ISSN 1745-8587

School of Economics, Mathematics and Statistics

BWPEF 1210

The Predictive Space or If x predicts y, what does y tell us about x?

Donald Robertson
University of Cambridge

Stephen Wright
Birkbeck, University of London

April 2012

Birkbeck Working Papers in Economics & Finance

The Predictive Space

or

If $\mathbf{x}$ predicts y, what does y tell us about $\mathbf{x}$?

Donald Robertson*and Stephen Wright†

April 5, 2012

Abstract

A predictive regression for y_t and a time series representation of the predictors, $\mathbf{x}_t$, together imply a univariate reduced form for y_t. In this paper we work backwards, and ask: if we observe y_t, what do its univariate properties tell us about any $\mathbf{x}_t$ in the "predictive space" consistent with those properties? We provide a mathematical characterisation of the predictive space and certain of its derived properties. We derive both a lower and an upper bound for the R^2 for any predictive regression for y_t. We also show that for some empirically relevant univariate properties of y_t, the entire predictive space can be very tightly constrained. We illustrate using Stock and Watson's (2007) univariate representation of inflation.

*Faculty of Economics, University of Cambridge, donald.robertson@econ.cam.ac.uk

†Corresponding author: Department of Economics, Maths & Statistics Birkbeck College, University of London, Malet Street, London W1E 7HX, UK. s.wright@bbk.ac.uk

1 Introduction

Assume we observe the history of some stationary time series process y_t. If there is some (possibly unobservable) $r \times 1$ vector $\mathbf{x}_t$ that predicts y_{t+1} up to a serially independent error, then the properties of $\mathbf{x}_t$ and of the predictive regression together determine the time series properties of y_t.

In this paper we work backwards, and ask what the observable univariate properties of y_t tell us about the "predictive space": the set of predictive models that are consistent with these properties. If we let $\mathcal{P}_r$ be the parameter space of all possible predictive models with r predictors, then if we observe some set of univariate properties $\mathbf{u}$, the predictive space $\mathbb{P}_{\mathbf{u}}$ is the pre-image of $\mathbf{u}$ in $\mathcal{P}_r$.

The more univariate properties we observe, the more tightly we can identify $\mathbb{P}_{\mathbf{u}}$. But even in the limiting case where we observe the vectors of true AR parameters $\boldsymbol{\lambda}$ and MA parameters $\boldsymbol{\theta}$, the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ is only set-identified. The ARMA parameters do however allow us to identify two points within the space, that turn out to be of particular interest. For any y_t that is an ARMA(p,q), we can always construct one element of $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ by a straightforward rewriting of the ARMA representation. A second element can be constructed from a "nonfundamental" (Lippi & Reichlin, 1994) ARMA representation in which all MA roots are replaced by their reciprocals. Although this is non-feasible as a predictive model for y_{t+1} if we condition only on the history of y_t, we can still derive its *properties* from those of the observable fundamental ARMA. These two particular predictive models provide us with important information about the characteristics of the entire predictive space. We show that we can use them to derive both a lower and an *upper* bound for the predictive R^2 of any predictive regression for y_t.

Previous research (Lippi & Reichlin, 1994; Fernandez-Villaverde et al, 2007,) has shown the link between nonfundamentalness and hidden state variables. Our R^2 bounds provide a new interpretation of nonfundamentalness by showing that the $\mathbf{x}_t$ derived from this particular nonfundamental representation is the best amongst all possible predictor vectors consistent with the history of y_t.

While calculation of the R^2 bounds requires knowledge of the true ARMA parameters, the predictive space can also be derived for a more restricted set of univariate properties. We focus in particular on the predictive space for the variance ratio $V = \sigma_P^2/\sigma_y^2$ of Cochrane (1988), where σ_P^2 is the innovation variance of the unit root (or Beveridge-Nelson, 1981) component in the cumulated process Σy_t. We show that for a commonly used class of predictive models, $V < 1$ imposes an upper bound both on R^2 and ρ, the correlation between predictive errors and innovations to long-horizon forecasts.

We illustrate our analysis with two examples.

Our first (analytical) example derives the predictive space for an ARMA(1,1) y_t process, which constrains the triplet (R^2, ρ, λ), where λ is the AR(1) parameter of the single predictor, x_t. If $V < 1$, but $\lambda > 0$, there is a mis-match between V and $V_{\widehat{y}}$ the variance ratios of y_t and of the predicted value $\widehat{y}_t = \beta x_{t-1}$, since the latter must be above unity. For a strongly persistent predictor, this constrains R^2 to lie in a narrow range quite close to zero, and ρ to be quite close to -1. We note that these univariate features seem to correspond fairly well with a number of observable time series, for example, GDP growth, stock returns and changes in exchange rates.

In our second (empirical) example we use our analysis to shed light on Stock and Watson's (2007) conclusion that inflation has become harder to forecast. We show that in recent data their preferred univariate representation implies that the upper and lower bounds for R^2 are very close to each other, so that there would be very limited scope for even the best possible predictive regression to outperform a univariate model. Furthermore, possible predictors must have low, or negative persistence, and ρ must again be very close to -1: neither of these features is observed in most commonly used observable predictors of inflation.

Our analysis is a reminder that time series prediction does not take place in an informational vacuum. Even a limited amount of information about the history of y_t can impose very tight restrictions on the predictive space that contains $\mathbf{x}_t$. For many observable processes it would be hard for even the best possible predictive regression to do much better than a univariate forecast. It may also imply that any predictive regression that *does* beat the ARMA will be likely to suffer from many of the small sample problems as does ARMA estimation.[1] We conjecture that the observable univariate properties of many of the y_t processes that economists wish to forecast may help to explain why economists appear to have such limited success at forecasting.

The paper is structured as follows. In Section 2 we define the properties of the underlying predictive system, and show how these can be related to both fundamental and nonfundamental ARMA representations of the reduced form process y_t. In Section 3 we formally define the predictive space, and derive our key results. Section 4 presents our illustrative examples, and in Section 5 we address some of the implications of our analysis for empirical research. Section 6 concludes. Appendices provide proofs and derivations, as well as extended analysis of our two examples.

[1] Since Stambaugh (1999) the literature on predictive regressions for financial returns has examined the inference problems that arise from a negative correlation between one-period-ahead prediction errors and innovations to predictions. Our analysis suggests that "Stambaugh Bias" may be a much more widespread feature of predictive models.

2 ARMA Representations and Predictive Systems

2.1 The Fundamental ARMA Representation

Assume that in population y_t admits a finite order ARMA(p,q) representation

$$y_t = \frac{\theta(L)}{\lambda(L)}\varepsilon_t \tag{1}$$

with $\theta(L) = \prod_{i=0}^{q}(1-\theta_i L)$; $\lambda(L) = \prod_{i=0}^{p}(1-\lambda_i L)$; $\theta_0 = \lambda_0 = 0$, where ε_t is IID.[2] Additionally we assume that a) $|\lambda_i| < 1$, y_t is stationary; b) $|\theta_i| < 1$, the representation in (1) is "fundamental" (Lippi & Reichlin, 1994) in terms of the history of y_t (the innovations, ε_t can be constructed from the history of y_t); and c) $\theta_i \neq 0$, $\lambda_j \neq 0$, $\theta_i \neq \lambda_j$, $\forall i > 0, \forall j > 0$, the representation has no redundant parameters. In principle we may have $q = 0$ or $p = 0$, or both, so that y_t may a pure AR or MA process, or may be IID.

The predictive R^2 of this representation satisfies

$$R_F^2(\boldsymbol{\lambda}, \boldsymbol{\theta}) = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2} \tag{2}$$

where $\boldsymbol{\lambda} = (0, \lambda_1 \ldots, \lambda_p)'$, $\boldsymbol{\theta} = (0, \theta_1 \ldots, \theta_q)'$. For an appropriate ordering, as $\lambda_i \to \theta_i$ $\forall i$, $R_F^2 \to 0$; however, given the restrictions above, R_F^2 can only be precisely zero for $q = p = 0$.

2.2 The Minimum Variance Nonfundamental Representation.

For $q > 0$ it is well-known that any representation in which one or more of the non-zero θ_i is replaced by its reciprocal generates identical autocorrelations to (1). Such representations are "nonfundamental" (Lippi & Reichlin, 1994), because their innovations cannot be recovered from the history of y_t;[3] they are therefore non-viable predictive models if we condition only on the history of y_t. However the *properties* of nonfundamental representations can be calculated from the parameters of the fundamental representation in (1).

[2] In Section 2.3 below we state the conditions under which this assumption will be valid.

[3] In this context (but not in general) fundamentalness corresponds to invertibility in terms of the history of y_t.

We focus on the particular nonfundamental representation

$$y_t = \frac{\theta^N(L)}{\lambda(L)}\eta_t \tag{3}$$

where $\theta^N(L) = \prod_{i=1}^{q}\left(1-\theta_i^{-1}L\right)$: for $i>0$ all the θ_i are replaced by their reciprocals, which lie outside the unit circle. It is straightforward to show (see Appendix A.1) that η_t thus defined has the minimum innovation variance amongst all fundamental or nonfundamental representations of the same order, with

$$\sigma_\eta^2 = \sigma_\varepsilon^2 \prod_{i=1}^{q}\theta_i^2 \tag{4}$$

and hence has predictive R^2 given by

$$R_N^2(\boldsymbol{\lambda},\boldsymbol{\theta}) = 1-\left(1-R_F^2(\boldsymbol{\lambda},\boldsymbol{\theta})\right)\prod_{i=1}^{q}\theta_i^2 \tag{5}$$

where R_F^2 is as defined in (2).

2.3 A Predictive System for y_t

The ARMA representation characterises the predictability of y_t conditional upon an information set that is restricted to the history of y_t itself. We now consider what the properties of y_t can tell us about any predictive model consistent with those properties, that may condition also on additional information. We first set out a general predictive system for y_t, and then show how it must relate to the ARMA reduced form.

Write a predictive regression for y_t in terms of a vector $\mathbf{x}_t$ of r predictors (all variables are normalised to have zero means):

$$y_t = \boldsymbol{\beta}'\mathbf{x}_{t-1} + u_t \tag{6}$$

Assumption A1: $\mathbf{x}_t$ *admits a stationary first-order vector autoregressive representation,*

$$\mathbf{x}_t = \Lambda\mathbf{x}_{t-1} + \mathbf{v}_t \tag{7}$$

The disturbances may be non-Gaussian so that (6) and (7) may represent a quite wide range of predictive models.[4]

[4]This framework can represent, at least to an approximation, predictive frameworks as diverse as, for example, vector autoregressions and cointegrating systems; unobserved components models;

Assumption A2: $\mathbf{\Lambda} = diag\left(\lambda_i^*\right)$, $i = 1, \ldots, r, |\lambda_i^*| < 1;\ \forall i;\ \lambda_i^* \neq \lambda_j^*, \forall i \neq j$

This assumption is relatively innocuous, as long as we admit complex x_{it}, since we can diagonalise any underlying VAR representation of the observables (it can also relatively easily be relaxed - see Section 3.7.1).[5] In this representation each of the r predictors is a stationary AR(1) . This may in principle represent cases where the true number of predictors, given by $s = rank\left(E\left(\mathbf{v}_t\mathbf{v}_t'\right)\right)$ is less than r, but each of the s predictors is a higher order ARMA process. (We shall show two particular examples below, in Section 3.1, in which $s = 1$)

Assumption A3: *The disturbances* $\mathbf{w}_t' = \left[\begin{array}{cc} u_t & \mathbf{v}_t' \end{array}\right]$ *are jointly IID, with covariance matrix* Ω.

The specification that the system disturbances $\mathbf{w}_t$ are serially independent, while standard, is crucial. It implies that the x_{it} are sufficient state variables for y_t. Crucially, therefore, (6) is not mis-specified. It also implies that conditional upon $\mathbf{x}_t$, the history of y_t is redundant - thus any predictive information in the history of y_t must already be contained within $\mathbf{x}_t$ (for example, if the $\mathbf{x}_t$ are derived from a diagonalisation of a vector autoregressive representation of y_t and some other set of directly observed predictors - see Appendix A.5). The assumption of a time-invariant distribution is however *not* crucial (see Section 3.7.3); it merely simplifies the exposition.

Considered in complete isolation, i.e., if we did *not* observe the history of y_t there would be only minimal constraints on the parameter space of equations of the form (6) and (7). The β_i could be of any sign, and of any magnitude; the λ_i could lie anywhere within the unit circle, and the only restriction on Ω would be that it be positive semi-definite. But we shall show that if we *do* observe the history of y_t, univariate properties constrain the entire "predictive space" consistent with those properties.

2.4 The Predictive System and the ARMA representation

Substituting from (7) into (6) and rewriting as

$$\det\left(\mathbf{I} - \mathbf{\Lambda}L\right) y_t = \beta' \mathrm{adj}\left(\mathbf{I} - \mathbf{\Lambda}L\right) \mathbf{v}_{t-1} + \det\left(\mathbf{I} - \mathbf{\Lambda}L\right) u_t \tag{8}$$

and, with sufficiently exotic errors, Markov switching models (see Hamilton, 1994, p. 679) . In Section 3.7.3 we discuss possible extensions to cases where the parameters of (6) and (7) may vary over time.

[5] Even if the predicted process y_t and all underlying predictors are real processes, some elements of $\mathbf{x}_t$ may be complex if some pairs of eigenvalues in $\mathbf{\Lambda}$ are complex conjugates.

the right-hand side is an MA(r) composite error process. Conditioning only on the history of y_t, (8) can be rewritten as the "structural" ARMA(r, r),

$$\lambda^*(L)\, y_t = \theta^*(L)\, \varepsilon_t \tag{9}$$

where $\lambda^*(L) \equiv \prod_{i=0}^{r}(1-\lambda_i^* L) \equiv \det(I-\Lambda L)$ is of order r given A2. Letting $\theta^*(L) \equiv \prod_{i=0}^{r}(1-\theta_i^* L)$, the θ_i^* must satisfy r moment conditions, such that the autocorrelations of the quasi-differenced dependent variable $\lambda^*(L)\, y_t$ match those implied by the underlying predictive system in (6) and (7).[6] The θ_i^* are functions of the full set of parameters $\boldsymbol{\Psi}$ of the predictive system, ie we have

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\boldsymbol{\Psi}) \tag{10}$$

where $\boldsymbol{\theta}^* = (\theta_0^*, \cdots, \theta_r^*)'$.

Consistency with the structural ARMA(r, r) representation in (9) therefore requires that the lag polynomials $\lambda(L)$ and $\theta(L)$ in the ARMA(p, q) representation in (1) must satisfy

$$\frac{\theta(L)}{\lambda(L)} = \frac{\theta^*(L)}{\lambda^*(L)} \tag{11}$$

The link between the dimensions of the two ARMA representations, in (1) and (9), is complicated by three possible cases. The AR dimension is reduced by one if one of the λ_i^* is zero;[7] both AR and MA dimensions will be reduced symmetrically if any of the elements of $\boldsymbol{\lambda}^*$ are also elements of $\boldsymbol{\theta}^*$ (so that there is cancellation of factors of the underlying AR and MA polynomials); and the order of the observed MA polynomial $\theta(L)$ will be reduced further if any of the θ_i^* are precisely equal to zero. Hence we have, in general,

$$p = r - \#\{\lambda_i^* = 0\} - \#\left\{\{\theta_i^* \neq 0\} \cap \left\{\lambda_j^* \neq 0\right\}\right\} \tag{12}$$

$$q = p - \#\{\theta_i^* = 0\} \tag{13}$$

3 The Predictive Space

Definition 1 ***(The parameter space of predictive systems with*** r ***predictors)*** *Let* $\mathcal{P}_r$ *be the set that (up to a scaling factor) defines all possible predictive*

[6] For details see Appendix A.2.

[7] Given A2, the λ_i^* are distinct, so at most one can be zero.

systems of the form (6) and (7), for any possible y_t*:*

$$\mathcal{P}_r = \left\{ \boldsymbol{\Psi} = (\boldsymbol{\lambda}^*, \boldsymbol{\omega}, \boldsymbol{\beta}) \subset \mathbb{C}^{\frac{r(r+5)}{2}} : |\lambda_i^*| < 1, \lambda_i^* \neq \lambda_j^*, \forall i \neq j,\ \boldsymbol{\Omega} = \boldsymbol{\Omega}(\boldsymbol{\omega})\ \textit{is p.s.d} \right\}$$

where $\boldsymbol{\lambda}^*$ *is an* $r \times 1$ *vector such that* $\Lambda = diag(\boldsymbol{\lambda}^*)$ *in (7);* $\boldsymbol{\omega}$ *contains the* $r(r+1)/2$ *above-diagonal elements of* $\boldsymbol{\Omega} = E(\mathbf{w}_t \mathbf{w}_t')$*, defined in A3;* $\boldsymbol{\beta}$ *is the* $r \times 1$ *vector of coefficients in (6).*

$\mathcal{P}_r$ defines the general parameter space of all possible r-predictor systems for which the parameters $\boldsymbol{\Psi} = (\boldsymbol{\lambda}^*, \boldsymbol{\omega}, \boldsymbol{\beta})$ satisfy the restrictions required for the predicted process to be stationary, and the innovation covariance matrix to be positive semi-definite. Note that we only need to consider off-diagonal elements of $\boldsymbol{\Omega}$ because the system in (6) and (7) is over-parameterised: ie., we could in principle either let $\boldsymbol{\Omega}$ be a correlation matrix, or let $\boldsymbol{\beta}$ be a vector of ones, without changing R^2 or the autocorrelation function of y_t, which is all that concerns us.

Definition 2 ***(The predictive space for*** $(\boldsymbol{\lambda}, \boldsymbol{\theta})$***)***

Let

$$f : \mathcal{P}_r \to \mathbb{C}^{p+q+2}$$

$$f(\boldsymbol{\Psi}) = (\boldsymbol{\lambda}, \boldsymbol{\theta})$$

where $\boldsymbol{\lambda} = (0, \lambda_1, \ldots, \lambda_p)'$ *and* $\boldsymbol{\theta} = (0, \theta_1, \ldots, \theta_q)'$*. contain the* $q + p$ *non-zero parameters in the ARMA representation (1), with* p *and* q *defined by (12) and (13). The predictive space* $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ *is the pre-image under* f *of* $(\boldsymbol{\lambda}, \boldsymbol{\theta})$*:*

$$f^{-1}(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}} \subset \mathcal{P}_r$$

thus each element of $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ *defines a predictive system consistent with the population ARMA representation, and hence the infinite history of a particular process* y_t*.*

Given the assumption of no redundancy in the ARMA representation, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are identified from the history of y_t; however the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ is in general only set-identified. Note that the definition of $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ is quite general, allowing cases in which p or q, or both, are less than r, or indeed are equal to zero.

We can also define the predictive space in terms of some other (generally restricted) set of univariate properties, denoted $\mathbf{u}$, as follows:

Definition 3 ***(The predictive space for*** $\mathbf{u} \in \mathbb{U}$***)***

Let

$$g : \mathcal{P}_r \to \mathbb{C}^m$$

$$g(\boldsymbol{\Psi}) = \mathbf{u}$$

where $\mathbf{u}$ *is an* $m \times 1$ *vector of univariate properties. The predictive space for* $\mathbf{u}$, $\mathbb{P}_{\mathbf{u}}$ *is the pre-image under* g *of* $\mathbf{u}$

$$g^{-1}(\mathbf{u}) = \mathbb{P}_{\mathbf{u}} \subseteq \mathcal{P}_r$$

and if $\mathbb{U}$ *is a set containing* $\mathbf{u}$,

$$\mathbb{P}_{\mathbb{U}} = \left\{ g^{-1}(\mathbf{u}) : \mathbf{u} \in \mathbb{U} \right\} = \bigcup_{\mathbf{u} \in \mathbb{U}} \mathbb{P}_{\mathbf{u}} \subseteq \mathcal{P}_r$$

thus each element of $\mathbb{P}_{\mathbb{U}}$ *defines a predictive system that maps to a* y_t *process with univariate properties* $\mathbf{u} \in \mathbb{U}$.

