


# Variable targeting and reduction in large vector autoregressions with applications to workforce indicators

T. S. McElroy & Thomas Trimbur

**To cite this article:** T. S. McElroy & Thomas Trimbur (2023) Variable targeting and reduction in large vector autoregressions with applications to workforce indicators, Journal of Applied Statistics, 50:7, 1515-1537, DOI: [10.1080/02664763.2022.2032619](https://doi.org/10.1080/02664763.2022.2032619)

**To link to this article:** <https://doi.org/10.1080/02664763.2022.2032619>

 View supplementary material 

 Published online: 07 Feb 2022.

 Submit your article to this journal 

 Article views: 46

 View related articles 

 View Crossmark data 



# Variable targeting and reduction in large vector autoregressions with applications to workforce indicators

T. S. McElroy<sup>a</sup> and Thomas Trimbur<sup>b</sup>

<sup>a</sup>Research and Methodology Directorate, U.S. Census Bureau, Washington, DC, USA; <sup>b</sup>Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA

## ABSTRACT

We develop statistical tools for time series analysis of large multivariate datasets, when a few core series are of principal interest and there are many potential auxiliary predictive variables. The methodology, based on Vector Autoregressions (VAR), handles the case where unrestricted fitting is precluded by a large number of series and a huge parameter space. In particular, we adopt a forecast error criterion and use Granger-causality tests in a sequential manner to build a VAR model that targets the main variables. This approach affects variable reduction (or equivalently, sparsity restrictions) in a computationally fast way that remains feasible for large dimensions. The search for the best model results in a VAR, fitted with a selection of supporting series, that has the best possible forecast performance with respect to the core variables. We apply the statistical methodology to model real Gross Domestic Product and the national Unemployment Rate, two time series widely monitored by economists and policy-makers, based on a large set of Quarterly Workforce Indicators comprising various major sectors of the economy and different measures of labor market conditions.

## ARTICLE HISTORY

Received 20 September 2019  
Accepted 13 January 2022

## KEYWORDS

Dimension reduction;  
macroeconomic forecasting;  
GDP; unemployment rate;  
variable selection; VAR


## 1. Introduction

### 1.1. Overview

This paper addresses a key application in economics: we examine a large dataset of over a hundred labor flow indicators developed at the U.S. Census Bureau. The objective is to forecast the national Unemployment Rate (UR) and real Gross Domestic Product (GDP), two major time series closely monitored by the Federal Reserve System and other policy-making entities. Specifically, the indicators from the Longitudinal Employer-Household Dynamics (LEHD) database [1] form a large pool of potential auxiliary variables, which provide extensive information on categorical transitions of the labor force – at both the firm and individual level – across major industries in the U.S. economy.

There are empirical studies and economic arguments behind the use of these kinds of auxiliary series in this context: see [23–25]. In particular, in recent macroeconomic work

**CONTACT** T. S. McElroy  [tucker.s.mcelroy@census.gov](mailto:tucker.s.mcelroy@census.gov)  Research and Methodology Directorate, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233-9100, USA.

 Supplemental data for this article can be accessed here: <https://doi.org/10.1080/02664763.2022.2032619>

This work was authored as part of the Contributor's official duties as an Employee of the United States Government and is therefore a work of the United States Government. In accordance with 17 USC. 105, no copyright protection is available for such works under US Law.

there has been surging interest in using indicators of labor market dynamics for forecasting national unemployment. Substantial gains were reported by Barnichon and Nekarda [3] and Barnichon and Garda [2] in forecast accuracy for UR relative to previous approaches based on traditional indicators or univariate models.

This paper proposes a method that expands the scope of the VAR framework to jointly and coherently address three major statistical challenges – dimension reduction, variable targeting, and variable selection, discussed in detail below – that arise when the number of variables present is relatively large. We define a model or case as ‘large’ in the variable count when either (i) the dimension  $N$  is on the order of, or exceeds the sample size  $T$ , or (ii) the value of  $N$  is sufficiently big to make standard modeling – a single-step or exact statistical treatment – difficult. (Now, many cases encountered in practice will satisfy both criteria; there can in addition be applications that fulfill (ii) but not (i), e.g. when 100 variables are considered at 5000 distinct time points.) We are principally interested in a context where it is important to screen out a large number of irrelevant variables in a short amount of time (e.g. a few minutes), thereby precluding more computationally demanding analyses.

