bivariate case the model would be in levels. We will examine both, similar issues arise.

For the first of these models, consider Equation (9),

$$\begin{pmatrix} \beta' W_t \\ \Delta W_{2t} \end{pmatrix} = \begin{pmatrix} \beta'\alpha + I_r \\ \alpha_2 \end{pmatrix} \beta' W_{t-1} + K\Phi(L)^{-1} u_t,$$

where the largest roots of the system for the cointegrating vectors $\beta' W_t$ are determined by the value for $\beta'\alpha + I_r$. For models where there are cointegrating vectors that are have near unit roots this means that eigen values of this term are close to one. The trending behavior of the cointegrating vectors thus depend on a number of parameters of the model. Also, trending behavior of the cointegrating vectors feeds back into the process for ΔW_{2t}. In a standard framework we would require that W_{2t} be $I(1)$. However, if $\beta' W_t$ is near $I(1)$ and $\Delta W_{2t} = \alpha_2 \beta' W_t +$ noise, then we would require that $\alpha_2 = 0$ for this term to be $I(1)$. If $\alpha_2 \neq 0$, then W_{2t} will be near $I(2)$. Hence under the former case the regression becomes

$$\begin{pmatrix} \beta' W_t \\ \Delta W_t \end{pmatrix} = \begin{pmatrix} \alpha_1 + I_r \\ 0 \end{pmatrix} \beta' W_t + K\Phi(L)^{-1} u_t$$

and $\beta' W_t$ having a trend is $\alpha_1 + I_r$ having roots close to one.

In the special case of a bivariate model with one possible cointegrating vector the autoregressive coefficient is given by $\rho_c = \alpha_1 + 1$. Hence modelling ρ_c to be local to one is equivalent to modelling $\alpha_1 = -\gamma/T$. The model without additional serial correlation becomes

$$\begin{pmatrix} \Delta c_t \\ \Delta x_t \end{pmatrix} = \begin{pmatrix} \rho_c - 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} - \theta u_{2t} \\ u_{2t} \end{pmatrix}$$

in triangular form and

$$\begin{pmatrix} \Delta y_t \\ \Delta x_t \end{pmatrix} = \begin{pmatrix} \rho_c - 1 \\ 0 \end{pmatrix} \begin{pmatrix} 1 & -\theta \end{pmatrix} \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$$

in the error correction form. We will thus focus on the simplified model for the object of focus

$$\Delta y_t = (\rho_c - 1) c_{t-1} + u_{1t} \tag{12}$$

as the forecasting model.

The model where we set ρ_c to unity here as an approximation results in the forecast equal to the no change forecast, i.e. $y_{T+h|T} = y_T$. Thus the unconditional forecast error is given by

$$E\big[y_{T+1} - y_T^f\big]^2 = E\big[(u_{1T+1}) - (\rho - 1)(y_T - \theta x_T)\big]^2$$
$$\approx \sigma_1^2\left(1 + T^{-1}\left\{\frac{\sigma_c^2}{\sigma_1^2}\right\}\frac{\gamma(1 - \mathrm{e}^{-2\gamma})}{2}\right),$$

where $\sigma_1^2 = \mathrm{var}(u_{1t})$ and $\sigma_c^2 = \mathrm{var}(u_{1t} - \theta u_{2t})$ is the variance of the shocks driving the cointegrating vector. This is similar to the result in the univariate model forecast when we use a random walk forecast, with the addition of the component $\{\sigma_c^2/\sigma_1^2\}$ which alters the effect of imposing the unit root. This ratio shows that the result depends greatly on the ratio of the variance of the cointegrating vector vis a vis the variance of the shock to y_t. When this ratio is small, which is to say that when the cointegrating relationship varies little compared to the variation in Δy_t, then the impact of ignoring the cointegrating vector is small for one step ahead forecasts. This makes intuitive sense – in such cases the cointegrating vector does not much depart from its mean and so has little predictive power in determining what happens to the path of y_t.

That the loss from imposing a unit root here – which amounts to running the model in differences instead of including an error correction term – depends on the size of the shocks to the cointegrating vector relative to the shocks driving the variable to be forecast means that the trade-off between estimation of the model and imposing the root will vary with this correlation. This adds yet another factor that would drive the choice between imposing the unit root or estimating it. When the ratio is unity, the results are identical to the univariate near unit root problem. Different choices for the correlation between u_{1t} and u_{2t} will result in different ratios and different trade-offs. Figure 11 plots, for $\{\sigma_c^2/\sigma_1^2\} = 0.56$ and 1 and $T = 100$ the average one step ahead MSE of the forecast error for both the imposition of the unit root and also the model where the regression (12) is run with a constant in the model and these OLS coefficients used to construct the forecast. In this model the cointegrating vector is assumed known with little loss as the estimation error on this term has a lower order effect.

The figure graphs the MSE relative to the model with all coefficients known to γ on the horizontal axis. The relatively flat solid line gives the OLS MSE forecast results for both models – there is no real difference between the results for each model. The steepest upward sloping line (long and short dashes) gives results for the unit root imposed model where $\sigma_c^2/\sigma_1^2 = 1$, these results are comparable to the $h = 1$ case in Figure 1 (the asymptotic results suggest a slightly smaller effect than this small sample simulation). The flatter curve corresponds to $\sigma_c^2/\sigma_1^2 < 1$ for the cointegrating vector chosen here ($\theta = 1$) and so the effect of erroneously imposing a unit root is smaller. However this ratio could also be larger, making the effect greater than the usual unit root model. The result depends on the values of the nuisance parameters. This model is however highly stylized. More complicated dynamics can make the coefficient on the cointegrating vector larger or smaller, hence changing the relevant size of the effect.

