National Institute Economic Review (2024), 1–25
doi:10.1017/nie.2024.1

NIER CAMBRIDGE UNIVERSITY PRESS

ARTICLE

HIGH-DIMENSIONAL FORECASTING WITH KNOWN KNOWNS AND KNOWN UNKNOWNS

M. Hashem Pesaran[1,2] and Ron P. Smith[3]

[1]University of Southern California, Los Angeles, CA, USA; [2]Trinity College, Cambridge, UK and [3]Birkbeck, University of London, London, UK
Corresponding author: M. Hashem Pesaran; Email: mhp1@cam.ac.uk

Abstract
Forecasts play a central role in decision-making under uncertainty. After a brief review of the general issues, this article considers ways of using high-dimensional data in forecasting. We consider selecting variables from a known active set, known knowns, using Lasso and **O**ne **C**ovariate at a time **M**ultiple **T**esting, and approximating unobserved latent factors, known unknowns, by various means. This combines both sparse and dense approaches to forecasting. We demonstrate the various issues involved in variable selection in a high-dimensional setting with an application to forecasting UK inflation at different horizons over the period 2020q1–2023q1. This application shows both the power of parsimonious models and the importance of allowing for global variables.

Keywords: forecasting; high-dimensional data; Lasso; OCMT; latent factors; principal components

JEL codes: C53; C55; E37; E52

> 'All the business of war, indeed all the business of life, is endeavour to find out what you don't know by what you do; that's what I called "*guess what was the other side of the hill*".
>
> Duke of Wellington

1. Introduction

Forecasts play a central role in decision-making under uncertainty. Good forecasts are those that lead to good decisions, in the sense that the expected payoff to the decision-maker using the forecast is greater than it would be otherwise.[1] In the case of inflation forecasts, which we consider below, the Bank of England makes forecasts to help it set monetary policy to keep inflation within a target range.[2] The payoff is the variation of inflation around the target. However, it is not clear how one would quantify the contribution of the forecast to the payoff in terms of a specific central bank loss function.

Since forecasts are designed to inform decisions, they are inherently linked to policy making. However, there is an issue as to whether one should use the same model for both forecasting and setting the policy instruments. Different questions require different types of models to answer them. A policy model might be quite large, while a forecasting model might be quite small. There is also an issue of how

[1]The linkage between forecasting and decision-making is discussed in Granger and Pesaran (2000a,b), who argue in favour of a closer link between the decision and forecast evaluation problems. Pesaran and Skouras (2004) provide a more general survey of decision-theoretic approaches to forecast evaluation.

[2]The letter of 26 June 2023 by Huw Pill, Bank chief economist to the chair of the House of Commons Treasury Committee sets out his assessment of the role played by the forecasts in the policy process.

transparent the model should be. It may be difficult to interpret why a machine learning statistical model makes the predictions it does, and this can be a major disadvantage when a policy requires communication of a persuasive narrative.

In recent years, forecasting has been influenced by the increasing availability of high-dimensional data, improvements in computational power and advances in econometrics and machine learning techniques. In some areas, such as meteorology, this has resulted in improved forecasts, increasing the number of hours ahead for which accurate predictions can be made. The improved forecasts lead to better decision-making as people change their behaviour in response to the predictions, and the effect of such responses on mortality from heat and cold is examined in Shrader *et al.* (2023). Despite advances in data, computation and technique, the improvement in accuracy of weather forecasts has not been matched by economic forecasts. This is a cause for concern, since as emphasised in the classic analysis of Whittle (1983), prediction and control are inherently linked and decisions over such elements of economic management such as monetary policy are dependent on a view of the future.

Macroeconomic forecasting is challenging because lags in responses to policies or shocks are long and variable and the economic system is responsive, events prompting changes in the structure of the economy. Forecasting tends to be relatively successful during normal times; however, in times of crises and change, in the face of large shocks or structural changes, when accurate predictions are most needed, forecasters tend to fail. For instance, inflation was 5.4% in December 2021. This was the last figure they had when, in February 2022, the Bank forecast that inflation would peak at 7.1% in 2022q2, and fall to 5.5% in 2023q1, and be back within target at 2.6% in 2024q1. The 2023q1 actual was 10.2%, almost 5 percentage points higher than the forecast. This burst in inflation was a global phenomenon and other central banks made similar errors.

Economic forecasters may use purely statistical models or more structural economic models that include the policy variables and important economic linkages. We will call these more structural models 'policy models', given the way structural has a number of interpretations. The statistical models will typically be conditional on information available at the time of the forecast, which may be inaccurate: knowing where one is at the time of forecast, nowcasting, is an important element. Policy models may also be conditional on the assumed future values. The Bank of England makes forecasts conditional on market expectations of future interest rates, assumptions about future energy prices and government announcements about future fiscal policy as well as other measures. Wrong assumptions about those future values may cause problems, as it did for the Bank in August 2022, when it anticipated a rise in energy prices but not the government response to them, so it overestimated inflation. There is also a policy issue as to whether fiscal and monetary policy should be determined independently by different institutions or jointly.

If both statistical and policy models are used, there is an issue as to how to integrate them. Forecast averaging has been widely shown to improve forecast performance, but forming averages on many variables may lack coherence and consistency. In September 2018, the Bank of England Independent Evaluation Office, IEO reported back to the Court of the Bank on the implementation and impact of the 2015 IEO review of forecasting performance. The IEO reported that some 'non-structural' models had been introduced as a source of challenge, and outputs were routinely shown to the Monetary Policy Committee (MPC) as a way of cross checking the main forecast. While some but not all members found them helpful, there was no desire to develop more models of this sort, and internally, they had not been integrated into the forecast process as a source of challenge.

Forecasts by central banks fulfil multiple purposes, including as a means of communication to influence expectations in the wider economy. This also makes it difficult to choose a loss function to evaluate forecasts. For instance, the Bank of England forecasts for inflation at a 2-year horizon are always close to the target of 2 per cent. Even were the Bank to think it unlikely that it could get back to target within 2 years, it might feel that its credibility might be damaged were it to admit that. An institutional issue is who 'owns' the forecast. The Bank of England forecast is the responsibility of the nine-member MPC; other central banks have different systems. For instance, the US Federal Reserve has a staff forecast not necessarily endorsed by the decision-makers.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

There is an issue about the optimal amount of information to use: both with respect to breadth, how many variables, and length, how long a run of data. With respect to breadth, in principle, one should use information on as many variables as possible and not just for the country being forecast, since in a networked world, foreign variables contain information. This is the information that is used in the Global vector autoregression (GVAR), whose use is surveyed in Chudik and Pesaran (2016). While the use of many variables might imply a large model, in practice quite parsimonious small models tend to be difficult to beat in forecasting competitions.

With respect to length, in a May 2023 hearing, the Chair of the House of Commons Treasury Select Committee asked the chief economist of the Bank of England, Huw Pill: 'Are you saying that, despite the Bank of England having been in existence for over 300 years, you look at only the last 30 years when you think about what the risks are to inflation?'. Pill emphasised the importance of the policy regime, which had been different in the past 30 years of inflation targeting than in earlier high inflation periods. The 30 years up to 2019 had also been different in terms of the absence of large real shocks, like Covid-19 and the effects of the Russian invasion of Ukraine.

Whether it is statistical or policy, the model will typically be supplemented by a judgemental input, justified by the argument that the forecaster has a larger information set than the model. In evidence to the Treasury Select Committee in September 2023, Sir Jon Cunliffe said: 'We start with the model. All models are caricatures of real life. There is a suite of models; that is the starting point. However, the MPC itself puts judgements that change the model, and we have made some quite big judgements in the past about inflation persistence and the like. Finally, when we have the best collective view of the committee, which is our judgement on top of the model, the model keeps us honest. It ensures that there is a general equilibrium and we cannot just move things around.'

In short, macroeconomic forecasting faces important challenges. It depends on how forecasts are announced and used in the decision-making process. To deal with a constantly changing economic environment, forecasts must continually adapt to new data sets, statistical techniques and theory-based economic insights, knowing that there are still key variables that might have been left out, either due to difficulties in measurement, oversight or ignorance. Forecasters must answer a range of difficult questions. What sample periods and potential variables to consider? How to decide which variables to use for forecasting, and whether to use the same sample periods for variable selection and for forecasting? Should one use ensemble forecasting from forecasts obtained either from different models or from the same model estimated over different sample sizes or with different degrees of down-weighting? One must only be humbled by the sheer extent of the uncertainty that these choices entail. It is within this wider context that this article tries to formalise some elements of the problem of forecasting with high-dimensional data and illustrates the various issues involved with an application to forecasting UK inflation.

The rest of this article is organised as follows. Section 2 sets out the high-dimensional forecasting framework we will be considering. Section 3 considers 'known knowns', selecting relevant variables from a known active set. Section 4 considers 'known unknowns' where there are known to be unobserved latent variables. Section 5 presents the empirical application on forecasting UK inflation. Section 6 contains some concluding comments.

2. The high-dimensional forecasting problem

Suppose the aim is to forecast a scalar target variable, denoted by y_{T+h}, at time T, for the future dates, $T+h$, $h=1,2,\ldots,H$. Given the historical observations, the optimal forecast of the target variable, y_{T+h}, depends on how the forecasts are used, namely the underlying decision problem. In practice, specifying loss functions associated with decision problems is difficult; hence, the tendency is to fall back on mean squared error loss. Under this loss function, the optimal forecasts are given by conditional expectations, $E\left(y_{T+h}|\mathcal{I}_T\right)$, where $\mathcal{I}_T$ is the set of available information, and expectations are formed with respect to the joint probability distribution of the target variable and the set of potential predictors under

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

consideration. But when the number of potential predictors, say K, is large, even this result is too general to be of much use in practice.

The high-dimensional nature of the forecasting problem also presents a challenge of its own when we come to multi-step ahead forecasting when forecasts of the target variable are required for different horizons, $h = 1, 2, \ldots, H$. Many decision problems require having forecasts many periods ahead, months, years and even decades ahead. Monetary policy is often conducted over the business cycle, at least 2–3 years ahead of the policy formulation. Climate change policy requires forecasts over many decades ahead. In interpreting Pharaoh's dreams, Joseph considered a two-period decision problem whereby 7 years of plenty are predicted to be followed by 7 years of drought. Multi-horizon forecasting is relatively straightforward when the number of potential predictors is small and a complete system of equations, such as a VAR, can be used to generate forecasts for different horizons from the same forecasting model in an iterative manner. Such an **iterated** approach is not feasible, and might not even be desirable, when the number of potential predictors is too large, since future forecasts of predictors are also needed to generate forecasts of y_{T+h} for $h \geq 2$. This is why in high-dimensional set-ups multi-period ahead forecasts are typically formed using different models for different horizons. This is known as the **direct** approach and avoids the need for forward iteration by directly regressing the target variable y_{t+h} on the predictors at time t, thus possibly ending up with different models and/or estimates for each h.[3]

To be more specific, ignoring intercepts and factors which we introduce below, suppose y_t is the first element of the high-dimensional vector $\boldsymbol{w}_t$, assumed to follow the first-order VAR model,

$$\boldsymbol{w}_t = \boldsymbol{\Phi}\boldsymbol{w}_{t-1} + \mathbf{u}_t. \tag{1}$$

Higher order VARs can be written as first-order VARs using the companion form. The error vector, $\mathbf{u}_t$, satisfies the orthogonality condition $E(\mathbf{u}_t|\mathcal{I}_{t-1}) = \mathbf{0}$, where $\mathcal{I}_{t-1} = (\boldsymbol{w}_{t-1}, \boldsymbol{w}_{t-2}, \ldots)$. Then

$$\boldsymbol{w}_{T+h} = \boldsymbol{\Phi}^h \mathbf{w}_T + \mathbf{u}_{h,T+h}, \tag{2}$$

where except for $h = 1$, the overlapping observations cause the error in (2) to have the moving average structure of order $h - 1$:

$$\mathbf{u}_{h,t+h} = \mathbf{u}_{t+h} + \boldsymbol{\Phi}\mathbf{u}_{t+h-1} + \boldsymbol{\Phi}^2\mathbf{u}_{t+h-2} + \ldots + \boldsymbol{\Phi}^{h-1}\mathbf{u}_{t+1}.$$