To the best of our knowledge, our methodology is a new ‘recipe’ and represents the first one to concurrently achieve our three aims. We start with a forecast error objective function and modify the function to focus on a small group of core variables. The elimination of candidate predictors occurs by finding those that do *not* ‘Granger cause’ the primary variables [21]; this is equivalent to using sparsity restrictions (i.e. imposing zeros in the coefficient matrices) determined according to predictive ability. The method is implemented with step-wise Wald statistics, which are used to cull the less important auxiliary variables while preserving forecast performance. Thus, the final model contains only the target series and their most effective predictors. Further zeros are imposed by a restricted least squares procedure [7] that enforces sparsity to the extent that the Whittle likelihood is not significantly worsened.

## 1.2. Background

In many sciences observational data can be subject to myriad complicated effects. Particularly when multiple variables are involved, the availability of well-founded prior knowledge – relevant to the variables’ collective properties and sufficiently detailed to inform quantitative statistical analysis – may be somewhat limited. For data recorded over time, Vector Autoregressions (VARs) are attractive as a compact, flexible, and empirically-oriented approach to describing dynamics and inter-relationships among a set of variables. There has been extensive work on VAR methodology applicable to small to moderate dimensions (on the order of 10 or less) and focusing on fluctuations in economic time series; some early examples are [12,18,26,37]. In more contemporary work, researchers have also used the VAR framework to model multivariate datasets in other scientific fields, such as sociology [16], neurology [14], and psychology [6].

In recent years with data availability becoming increasingly comprehensive, researchers have considered the statistical modeling of systems that can have ever larger numbers of variables. ‘Machine-Learning’ procedures can be applied to process large datasets and recognize highly general and flexible patterns that may be used for out-of-sample predictions; their budding use in the context of economic data is summarized and discussed in [33]. The key aspect of such procedures is that they involve ingredients – rudimentary from

a statistical perspective – that can be implemented with a very low computational burden relative to model dimension, making it possible to handle extremely large numbers of variables. Another summary and description of machine-learning techniques are found in [9], which also reports applications to central bank supervisory behavior and to inflation forecasting around the Great Recession. [10] describe applications to human capital, such as predicting the propensity of police academy graduates to maintain proper conduct (e.g. use minimal force when making arrests) based on a large panel of individual characteristics.

In Big Data applications involving multivariate time series analysis, VAR models in particular have seen widespread use in recent years as an effective empirical tool, as they can be treated with simple procedures like regression and method of moments. Various methods, such as LASSO [38] and Bayesian techniques, have been proposed for estimating high-dimensional VARs; recent examples are [8,11,20,22]. The Bayesian VAR methods avoid maximization over a high-dimensional parameter space by instead computing posterior estimates, and utilizing prior distributions that place prior mass on sparse configurations – thereby effecting a shrinkage towards zero (see the overview in [28]) together with model averaging. However, such approaches do not provide a single model that contains identified explanatory variables.

The first challenge we address is that the presence of many irrelevant variables and parameters in the VAR model can hinder the estimation of actual linkages and, related to this drawback, can negatively affect forecasting ability. Moreover, critical obstacles arise from a computational stand-point for very large dimensions with  $N > T$ . When the number of variables exceeds sample size, for instance, traditional Ordinary Least Squares and Yule-Walker estimation become infeasible [31]. Where possible, the LASSO approach attempts parameter reduction for large VAR systems like the ones permitted here by imposing sparsity restrictions. However, it can be problematic to find sparse structures that are both plausible and not over-constrictive. For example, Koop and Korobilis [29] employ a restricted parameter space, and similar devices are used in the LASSO literature [34]. Yet without reducing the massive number of parameters, one is left with a highly nonlinear and intractable likelihood surface, whose maximization is computationally impossible; cf. the discussion in [13]. Also, model averaging cannot achieve the objective of variable selection – one is left with a combination of models that each may feature diverse variable sets.

A second challenge is to allow for targeting specific variables of primary interest. Whereas most standard methods treat all series generically, it often occurs that only a limited set of variables represent focal points of the analysis, such as trends in aggregated measures as opposed to sectoral ones. For instance, in the analysis of macroeconomic data, time series like aggregate real GDP and employment are frequently considered as ‘targets’ to interpret and forecast; there is typically little interest in studying auxiliary variables such as output in small sectors in any detail, except for analysts or practitioners handling a specific industry. This kind of situation arises often in economic applications.

Hence, a second aim for our methodology is to focus on forecasting performance for the core series and to clearly distinguish between core and auxiliary variables. However, we wish to build a VAR model containing a subset of principal variables; we cannot employ a simple VAR-X model (i.e. regressing core variables on their own lags, as well as those of auxiliary variables), since the dynamics of the ancillary variables should also be considered. Although variable targeting can be achieved in the Bayesian [27] or LASSO [19]

approaches, for such methods the implementation would be difficult (the use of Dirac spikes, for example, would necessitate many Monte Carlo draws). In contrast, the dynamic factor analysis in [39] does not make any distinctions between primary variables and the numerous auxiliary indicators, even though the discussion of empirical results draws special attention to series (like GDP) versus other series viewed as being comparatively minor. Our approach achieves this second aim and, in doing so, the necessary computing time is greatly reduced.