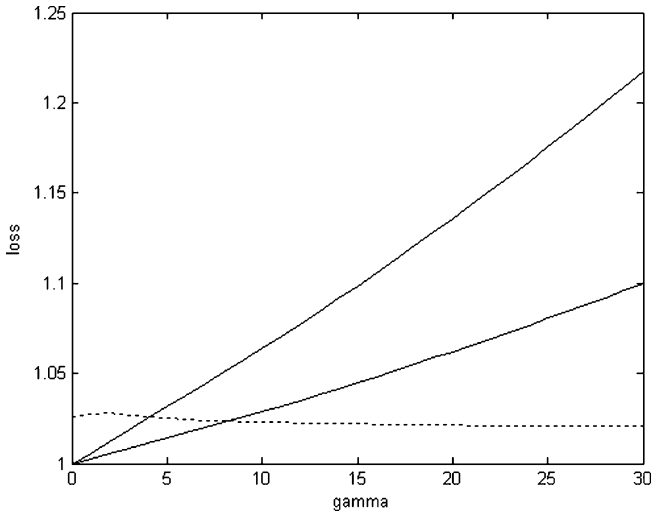

Figure 11. The upward sloping lines show loss from imposing a unit root for $\sigma_1^{-2}\sigma_c^2 = 0.56$ and 1 for steeper curves, respectively. The dashed line gives the results for OLS estimation (both models).

In the alternate case, where we are sure the cointegrating vector does not have too much persistence however we are unsure if there are unit roots in the underlying data, the model is close to one in differences. This can be seen in the general case from the general VAR form

$$W_t = A(L)W_{t-1} + u_t,$$
$$\Delta W_t = \big(A(1) - I_n\big)W_{t-1} + A^*(L)\,\Delta W_{t-1} + u_t$$

through using the Beveridge Nelson decomposition. Now let $\Psi = A(1) - I_n$ and consider the rotation

$$\begin{aligned}\Psi W_{t-1} &= \Psi K^{-1} K W_{t-1} \\ &= [\Psi_1, \Psi_2] \begin{pmatrix} I_r & \theta \\ 0 & I_{n-r} \end{pmatrix} \begin{pmatrix} I_r & \theta \\ 0 & I_{n-r} \end{pmatrix} \begin{pmatrix} \beta' W_t \\ W_{2t} \end{pmatrix} \\ &= \Psi_1 \beta' W_{t-1} + (\Psi_2 + \theta\Psi_1) W_{2t-1},\end{aligned}$$

hence the model can be written as

$$\Delta W_t = \Psi_1 \beta' W_{t-1} + (\Psi_2 + \theta\Psi_1) W_{2t-1} + A^*(L)\,\Delta W_{t-1} + u_t,$$

where the usual ECM arises if $(\Psi_2 + \theta\Psi_1)$ is zero. This is the zero restriction implicit in the cointegration model. Hence in the general case the 'near to unit root' of the right-hand side variables in the cointegrating framework is modelling this term to be near to zero.

This model has been analyzed in the context of long run forecasting in very general models by Stock (1996). To capture these ideas consider the triangular form for the model without serial correlation

$$\begin{pmatrix} y_t - \varphi' z_t - \theta x_t \\ (1-\rho_x L)(x_t - \phi' z_t) \end{pmatrix} = K u_t = \begin{pmatrix} u_{1t} - \theta u_{2t} \\ u_{2t} \end{pmatrix}$$

so we have $y_{T+h} = \varphi' z_{T+h} + \theta x_{T+h} + u_{1T+h} - \theta u_{2T+h}$. Combining this with the model of the dynamics of x_t gives the result for the forecast model. We have

$$\begin{aligned} &x_t = \phi z_t + u_{2t}^*, \quad t = 1, \ldots, T, \\ &(1-\rho_x L) u_{2t}^* = u_{2t}, \quad t = 2, \ldots, T, \\ &u_{21}^* = \xi, \end{aligned}$$

and so as

$$x_{T+h} - x_T = \sum_{i=1}^{h} \rho_x^{h-i} u_{2T+i} + \big(\rho^h - 1\big)\big(x_T - \phi' z_T\big) + \phi'(z_{T+h} - z_T),$$

then

$$\begin{aligned} y_{T+h} - y_T = \theta &\left(\sum_{i=1}^{h} \rho^{h-i} u_{2T+i} + \big(\rho^h - 1\big)\big(x_T - \phi' z_T\big) + \phi'(z_{T+h} - z_T) \right) \\ &- c_T + \varphi'(z_{T+h} - z_T) + u_{1T+h} - \theta u_{2T+h}. \end{aligned}$$

From this we can compute some distributional results.

If a unit root is assumed (cointegration 'wrongly' assumed) then the forecast is

$$\begin{aligned} y_{T+h|T}^{\mathrm{R}} - y_T &= \theta\phi'(z_{T+h} - z_T) - c_T + \varphi'(z_{T+h} - z_T) \\ &= (\theta\phi + \varphi)'(z_{T+h} - z_T) - c_T. \end{aligned}$$

In the case of a mean this is simply

$$y_{T+h|T}^{\mathrm{R}} - y_T = -(y_T - \varphi_1 - \gamma x_T)$$

and for a time trend it is

$$\begin{aligned} y_{T+h|T}^{\mathrm{R}} - y_T &= \theta\phi'(z_{T+h} - z_T) - c_T + \varphi'(z_{T+h} - z_T) \\ &= (\theta\phi_2 + \varphi_2)h - (y_T - \varphi_1 - \varphi_2 T - \theta x_T). \end{aligned}$$

If we do not impose the unit root we have the forecast model

$$\begin{aligned} y_{T+h|T}^{\mathrm{UR}} - y_T &= \theta\big(\rho^h - 1\big)\big(x_T - \phi' z_T\big) + \phi'(z_{T+h} - z_T) - c_T + \varphi'(z_{T+h} - z_T) \\ &= (\theta\phi + \varphi)'(z_{T+h} - z_T) - c_T - \theta\big(\rho^h - 1\big)\big(x_T - \phi' z_T\big). \end{aligned}$$

This allows us to understand the costs and benefits of imposition. The real discussion here is between imposing the unit root (modelling as a cointegrating model) and not

imposing the unit root (modelling the variables in levels). Here the difference in the two forecasts is given by

$$y^{\mathrm{UR}}_{T+h|T} - y^{\mathrm{R}}_{T+h|T} = -\theta\big(\rho^h - 1\big)\big(x_T - \phi' z_T\big).$$

We have already examined such terms. Here the size of the effect is driven by the relative size of the shocks to the covariates and the shocks to the cointegrating vector, although the effect is the reverse of the previous model (in that model it was the cointegrating vector that is persistent, here it is the covariate). As before the effect is intuitively clear, if the shocks to the near nonstationary component are relatively small then x_T will be close to the mean and the effect is reduced. An extra wedge is driven into the effect by the cointegrating vector θ. A large value for this parameter implies that in the true model that x_t is an important predictor of y_{t+1}. The cointegrating term picks up part of this but not all, so ignoring the rest becomes costly.