Under the VAR specification $E(\mathbf{u}_{h,T+h}|\mathcal{I}_T) = 0$, for $h = 1, 2, \ldots$ and the optimal (in the mean squared error sense) h-step ahead forecast of $\boldsymbol{w}_{T+h}$ is $E(\boldsymbol{w}_{T+h}|\mathcal{I}_T) = \boldsymbol{\Phi}^h\boldsymbol{w}_T$. But given that in most forecasting applications the dimension of $\boldsymbol{w}_t$ is large, it is not feasible to estimate $\boldsymbol{\Phi}$ directly without imposing strong sparsity restrictions. Instead, we take the target variable, y_{T+h}, to be the first element of $\boldsymbol{w}_{T+h}$ and consider the direct regression

$$y_{t+h} = \boldsymbol{\phi}_h'\boldsymbol{w}_t + u_{h,t+h},$$

where $\boldsymbol{\phi}_h'$ is the first row of $\boldsymbol{\Phi}^h$, and $u_{h,t+h}$ is the first element of $\mathbf{u}_{h,t+h}$. We still face a high-dimensional problem since there are a large number of potential covariates in $\boldsymbol{w}_t$. We consider the implementation of the direct approach under two scenarios concerning the potential predictors. First, when it is known that the target variable y_{t+h} is a sparse linear function of a large set of observed variables $\boldsymbol{x}_t$ (a subset of $\boldsymbol{w}_t$) known as the 'active set'. The model is sparse in the sense that y_{t+h} depends on a small number of

[3]Marcellino *et al.* (2006) discuss the pros and cons of iterated and direct approaches to forecasting when K is small, and the target variable and the predictors can be jointly modelled as low-dimensional VARs or VARMAs. It is shown that if the underlying VAR model is correctly specified, then iterated forecasts, being coherent, are preferred to direct forecasts. However, under misspecification, direct forecasts could perform better. Pesaran *et al.* (2011) reconsider the comparison of iterated and direct forecasts to factor augmented VARs.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

covariates, that are *known* to be a subset of the much larger active set. The machine learning literature focuses on this case, which we refer to as the case of 'known knowns'. Second, when y_{t+h} could also depend on a few latent (unobserved) factors, $\mathbf{f}_t$, not directly included in the active set, which we call the case of 'known unknowns'.

Specifically, we suppose that for each h, y_{t+h} can be approximated by the following linear model, where the predictors are also elements of $\boldsymbol{w}_t$ in the high-dimensional VAR (1),

$$y_{t+h} = c_h + \mathbf{a}_h'\mathbf{z}_t + \sum_{j=1}^{K}\beta_{jh}I(j \in DGP)x_{jt} + \boldsymbol{\psi}_h'\mathbf{f}_t + u_{h,t+h}, \tag{3}$$

for $t = 1,2,\ldots,T-h$, where c_h is the intercept, $\mathbf{z}_t$ is a vector of small number, p, of preselected covariates included across all horizons h. Obvious examples include lagged values of the target variable $(y_t, y_{t-1}, \ldots)$. Other variables can also be included in $\mathbf{z}_t$ on the basis of *a priori* theory or strong beliefs. The third component of y_{t+h} specifies the subset of variables in the active set $\boldsymbol{x}_{Kt} = (x_{1t}, x_{2t}, \ldots, x_{Kt})'$. $I(j \in DGP)$ is an indicator variable that takes the value of unity if x_{jt} is included in the data generating process (DGP) for y_{t+h} and zero otherwise. It is only if $I(j \in DGP) = 1$ that β_{jh} will be identified. We discuss ways to determine the selection indicator $I(j \in DGP)$ below. The number of variables included in the DGP is given by $k = \sum_{j=1}^{K} I(j \in DGP)$, which is supposed to be small and fixed as T (and possibly K) becomes large. This assumption imposes sparsity on the relationship between the target and the variables in the active set. In addition, we allow for a small number of latent factors, $\mathbf{f}_t$, that represent other variables influencing y_{t+h} that are not observed directly, but known to be present—the known unknowns.

Giannone *et al.* (2021) contrast **sparse methods**, that select a few variables from the active set as predictors, such as Lasso and **O**ne **C**ovariate at a time **M**ultiple **T**esting (OCMT) discussed below, and **dense methods**, that select all the variables in the active set but attach small weights to many of them, such as principal components (PCs), ridge regression and other shrinkage techniques. Rather than having to choose between sparse and dense predictors, we consider approaches that combine the two. We apply sparse selection methods to the variables in the active set, and use dense shrinkage methods to approximate $\mathbf{f}_t$ from a wider set of variables with $\boldsymbol{x}_t$ included as a subset. We first consider the selection problem, known knowns, where we know the active set of potential covariates, and we then consider known unknowns where there are unobserved factors. The elastic net regression of Zou and Hastie (2005) discussed below also combines sparse and dense techniques.

Throughout, we shall assume that the errors, $u_{h,t+h}$, in (3) satisfy the orthogonality condition $E(u_{h,t+h} \mid \mathbf{z}_t, \boldsymbol{x}_t, \mathbf{f}_t,) = 0$, for $h \geq 1$. In the context of the high-dimensional VAR model discussed above, this orthogonality condition holds so long as the underlying errors, $\mathbf{u}_t$, are serially uncorrelated. This is so despite the fact that due to the use of overlapping observations $u_{h,t+h}$ will be serially correlated when $h > 1$. This is an important consideration when high-dimensional techniques are applied to select predictors for multi-step ahead forecasting; an issue to which we will return.

3. Known knowns

In the case of known knowns, forecasts are obtained assuming that y_{t+h} is a linear function of $\boldsymbol{x}_t$

$$y_{t+h} = c_h + \mathbf{a}_h'\boldsymbol{z}_t + \sum_{j=1}^{K}\beta_{hj}x_{jt} + u_{h,t+h}, \text{for } t = 1,2,\ldots,T, \tag{4}$$

subject to some penalty condition on $\left\{\beta_{hj}\right\}$. Some of the covariates, x_{jt}, could be transformations of other covariates, such as interaction terms. It is assumed that the model is correctly specified, in the sense that, apart from $\boldsymbol{z}_t$, the variables that drive y_{t+h} are all included in the active set, $\mathbf{x}_{Kt}$.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

Penalised regressions estimate β by solving the following optimisation problem:

$$\min_{\mathbf{a}_h,\beta_h}\left\{\sum_{t=1}^{T}\left(y_{t+h}-\mathbf{a}_h'\mathbf{z}_t-\beta_h'\mathbf{x}_{Kt}\right)^2+\lambda_{hT}\sum_{i=1}^{K}\left[(1-\alpha)|\beta_{hi}|+\alpha\beta_{hi}^2\right]\right\},$$

where $\beta_h=(\beta_{h1},\beta_{h2},\ldots,\beta_{hK})'$, for given values of the 'tuning' parameters λ and α. When $\alpha=1$, we have ridge regression. When $\alpha=0$ and $\lambda_{hT}\neq 0$, we have the **Lasso** regression, which is better suited for variable selection. When $\lambda_{hT}\neq 0$ and $\alpha\neq 0$, we have the Zou and Hastie (2005) **elastic net** regression, which also mixes sparse and dense approaches.

Many standard forecasting techniques result from a particular choice of the penalty function. Shrinkage estimators such as ridge or some Bayesian forecasts can be derived using the ℓ_2 norm $\sum_{i=1}^{K}\beta_{hi}^2<C_h<\infty$. Lasso (least absolute shrinkage and selection operator) follows when the ℓ_1 norm is used $\sum_{i=1}^{K}|\beta_{hi}|<C_h<\infty$. The difference is shown in Figure 1 below in Tibshirani (1996) where the ℓ_1 norm yields corner solutions with many of the coefficients, β_{hj}, estimated to be zero. In contrast, the use of ℓ_2 norm yields non-zero estimates for all the coefficients with many very close to zero.

There are also a large number of variants of Lasso, including adaptive Lasso, group Lasso, double Lasso, fused Lasso and prior Lasso. We will focus on Lasso itself, which we use in our empirical application and which we will compare to OCMT as an alternative procedure which is based on inferential rather than penalised procedures.

3.1. Lasso

In this article, we focus on Lasso, but acknowledge that are many variations on Lasso such as adaptive Lasso, group Lasso, fused Lasso and prior Lasso. Lasso estimates β_h by solving the following optimisation problem:

$$\min_{\beta_h}\left\{\sum_{t=1}^{T}\left(y_{t+h}-c_h-\beta_h'\boldsymbol{x}_t\right)^2+\lambda_{hT}\sum_{i=1}^{K}|\beta_{hi}|\right\}, \tag{5}$$

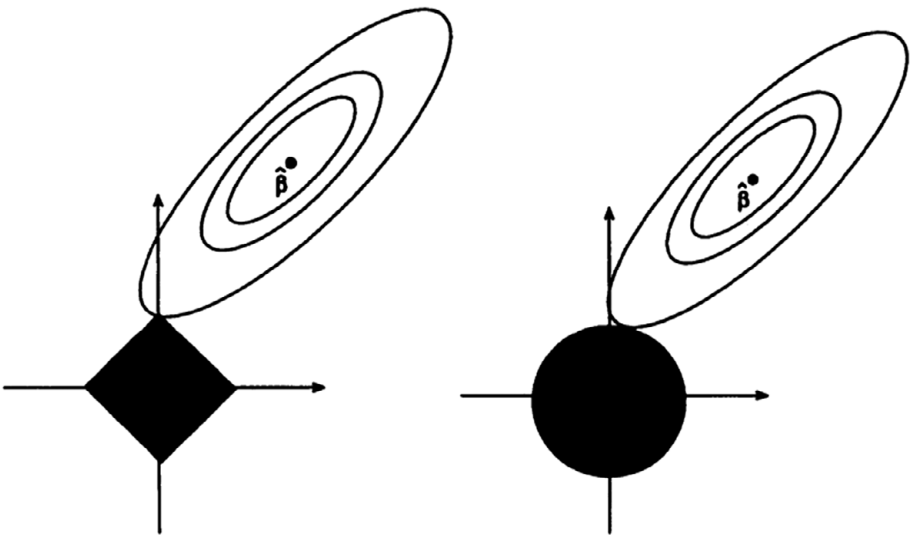

Figure 1. Estimation for the Lasso (left) and ridge (right) regression.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

where $\beta_h = (\beta_{h1}, \beta_{h2}, \ldots, \beta_{hK})'$, $\boldsymbol{x}_{Kt} = (x_{1t}, x_{2t}, \ldots, x_{Kt})'$ for a given choice of the 'tuning' parameter, λ_{hT}. The variable selection consistency of Lasso has been investigated by Meinshausen and Bühlmann (2006), Zhao and Yu (2006) and more recently by Lahiri (2021). The key condition is the so-called 'Irrepresentable Condition (IRC)' that places restrictions on the magnitudes of the correlations between the signals ($\mathbf{X}_{1h}$, standardised) and the rest of the covariates ($\mathbf{X}_{2h}$, standardised), taken as given (deterministic). The IRC is:

$$\text{IRC}: \left\| \left(T^{-1}\mathbf{X}_{2h}'\mathbf{X}_{1h}\right)\left(T^{-1}\mathbf{X}_{1h}'\mathbf{X}_{1h}\right)^{-1} sign\left(\beta_h^0\right) \right\|_{\infty} < 1, \tag{6}$$

where $\beta_h^0 = \left(\beta_{1h}^0, \beta_{2h}^0, \ldots, \beta_{k_h h}^0\right)'$ denotes the vector of true signal coefficients.[4] The IRC condition is met for pure noise variables, but need not hold for proxy variables, noise variables that are correlated with the true signals.

