Why do nonlinear models provide poor macroeconomic forecasts?*

Gray Calhoun
Iowa State University

Graham Elliott
University of California, San Diego
9500 Gilman Drive
LA JOLLA, CA, 92093-0508

April 20, 2012

Abstract

*Corresponding author : Graham Elliott, Department of Economics, 9500 Gilman Drive, LA JOLLA, CA, 92109, phone (858) 534 4481, fax (858) 534 7040, email:grelliott@ucsd.edu.

1 Introduction

Extending forecasting models beyond linear models has for some time been considered one of the directions in which macroeconomic forecasting models would take resulting in improvements in forecasting these variables.

There are good reasons to think that the true data generating processes of macroeconomic data are nonlinear. Nearly all economic models of the macroeconomy are very nonlinear. For example DSGE models of the economy predict complicated nonlinear relationships between the variables and between past and future variables. Indeed it would be highly surprising if the relationships between current and future outcomes of variables were truly linear in the variables.

From the perspective of forecasting, this future of nonlinear models used to predict outcomes has not materialized despite the availability of relatively long sets of data for estimating the models and the feasibility of computational methods to estimate them. In forecasting macroeconomic data, a number of studies over the last decade or so have found a very small role for parametric nonlinear models. Nonparametric models have also only found use for small niche variables, typically not for macroeconomic forecasting applications. Recent surveys of forecasting inflation (Wright (2012) show little influence of nonlinear models in forecasting these variables.

This paper seeks to go beyond documenting the performance of linear versus nonlinear models and ask why it is that these models do not appear to forecast well. Such an understanding would help in a number of directions. First, understanding why currently employed nonlinear ,models do not work well will help with understanding the correct direction for improving forecasting. Given our expectation that the economy is nonlinear it will help in directing choices over modelling nonlinear behavior.

We suggest and examine a number of possibilities for the relatively poor performance of nonlinear methods. These are (a) there is not really any evidence of nonlinearity in the data, it could be that the data just does not show that nonlinear models are appropriate; (b) the typical choices of functional form are not the correct ones for picking up the types of nonlinearity in the data. (c) the functional forms might be appropriate however the gains over standard linear models might be too small for them to be useful in forecasting macroeconomic variables; (d) the models may be appropriate however estimation issues may indicate that the models could perform poorly and finally (e) the models may perform poorly overall however the period of evaluation might be critical in choosing a nonlinear forecasting model.

In this study we consider only univariate forecasting methods, and focus on parametric nonlinear forecasting methods. We choose to include only two simple linear models for comparison, given previous results that show that using autoregressive models as a linear model baseline are appropriate.

There are some caveats to the results. First, they only extend to the parametric models considered. Second, they only extend to univariate models, it may well be that for multivariate models the results are different. Finally, there is some evidence that for non aggregated price data (for example electricity prices) that nonlinear models perform well. Such data is usually better measured and available at higher frequencies than the data examined here.

2 Overview of Models and Previous Empirical Results

From a parametric modelling perspective, there are only a few nonlinear models typically considered in practice. These take the form of allowing a number of regimes (autoregressive models) and differ on how the model moves between them. Such models are detailed in the next subsection. We also review how well these models have been found to perform for macroeconomic data in a subsequent subsection.

2.1 Overview of Models

The most common form of parametric nonlinear models used in macroeconomic forecasting are Threshold autoregressive models or Smooth Transition Threshold autoregressive models. These models can be written as

$$y_{t+1} = \phi_0 + \sum_{j=1}^{k} \phi_j y_{t-j+1} + \left(\theta_0 + \sum_{j=1}^{k} \theta_j y_{t-j+1} \right) G(s_t; \gamma, c) + \varepsilon_{t+1}$$

for various specifications of $G(s_t; \gamma, c)$ where s_t is an observed variable and $\{\phi_i, \theta_i\}_{j=0}^{k}$ and $\{\gamma, c\}$ are unknown parameters. The various forms of the model are due to different specifications of the function $G(.)$. If $G(s; \gamma, c) = 1(s_t > c)$ then the model is a threshold autoregressive model (TAR) with the autoregressive model having different coefficients whenever

s_t is above or below c.(γ is undefined here). When s_t is set to a lag of y_t this is known as a self exciting (SETAR) autoregressive model. The TAR model has an abrupt change in the coefficients at $s_t = c$, smooth transition autoregressive (STAR) models avoid this by parameterizing $G(.)$ such that it smoothly moves between zero and one as s_t varies, with the parameters γ and c controlling the speed and position of the transition. Clearly any cumulative distribution function for $G(.)$ will achieve this goal. Most popular in practice have been the logistic distribution, so $G(s_t, \gamma, c) = (1+\exp(-\gamma(s_t-c)))^{-1}$ or the exponential $G(s_t, \gamma, c) = 1 - \exp\{-\gamma(s_t - c)^2\}$. A typical choice for s_t is a lag of y_t, we will use $s_t = y_t$.