As in the case of the near unit root cointegrating vector this model is quite stylized and models with a greater degree of dynamics will change the size of the results, however the general flavor remains.

6. Predicting noisy variables with trending regressors

In many problems the dependent variable itself displays no obvious trending behavior, however theoretically interesting covariates tend to exhibit some type of longer run trend. For many problems we might rule out unit roots for these covariates, however the trend is sufficiently strong that often tests for a unit root fail to reject and by implication standard asymptotic theory for stationary variables is unlikely to approximate well the distribution of the coefficient on the regressor. This leads to a number of problems similar to those examined in the models above.

To be concrete, consider the model

$$y_{1t} = \beta_0' z_t + \beta_1 y_{2t-1} + v_{1t} \tag{13}$$

which is to be used to predict y_{1t}. Further, suppose that y_{2t} is generated by the model in (1) in Section 3. The model for $v_t = [v_{1t}, v_{2t}]'$ is then $v_t = b^*(L)\eta_t^*$ where $E[\eta_t^* \eta_t^{*\prime}] = \Sigma$ where

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \delta\sigma_{11}\sigma_{22} \\ \delta\sigma_{11}\sigma_{22} & \sigma_{22}^2 \end{pmatrix}$$

and

$$b^*(L) = \begin{pmatrix} 1 & 0 \\ 0 & c(L) \end{pmatrix}.$$

The assumption that v_{1t} is not serially correlated accords with the forecasting nature of this regression, if serial correlation were detected we would include lags of the dependent variable in the forecasting regression.

This regression has been used in many instances for forecasting. First, in finance a great deal of attention has been given to the possibility that stock market returns are predictable. In the context of (13) we have y_t being stock returns from period $t-1$ to t and y_{2t-1} is any predictor known at the time one must undertake the investment to earn the returns y_{1t}. Examples of predictors include dividend–price ratio, earnings to price ratios, interest rates or spreads [see, for example, Fama and French (1998), Campbell and Shiller (1988a, 1988b) Hodrick (1992)]. Yet each of these predictors tends to display large amounts of persistence despite the absence of any obvious persistence in returns [Stambaugh (1999)]. The model (13) also well describes the regression run at the heart of the 'forward market unbiasedness' puzzle first examined by Bilson (1981). Typically such a regression regresses the change in the spot exchange rate from time $t-1$ to t on the forward premium, defined as the forward exchange rate at time $t-1$ for a contract deliverable at time t less the spot rate at time $t-1$ (which through covered interest parity is simply the difference between the interest rates of the two currencies for a contract set at time $t-1$ and deliverable at time t). This can be recast as a forecasting problem through subtracting the forward premium from both sides, leaving the uncovered interest parity condition to mean that the difference between the realized spot rate and the forward rate should be unpredictable. However the forward premium is very persistent [Evans and Lewis (1995) argue that this term can appear quite persistent due to the risk premium appearing quite persistent]. The literature on this regression is huge. Froot and Thaler (1990) give a review. A third area that fits this regression is use of interest rates or the term structure of the interest rates to predict various macroeconomic and financial variables. Chen (1991) shows using standard methods that short run interest rates and the term structure are useful for predicting GNP.

There are a few 'stylized' facts about such prediction problems. First, in general the coefficient β often appears to be significantly different from one under the usual stationary asymptotic theory (i.e. the t statistic is outside the ± 2 bounds). Second, R^2 tends to be very small. Third, often the coefficient estimates seem to vary over subsamples more than standard stationary asymptotic theory might predict. Finally, these relationships have a tendency to 'break down' – often the in sample forecasting ability does not seem to translate to out of sample predictive ability. Models where β is equal to or close to zero and regressors that are nearly nonstationary combined with asymptotic theory that reflects this trending behavior in the predictor variable can to some extent account for all of these stylized facts.

The problem of inference on the OLS estimator $\hat{\beta}_1$ in (13) has been studied in both cases specific to particular regressions and also more generally. Stambaugh (1999) examines inference from a Bayesian viewpoint. Mankiw and Shapiro (1986), in the context of predicting changes in consumption with income, examined these types of regressions employing Monte Carlo methods to show that t statistics overreject the null hypothesis that $\beta = 0$ using conventional critical values. Elliott and Stock (1994) and Cavanagh, Elliott and Stock (1995) examined this model using local to unity asymptotic theory to understand this type of result. Jansson and Moriera (2006) provide methods to test this hypothesis.

First, consider the problem that the t statistic overrejects in the above regression. Elliott and Stock (1994) show that the asymptotic distribution of the t statistic testing the hypothesis that $\beta_1 = 0$ can be written as the weighted sum of a mixed normal and the usual Dickey and Fuller t statistic. Given that the latter is not well approximated by a normal, the failure of empirical size to equal nominal size will result when the weight on this nonstandard part of the distribution is nonzero.

To see the effect of regressing with a trending regressor we will rotate the error vector v_t through considering $\eta_t = Rv_t$ where

$$R = \begin{pmatrix} 1 & -\delta\frac{\sigma_{11}}{c(1)\sigma_{22}} \\ 0 & 1 \end{pmatrix}$$

so $\eta_{1t} = v_{1t} - \delta\frac{\sigma_{11}}{c(1)\sigma_{22}}v_{2t} = v_{1t} - \delta\frac{\sigma_{11}}{c(1)\sigma_{22}}\eta_{2t}$. This results in the spectral density of η_t at frequency zero scaled by 2π equal to $Rb^*(1)\Sigma b^*(1)R'$ which is equivalent to

$$\Omega = Rb^*(1)\Sigma b^*(1)R' = \begin{pmatrix} \sigma_{22}^2(1-\delta^2) & 0 \\ 0 & c(1)^2\sigma_{11}^2 \end{pmatrix}.$$

Now consider the regression

$$\begin{aligned} y_{1t} &= \beta_0' z_t + \beta_1 y_{1t-1} + v_{1t} \\ &= \left(\beta_0' + \phi'\right)z_{t-1} + \beta_1\left(y_{2t-1} - \phi' z_{t-1}\right) + v_{1t} \\ &= \tilde{\beta}_0' z_{t-1} + \beta_1\left(y_{2t-1} - \phi' z_{t-1}\right) + v_{1t} \\ &= \beta' X_t + v_{1t}, \end{aligned}$$

where $\beta = (\tilde{\beta}_0', \beta_1)'$ and $X_t = (z_t', y_{1t-1} - \phi' z_{t-1})'$.