To appreciate the significance of the IRC, suppose the DGP contains x_{1t} and x_{2t} and the rest of the covariates in the active set are $x_{3t}, x_{4t}, \ldots, x_{Kt}$. Denote the sample correlation coefficient between x_{1t} and x_{2t} by $\widehat{\rho}$ $\left(\widehat{\rho}^2 < 1\right)$ and the sample correlation coefficient of x_{1t} and x_{2t} with the rest of the covariates in the active set by $\widehat{\rho}_{1s}, \widehat{\rho}_{2s}$, for $s = 3, 4, \ldots, K$. Then, dropping the subscript h, the IRC for the s^{th} covariate is given by

$$\left| [\text{sign}(\beta_{01}),\ \text{sign}(\beta_{02})]' \begin{pmatrix} 1 & \widehat{\rho} \\ \widehat{\rho} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\rho}_{1s} \\ \widehat{\rho}_{2s} \end{pmatrix} \right| < 1$$

which yields

$$|sign(\beta_{01})(\widehat{\rho}_{1s} - \widehat{\rho}\widehat{\rho}_{2s}) + sign(\beta_{02})(\widehat{\rho}_{2s} - \widehat{\rho}\widehat{\rho}_{1s})| < 1 - \widehat{\rho}^2$$

for $s = 3, 4, \ldots, K$. In this example, there are two cases to consider: A: $sign(\beta_{02}) = sign(\beta_{01})$; and B: $sign(\beta_{02}) = -sign(\beta_{01})$. For case A $\sup_s |\widehat{\rho}_{1s} + \widehat{\rho}_{2s}| < 1 + \widehat{\rho}$, and for case B $\sup_s |\widehat{\rho}_{1s} - \widehat{\rho}_{2s}| < 1 - \widehat{\rho}$. Since the signs of the coefficients are unknown, for all possible values of $\widehat{\rho}, \widehat{\rho}_{1s}$ and $\widehat{\rho}_{2s}$, we can ensure the IRC condition is met if $|\widehat{\rho}| + \sup_s|\widehat{\rho}_{1s}| + \sup_s|\widehat{\rho}_{2s}| < 1$. This example shows the importance of the correlations between the true covariates in the DGP as well as between the true covariates and the other members of the active set that do not belong to the DGP. The IRC is quite a stringent condition and it is not just when one has proxies in the active set that are highly correlated with the true covariates that Lasso will tend to choose too many variables. In practice, one cannot check the IRC condition since one does not know which variables are the true signals.

In addition to the IRC, it is also required that

$$\text{MinC}: \min_{j=1,2,\ldots,k} \left|\beta_{jh}^0\right| > (2T)^{-1}\lambda_{hT} \left| \left(T^{-1}\mathbf{X}_{1h}'\mathbf{X}_{1h}\right)^{-1} sign\left(\beta_h^0\right) \right|_j,$$

$$\text{Penalty Condition}: T^{-1}\lambda_{hT} = o(1).$$

The penalty condition, which follows from MinC, says that the penalty has to rise with T, but not too fast and not too slowly. The expansion rate of λ_{hT} depends on the magnitude and the sign of β_{jh}^0, and the correlations of signals with the proxy variables. Lahiri (2021) shows that the penalty condition can be relaxed to $\lim_{T\to\infty} T^{-1}\lambda_{hT} < \liminf_{T\to\infty} d_{hT}$, where

[4]The number of signals, k_h, could vary with h.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

$$d_{hT} = 2\min_j \left|\beta_{jh}^0\right| \Big/ \left|\left(T^{-1}\mathbf{X}_{1h}'\mathbf{X}_{1h}\right)^{-1} sign\left(\beta_h^0\right)\right|_j.$$

The above conditions do not restrict the choice of λ_{hT} very much, hence the recourse to cross-validation (CV) to determine it. In practice, λ_{hT} is calibrated using M-fold CV techniques. The observations, $t = 1, 2, \ldots, T$, are partitioned into M disjoint subsets (folds), of size approximately $m = T/M$. Then, $M-1$ subsets are used for training and one for evaluation. This is repeated with each fold being used in turn for evaluation. M is typically set to 5 or 10. CV methods are often justified in machine learning literature under strong assumptions, such as independence and parameter stability across the sub-samples used in CV. These assumptions are rarely met in the case of economic time series data, an issue that is discussed further in the context of the empirical example in Section 5.

3.2. One Covariate at a time Multiple Testing

The need for CV is avoided in the procedure proposed by Chudik *et al.* (2018), (CKP). This is the **OCMT** procedure, where covariates are selected **one at a time**, using the t-statistic for testing the significance of the variables in the active set, **individually**.[5] Ideas from the multiple testing literature are used to control the false discovery rate, and ensure that the selected covariates encompass the true covariates (signals) with probability tending to unity, under certain regularity conditions. Like Lasso, OCMT has no difficulty in dealing with (pure) noise variables, and is very effective at eliminating them. Also, like Lasso, it requires some *min* condition such as $\left|\beta_{jh}^0\right| >> \sqrt{\frac{k\log(K)}{T}}$, for $j = 1, 2, \ldots, k$.[6] But because it considers a single variable at a time, OCMT does not require the IRC condition to hold and is not affected by the correlation between the members of the DGP as Lasso is. Instead, it requires the number of proxies signals, say k_T^*, to rise no faster than $\sqrt{T}$. Chudik *et al.* (2023), discussed below, is primarily concerned with parameter instability; however, Section 4 of that paper has a detailed comparison of the assumptions required for Lasso and OCMT under parameter stability.

OCMT's condition on k_T^* has been recently relaxed by Sharifvaghefi (2023) who allows $k_T^* \to \infty$, with T. He considers the following DGP:

$$y_{t+h} = c_h + \mathbf{a}_h'\mathbf{z}_t + \sum_{j=1}^{K} \beta_{jh} I(j \in DGP) x_{jt} + u_{h,t+h}, \tag{7}$$

where as before $\mathbf{z}_t$ is a known vector of preselected variables, and it is assumed that the k signals are contained in the known active set $\mathcal{S}_{K,t} = \{x_{jt}, j = 1, 2, \ldots, K\}$. Note that for now the DGP in (7) does not include the additional latent factors, $\mathbf{f}_t$, introduced in (3). Without loss of generality, consider the extreme case where there are no noise variables and **all** proxy or pseudo signal variables $(x_{jt},$ for $j = k+1, k+2, \ldots, K)$ are correlated with the signals, $\mathbf{x}_{1t} = (x_{1t}, x_{2t}, \ldots, x_{kt})'$. In this case, k_T^* rises with K and OCMT is no longer applicable. However, in this case because of the correlation with the proxies, the signals, $\mathbf{x}_{1t}$, become latent factors for the proxy variables and we have

$$x_{jt} = \phi_{j0} + \sum_{i=1}^{k} \phi_{ji} x_{it} + \varepsilon_{jt} = \phi_{j0} + \boldsymbol{\phi}_j'\mathbf{x}_{1t} + \varepsilon_{jt},$$

[5]Since t-ratios are invariant to scale, no pre-standardization of the covariates in the active set is required. This is in contrast to Lasso, which is typically implemented after in-sample standardization of the covariates.

[6]To simplify the exposition, we have dropped explicit reference to the forecast horizon, h. However, in practice, and as we shall see from the empirical applications below, the number and identity of selected signals could differ with h.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

for $j=k+1,k+2,\ldots,K$. Although the identity of these common factors is unknown, because we do not know the true signals, they can be approximated by the PCs of the variables in the active set.

Specifically, following Sharifvaghefi (2023), denote the latent factors that result in non-zero correlations between the noise variables in the active set and the signals by $\boldsymbol{\varkappa}_t$ and consider the factor model

$$x_{jt}=\boldsymbol{\kappa}_j'\boldsymbol{\varkappa}_t+v_{jt},\ \text{for } j=1,2,\ldots,K, \tag{8}$$

where $\boldsymbol{\kappa}_j$, for $j=1,2,\ldots,K$ are the factor loadings and the errors, v_{jt}, are weakly cross-correlated and distributed independently of the factors and their loadings. Under (8), the DGP, (7), can be written equivalently as

$$y_{t+h}=c_h+\mathbf{a}_h'\mathbf{z}_t+\mathbf{b}_h'\boldsymbol{\varkappa}_t+\sum_{j=1}^{K}\beta_{jh}I(j\in DGP)v_{jt}+u_{h,t+h}, \tag{9}$$

where $\mathbf{b}_h=\sum_{j=1}^{K}I(j\in DGP)\beta_{jh}\kappa_j$. When $\boldsymbol{\varkappa}_t$ and v_{jt} are known, the problem reduces to selecting v_{jt} from $\mathcal{S}_{K,t}^{v}=\{v_{jt},j=1,2,\ldots,K\}$, conditional on $\mathbf{z}_t$ and $\boldsymbol{\varkappa}_t$. Sharifvaghefi shows that the OCMT selection can be carried out using the PC estimators of $\boldsymbol{\varkappa}_t$ and $\mathbf{v}_{jt}$—denoted by $\widehat{\boldsymbol{\varkappa}}_t$ and $\widehat{\mathbf{v}}_{jt}$, if both K and T are large. He labels this procedure as generalised OCMT (GOCMT). Note that at the moment, it is assumed that $\boldsymbol{\varkappa}_t$ does not **directly** affect y_{t+h}, it only enters through the x_{jt}. It represents the signals, the common factors correlated with the proxies, and provides a way of filtering the correlations in the first step.

3.3. Generalised OCMT

The GOCMT procedure simply augments the OCMT regressions with the PCs, $\widehat{\boldsymbol{\varkappa}}_t$, and considers the statistical significance of $\widehat{\mathbf{v}}_{jt}$ for each j, *one at the time*. Lasso-factor models have also been considered by Hansen and Liao (2019) and Fan *et al.* (2020). In practice, since $\mathbf{x}_j=\widehat{\boldsymbol{\Xi}}\widehat{\boldsymbol{\psi}}_j+\widehat{\mathbf{v}}_j$, where $\widehat{\boldsymbol{\Xi}}=(\widehat{\boldsymbol{\varkappa}}_1,\widehat{\boldsymbol{\varkappa}}_2,\ldots,\widehat{\boldsymbol{\varkappa}}_T)'$, then $\mathbf{M}_{\widehat{\boldsymbol{\Xi}}}\mathbf{x}_j=\mathbf{M}_{\widehat{\boldsymbol{\Xi}}}\widehat{\mathbf{v}}_j$, where $\mathbf{M}_{\widehat{\boldsymbol{\Xi}}}=\mathbf{I}_T-\widehat{\boldsymbol{\Xi}}\left(\widehat{\boldsymbol{\Xi}}'\widehat{\boldsymbol{\Xi}}\right)^{-1}\widehat{\boldsymbol{\Xi}}'$, and GOCMT reduces to OCMT when $\mathbf{z}_t$ is augmented with $\widehat{\boldsymbol{\varkappa}}_t$, where the statistical significance of x_{jt} as a predictor of y_{t+h} is evaluated for each j, one at a time. Like OCMT, GOCMT allows for the multiple testing nature of the procedure (K separate tests—with K large) by increasing the level of significance with K. The number of PCs, $dim(\widehat{\boldsymbol{\varkappa}}_t)$, can be determined using one of the criteria suggested in the factor literature.