Primary justifications for these types of models are that they are theoretically reasonable in the sense that we could imagine that the model is different for a different part of the sample space (for example when s_t is small there is little government intervention, but when it is larger the government intervenes), the model remains stationary for many parameterizations, and the models are not too different from linear ones in that they nest linear models if there are parameters $\{\gamma, c\}$ such that $G(.) = 0$ and hence provide a direct extension of the linear autoregressive model. However such models are usually not grounded in economic theory and are essentially ad hoc.

From a forecasting perspective, the typical approach is to first test for nonlinearity, and on rejection of such a test then use the parametric nonlinear form of ones preference. Whilst tests that are directed at a particular form of $G(.)$ exist (Hansen (19XX) for the TAR model and Franq et. al. (2010) for the smooth models) a more standard approach is to use a version of an LM test for the null hypothesis of linearity (following Saikkonen and Luukkonen (1988)). These are motivated by taking a first or second order expansion of $G(.)$ around γ and evaluating at $\gamma = 0$, which yields either y_t^2 or y_t^2 and y_t^3 (see van Dijk et. al (2002) for a review, also Terasvirta et. al. (2010) for a derivation and history of these tests)). The tests often considered are tests on coefficients from regressions of the form

$$y_{t+1} = \delta_0 + \sum_{j=1}^{k} \delta_j y_{t-j+1} + \sum_{j=1}^{k} \beta_j y_{t-j+1}^2 + u_{1t}$$

or

$$y_{t+1} = \delta_0 + \sum_{j=1}^{k} \delta_j y_{t-j+1} + \sum_{j=1}^{k} \beta_{1j} y_{t-j+1}^2 + \sum_{j=1}^{k} \beta_{2j} y_{t-j+1}^3 + u_{2t}.$$

The test examines $H_0 : \beta_{ij} = 0$ for all i,j vs $H_a : \beta_{ij} \neq 0$ for some {i,j}. Nonlinearity is detected when these tests reject. It is also possible to include cross products of the lagged values for y_t, however we do not include them in this study.

(we have not included Markov Switching so far).

The family of models described above are time varying parameter models where the time variation of the parameters is endogenous. An alternative set of models of time variation, which are essentially nonlinear even though not considered often directly as such, are models where the parameters break either abruptly or follow an exogenous process. The first of these comes under the name of breaks, the second often termed simply time varying parameter models.

- Breaks

-Time Varying parameters

2.2 Previous Empirical Results

In forecasting monthly US data, there is enough evidence to conclude that nonlinear models do not forecast well. Stock and Watson (2001) examined a large set of monthly US macroeconomic variables up to 1996 (the data used in this paper is the same data updated to December 2011, currently for 115 series of their 215). They compared a great number of forecasting methods that included LSTAR models as well as Neural Net Models. In all of their rankings (different horizons, different loss functions), LSTAR models typically performed worse than the best linear models (as did for the most part ANN models) and worse than the baseline autoregressive models. (more detail). They conclude that "most of the nonlinear forecast methods produce worse forecasts than the linear methods". (more detail).

In a similar study for European monthly data series Marcellino (2004) found better evidence for the usefulness of STAR and ANN models as well as random coefficient models (the latter of which did not appear in Stock and Watson's (2001) study as a forecasting method). He reports that for this data STAR and ANN models are preferred for forecasting for 40% of the series. There appears to be no relationship between forecast horizon and the performance of the nonlinear models, although within the linear models the LSTAR performs relatively well at the shortest horizon and the ANN model at the longest forecast horizon. As with most studies Marcellino (2004) finds that when the nonlinear models are dominated by the linear models they are often quite poor predictors (the difference between the performance of the models is large).

In summaries of these results, Terastvirta (2006) and Terasvirta et. al. (2010) express that whilst the STAR methods do not appear to be overall good forecasters, there is still some chance that for very particular series (European unemployment) they might still be worth examining and also that they seem to be useful in forecast combinations. However the general conclusion is that the STAR methods do not perform any better and often worse than linear methods.

One can also examine this from a different perspective — looking at which models are favored by various forecasting literatures for different variables. For example for inflation forecasting a recent survey by Faust and Wright (2012) does not consider STAR models (due to the lack of a successful forecasting literature for this variable) however do include two time varying coefficient models. (Stock and Watson reference).