In Definition 3 the predictive space is the pre-image of a vector of univariate properties that may in principle be measured precisely, or may satisfy some set of inequalities. In Section 3.6 we shall consider the predictive space $\mathbb{P}_V$ for a particular univariate property, V, the limiting variance ratio (Cochrane, 1988). In Appendix B.2 we also consider, in relation to our empirical example, discussed in Section 4.2, the predictive space $\mathbb{P}_{\mathbb{U}}$ for $u = (V, R_F^2) \in \mathbb{U}$ where the set of feasible values captures sampling uncertainty in finite samples.

Clearly $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ is a special case of Definition 3 and we must have

$$\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}} \subseteq \mathbb{P}_{\mathbf{u}} \subseteq \mathcal{P}_r \tag{14}$$

The more we know about univariate properties, the more restricted is the predictive space; except in the limiting case that $\mathbf{u} = (\boldsymbol{\lambda}, \boldsymbol{\theta})$ (or is some invertible function thereof that uniquely identifies $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$), in which case the predictive space cannot be reduced further.[8] We initially focus on the properties of this minimal set, $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$.

3.1 ARMA Representations as Elements of the Predictive Space

It is straightforward to show that the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ is non-empty. Using (1) and (3), define the $r \times 1$ coefficient vectors $\boldsymbol{\beta}_F = \left(\beta_{F,1}, \ldots, \beta_{F,r}\right)'$; $\boldsymbol{\beta}_N =$

[8] There might also in principle be special cases of $\mathbf{u}$ that do nothing to restrict $\mathbb{P}_{\mathbf{u}}$ relative to $\mathcal{P}_r$, hence the second relationship in (14) may hold with equality: for example, if $\mathbf{u}$ takes the value 1 if y_t is stationary, and 0 otherwise.

$\left(\beta_{N,1},\ldots,\beta_{N,r}\right)'$ that satisfy[9]

$$1+\sum_{i=1}^{r}\frac{\beta_{F,i}L}{1-\lambda_i L} = \frac{\theta\left(L\right)}{\lambda\left(L\right)} \tag{15}$$

$$1+\sum_{i=1}^{r}\frac{\beta_{N,i}L}{1-\lambda_i L} = \frac{\theta^N\left(L\right)}{\lambda\left(L\right)} \tag{16}$$

We can then define two $r\times 1$ vectors of "univariate predictors"

$$\mathbf{x}_t^F = \Lambda\mathbf{x}_{t-1}^F+\mathbf{1}\varepsilon_t \tag{17}$$

$$\mathbf{x}_t^N = \Lambda\mathbf{x}_{t-1}^N+\mathbf{1}\eta_t \tag{18}$$

and hence by construction we have two predictive regressions

$$y_t = \boldsymbol{\beta}_F'\mathbf{x}_{t-1}^F+\varepsilon_t \tag{19}$$

$$y_t = \boldsymbol{\beta}_N'\mathbf{x}_{t-1}^N+\eta_t \tag{20}$$

The predictive regressions in (19) and (20) together with the processes for the two univariate predictor vectors in (17) and (18) are both special cases of the general predictive system in (6) and (7), but with rank 1 covariance matrices, $\Omega_F=\sigma_\varepsilon^2\mathbf{1}\mathbf{1}'$, and $\Omega_N=\sigma_\eta^2\mathbf{1}\mathbf{1}'$.[10] Hence by construction both are elements of $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$. We shall show below that the properties of the two special cases provide us with important information about *all* predictive systems consistent with the history of y_t.

3.2 Simplifying Assumptions

We noted above that the link between r, the number of AR(1) predictors, and the order of the observable ARMA representation, as set out in (12) and (13), is potentially complicated when p or q, or both, are less than r, Since this can only occur by some measure zero combination of the structural parameters, we derive the first of our core results under assumptions that rule such cases out.

Assumption A4: $\lambda_i^*\neq 0,\forall i$

Assumption A5: $\theta_j^*\left(\boldsymbol{\Psi}\right)\neq\lambda_i^*,\forall i,\ \forall j.$

[9] If $p<r$,then $r-p$ elements of $\boldsymbol{\beta}_F$ and $\boldsymbol{\beta}_N$ will be zero; additionally if $p<q$ there will be $p-q$ restrictions on the β_i such that the MA order is matched.

[10] Note that we could also write (19) as $y_t=\boldsymbol{\beta}'\widehat{\mathbf{x}}_{t-1}+\varepsilon_t$; where $\widehat{\mathbf{x}}_t=E\left(\mathbf{x}_t|\left\{y_i\right\}_{i=-\infty}^{t}\right)$ is the optimal estimate of the predictor vector given the single observable y_t and the state estimates update by $\widehat{x}_t=\boldsymbol{\Lambda}\widehat{x}_{t-1}+\mathbf{k}\varepsilon_t$, where $\mathbf{k}$ is a vector of steady-state Kalman gain coefficients (using the Kalman gain definition as in Harvey, 1981). The implied reduced form process for y_t must be identical to the fundamental ARMA representation (Hamilton, 1994) hence we have $\beta_{F,i}=\beta_i k_i$.

Assumption A6: $\theta_i^*(\boldsymbol{\Psi}) \neq 0, \forall i$

Taken together with Assumption A2, that all the λ_i^* in (9) are distinct, Assumptions A4 and A5 imply, from (12), that $p = r$, while, from (13) Assumption A6 implies that $q = p$. Assumptions A4 to A6 thus together imply that $p = q = r$: the AR and MA orders are both equal to the number of AR(1) predictors, and hence $\lambda(L) = \lambda^*(L); \theta(L) = \theta^*(L)$. Since Assumption 5 rules out cancellation of AR and MA polynomials, it also follows that there must be at least some degree of univariate predictability: ie, in (2), $R_F^2 > 0$.

In Section 3.7.1 we discuss the impact of relaxing these assumptions, and also discuss other generalisations to cases where $p \neq q$.

3.3 Univariate bounds for the predictive R^2

The two predictive systems derived in Section 3.1 are not simply special cases: their properties provide bounds that apply to *any* predictive system in the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$.

Proposition 1 ***(Bounds for the Predictive*** R^2*)** Let*

$$R^2(\boldsymbol{\Psi}) = 1 - \sigma_u^2/\sigma_y^2 \tag{21}$$

be the predictive R^2 for a predictive system of the form (6) and (7), with parameters $\boldsymbol{\Psi} \in \mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$. We have

$$R_{\min}^2 \leq R^2(\boldsymbol{\Psi}) \leq R_{\max}^2$$

where under Assumptions A4 to A6,

$$\begin{aligned} R_{\min}^2 &= R_F^2(\boldsymbol{\lambda}, \boldsymbol{\theta}) > 0 \\ R_{\max}^2 &= R_N^2(\boldsymbol{\lambda}, \boldsymbol{\theta}) < 1 \end{aligned}$$

where R_F^2 and R_N^2, defined in (2) and (5), are the predictive R^2s from the ARMA representations in (1) and (3).

Proof. See Appendix A.3. ■

To provide some intuition for this result it is helpful to relate it to the predictive systems in terms of $\mathbf{x}_t^F$ and $\mathbf{x}_t^N$, derived in Section 3.1, that are themselves reparameterisations of the ARMA representations that provide the lower and upper bounds in Proposition 1.

The fundamental univariate predictor vector, $\mathbf{x}_t^F$, defined in (17) is the worst predictor vector (in terms of R^2) consistent with univariate properties; the non-

fundamental predictor vector $\mathbf{x}_t^N$, defined in (18) is the best. While the predictive regression (20) in terms of $\mathbf{x}_t^N$ is not a viable predictive model, since $\mathbf{x}_t^N$ cannot be recovered from the history of y_t, we know that, the better any predictor vector predicts, the more it must resemble $\mathbf{x}_t^N$. Conversely, the less well $\mathbf{x}_t$ predicts, the more it must resemble the fundamental univariate predictor vector $\mathbf{x}_t^F$ (which *can* be recovered from the history of y_t).

The intuitive basis for the lower bound in the inequality in Proposition 1 is quite straightforward. The predictions generated by $\mathbf{x}_t^F$ (or equivalently, by the ARMA representation) condition only on the history of y_t, so they cannot be worsened by increasing the information set to include the true predictor vector. Furthermore, since Assumptions A4 to A6 imply at least some univariate predictability, the lower bound is strictly positive.

The intuition for the upper bound arises from a key (and well-known) feature of any nonfundamental representation: that the innovations cannot be recovered from the history of y_t. In general however they *can* be derived as a linear combination of the history and the *future* of y_t. Since future values of y_t can be expressed in terms of current and future values of the true predictor vector $\mathbf{x}_t$, it follows that any set of nonfundamental innovations must also have predictive power for $\mathbf{x}_t$. Thus far the intuition is relatively straightforward; but the key additional feature of the proof of the proposition follows directly from the distinctive properties of the minimum variance nonfundamental representation. The proof shows that under A4 to A6 the innovations to (3), η_t, can be derived solely from future and current, but *not* lagged values of y_t. As a result it follows straightforwardly that there must be one-way Granger Causality from the nonfundamental predictor vector, $\mathbf{x}_t^N$ to $\mathbf{x}_t$. Hence $\mathbf{x}_t^N$ must always outpredict $\mathbf{x}_t$ except in the limiting case that the two predictor vectors are identical. Furthermore, since, under A6, none of the θ_i^* is equal to zero, from (5) and (4) the upper bound for R^2 is strictly less than unity.

Proposition 1 implicitly constrains the entire predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$. While R^2 maps from $\mathcal{P}_r$, the set of all logically possible predictive systems, to the interval $[0,1]$, under A4 to A6 R^2 maps from $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ to the interval $[R_F^2, R_N^2]$ which is contained strictly within $[0,1]$. Thus univariate properties not only constrain elements of $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ to occur in particular combinations, but may also entirely exclude large parts of the potential parameter space $\mathcal{P}_r$.

3.4 Limiting Cases

Under Assumptions A4 to A6, the upper and lower bounds for R^2 from Proposition 1 lie strictly within $[0,1]$. We also have the following important limiting cases:

Corollary 1 ***(Limiting Cases under A4 to A6)***

a) As $\theta_i \to 0$ *for some* i, $R^2_{\max} = R^2_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) \to 1$

b) As $\theta_i \to \lambda_i \ \forall i$, $R^2_{\min} = R^2_F(\boldsymbol{\theta}, \boldsymbol{\lambda}) \to 0$

c) As θ_i *and* $\lambda_i \to 0 \ \forall i$, $R^2_{\max} = R^2_N(\boldsymbol{\theta}, \boldsymbol{\lambda}) \to 1$ *and* $R^2_{\min} = R^2_F(\boldsymbol{\theta}, \boldsymbol{\lambda}) \to 0$

d) As $|\theta_i| \to 1 \ \forall \ i$, $R^2_{\max} - R^2_{\min} \to 0$, $\boldsymbol{\beta} \to \boldsymbol{\beta}_F$, $\boldsymbol{\Omega} \to \sigma^2_\varepsilon \mathbf{1}\mathbf{1}'$

e) As $\lambda_i \to \theta_i$, $|\theta_i| \to 1 \ \forall \ i$, $R^2_{\max} \to R^2_{\min} \to 0$ $\boldsymbol{\beta} \to \mathbf{0}$, $\boldsymbol{\Omega} \to \sigma^2_\varepsilon \mathbf{1}\mathbf{1}'$

In cases a) to c) the predictive system is tending towards limiting cases that are ruled out by Assumptions A4 to A6. In the neighbourhood of case b) y_t is nearly IID. In Case c), both y_t and all the x_{it} are nearly IID. Note that this is the only limiting case in which the inequality in Proposition 1 is entirely devoid of content. In contrast, in both cases d) and e) the space that R^2 can occupy, and thus the entire Predictive Space, is contracting towards a single point.

We illustrate the limiting cases in Corollary 1 in relation to our analytical example in Section 4.1.

3.5 A Caveat: Imperfect Predictors

Our lower bound for R^2 from Proposition 1 tells us that $\mathbf{x}_t$, the vector of true state variables for y_t must predict at least as well as the univariate predictor vector $\mathbf{x}^F_t$. This does *not* tell us that if we simply run a regression of the form $y_t = \boldsymbol{\gamma}'\mathbf{q}_{t-1} + \omega_t$ for some vector of predictors $\mathbf{q}_t = \mathbf{B}\mathbf{q}_{t-1} + \boldsymbol{\zeta}_t$, that may have some predictive power for y_t, this must imply $R^2_{\mathbf{q}} > R^2_F$. If $\mathbf{q}_t \neq \mathbf{x}_t$, but is some imperfect predictor, correlated with $\mathbf{x}_t$, any such regression will in general be mis-specified, hence ω_t and $\boldsymbol{\zeta}_t$ cannot be jointly IID. However, R^2_F *will* be a lower bound, if information from $\mathbf{q}_t$ is used *efficiently*. Consider some set of state estimates $\widehat{\mathbf{x}}_t = E\left(\mathbf{x}_t | \{\mathbf{q}_i, y_i\}^t_{i=-\infty}\right)$ derived by the Kalman Filter. Under A4 to A6 the resulting vector of state estimates will have the same autoregressive form as the true state variables, with innovations $\widehat{\mathbf{v}}_t$ that are jointly IID with the innovations to the associated predictive regression $y_t = \boldsymbol{\beta}'\widehat{\mathbf{x}}_{t-1} + \widehat{u}_t$,[11] and hence a representation with $\widehat{\mathbf{x}}_t$ is also nested within the general predictive system. If $\mathbf{q}_t$ has any informational content about $\mathbf{x}_t$ independent of the history of y_t, then $R^2_{\widehat{\mathbf{x}}}$ must be strictly greater than R^2_F. If it is *not*, then $\mathbf{q}_t$ must be predictively redundant. It may be correlated with $\mathbf{x}_t$, but this correlation must be solely due to a correlation with the history of y_t, or equivalently with $\mathbf{x}^F_t$.[12]

[11] This is the "Innovations Representation" of Hansen & Sargent, 2007, Chapter 9.

[12] This is implicitly the null hypothesis of no Granger Causality from $\mathbf{q}_t$ as originally formulated in Granger (1969).

3.6 The Predictive Space for the Variance Ratio

Proposition 1 implies that at points corresponding to the upper and lower bounds for R^2, the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ collapses to two distinct points: at $R^2 = R^2_{\min}$, $\boldsymbol{\beta} = \boldsymbol{\beta}_F, \boldsymbol{\Omega} = \boldsymbol{\sigma}^2_\varepsilon \mathbf{1}\mathbf{1}'$, and at $R^2 = R^2_{\max}$, $\boldsymbol{\beta} = \boldsymbol{\beta}_N, \boldsymbol{\Omega} = \boldsymbol{\sigma}^2_\eta \mathbf{1}\mathbf{1}'$. Thus in both limiting cases of the predictive system, $\text{rank}(\Omega) = 1$, so all predictors have innovations that are perfectly correlated both with each other, and with innovations to y_t. Hence for any predictive system sufficiently close to these limiting cases Ω will be close to being singular.

The history of y_t also of course constrains Ω in general, since for all parameter combinations $\boldsymbol{\Psi}$ within the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ all autocorrelations of y_t generated by the predictive system must match those from the ARMA representation. The full set of restrictions requires knowledge of the true ARMA parameters. However, exploiting Definition 3, we can also define the predictive space $\mathbb{P}_{\mathbf{u}}$ in terms of some other set of univariate properties. Our second result shows that knowledge of just one summary univariate property of y_t, the limiting variance ratio of Cochrane (1988) also puts significant constraints on the entire predictive space consistent with that property. It also highlights the implied restrictions on the innovation covariance matrix $\boldsymbol{\Omega}$.

For many, if not most, predictive systems, the stationary predicted variable y_t will itself be the first difference in some underlying process, ie, let

$$y_t = \Delta Y_t \tag{22}$$

where Y_t might for example be the level of real GNP; some measure of real stock prices or cumulative returns; the level of the real exchange rate; the nominal interest rate or inflation rate. Since Cochrane (1988) (and many others) a commonly used univariate statistic is the variance ratio

$$VR(h) = \frac{var\left(\sum_{i=1}^{h} y_{t+i}\right)}{h\sigma_y^2} = \frac{var\left(\Delta_h Y_{t+h}\right)}{h\sigma^2_{\Delta Y}} \tag{23}$$

where if y_t is IID, $VR(h) = 1$ for all h, while if $VR(h)$ is asymptotically decreasing in h the nonstationary process Y_t has a random walk (or Beveridge-Nelson (1981) permanent) component with lower innovation variance than $y_t = \Delta Y_t$ itself.[13] In

[13] For many series the issue of whether the variance ratio slopes downward has been widely debated. Eg for GNP the debate initiated by Cochrane (1988) vs Campbell & Mankiw (1987); and for stock returns the literature arising out of Fama & French (1988) vs Kim et al (1991). Note that Pastor & Stambaugh's (2011) most recent contribution to this literature focuses on the properties of the *conditional* variance ratio, which may slope upwards even when (as their dataset shows) the

such cases Y_t is often referred to as "mean-reverting".[14]

Cochrane (1988) showed that, as a population property, the limiting variance ratio $V = \lim_{h\to\infty} VR(h)$ must be equal to the ratio σ_P^2/σ_y^2, where σ_P^2 is the innovation variance of the random walk component in Y_t. He also showed that this ratio must equal the innovation variance of the random walk component implied by *any* predictive model, whether univariate or multivariate.

The multivariate Beveridge-Nelson(1981)/Stock-Watson (1988) decomposition for Y_t is[15]

$$\begin{aligned} Y_t &= Y_t^P + Y_t^T \qquad (24) \\ &\text{where} \\ Y_t^P &= \lim_{h\to\infty} E_t Y_{t+h} = Y_t + \lim_{h\to\infty} E_t \Delta_h Y_{t+h} \qquad (25) \end{aligned}$$

It is straightforward to show (see Appendix A.4) that the predictive system in (6) and (7) implies

$$Y_t^P = \frac{u_t + \boldsymbol{\delta}'\mathbf{v}_t}{1-L} \qquad (26)$$

where $\boldsymbol{\delta}' = \boldsymbol{\beta}' [I - \Lambda]^{-1}$ is a vector of "long-run multipliers". The innovation to Y_t^P in (26) can be split conceptually into two components, corresponding to the partition of Y_t^P in (25). The first is the prediction error in the predictive regression (6) (which, given that it is IID, will in expectation persist indefinitely in Y_t); the second is the innovation to expected growth in Y_t over an infinite horizon, which is a linear combination of innovations to the x_{it}. We shall denote the correlation between these two components as the "Beveridge Nelson Correlation",[16] defined as

$$\rho = corr\left(u_t, \boldsymbol{\delta}'\mathbf{v}_t\right) \qquad (27)$$

Since the limiting variance ratio V must be identical to the ratio σ_P^2/σ_y^2 from any predictive system, knowing V restricts the entire predictive space, $\mathbb{P}_V$ consistent with a given value of V (i.e., setting $\mathbf{u} = V = g(\Psi)$ in Definition 3). The following result expresses this restriction in terms of three summary features of any predictive

unconditional ratio, which we consider, slopes downwards.