In the first stage, K **separate** OLS regressions are computed, where the variables in the active set are entered one at a time:

$$y_{t+h}=c_h+\mathbf{a}_h'\mathbf{z}_t+\mathbf{b}_h'\widehat{\boldsymbol{\varkappa}}_t+\phi_{jh}x_{jt}+e_{j,h,t+h},t=1,2,\ldots,T,\ \ \text{for } j=1,2,\ldots,K. \tag{10}$$

Denote the t-ratio of ϕ_{jh} by $t_{\hat{\phi}_{j,(1)}}$. Then, variable j is selected if

$$\widehat{\mathcal{J}}_{j,(1)}=I\left[\left|t_{\hat{\phi}_{j,(1)}}\right|>c_p(K,\delta)\right],\text{for } j=1,2,\ldots,K, \tag{11}$$

where $c_p(K,\delta)$ is a critical value function given by

$$c_p(K,\delta)=\Phi^{-1}\left(1-\frac{p}{2K^{\delta}}\right). \tag{12}$$

p is the nominal size (usually set to 5%), $\Phi^{-1}(\cdot)$ is the inverse of a standard normal distribution function and δ is a fixed constant set in the interval [1,1.5]. In the second step a multivariate regression of y_{t+h} on $\mathbf{z}_t$

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

and all the selected regressors is considered for inference and forecasting. Serial correlation will arise with OCMT when selection is based on one variable at the time, and the omitted variables are mixing (serially correlated). CKP discuss this in section C of the online theory supplement to their paper and suggest using a more conservative (higher) critical value—namely using $\delta = 1.5$ rather than $\delta = 1.0$.

When the covariates are **not** highly correlated, OCMT applies irrespective of whether K is small or large relative to T, so long as $T = \ominus(K^c)$, for some finite $c > 0$. But to allow for highly correlated covariates, GOCMT requires K to be sufficiently large to enable the identification of the latent factor, $\boldsymbol{\varkappa}_t$. In cases where K is not that large, it might be a good idea to augment the active set for the target variable, y_{t+h}, $\mathcal{S}_{K,t} = \{x_{jt}, j = 1, 2, \ldots, K\}$, with covariates for other variables determined simultaneously with y_{t+h}, ending up with $\overline{K} > K$ covariates for identification of $\boldsymbol{\varkappa}_t$. GOCMT does not impose any restriction on the correlations between the variables other than that they cannot be perfectly collinear.

3.4. High-dimensional variable selection in presence of parameter instability

OCMT has also been recently generalised by Chudik *et al.* (2023) to deal with parameter instability. Under parameter instability, OCMT correctly selects the covariates with non-zero average (over time) effects, using the full sample. However, the adverse effects of changing parameters on the forecast may mean that while the full sample is the best to use for selection, it need not be the best to use for estimating the forecasting model. Instead, it may be better to use shorter windows or weight the observations in the light of the evidence on break points and break sizes.

Determining the appropriate window or weighting for the observations before estimation is a difficult problem and no fully satisfactory procedure seems to be available. It is common in finance to use rolling windows of 60 or 120 months, but one problem with shorter windows is that if you have periods of instability interspersed with periods of stability, like the Great Moderation, estimates using a short window from the stable period may understate the degree of uncertainty. This happened during the financial crisis when the short windows used for estimation did not reflect past turbulence. Similarly, the Bank of England estimating their models using the low inflation regime of the past 30 years discounted the evidence from the high inflation regime of the 1970s and 1980s.

While identifying the date of a break might not be difficult, identifying the size of the break may be problematic if the break point is quite recent. If there is a short time since the break, there is little data on which to estimate the post-break coefficient with any degree of precision. If there is a long time since the break, then using post break data is sensible. Pesaran *et al.* (2013) examine optimal forecasts in the presence of continuous and discrete structural breaks. These present quite different sorts of challenges. With continuous breaks, the parameters change often by small amounts. With discrete breaks, the parameters change rarely but by large amounts. They propose weighting observations to obtain optimal forecasts in the MSFE sense and derive optimal weights for one-step ahead forecasts for the two types of break. Under continuous breaks, their approach largely recovers exponential smoothing weights. Under discrete breaks between two regimes, the optimal weights follow a step function that allocates constant weights within regimes but different weights in different regimes. In practice, the time and size of the break are uncertain, and they investigate robust optimal weights. Averaging forecasts with different weighting schemes, for instance, with exponential smoothing parameters between 0.96 and 0.99, may also be a way to produce more robust forecasts.

4. Known unknowns

So far, we have considered techniques (penalised regressions and OCMT) that assume y_{t+h} depends on $\mathbf{z}_t$ and a subset of a set of covariates—the active set—which is assumed **known**. In contrast, shrinkage type techniques such as PCs, (implicitly) assume that y_{t+h} depends on $\mathbf{z}_t$ and the $m \times 1$ vector of **unknown** factors $\mathbf{f}_t$

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

$$y_{t+h} = c_h + \mathbf{a}_h' \mathbf{z}_t + \boldsymbol{\theta}_h' \mathbf{f}_t + u_{h,t+h}.$$

This is a simple example of techniques that in our terminology can be viewed as belonging to a class of forecasting models based on **known unknowns**. The uncertainty about $\mathbf{f}_t$ is resolved assuming it can be identified from a **known** active set, such as $\mathcal{S}_{K,t} = \{x_{jt}, j = 1,2,\ldots,K\}$. Individual covariates in $\mathcal{S}_{K,t}$ are not considered for selection (although a few could be preselected and included in $\boldsymbol{z}_t$). To forecast y_{t+h} one still requires to forecast the PCs and to allow for the uncertainty regarding $m = \dim(\mathbf{f}_t)$.

Factor augmented VARs (FAVAR), initially proposed by Bernanke *et al.* (2005, BBE), augment the standard VAR models with a set of unobserved common factors. In the context of our set up, FAVAR can be viewed as a generalised version of (3), where $\mathbf{y}_{t+h}$ is a vector and $\boldsymbol{z}_t = \{\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-p}\}$. BBE argue that small VARs gave implausible impulse response functions, such as the 'price puzzle', which were interpreted as reflecting omitted variables. One response was to add variables and use larger VARs, but this route rapidly runs out of degrees of freedom, since Central Bankers monitor hundreds of variables. The FAVAR was presented as a solution to this problem. Big Bayesian VARs are an alternative solution.

The assumptions that underlie both penalised regression and PC shrinkage are rather strong. The former assumes that $\mathbf{f}_t$ can affect y_{t+h} only indirectly through x_{jt}, $j = 1,2,\ldots,K$, and the latter does not allow for individual variable selection. Suppose that $\mathbf{f}_t$ also enters (7) then the model can be written as (3) above, repeated here for convenience:

$$y_{t+h} = c_h + \mathbf{a}_h' \mathbf{z}_t + \sum_{j=1}^{K} \beta_{jh} I(j \in DGP) x_{jt} + \boldsymbol{\psi}_h' \mathbf{f}_t + u_{h,t+h}.$$

The forecasting problem now involves both selection and shrinkage. The $\mathbf{f}_t$ can be identified by, for instance, the PCs of the augmented active set x_{jt}, $j = 1,2\ldots,K,K+1,\ldots,\overline{K}$ which can be wider than the active set of covariates used to predict y_t. There are various other ways that the unobserved $\mathbf{f}_t$ could be estimated, but we use PCs as an example, since they are widely used.

The latent factors are unlikely to be only specific to the target variable under consideration. Observed global factors, such as oil and raw material prices or inflation and output growth of major countries such as United States can be included in the active set. The main issue is how to deal with global factors, such as technology, political change and so on that are unobserved and tend to affect many countries in the world economy. Call this vector of global factors $\mathbf{g}_t$. A natural extension is to introduce forecast equations for other countries (entities) who have close trading relationships with United Kingdom and use penalised panel regressions, where the panel dimension allows identification of the known unknowns.

More specifically, suppose there are N other units (countries) that are affected by observed country-specific covariates, $\mathbf{z}_{it}$, $i = 1,2,\ldots,N$, and $x_{i,jt}$ for $j = 1,2,\ldots,K_i$, plus domestic latent factors $\mathbf{f}_{it}$, and global latent factors, $\mathbf{g}_t$. The forecasting equations are now generalised as

$$y_{i,t+h} = c_h + \mathbf{a}_{ih}' \mathbf{z}_{it} + \sum_{j=1}^{K_i} \beta_{ijh} I(j \in DGP_i) x_{i,jt} + \theta_{ih}' \mathbf{f}_{it} + \psi_{ih}' \mathbf{g}_t + u_{i,h,t+h},$$

where $k_i = \sum_{j=1}^{K_i} I(j \in DGP_i)$ is finite as $K_i \to \infty$, for $i = 0,1,2,\ldots,N$. For the country-specific covariates, we postulate that there is an augmented active set

$$x_{i,jt} = \gamma_{ij}' \mathbf{f}_{it} + v_{i,jt}, j = 1,2,\ldots,K_i, K_{i+1}, \ldots, \overline{K}_i,$$

where $\mathbf{f}_{it}$ are the latent factors. The global factors are then identified as the common components of the country-specific factors, namely

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

$$\mathbf{f}_{it} = \mathbf{\Psi}_i \mathbf{g}_t + \boldsymbol{\xi}_{it},$$

for $i = 1,2,..,N$, with N large.

Variable selection for the target variable (say UK inflation) can now proceed by applying GOCMT, with the UK model augmented with UK-specific PCs, $\widehat{\mathbf{f}}_{it}$ as well as the PC estimator of the global factor, $\mathbf{g}_t$, that drives the country specific factors. This can be extracted from $\widehat{\mathbf{f}}_{it}$ as PCs of the country-specific PCs. In addition to common factor dependence, countries are also linked through trade and other more local features (culture, language). Such 'network' effects can be captured by using 'starred' variables, to use the GVAR terminology. A simple example would be (for $i = 0,1,\ldots,N$)

$$\begin{aligned} y_{i,t+h} = c_{ih} + \delta_{ih} y_{it}^* + \mathbf{a}_{ih}' \boldsymbol{z}_{it} + \sum_{j=1}^{K_i} \beta_{ijh} I(j \in DGP_i) x_{i,jt} + \\ \theta_{ih}' \mathbf{f}_{it} + \psi_{ih}' \mathbf{g}_t + u_{i,h,t+h}, \end{aligned} \tag{13}$$

where $i = 0$ represents United Kingdom, and $y_{it}^* = \sum_{j=1}^{N} w_{ij} y_{jt}$, w_{ij} (trade weights) measures the relative importance of country j in determination of country i^{th} target variable. Similarly, $\boldsymbol{z}_{it}^* = \sum_{j=1}^{N} w_{ij}^* \mathbf{z}_{jt}$ can also be added to the model if deemed necessary.

The network effects can be included either as an element of $\boldsymbol{z}_{it}$ or could be made subject to variable selection. The problem becomes much more complicated if we try to relate $y_{i,t+h}$ simultaneously to $y_{i,t+h}^*$. Further, for forecasting, following Chudik *et al.* (2016), one might also need to augment the UK regressions with time series, forecasting models for the common factors.

Equation (13) allows for a number of different approaches to dimension reduction. As has been pointed out by Wainwright (2019): 'Much of high-dimensional statistics involves constructing models of high-dimensional phenomena that involve some implicit form of low-dimensional structure, and then studying the statistical and computational gains afforded by exploiting this structure'. Shrinkage methods, like PCs, assume a low dimensional factor structure. The two selection procedures that we have considered, Lasso and OCMT, exploit different aspects of the low-dimensional sparsity structure assumed for the underlying data generating process. Lasso restricts the magnitude of the correlations within and between the signals and the noise variables. OCMT limits the rate at which the number of proxy variables rises with the sample size. GOCMT relaxes this restriction by filtering out the effects of latent factors that bind the proxies to the true signals before implementing the OCMT procedure.