As a third challenge, it is often desirable to work directly with the system variables rather than with indirect constructs (e.g. linear combinations of variables) because of the ease in interpreting (employing) explicit relationships. Widely-used approaches such as factor representations [17,39,40] and random projections [30] produce reductions in the parameter count by expressing the observation vector in terms of a smaller number of processes. However, our objective is a VAR model for the original variables, not a projection of them. Our method reduces dimension without transforming the variables: a substantial dimension reduction can be realized and, at the same time each potential explanatory or auxiliary variable is allowed to enter directly and transparently into the stochastic formulation of the targeted series.

Table 1 provides a summary of popular methods, classified according to our desired criteria. Whereas all the common methods (except the VAR-X) achieve some kind of dimension reduction, variable targeting (i.e. concentrating the model fitting so as to optimize the prediction performance of certain core variables that are of chief interest) is not achieved by Dynamic Factors or Random Projections. Although the VAR-X approach can indeed be used to target core variables, this comes at the cost of including many auxiliaries (whose dynamics are not modeled), and dimension reduction is obstructed unless one also introduces a LASSO or Bayesian methodology. As for variable selection, common implementations of VAR-LASSO do not utilize a parameter penalization structure that allows for the elimination of entire sets of variables (and this point is further discussed in Section 2.3), although some customization may be able to achieve this desideratum. Bayesian VAR methodology typically involves model averaging, so that there is no single model declaring whether variables are included or excluded; neither do Dynamic Factors or Random Projections provide variable selection, since both achieve dimension reduction by producing new functions of the input variables. The computational burden for VAR-LASSO and Bayesian VAR is moderate to high, whereas the other proposed methods have a low cost.

In contrast, the method of this article (denoted by WALD in Table 1) is able to achieve our three desiderata while keeping the computational cost low. However, we recognize

**Table 1.** Overview of popular available methods for dimension reduction, variable targeting, and variable selection for VAR models.

Method	Dimension Reduction	Variable Targeting	Variable Selection	Computation
VAR-LASSO	Yes	Yes	Yes <sup>a</sup>	Moderate to High
Bayesian VAR	Yes	Yes	No	High
VAR-X	No	Yes	Yes	Low
Dynamic Factors	Yes	No	No	Low
Random Projections	Yes	No	No	Low
WALD	Yes	Yes	Yes	Low

Note: <sup>a</sup>Off-the-shelf implementation puts zeros on parameters, and not entire variables; a customized modification can provide variable selection.

that there are other empirical contexts where analysts may wish instead to use previously established methods. If computation speed is not a consideration, and one has the time to design and tune a customized penalization structure, then the VAR-LASSO would be a good choice. Furthermore, if variable selection is of no concern, then the Bayesian VAR is very attractive due to the substantial amount of information provided through posterior distributions about parameter estimates; however, this method is expensive – so if there is an urgency to the analysis, then Dynamic Factors or Random Projections would be preferable. These latter two methods would also be attractive if variable targeting is unimportant to the analyst.

The rest of the paper is arranged as follows. Section 2 sets out the methodology in detail and gives the key statistical result with proof. Section 3 presents an application with real GDP and UR as targeted aggregate series, and Section 4 concludes. Supplementary material contains a background discussion on VAR methodology in econometrics (Appendix A), a description of software and variable selection methodology (Appendix B), and further empirical results (Appendix C).

## 2. Statistical methodology

This section sets forth the methodology and begins with a new theoretical result formulated as a proposition. The method then applies this result in an iterative framework. The basic idea is to start with a forecast error objective function and modify the function to focus on a small group of core variables. This is related to the procedure used in [5] who, however, use a different modeling strategy.

Our method can handle situations with up to hundreds of potential auxiliary variables by partitioning the forecast error variance matrix and focusing on the section corresponding to core variables. Candidate predictors that do not improve the relevant forecasts of variables are eliminated. To make this approach operational, a ‘marginal’ algorithmic strategy is used. Thus, starting with the smallest possible model, we successively add in supporting series that improve forecast performance for the core group, in the sense of rejecting the null hypothesis of Granger non-causality. At each stage, this side-steps computational hurdles involving large ill-conditioned matrices.