[14] This term is actually a misnomer except in the special case where $VR(h)$ asymptotes to zero, implying that Y_t, rather than y_t is the underlying stationary process, and hence has been over-differenced. For *any* y_t process that has some serial correlation structure, Y_t will have a mean-reverting transitory component, whatever the slope of the variance ratio.

[15] We neglect constants and hence deterministic growth in Y_t. For a more general definition see, eg, Garratt, Robertson & Wright, 2006.

[16] With apologies to Stock & Watson (1988) who generalised the original Beveridge-Nelson (1981) univariate decomposition to the multivariate version used here.

system.

Proposition 2 ***(The Predictive Space for the Variance Ratio)*** *Let V be the limiting variance ratio of Cochrane(1988), defined by*

$$V = \lim_{h\to\infty} VR(h) \equiv 1 + 2\sum_{i}^{\infty} corr(y_t, y_{t-i}) \tag{28}$$

For any given V, using Definition 3, the parameters $\boldsymbol{\Psi} = (\boldsymbol{\lambda}^*, \boldsymbol{\beta}, \boldsymbol{\omega}) \in \mathbb{P}_V$ must satisfy

$$\begin{aligned} g(\boldsymbol{\Psi}) &= V \\ \text{where } g(\boldsymbol{\Psi}) &= 1 + R^2(V_{\widehat{y}} - 1) + 2\rho\sqrt{V_{\widehat{y}} R^2(1 - R^2)} \end{aligned} \tag{29}$$

where $R^2(\boldsymbol{\Psi})$ is the predictive R^2 from (6); $\rho(\boldsymbol{\Psi}) = corr(u_t, \boldsymbol{\delta}'\mathbf{v}_t)$, is the Beveridge-Nelson Correlation, defined in (27); and $V_{\widehat{y}}(\boldsymbol{\Psi})$ is the variance ratio of the predicted value $\widehat{y}_t \equiv \boldsymbol{\beta}'\mathbf{x}_{t-1}$, calculated by replacing y_t with $\widehat{y}_t$ in (28).

Proof. See Appendix A.4. ■

The definition of the limiting variance ratio V in (28) shows that if V differs from unity the sum of all autocorrelations of y_t must differ from zero. The counterpart to this for any predictive system within $\mathbb{P}_V$ is that this serial correlation in y_t must come from somewhere. Equation (29) in the proposition shows that this imposes a restriction on the triplet $(R^2, \rho, V_{\widehat{y}})$ that must be satisfied for any predictive model consistent with a y_t process with limiting variance ratio V.

One key feature of this restriction arises by inspection of (29). Trivially $R^2 = 1$ must imply $\widehat{y}_t = y_t$ and hence $V_{\widehat{y}}$, the variance ratio of the predictions, must equal V. But as a direct implication, for any predictive model for which $V_{\widehat{y}}$ does *not* equal V, R^2 must be bounded away from unity. Since $V \neq 1$ must also imply some univariate predictability, it follows that for for any y_t process with $V \neq 1$, and any predictive system with $V_{\widehat{y}} \neq V$, R^2 must be bounded strictly within $[0, 1]$.

In contrast to Proposition 1, which showed that R^2 bounds can be derived directly from the parameters of the univariate representation, the more limited univariate information that defines $\mathbb{P}_V$ means that the implied R^2 bounds also depend on $V_{\widehat{y}}$, and hence in general on the properties of the predictive system. However, two corollaries to Proposition 2 show that certain general features of predictive systems, or of the reduced form process itself, put a lower bound on $V_{\widehat{y}}$. For the case $V < 1$ this in turn bounds both R^2 and ρ.

Corollary 2 *Let* $\rho_{ij} = corr\left(\beta_i v_{it}, \beta_j v_{jt}\right)$. *If* $\lambda_i \geq 0$ *for all* i; $\rho_{ij} \geq 0\ \forall j \neq i$ *then* $V_{\widehat{y}} \geq 1$, *and hence if* $V < 1$,

$$\begin{aligned} \rho &\leq V - 1 < 0 \\ R^2 &\leq \frac{1 + \sqrt{V\left(2 - V\right)}}{2} < 1 \end{aligned}$$

Corollary 2 is of particular interest given the common *a priori* restriction that all λ_i are non-negative, and innovations to predictors are orthogonal. Any such predictive system must generate predictions with an upward-sloping variance ratio ($V_{\widehat{y}} > 1$). In such cases, by inspection of (29), there *must* be a negative correlation of prediction errors with long-run forecasts ($\rho < 0$), and both ρ and R^2 have upper bounds that depend only on V. The proof of both bounds also shows that, the higher is $V_{\widehat{y}}$, the more tightly both ρ and R^2 are constrained.

Corollary 3 *Let* y_t *be an ARMA*$(1,1)$ *with AR and MA parameters* λ *and* θ. *Under A4 to A6,* $r = 1$, *and hence* $V_{\widehat{y}} = (1 + \lambda) / (1 - \lambda)$ *is a strictly univariate property. Thus in this case, for* $V \neq 1, V_{\widehat{y}} \neq V$, *observing* $(V, V_{\widehat{y}})$ *is equivalent to observing* (λ, θ), *and hence the implied* R^2 *bounds are identical to those given by Proposition 1. Furthemore, if* $V < 1$ *and* $V_{\widehat{y}}(\lambda) > 1$ *then*

$$\rho \leq \rho_{\max} = -\sqrt{\frac{(1 - V)\left(V_{\widehat{y}} - V\right)}{V_{\widehat{y}}}} < 0$$

Corollary 3 provides an illustration of how, as discussed in relation to Definitions 2 and 3, the more properties we observe (or assume) for the process y_t, the more tightly we can identify the predictive space. For the $r = 1$ case, if we only observe V, then the restriction on the triplet $(R^2, \rho, V_{\widehat{y}})$ in Proposition 2 is equivalent to a restriction on R^2, ρ and λ. If $V < 1$, and we impose the *a priori* restriction that λ is positive, hence $V_{\widehat{y}} > 1$, then we have a special case of Corollary 2, which puts (fairly weak) upper bounds on both ρ and R^2. But if we observe both V and $V_{\widehat{y}}$ this is equivalent to observing the ARMA parameters (λ, θ), and we can derive bounds for both R^2, as in Proposition 1, and ρ, as in Corollary 3, that define a strictly narrower space than those given in Corollary 2. We discuss this special case further in Section 4.1.

3.7 Extensions and Generalisations

3.7.1 ARMA(p,q) reduced forms with $p \neq q = r$

It is relatively easy to accommodate cases where $p \neq q$. The common feature of these extensions to our framework is however that q, the moving average order, is always equal to r, the number of AR(1) predictors.

It is straightforward to relax Assumptions A2 and A4, which together imply that $\mathbf{\Lambda}$, the autoregressive matrix of the predictors, is diagonal, with distinct non-zero elements. If there is some set of observable underlying predictors $\mathbf{z}_t = \mathbf{A}\mathbf{z}_{t-1} + \mathbf{w}_t$ then we can always write $\mathbf{A} = \mathbf{S}^{-1}\mathbf{\Lambda}\mathbf{S}$, where $\mathbf{\Lambda}$ takes the Jordan Normal Form, and define $\mathbf{x}_t = \mathbf{S}\mathbf{z}_t$. Assumptions A2 and A4 are therefore just restrictions on the eigenvalues of $\mathbf{A}$. The more general form, with some zeroes on the diagonal or ones on the super-diagonal of $\mathbf{\Lambda}$, arises if $\mathbf{A}$ has zero or repeated eigenvalues. The former case, which implies that some of the x_{it} have no persistence, lowers the order of $\det(I - \mathbf{\Lambda}L)$ and hence p, and thus may result in $p < q = r$ under A5 and A6. Thus q still reveals the number of predictors. The latter case does not change p, and has no impact on our results other than to complicate the algebra.

It is also straightforward to generalise to cases with $p > q$. If y_t is an ARMA(p,q) with $p > q$, this can be reduced to an ARMA(q,q) for some process $\phi(L)\, y_t$, with the same univariate innovations, where $\phi(L)$ is a $(p-q)$th order polynomial. Having done so, under A4 to A6, q reveals r, the number of AR(1) predictors in a predictive system for $\phi(L)\, y_t$, but by implication also for y_t itself. The definitions of R_F^2 and R_N^2 for y_t remain unchanged, since they apply for any p and q. However, they are now the lower and upper bounds for R^2 in a predictive regression for y_t that conditions on q predictors and $p-q$ lags of y_t.[17] We provide an illustrative example in Appendix B.1, in which y_t is an ARMA$(2,1)$, which implies that the predictive regression is one of the equations in a bivariate VAR(1).

3.7.2 Relaxing Assumptions A4 to A6

While it is convenient for our analysis to maintain A5 (no cancellation of AR and MA terms) and A4 and A6 (no AR or MA terms precisely equal to zero) in deriving Proposition 1, not much actually hinges on this. The limiting cases of Proposition 1 described in Corollary 1 describe the nature of predictive systems that are close to violating A4 to A6. The definitions of the Predictive Space do not rely on these assumptions, and hold for any values of r, p and q, as does Proposition 2.

[17] Appendix A.5 derives a special case of particular interest, if y_t is an element of a vector autoregression,

Furthermore, as our discussion of empirical implications (see in particular Section 5.2) makes clear, arguably the key issue is not whether predictive systems actually violate these assumptions, but whether they are sufficiently close to doing so that in a finite dataset it may be impossible to tell.[18]

3.7.3 Time-varying parameters

In general, if any of the parameters in the predictive model (including elements of $\boldsymbol{\Omega}$) are non-constant over time, this will translate into time variation in the parameters of the univariate representation for y_t. However, this does not of itself detract from the key insights that our analysis provides: it merely complicates the algebra. The proof of the R^2 bounds in Proposition 1, for example, relies on the assumption of independence of the underlying innovations; not on their having a time-invariant distribution, nor on the constancy of $\boldsymbol{\lambda}$ or $\boldsymbol{\beta}$. Thus even with time-varying parameters there will still be upper and lower bounds for the predictive error variance, but these would themselves be derived from time varying ARMA representations. We discuss this issue further in the context of our empirical application in Section 6.

4 Illustrative Examples

4.1 An Analytical Example: The Predictive Space for an ARMA(1,1)

Assume that y_t has a fundamental ARMA(1,1) representation with a white noise innovation

$$y_t = \left(\frac{1-\theta L}{1-\lambda L}\right)\varepsilon_t \tag{30}$$

Under Assumptions A1 to A6 this implies that there must be some underlying predictive system with $r = 1$, of the general form

$$y_t = \beta x_{t-1} + u_t \tag{31}$$

$$x_t = \lambda x_{t-1} + v_t \tag{32}$$

[18] Lippi & Reichlin (1994) analyse the implications of violations of A5, which lead to "nonbasic" nonfundamental representations, for which some of the parameters cannot be recovered even in an infinite sample. While Lippi & Reichin assert that the nonbasic property is "not likely to occur in models based on economic theory" (Lippi & Reichlin, 1994, p 315), Baxter, Graham & Wright (2011) show that it will arise naturally in models where agents have imperfect information. Nonetheless, even in cases where the nonbasic property may arise, there will always be a basic nonfundamental representation of the form in (3) the properties of which can be derived from the history of y_t.

where x_t is scalar and $\mathbf{w}_t = (u_t, v_t)'$ is vector white noise.

This very simple framework provides a wide range of insights into more general cases. It is also a framework that has been extensively employed in empirical prediction problems, particularly in the literature on predictive return regressions in empirical finance. One notable feature is that here the Beveridge-Nelson Correlation $\rho = corr\left(u_t, \frac{\beta}{1-\lambda}v_t\right)$ is identical to the "Stambaugh Correlation" $\rho_S = corr\left(u_t, \beta v_t\right)$, since forecasts of y_t at all horizons are simply scalings of one-period-ahead forecasts. As noted in the introduction, the Stambaugh Correlation has, since Stambaugh (1999), been the focus of a a large literature on inference problems in predictive regressions.

We can straightforwardly reparameterise the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ to be the set of all possible values of the triplet (R^2, ρ, λ) that generate the reduced form (30).[19] The predictive space $\mathbb{P}_{\lambda,\theta}$ is then the pre-image in $\mathbb{R}^3$ of the ARMA coefficients (λ, θ) :

$$\mathbb{P}_{\lambda,\theta} = \left\{\left(R^2, \rho, \lambda\right) : f\left(R^2, \rho, \lambda\right) = (\lambda, \theta)\right\} \subseteq \mathcal{P}_1 \tag{33}$$

The result in Propositions 1 and 2 place significant restrictions on all three elements in $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$.

Most straightforwardly, λ, the AR parameter of the predictor in (32), must be equal to the AR parameter in the ARMA representation (30).

In the case of R^2, the lower and upper bounds in Proposition 1 are, given $r = p = q = 1$,

$$R^2_{\min} = R^2_F(\lambda, \theta) = \frac{(\theta - \lambda)^2}{1 - \lambda^2 + (\theta - \lambda)^2} \tag{34}$$

$$R_{\max} = R^2_N(\lambda, \theta) = 1 - \theta^2\left(1 - R^2_F(\lambda, \theta)\right) \tag{35}$$

These formulae provide simple illustrations of each of the limiting cases described by Corollary 1.

If θ is close to zero (hence y_t is close to an AR(1)) then $R^2_{\max}$ is close to unity (illustrating case a) of Corollary 1). If θ is close to λ, $R^2_{\min} = R^2_F$ is close to zero, so y_t is close to being white noise (illustrating case b). But only if θ *and* λ are sufficiently close to zero (implying that *both* y_t and the single predictor x_t are close to white noise), does the inequality for R^2 open up to include the entire range from zero to unity (illustrating case c).

In marked contrast, as $|\theta|$ tends to unity the space that R^2 can inhabit (and indeed the entire predictive space) collapses to a single point (illustrating case d) of

[19] See Appendix A.6 for full derivation.

Corollary 1. Thus any ARMA(1, 1) process with high $|\theta|$ will have a very limited range of values of R^2 for a single predictor model: ie, there is very little scope for any predictive model to outperform the ARMA. If, additionally, $\theta \approx \lambda$, ie, y_t is close to univariate white noise, then the upper bound for R^2 will also be quite close to zero, so that both in relative and absolute terms all predictive models must predict badly (illustrating the final limiting case, e) of Corollary 1).

A further convenient feature of the ARMA(1,1) is that, as Corollary 3 shows, observing λ and θ is equivalent to observing the limiting variance ratios V and $V_{\widehat{y}} = \frac{1+\lambda}{1-\lambda}$ so that for this special case the R^2 bounds implied by Propositions 1 and 2 are identical. It is also straightforward to show that

$$\theta > \lambda > 0 \Longleftrightarrow V < 1 \tag{36}$$

thus under this condition we can also explicitly derive an upper bound for ρ, $\rho_{\max}$, as given in Corollary 3 which is strictly negative. Since both V and $V_{\widehat{y}}$ can be written as functions of (λ, θ) we can also write $\rho_{\max}$ explicitly in terms of the ARMA parameters, giving

$$\theta > \lambda > 0 \Longleftrightarrow \rho \leq \rho_{\max} = -\left(\frac{2\sqrt{(\theta-\lambda)(1-\theta\lambda)\theta}}{1-\lambda^2+(\theta-\lambda)^2}\right) < 0 \tag{37}$$

Thus for all ARMA(1,1) processes with $\theta > \lambda > 0$, and hence a declining variance ratio, *any* single predictor system for such a process must have errors in forecasting y_t that are negatively correlated with revisions to forecasts of y_{t+k}.for $k \geq 1$.[20]

Figure 1[21] illustrates the predictive space for two processes, y_{1t} and y_{2t}. For both we set the AR parameter λ to be 0.8. Using A4 to A6, this must be the AR parameter of the single predictor in (31), which is therefore strongly persistent. For y_{1t} we set $\theta = 0.9$, while for y_{2t}, we set $\theta = 0.7$. Given this parameterisation both have identical, and very limited, univariate predictability (from (34), both have $R^2_F = 0.027$); but, from (36) y_{1t} has a variance ratio that monotonically declines, while that for y_{2t} monotonically increases (with asymptotes of $V_1 = 0.24$ and $V_2 = 2.2$ respectively). Thus the two cumulated processes $Y_{1t} = (1-L)^{-1}y_{1t}$ and $Y_{2t} = (1-L)^{-1}y_{2t}$ have permanent components with relatively low, and relatively high innovation variances, respectively.

With $p = q = r$ the predictive space $\mathbb{P}_{\lambda,\theta}$ is a curve in $\mathbb{R}^3$. Given that for both processes we fix $\lambda = 0.8$, in Figure 1 we plot $\mathbb{P}_{\lambda,\theta}$ for y_{1t} and y_{2t}, as lines in $\mathbb{R}^2$ that

[20] By implication the very common identification assumption in simple state space representations that $\rho = 0$ must generate $V > 1$ if $\lambda > 0$.

[21] All tables and figures are appended to the paper.

satisfy (33), given $\lambda = 0.8$.

The leftmost point of each line in Figure 1 pins down $R^2_{\min} = R^2_F = 0.027$, which is identical for both processes. The rightmost point pins down $R^2_{\max} = R^2_N$. The lower value of θ for y_{2t} implies a higher upper bound for R^2, with $R^2_{\max} = 1 - 0.7^2 (1 - .027) = 0.52$, compared to $1 - 0.9^2 (1 - .027) = 0.21$ for y_{1t}. At both extremes, the Beveridge-Nelson-Stambaugh correlation, $\rho = \pm 1$: i.e., u_t and v_t, in (31) and (32) are perfectly correlated. For y_{1t}, with $\theta > \lambda$, from Corollary 3, ρ must be negative throughout, with a turning point at $\rho = \rho_{\max}$ as given in (37). In contrast for y_{2t}, ρ is monotonically decreasing in R^2.

Figure 1 also illustrates the feature we noted in our discussion of Proposition 2: that a negative Beveridge-Nelson-Stambaugh Correlation is *not* just a property of y_t processes with $V < 1$. The intuition for this feature can be related straighforwardly to the general restrictions on the predictive space implied by (29) in Proposition 2. In this special case the predicted value $\widehat{y}_t = \beta x_{t-1}$ is an AR(1) with high persistence. As a result for both processes, from Corollary 3 the variance ratio for $\widehat{y}_t$ is very high: $V_{\widehat{y}} = \frac{1+\lambda}{1-\lambda} = 9$. Since even for y_{2t} this is much higher than the limiting variance ratio for the process itself, from (29), for sufficiently high R^2, the predictive system can only match the lower V if $\rho < 0$.

Figure 1 shows that different univariate properties can have distinctly different implications for the range of possible values each parameter can take within the predictive space. If $\theta < \lambda$ (the variance ratio slope upwards), the range of possible values of R^2 is larger than if $\theta > \lambda$; and (for some value of R^2) ρ can lie anywhere in $[-1, 1]$. In contrast if $\theta > \lambda$ (the variance ratio slopes downwards) all possible values of ρ lie in a narrow range close to -1, and the gap between $R^2_{\min}$ and $R^2_{\max}$ is quite narrow. Thus for this case the range of the inverse function $f^{-1}(\lambda, \theta)$ that, from (33), defines $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ lies element-by-element strictly within $\mathcal{P}_1$: ie, λ is pinned down directly by ARMA properties, $R^2 \in [R^2_F, R^2_N]$ and $\rho \in [-1, \rho_{\max}]$.