5. Forecasting UK inflation

5.1. Introduction

We apply the procedures proposed above to the problem of forecasting quarterly UK inflation at horizons $h = 1,2$ and 4. The target variable is the headline rate, average annual UK inflation, which is also forecast by the Bank of England. It is labelled DPUK4, defined as $\pi_{t+h} = 100 \times \log\left(p_{t+h}/p_{t+h-4}\right)$, where p_t is the UK consumer price index taken from the IMF International Financial Statistics. Forecasting annual rates of inflation at quarterly frequencies is subject to the overlapping observations problem when $h > 1$, and it is important that the preselected variables in $\boldsymbol{z}_t$, or the variables include in the active set $\mathcal{S}_{K,t} = \{x_{jt}, j = 1,2,\ldots,K\}$ are all predetermined (known) at time t. Furthermore, as discussed earlier, the variables selected for forecasting inflation at different horizons need not be the same, and the selected variables are also likely to change over time.

Since we have emphasised the importance of international network effects, we need to use a quarterly data set that includes a large number of countries to estimate global factors and to allow the construction of the y_{it}^* variables that appear in (13). The GVAR data set provides such a source. The publicly available data set compiled by Mohaddes and Raissi (2024) covers 1979q1–2023q3. We are very grateful to them

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

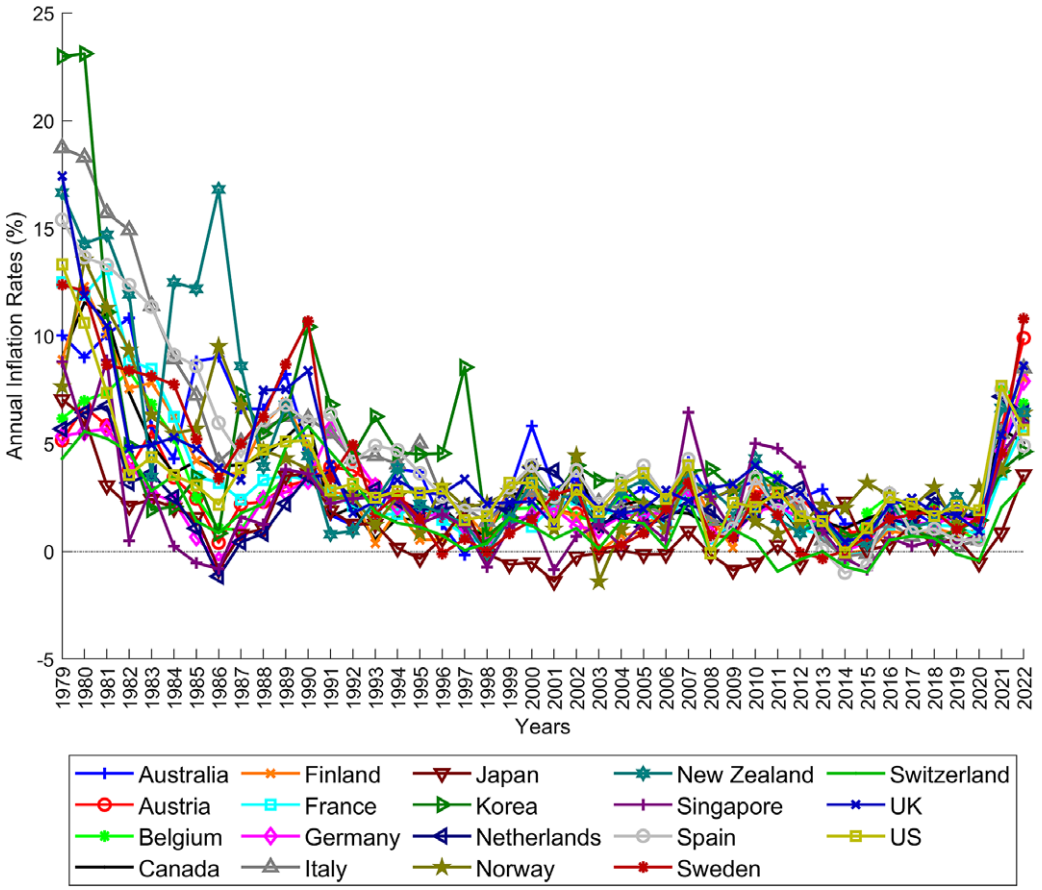

Figure 2. Inflation across advanced economies.

for extending the data. While the latest GVAR data set goes up to 2023q3, we only had access to data till 2023q1 when we started the forecasting exercise, the results of which are reported in this article, but the data that we used matches that GVAR 2023 vintage, which was released in January 2024.[7]

The database includes quarterly macroeconomic data for 6 variables (log real GDP, y; the rate of inflation, dp; short-term interest rate, r; long-term interest rate, lr; the log deflated exchange rate, ep and log real equity prices, eq), for 33 economies as well as data on commodity prices (oil prices, poil, agricultural raw material, pmat and metals prices, pmetal). These 33 countries cover more than 90% of world GDP. The GVAR data were supplemented with other specific UK data on money, wages, employment and vacancies, in the construction of the active set discussed below.

In the light of the argument in Chudik *et al.* (2023), we use the full sample beginning in 1979q1 for variable selection. There are arguments for down-weighting earlier data for estimation when there have been structural changes, as discussed by Pesaran *et al.* (2013). However, the full sample was used both for variable selection and estimation of the forecasting model in order to allow evidence from the earlier higher inflation regime to inform both aspects.

Two sets of variables are considered for inclusion in $\boldsymbol{z}_t$. The first set, which we label AR2, includes lags of the target variable π_t, π_{t-1} (or equivalently π_t and $\Delta\pi_t$). Given the importance we attach to global

[7]GVAR Data 1979q1–2023q3 (2023 Vintage) is available at https://www.mohaddes.org/gvar. Further material on the GVAR is provided at https://sites.google.com/site/gvarmodelling/gvar-toolbox.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

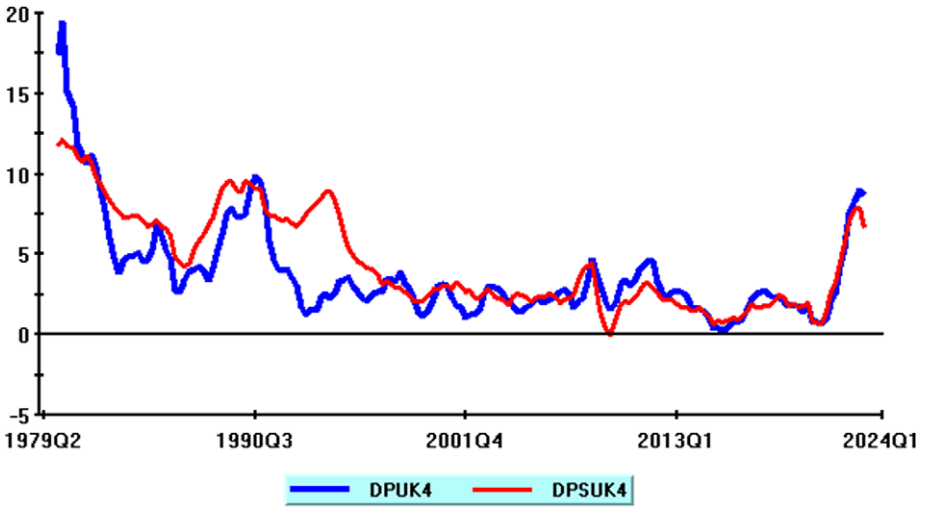

Figure 3. UK and UK-specific global inflation.

variables and network effects, the second set, which we label ARX2, also includes π_t^*, π_{t-1}^* (or equivalently π_t^* and $\Delta\pi_t^*$) where π_t^* is a measure of UK-specific foreign inflation constructed using UK trade weights with the other countries.[8]

If there is a global factor in inflation, the inflation rates of different countries will be highly correlated and tend to move together. Figure 2 demonstrates that this is in fact the case. It plots the inflation rates for 19 countries over the period 1979–2022. It is clear that they do move together, reflecting a strong common factor. The dispersion is somewhat greater in the high inflation period of the 1980s. At times, individual countries break away from the herd with idiosyncratic bursts of inflation, like New Zealand in the mid-1980s. However, it is striking that inflation in every country increases from 2020.

To demonstrate the importance of the global factor for the United Kingdom, Figure 3 plots π_t, and π_t^*, UK inflation and UK-specific foreign inflation. The two series move together, and from the mid-1990s they are very close. This indicates that not only is one unlikely to be able to explain UK inflation just by UK variables but that there are good reasons to include this UK-specific measure of foreign inflation in one of our specifications for $\boldsymbol{z}_t$. The GVAR estimates also indicate the higher sensitivity of the UK to foreign variables than the US or euro area. This is not surprising, they are larger, less open, economies.

5.2. Active set

We now turn to the choice of the members of the active set, $\mathcal{S}_{K,t} = \{x_{jt}, j = 1,2,\ldots,K\}$, some of which may be included in $\boldsymbol{z}_t$. While our focus is on forecasting not on building a coherent economic model, our choice of the covariates in the active set is motivated by the large Phillips curve literature, which suggests important roles for demand, supply and expectations variables. The demand and supply variables in the active set are both domestic and foreign from both product and labour markets. Expectations are captured by financial variables. Interaction terms were not included, but the non-linearities that have been investigated in the literature may be picked up by the latent foreign variables. These are represented by UK-specific measures of foreign inflation and output that could be viewed as estimates of $\mathbf{g}_t$ that are tailored to the UK in relation to her trading partners.

[8]Specifically, $\pi_t^* = \sum_j w_j \pi_{jt}$, where π_{jt} is the inflation rate in country j and w_j is the trade weight of country j with United Kingdom.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

Table 1. List of covariates for UK inflation forecasting

Variable	Description
DPUK4	Four-quarter UK rate of inflation
DYUK4	Four-quarter rate of change of UK real GDP
GAPUK8	UK log real GDP relative to its 8-quarter moving average
GAPUK12	UK log real GDP relative to its 12-quarter moving average
DEMUK4	Four-quarter rate of change of UK employment
DVUK4	Four-quarter rate of change of UK vacancies
DUUK	Four-quarter change in UK unemployment
DWUK4	Four-quarter rate of change of average weekly earnings
RUK4	Four-quarter average UK short interest rate
LRUK4	Four-quarter average UK long interest rate
DMUK4	Four-quarter rate of change of UK M4 money
DEQUK4	Four-quarter rate of change of UK real equity prices
DPOIL4	Four-quarter rate of change of oil prices
DPMAT4	Four-quarter rate of change of material prices
DPMETAL4	Four-quarter rate of change of metal prices
DPMUK4	Four-quarter rate of change of UK import prices
DEPUK4	Four-quarter rate of change of UK deflated dollar exchange rate
DPSUK4	UK-specific measure of four-quarter foreign inflation
DYSUK4	Four-quarter rate of change of UK-specific foreign real GDP
RUS4	Four-quarter average US short interest rate
LRUS4	Four-quarter average US long interest rate
DYCHINA4	Four-quarter rate of change of Chinese real GDP
GAPSUK8	UK-specific log foreign real GDP relative to 8 quarter moving average
GAPSUK12	UK-specific log foreign real GDP relative to the 12-quarter moving average
DYUS4	Four-quarter rate of change of US real GDP
DPUS4	Four-quarter US rate of inflation

Accordingly, we consider 26 covariates $\left(x_{jt}\right)$ listed in Table 1, and their changes $\Delta x_{jt} = x_{jt} - x_{j,t-1}$, giving an active set with $K = 52$ variables to select from. Whereas in a regression including current and lagged values of a regressor (say x_t and x_{t-1}) is equivalent to including current and change (namely x_t and Δx_t), in selection the two specifications can result in different outcomes. Including Δx_t is better since, as compared to x_{t-1}, it is less correlated with the level of the other variables in the active set. The 26 included covariates are measured as four-quarter rates of change, changes or averages to match the definition of the target variable. The rates of change are per cent per annum.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

UK goods market demand indicators: rate of change of output, two measures of the output gap: log output minus either a $P=8$ or $P=12$ quarter moving average of log output, $Gap(y_t,P)=y_t-P^{-1}\sum_{p=1}^{P}y_{t-p}$;

UK labour market demand indicators: rate of change of UK employment, vacancies and average weekly earnings and the change in unemployment;

UK financial indicators: annual averages of UK short and long interest rates, the rate of change of money, UK M4 and of UK real equity prices;

Global cost pressures on the UK: rate of change of the price of oil, metals, materials, UK import prices and deflated dollar exchange rate;

Foreign demand and supply variables: UK-specific global measures, foreign inflation, rate of change of foreign output and two measures of the foreign output gap: log foreign output minus either an 8- or 12-quarter moving average of log foreign output. In addition, large country variables were added: annual average of US short and long interest rates, rates of change of US output and prices, and of Chinese output.