### 2.1. Mathematical formulation of the main result

Let  $y'_t = [x'_t, z'_t]$  be an  $N \times 1$  vector time series, where  $x_t$  is ‘low-dimensional’ and corresponds to the current model (including all the core variables, but possibly some auxiliary variables as well). New variables  $z_t$  are considered, consisting of additional auxiliary variables that are candidates for predicting  $x_t$ . Henceforth, we refer to the current model as the nested model, whereas the nesting model includes the additional auxiliaries.

Our approach to dimension reduction can be used for various kinds of multivariate models for  $\{y_t\}$ . Here, the fitting strategy is developed for the VAR class. Given a choice of order  $p$ , a positive integer giving the maximum lag considered, a VAR( $p$ ) process for  $\{y_t\}$  is expressed as

$$\Phi(L) y_t = \epsilon_t, \quad \epsilon_t \sim \text{WN}(0, \Sigma)$$

with  $L$  the lag operator defined via  $L y_t = y_{t-1}$ . The matrix AR polynomial in  $L$  is given by  $\Phi(L) = I_N - \sum_{j=1}^p \Phi^{(j)} L^j$ , where  $I_N$  is an  $N \times N$  identity matrix, and each  $\Phi^{(j)}$  denotes

a coefficient matrix with real-valued entries. It is assumed that  $\{y_t\}$  is stationary [31]; in practice, this requires appropriate differencing of trending series, such as real GDP. Our aim is to determine whether  $\{z_t\}$  should be included in the model. For each  $1 \leq j \leq p$ , partition  $\Phi^{(j)}$  conformably with  $y_t = [x'_t, z'_t]'$  so that

$$\Phi^{(j)} = \begin{bmatrix} \Phi_{xx}^{(j)} & \Phi_{xz}^{(j)} \\ \Phi_{zx}^{(j)} & \Phi_{zz}^{(j)} \end{bmatrix}.$$

We consider the null hypothesis that  $\{z_t\}$  does not Granger-cause  $\{x_t\}$ , i.e.

$$H_0 : \Phi_{xz}^{(j)} \equiv 0 \quad \forall 1 \leq j \leq p.$$

Rejections of this null will favor  $\Phi_{xz}^{(j)}$  having some nonzero entries. The main result below states that if auxiliary variables do not Granger-cause core variables, and we restrict our attention to forecast performance for  $\{x_t\}$ , then parameter estimates can be obtained via the Yule-Walker equations for core variables alone. That is, the sparsity restriction and the fitting criterion imply that, for the purposes of parameter estimation, the auxiliary variables can be omitted from the statistical analysis; this facilitates a block-recursive estimation.

**Proposition 2.1:** *Consider the VAR( $p$ ) model for  $\{y_t\}$  consisting of current  $\{x_t\}$  and auxiliary  $\{z_t\}$  variables. Suppose that  $\{z_t\}$  does not Granger-cause  $\{x_t\}$ , and suppose that we fit the model so as to minimize the determinant of the mean square forecast error variance matrix of the current variables. Then the parameter estimates are given by the solution to the Yule-Walker equations arising from  $\{x_t\}$  alone.*

**Proof:** From [32], we know that the log determinant of the mean square forecast error variance matrix equals the concentrated Whittle likelihood for a VAR( $p$ ) model. This takes the form

$$\begin{aligned} \Omega &= \Gamma(0) - [\Phi^{(1)}, \dots, \Phi^{(p)}] [\Gamma(1), \dots, \Gamma(p)]' - [\Gamma(1), \dots, \Gamma(p)] [\Phi^{(1)}, \dots, \Phi^{(p)}]' \\ &\quad + [\Phi^{(1)}, \dots, \Phi^{(p)}] \Gamma^{(p)} [\Phi^{(1)}, \dots, \Phi^{(p)}]', \end{aligned}$$

where  $\Gamma(h) = \text{Cov}(y_{t+h}, y_t)$  and  $\Gamma^{(p)}$  is a block matrix with  $jk$ th block  $\Gamma(k-j)$ . Let  $\Gamma(h)$  be partitioned conformably to  $y_t = [x'_t, z'_t]'$ , so that  $\Gamma_{xx}(h) = \text{Cov}(x_{t+h}, x_t)$ ,  $\Gamma_{zx}(h) = \text{Cov}(z_{t+h}, x_t)$ , etc. Under  $H_0$  we have  $\Phi_{xz}^{(j)} = 0$  for all  $1 \leq j \leq p$ , and hence the upper left block of  $\Omega$  is