Exploiting Definition 3 we can also use Figure 1 to illustrate the properties of the predictive space for single predictor models that generate values of V and R^2_F within a particular range. To simplify we again fix the AR parameter of the predictor, $\lambda = 0.8$. It is straightforward to show that the predictive space satisfying $V \leq 0.24$ (the limiting variance ratio for y_{1t}) is then simply the area below the lower of the two lines shown in Figure 1. As $V \to 0$ the predictive space contracts towards a single point where $\rho = -1$, and $R^2_F(\lambda, \theta) = R^2_{\min} = R^2_{\max} = 0.1$. Since $R^2_{\min} = R^2_F(\lambda, \theta)$ must be strictly less than 0.1 for any $V \in (0, 24)$ the inequality on V also imposes at most a very limited degree of univariate predictability.

For a given value of λ, using Definition 3, the area under the curve thus defines

$\mathbb{P}_{\mathbb{U}} \subset \mathcal{P}_1$ for $\mathbb{U} = \{(V, R_F^2) : V \in [0, 0.24], R_F^2 \in [0.027, 0.1]\}$. All predictive models within this space imply low univariate predictability ($R_F^2 \leq 0.1$); upper and lower bounds for R^2 that are fairly close together ($R^2_{\max} - R^2_{\min} \leq 0.18$); and a Beveridge-Nelson-Stambaugh correlation very close to -1 ($\rho \leq \rho_{\max} = -0.86$). Thus the properties of this space illustrate that for any ARMA(1,1) y_t process, the lower is the limiting variance ratio V, the more closely any single predictor model with a persistent predictor must resemble a univariate model.

4.2 An empirical application: Stock & Watson's (2007) model of inflation

Stock & Watson (2007, henceforth SW) show that for a range of widely used predictors of inflation there is little evidence that it is possible to out-forecast a very simple univariate model, particularly in recent data. This application provides a simple and powerful illustration of our analysis, which in turn sheds light on SW's results.

SW's preferred univariate representation uses unobserved components estimation, but is equivalent to an IMA(1,1) model of inflation, π_t. Hence, if we let $y_t = \Delta\pi_t$, this is a special case of the ARMA(1,1) example in Section 4.1, setting $\lambda = 0$.[22]

This immediately provides a crucial piece of information about the predictive space for inflation: that, at least in a single predictor model, any such predictor must be (or be indistinguishable from) an IID process. This puts very strong restrictions on candidate predictors.

The remaining two elements of the predictive space (parameterised, as in the example of Section 4.1, as (R^2, ρ, λ), depend on the estimated value of θ. The first two columns of Table 1 show SW's estimates of θ in two subsamples. Shown below these estimates are the implied values of $R^2_{\min}$ and $R^2_{\max}$ using (34) and (35) as well as of $\rho_{\max}$, using (37), setting $\lambda = 0$.[23]

The two sub-samples show a distinct contrast. In the earlier sample, $\widehat{\theta}$ is relatively close to zero, with the result that the implied range of values of R^2 for any predictor is barely constrained, with $R^2_{\min}$ very close to zero, and $R^2_{\max}$ very close to unity. This is close to the limiting case c) of Corollary 1. However, even in this sample, univariate properties still impose restrictions on the predictive space. The final

[22] The unobserved components estimation technique also constrains θ to lie in $(0, 1)$

[23] Standard errors are derived using the delta method. Note that here we treat it as known that $\lambda = 0$, and hence only consider the impact of sampling variation in $\widehat{\theta}$. In Appendix B.2 we consider the case where the true value of λ differs from zero; but we show that this does little to change the conclusions presented here.

row of Table 1 also shows that, using (37), $\rho_{\max}$, is around -0.5 : so any predictor consistent with univariate properties must not only be (or be indistinguishable from) an IID process (given $\lambda = 0$), but must also be quite strongly negatively correlated with the prediction error for inflation.

The restriction on the predictive space for the variance ratio, $\mathbb{P}_V$, given in (29) in Proposition 2 helps to provide intuition. Here, with an IID predictor, the limiting variance ratio of the predictions, $V_{\widehat{y}}$ is unity, while the limiting variance ratio for y_t itself is given by $V(\theta) = (1-\theta)^2 / (1+\theta^2) < 1$. Thus to generate enough negative serial correlation in y_t to match the point estimate $V\left(\widehat{\theta}\right) = 0.49$ in this sample, ρ must be sufficiently negative.

In the second of the two samples, $\widehat{\theta}$ is distinctly closer to unity, and hence we are closer to the limiting case d) of Corollary 1, so that the predictive space is more constrained. The range of possible values of R^2 is considerably narrower, and the implied value of $\rho_{\max}$ is now much closer to -1.

This contraction of the predictive space is even more marked if, using SW's preferred univariate representation, we generalise our analysis to allow explicitly for time variation in θ (which SW model indirectly by allowing the variances of the permanent and transitory innovations to inflation to vary over time).[24] The last three columns of Table 1 show 16.5%, 50% and 83.5% quantiles of the posterior distribution for $\widehat{\theta}_t$ in the last observation of SW's sample, 2004:IV. On the basis of the median estimate of $\widehat{\theta} = 0.85$ at this point the predictive space is extremely compressed, with R^2 lying in a narrow range between 0.42 and 0.58; but, most notably, $\rho_{\max}$ is essentially indistinguishable from -1.[25]

In the light of these calculations, SW's conclusion that inflation has become much harder to forecast in recent data becomes readily interpretable in terms of the univariate representation. Essentially in recent inflation data there is quite limited scope for even the best possible predictor of inflation consistent with the properties of inflation to out-predict the fundamental ARMA representation. Any such predictor must be close to IID, with innovations that are nearly perfectly negatively correlated with innovations to inflation. The predictions it generates, $\widehat{y}_{t+1} = \beta x_t$, must also very closely resemble those of the "univariate predictor" x_t^F defined in Section 3.1 (which here is simply ε_t) since $\text{corr}\left(x_t, x_t^F\right)$ is bounded below by $\sqrt{R_F^2/R_N^2}$,[26] which in this simple case is just equal to θ. Thus all possible single predictor models of

[24] As noted in Section 3.7.3, we can generalise our analysis fairly straightforwardly to accommodate time-varying parameters.

[25] These implied figures are themselves time-varying, and can, given SW's representation, only be defined locally.

[26] Since by the missing variables formula $R_F^2 = corr\left(x_t^F, x_t\right)^2 R^2$, and, from Proposition 1, $R^2 \leq R_N^2$.

inflation must also closely resemble the fundamental ARMA representation. We shall see below that this conclusion also applies, with only minor qualifications, to multiple predictor models.

5 Discussion and implications for empirical prediction problems

5.1 What can the history of y_t tell us in a finite sample?

At the start of this paper we posed the question: what does the history of a process y_t tell us about the nature of any possible predictive system in terms of a vector of predictors, $\mathbf{x}_t$ that can have generated the univariate reduced form? The answer depends on how much we know (or assume) about univariate properties.

Our core results have been derived in terms of population properties. If we had an infinite history of y_t we would know the true population parameters in the ARMA(p, q) representation. Under A4 to A6, the MA parameter, q equals r, the number of predictors in the predictive system, and the ARMA parameters define the predictive space $\mathbb{P}_{\boldsymbol{\lambda},\boldsymbol{\theta}}$ of all possible systems consistent with the reduced form, which is a strict subset of $\mathcal{P}_r$, the parameter space of all possible r-predictor systems.

Clearly, in a finite sample things are not so straightforward. In general we cannot know the true ARMA parameters. Diagnostic tests will be unable to distinguish between an ARMA(p, q) and an ARMA$(p + 1, q + 1)$ for θ_{q+1} and λ_{p+1} sufficiently close either to zero or to each other. However, finite sample properties can still provide us with important information about the predictive space.

First, standard model selection criteria on a finite sample of data will at least provide us with an estimated ARMA(p, q) representation that cannot be rejected against higher order alternatives. Under our assumptions, such a representation implies that there must be some predictive system for y_t with *at least* q predictors.[27] Our empirical application provides a simple example: Stock and Watson's (2007) IMA(1) representation of inflation tells us that there must be at least one predictor of inflation, and that any single predictor must itself be close to being IID.[28]

Second, even a more limited set of observable univariate properties - for example, a declining variance ratio and a given degree of univariate predictability - may also tell us a lot about the predictive space. For many y_t processes we can feel reasonably

[27] To be more precise, following the logic of Section 3.7.1 there must be some predictive system for y_t with at least q predictors and $\max(p - q, 0)$ additional lags of y_t

[28] We neglect here the issue of time-varying parameters, which as we noted above does not change the key elements of our arguments.

confident that the univariate R^2, R_F^2 does not exceed some value. This may in some cases be quite close to zero (for example, stock returns, GNP growth and real exchange rate changes). For other processes with stronger univariate predictability (for example changes in inflation, as analysed in our example) we may be able to identify a reasonably narrow confidence interval for R_F^2. We may also be able derive confidence intervals for the limiting variance ratio;[29] or may at least be able to reject $VR(h) = 1$ with high probability on a one-tailed test, for some large h.[30] Even this limited information tells us that y_t must, at a minimum, have an ARMA(1,1) representation, and hence must be generated by a predictive system with at least one predictor, with a Stambaugh/Beveridge-Nelson Correlation ρ that is likely to be close to -1.

Third, the necessary link between the ARMA representation and the predictive system can also have important implications for the way Granger Causality tests are carried out. For example, a conventional test of one-way Granger Causality in a VAR(1) in terms of y_t and some observable predictor z_t sets the coefficient on z_{t-1} to zero, thus forcing y_t to be an AR(1) under the null. But if the data point to, for example, an ARMA(1,1) as a minimal representation of y_t, the conventional null is clearly mis-specified, since the ARMA representation tells us that there must be *some* true predictor of y_t, even if it is not the observable predictor z_t.[31] Robertson & Wright (2011) propose an alternative test procedure, consistent with the original Granger (1969) definition of causality, that avoids this pitfall, by ensuring that y_t has the same order ARMA representation under both the null and the alternative.[32]

Fourth, Proposition 2 and its corollaries showed that there will be a range of y_t processes for which any predictive model in the predictive space must have a Beveridge-Nelson Correlation, ρ that is strongly negative, and hence an innovation covariance matrix Ω that is close to singular. Since Stambaugh (1999) the literature on predictive return regressions, in particular, has addressed the inference problems that arise when the Stambaugh Correlation $\rho_S = corr(\beta' v_t, u_t)$ (which for a single predictor model is identical to ρ, and in many predictive models will be of a very

[29] Or, as we show in the case of the Stock-Watson (2007) example, some transform thereof. See Appendix B.2.

[30] As another concrete example Pastor & Stambaugh's (2011) long annual dataset shows that the sample variance ratio for real stock returns asymptotes to a value sufficiently far below unity that the null $V \geq 1$, can be rejected with high probability.

[31] Equivalently, the conventional Granger causality test implies the joint null, that z_t does not predict y_t, *and* that y_t is an AR(1) in reduced form.

[32] In much recent literature the focus has switched from within-sample causality testing to out-of-sample predictive testing. But very similar considerations apply : the standard univariate benchmark employed for comparisons is usually a finite order AR representation, whereas if the true DGP is a predictive system it should be an ARMA.

similar magnitude) is close to -1, which implies that estimated values of β are biased away from zero in finite samples. "Stambaugh Bias" arises because such predictive models are "ARMA-like" (since any ARMA model has $\rho = \pm 1$). But our results imply that for any y_t process for which $\rho_{\max}$ is close to -1 (eg, the Stock-Watson (2007) model of inflation analysed in Section 4.2), Stambaugh Bias must be an endemic problem for *any* predictive regression. For such y_t processes the presumed advantage of predictive regressions over ARMA models, due to the well-known difficulties in ARMA estimation, are therefore largely illusory.

5.2 A puzzle: low order ARMAs vs high order predictive models

One puzzle opened up by our analysis is an apparent disconnect between univariate models and predictive regressions. Typically in estimated ARMA representations, q will be small. How can we reconcile ARMA models with low q with the much larger numbers of predictors we observe in many predictive regressions?

Clearly if we take predictive regressions seriously, as our analysis does, then under our maintained assumptions if there are r predictors in an estimated predictive regression as in (6), with an autoregressive matrix with r distinct eigenvalues, the parameters of the predictive system must map to an ARMA(r, r), as in (9). The only possible reconciliation with low q in the observable ARMA is that there must be $r - q$ pairs of AR and MA parameters that are sufficiently close to cancellation that a lower order ARMA representation cannot be rejected.[33] But this in itself *still* provides us with important information about the the predictive space.

The Stock & Watson (2007) model of inflation analysed in Section 4.2 again provides a simple illustration. While their MA(1) representation of $y_t = \Delta\pi_t$ is consistent with consistent with data in each of their sub-samples, clearly another possible representation might be, for example, an ARMA(2,2), as long as the additional AR and MA parameters are sufficiently close to zero or cancellation. For any such univariate representation the predictive space is the set of all 2 predictor models with parameters $\mathbf{\Psi}$ that map to a y_t process that is indistinguishable from SW's MA(1) in the given sample. This could for example contain predictive models in which one of the predictors, say x_{1t}, is close to IID ($\lambda_1 \approx 0$), while the second has an AR parameter, say λ_2 anywhere within $[-1, 1]$ as long as long as $\theta_1(\mathbf{\Psi}) \approx \widehat{\theta}$, SW's point estimate, and $\theta_2(\mathbf{\Psi}) \approx \lambda_2$. This might appear in principle to open up

[33] Adapting Lippi & Reichlin's (1994) terminology, this means that the nonfundamental representation (3) that determines $R^2_{\max}$ is "nearly nonbasic": i.e., its properties can barely be detected from the history of y_t.

the predictive space considerably. However, we show in Appendix B.2 that if on *a priori* grounds we assume that both predictors have positive persistence ($\lambda_1, \lambda_2 > 0$) (Stock & Watson note that this is a feature of many commonly used candidate predictors) this expansion of the predictive space is largely illusory: the feasible range of (R^2, ρ) combinations is essentially the same as in a single predictor model, as shown in Table 1. We show that for any multiple predictor model to predict *better* than a single predictor model requires (as a necessary but not sufficient condition) that $\lambda_i < 0$ for at least some i, and also puts significant constraints on Ω, the innovation correlation matrix. Thus multiple predictor models are still very tightly constrained by univariate properties.

An alternative explanation for large numbers of predictors in predictive regressions is that they are not the true state variables, but a set of imperfect predictors as discussed in Section 3.5 that merely provide a noisy signal of the much smaller set of true state variables. But in such cases the predictive regression must be mis-specified, and the appropriate estimation methodology is one of signal extraction rather than regression (*cf* Stock & Watson, 2002). Furthermore, as we noted in Section 3.5, while the true predictor vector $\mathbf{x}_t$ must predict at least as well the true ARMA representation, this need not be the case for a set of imperfect predictors. But if a set of observable predictors does *not* predict as well as the ARMA, or even if it does not predict significantly better, this tells us either that there is mis-specification of the predictive regression, or that the apparent predictive power is spurious.[34]

One argument for a relatively small number of true state variables, which is therefore more consistent with low order ARMA models, is if different predictors have common AR roots, and hence can be aggregated together. This property can arise out of theoretical models that are driven by a relatively small number of exogenous stochastic processes. The original stochastic growth model was usually assumed to be driven by a single AR(1) technology process, and one persistent pre-determined state variable, the capital stock, so that all elements of the model were ARMA(2,1) processes with perfectly correlated innovations and common AR roots.[35] More recently much of the focus on estimated DSGE models has followed the example of Smets & Wouters (2007) in assuming a relatively large number of underlying driving stochastic processes (which would imply high r in our framework). But it is noteworthy that Boivin and Giannoni (2006), who take a signal extraction approach to DSGE estimation, based on a large number of indicators, conclude in

[34] The alternative tests of Granger Causality in Robertson & Wright (2011) discussed in the previous sub-section provide one way to address this issue.

[35] See for example the exposition in Campbell, (1994).

favour of a relatively small number of driving processes, consistent with (relatively) low order ARMA reduced forms.

An almost diametrically opposite argument is implicit in the long memory literature. A limiting case of our predictive system arises as r, the number of AR(1) predictors with distinct AR(1) parameters, tends to infinity. In such cases there is no viable finite order ARMA representation; however, as Granger's (1980) original derivation showed, assumptions about the nature of the limiting distribution of the λ_i as $r \to \infty$ may allow a univariate representation of y_t as a long memory process with a relatively small number of parameters. In such cases, therefore, finite sample univariate representations with a small number of parameters need not necessarily be finite order ARMAs, and hence need not of themselves imply a small number of state variables. But if the history of y_t does lead us to the conclusion that it is a long memory process, this *still* has strong implications for the predictive space: either predictors must themselves be long memory processes; or y_t must have an infinite or extremely high dimension predictor vector, with a particular limiting distribution of the λ_i.

5.3 Limits to predictability?

Does our analysis help to explain why economists appear to have only rather limited success at forecasting? The analytical and empirical examples analysed in Sections 4.1 and 4.2 illustrated cases where the predictive space is very tightly constrained, and also illustrated that for some series, even if we observed the true state variable, it would make a marginal contribution to improving R^2 beyond the univariate lower bound. It is notable that the y_t processes in such examples capture the univariate properties of a quite wide range of observable economic time series.

Our framework does not however justify unqualified pessimism about prediction:

- In the ARMA(1,1) example in Section 4.1 we showed that, of two ARMA(1,1) processes that were both close to univariate white noise, the predictive space for one was distinctly less constrained. The key feature determining the difference was whether the variance ratio sloped upwards (the predictive space is relatively unconstrained) or downwards (the predictive space is tightly constrained). But the upward sloping variance ratio means that the flipside of this greater scope for forecastability in the short run is greater uncertainty in the long run. For example, Campbell & Viceira (1999) show that short-term returns on cash have increasing variance ratios, and hence should be much more forecastable than, for example, short-term returns on stocks. But their

increasing variance ratio also means that long-term cash returns have much higher unconditional uncertainty.

- Even when the predictive space is very tightly constrained, the nature of these constraints can still give guidance on what kind of predictors are likely to give predictive power for y_t. Thus, our empirical example in Section 4.2 showed that a single predictor model of inflation has a very tightly constrained predictive space, but also pointed to necessary characteristics of both single and multiple predictors that might in principle offer scope for improved predictions.

6 Conclusions

Prediction of time series processes is not carried out in an informational vacuum. Our analysis has shown that what we know (or assume) about the time series properties of some process y_t can tell us a lot about the properties of any predictor vector $\mathbf{x}_t$ and predictive regression consistent with those properties. For some (possibly many) y_t processes this may well imply that it will be hard to find predictive regressions that predict much better than an ARMA model, and that may share many of the finite sample problems of ARMA estimation. But, to the extent that we can find models that approach the limits of predictability, we are more likely to do so by being aware of the constraints that the time series properties of y_t put on the predictive space containing $\mathbf{x}_t$.

References

Baxter, B, Graham, L and Wright, S "Invertible and non-invertible information sets in linear rational expectations models" *Journal of Economic Dynamics and Control*, vol. 35(3) (2011) pages 295-311,

Beveridge, S. and C.R. Nelson (1981), "A New Approach to the Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the "Business Cycle"," *Journal of Monetary Economics*, 7(2), 151-174.