5.3. Variable selection

5.3.1. Variable selection procedures

We consider Lasso, Lasso conditional on $\boldsymbol{z}_t$ and GOCMT conditional on $\boldsymbol{z}_t$. With Lasso, the variables are standardised in-sample before implementing variable selection. The Lasso penalty parameter, λ_T, is estimated using 10 fold CV, across subsets of the observations. As noted above, the assumptions needed for standard CV procedures, for instance those used in the program cv.glmnet, are not appropriate for time series. Time series show features such as persistence and changing variance that are incompatible with those assumptions. In the standard procedure the CV subsets (folds) are typically chosen randomly. This is appropriate if the observations are independent draws from a common distribution, but this is not the case with time series. Since order matters in time series, we retain the time order of the data within each subset. See Bergmeir *et al.* (2018) who provide Monte Carlo evidence on various procedures suggested for the case of serially correlated data. We use all the data, and do not leave gaps between subsets. In addition, the standard procedure chooses the $\widehat{\lambda}_{hT}$ that minimises the pooled MSE over the 10 subsets. But when variances differ substantially over subsets pooling is not appropriate, instead we follow Chudik *et al.* (2018, CKP), and use the average of the $\widehat{\lambda}_{hT}$ chosen in each subset. Full details are provided in the Appendix to CKP (2018).

As well as standard Lasso, for consistency with OCMT, we also generated Lasso forecasts conditional on $\boldsymbol{z}_t$ by including a preselected set of variables $\boldsymbol{z}_t$ in the optimisation problem (5). This generalised Lasso procedure solves the following optimisation problem:

$$\min_{\mathbf{a}_h,\beta_h}\left\{\sum_{t=1}^{T}\left(y_{t+h}-c_h-\mathbf{a}_h'\mathbf{z}_t-\beta_h'\boldsymbol{x}_t\right)^2+\lambda_T\sum_{i=1}^{K}|\beta_{hi}|\right\}, \tag{14}$$

where the penalty is applied only to the variables in the active set, $\boldsymbol{x}_t$, and not to the preselected variables, $\boldsymbol{z}_t$. The above optimisation problem can be solved in two stages. In the first stage, the common effects of $\boldsymbol{z}_t$ are filtered out by regressing y_{t+h} and $\boldsymbol{x}_t$ on the preselected variables $\boldsymbol{z}_t$ and saving the residuals $e_{y.z}$ and $e_{xj.z}$, $j=1,2,..,K$. In the second stage, Lasso is applied to these residuals. A proof that this two-step procedure solves the constrained minimisation problem in (14) is provided by Sharifvaghefi and reproduced in the Appendix.

In the OCMT critical value function, $c_p(K,\delta)=\Phi^{-1}\left(1-\frac{p}{2K^{\delta}}\right)$, we set $p=0.05$ and $\delta=1$. With $K=52$, this means that we only retain variables with t-ratios (in absolute value) exceeding $c_{0.05}(52,1)=3.3$. To allow for possible serial correlation, we also experimented with setting $\delta=1.5$, which yields,

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

$c_{0.05}(52,1.5) = 3.82$. The results were reasonably robust and we focus on the baseline choice of $\delta = 1$, also recommended by CKP.

We implement Lasso and OCMT conditional on two preselected sets of variables, either an AR2 written as level and change $\mathbf{z}_t = (\pi_t, \Delta\pi_t)'$ or given the role of foreign inflation, shown above, the AR2 augmented by the level and change of the UK-specific measure of foreign inflation, denoted ARX, $\mathbf{z}_t = (\pi_t, \Delta\pi_t, \pi_t^*, \Delta\pi_t^*)'$. As noted above, for selection including current and change is better than including current and lag. For comparative purposes, we also generated forecasts with the preselected variables only, namely the AR2 forecasts generated from the regressions

$$AR2: \ \pi_{t+h} = c_h + a_1\pi_t + a_{2h}\Delta\pi_t + u_{h,t+h},$$

and the ARX forecasts generated from

$$ARX: \ \pi_{t+h} = c_h + a_{1h}\pi_t + a_{2h}\Delta\pi_t + a_{3h}\pi_t^* + a_{4h}\Delta\pi_t^* + u_{h,t+h}.$$

Variable selection is carried out recursively, for each forecast horizon h separately, using an expanding windows approach. All data samples start in 1979q2 and end in the quarter that forecasts are made. To forecast the average inflation over the four quarters to 2020q1 using a forecast horizon of $h = 4$, the sample used for selection and estimation ends in 2019q1. The end of the sample is then moved to 2019q2 to forecast the average inflation over the four quarters to 2020q2, and so on. Similarly, to forecast the average inflation over the four quarters to 2020q1 using $h = 2$, the sample ends in 2019q3, and using $h = 1$, the sample ends in 2019q4. These sequences continue one quarter at a time until the models are selected and estimated to forecast inflation over the four quarters to 2023q1. Thus, for $h = 4$, there are 17 samples used for variable selection, while for $h = 2$ and $h = 1$ there are 15 and 14 such variable selection samples. This process of recursive model selection and estimation means that the variables selected can change from quarter to quarter, and for each forecast horizon, h.

Section S-3 of the online supplement list the variables selected by each of the procedures, for each quarter and each forecast horizon. The main features are summarised here.

5.3.2. Number of variables selected

Table 2 gives the minimum, maximum, and average number of variables selected for the three forecast horizons and five variable selection procedures. Except for AR2-OCMT, at $h = 4$, OCMT chooses fewer variables than Lasso. Lasso conditional on the preselected variables selects a larger number of variables in

Table 2. Number of variables selected by Lasso and OCMT including preselected

	Forecast horizon, h, in quarters								
	Total number of preselected and selected variables								
	$h=1$			$h=2$			$h=4$		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Lasso	7	12	8.1	5	9	6.1	3	6	5.2
AR2-Lasso	5	11	8.2	9	16	13.5	8	11	9.5
AR2-OCMT	2	3	2.2	4	5	4.5	5	14	6.2
ARX-Lasso	8	16	12.4	12	19	16.3	2	15	9.9
ARX-OCMT	4	4	4	5	5	5	5	8	5.8

Note: The reported results are based on 14, 15 and 17 variable selection samples for 1-, 2- and 4-quarter ahead models, respectively. The AR2 and ARX components include two and four preselected variables, respectively.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

Table 3. Estimates of the Lasso penalty parameter computed by 10-fold cross-validation procedure

	Forecast horizon, h, in quarters								
	$h=1$			$h=2$			$h=4$		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Lasso	0.07	0.12	0.11	0.22	0.31	0.26	0.33	0.47	0.41
AR2–Lasso	0.08	0.10	0.09	0.09	0.14	0.11	0.18	0.27	0.22
ARX–Lasso	0.04	0.08	0.06	0.07	0.12	0.09	0.15	0.33	0.22

Note: The reported estimates are based on Lasso penalty estimates (obtained from 10-fold cross-validation) for 14, 15 and 17 variable selection samples for 1-, 2- and 4-quarter ahead models, respectively.

total than the standard Lasso without conditioning. Conditioning on preselected variables is much more important for OCMT as compared to Lasso. This finding is in line with the theoretical results obtained by Sharifvaghefi (2023) who establishes the importance of conditioning on the latent factors when applied to an active set with highly correlated covariates. The number of variables Lasso selects falls with the forecast horizon,[9] while the number of variables selected by OCMT rises with the horizon. These results show that Lasso and OCMT could select very different models for forecasting.

As expected, the number of variables selected by Lasso correlates with the estimates of the penalty parameter $\widehat{\lambda}_{hT}$, computed by CV. These values are summarised in Table 3. For all three Lasso applications, the mean of the estimated penalty parameter increases with the forecast horizon, though more slowly for the specifications that include preselected variables. For Lasso (without preselection), the number of variables selected falls with the forecast horizon because of the increasing penalty parameter. This is not as clear-cut for the specifications including preselected variables.

5.3.3. OCMT: selected variables by horizon

OCMT selects only a few variables in addition to the preselected UK and UK-specific foreign inflation ($\pi_t, \Delta\pi_t$, π_t^* and $\Delta\pi_t^*$).[10] For $h=1$, the variables selected are given in sub-section S-3.1.1 of the online supplement. In addition to the two preselected variables ($\pi_t, \Delta\pi_t$), AR2-OCMT selects the rate of change of wages (DWUK4) for samples ending in 2021q2, 2021q4 and 2022q1, and no other variables. ARX-OCMT does not select any additional variables for any of the 14 variable selection samples!

For $h=2$, the variables selected are given in Section S-3.1.2 of the online supplement. AR2-OCMT selects the rates of change of money (DMUK4) and exchange rate (DEPUK4) from samples ending in 2019q3 to 2020q3, then the rate of change of wages is added till the sample ending in 2022q2, from then the rates of change of money (DMUK4) and wages (DWUK4) are selected. ARX-OCMT selects just the rate of change of money (DMUK4) as an additional variable in every sample for $h=2$.

For $h=4$, the variables selected are given in Section S-3.1.3 of the online supplement. AR2-OCMT chooses the same 3 extra variables—the rate of change of money (DMUK4) and exchange rate (DEPUK4) as well as the UK-specific measure of foreign inflation (DPSUK4, π_t^*)—for every sample ending from 2019q1 to 2021q1. Then, in the sample ending in 2021q2, AR2-OCMT chooses 12 extra variables. The number of variables selected then falls to 7 in 2021q3, 6 in 2021q4, 5 in 2022q1 and 4 in 2022q2 − 2022q4. These four are the rate of change of money (DMUK4), of material prices (DPMAT4), of wages (DWUK4), and π_t^* (DPSUK4). The number of variables selected falls to 3 in the sample ending

[9]This is possibly because, as h increases the β_h^0 get smaller in (6), the IRC is more likely to be satisfied and Lasso is less likely to falsely select additional variables.

[10]The OCMT results reported here are based on the critical value function given by (12) with $\delta=1$. Using the larger value of $\delta=1.5$ reduces the number of selected variables for a few sample periods, but the outcomes are generally robust to the choice on δ on the interval [1,1.5]. The selection results for OCMT with $\delta=1.5$ are reported in the online supplement.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

Table 4. The number of times covariates from the active set are selected by Lasso at different forecast horizons

	Horizon			
	$h=1$	$h=2$	$h=4$	Selected covariates
DPUK4	14	15	17	Four-quarter UK rate of inflation
DPSUK4	14	15	17	UK-specific measure of four-quarter foreign inflation
DWUK4	14	15	14	Four-quarter rate of change of average weekly earnings
DDPUK4	14	15	0	Change in four-quarter UK rate of inflation
DPMUK	14	8	0	Four-quarter rate of change of UK import prices
DLRUK4	14	4	0	Change in four-quarter average UK long interest rate
DDPSUK4	14	3	0	Change in UK-specific measure of four-quarter foreign inflation
DMUK	10	15	17	Four-quarter rate of change of UK M4 money
DRUK4	2	0	0	Change in four-quarter average UK short interest rate
DEMUK4	1	0	6	Four-quarter rate of change of UK employment
DDPOIL4	1	0	0	Change in four-quarter rate of change of the oil price
DDEQUK4	1	0	0	Change in four-quarter rate of change of UK equity prices
DDYSUK4	1	0	0	Change in four-quarter rate of change of UK foreign real GDP
DGAPSUK12	1	0	0	UK log foreign real GDP relative to eight-quarter moving average
DPUS4	0	2	2	Four-quarter US rate of inflation

Note: The number of variable selection samples are 14, 15 and 17 for $h=1$, 2 and 4 quarter ahead models, respectively.

in 2023q1 when the foreign inflation measure is no longer selected. ARX-OCMT chooses the rate of change of employment (DEMUK4) and of money for samples ending in 2019q1 − 2021q4, then adds material prices in 2022q1, and selects just the rate of change of money for the last four samples.