$$\begin{aligned} \Omega_{xx} &= \Gamma_{xx}(0) - [\Phi_{xx}^{(1)}, 0, \dots, \Phi_{xx}^{(p)}, 0] [\Gamma_{xx}(1), \Gamma_{xz}(1), \dots, \Gamma_{xx}(p), \Gamma_{xz}(p)]' \\ &\quad - [\Gamma_{xx}(1), \Gamma_{xz}(1), \dots, \Gamma_{xx}(p), \Gamma_{xz}(p)] [\Phi_{xx}^{(1)}, 0, \dots, \Phi_{xx}^{(p)}, 0]' \\ &\quad + [\Phi_{xx}^{(1)}, 0, \dots, \Phi_{xx}^{(p)}, 0] \Gamma^{(p)} [\Phi_{xx}^{(1)}, 0, \dots, \Phi_{xx}^{(p)}, 0]' \\ &= \Gamma_{xx}(0) - \sum_{j=1}^p \Phi_{xx}^{(j)} \Gamma_{xx}(j) - \sum_{k=1}^p \Gamma_{xx}(-k) \Phi_{xx}^{(k)'} + \sum_{j,k=1}^p \Phi_{xx}^{(j)} \Gamma_{xx}(k-j) \Phi_{xx}^{(k)'}, \end{aligned}$$

which is the mean square forecast error variance matrix for the core variables alone. It was shown in [32] that the solution to the Yule-Walker equations (for  $\{x_t\}$ ) minimize each entry of  $\Omega_{xx}$ , and hence minimize  $\log \det \Omega_{xx}$ . ■



**Remark 2.1:** The full objective function  $\log \det \Omega$ , minimized under the constraint offered by  $H_0$ , has formulas given in [32] that require computations involving the autocovariances of  $\{z_t\}$ . These computations are problematic when  $T < N$ . Hence, adoption of the objective function  $\log \det \Omega_{xx}$  leads to a substantial simplification, essentially allowing us to proceed even when  $T < N$ .

As a consequence, in order to fit the nested model, we can just solve the Yule-Walker equations for  $\{x_t\}$ . For the nesting model, we can solve the Yule-Walker equations for  $\{x_t, z_t\}$  and focus our attention on the upper block, namely  $[\Phi_{xx}^{(j)} \Phi_{xz}^{(j)}]$ . By solving the Yule-Walker equations expressed in terms of the sample autocovariances (based on a sample  $x_1, \dots, x_T$ ) we obtain estimates for the VAR coefficients, which are expressed as  $\hat{\Phi}^{(j)}$ . The Wald statistic involves testing  $H_0$  by measuring the discrepancy of  $\text{vec } \hat{\Phi}_{xz}$  from zero. Under regularity conditions on the time series (see the discussion in Section 3.1 of [41]) there is a central limit theorem

$$\sqrt{T} \left( \text{vec} \left[ \hat{\Phi}^{(1)}, \dots, \hat{\Phi}^{(p)} \right] - \text{vec} \left[ \Phi^{(1)}, \dots, \Phi^{(p)} \right] \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma^{(p)-1} \otimes \Sigma).$$

Because  $\Phi_{xz}^{(j)} = E_1 \Phi^{(j)} E_2$  for appropriate block matrices  $E_1$  and  $E_2$ , we obtain  $\text{vec } \Phi_{xz}^{(j)} = (E_2' \otimes E_1) \text{vec } \Phi^{(j)}$ , and hence

$$\begin{aligned} \sqrt{T} \left( \text{vec} \left[ \hat{\Phi}_{xz}^{(1)}, \dots, \hat{\Phi}_{xz}^{(p)} \right] - \text{vec} \left[ \Phi_{xz}^{(1)}, \dots, \Phi_{xz}^{(p)} \right] \right) &\xrightarrow{\mathcal{L}} \\ \mathcal{N} \left( 0, [I_p \otimes E_2'] \Gamma^{(p)-1} [I_p \otimes E_2] \otimes E_1 \Sigma E_1' \right), \end{aligned}$$

where  $I_p$  is a  $p$ -dimensional identity matrix. This limiting variance matrix, denoted  $V$ , is easy to compute. It is estimated by  $\hat{V}$ , given by inserting the sample estimates for  $\Gamma^{(p)}$  and  $\Sigma$  – the latter is immediately obtained after fitting the nesting model. The Wald statistic is then defined by

$$W = T \text{vec} \left[ \hat{\Phi}_{xz}^{(1)}, \dots, \hat{\Phi}_{xz}^{(p)} \right]' \hat{V}^{-1} \text{vec} \left[ \hat{\Phi}_{xz}^{(1)}, \dots, \hat{\Phi}_{xz}^{(p)} \right],$$

which under  $H_0$  has a  $\chi_r^2$  distribution, where  $r$  is the length of  $\text{vec}[\Phi_{xz}^{(1)}, \dots, \Phi_{xz}^{(p)}]$ . A significant value of  $W$  indicates rejection of Granger non-causality of the auxiliary variables, and favors the nesting model over the nested model.