Boivin, J, and Giannoni, M (2006) "DSGE Models in a Data-Rich Environment" NBER *Working Paper* no. 12772

Campbell, John Y, (1994), "Inspecting the mechanism: an analytical approach to the stochastic growth model", Journal of Monetary Economics 33, pp.463-506.

Campbell, John Y and Mankiw, N Gregory (1987) "Are Output Fluctuations Transitory", *Quarterly Journal of Economics,* 102, pp 857-80

Campbell J Y & L M. Viceira, (1999). "Consumption And Portfolio Decisions When Expected Returns Are Time Varying," *The Quarterly Journal of Economics*, MIT Press, vol. 114(2), pages 433-495, May

Cochrane, J.H (1988) "How big is the random walk in GDP?. "*Journal of Political Economy*, v. 96, n. 5, p. 893-92

Cochrane J H (2008) The Dog that Did Not Bark: A Defense of Return Predictability *Review of Financial Studies*, 2008, pp 1533-1575.

Fama, Eugene F and French, Kenneth R (1988) "Permanent and Temporary Components of Stock Prices", *Journal of Political Economy,* 96, pp 246-73

Fernandez-Villaverde, J, Rubio-Ramirez, F, Sargent, T and Watson, M (2007) "ABCs (and Ds) of Understanding VARs", American Economic Review, pp 1021-1026

Granger, C (1969) Investigating Causal Relations by Econometric Models and Cross-spectral Methods, *Econometrica,* Vol. 37, No. 3 (Aug., 1969), pp. 424-438

Garratt, A, Robertson, D and Wright, S (2006) "Permanent vs Transitory Components and Economic Fundamentals" *Journal of Applied Econometrics* 21 (4) 521-542

Hamilton J D (1994) *Time Series Analysis* 1994 Princeton University Press

Hansen, L P and Sargent, T J, *Recursive Linear Models of Dynamic Economies*, unpublished

Harvey, Andrew (1981) *Time Series Models* Philip Allan, London

Kim, M J, Nelson, C R and Startz, R (1991) "Mean Reversion in Stock Prices? A Reappraisal of the Empirical Evidence". *Review of Economic Studies*, 58(3), 515-528

Lippi, Marco and Reichlin, Lucrezia (1994) "VAR analysis, nonfundamental representations, Blaschke matrices", Journal of Econometrics, 63 pp 307-325

Pastor L and R F Stambaugh (2009) "Predictive Systems: Living with Imperfect Predictors" *Journal of Finance* 64(4) p1583-1628

Pastor L and R F Stambaugh (2011) "Are stocks really less volatile in the long run?" *Journal of Finance,* forthcoming

Robertson D and S Wright (2011) Stambaugh Correlations, Monkey Econometricians and Redundant Predictors, Birkbeck College *Working Paper*

Smets, F and Wouters, R (2007) "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach", *American Economic Review,* 97 , pp 586-606

Stambaugh R F (1999) "Predictive Regressions", *Journal of Financial Economics* 54: 375–421.

Stock, J.H. and M.W. Watson (1988), 'Testing for Common Trends', *Journal of the American Statistical Association*, 83, 1097–1107.

Stock, James H and Watson, Mark W (2002), "Macroeconomic Forecasting Using Diffusion Indices", *Journal of Business & Economic Statistics*

Stock, James H and Watson, Mark W (2007) "Why Has U.S. Inflation Become Harder to Forecast?" (with James H. Stock), *Journal of Money Credit and Banking*, Vol. 39, No. 1

Appendix

The Appendix is structured as follows:

Appendix A provides proofs of propositions and corollaries, together with derivations exploited in the main text of the paper. For ease of checking by referees, we provide considerably more detail than we would expect to be included in any published version of the paper.

Appendix B provides extended versions of our two examples that illustrate the properties of the predictive space in more complex models than those included in the main text. At various points in the main paper we make reference to key features of these additional examples, hence we felt referees would wish to be able to see substantiation of the points made. We would assume that any published version of the paper could make reference to a working paper that would include this additional material.

A Proofs and Derivations

A.1 Properties of the Minimum Variance Nonfundamental Representation

The following result summarises the key properties of the representation in (3):

Lemma 1 *In the set of all possible nonfundamental ARMA(p,q) representations consistent with (1) in which, for $q>0$, θ_i is replaced with θ_i^{-1} for at least some i, the moving average polynomial $\theta^N(L)$ in (3) in which θ_i is replaced with θ_i^{-1} for all i, has innovations η_t with the minimum variance, with*

$$\sigma_\eta^2 = \sigma_\varepsilon^2 \prod_{i=1}^{q} \theta_i^2 \tag{38}$$

and hence

$$R_N^2(\boldsymbol{\lambda},\boldsymbol{\theta}) = 1 - \left(1 - R_F^2(\boldsymbol{\lambda},\boldsymbol{\theta})\right) \prod_{i=1}^{q} \theta_i^2 \tag{39}$$

Proof. Equating (1) to (3) the non-fundamental and fundamental innovations are related by

$$\varepsilon_t = \prod_{i=1}^{q} \left(\frac{1-\theta_i^{-1}L}{1-\theta_i L}\right) \eta_t = \sum_{i=0}^{\infty} c_i \eta_{t-i} \tag{40}$$

for some square summable c_i. Therefore, since η_t is itself IID,

$$\sigma_\varepsilon^2 = \sigma_\eta^2 \sum_{i=0}^{\infty} c_i^2 \tag{41}$$

Now define

$$c(L) = \sum_{i=0}^{\infty} c_i L^i = \prod_{i=1}^{q} \left(\frac{1 - \theta_i^{-1} L}{1 - \theta_i L} \right) \tag{42}$$

so

$$c(1) = \prod_{i=1}^{q} \left(\frac{1 - \theta_i^{-1}}{1 - \theta_i} \right) = \prod_{i=1}^{q} \left(\frac{-1}{\theta_i} \right) \tag{43}$$

and

$$c(1)^2 = \prod_{i=1}^{q} \frac{1}{\theta_i^2} = \left(\sum_{i=0}^{\infty} c_i \right)^2 = \sum_{i=0}^{\infty} c_i^2 + \sum_{j \neq i} c_i c_j \tag{44}$$

Sine ε_t is IID we have

$$E(\varepsilon_t \varepsilon_{t+j}) = 0 \;\; j = 1, \ldots, \infty$$

implying

$$\sum_{i=0}^{\infty} c_i c_{i+j} = 0 \;\; j = 1, \ldots, \infty \tag{45}$$

Hence we have

$$\sum_{j=1}^{\infty} \sum_{i=0}^{\infty} c_i c_{i+j} = \sum_{j \neq i} c_i c_j = 0 \tag{46}$$

thus

$$\sum_{i=0}^{\infty} c_i^2 = c(1)^2 = \prod_{i=1}^{q} \frac{1}{\theta_i^2} \tag{47}$$

Thus using (47) and (41) we have (38) and hence (39).

To show that this is the nonfundamental representation with the minimum innovation variance, consider the full set of nonfundamental ARMA(p, p) representations, in which, for each representation k, $k = 1, \ldots, 2^q - 1$, there is some ordering such that, θ_i is replaced with θ_i^{-1}, $i = 1, \ldots, s\,(k)$, for $s \leq q$. For any such representation, with innovations $\eta_{k,t}$, we have

$$\sigma_{\eta,k}^2 = \sigma_\varepsilon^2 \prod_{i=1}^{s(k)} \theta_i^2 \tag{48}$$

This is minimised for $s\,(k) = q$, which is only the case for the single representation in which θ_i is replaced with θ_i^{-1} for all i, and thus this will give the minimum variance nonfundamental representation. ■

A.2 Moment Conditions

After substitution from (7) the predictive regression (6) can be written as in (8), restated here, as

$$\det\left(\mathbf{I}-\mathbf{\Lambda}L\right)y_t=\beta'\text{adj}\left(\mathbf{I}-\mathbf{\Lambda}L\right)\mathbf{v}_{t-1}+\det\left(\mathbf{I}-\mathbf{\Lambda}L\right)u_t \tag{49}$$

Given diagonality of $\mathbf{\Lambda}$, from A1, we can rewrite this as

$$\widetilde{y}_t\equiv\prod_{i=1}^{r}\left(1-\lambda_i^*L\right)y_t=\sum_{i=1}^{r}\beta_i\prod_{j\neq i}\left(1-\lambda_j^*L\right)Lv_{it}+\prod_{i=1}^{r}\left(1-\lambda_i^*L\right)u_t\equiv\sum_{i=0}^{r}\boldsymbol{\gamma}_i'L^i\mathbf{w}_t \tag{50}$$

where $\widetilde{y}_t$ is an MA(r), $\mathbf{w}_t=\begin{pmatrix}u_t & \mathbf{v}_t'\end{pmatrix}'$, and the final equality implicitly defines a set of $(r+1)\times 1$ vectors, $\boldsymbol{\gamma}_i$, $i=0,\ldots,r$, $\boldsymbol{\gamma}_i=\boldsymbol{\gamma}_i\left(\boldsymbol{\beta},\boldsymbol{\lambda}^*\right)$.

Let A_i be the ith order autocorrelation of $\widetilde{y}_t$ implied by the predictive system. We have

$$A_i\left(\boldsymbol{\beta},\boldsymbol{\lambda}^*,\mathbf{\Omega}\right)=\frac{\sum_{j=0}^{r-i}\boldsymbol{\gamma}_j'\mathbf{\Omega}\boldsymbol{\gamma}_{j+i}}{\sum_{j=0}^{r}\boldsymbol{\gamma}_j'\mathbf{\Omega}\boldsymbol{\gamma}_j} \tag{51}$$

Let κ_i be the ith order autocorrelation of $\widetilde{y}_t$ implied by the ARMA(r,r) representation in (9), given by (Hamilton, 1994, p51)

$$\kappa_i\left(\boldsymbol{\theta}^*\right)=\frac{\psi_i+\psi_{i+1}\psi_1+\psi_{i+2}\psi_2+....+\psi_r\psi_{r-i}}{1+\psi_1^2+\psi_2^2+...+\psi_r^2} \tag{52}$$

where the $\psi_i\left(\boldsymbol{\theta}^*\right)$ satisfy

$$\prod_{i=1}^{r}\left(1-\theta_i^*L\right)=\sum_{i=1}^{r}1+\psi_1L+\psi_2L^2+...+\psi_rL^r$$

Thus the θ_i^* are the solutions to the moment conditions

$$\kappa_i\left(\boldsymbol{\theta}^*\right)=A_i\left(\boldsymbol{\beta},\boldsymbol{\lambda}^*,\mathbf{\Omega}\right),\quad i=1..r \tag{53}$$

such that $\theta_i^*\in(-1,1)\ \forall i$.

A.3 Proof of Proposition 1

We proceed by proving two sub-results that lead straightforwardly to the result in the Proposition itself.

Lemma 2 *In the population regression*

$$y_t = \boldsymbol{\nu}_{\mathbf{x}}' \mathbf{x}_{t-1} + \boldsymbol{\nu}_F' \mathbf{x}_{t-1}^F + \xi_t \tag{54}$$

where the true process for y_t is as in (6), and $\mathbf{x}_t^F$ is the vector of fundamental univariate predictors defined in (17), all elements of the coefficient vector $\boldsymbol{\nu}_F$ are zero.

Proof. The result will follow automatically if we can show that the x_{it-1}^F are all orthogonal to $u_t \equiv y_t - \beta' \mathbf{x}_{t-1}$. Equalising (1) and (6), and substituting from (7), we have (noting that under A1 to A3 $p = q = r$)

$$y_t = \frac{\prod_{i=1}^r (1-\theta_i L)}{\prod_{i=1}^r (1-\lambda_i L)} \varepsilon_t = \frac{\beta_1 v_{1t-1}}{1-\lambda_1 L} + \frac{\beta_2 v_{2t-1}}{1-\lambda_2 L} + \ldots + \frac{\beta_r v_{rt-1}}{1-\lambda_r L} + u_t \tag{55}$$

So we may write, using (17),

$$\begin{aligned} x_{jt-1}^F &= \frac{\varepsilon_{t-1}}{1-\lambda_j L} \\ &= \left(\frac{L}{1-\lambda_j L}\right) \frac{\prod_{i=1}^r (1-\lambda_i L)}{\prod_{i=1}^r (1-\theta_i L)} \left(\frac{\beta_1 L v_{1t-1}}{1-\lambda_1 L} + \frac{\beta_2 L v_{2t-1}}{1-\lambda_2 L} + \ldots + \frac{\beta_r L v_{rt-1}}{1-\lambda_r L} + u_t\right) \end{aligned} \tag{56}$$

Given the assumption that u_t and the v_{it} are jointly IID, u_t will indeed be orthogonal to x_{jt-1}^F, for all j, since the expression on the right-hand side involves only terms dated $t-1$ and earlier, thus proving the Lemma. ■

Lemma 3 *In the population regression*

$$y_t = \boldsymbol{\phi}_{\mathbf{x}}' \mathbf{x}_{t-1} + \boldsymbol{\phi}_N' \mathbf{x}_{t-1}^N + \zeta_t \tag{57}$$

where $\mathbf{x}_t^N$ is the vector of nonfundamental univariate predictors defined in (18), all elements of the coefficient vector $\boldsymbol{\phi}_{\mathbf{x}}$ are zero.

Proof. The result will again follow automatically if we can show that the x_{it-1} are all orthogonal to $\eta_t \equiv y_t - \beta_N' \mathbf{x}_{t-1}^N$. Equating (3) and (6), and substituting from (7), we have

$$y_t = \frac{\prod_{i=1}^r (1-\theta_i^{-1} L)}{\prod_{i=1}^r (1-\lambda_i L)} \eta_t = \beta_1 \frac{v_{1t-1}}{1-\lambda_1 L} + \beta_2 \frac{v_{2t-1}}{1-\lambda_2 L} + \ldots + \beta_r \frac{v_{rt-1}}{1-\lambda_r L} + u_t \tag{58}$$

Using

$$\frac{1}{1-\theta_i^{-1} L} = \frac{-\theta_i F}{1-\theta_i F}$$

where F is the forward shift operator, $F = L^{-1}$, we can write

$$\eta_t = F^r \prod_{i=1}^{r} (-\theta_i) \left(\frac{\prod_{i=1}^{r}(1-\lambda_i L)}{\prod_{i=1}^{r}(1-\theta_i F)} \right) \left(\beta_1 \frac{v_{1t-1}}{1-\lambda_1 L} + \beta_2 \frac{v_{2t-1}}{1-\lambda_2 L} + \ldots + \beta_r \frac{v_{rt-1}}{1-\lambda_r L} + u_t \right) \tag{59}$$

Now

$$\begin{aligned} F^r \frac{\prod_{i=1}^{r}(1-\lambda_i L)}{\prod_{i=1}^{r}(1-\theta_i F)} \frac{v_{kt-1}}{(1-\lambda_k L)} &= F^r \left(\frac{\prod_{i\neq k}(1-\lambda_i L)}{\prod_{i=1}^{r}(1-\theta_i F)} \right) v_{kt-1} \\ &= v_{kt} + c_1 v_{kt+1} + c_2 v_{kt+2+} \ldots \end{aligned}$$

for some $c_1, c_2, ...$ since the highest order term in L in the numerator of the bracketed expression is of order $r-1$, and

$$F^r \left(\frac{\prod_{i=1}^{r}(1-\lambda_i L)}{\prod_{i=1}^{r}(1-\theta_i F)} \right) u_t = u_t + b_1 u_{t+1} + b_2 u_{t+2} \ldots$$

for some $b_1, b_2, \ldots$, since the highest order term in L in the numerator of the bracketed expression is of order r. Hence η_t can be expressed as a weighted average of current and forward values of u_t and v_{it} and will thus be orthogonal to $x_{it-1} = \frac{v_{it-1}}{1-\lambda_i L}$ for all i, by the assumed joint IID properties of u_t and the v_{it}, proving the Lemma. ■

Now let $R_1^2 \equiv 1 - \sigma_\xi^2/\sigma_y^2$ be the predictive R^2 of the predictive regression (54) analysed in Lemma 2. Since the predictive regressions in terms of $\mathbf{x}_t$ in (6) and in terms of $\mathbf{x}_t^F$ in (19) are both nested in (54) we must in general have $R_1^2 \geq R^2$ and $R_1^2 \geq R_F^2$. But Lemma 2 implies that, given $\boldsymbol{\nu}_F = 0$ we must have $R_1^2 = R^2$, hence $R^2 \geq R_F^2$.

By similar argument,let $R_2^2 \equiv 1 - \sigma_\zeta^2/\sigma_y^2$ be the predictive R^2 of the predictive regression (57) analysed in Lemma 3. Since the predictive regressions in terms of $\mathbf{x}_t$ in (6) and in terms of $\mathbf{x}_t^N$ in (20) are both nested in (57) we must in general have $R_2^2 \geq R^2$ and $R_2^2 \geq R_N^2$. But Lemma 3 implies that, given $\boldsymbol{\phi}_{\mathbf{x}} = 0$ we must have $R_2^2 = R_N^2$, hence $R_N^2 \geq R^2$.

Hence

$$R_F^2(\boldsymbol{\theta}, \boldsymbol{\lambda}) \leq R^2 \leq R_N^2(\boldsymbol{\theta}, \boldsymbol{\lambda})$$

thus proving the Proposition.■

A.4 Proof of Proposition 2 and Corollaries 2 and 3.

The predictive system in (6) and (7) implies the multivariate Beveridge Nelson(1981)/Stock Watson (1988) decomposition

$$\begin{aligned}
\begin{bmatrix} \Delta Y_t \\ \mathbf{x}_t \end{bmatrix} &= \mathbf{C}(L) \begin{bmatrix} u_t \\ \mathbf{v}_t \end{bmatrix} \\
&= \begin{bmatrix} 1 & \boldsymbol{\beta}' L [I - \boldsymbol{\Lambda} L]^{-1} \\ 0 & [I - \boldsymbol{\Lambda} L]^{-1} \end{bmatrix} \begin{bmatrix} u_t \\ \mathbf{v}_t \end{bmatrix} \\
&= [\mathbf{C}(1) + \mathbf{C}^*(L)(1 - L)] \begin{bmatrix} u_t \\ \mathbf{v}_t \end{bmatrix} \\
&= \left\{ \begin{bmatrix} 1 & \boldsymbol{\beta}' [I - \boldsymbol{\Lambda}]^{-1} \\ 0 & [I - \boldsymbol{\Lambda}]^{-1} \end{bmatrix} + \begin{bmatrix} 0 & \boldsymbol{\beta}' \left(L [I - \boldsymbol{\Lambda} L]^{-1} - [I - \boldsymbol{\Lambda}]^{-1} \right) \\ 0 & [I - \boldsymbol{\Lambda} L]^{-1} - [I - \boldsymbol{\Lambda}]^{-1} \end{bmatrix} \right\} \begin{bmatrix} u_t \\ \mathbf{v}_t \end{bmatrix}
\end{aligned} \tag{60}$$

for which the equation for ΔY_t in the last line can be written, as in (24) and (26) in the main text, restated here

$$Y_t = Y_t^P + Y_t^T \tag{61}$$

where

$$Y_t^P = \frac{u_t + \boldsymbol{\delta}' \mathbf{v}_t}{1 - L} \tag{62}$$

where $\boldsymbol{\delta} = \boldsymbol{\beta}' [I - \boldsymbol{\Lambda}]^{-1}$.