5.3.4. Lasso: selected variables by horizon

Lasso selections for each sample and horizon are given in the online supplement, Section S-3.2. Lasso tends to select more variables than OCMT so we give less detail. Table 4 lists the variables chosen by standard Lasso at each horizon and the number of times they were chosen out of the maximum number of possible samples: 14, for $h=1$, 15 for $h=2$ and 17 for $h=4$. UK inflation, π_t (DPUK4) is always chosen in every sample at every horizon as is the UK measure of foreign inflation, π_t^* (DPSUK4). The change in UK inflation, $\Delta\pi_t$, (DDPUK4) is chosen in every sample in the case of models for $h=1$, and $h=2$, but never for $h=4$. The change in foreign inflation $\Delta\pi_t^*$ (DDPSUK4) is chosen in every sample at $h=1$, in 3 samples at $h=2$ but never at $h=4$. Thus, Lasso provides considerable support for the choice of preselected variables in $\boldsymbol{z}_t$ that include foreign inflation as well as the two lagged inflation variables.

Apart from these variables, the rate of change of wages and of money figure strongly when using Lasso. The rate of change of wages (DWUK4) is chosen in all the samples for $h=1$ and $h=2$ and 14 of the 17 samples for $h=4$. The rate of change of money (DMUK) is chosen in 10 of the 14 samples for $h=1$, and in every sample for $h=2$ and $h=4$. Money and wages are also chosen by OCMT but the rate of change of the exchange rate selected by OCMT is never chosen by Lasso.

When $h=1$, Lasso also always selects two other variables, namely the change in long interest rates (DLRUK4), and import price inflation (DPMUK).

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

5.4. Forecasts

The point forecasts of inflation for $h = 1, 2$ and $h = 4$ for the various selection procedures are summarised in Section S-4 of the online supplement. For each forecast horizon, we have 13 forecasts and their realisations for the quarters 2020q1 to 2023q1 inclusive. These are summarised in Table 5. We use the root mean square forecast error (RMSFE) as our forecast evaluation criterion. Since 13 forecast errors represent a very short evaluation sample with considerable serial correlation, testing for the significance of the loss differences using the Diebold and Mariano (1995) test would not be reliable and is not pursued here.

5.4.1. One quarter ahead forecasts

Figure 4 gives the plots of actual inflation and forecasts one quarter ahead. Section S-4.1 of the online supplement gives the point forecasts. For $h = 1$, ARX has the lowest RMSFE, the π_t^* and $\Delta\pi_t^*$ improve forecast performance relative to the AR2. AR2-OCMT adds wage growth in three periods. Lasso suffers from choosing too many variables relative to OCMT. The forecasts are very similar, except Lasso predicted a large drop in 2020q3 with a subsequent rebound. This results from selecting an output gap

Table 5. Root mean square forecast errors by forecast horizons, h = 1, 2 and 4 over the period 2020q1–2023q1

Forecast source	h = 1	h = 2	h = 4
AR2	0.9141	1.6039	3.2524
ARX	**0.7884**	**1.3813**	2.9883
Lasso	0.9696	1.4109	3.0131
AR2-Lasso	0.9617	2.8719	4.3440
ARX-Lasso	0.8750	2.9231	4.5021
AR2-OCMT	0.9233	1.6800	**2.4643**
ARX-OCMT	**0.7884**	1.4217	3.2470
Bank of England	1.3288	1.6959	3.0042

Note: The RMSFE figures are taken from online supplement Tables S-4.1–S-4.3. The least value for RMSFE for each forecast horizon is shown in bold.

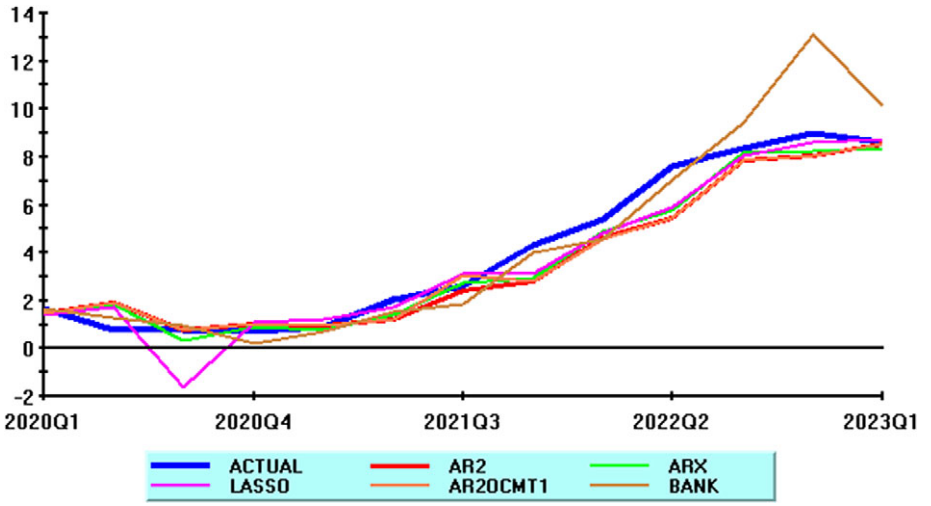

Figure 4. One quarter ahead forecasts.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

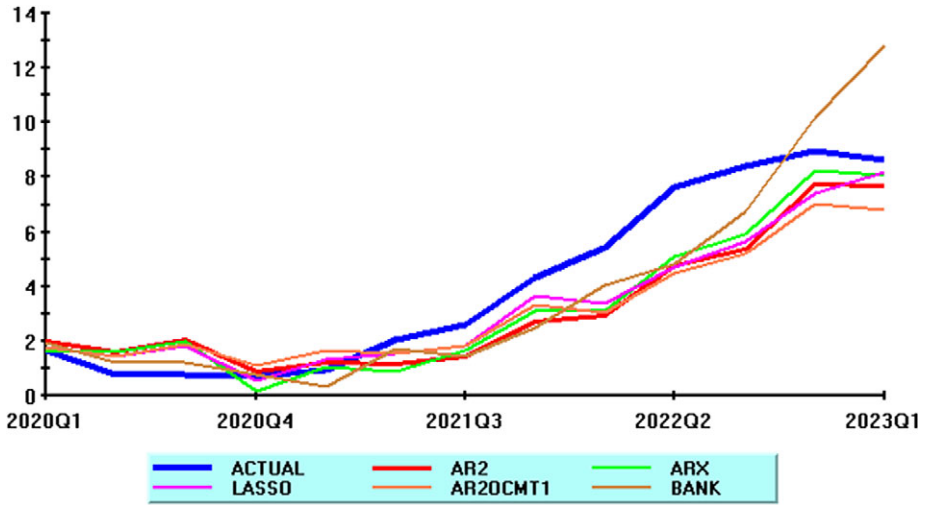

Figure 5. Plot of forecasts two quarters ahead.

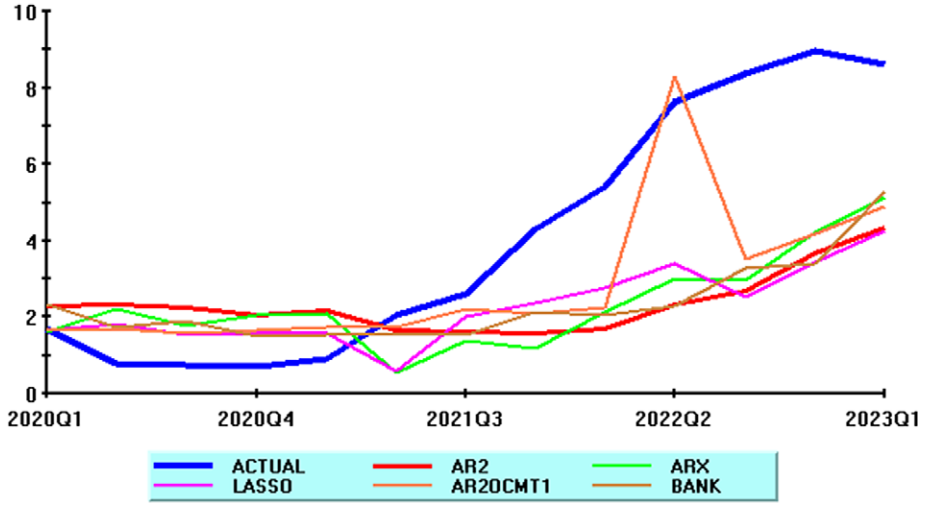

Figure 6. Plot of forecasts four quarters ahead.

measure, when UK output dropped sharply in 2020*q*2. This sharp drop and rebound was also a feature of Lasso forecasts at other horizons. The Bank of England overestimated inflation in 2022*q*4, correctly anticipating higher energy prices but not anticipating the government energy price guarantees.

5.4.2. Two quarter ahead forecasts

Figure 5 gives the plots of actual inflation and forecasts two quarters ahead. Section S-4.2 of the online supplement gives the values. For $h = 2$, ARX again has the lowest RMSFE. ARX-OCMT selects money growth in every period. Lasso selects between five and nine variables.

5.4.3. Four quarter ahead forecasts

Figure 6 gives the plots of actual inflation and forecasts four quarters ahead. Section S-4.3 of the online supplement gives the values. The case of $h = 4$ is the only one where the ARX does not have the lowest

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

Table 6. Contemporaneous and lagged effects of oil price changes on UK inflation (π_{t+1})

Covariates	1979q2–2019q4		1979q2–2022q4	
π_t	**0.936**	**0.936**	**0.957**	**0.962**
	(13.40)	(12.97)	(13.70)	(13.23)
π_{t-1}	−0.134	−0.136	−0.145	−0.151
	(−2.05)	(−2.02)	(−2.20)	(−2.21)
π_t^*	0.480	0.512	0.600	0.524
	(4.14)	(3.31)	(5.50)	(4.39)
π_{t-1}^*	−0.381	−0.316	−0.504	−0.432
	(−3.28)	(−2.52)	(−4.64)	(−3.61)
$\Delta poil_{t+1}$		–	0.012	–
	(3.40)	–	(3.91)	–
$\Delta poil_t$	–	0.006	–	0.005
	–	(1.47)	–	(1.37)
$\bar{R}^2$	0.953	0.950	0.952	0.948
SER	0.612	0.631	0.626	0.651

RMSFE. The lowest RMSFE is obtained by AR2-OCMT. It does well by having a very high inflation forecast in 2022*Q*2. This corresponds to the selection of 12 extra variables in the sample ending in 2021*q*2. It then rejoins the pack in 2022*q*3.

5.4.4. Summary

Table 5 brings together the RMSFE for each of the selection methods at different horizons. Both the variable selection and forecasting exercises highlight the importance of taking account of persistence and foreign inflation for UK inflation forecasting. Lasso selects π_t and π_t^* in all three forecast horizon models. ARX, which includes UK and foreign inflation as preselected variables, tends to perform best in forecasting, but in the present application, the OCMT component does not seem to add much once the preselected variables are included. However, Lasso performs rather poorly when it is conditioned on the preselected variables.