- Handbook chapters, role of nonlinear models e.g. Wright inflation chapter simple models do better.

- Individual papers, any interest here?

3 Results

A number of models are examined

(a) We have two simple baseline linear models, an AR(1) model and also an AR(p) model where p is chosen using the BIC criterion. The first of these is included as a very parsimonious linear model, the second to capture a reasonable parametric linear model. We compare a number of nonlinear models and models approximating nonlinear models.

(b) The main approximation models are the 'auxiliary' models typically used for testing for nonlinearity. There are two of these models. The first augments the AR(p) model (where p is the same lag length as chosen by the BIC for the AR(p) model) with squared terms for the lags. So if y_{t-k} is in the model then so is y_{t-k}^2. This the the regression often used to test for nonlinearity. The reasoning for including this model is threefold. First, it is somewhat odd to find that these predictors are important in sample (so we reject linearity) only to discard them and put in a particular nonlinear component, as suggested in Van Dijk et. al for STAR modelling. Second, there are quite a number of possible functional forms that are suggested by these terms entering the model, and it is unclear which to include. Finally, if this nonlinear approximation model performs well but the STAR models do not then it provides direct evidence that there is nonlinearity but not of the forms typically considered. Hence it is direct evidence for point XX.

(c) We examine LSTAR and ESTAR models, as well as SETAR models. (issues with estimation etc.)

(d) To be done

The data is an extended version of the Stock and Watson (2001) dataset (only partially updated at this point). The data is updated to October 2011, using the same start dates as the Stock and Watson study. We also employ the same transformations as the original study.

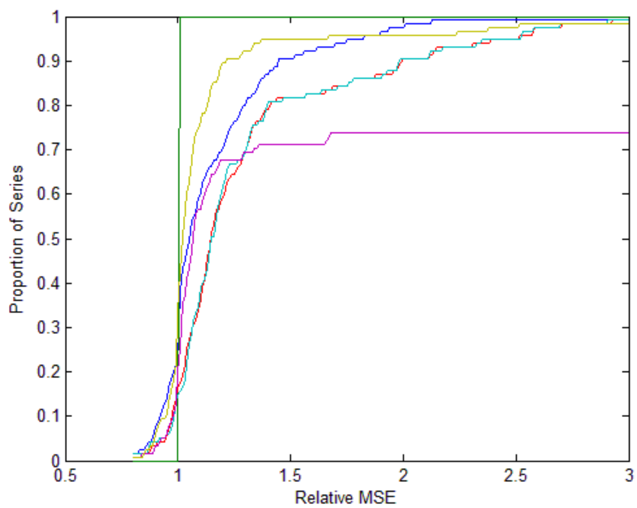
3.1 Do Parametric Models Outperform or Underperform?

Figure 2 gives the performance of the various methods relative to the autoregressive model with a lag selected by BIC (for each series and method we compute the squared loss relative to the autoregressive model, so a number less than one indicates a smaller MSE than the autoregressive model). On the horizontal axis we present various values for this ratio, and graph the percentage of series for each method that has a smaller relative MSE than that number. The point on the vertical axis where these lines cut the vertical line where the x axis equals one gives the proportion of series for which the method outperforms the autoregressive model.

A number of points are evident in this picture. First, the impact of the insanity filter for the LSTAR is very large. Without the insanity filter LSTAR is better than the baseline model 20% of the time but when it is not better it is often a lot worse. For almost 30% of the series the MSE is three or more times that of the baseline model. The reason is the occasional ridiculous forecast (more info on this) However with the insanity filter, the result is quite different. Now for a third of the series LSTAR has a smaller MSE than the baseline. That the curve for the filtered LSTAR is highest indicates that it has more series with lower MSE's than the other methods for a great range of relative MSE's. However it is also clear that when the LSTAR outperforms the baseline, it does so by a small improvement over the baseline whereas when it is worse, it is often much worse. (numbers to get at this).