Cochrane (1988, equation (10)) shows that $V = \lim_{h \to \infty} VR(h)$ as defined in (23) must, letting $\sigma_p^2 \equiv var\left(\Delta Y_t^P\right)$, satisfy

$$V = \frac{\sigma_p^2}{\sigma_y^2} \tag{63}$$

since σ_p^2 must be equal in population whether it is derived from the univariate or multivariate representation. Thus straightforwardly we have, from (62)

$$V = \frac{var(u_t + \boldsymbol{\delta}' \mathbf{v}_t)}{var(u_t + \boldsymbol{\beta}' \mathbf{x}_{t-1})} \tag{64}$$

By setting u_t to zero in (64) we can also derive the variance ratio of the predicted value for y_t, $\widehat{y}_t = \beta' x_{t-1}$ i.e.

$$V_{\widehat{y}} \equiv \frac{var(\boldsymbol{\delta}' \mathbf{v}_t)}{var(\beta' \mathbf{x}_t)} \equiv \frac{\sigma_{\delta' v}^2}{\sigma_{\widehat{y}}^2} \tag{65}$$

Using this definition, and the definitions of R^2 and ρ in Propositions 1 and 2 we have

$$\begin{aligned} V &= \frac{var\left(u_t + \boldsymbol{\delta}'\mathbf{v}_t\right)}{\sigma_y^2} = \frac{\sigma_u^2 + \sigma_{\boldsymbol{\delta}'\mathbf{v}}^2 + 2\boldsymbol{\delta}'\Omega_{uv}}{\sigma_y^2} \\ &= \frac{\sigma_u^2}{\sigma_y^2} + \frac{\sigma_{\boldsymbol{\delta}'\mathbf{v}}^2}{\sigma_{\widehat{y}}^2}\frac{\sigma_{\widehat{y}}^2}{\sigma_y^2} + \frac{2.\boldsymbol{\delta}'\Omega_{uv}}{\sigma_{\boldsymbol{\delta}'\mathbf{v}}.\sigma_u}.\frac{\sigma_{\boldsymbol{\delta}'\mathbf{v}}.\sigma_u}{\sigma_y^2} \\ &= 1 - R^2 + V_{\widehat{y}}R^2 + 2\rho\frac{\sigma_{\boldsymbol{\delta}'\mathbf{v}}}{\sigma_{\widehat{y}}}.\frac{\sigma_{\widehat{y}}}{\sigma_y}\frac{\sigma_u}{\sigma_y} \\ &= 1 + R^2\left(V_{\widehat{y}} - 1\right) + 2\rho\sqrt{V_{\widehat{y}}R^2\left(1 - R^2\right)} \end{aligned}$$

as given in (29) in the proposition.■

To prove Corollary 2, partition Ω as

$$\Omega = \begin{bmatrix} \sigma_u^2 & \Omega_{uv}' \\ \Omega_{uv} & \Omega_v \end{bmatrix} \tag{66}$$

hence

$$\begin{aligned} \sigma_{\delta' v}^2 &= \delta'\Omega_v\delta = \sum_{i=1}^{r}\sum_{j=1}^{r} D_{ij}\Omega_{ij}; \quad = tr\left(D\Omega_v\right) \\ \text{where } D_{ij} &= D_{ji} = \delta_i\delta_j = \frac{\beta_i\beta_j}{\left(1 - \lambda_i\right)\left(1 - \lambda_j\right)} \end{aligned}$$

$$\begin{aligned} \sigma_{\widehat{y}}^2 &= \beta' E\left(\mathbf{x}_t\mathbf{x}_t'\right)\beta = \sum_{i=1}^{r}\sum_{j=1}^{r} B_{ij}\Omega_{ij} = tr\left(B\Omega_v\right) \\ \text{where, using A1, } E\left(\mathbf{x}_t\mathbf{x}_t'\right) &= \left[\frac{\Omega_{ij}}{1 - \lambda_i\lambda_j}\right]_{ij} \Rightarrow B_{ij} = B_{ji} = \frac{\beta_i\beta_j}{1 - \lambda_i\lambda_j} \end{aligned}$$

Manipulation of the definitions of D and B gives

$$D_{ij} - B_{ij} = \frac{\beta_i\beta_j}{\left(1 - \lambda_i\lambda_j\right)}\left(\frac{\lambda_i}{1 - \lambda_i} + \frac{\lambda_j}{1 - \lambda_j}\right) \tag{67}$$

hence, since we can write (65) as

$$V_{\widehat{y}} = \frac{tr\left(D\Omega_v\right)}{tr\left(D\Omega_v\right)} \tag{68}$$

under the conditions stated the numerator of (68) is element-by-element larger than the denominator, hence $V_{\widehat{y}} > 1$.

To derive the bounds for ρ and R^2, solving (29) for ρ we obtain

$$\rho = h\left(R^2, V, V_{\widehat{y}}\right) = -\frac{1}{2}\left(\frac{1 - V - R^2\left(1 - V_{\widehat{y}}\right)}{\sqrt{V_{\widehat{y}} R^2\left(1 - R^2\right)}}\right) \tag{69}$$

which describes a surface in three dimensions that satisfies $g(\Psi) = V$. Taking $V_{\widehat{y}}$ as given, we then have

$$\frac{\partial h}{\partial R^2} = -\frac{1}{4}\left(\frac{1 - V + V_{\widehat{y}} - V}{\left(R^2\left(1 - R^2\right)\right)^{\frac{3}{2}} V_{\widehat{y}}^{\frac{1}{2}}}\right)\left(R^2 - \frac{1 - V}{1 - V + V_{\widehat{y}} - V}\right) \tag{70}$$

so ρ has at most one stationary point $\widehat{\rho}$, within $[-1, 1]$, which is given by

$$\widehat{\rho} = \sqrt{\frac{\left(1 - V\right)\left(V_{\widehat{y}} - V\right)}{V_{\widehat{y}}}} sgn\left(V - 1 + V - V_{\widehat{y}}\right) \tag{71}$$

For $V < 1, V_{\widehat{y}} > V$, by inspection of (70) and (71), this is a maximum, and is a decreasing function of $V_{\widehat{y}}$, hence, giving the upper bound for ρ in Corollary 2, by setting $V_{\widehat{y}} = 1$ in (71).

For given $V_{\widehat{y}}$, and $V < 1$, there are two values of R^2 satisfying $h\left(R^2, V, V_{\widehat{y}}\right) = -1$. The upper of these two solutions (which yields the maximum possible R^2 for given V and $V_{\widehat{y}}$) is

$$\widehat{R}^2\left(V, V_{\widehat{y}}\right) = \frac{1 - V + V_{\widehat{y}}\left(1 + V\right) + 2\sqrt{V_{\widehat{y}} V\left(1 + V_{\widehat{y}} - V\right)}}{\left(1 + V_{\widehat{y}}\right)^2} \tag{72}$$

and, by inspection,

$$\widehat{R}^2\left(V, 1\right) = \frac{1 + 2\sqrt{V\left(2 - V\right)}}{2} \tag{73}$$

To show that this is an upper bound under the assumptions of Corollary 2 we need to show that $\widehat{R}^2\left(V, V_{\widehat{y}}\right)$ is a strictly decreasing function of $V_{\widehat{y}}$ for $V_{\widehat{y}} > 1$. Treating V as a fixed parameter we can write

$$\widehat{R}^2\left(V_{\widehat{y}}\right) = \frac{f\left(V_{\widehat{y}}\right)}{g\left(V_{\widehat{y}}\right)} \tag{74}$$

where, by inspection of (72), we have $R^2\left(V, V_{\widehat{y}}\right) = 1$, hence $f(V) = g(V)$. We also have

$$f'\left(V_{\widehat{y}}\right) - g'\left(V_{\widehat{y}}\right) = \left(V - 1 + V - V_{\widehat{y}}\right)\left(1 - H\left(V_{\widehat{y}}\right)\right) \tag{75}$$

where

$$H\left(V_{\widehat{y}}\right)=\sqrt{\frac{V}{V_{\widehat{y}}\left(1+V_{\widehat{y}}-V\right)}} \tag{76}$$

By inspection $H\left(V\right)=1$, and hence $f'\left(V\right)=g'\left(V\right)$. For $V<1, V_{\widehat{y}}>1$, $H<1$, and so $f'<g'$, hence $\widehat{R}^2\left(V_{\widehat{y}}\right)$ is indeed a strictly decreasing function, and hence (73) is an upper bound.[36]■

The upper bound for ρ in Corollary 3 follows directly from (71). Substituting for $V\left(\lambda,\theta\right)$ and $V_{\widehat{y}}\left(\lambda\right)$ in (69) using (84) (see Appendix A.6) the two solutions to $h\left(R^2,V\left(\lambda,\theta\right),V_{\widehat{y}}\left(\lambda\right)\right)$ yield $R^2_{\min}\left(\lambda,\theta\right)$ and $R^2_{\max}\left(\lambda,\theta\right)$ as in (34) and (35), which are special cases of the general formula in Proposition 1.■

A.5 Derivation of the predictive regression from a vector autoregression

Assume the underlying VAR is

$$\begin{aligned}\left[\begin{array}{c}\mathbf{z}_t\\ y_t^*\end{array}\right] &= \mathbf{A}\left[\begin{array}{c}\mathbf{z}_{t-1}\\ y_{t-1}^*\end{array}\right]+\left[\begin{array}{c}\boldsymbol{\zeta}_t\\ u_t\end{array}\right]\\ &= \mathbf{S}\boldsymbol{\Lambda}^*\mathbf{S}^{-1}\left[\begin{array}{c}\mathbf{z}_{t-1}\\ y_{t-1}^*\end{array}\right]+\left[\begin{array}{c}\boldsymbol{\zeta}_t\\ u_t\end{array}\right]\end{aligned}$$

where y_t^* is the variable of interest, and $\mathbf{z}_t$ is a vector with r elements, hence the system has $r+1$ elements, and $\boldsymbol{\Lambda}^*$=$diag\left(\lambda_1,\ldots,\lambda_{r+1}\right)$. Let

$$\mathbf{x}_t^*=\left[\begin{array}{c}x_{1t}\\ ..\\ x_{r+1t}\end{array}\right]=S^{-1}\left[\begin{array}{c}\mathbf{z}_t\\ y_t^*\end{array}\right];v_t^*=S^{-1}\left[\begin{array}{c}\boldsymbol{\zeta}_t\\ u_t\end{array}\right]$$

then we can write, with $h'=$ the bottom row of S,

$$\begin{aligned}y_t^* &= h'\mathbf{x}_t^*=h'\left[I-\Lambda^*L\right]^{-1}v_t^*\\ &\quad \text{where we must have}\\ u_t &= h'v_t^*\end{aligned}$$

[36]Not that the other solution to $h\left(R^2,V,V_{\widehat{y}}\right)=-1$, (which yields the minimum R^2 for given V and $V_{\widehat{y}}$) is a strictly decreasing function of $V_{\widehat{y}}$, hence under the assumptions of the Corollary, we cannot derive a lower bound for R^2.

hence

$$
\begin{aligned}
y_t^* &= h' \begin{bmatrix} \frac{1}{1-\lambda_1 L} & & & \\ & \frac{1}{1-\lambda_2 L} & & \\ & & \ldots & \\ & & & \frac{1}{1-\lambda_{r+1} L} \end{bmatrix} v_t^* \\
(1-\lambda_{r+1}L)\, y_t^* &= h' \begin{bmatrix} \frac{1-\lambda_{r+1} L}{1-\lambda_1 L} & & & \\ & \frac{1-\lambda_{r+1} L}{1-\lambda_2 L} & & \\ & & \ldots & \\ & & & 1 \end{bmatrix} v_t^* \\
&= h' \begin{bmatrix} \frac{(\lambda_1-\lambda_{r+1})L}{1-\lambda_1 L}+1 & & & \\ & \frac{(\lambda_2-\lambda_{r+1})L}{1-\lambda_2 L}+1 & & \\ & & \ldots & \\ & & & 1 \end{bmatrix} v_t^* \\
&= h' \left\{ \begin{bmatrix} \frac{\lambda_1-\lambda_{r+1}}{1-\lambda_1 L} & & & \\ & \frac{\lambda_2-\lambda_{r+1}}{1-\lambda_2 L} & & \\ & & \ldots & \\ & & & 0 \end{bmatrix} v_{t-1}^* + v_t^* \right\} \\
&= h' \begin{bmatrix} \lambda_1-\lambda_{r+1} & & & \\ & \lambda_2-\lambda_{r+1} & & \\ & & \ldots & \\ & & & 0 \end{bmatrix} x_{t-1}^* + u_t \\
y_t &= \beta' \mathbf{x}_{t-1} + u_t
\end{aligned}
$$

where $y_t = (1-\lambda_{r+1}L)\, y_t^*$; $\mathbf{x}_t = (x_{1t}, \ldots, x_{rt})'$.

Thus our general predictive regression with r predictors can arise from a VAR in $r+1$ underlying variables, where y_t is the underlying variable we wish to predict, y_t^*, in quasi-differenced form. Since u_t is the innovation to both y_t and y_t^* it is straightforward to amend our R^2 formulae in terms of y_t^*. It is also evident that, since y_t must be ARMA(r,r) in reduced form, y_t^* will be ARMA$(r+1,r)$, with identical parameters and innovation, ε_t, as y_t, but with an additional AR parameter.

Example S1 in Appendix B.1 follows by letting $\mathbf{z}_t$, and hence $\mathbf{x}_t$ be scalar processes.

A.6 Derivations for ARMA(1,1) example

Substituting from (32) into (31) gives the equivalent of (8),

$$(1-\lambda L)\, y_t = \beta v_{t-1} + (1-\lambda L)\, u_t \tag{77}$$

where the right-hand side is a composite MA(1) error process. Equating the first-order autocorrelation of $(1-\theta L)\,\varepsilon_t$ to that of the process on the right-hand side of (8) gives the single moment condition

$$\frac{\theta}{1+\theta^2} = \frac{\lambda - \rho F}{1+\lambda^2+F^2-2\lambda\rho F} \tag{78}$$

$$\text{where } F(R^2,\lambda) = \sqrt{\left(1-\lambda^2\right)\frac{R^2}{1-R^2}}$$

The MA parameter $\theta\left(R^2,\rho,\lambda\right)$ is then the solution to (78) in $(-1,1)$), and the predictive space can thus be reparameterised in terms of the scale-independent triplet $\left(R^2,\rho,\lambda\right)$.

The moment condition is also satisfied by θ^{-1}. The (unique) nonfundamental ARMA(1,1) representation is

$$y_t = \left(\frac{1-\theta^{-1}L}{1-\lambda L}\right)\eta_t \tag{79}$$

which has the standard property that the nonfundamental innovation η_t can only be recovered from current and future values of y_t, i.e., we have

$$\varepsilon_t = \left(\frac{1-\lambda L}{1-\theta L}\right) y_t = \sum_{i=0}^{\infty}\theta^i\left[y_{t-i}-\lambda y_{t-i-1}\right] \tag{80}$$

$$\eta_t = \left(\frac{1-\lambda L}{1-\theta^{-1}L}\right) y_t = -\theta L^{-1}\left(\frac{1-\lambda L}{1-\theta L^{-1}}\right) y_t = -\sum_{i=1}^{\infty}\theta^i\left[y_{t+i}-\lambda y_{t+i-1}\right] \tag{81}$$

thus as noted above, the nonfundamental representation is not a viable predictive model. However, its properties can be derived straightforwardly from those of the fundamental representation, with $\sigma_\eta^2 = \theta^2\sigma_\varepsilon^2$ (a special case of the general formula in (4)).

We can write the two ARMA representations as special cases of the system in

(32) and (31), as in Section 3.1, using

$$x_t^F = \frac{\varepsilon_t}{1-\lambda L}, \quad \beta_F = \lambda - \theta; \tag{82}$$

$$x_t^N = \frac{\eta_t}{1-\lambda L}, \quad \beta_N = \lambda - \theta^{-1} \tag{83}$$

The upper and lower bounds in (34) and (35) can then be derived straightforwardly from the standard R^2 formula applied to the predictors in (82) and (83).

To derive the limiting variance ratio, V, if we write the ARMA$(1,1)$ representation in (30) as $y_t = C(L)\varepsilon_t$, then the Beveridge-Nelson permanent component has variance $\sigma_P^2 = C(1)^2 \sigma_\varepsilon^2$, and hence

$$V(\lambda,\theta) = \left(\frac{1-\theta}{1-\lambda}\right)^2 \left(1 - R_F^2(\lambda,\theta)\right)$$

substituting from (34) gives

$$V(\lambda,\theta) = \frac{(1-\theta)^2(1+\lambda)}{\left(1-\lambda^2+(\theta-\lambda)^2\right)(1-\lambda)} \tag{84}$$

The formula for $V_{\widehat{y}}(\lambda)$ in Corollary 3 follows automatically by setting $\theta = 0$, since $\widehat{y}_t = \beta x_{t-1}$ is an AR(1).

B Supplementary Material for Examples

B.1 The predictive space for an ARMA(2,1)

The analytical example of an ARMA(1,1) in Section 4.1 can be straightforwardly extended to illustrate the case where $p > q$, as discussed Section 3.7.1. Assume that y_t is an ARMA$(2,1)$ driven by a white noise innnovation,

$$y_t = \left(\frac{1-\theta L}{(1-\lambda_1 L)(1-\lambda_2 L)}\right)\varepsilon_t \tag{85}$$

Under Assumptions A4 to A6 there is some underlying predictive system with $r = q = 1$, of the general form

$$y_t = \lambda_1 y_{t-1} + \beta x_{t-1} + u_t \tag{86}$$

$$x_t = \lambda_2 x_{t-1} + v_t \tag{87}$$

since, by quasi-differencing both (85) and (86)[37] it is evident that $(1-\lambda_2 L)\, y_t$ is an ARMA(1,1), hence, by application of Proposition 1 the predictive regression for $(1-\lambda_2 L)\, y_t$ (86) must satisfy $\sigma_u^2 \in \left[\theta^2\sigma_\varepsilon^2, \sigma_\varepsilon^2\right]$. Thus for (86) we have $R^2 = 1-\sigma_u^2/\sigma_y^2 \in [R_F^2, R_N^2]$ where the definitions of both upper and lower bounds are given by (2) and (5), as applied to an ARMA(2,1). The only modification of our earlier analysis is thus that these now represent bounds for any equation that conditions on a single lag of y_t as well as a single lagged predictor.

By inspection the system in (86) and (87) is a bivariate first order vector autoregression, with one-way Granger causality from x_{t-1} to y_t. This might appear to restrict the predictive space for y_t to VARs with one-way causality; but this is not the case. Any underlying bivariate VAR in y_t and some scalar predictor z_t, with autoregressive matrix $\mathbf{A}$, such that $a_{12} \neq 0, a_{21} \neq 0$ (implying two-way Granger Causality) can be expressed in the restricted form of (86) and (87), with $(\lambda_1, \lambda_2) = eig(\mathbf{A})$, and v_t is some combination of both underlying innovations, without changing the properties of the predictive error, u_t.[38] Thus for an ARMA(2,1) both Propositions 1 and 2 apply to the equation for y_t in *any* bivariate VAR(1).

Unsurprisingly, the predictability of y_t from the VAR may be very different from the ARMA(1,1) case, to the extent that the lagged dependent variable increases the predictive power of the predictive regression. The lower and upper bounds, $R^2_{\min} = R^2_F(\lambda_1, \lambda_2, \theta)$ and $R^2_{\max} = R^2_N(\lambda_1, \lambda_2, \theta)$ are more complicated functions of the ARMA parameters, but are still linked by $R^2_N = 1 - \theta^2(1 - R^2_F)$, hence the closer $|\theta|$ is to unity, the narrower is the space that R^2 can inhabit.