Typically OLS is used to examine this regression. We have that

$$\begin{aligned} \hat{\beta} - \beta &= \left(\sum_{t=2}^{T} X_t X_t'\right)^{-1} \sum_{t=2}^{T} X_t v_{2t} \\ &= \left(\sum_{t=2}^{T} X_t X_t'\right)^{-1} \sum_{t=2}^{T} X_t \eta_{2t} + \delta\frac{\sigma_{22}}{c(1)\sigma_{11}}\left(\sum_{t=2}^{T} X_t X_t'\right)^{-1} \sum_{t=2}^{T} X_t \eta_{1t} \end{aligned}$$

since $v_{2t} = \eta_{2t} + \delta\frac{\sigma_{22}}{c(1)\sigma_{11}}\eta_{1t}$. What we have done is rewritten the shock to the forecasting regression into orthogonal components describing the shock to the persistent regressor and the shock unrelated to y_{2t}.

To examine the asymptotic properties of the estimator, we require some additional assumptions. Jointly we can consider the vector of partial sums of η_t and we assume that this partial sum satisfies a functional central limit theorem (FCLT)

$$T^{-1/2}\sum_{t=1}^{[T.]} \eta_t \Rightarrow \Omega^{1/2}\begin{pmatrix} W_{2.1}(\cdot) \\ M(\cdot) \end{pmatrix},$$

where $M(\cdot)$ is as before and is asymptotically independent of the standard Brownian motion $W_{2.1}(\cdot)$.

Now the usefulness of the decomposition of the parameter estimator into two parts can be seen through examining what each of these terms look like asymptotically when suitably scaled. The first term, by virtue of η_{1t} being orthogonal to the entire history of x_t, will when suitably scaled have an asymptotic mixed normal distribution. The second term is exactly what we would obtain, apart from being multiplied at the front by $\delta\frac{\sigma_{22}}{\sigma_{11}}$, in the Dickey and Fuller (1979) regression of x_t on a constant and lagged dependent variable. Hence this term has the familiar nonstandard distribution from that regression when standardized in the same way as the first term. Also by virtue of the independence of η_{1t} and ε_{2t} each of these terms is asymptotically independent. Thus the limit distribution for the standardized coefficients is a weighted sum of a mixed normal and a Dickey and Fuller (1979) distribution, which will not be well approximated by a normal distribution.

Now consider the t statistic testing $\beta = 0$. The t statistic testing the hypothesis that $\beta_1 = 0$ when this is the null is typically employed to justify the regressors inclusion in the forecasting equation. This t statistic has an asymptotic distribution given by

$$t_{\hat{\beta}_1=0} \Rightarrow \left(1-\delta^2\right)^{1/2} z^* + \delta \mathrm{DF},$$

where z^* is distributed as a standard normal and DF is the usual Dickey and Fuller t distribution when $c(1) = 1$ and $\gamma = 0$ and a variant of it otherwise. The actual distribution is

$$\mathrm{DF} = \frac{0.5(M^d(1)^2 - M^d(0)^2 - c(1)^2)}{\int M^d(s)\,\mathrm{d}s},$$

where $M^d(s)$ is the projection of $M(s)$ on the continuous analog of z_t. When $\gamma = 0$, $c(1) = 1$ and at least a constant term is included this is identical to the usual DF distribution with the appropriate order of deterministic terms. When $c(1)$ is not one we have an extra effect through the serial correlation [cf. Phillips (1987)].

The nuisance parameter that determines the weights, δ, is the correlation between the shocks driving the forecasting equation and the quasi difference of the covariate to be included in the forecasting regression. Hence asymptotically, this nuisance parameter along with the local to unity parameter describe the extent to which this test for inclusion over rejects.

The effect of the trending regressor on the type of R^2 we are likely to see in the forecasting regression (13) can be seen through the relationship between the t statistic and R^2 in the model where only a constant is included in the regression. In such models we have that the R^2 for the regression is approximately $T^{-1}t^2_{\beta_1=0}$. In the usual case of including a stationary regressor without predictive power we would expect that TR^2 is approximately the square of the t statistic testing exclusion of the regressor, i.e. is distributed as a χ^2_1 random variable, hence on average we expect R^2 to be T^{-1}. But in the case of a trending regressor $t^2_{\beta_1=0}$ will not be well approximated by a χ^2_1 as the

Table 1
Overrejection and R^2 as a function of endogeneity

		$\delta = 0.1$	0.3	0.5	0.7	0.9
0	% rej	0.058	0.075	0.103	0.135	0.165
	ave R^2	0.010	0.012	0.014	0.017	0.019
5	% rej	0.055	0.061	0.070	0.078	0.087
	ave R^2	0.010	0.011	0.011	0.012	0.013
10	% rej	0.055	0.058	0.062	0.066	0.071
	ave R^2	0.010	0.010	0.011	0.011	0.012
15	% rej	0.056	0.057	0.059	0.062	0.065
	ave R^2	0.010	0.010	0.011	0.011	0.011
20	% rej	0.055	0.057	0.059	0.060	0.063
	ave R^2	0.010	0.010	0.010	0.011	0.011

t statistic is not well approximated by a standard normal. On average the R^2 will be larger and because of the long tail of the DF distribution there is a larger chance of having relatively larger values for R^2. However, we still expect R^2 to be small most of the time even though the test of inclusion rejects.