We note in passing that the conceptual framework, which involves minimizing the average squared forecast errors for the core series – localizing the fitting criterion for this subset of interest – can be used more broadly for other types of multivariate models. For example, a general multivariate criterion is minimization of the determinant of  $\sum_t (y_t - \hat{y}_t)(y_t - \hat{y}_t)'$ , where  $\hat{y}_t$  is a one-step ahead forecast, and depends on the model parameters. We can alter this criterion to  $\sum_t (x_t - \hat{x}_t)(x_t - \hat{x}_t)'$ , where  $\hat{x}_t$  is a forecast of the core series generated by past observed  $y_1, \dots, y_{t-1}$ . We can then test whether there is a benefit to using  $z_1, \dots, z_{t-1}$  in these forecasts, by comparing criterion values to  $\sum_t (x_t - \tilde{x}_t)(x_t - \tilde{x}_t)'$ , where  $\tilde{x}_t$  is a forecast that only uses  $x_1, \dots, x_{t-1}$ .

In the case of a VAR model, such tests reduce to a test of whether a specific block of parameters is zero, whereas in the general case there may be no such association to specific sets of zero parameter values. It is possible to formulate the problem in the frequency



domain, as a decomposition of the spectral density of  $\{y_t\}$ , but such a discussion is far outside the scope of this article.

## 2.2. Procedure for variable selection

We propose a variable selection procedure, starting with  $\{x_t\}$  given by the core variables, and in each step with  $\{z_t\}$  corresponding to a single auxiliary variable, ordered by their predictive content for the core set. Hence, as a preliminary step we apply Proposition 2.1 to  $\{y_t\}$  consisting of  $N_C$  core variables and a single auxiliary, and record the Wald statistic's  $p$ -value. Repeating this procedure for all  $N_A$  auxiliaries, we obtain  $N_A$   $p$ -values, one for each model of dimension  $N_C + 1$ . Then as a second step, having sorted the auxiliaries according to their  $p$ -values, we iteratively apply Proposition 2.1, starting with the auxiliary that has the lowest  $p$ -value (i.e. the most predictive content). This adds variables until there is no further benefit. The third step refines these variables, removing those that do not help to predict the core variables.

Let us consider the second step in more detail. In the  $i$ th stage (with  $1 \leq i \leq N_A$ ) we consider  $\{x_t\}$  as the model of the previous stage (or just the set of core variables, when  $i = 1$ ), and let  $\{z_t\}$  be the  $i$ th best (assessed according to lower  $p$ -values from the first step) auxiliary for predicting the core variables. We then apply the Wald test again, and add the auxiliary to the model if we reject  $H_0$  (at some specified Type I error rate  $\alpha$ ), then proceeding to stage  $i + 1$ . (Note that significance values ignore the randomness inherent in model selection that has already occurred – this is a conscious choice based on considerations of computational speed, which preclude VAR-LASSO or Bayesian VAR.) However, if we fail to reject  $H_0$  then there is not enough predictive content, and we move to the next stage. In this manner our initial subset VAR model is obtained.

For the third step we further refine this initial model, which may contain some spurious variables (but hopefully has not excluded any pertinent variables). We proceed to test whether each auxiliary variable Granger causes the core variables. That is, if we have  $N_C$  core variables and  $N_A$  auxiliary variables at the end of the second step, then for  $1 \leq i \leq N_A$  we test whether  $\Phi_{xz}^{(j)} = 0$  for all  $1 \leq j \leq p$  (where  $x$  corresponds to core variables, and  $z$  is the  $i$ th auxiliary) in the model consisting of all  $N_C + N_A$  variables. Any auxiliaries failing to reject the null hypotheses are discarded. By this means we arrive at our final model.

When applying this forward selection procedure to the entire set of core variables, auxiliary variables are included according to overall forecast performance assessed via the log determinant of the forecast error variance matrix. Through experience running the algorithm, we found that this can give adverse outcomes whereby only variables are selected that minimize the forecast errors for one particular core variable that dominates others in terms of scale. A remedy is to apply the variable selection procedure to each core variable individually, obtaining distinct sets of explanatory variables. Then to arrive at a single VAR model including all core and auxiliaries, we take the union of the various sets of explanatory variables. This has the advantage of discerning the particular supporting variables that are useful for each core variable (which may be of interest in subsequent analysis), while still permitting a complete model that is likely to be effective in jointly predicting the core variables.