Figure A1 illustrates the impact of the additional AR parameter using the same parameterisations for θ and λ_1 as in Figure 1, ie we set $\lambda_1 = 0.8$, with $\theta = 0.9$ in Panel A (as for y_{1t} in the first example) and θ =0.7 in Panel B (as for y_{2t}). We then vary λ_2 between zero and unity: thus the intercepts on the x-axis in the two panels of Figure A1 are equal to those on the y-axis in Figure 1. Figure A1 shows that, except for very low values of λ_2 for y_{1t}, the gap between the upper and lower bounds for R^2 for both processes falls monotonically as λ_2 rises.[39]

Figure A1 illustrates, particularly in the case shown in Panel A, that the pre-

[37] Note that the choice of AR parameter in quasi-differencing, whilst it clearly changes the properties of the system in (86) and (87) has no impact on our results for the underlying process.

[38] See Appendix A.5, which derives this feature for a general VAR(1) with r predictors, with unrestricted Granger Causality.

[39] For sufficiently low λ_2 both upper and lower bounds for R^2 for y_{1t} illustrated in Panel A of Figure 2 initially *fall* as λ_2 increases. This may appear paradoxical, but this feature arises because, by holding λ_1 and θ constant, and raising λ_2, we are in effect considering a range of different histories of y, as captured by the ARMA representation. As λ_2 rises above zero, y_t initially becomes closer to white noise. Straightforwardly this must push down $R^2_{\min}$, and, since we are holding θ constant, it must also push down $R^2_{\max}$.

dictive space may be very tightly constrained even when, in contrast to our first example, a process may be strongly predictable. The key issue is the extent to which that additional predictability arises from the history of the process itself - if so, the marginal contribution to predictive power of observing x_t, the single predictor in (86) and (87), compared to simply forecasting using the fundamental ARMA representation, may still be very limited.

B.2 Stock & Watson's (2007) model of inflation: the predictive space with multiple predictors

As discussed in Section 5.2, in finite samples we cannot rule out the possibility that, while a representation of the process $y_t = \Delta\pi_t$ as an MA(1) may, as SW find, match the data in their sub-samples, this may be consistent with the true ARMA process being higher order, as long as the additional AR and MA parameters are sufficiently close either to zero or cancellation. To illustrate, we first consider the case that the true univariate process for y_t is an ARMA(2,2), implying, under our maintained assumptions, a predictive model with two predictors. We also briefly consider the implications of this analysis for $r > 2$ predictors

Since, by assumption, the true ARMA representation is in effect unobservable, we need to consider the predictive space for the univariate properties that we can actually observe, taking into account the range of variability of parameter estimates in finite samples. We focus on two univariate properties that summarise SW's univariate representation.

The first is the univariate R-squared, R_F^2, which, from Proposition 1, is the lower bound for R^2 in any predictive model. In Table 1 we reported implied values $R_F^2\left(\widehat{\theta}\right)$ and asymptotic standard errors, given SW's estimates of the single MA parameter, $\widehat{\theta}$. Since SW found that this representation could not be rejected against higher alternatives we assume that the true higher order ARMA representation must still have a value of R_F^2 within the 90% confidence interval for $R_F^2\left(\widehat{\theta}\right)$.[40]

However, in principle simply matching univariate R-squareds could admit ARMA representations that imply distinctly different long-run properties from SW's representation. As noted in Section 4.2, SW do not estimate the MA(1) representation directly, but instead estimate an unobserved components representation for $\pi_t = Y_t = (1-L)^{-1} y_t$ that fits within the general form of the Beveridge-Nelson(1981)

[40] Here we use true $R_F^2(\boldsymbol{\lambda}, \boldsymbol{\theta})$ values. Clearly there are additional sampling issues but this exercise is intended only for illustrative purposes.

permanent-transitory decomposition in (24) in the main text, restated here,

$$Y_t = Y_t^P + Y_t^T \tag{88}$$

where SW specify as an identifying assumption that the transitory component of inflation Y_t^T is white noise, orthogonal to the permanent innovation: this is equivalent to an MA(1) representation of $y_t = \Delta Y_t$, with θ constrained to be non-negative.[41] (In SW' time-varying representation both innovation variances are themselves modelled as random walks in log terms, but we focus here for simplicity on the fixed coefficient representations). It is straightforward to show that for this representation the limiting variance ratio is given by

$$V = \frac{\varsigma}{2+\varsigma}$$

where $\varsigma = var\left(\Delta Y_t^P\right)/var\left(Y_t^T\right) > 0$ is the signal to noise ratio. By construction therefore $V < 1$ for any value of ς. The point estimates of θ reported in Columns 1 and 2 of Table 1 correspond to point estimates of ς of 1.9 and 0.18 respectively, and hence implied values of V of 0.49 and 0.08.

We can thus exploit Definition 3 and consider the predictive space

$$\mathbb{P}_{\mathbb{U}} \subset \mathcal{P}_r : \mathbb{U} = \left\{ V \in \left(\widehat{V} \pm 1.65 s.e.\left(\widehat{V}\right)\right), R_F^2 \in \left(\widehat{R}_F^2 \pm 1.65 s.e.\left(\widehat{R}_F^2\right)\right)\right\} \tag{89}$$

where the range of values of V and R_F^2 implied by the predictive model must lie within the 90% confidence intervals implied by SW's point estimates of θ, which we take as $\widehat{V} = V\left(\widehat{\theta}\right)$ and $\widehat{R}_F^2 = R_F^2\left(\widehat{\theta}\right)$ (in both cases, standard errors are approximated using the delta method). To illustrate, we use the confidence intervals for V and R_F^2 implied by SW's estimates in their second sub-sample, shown in Column 2 of the Table 1, given by $\widehat{\theta} = 0.656$, $s.e.\left(\widehat{\theta}\right) = 0.088$, since in this sample we have a very clear rejection of the null that y_t is IID.

To derive the properties of the predictive space $\mathbb{P}_{\mathbb{U}}$ we first construct a large number of predictive models within the parameter space $\mathcal{P}_r$ of all possible predictive models with r predictors, as defined in Definition 1, by sampling from independent uniform distributions for each of the parameters in $\mathbf{\Psi}$ over their permissible ranges (for precise details see Appendix B.3). If the resulting parameter vector in draw s, $\mathbf{\Psi}_s$ satisfies $\left(V\left(\mathbf{\Psi}_s\right), R_F^2\left(\mathbf{\Psi}_s\right)\right) \in \mathbb{U}$, then $\mathbf{\Psi}_s \in \mathbb{P}_{\mathbb{U}}$. We can then consider the

[41] In contrast to the original Beveridge-Nelson (1981) derivation from an estimated ARMA model, in which the identification assumption was that the two innovations are perfectly correlated. However we follow Cochrane (1988) in taking the key property of the decomposition to be that the permanent component is a random walk.

properties of predictive systems that do satisfy these restrictions, and hence, to within a reasonable range of sampling variation, are consistent with SW's estimated univariate representation.

Figure A2 illustrates the relatively simple case with $r = 1$. In terms of SW's unobserved components representation the only modification this implies is that in (61) we allow Y_t^T to be an AR(1), rather than pure white noise, and hence y_t is now an ARMA(1,1), as analysed in Section 4.1.

A convenient feature of the $r = 1$ case is that the predictive space $\mathbb{P}_{\mathbb{U}} \subset \mathcal{P}_1$ has positive volume in three dimensions; thus our simulation methodology is here equivalent to Monte Carlo integration. We take 10 million draws from uniform distributions of the 3 underlying parameters, parameterised as in Section 4.1 as $\mathbf{\Psi} = (R^2, \rho, \lambda)$. Panel A shows a three-dimensional view of the predictive space thus derived. Given the limitations of three-dimensional graphic displays, Panels B to D give 2 dimensional representations that clarify certain features of the space.

It is helpful in considering the information shown in Figure A2 to define

$$v = \frac{V}{1+V} \in (0,1) \tag{90}$$

since then the set containing all logically possible (R_F^2, V) combinations can be represented by (R_F^2, v) which must lie within the unit square. On this transformed basis the set $\mathbb{U}$ defined in (89) is a rectangle containing roughly 2.5% of the total possible space: thus even allowing for sampling variation we would also expect to rule out a significant proportion of the potential parameter space of single predictor models, $\mathcal{P}_1$. But it turns out that the predictive space $\mathbb{P}_{\mathbb{U}}$ actually represents a distinctly smaller fraction (only around 0.75%) of $\mathcal{P}_1$ (here given by $\mathcal{P}_1 = (-1,1) \times [-1,1] \times [0,1]$, the cuboid contained within the 3 axes).

Panel A shows a 3-dimensional scatter plot of the predictive space $\mathbb{P}_{\mathbb{U}} \subset \mathcal{P}_1$. Each of the simulated predictive models that satisfy the restrictions in (89) is shown as a single point. All three parameters (R^2, ρ, λ) are bounded in at least one direction. These bounds can be best understood by first considering the permissible space of the true ARMA$(1,1)$ parameters, (λ, θ). In Stock and Watson's representation, λ is constrained to be precisely zero. Figure A2 shows that, given sampling variation, this is consistent with the true value of λ differing from zero, but that all possible values must lie roughly within $[-0.5, 5]$. This feature arises from the expressions for $V(\lambda, \theta)$ and $R_F^2(\lambda, \theta)$ given in (84) and (34). For a given value of $V < 1$, the higher is λ, the *lower* is R_F^2; thus for sufficiently positive λ, the implied univariate properties will be inconsistent with the degree of univariate predictability (a point

estimate $R_F^2\left(\widehat{\theta}\right) = 0.3$) implied by SW's estimates. Thus a strongly persistent single predictor of inflation can be categorically ruled out.[42] Equally, if λ were too far below zero, there would be too *much* implied univariate predictability. As a result any AR(1) single predictor of inflation must itself have fairly low univariate predictability, with a maximum univariate R^2 of around 0.25 (with the IID predictor implied by SW's restricted representation as a special case). Crucially, also, for the full range of feasible values of λ, the MA parameter θ must be strictly positive to satisfy the two restrictions.

Given the permissible space for (λ, θ), the restrictions on the remaining two parameters in the predictive space, R^2 and ρ, are easy to interpret in terms of Propositions 1 and 2 (which, as Corollary 3 showed, give identical results for the ARMA(1,1) case). For any permissible (λ, θ) pair, R^2 must lie within the upper and lower bounds of Proposition 1, which are in turn strictly within $[0, 1]$ given that $\theta \neq \lambda$ and θ cannot be zero (i.e., a true AR(1) representation is ruled out); ρ also has an absolute upper bound well below zero.

Panels B to D of Figure A2, give 2 dimensional views that bring out more clearly the restrictions on each of the three parameters. While Panel B[43] shows that the range of feasible (R^2, ρ) combinations within $\mathbb{P}_{\mathbb{U}}$ is wider than that implied by Table 1, Panels C and D show that this wider range arises largely due to *negative* values of λ. If on *a priori* grounds we wished to consider only cases with $\lambda \geq 0$, Panel C shows that the absolute upper bounds for R^2 and ρ, even given sampling variation, would be very similar to those given in Table 1.[44]

Consider now the case with $r = 2$ predictors. We again make 10 million random draws from the parameter space of all possible 2 predictor models, $\mathcal{P}_2$. Only those predictive models with parameters $\mathbf{\Psi}$ that map to $\mathbf{u} = (R_F^2, V) \in \mathbb{U}$ defined in (89) lie within the predictive space $\mathbb{P}_{\mathbb{U}} \subset \mathcal{P}_2$.

While the predictive space is now in seven dimensions (see Appendix B.3 for precise details), we can still represent it in 3 dimensions, in terms of the same combination of summary properties of predictive systems, $(R^2, \rho, V_{\widehat{y}})$ that we showed must be constrained by Proposition 2. Panel A of Figure A3 shows a three dimensional

[42] A high λ would be more consistent with, for example, data on stock returns or real exchange rate changes, which also appear to have $V < 1$, but have minimal univariate predictability.

[43] This is directly comparable with Figure 1, which illustrated our ARMA(1,1) example in Section 4.1.

[44] Corollaries 2 and 3 provides a straightforward explanation of both features. SW's restriction that $\lambda = 0$ imposes the property that all predictors must be IID, and hence that the variance ratio of the predicted value, $V_{\widehat{y}} = 1$. For $\lambda > 0$, $V_{\widehat{y}}(\lambda) = \frac{1+\lambda}{1-\lambda} > 1$. By inspection of the expression for $\rho_{\max}$ in Corollary 3, the upper bound for ρ is a decreasing function of $V_{\widehat{y}}$, and hence of λ. The upper bound for R^2 in Corollary 2 also implies that, for any $V_{\widehat{y}} > 1$, the implied upper bound for R^2 falls.

scatter plot of the triplet $(R^2(\mathbf{\Psi}), \rho(\mathbf{\Psi}), v_{\widehat{y}}(\mathbf{\Psi}))$ implied by every predictive model with parameters $\mathbf{\Psi}$ that lie within the predictive space, where we define $v_{\widehat{y}} = \frac{V_{\widehat{y}}}{1+V_{\widehat{y}}}$, consistently with v, defined in (90) so that all three elements are bounded. Panels B to D give alternative 2-dimensional views to clarify certain characteristics.

The comparison with Figure A2 is relatively straightforward, since for that case, with $r = 1$, we have $v_{\widehat{y}} = \frac{1+\lambda}{2}$, hence Panels A, C and D of Figure A2 are directly comparable with their equivalents in Figure A3, but for a relabelling of the axes. On the basis of this comparison there are evident similarities in the shape of the predictive space, but also some clear differences.

A first, and crucial, similarity is that the predictive space, as summarised in these three properties, again occupies only a very small proportion of the potential parameter space, $\mathcal{P}_2$ of all predictive models, which maps to any point in the cuboid contained within the axes of Panel A. There are also clear similarities in the pattern of feasible $(R^2, \rho, v_{\widehat{y}})$ combinations for most values of R^2 within its feasible range. Thus, for virtually all predictive models with an R^2 within the feasible range for single predictor models, as shown in Table 1, ρ must also lie within a similar range to the single predictor case.

However in contrast to the absolute upper bound for R^2 shown in Figure A2 for the single predictor case, Figure A3 shows that in the 2 predictor case, for some $(R^2, \rho, v_{\widehat{y}})$ combinations, R^2 can be arbitrarily close to unity, ρ can lie anywhere in $[-1, 1]$, and $v_{\widehat{y}}$ can be arbitrarily close to zero. The necessary properties of the ARMA$(2, 2)$ representations of all the y_t processes implied by Figure A3 again provide the explanation.

The necessary link with ARMA properties arises straightforwardly from Propositions 1 and 2. All predictive models in $\mathbb{P}_{\mathbf{U}} \subset \mathcal{P}_2$ must map to ARMA representations that in turn imply (V, R_F^2) combinations consistent with the properties of inflation. This clearly restricts the permissible space of the additional ARMA parameters. But, in contrast to the $r = 1$ case, it does *not* bound both the MA parameters away from zero. Thus for some predictive systems within the predictive space the ARMA representation can in principle be arbitrarily close to the limiting case a) of Corollary 1 in which $\theta_i \to 0$ for some i, and hence $R^2_{\max} \to 1$. From Proposition 2, this *requires* $V_{\widehat{y}}$ to approach V, which here must lie in a range close to zero. Furthermore, for sufficiently high R^2, the properties of the predictive error u_t become irrelevant, and hence ρ can live anywhere in $[-1, 1]$.

However, this does *not* remove the restrictions on the predictive space. To attain (R^2, ρ) combinations outside the feasible range of single predictor models puts extremely tight restrictions on the underlying predictive model.

Panels E and F of Figure A3 show, first, that a very large proportion of the predictive space contains predictive models in which at least one, and in many cases both of the predictors are AR(1) processes with $\lambda_i < 0$. This is particularly the case for predictive models with relatively high R^2, which Panel D shows *require* at least one negative λ_i. Nor is this feature surprising in the light of our discussion of Proposition 2, where we noted that for a predictive regression to achieve a high R^2 requires that $V_{\widehat{y}}$, the limiting variance ratio of the predictive values, be close to V, the limiting variance ratio for y_t itself. This feature is evident in Panels C and D of Figure A3, which illustrate the upper bounds for both R^2 and ρ implied by any predictive model with $V_{\widehat{y}} \geq 1$ (and hence $v_{\widehat{y}} > \frac{1}{2}$) given in Corollary 2.[45] While a predictive model with positive values of both λ_1 and λ_2 does not of itself rule out $V_{\widehat{y}} < 1$ it *does* require that ρ_{12}, the correlation between predictor innovations, be sufficiently negative to offset the impact of positive persistence of both predictors.[46] As a result the common *a priori* assumption that predictors have positive persistence (as discussed by SW this is also a feature also of many observed candidate predictors) would rule out much of the predictive space, and leave R^2 and ρ constrained to lie in a very similar range to the single predictor case.

The apparent expansion of the predictive space that allows R^2 to be arbitrarily close to unity for some parameter combinations also turns out to be largely illusory. Panels E to G of Figure A3 show that, as R^2 increases, λ_1 and λ_2 and ρ_{12} all become increasingly constrained. Most strikingly, as $R^2 \to 1$, $\rho_{12} \to -1$. Since, for sufficiently high R^2, the properties of u_t become irrelevant, this feature tells us that all such models must closely resemble ARMA models. This is again what we would expect from Proposition 1, since for such cases we must also have $R^2 \to R^2_{\max} = R^2_N$, the R^2 of the nonfundamental ARMA representation. But in the particular case where $R^2_N \to 1$, because $\theta_i \to 0$ for some i, it is straightforward to show that, in the limit, η_t, the nonfundamental innovation, is simply equal to ε_{t+1}, the fundamental innovation in the next period. It is hardly surprising that, if we could observe ε_{t+1} in period t (whether by divine intervention or by use of a a time machine) we would be able predict y_{t+1} with $R^2 = 1$; but this is of not much practical benefit of forecasting. But Figure A3 tells us that, for inflation, at least, this would, to a quite close approximation, be the *only* way to achieve an R^2 approaching unity.

[45] The range of permissible values of V given in (89) gives a maximum possible value of around 0.17, to be consistent with SW's representation. Using Corollary 2 this implies that, for any predictive model with $V_{\widehat{y}} \geq 1$, the absolute maximum R^2 is just under 0.8, as illustrated in Panel D of Figure 2, and the absolute maximum $\rho = -0.83$.

[46] The proof of Corollary 2 shows that a sufficient condition for $V_{\widehat{y}}$ to be strictly greater than unity is if just one of the inequalities on λ_1, λ_2 and ρ_{12} holds in strong form. Hence if $\lambda_1 > 0, \lambda_2 > 0$, then $V_{\widehat{y}} < 1$ requires $\rho_{12} < 0$.

It seems unlikely that the message of this exercise would change if we increased the number of predictors further, and considered general models with $r \geq 2$. The constraints on the three summary properties $(R^2, \rho, v_{\widehat{y}})$ illustrated in Figure A3 would still apply for any r. And any additional predictors would still need to satisfy restrictions such that additional implied MA and AR parameters would be close to cancellation or zero. Thus Stock and Watson's results imply that the predictive space must be very tightly constrained for any r.