The extent of overrejection and the average R^2 for various values of δ and γ are given in Table 1 for a test with nominal size equal to 5%. The sample size is $T = 100$ and zero initial condition for y_{1t} was employed.

The problem is larger the closer y_{1t} is to having a unit root and the larger is the long run correlation coefficient δ. For moderate values of δ, the effect is not great. The rejection rate numbers mask the fact that the $t_{\beta_1=0}$ statistics can on occasion be far from ± 2. A well-known property of the DF distribution is a long tail on the left-hand side of the distribution. The sum of these distributions will also have such a tail – for $\delta > 0$ it will be to the left of the mean and for $\delta > 0$ to the right. Hence some of these rejections can appear quite large using the asymptotic normal as an approximation to the limit distribution. This follows through to the types of values for R^2 we expect. Again, when γ is close to zero and δ is close to one the R^2 is twice what we expect on average, but still very small. Typically it will be larger than expected, but does not take on very large values. This conforms with the common finding of trending predictors appearing to be useful in the regression through entering the forecasting regression with statistically significant coefficients however they do not appear to pick up much of the variation in the variable to be predicted.

The trending behavior of the regressor can also explain greater than expected variability in the coefficient estimate. In essence, the typically reported standard error of the estimate based on asymptotic normality is not a relevant guide to the sampling variability of the estimator over repeated samples and hence expectations based on this will mislead. Alternatively, standard tests for breaks in coefficient estimates rely on the stationarity of the regressors, and hence are not appropriate for these types of regressions.

Hansen (2000) gives an analysis of break testing when the regressor is not well approximated by a stationary process and provides a bootstrap method for testing for breaks.

In all of the above, I have considered one step ahead forecasts. There are two approaches that have been employed for greater than one step ahead forecasts. The first is to consider the regression $y_{1t} = \beta_0' z_t + \beta_1 y_{2t-h} + \tilde{v}_{1t}$ as the model that generates the h step ahead forecast where $\tilde{v}_{1t}$ is the iterated error term. In this case results very similar to those given above apply.

A second version is to examine the forecastability of the cumulation of h steps of the variable to be forecast. The regression is

$$\sum_{i=1}^{h} y_{1t+i} = \beta_0' z_t + \beta_1 y_{2t} + \tilde{v}_{2t+h}.$$

Notice that for large enough h this cumulation will act like a trending variable, and hence greatly increase the chance that such a regression is really a spurious regression. Thus when y_{2t} has a unit root or near unit root behavior the distribution of $\hat{\beta}_1$ will be more like that of a spurious regression, and hence give the appearance of predictability even when there is none there. Unlike the results above, this can be true even if the variable is strictly exogenous. These results can be formalized analytically through considering the asymptotic thought experiment that $h = [\lambda T]$ as in Section 3 above. Valkenov (2003) explicitly examines this type of regression for $z_t = 1$ and general serial correlation in the predictor and shows the spurious regression result analytically.

Finally, there is a strong link between these models and those of Section 5 above. Compare Equation (12) and the regression examined in this section. Renaming the dependent variable in (12) as y_{2t} and the 'cointegrating' vector y_{1t} we have the model of this section.

7. Forecast evaluation with unit or near unit roots

A number of issues arise here. In this handbook West examines issues in forecast evaluation when the model is stationary. Here, when the data have unit root or near unit root behavior then this must be taken into account when conducting the tests. It will also affect the properties of constructed variables such as average loss depending on the model. Alternatively, other possibilities arise in forecast evaluation. The literature that extends these results to use of nonstationary data is much less well developed.

7.1. Evaluating and comparing expected losses

The natural comparison between forecasting procedures is to compare the procedures based on 'holdout' samples – use a portion of the sample to estimate the models and a portion of the sample to evaluate them. The relevant statistic becomes the average 'out of sample' loss. We can consider the evaluation of any forecasting model where either (or

both) the outcome variable and the covariates used in the forecast might have unit roots or near unit roots. The difficulty that typically arises for examining sample averages and estimator behavior when the variables are not obviously stationary is that central limit theorems do not apply. The result is that these sample averages tend to converge to nonstandard distributions that depend on nuisance parameters, and this must be taken into account when comparing out of sample average MSE's as well as in understanding the sampling error in any given average MSE.

Throughout this section we follow the majority of the (stationary) literature and consider a sampling scheme where the T observations are split between a model estimation sample consisting of the observations $t = 1, \ldots, T_1$, and an evaluation sample $t = T_1 + 1, \ldots, T$. For asymptotic results we allow both samples to get large, defining $\kappa = T_1/T$. Further, we will allow the forecast horizon h to remain large as T increases, we set $h/T = \lambda$. We are thus examining approximations to situations where the forecast horizon is substantial compared to the sample available. These results are comparable to the long run forecasting results of the earlier sections.

As an example of how the sample average of out of sample forecast errors converges to a nonstandard distribution dependent on nuisance parameters, we can examine the simple univariate model of Section 3. In the mean case the forecast of y_{t+h} at time t is simply y_t and so the average forecast error for the holdout sample is

$$\mathrm{MSE}(h) = \frac{1}{T - T_1 - h} \sum_{t=T_1+1}^{T-h} (y_{t+h} - y_t)^2.$$

Now allowing $T(\rho - 1) = -\gamma$ then using the FCLT and continuous mapping theorem we have that after rescaling by T^{-1} then

$$\begin{aligned} T^{-1}\mathrm{MSE}(h) &= \frac{T}{T - T_1 - h} T^{-1} \sum_{t=T_1+1}^{T-h} \left(T^{-1/2} y_{t+h} - T^{-1/2} y_t\right)^2 \\ &\Rightarrow \sigma_\varepsilon^2 \frac{1}{1 - \lambda - \kappa} \int_\kappa^{1-\lambda} \left(M(s + \lambda) - M(s)\right)^2 \mathrm{d}s. \end{aligned}$$

The additional scaling by T gives some hint to understanding the output of average out of sample forecast errors. The raw average of out of sample forecast errors gets larger as the sample size increases. Thus interpreting directly this average as the likely forecast error using the model to forecast the next h periods is misleading. However on rescaling, it can be considered in this way. In the case where the initial value for the process y_t comes from its unconditional distribution, i.e. $\alpha = 1$, the limit distribution has a mean that is exactly the expected value for the expected MSE of a single h step ahead forecast.