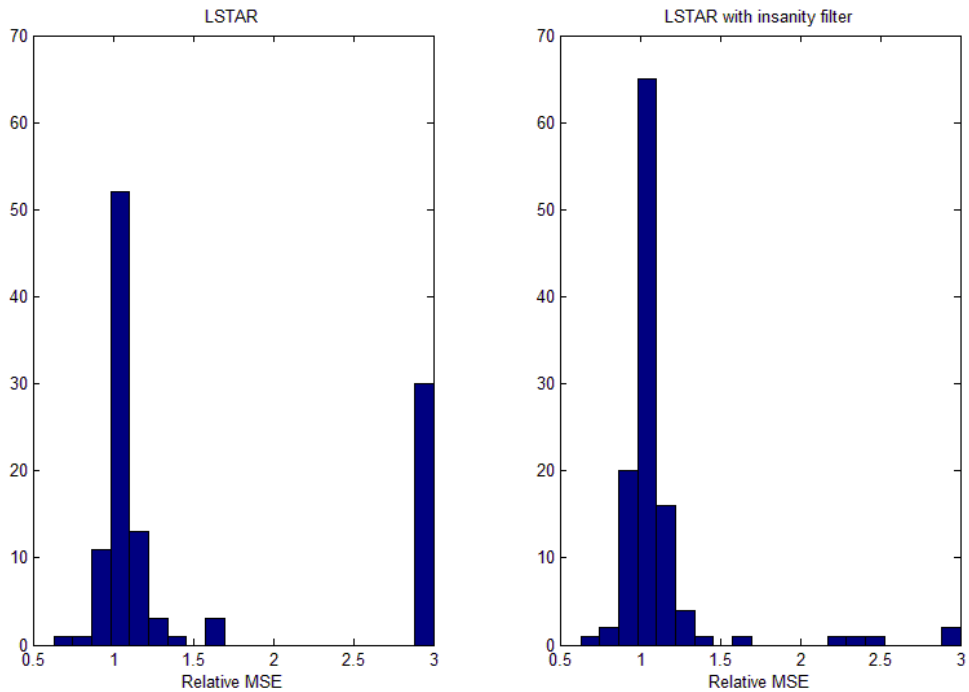
To see more clearly how the LSTAR performs only slightly better when it is better and often much worse when it is not we report a histogram of the relative MSE's of the unfiltered

Figure 1: Performance of Forecast Methods



Notes: We plot the proportion of series with a relative MSE for each of the methods relative to the baseline AR model less than the value given on the x axis. The green line is for the baseline model (so is a step function), yellow is the (insanity filtered) STAR, purple the unfiltered STAR, blue is AR(1) model, red and aqua are the approximate nonlinear models

Figure 2: Relative Performance of STAR models



Notes: The histograms show the average MSE relative to the baseline autoregressive model (top coded to 3).

and insanity filtered methods. In each case we top code the relative MSE so that any value above 3 is set to three. The results are shown in Figure 3. The mode is above one but close to one. However the relative losses are skewed somewhat to larger numbers and the number of models for which there are large gains over the linear model is very small. This is true of the LSTAR using the insanity filter and without.

3.2 Evidence for nonlinearity

A first explanation of the relatively poor general performance of nonlinear forecasting models is that there is no detectable evidence for nonlinearity in the data. Terasvirta et. al. (2010) argues that the Stock and Watson results should be taken with a grain of salt since they average over data that shows evidence for nonlinearity and series that do not (as we do above as well). There are a number of aspects to this — first, are tests for nonlinearity likely to

Figure 3: Power of Tests for nonlinearity

Notes: The highest line is an infeasible upper bound for the test power. The light blue line is the upper bound for power (infeasible test), Green line is the test based on (2), Red line is test based on (3) and dark blue line is the test based on testing $H_0 : \theta_0 = \theta_1 = 0$ when c is set to zero and γ to 2 divided by the standard deviation of y_t.

be useful for finding the types of nonlinearity we are modelling and second do we actually detect the nonlinearity in the data.

The power of tests for nonlinearity when the true model is a STAR model are somewhat odd as a result of the properties of the models themselves. As one might expect, power is generally larger when the coefficients in each regime are more distinct, however for many samples sizes this relationship is not monotonic. However if one is considering stationary models there are upper bounds as to the differences between the autoregressive coefficients. It is also unreasonable to consider massive changes in the constant term, since it is unlikely that long term averages differ greatly. Power is also affected by the baseline model since it affects the variation of the regressors. Power also depends strongly on the value for γ - for small γ the model is close to linear and power is low. However for large γ power is also low as this tends to make the G(.) function such that most of the sample is in a single regime. To our knowledge there are no analytical local power results for these tests.

Without analytical local power results, we resort to Monte Carlo results to show these effects. Figure 1 shows power as we change the difference between the AR coefficients for various values for γ. The Monte Carlo design is a STAR model with a single lag, $c = 0$. Power curves for four tests are examined — the highest curve in each figure is the infeasible test of the null that $\theta_0 = \theta_1 = 0$ where $G(.)$ is known (this is included to provide an upper bound on possible power). In addition power curves for feasible tests discussed above are included. There is not a lot of difference between the power of the feasible tests and the upper bound. There is also not much difference between the tests. However the points mentioned above are clear. Reasonable power requires a relatively large gap between the models in the two regimes. As γ gets large we see that for large enough gaps between the regimes power actually begins to decline again.