5.5. Contemporaneous drivers

Our forecasts of π_{t+h} are based on variables observed at time t, and do not depend on any conditioning. But, as noted above with respect to the Bank of England, it is common to condition on contemporaneous values of variables that are considered as *proximate* causes of the variable to be forecast. Even if such causal variables can be identified; however, it does not mean that they help with forecasting—often such causal variables are themselves difficult to be forecast. The Bank of England overestimated inflation in 2022q4, correctly anticipating higher energy prices but not anticipating the government energy price guarantees. This illustrates the dangers of conditioning on variables that cannot be forecasted. Understanding does not necessarily translate into better forecasts. For example, knowing the causes of earthquakes does not necessarily help in predicting them in a timely manner.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

This point can be illustrated by including contemporaneous changes in oil prices, in the UK inflation equation (over the period pre Covid-19 and the full sample) for the case $h = 1$. For both samples, $\Delta poil_{t+1}$ are highly statistically significant, but their lagged values $\Delta poil_t$ are not (Table 6).

6. Conclusion

High-dimensional data are not a panacea; the data must have some predictive content that might come from spatial or temporal sequential patterns. Forecasting is particularly challenging either if there are unknown unknowns (factors that are not even thought about) or if there are known factors that are falsely believed to be important. When there are new global factors, like Covid-19, or when a relevant variable has shown little variation over the sample period, forecasting their effect is going to be problematic.

Many forecasting problems require a hierarchical structure where latent factors at local and global levels are explicitly taken into account. This is particularly relevant for macro forecasting in an increasingly interconnected world. It is important that we allow for global factors in national forecasting exercises—and GVAR was an attempt in this direction.

A number of key methodological issues were illustrated with a simple approach to forecasting UK inflation, which has become a topic of public discussion. This example showed both the power of parsimonious models and the importance of global factors. There remain many challenges. How to allow for regime change and parameter instability in the case of high-dimensional data analysis? How to choose data samples? Our recent research suggests that it is best to use long time series samples for variable selections, but consider carefully what sample to use for forecasting. Given a set of selected variables, parameter estimation can be based on different window sizes or down-weighting. Should we use ensemble or forecast averaging? Forecast averaging will only work if the covariates used to forecast the target variable are driven by strong common factors; otherwise, one will be averaging over noise.

There are some more general lessons. Econometric and statistical models must not become a straightjacket. Forecasters should be open minded about factors not included in their model, and acknowledge that forecasts are likely to be wrong if unexpected shocks hit.

Acknowledgements. This article was developed from Pesaran's Deane-Stone Lecture at the National Institute of Economic and Social Research, London. The authors greatly benefitted from comments when earlier versions of this article were presented in 2023 at NIESR on 21 June, the International Association of Applied Econometrics Annual Conference in Oslo, June 27-30, 2023, Bayes Business School, City University, 22 November 2023, and at the Economics Department-Wide Seminar, Emory University, 16 February, 2024. The authors are also grateful for comments from Alex Chudik, Anthony Garratt, George Kapetanios, Essie Maasoumi, Alessio Sancetta, Mahrad Sharifvaghefi, Allan Timmermann and Stephen Wright. The authors particularly thank Hayun Song, their research assistant, for his work in coding, empirical implementation, and the tabulation of results, along with his invaluable research assistance.

References

Bergmeir, C., Hyndman, R.J. and Koo, B. (2018), 'A note on the validity of cross-validation for evaluating autoregressive time series prediction', *Computational Statistics & Data Analysis*, **120**, pp. 70–83.

Bernanke, B.S., Boivin, J. and Eliasz, P. (2005), 'Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach', *The Quarterly Journal of Economics*, **120**, pp. 387–422.

Chudik, A., Grossman, V. and Pesaran, M.H. (2016), 'A multi-country approach to forecasting output growth using PMIs', *Journal of Econometrics*, **192**, pp. 349–365.

Chudik, A., Kapetanios, G. and Pesaran, M.H. (2018), 'A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models', *Econometrica*, **86**, pp. 1479–1512.

Chudik, A. and Pesaran, M.H. (2016), 'Theory and practice of GVAR modelling', *Journal of Economic Surveys*, **30**, pp. 165–197.

Chudik, A., Pesaran, M.H. and Sharifvaghefi, M. (2023), "Variable selection in high dimensional linear regressions with parameter instability," arXiv:2312.15494 [econ.EM] 24 Dec. 2023, available online at https://arxiv.org/abs/2312.15494.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

Diebold, F.X. and Mariano, R.S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics*, **13**, pp. 253–263.
Fan, J., Ke, Y. and Wang, K. (2020), 'Factor-adjusted regularized model selection', *Journal of Econometrics*, **216**, pp. 71–85.
Giannone, D., Lenza, M. and Primiceri, G.E. (2021), 'Economic predictions with big data: The illusion of sparsity', *Econometrica*, **89**, pp. 2409–2437.
Granger, C.W. and Pesaran, M.H. (2000a), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting*, **19**, pp. 537–560.
Granger, C.W.J. and **Pesaran, M.H.** (2000b), 'A decision theoretic approach to forecast evaluation', in *Statistics and Finance: An Interface*, London, World Scientific, pp. 261–278.
Hansen, C. and Liao, Y. (2019), 'The factor-lasso and k-step bootstrap approach for inference in high-dimensional economic applications', *Econometric Theory*, **35**, pp. 465–509.
Lahiri, S.N. (2021), 'Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions', *The Annals of Statistics*, **49**, pp. 820–844.
Marcellino, M., Stock, J.H. and Watson, M.W. (2006), 'A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series', *Journal of Econometrics*, **135**, pp. 499–526.
Meinshausen, N. and Bühlmann, P. (2006), 'Variable selection and high-dimensional graphs with the lasso', *Annals of Statistics*, **34**, pp. 1436–1462.
Mohaddes, K. and Raissi, M. (2024), 'Compilation, Revision and Updating of the Global VAR (GVAR) Database, 1979Q2-2023Q3', *Mendeley Data*, **V1**, pp. 1–9.
Pesaran, M.H., Pick, A. and Pranovich, M. (2013), 'Optimal forecasts in the presence of structural breaks', *Journal of Econometrics*, **177**, pp. 134–152.
Pesaran, M.H., Pick, A. and Timmermann, A. (2011), 'Variable selection, estimation and inference for multi-period forecasting problems', *Journal of Econometrics*, **164**, pp. 173–187.
Pesaran, M.H and **Skouras, S.** (2004), 'Decision-based methods for forecast evaluation', in Michael, D.F.H. and Clements, P. (eds), *A Companion to Economic Forecasting, Chap. 11*, Oxford, Wiley Online Library, pp. 241–267.
Sharifvaghefi, M. (2023), 'Variable selection in linear regressions with many highly correlated covariates', available online at https://ssrn.com/abstract=4159979.
Shrader, J.G., Bakkensen, L. and Lemoine, D. (2023), 'Fatal Errors: The Mortality Value of Accurate Weather Forecasts', Working Paper Series, National Bureau of Economic Research, Number 31361.
Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**, pp. 267–288.
Wainwright, M.J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge: Cambridge University Press.
Whittle, P. (1983), *Prediction and Regulation by Linear Least-Square Methods*, Minneapolis, University of Minnesota Press.
Zhao, P. and Yu, B. (2006), 'On model selection consistency of Lasso', *The Journal of Machine Learning Research*, **7**, pp. 2541–2563.
Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society B*, **67**, pp. 301–320.

Appendix

The Lasso procedure with a set of preselected variables[11]

Let $\mathbf{y} = (y_1, y_2, \cdots, y_T)'$ be the vector of observations for the target variable. Suppose we have a vector of pre-selected covariates denoted by $\mathbf{z}_t = (z_{1t}, z_{2t}, \cdots, z_{mt})'$. Additionally, there is a vector of covariates denoted by $\mathbf{x}_t = (x_{1t}, x_{2t}, \cdots, x_{nt})'$, from which we aim to select the relevant ones for the target variable using the Lasso procedure. We can further stack the observations for $\mathbf{z}_t$ and $\mathbf{x}_t$ in matrices $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T)'$ and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T)'$, respectively. For a given value of the tuning parameter, λ, the Lasso problem can be written as:

$$\left(\widehat{\delta}(\lambda), \widehat{\beta}(\lambda)\right)' = \text{argmin}_{\mathbf{b}_z, \mathbf{b}} \left\{ (\mathbf{y} - \mathbf{Z}\mathbf{b}_z - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{Z}\mathbf{b}_z - \mathbf{X}\mathbf{b}) + \lambda \|\mathbf{b}\|_1 \right\}.$$

[11]We are grateful to Dr. Mahrad Sharifvaghefi for providing the proofs in this appendix.

https://doi.org/10.1017/nie.2024.1 Published online by Cambridge University Press

Partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ is the matrix of covariates with the corresponding vector of estimated coefficients, $\widehat{\boldsymbol{\beta}}_1(\lambda)$, different from zero and $\mathbf{X}_2$ is the matrix of covariates with the corresponding vector of estimated coefficients, $\widehat{\boldsymbol{\beta}}_2(\lambda)$, equal to zero. So, $\mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda) = \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda)$. By the first order conditions we have:

$$\mathbf{X}_1'\left(\mathbf{y} - \mathbf{Z}\widehat{\delta}(\lambda) - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right) - \lambda \text{sign}\left(\widehat{\boldsymbol{\beta}}_1(\lambda)\right) = 0, \tag{A.15}$$

$$\mathbf{Z}'\left(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\delta}}(\lambda) - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right) = 0. \tag{A.16}$$

and

$$-\lambda\mathbf{1} \leq \mathbf{X}_2'\left(\mathbf{y} - \mathbf{Z}\widehat{\boldsymbol{\delta}}(\lambda) - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right) \leq \lambda\mathbf{1}, \tag{A.17}$$

Where $\mathbf{1}$ represents a vector of ones. We can further conclude from Equation (A.16) that:

$$\widehat{\boldsymbol{\delta}}(\lambda) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right). \tag{A.18}$$

By substituting $\widehat{\boldsymbol{\delta}}(\lambda)$ from (A.18) into (A.15), we have

$$\mathbf{X}_1'\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda) - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right)\right) - \lambda \text{sign}\left(\widehat{\boldsymbol{\beta}}_1(\lambda)\right) = 0.$$

We can further write this as:

$$\mathbf{X}_1'\left(\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\right)\left(\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right) - \lambda \text{sign}\left(\widehat{\boldsymbol{\beta}}_1(\lambda)\right) = 0.$$

Therefore,

$$\tilde{\mathbf{X}}_1'\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right) - \lambda \text{sign}\left(\widehat{\boldsymbol{\beta}}_1(\lambda)\right) = 0, \tag{A.19}$$

where $\tilde{\mathbf{X}}_1 = \mathbf{M}_Z\mathbf{X}_1$, $\tilde{\mathbf{y}} = \mathbf{M}_Z\mathbf{y}$ and $\mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$.

Similarly, by substituting $\widehat{\delta}(\lambda)$ from (A.18) into (A.17), we have

$$-\lambda\mathbf{1} \leq \tilde{\mathbf{X}}_2'\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}_1\widehat{\boldsymbol{\beta}}_1(\lambda)\right) \leq \lambda\mathbf{1}. \tag{A.20}$$

Note that (A.19) and (A.20) are the first order conditions of the following Lasso problem:

$$\widehat{\beta}(\lambda) = \text{argmin}_{\mathbf{b}}\left\{\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{b}\right)'\left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\mathbf{b}\right) + \lambda\|\mathbf{b}\|_1\right\}. \tag{A.21}$$

Therefore, we can first obtain the estimator of the vector coefficients for $\mathbf{X}$, $\widehat{\boldsymbol{\beta}}(\lambda)$, by solving the Lasso problem given by (A.21) and then estimate the vector of coefficients for $\mathbf{Z}$, $\widehat{\boldsymbol{\delta}}(\lambda)$, by using Equation (A.18).

Cite this article: Pesaran, M. H. and Smith, R. P. (2024), 'High-dimensional forecasting with known knowns and known unknowns', *National Institute Economic Review*, pp. 1–25. https://doi.org/10.1017/nie.2024.1