When the largest root is estimated these expressions become even more complicated functions of Brownian motions, and as earlier become very difficult to examine analytically.

When the forecasting model is complicated further, by the addition of extra variables in the forecasting model, asymptotic approximations for average out of sample

forecast error become even more complicated, typically depending on all the nuisance parameters of the model. Corradi, Swanson and Olivetti (2001) extend results to the cointegrated case where the rank of cointegration is known. In such models the variables that enter the regressions are stationary, and the same results as for stationary regression arise so long as loss is quadratic or the out of sample proportion grows at a slower rate than the in sample proportion (i.e. κ converges to one). Rossi (2005) provides analytical results for comparing models where all variables have near unit roots against the random walk model, along with methods for dealing with the nuisance parameter problem.

7.2. Orthogonality and unbiasedness regressions

Consider the basic orthogonality regression for differentiable loss functions, i.e. the regression

$$L'(e_{t+h}) = \beta' X_t + \varepsilon_{t+h}$$

(where X_t includes any information known at the time the forecast is made and $L'(\cdot)$ is the first derivative of the loss function) and we wish to test the hypothesis H_0: $\beta = 0$. If some or all of the variables in X_t are integrated or near integrated, then this affects the sampling distribution of the parameter estimates and the corresponding hypothesis tests.

This arises in practice in a number of instances. We have earlier noted that one popular choice for X_t, namely the forecast itself, has been used in testing what is known as 'unbiasedness' of the forecasts. In the case of MSE loss, where $L'(e_{t+h}) = e_{t+h}/2$ then unbiasedness means that on average the forecast is equal to the outcome. This can be done in the context of the regression above using

$$y_{t+h} - y_{t,t+h} = \beta_0 + \beta_1 y_{t+h,t} + \varepsilon_{t+h}.$$

If the series to be forecast is integrated or near integrated, then the predictor in this regression will have these properties and standard asymptotic theory for conducting this test does not apply.

Another case might be a situation where we want to construct a test that has power against a small nonstationary component in the forecast error. Including only stationary variables in X_t would not give any power in that direction, and hence one may wish to include a nonstationary variable. Finally, many variables that are suggested in theory to be potentially correlated with outcomes may exhibit large amounts of persistence. Such variables include interest rates etc. Again, in these situations we need to account for the different sampling behavior.

If the variables X_t can be neatly split (in a known way) between variables with unit roots and variables without and it is known how many cointegrating vectors there are amongst the unit root variables, then the framework of the regression fits that of Sims, Stock and Watson (1990). Under their assumptions the OLS coefficient vector $\hat{\beta}$ converges to a nonstandard distribution which involves functions of Brownian motions and

normal variates. The distribution depends on nuisance parameters and standard tabulation of critical values is basically infeasible (the number of dimensions would be large). As a consequence, finding the critical values for the joint test of orthogonality is quite difficult.

This problem is of course equivalent to that of the previous section when it comes to distribution theory for $\hat{\beta}$ and consequently on testing this parameter. The same issues arise. Thus orthogonality tests with integrated or near integrated regressors are problematic, even without thinking about the construction of the forecast errors. Failure to realize the impacts of these correlations on the hypothesis test (i.e. proceeding as if the t statistics had asymptotic normal distributions or that the F statistics have asymptotic chi-square distributions) results in overrejection. Further, there is no simple method for constructing the alternate distributions, especially when there is uncertainty over whether or not there is a unit root in the regressor [see Cavanagh, Elliott and Stock (1995)].

Additional issues also arise when X_t includes the forecast or other constructed variables. In the stationary case results are available for various construction schemes (see Chapter 3 by West in this Handbook). These results will not in general carry over to the problem here.

7.3. Cointegration of forecasts and outcomes

An implication of good forecasting when outcomes are trending would be that forecasts and outcomes of the variable of interest would have a difference that is not trending. In this sense, if the outcomes have a unit root then we would expect forecasts and outcomes to be cointegrated. This has led some researchers to examine whether or not the forecasts made in practice are indeed cointegrated with the variable being forecast. The expected cointegrating vector is $\beta = (1, -1)'$, implying that the forecast error is stationary. This has been undertaken for exchange rates [Liu and Maddala (1992)] and macroeconomic data [Aggarwal, Mohanty and Song (1995)]. In the context of macroeconomic forecasts, Cheung and Chinn (1999) also relax the cointegrating vector assumption that the coefficients are known and estimate these coefficients.

The requirement that forecasts be cointegrated with outcomes is a very weak requirement. Note that the forecasters information set includes the current value of the outcome variable. Since the current value of the outcome variable is trivially cointegrated with the future outcome variable to be forecast (they differ by the change, which is stationary) then the forecaster has a simple observable forecast that satisfies the requirement that the forecast and outcome variable be cointegrated. This also means that forecasts generated by adding any stationary component to the current level of the variable will also satisfy the requirement of cointegration between the forecasts and the outcome. Thus even forecasts of the change that are uncorrelated with the actual change provided they are stationary will result in cointegration between forecasts and outcomes.

We can also imagine what happens under the null hypothesis of no cointegration. Under the null, forecast errors are $I(1)$ and hence become arbitrarily far from zero with

probability one. It is hard to imagine that a forecaster would stick with such a method when the forecast becomes further from the current value of the outcome than typical changes in the outcome variable would suggest are plausible.

That this weak requirement obviously holds in many cases has not meant that the hypothesis has not been rejected. As with all testing situations, one must consider the test a joint test of the proposition being examined and the assumptions under which the test is derived. Given the unlikely event that forecasts and outcomes are truly becoming arbitrarily far apart, as would be suggested by a lack of cointegration, perhaps the problem is in the assumption that the trend is correctly characterized by a unit root. In the context of hypothesis testing on the β parameters Elliott (1998) shows that near unit roots causes major size distortions for tests on this parameter vector.

Overall, these tests are not likely to shed much light on the usefulness of forecasts.