It is also the case that tests for linearity have power against a wide set of alternatives, not just the nonlinearity that is being looked at. For example, depending on the nature of the time variation in parameters, these tests can also have power against time varying parameter models even though the variation is not of the form of STAR models.

Empirically, we find evidence of nonlinearity for 47-60% of the sample depending on the test employed. Correcting for heteroskedasticity and possible time dependence these numbers fall a bit, but are still showing evidence of nonlinearity for about half the series. The 60% of rejections arises using the LM test with squared and cubed terms and no standard error corrections. For the noncorrected tests, all three reject for 40% of the series and all three fail to reject for 33% of the series, so there is strong agreement between the three tests employed. (Add in break tests as well). We can compare this with the sample period examined in the Stock and Watson (2001) study, where there are rejections for 32-52% of the series.

Hence there appears to be relatively strong evidence of nonlinearity in many of the series. However the precise form of course is difficult to pin down. Rejections do not imply directly the types of functional forms often chosen in practice (and used here).

Table 1: Tests for Nonlinearity

	Full Dataset			Up to Dec 1996		
	OLS	OLS(White)	OLS(NW)	OLS	OLS(White)	OLS(NW)
Squared terms	47	20	27	32	17	22
Squared and cubed	61	45	52	52	42	43
Approximate logistic	56	41	43	49	32	39

3.3 Functional Form

A second explanation could come down to functional form choices. It may well be that although, as per the previous subsection, there is evidence of nonlinearity that the particular functional forms chosen by forecasters are poor approximations to the true unknown functional forms. Evidence on this arises from various perspectives. In Monte Carlo analysis we can examine how well STAR models perform when the nonlinearity is of an alternate form. Empirically we can examine the relative performance of different functional forms, as well as examine how well the nonlinear methods do in outperforming baseline models when there is evidence of nonlinearity vs when there is little evidence for nonlinearity.

In a criticism of the Stock and Watson (2001) results, Terasvirta et. al. (2010) notes that they, like the numbers presented above, ignore whether or not the tests for nonlinearity reject or not and hence pool results over series that appear nonlinear with those that do not. However for this dataset, rejections of tests for nonlinearity do not appear to be particularly useful for predicting which variables have a better forecasting performance with LSTAR over the baseline. The following table gives the proportions of series for various combinations of rejecting or failing to reject the null hypothesis of linearity and whether or not the relative MSE is above or below one (so for example for 24% of the series, Relative MSE is less than one and we reject linearity).

Table 2: LSTAR Forecasts and Nonlinearity Tests

	Rel. MSE<1	Rel. MSE≥ 1	Test for Linearity
Reject Linearity	0.24	0.37	0.61
Fail to Reject	0.09	0.30	0.39
Rel. MSE	0.33	0.67	

Notes: Four upper left entries are the proportion of series for which the relative MSE and nonlinearity test outcomes are as indicated. Remaining entries are sums of the rows or columns.

The table shows that although most of the times the relative MSE is less than one that linearity is rejected, the chance that this occurs given that linearity is rejected is less than the chance that the relative MSE is greater than one given that linearity is rejected. Indeed, better performance of the nonlinear models over the baseline models conditional on rejection is only slightly larger than the proportion over all the samples. This means that relying on full sample rejections to guide use of LSTAR would not have been a useful approach. On the other hand failure to reject is a strong predictor that the relative MSE is likely to be above one. This is strongly indicating that the functional form is not particularly relevant for these series.

We can also examine if approximations to nonlinear models outperform the chosen nonlinear functional form. The general approach to employing the nonlinear models such as LSTAR is to run an auxiliary regression and test the usefulness of squared and potentially cubed lagged values of the data as predictors. On finding that these predictors are useful, they are then discarded for the chosen nonlinear model. Rather than discard them, we could use these approximate models for forecasting since they have been found to be predictive in sample. An argument against these variables is that the models they imply are typically nonstationary, although from a purely predictive perspective this is not a real problem. Another reason for considering these variables is that rejections using this type of LM test do not imply a particular functional form for the nonlinearity — many different functions are potentially consistent with the rejection. Finally, polynomials in the predictors are one form of a sieve type specification.

Comparing this forecasting approach with LSTAR when the true model is an LSTAR model also gives some indication as to (a) how big the losses are when using an approximate specification and (b) how likely it is that estimation error is playing a role in the forecasting performance of the LSTAR model. We can examine this in Monte Carlo simulations for a variety of sets of parameters specifying the LSTAR models. Results ...