B.3 Background material for extended Stock-Watson example

The space of all possible predictive systems for a general 2 predictor model, $\mathcal{P}_2$, can be represented in terms of the seven parameters $(\lambda_1, \lambda_2, \beta_1, \beta_2, \rho_{1u}, \rho_{2u}, \rho_{12}) \in \mathcal{P}_2$ that satisfy

$$\begin{aligned}
\beta_1, \beta_2 &\in [0, \infty) && (91)\\
\rho_{1u}, \rho_{12} &\in [-1, 1] && (92)\\
\rho_{2u} &\in \left[\rho_{1u} \cdot \rho_{12} \pm \sqrt{\left(1-\rho_{1u}^2\right)\left(1-\rho_{12}^2\right)}\right] && (93)\\
\lambda_1, \lambda_2 &\in (-1, 1) && (94)
\end{aligned}$$

These parameters jointly determine the MA parameters θ_1 and θ_2, which satisfy two moment conditions derived from the general form of (53):

$$\begin{aligned}
\frac{-\left(1+\theta_1\theta_2\right)\left(\theta_1+\theta_2\right)}{1+\left(\theta_1+\theta_2\right)^2+\theta_1^2\theta_2^2} &= \frac{\gamma_0'\mathbf{\Omega}\gamma_1+\gamma_1'\mathbf{\Omega}\gamma_2}{\gamma_0'\mathbf{\Omega}\gamma_0+\gamma_1'\mathbf{\Omega}\gamma_1+\gamma_2'\mathbf{\Omega}\gamma_2} && (95)\\
\frac{\theta_1\theta_2}{1+\left(\theta_1+\theta_2\right)^2+\theta_1^2\theta_2^2} &= \frac{\gamma_0'\mathbf{\Omega}\gamma_2}{\gamma_0'\mathbf{\Omega}\gamma_0+\gamma_1'\mathbf{\Omega}\gamma_1+\gamma_2'\mathbf{\Omega}\gamma_2} && (96)
\end{aligned}$$

where the γ_i, as defined implicitly for the general case in (50), are given by

$$\gamma_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}; \quad \gamma_1 = \begin{bmatrix} -\left(\lambda_1+\lambda_2\right) \\ \beta_1 \\ \beta_2 \end{bmatrix}; \quad \gamma_2 = \begin{bmatrix} \lambda_1\lambda_2 \\ -\beta_1\lambda_2 \\ -\beta_2\lambda_1 \end{bmatrix}$$

We constrain the β_i to be positive, since we allow correlation coefficients to be both positive and negative (this is a pure normalisation; we could equally well leave the β_i unconstrained and constrain the correlation coefficents to be non-negative). We can also assume $\sigma_u = \sigma_1 = \sigma_2 = 1$ without loss of generality since the units of y_t are irrelevant to our results, so that the off–diagonal elements of Ω can be correlation

coefficients. One of the correlation coefficients must live in a restricted space, relative to the other two, to ensure that Ω is positive semi-definite; for convenience we choose this to be ρ_{2u}.

Exploiting the normalisation that $\sigma_u^2 = \sigma_v^2 = 1$,the predictive R^2 and ρ are given by

$$\begin{aligned} R^2 &= R^2\left(\beta_1, \beta_2, \lambda_1, \lambda_2, \sigma_{12}\right) = \frac{\sigma_{\widehat{y}}^2}{1+\sigma_{\widehat{y}}^2} \text{ where} \qquad (97) \\ \sigma_{\widehat{y}}^2 &= \frac{\beta_1^2}{1-\lambda_1^2} + \frac{\beta_2^2}{1-\lambda_2^2} + \frac{2\beta_1\beta_2\rho_{12}}{1-\lambda_1\lambda_2} \end{aligned}$$

$$\rho = \left[\begin{array}{cc} \boldsymbol{\delta}' & 0 \end{array}\right] \Omega \left[\begin{array}{c} \mathbf{0} \\ 1 \end{array}\right] \left(\left[\begin{array}{cc} \boldsymbol{\delta}' & 0 \end{array}\right] \Omega \left[\begin{array}{c} \boldsymbol{\delta} \\ 0 \end{array}\right]\right)^{-1} \qquad (98)$$

To generate Figures A2 and A3, we randomly sample all parameters from uniform distributions, or transformations thereof, to ensure a bounded parameter space, as follows:

1. We draw λ_1 and λ_2 from independent $U\left(-1,1\right)$ distributions;

2. Two of the correlations, ρ_{12} and ρ_{1u} are drawn from independent $U\left(-1,1\right)$ distribution

3. For the third correlation, ρ_{2u}, (93) implies

$$\rho_{2u} = \rho_{1,u}\rho_{12} + k\sqrt{\left(1-\rho_{1,u}^2\right)\left(1-\rho_{12}^2\right)}, k \in [-1,1]$$

   thus we also draw k from a $U\left(-1,1\right)$.

4. To ensure a bounded parameter space we derive the two regression coefficients β_1 and β_2 as monotonically increasing transformations of "semi-R^2s": notional R^2 values if each predictor was the sole predictor in the regression, ie, given the normalisations we have

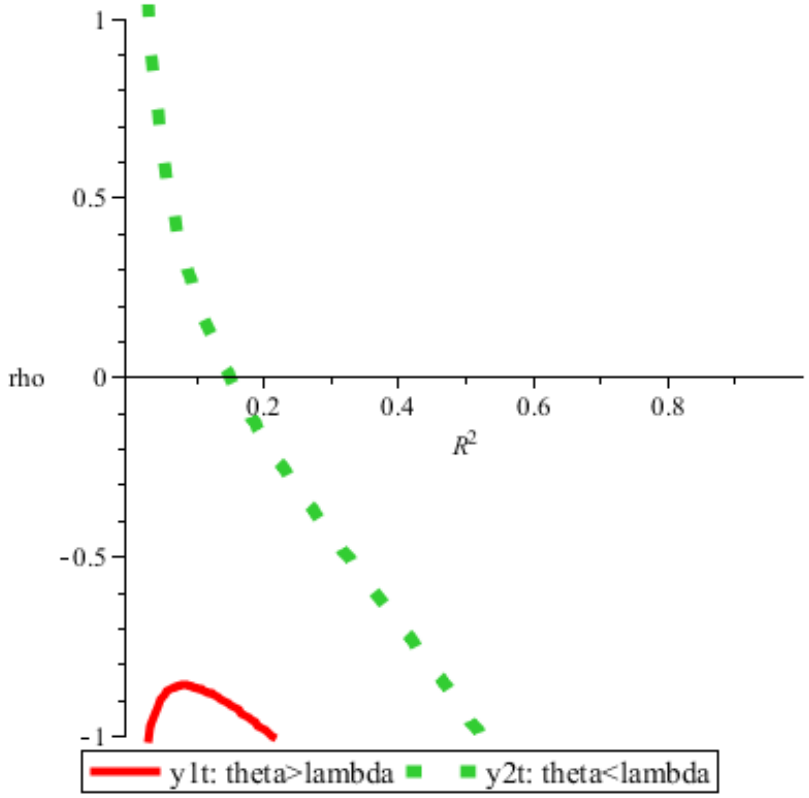
$$R_i^2 = \frac{\beta_i^2}{1-\lambda_i^2+\beta_i^2};\ i = 1,2$$

   and hence we draw R_1^2 and R_2^2 from $U\left(0,1\right)$ distributions, and then derive

$$\beta_i = \sqrt{(1-\lambda_i^2)\frac{R_i^2}{1-R_i^2}}; i = 1,2$$

In Figure A2, we set $\lambda_2 = \beta_2 = \rho_{12} = \rho_{2u} = 0$, thus restricting ourselves to ARMA(1,1) models; in Figure A3 we draw from the full range of values in $\mathcal{P}_2$, as defined above.

Figure 1 The Predictive Space $\mathbb{P}_{\lambda,\theta}$ for two ARMA(1,1) processes



Notes to Figure 1:
Figure 1 illustrates the predictive space $P_{\lambda,\theta}$: combinations of predictive R^2 and ρ (the Stambaugh-Beveridge-Nelson Correlation) for single predictor models consistent with an ARMA(1,1) reduced form with AR parameter λ and MA parameter θ (see Section 4.1 for full details)
The solid line represents the process y_{1t}, with $\lambda = 0.8$, $\theta = 0.9$ (hence $V < 1$)
The dashed line represents the process y_{2t}, with $\lambda = 0.8$, $\theta = 0.7$ (hence $V > 1$)

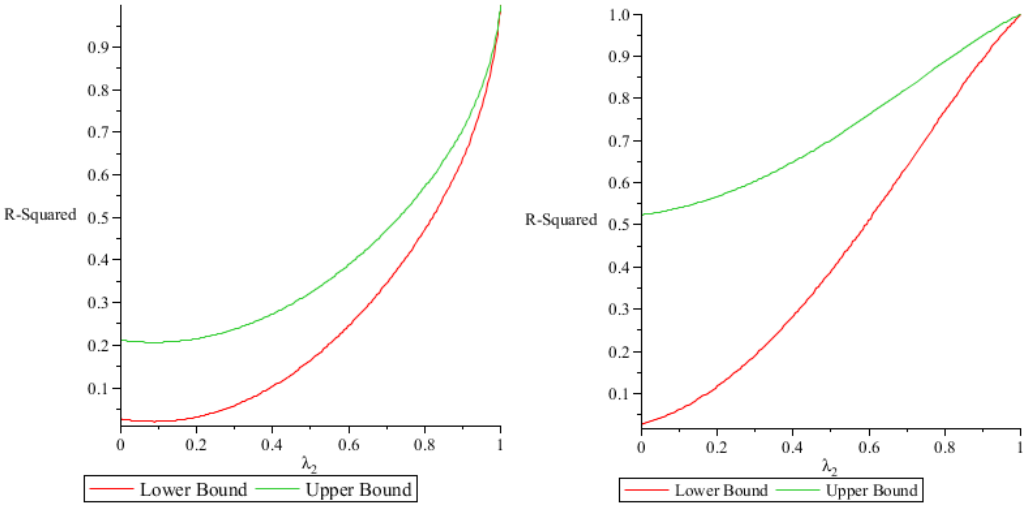
Table 1. The Predictive Space for US GDP Inflation implied by Stock & Watson's (2007) Univariate Representations.

Estimates	ARMA estimation in subsamples (standard errors in brackets)		Time-varying estimation by unobserved components: quantiles of posterior distribution, 2004:IV		
	60:1-83:IV	84:1-2004:IV	16.5%	50%	83.5%
$\hat{\theta}$	0.275 (0.085)	0.656 (0.088)	0.70	0.85	0.94
$R^2_{\min}\left(\hat{\theta}\right)$	0.070 (0.040)	0.301 (0.042)	0.329	0.419	0.469
$R^2_{\max}\left(\hat{\theta}\right)$	0.930 (0.040)	0.699 (0.042)	0.671	0.581	0.531
$\rho_{\max}\left(\hat{\theta}\right)$	-0.511 (0.136)	-0.917 (0.049)	-0.940	-0.987	-0.998

Notes to Table 1. Estimates in first row of Table 1 are derived from Stock & Watson (2007), Table 3 (columns and 2) and Figure 2 (columns 3 to 5). Remaining rows use formulae in equations (35), (36) and (38), setting λ=0, and θ equal to the estimated value in the relevant column of the first row. Standard errors for $\hat{\theta}$ are as reported by Stock and Watson; standard errors in remaining rows of columns 1 and 2 are approximated using the delta method.

Figure A1. Univariate bounds for R^2 for an element of a bivariate first-order VAR

(y_t is ARMA(2,1) in reduced form)



Panel A: $\lambda_1 = 0.8$, $\theta = 0.9$ **Panel B**: $\lambda_1 = 0.8, \theta = 0.7$

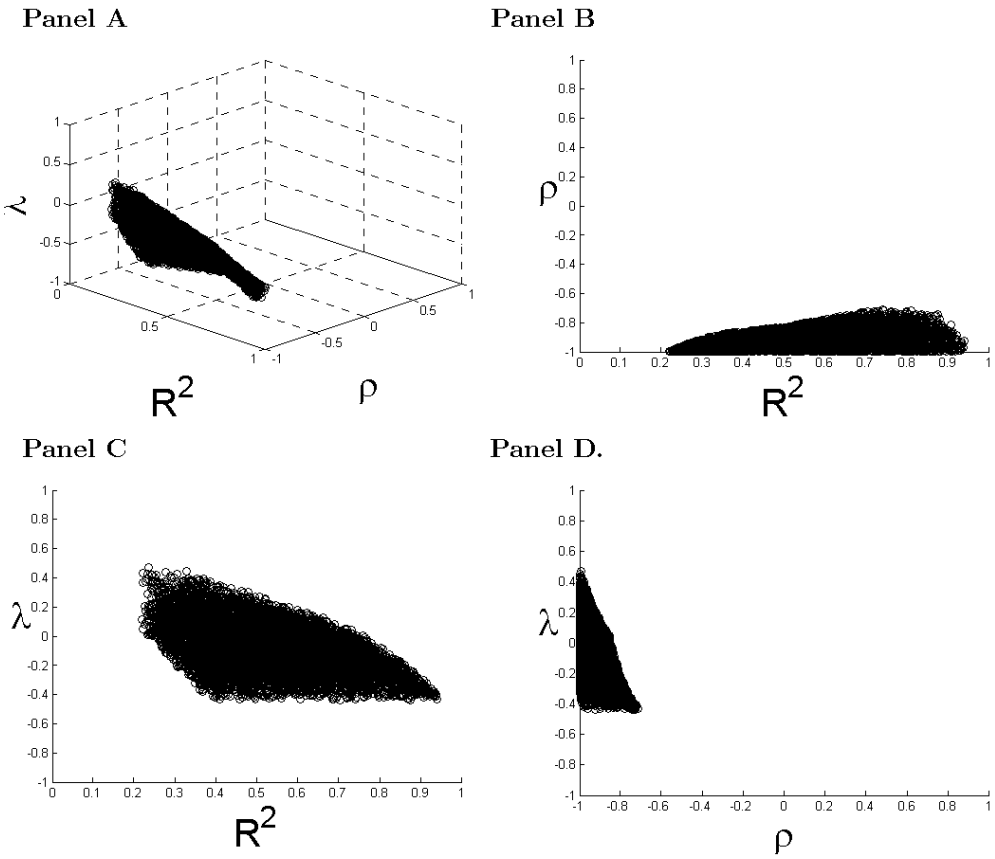
Notes to Figure A1

Figure A2 illustrates the nature of the R^2 bounds from Proposition 2 for a y_t process that is an ARMA(2,1) process

$$y_t = \left(\frac{1 - \theta L}{(1 - \lambda_1 L)(1 - \lambda_2 L)} \right) \varepsilon_t$$

as described in Appendix B1. We set θ and λ_1 equal to the values in the example in Section 4.1 of the main paper: Panel A corresponds to the y_{1t} process in Figure 1; Panel B corresponds to the y_{2t} process. Panels A and B illustrate how the R^2 bounds vary as the second AR parameter, λ_2, varies between zero and 1.

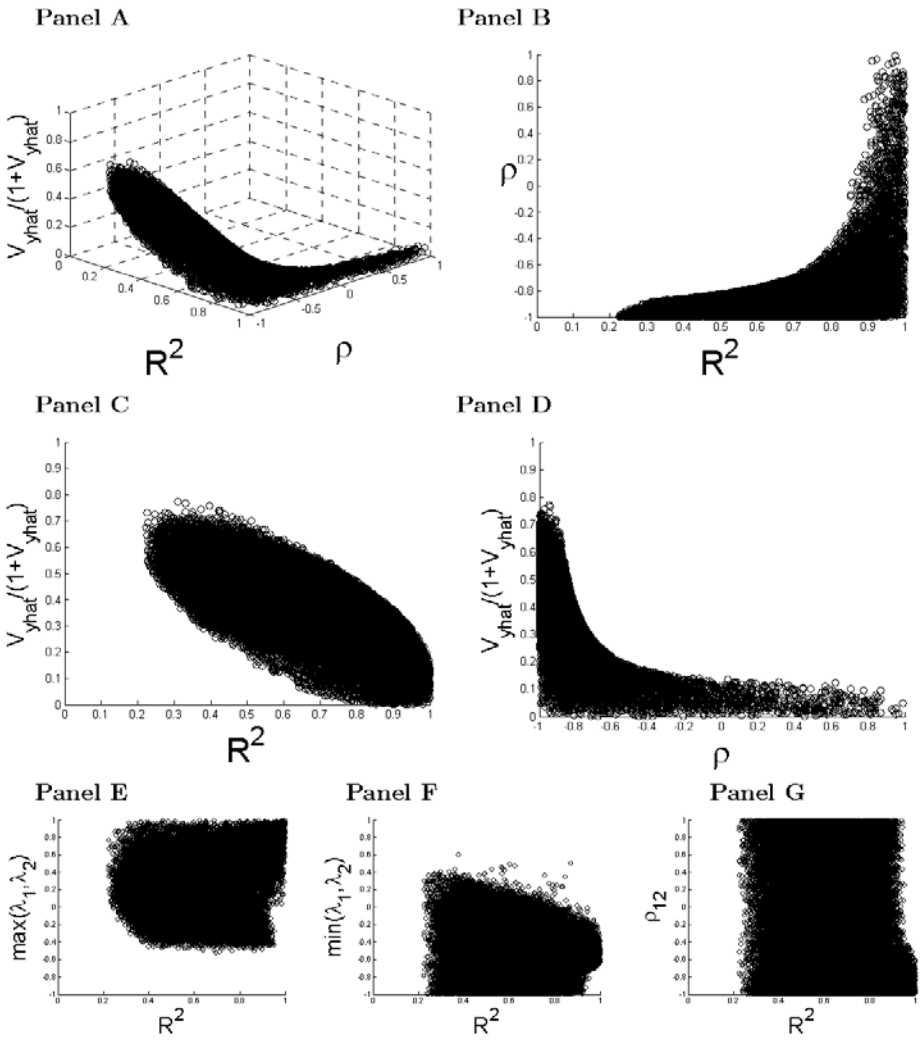
Figure A2: The predictive space $\mathbb{P}_{\mathbb{U}} \subset \mathcal{P}_1$ for a single predictor model of $y_t = \Delta\pi_t$, consistent with Stock and Watson's (2007) univariate representation



Notes to Figure A2

Figure A2 shows alternative 3-dimensional and 2-dimensional views of the predictive space $\mathbb{P}_{\mathbb{U}}$ as defined in (89) in Appendix B2: parameter combinations $\left(\lambda, \rho, R^2\right)$ of single predictor models consistent with Stock & Watson's (2007) univariate representation of inflation. Each point in the scatterplot represents the parameters of a predictive model within the predictive space.

Figure A3. The predictive space $\mathbb{P}_{\mathbb{U}} \subset \mathcal{P}_2$ for a two predictor model of $y_t = \Delta\pi_t$, consistent with Stock and Watson's (2007) univariate representation



Notes to Figure A3

Figure A3 shows alternative 3-dimensional and 2-dimensional views of the predictive space $P_{\mathbb{U}} \subset P_2$ as defined in (89) in Appendix B2: combinations of the three summary properties $\left(v_{\widehat{y}}\left(\mathbf{\Psi}\right), \rho\left(\mathbf{\Psi}\right), R^2\left(\mathbf{\Psi}\right)\right)$ of predictive models with two predictors, with parameters Ψ, consistent with Stock & Watson's (2007) univariate representation of inflation (where $v_{\widehat{y}} = V_{\widehat{y}}/(1 + V_{\widehat{y}})$). Each point in the scatterplot represents the parameters of a predictive model within the predictive space.