8. Conclusion

Making general statements as to how to proceed with forecasting when there is trending behavior is difficult due to the strong dependence of the results on a myriad of nuisance parameters of the problem – extent of deterministic terms, initial values and descriptions of serial correlation. This becomes even more true when the model is multivariate, since there are many more combinations of nuisance parameters that can either reduce or enhance the value of estimation over imposition of unit roots.

Theoretically though a number of points arise. First, except for roots quite close to one estimation should outperform imposition of unit roots in terms of MSE error. Indeed, since estimation results in bounded MSE over reasonable regions of uncertainty over the parameter space whereas imposition of unit roots can result in very large losses it would seem to be the conservative approach would be to estimate the parameters if we are uncertain as to their values. This goes almost entirely against current practice and findings with real data. Two possibilities arise immediately. First, the models for which under which the theory above is useful are not good models of the data and hence the theoretical size of the trade-offs are different. Second, there are features of real data that, although the above models are reasonable, they affect the estimators in ways ignored by the models here and so when parameters are estimated large errors make the results less appropriate. Given that tests designed to distinguish between various models are not powerful enough to rule out the models considered here, it is unlikely that these other functions of the data – evaluations of forecast performance – will show the differences between the models.

For multivariate models the differences are exacerbated in most cases. Theory shows that imposing cointegration on the problem when true is still unlikely to help at longer horizons despite its nature as a long run restriction on the data. A number of authors have sought to characterize this issue as not one of imposing cointegration but imposing the correct number of unit roots on the model, however these are of course equivalent. It is true however that it is the estimation of the roots that can cause MSE to be larger,

they can be poorly estimated in small samples. More directly though is that the trade-offs are similar in nature to the univariate model. Risk is bounded when the parameters are estimated.

Finally, it is not surprising that there is a short horizon/long horizon dichotomy in the forecasting of variables when the covariates display trending behavior. In the short run we are relating a trending variable to a nontrending one, and it is difficult to write down such a model where the trending covariate is going to explain a lot of the nontrending outcome. At longer horizons though the long run prediction becomes the sum of stationary increments, allowing trending covariates a greater opportunity of being correlated with the outcome to be forecast.

In part a great deal of the answer probably lies in the high correlation between the forecasts that arise from various assumptions and also the unconditional nature of the results of the literature. On the first point, given the data the differences just tend not to be huge and hence imposing the root and modelling the variables in differences not greatly costly in most samples, imposing unit roots just makes for a simpler modelling exercise. This type of conditional result has not been greatly examined in the literature. Things brings the second point – for what practical forecasting problems does the unconditional, i.e. averaging over lots of data sets, best practice become relevant? This too has not been looked at deeply in the literature. When the current variable is far from its deterministic component, estimating the root (which typically means using a mean reverting model) and imposing the unit root (which stops mean reversion) have a bigger impact in the sense that they generate very different forecasts. The modelling of the trending nature becomes very important in these cases even though on average it appears less important because we average over these cases as well as the more likely case that the current level of the variable is close to its deterministic component.

References

Abidir, K., Kaddour, H., Tzavaliz, E. (1999). "The influence of VAR dimensions on estimator biases". Econometrica 67, 163–181.

Aggarwal, R., Mohanty, S., Song, F. (1995). "Are survey forecasts of macroeconomic variables rational?" Journal of Business 68, 99–119.

Andrews, D. (1993). "Exactly median-unbiased estimation of first order autoregressive/unit root models". Econometrica 61, 139–165.

Andrews, D., Chen, Y.H. (1994). "Approximately median-unbiased estimation of autoregressive models". Journal of Business and Economics Statistics 12, 187–204.

Banerjee, A. (2001). "Sensitivity of univariate AR(1) time series forecasts near the unit root". Journal of Forecasting 20, 203–229.

Bilson, J. (1981). "The 'speculative efficiency' hypothesis". Journal of Business 54, 435–452.

Box, G., Jenkins, G. (1970). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Campbell, J., Perron, P. (1991). "Pitfalls and opportunities: What macroeconomists should know about unit roots". NBER Macroeconomics Annual, 141–201.

Campbell, J., Shiller, R. (1988a). "The dividend–price ratio and expectations of future dividends". Review of Financial Studies 1, 195–228.

Campbell, J., Shiller, R. (1988b). "Stock prices, earnings and expected dividends". Journal of Finance 43, 661–676.
Canjels, E., Watson, M. (1997). "Estimating deterministic trends in the presence of serially correlated errors". Review of Economics and Statistics 79, 184–200.
Cavanagh, C., Elliott, G., Stock, J. (1995). "Inference in models with nearly integrated regressors". Econometric Theory 11, 11231–11247.
Chen, N. (1991). "Financial investment opportunities and the macroeconomy". Journal of Finance 46, 495–514.
Cheung, Y.-W., Chinn, M. (1999). "Are macroeconomic forecasts informative? Cointegration evidence from the ASA-NBER surveys". NBER Discussion Paper 6926.
Christoffersen, P., Diebold, F. (1998). "Cointegration and long-horizon forecasting". Journal of Business and Economic Statistics 16, 450–458.
Clements, M., Hendry, D. (1993). "On the limitations of comparing mean square forecast errors". Journal of Forecasting 12, 617–637.
Clements, M., Hendry, D. (1995). "Forecasting in cointegrated systems". Journal of Applied Econometrics 11, 495–517.
Clements, M., Hendry, D. (1998). Forecasting Economic Time Series. Cambridge University Press, Cambridge, MA.
Clements, M., Hendry, D. (2001). "Forecasting with difference-stationary and trend-stationary models". Econometrics Journal 4, s1–s19.
Cochrane, D., Orcutt, G. (1949). "Applications of least squares regression to relationships containing autocorrelated error terms". Journal of the American Statistical Association 44, 32–61.
Corradi, V., Swanson, N.R., Olivetti, C. (2001). "Predictive ability with cointegrated variables". Journal of Econometrics 104, 315–358.
Dickey, D., Fuller, W. (1979). "Distribution of the estimators for autoregressive time series with a unit root". Journal of the American Statistical Association 74, 427–431.
Diebold, F., Kilian, L. (2000). "Unit-root tests are useful for selecting forecasting models". Journal of Business and Economic Statistics 18, 265–273.
Elliott, G. (1998). "The robustness of cointegration methods when regressors almost have unit roots". Econometrica 66, 149–158.
Elliott, G., Rothenberg, T., Stock, J. (1996). "Efficient tests for and autoregressive unit root". Econometrica 64, 813–836.
Elliott, G., Stock, J. (1994). "Inference in models with nearly integrated regressors". Econometric Theory 11, 1131–1147.
Engle, R., Granger, C. (1987). "Co-integration and error correction: Representation, estimation, and testing". Econometrica 55, 251–276.
Engle, R., Yoo, B. (1987). "Forecasting and testing in co-integrated systems". Journal of Econometrics 35, 143–159.
Evans, M., Lewis, K. (1995). "Do long-term swings in the dollar affect estimates on the risk premium?" Review of Financial Studies 8, 709–742.
Fama, E., French, F. (1998). "Dividend yields and expected stock returns". Journal of Financial Economics 35, 143–159.
Franses, P., Kleibergen, F. (1996). "Unit roots in the Nelson–Plosser data: Do they matter for forecasting". International Journal of Forecasting 12, 283–288.
Froot, K., Thaler, R. (1990). "Anomalies: Foreign exchange". Journal of Economic Perspective 4, 179–192.
Granger, C. (1966). "The typical spectral shape of and economic variable". Econometrica 34, 150–161.
Hall, R. (1978). "Stochastic implications of the life-cycle-permanent income hypothesis: Theory and evidence". Journal of Political Economy 86, 971–988.
Hansen, B. (2000). "Testing for structural change in conditional models". Journal of Econometrics 97, 93–115.
Hodrick, R. (1992). "Dividend yields and expected stock returns: Alternative procedures for inference measurement". Review of Financial Studies 5, 357–386.