Empirically, we find that LSTAR is for the most part a better predictor than the approximate nonlinear models. We can see this in two ways. First, consider the results presented in Figure 2 earlier. Here we see that the (insanity filtered) LSTAR model tends to have a greater proportion of series with relative MSE at any point compared to the approximate methods (the curve for the LSTAR is higher than that for the approximate methods). It is also clear from this Figure that there is no difference between the approximate method that includes the cubed terms and the one that does not.

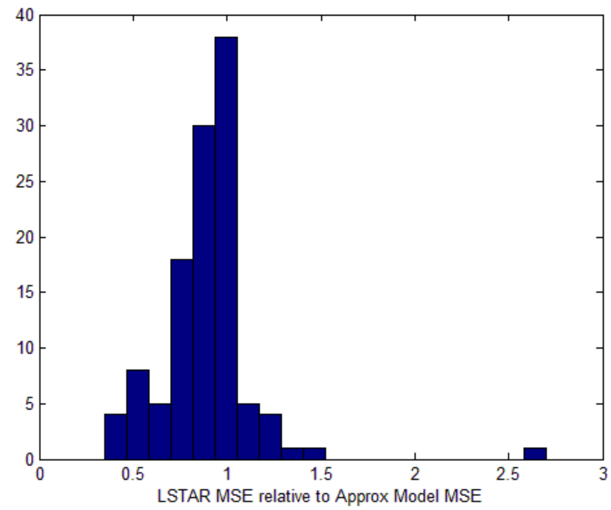
A second way to see the result is presented in Figure 4. Using the baseline of the approximation method with cubed terms included a histogram of the LSTAR MSE relative to this baseline is presented. Clearly LSTAR often improves upon the approximation method, and when it does the improvement is larger than that the loss when it fails to improve.

A similar table as in Table 2 using the cubic approximation model shows that full sample rejection with this statistic has an even smaller predictive power in terms of when the method will outperform the autoregressive model than it does for the LSTAR model.

3.4 Gains too small for effect to make a difference

It is also possible that we have reasonable evidence of nonlinearity in the data, and that also the LSTAR is a reasonable approximation to this nonlinearity, however the gains in MSE from using LSTAR model are just relatively small and overwhelmed by sampling error. In such cases we might see results such as those above (tests for nonlinearity reject) however gains are small and elusive from using nonlinear models.