To further support model parsimony and forecasting ability, we can fit the final VAR with zeros imposed by constrained Yule-Walker, as described in [32], placing zeros in the

coefficient matrices for auxiliary variables that do not Granger cause a particular core variable. The identification of zero-placements can proceed via sorted t-statistics, as in [7], though other options are discussed in [35,36]. We sort the t-statistics computed from the unrestricted final VAR; beginning with the smallest absolute t-statistic, we proceed to insert zeros one by one, checking whether the Whittle likelihood changes significantly – as compared to the original, unrestricted final VAR model – when the model is refitted with restrictions. Because the likelihood is not significantly changed at each step, we expect the impact on forecast performance to be slight. We emphasize that imposing these zero constraints alters all the remaining coefficients estimates.

In this way, the final best set of predictor auxiliaries is determined via simple Wald tests performed in an algorithmic framework. This shows how the large-scale culling preserves the forecast performance of the model as the dimension and number of variables are reduced. We remark that auxiliary variables  $z$  can be correlated with one another, as they are assumed to have a VAR structure. A particular variable is included in the current model whenever  $H_0$  is rejected, where the null corresponds to the absence of Granger causality. So a variable must be able to add predictive content to the core variables that has not already been fully realized by other variables already in the model. If a new variable is highly correlated with previously included variables, then it is less likely to improve predictions, and may be excluded due to redundancy. More details are provided in Appendix B of the Supplementary material, which also discusses the relationship between the Type I error rate and the Family-Wise Error Rate of the sequential procedure.

### 2.3. Simulation studies

We evaluate the proposed method in comparison to a custom implementation of VAR-LASSO, which we first describe. A particular variable of index  $i$  ( $1 \leq i \leq N$ ) is included in the VAR model if and only if it Granger causes the other variables in the model; hence it should be excluded if column  $i$  of  $\Phi^{(j)}$  (denoted  $\Phi_{:,i}^{(j)}$ ) is zero for all  $1 \leq j \leq p$ . So we can associate a LASSO penalty  $\lambda_i$  to the corresponding parameters via the quantity  $\sum_{j=1}^p \|\Phi_{:,i}^{(j)}\|_1$ , where  $\|\cdot\|_1$  denotes the vector  $L_1$ -norm. Hence, we obtain a VAR-LASSO objective function by adding the penalty

$$\sum_{i=1}^N \lambda_i \sum_{j=1}^p \|\Phi_{:,i}^{(j)}\|_1$$

to the Whittle likelihood. This is a straight adaptation of LASSO variable selection from linear models to the case of multiple regressions occurring in VAR modeling. Other types of group LASSO are described and implemented in packages such as BigVAR [34], but these do not allow for a group penalty structure of the type that our problem requires. For this reason we have constructed our own implementation (see Appendix B). From some experimentation, it is important to initialize the VAR-LASSO properly, and we use the Yule-Walker estimates as starting values whenever this is computationally feasible ( $T > N$ ). We have experimented with various values of the LASSO penalty; in our simulations we consider three possible values.

We assess the methods through computational speed as well as Family-Wise Error Rate (FWER), False Discovery Rate (FDR), and True Discovery Rate (TDR). These are defined

[4] as follows. Let  $H_0$  denote the hypothesis that a particular variable is not in the model, whereas  $H_a$  says that it is in the model. The null hypothesis is rejected by a large Wald statistic, thereby indicating that the variable should be included. If we have a total of  $N$  variables being tested, let  $\mathcal{U}$  count the number of variables not in the model that are excluded (a correct decision),  $\mathcal{V}$  is the count of variables not in the model that are included (a Type I error),  $\mathcal{T}$  is the number of variables in the model that are excluded (a Type II error), and  $\mathcal{S}$  is the number of variables in the model that are included (a correct decision). So  $\mathcal{U} + \mathcal{V}$  is the number of ancillary variables (i.e. those not in the model), while  $\mathcal{T} + \mathcal{S}$  is the number of core plus true auxiliary variables. Also  $\mathcal{U} + \mathcal{T}$  is the number of variables we exclude, whereas  $\mathcal{V} + \mathcal{S}$  is the number we include. Then (so long as  $\mathcal{V} + \mathcal{S} > 0$  and  $\mathcal{U} + \mathcal{T} > 0$ )

$$\text{FWER} = \mathbb{P}[\mathcal{V} \geq 1] \quad \text{FDR} = \mathbb{E}[\mathcal{V}/(\mathcal{V} + \mathcal{S})] \quad \text{TDR} = \mathbb{E}[\mathcal{U}/(\mathcal{U} + \mathcal{T})].$$

Whereas FWER and FDR are measures of Type I error rate (i.e. it assesses whether we are including spurious variables), TDR is a measure of power (i.e. assessing whether we are excluding spurious variables). In a simulation setting these three measures are estimated over the Monte Carlo repetitions simply by tallying and averaging results.