Hoffman, D., Rasche, R. (1996). "Assessing forecast performance in a cointegrated system". Journal of Applied Econometrics 11, 495–516.
Jansson, M., Moriera, M. (2006). "Optimal inference in regression models with nearly integrated regressors". Econometrica. In press.
Johansen, S. (1991). "Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models". Econometrica 59, 1551–1580.
Kemp, G. (1999). "The behavior of forecast errors from a nearly integrated $I(1)$ model as both the sample size and forecast horizon gets large". Econometric Theory 15, 238–256.
Liu, T., Maddala, G. (1992). "Rationality of survey data and tests for market efficiency in the foreign exchange markets". Journal of International Money and Finance 11, 366–381.
Magnus, J., Pesaran, B. (1989). "The exact multi-period mean-square forecast error for the first-order autoregressive model with an intercept". Journal of Econometrics 42, 238–256.
Mankiw, N., Shapiro, M. (1986). "Do we reject too often: Small sample properties of tests of rational expectations models". Economic Letters 20, 139–145.
Meese, R., Rogoff, K. (1983). "Empirical exchange rate models of the seventies: Do they fit out of sample?" Journal of International Economics 14, 3–24.
Müller, U., Elliott, G. (2003). "Tests for unit roots and the initial observation". Econometrica 71, 1269–1286.
Nelson, C., Plosser, C. (1982). "Trends and random walks in macroeconomic time series: Some evidence and implications". Journal of Monetary Economics 10, 139–162.
Ng, S., Vogelsang, T. (2002). "Forecasting dynamic time series in the presence of deterministic components". Econometrics Journal 5, 196–224.
Phillips, P.C.B. (1979). "The sampling distribution of forecasts from a first order autoregression". Journal of Econometrics 9, 241–261.
Phillips, P.C.B. (1987). "Time series regression with a unit root". Econometrica 55, 277–302.
Phillips, P.C.B. (1998). "Impulse response and forecast error variance asymptotics in nonstationary VARs". Journal of Econometrics 83, 21–56.
Phillips, P.C.B., Durlauf, S.N. (1986). "Multiple time series regression with integrated processes". Review of Economic Studies 52, 473–495.
Prais, S., Winsten, C.B. (1954). "Trend estimators and serial correlation". Cowles Foundation Discussion Paper 383.
Rossi, B. (2005). "Testing long-horizon predictive ability with high persistence, and the Meese–Rogoff puzzle". International Economic Review 46, 61–92.
Roy, A., Fuller, W. (2001). "Estimation for autoregressive time series with a root near one". Journal of Business and Economic Studies 19, 482–493.
Sampson, M. (1991). "The effect of parameter uncertainty on forecast variances and confidence intervals for unit root and trend stationary time series models". Journal of Applied Econometrics 6, 67–76.
Sanchez, I. (2002). "Efficient forecasting in nearly non-stationary processes". Journal of Forecasting 21, 1–26.
Sims, C., Stock, J., Watson, M. (1990). "Inference in linear time series models with some unit roots". Econometrica 58, 113–144.
Stambaugh, R. (1999). "Predictive regressions". Journal of Financial Economics 54, 375–421.
Stock, J.H. (1987). "Asymptotic properties of least squares estimators of cointegrating vectors". Econometrica 55, 1035–1056.
Stock, J.H. (1991). "Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series". Journal of Monetary Economics 28, 435–459.
Stock, J.H. (1994). "Unit roots, structural breaks and trends". In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. Elsevier, Amsterdam, pp. 2740–2841.
Stock, J.H. (1996). "VAR, error correction and pretest forecasts at long horizons". Oxford Bulletin of Economics and Statistics 58, 685–701.
Stock, J.H., Watson, M.W. (1999). "A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series". In: Engle, R., White, H. (Eds.), Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger. Oxford University Press, Oxford, pp. 1–44.

Turner, J. (2004). "Local to unity, long horizon forecasting thresholds for model selection in the AR(1)". Journal of Forecasting 23, 513–539.

Valkenov, R. (2003). "Long horizon regressions: Theoretical results and applications". Journal of Financial Economics 68, 201–232.

Watson, M. (1994). "Vector autoregression and cointegration". In: Engle, R., McFadden, D. (Eds.), Handbook of Econometrics, vol. 4. Elsevier, Amsterdam, pp. 2843–2915.