Figure 4: Relative Performance of Approximate Nonlinear model to STAR model

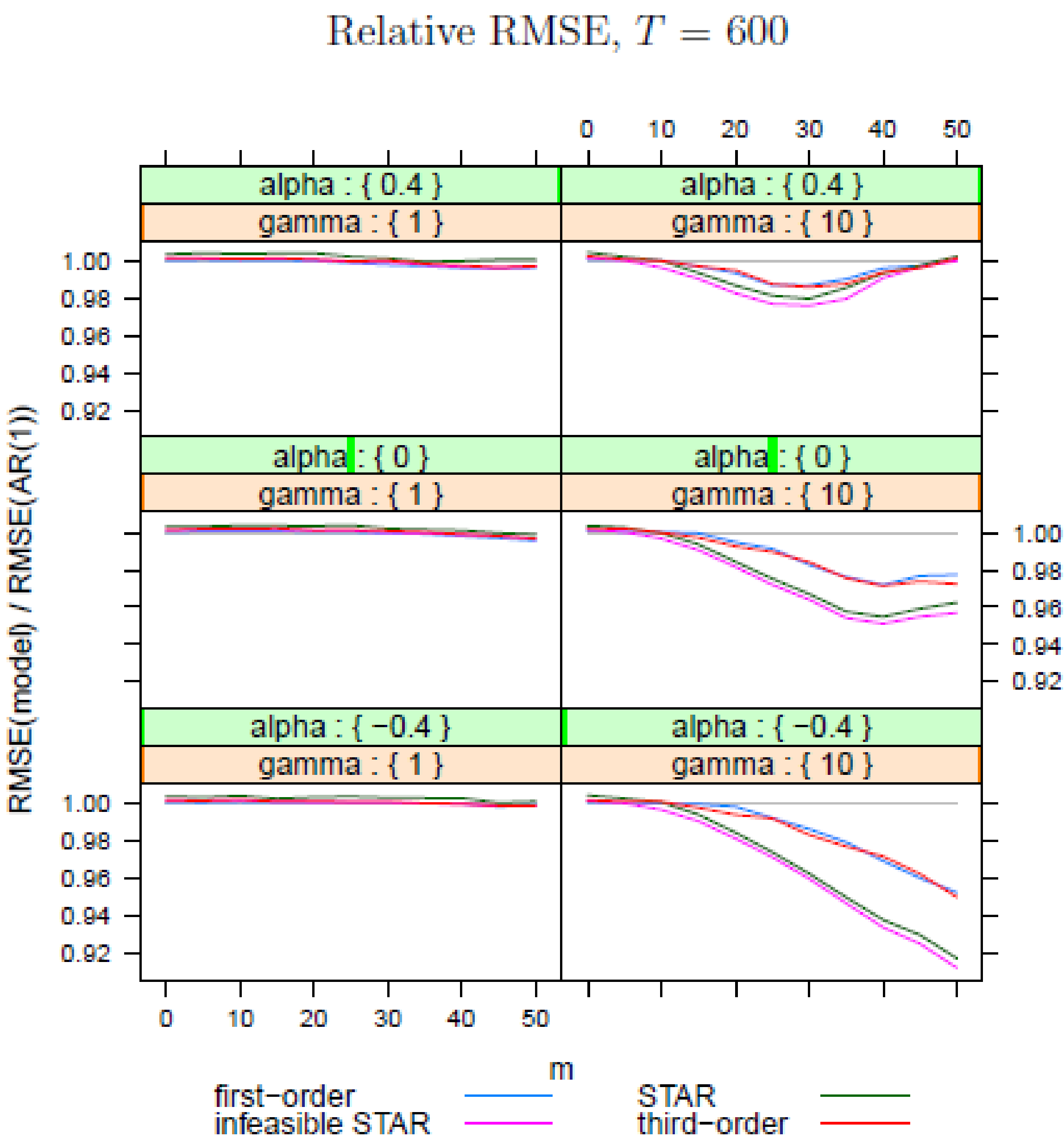


Notes: The histogram shows the relative MSE of the approximate nonlinear method (cubed term version) relative to the (insanity filtered) LSTAR model.

One approach to examining this is to examine the size of the gains in a Monte Carlo experiment. Figure XX reports results comparing the baseline autoregressive model (with a single lag) against the feasible and infeasible STAR models (for the infeasible STAR, we assume that γ and c are set to their known values, so all estimates come from a linear regression), as well as the approximate nonlinear models that use squared and cubed lags as predictors. We set $T = 600$, let $\phi_1 = \alpha$ for various values of α, and $\theta_0 = m/\sqrt{T}$ leaving $\phi_0 = \theta_1 = 0$. So the difference between the autoregressive models in each of the regimes for the STAR model is through differences in the constant term. What we see is that for γ small (here 1), there are basically no differences between the forecasting performance of the various methods (these are the results in each of the left hand side panels). This is despite the fact that tests for nonlinearity here when the STAR model is the true dgp have considerable power. For much larger γ the nonlinear models do offer some improvement over the baseline autoregressive models, with the difference between the various nonlinear models being small.

- require information on the values of γ in the data. Small or large, how do they relate to when the LSTAR does well or poorly

Figure 5: Relative Performance of methods when dgp is STAR



Notes: Monte Carlo results for the relative RMSE for the various nonlinear models relative to the AR(1) model. Numbers less than one indicate improvement over the baseline AR(1) model.

3.5 Estimation Issues

Another potential explanation is that even if the functional form is reasonable, issues with the estimation of the parameters may be such that the models are still outperformed by the more parsimonious linear models. There are a number of theoretical issues. First, if γ is close to zero then even if the STAR model is correctly specified it is nearly not identified and we can expect unusual behavior of the estimates. This may cause the forecasts to be poor. In the case that γ is large there are also problems, since the likelihood becomes flat in some of the parameters. Terasvirta et al. (2010) argue that this is not so much a problem as the model produces similar results however it is also closer to a linear model in the sense that most of the observations are near one regime.

We can examine for reasonable models in Monte Carlo whether or not the STAR models outperform simpler alternatives when the STAR model is correctly specified, and if so how large the difference is. To do this we refer to the experiment in Figure 4 in the previous subsection. The results show that for $T = 600$, estimation issues do not appear to be a major problem for estimating the STAR model. Results for the feasible and infeasible STAR models (which differ only in the estimation of the nonlinear parameters) are almost equivalent for all of the experiments.

The empirical results give some credence to the possibility that the results are due to estimation issues. In the shorter Stock and Watson (2001) sample, STAR models outperform linear models for very few series. In the updated dataset STAR models outperform for 33% of the series. These measures rely on a much longer estimation sample (with a minimum of 493 observations rather than 171 in the earlier work). But they also evaluate over a different period. To compare on a single evaluation period the effect of the estimation results, we employ for the full dataset a rolling regression estimation of the forecasting models rather than the recursive regressions reported above. So for the sample evaluation period as examined in Section 3.1, we change the estimation only in that we use samples of the same length regardless of the available data (dropping earlier observations as we forecast at later dates). The number of observations in each estimation sample for constructing the forecast is 493 observations, still a relatively large sample. The results are qualitatively similar to those obtained using the recursive method, and quantitatively only a little worse. The STAR model (with the insanity filter) now outperforms the baseline autoregressive model for 31%

of the sample. This is some but weak evidence that estimation issues are the reason for the generally poor performance of these nonlinear models.

3.6 Period of Evaluation

A final possible issue might be the period of evaluation. It may well be that there are periods where the nonlinearity is relatively useful for forecasting even though over long periods the relative MSE to the baseline is not too different from one.

Empirical: To be done

4 Conclusions

Whilst in the extended dataset, parametric nonlinear models prove to be better forecasters for monthly macroeconomic variables than studies based on earlier data suggest, overall they are not good in a number of senses. First, they are better than the baseline linear autoregressive model only 33% of the time. Second, the finding of nonlinearity does not seem to predict well for which series the nonlinear models outperform linear models. For the majority of series identified as nonlinear by the test for nonlinearity the parametric nonlinear model does not outperform the baseline autoregressive model. Third, when they are better predictors, the gain is relatively small whereas the losses when they are worse are more often large.

In terms of explaining why the models do not appear to perform well, we identified a number of potential reasons. First, it could be that the data simply does not display any indications that a nonlinear model is appropriate. This argument does not appear to hold up — for about half the series we do indeed reject nonlinearity. Second, it could be that the functional forms are not reasonable. This argument appears to have merit. Tests for nonlinearity were not particularly predictive as to which series the nonlinear model would outperform the linear model. The parametric nonlinear model failed to outperform the linear model for the majority of series that tests suggested would display nonlinear behavior. A third possibility is that these models have little potential for gain. This is true for a great number of parameterizations of the STAR model (still needs work). There is also the possibility that estimation error overwhelms the gain. This does not appear to be going on, the size of the samples appears large enough for relatively good estimation of the models.

5 Appendix

Details regarding the data

The data is an updated version of the same dataset used in Stock and Watson (2001), currently however only 115 of their 215 variables are updated. Transformations of the data (takings logs, differencing) are also as in that paper. The data is updated to November 2011.

Details for Tests for Nonlinearity.

The two popular tests for nonlinearity employed in the paper are based on regressions of y_t on a constant, lags of y_t, squares of lags of y_t and (for the second version) cubes of lags of y_t. Then a Wald test using critical values from the asymptotic χ^2_d distribution are employed where d is the number of parameters set to zero for the test. For the test including squared terms, this is equal to the number of squared terms included, for model including cubed terms it is twice that number. In addition we report results from a test based on the STAR model setting $\gamma = 2/\sigma_y$ and $c = 0$ where σ_y^2 is the variance of the y_t variable (estimated from the data). This test has a χ^2_d limit distribution where d is equal to one plus the number of lags in the regression. For the Monte Carlo results the number of lags is set to one, for the empirical work it is determined by a BIC criterion with a maximum lag of 6.

Details for Forecasting methods employed.

We report two sets of results for the autoregressive baseline model, the first is a simple AR(1) with constant and the second an AR(1) with the lag length selected by BIC. This lag length is then imposed on the other models (so the autoregressive model and the STAR models have the same number of lags always).

References

Dijk, D., T. Terasvirta, and P. Franses (2002): "Smooth transition autoregressive models - a survey of recent developments," *Econometric Reviews*, 21, 1–47.

Francq, C., L. Horvath, and J.-M. Zakoin (2010): "Sup-Tests for linearity ina general nonlinaer AR(1) model," *Econometric Review*, 26, 965–993.

Saikkonen, P., and P. Luukkonen (1988): "Lagrange multiplier tests for testing nonlinearities in time series models," *Scandinavian Journal of Statistics*, 15, 55–68.

Stock, J., and M. Watson (1999): "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive Granger*, ed. by R. Engle, and H. White, pp. 1–44. Oxford University Press, Oxford.

Terasvirta, T. (2006): "Forecasting Economic Variables with Nonlinear Models," in *Handbook of Forecasting Volume 1*, ed. by A. T. G. Elliott, and C. Granger, pp. 413–460. Elsevier, Amsterdam.