Our first study is a VAR(1) with  $N = 10$ : there is a single core series (which is known), along with 3 auxiliaries and 6 ancillary variables (not in the model). We randomly generate a stable matrix  $\Phi^{(1)}$  with upper right  $4 \times 6$  block equal to zero, obtaining

$$\begin{bmatrix} 0.4797 & -0.5983 & -1.9193 & 0.9214 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ -0.1885 & -0.0417 & 0.0481 & 0.3317 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.2791 & -1.0336 & -0.5497 & 0.5732 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ -0.1392 & -0.3210 & 0.4331 & -0.1553 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 1.6387 & 2.3115 & 1.4377 & 1.5232 & 0.2771 & -0.3079 & 0.7505 & -0.9826 & -0.0552 & 0.9051 \\ 0.6213 & 0.9723 & 0.4270 & -1.2360 & 0.3354 & -0.2831 & -0.1129 & 0.2341 & 0.7441 & -0.6435 \\ 0.2027 & 0.9646 & -1.7445 & 1.6659 & -1.1264 & -0.3722 & -0.3237 & -0.4143 & -0.4983 & -0.1344 \\ 1.1089 & -0.5441 & -0.0253 & -0.5036 & -0.6673 & -0.0860 & 0.0951 & -0.2642 & -0.7872 & 0.7475 \\ -0.2062 & 0.6712 & -1.4883 & -0.5785 & -0.2112 & -1.6213 & -0.0009 & -0.8632 & 0.7421 & -0.0696 \\ -0.3790 & 0.5008 & -0.5416 & 0.2718 & 1.2335 & 1.0669 & 0.1703 & 0.4655 & -1.1292 & 0.7296 \end{bmatrix}.$$

The innovation covariance matrix is set to the identity. We generate 100 Gaussian samples of size  $T = 50, 200, 800$  using the stationary initialization of the VAR(1). Our proposed method (denoted WALD) with threshold values  $\alpha = .05, .10, .20$  is compared to VAR-LASSO with penalties  $\lambda_i = 1, 10, 100$  (the same penalty is used for all variables). The results of the first study are summarized in Table 2. The VAR-LASSO method puts all 10 variables into the model, no matter the penalty settings, indicating that this method will not work well for variable selection without additional tuning. The timing is not exorbitant, and yet far slower than the WALD method. As expected, for the WALD method the FWER and FDR increases with  $\alpha$ , and is also fairly steady with respect to sample size (except with  $\alpha = .05$ ). Power approaches perfect selection even with moderate sample size; by lowering  $\alpha$  to .05 we can still achieve TDR equal to one (i.e. all true variables are included) while keeping FWER and FDR fairly low (this is a less egregious error, to sometimes include spurious variables).

Our second study is a VAR(1) with  $N = 100$  variables: a single core series, 9 auxiliaries, and 90 spurious ancillary variables. The randomly generated coefficient matrix is stable, and has a  $10 \times 90$  upper right block of zeros (it is too large to display). We generate 100 samples in the same manner, but use sample sizes  $T = 40, 60, 80$  to test the high-dimensional case where  $T < N$ . These choices preclude using the Yule-Walker as an initialization to the

**Table 2.** Simulation results for 10-dimensional VAR(1) process. Sample sizes are  $T = 50, 200, 800$ . LASSO and WALD methods are compared, assessed by Time (in seconds), FWER, FDR, and TDR.

$T = 50$	Time	FWER	FDR	TDR
LASSO ( $\lambda = 1$ )	2.10	1.00	0.600	1.000
LASSO ( $\lambda = 10$ )	2.08	1.00	0.600	1.000
LASSO ( $\lambda = 100$ )	2.07	1.00	0.600	1.000
WALD ( $\alpha = .05$ )	0.0298	0.16	0.042	0.979
WALD ( $\alpha = .10$ )	0.0301	0.27	0.064	0.987
WALD ( $\alpha = .20$ )	0.0308	0.56	0.141	0.994
$T = 200$	Time	FWER	FDR	TDR
LASSO ( $\lambda = 1$ )	2.06	1.00	0.600	1.000
LASSO ( $\lambda = 10$ )	2.29	1.00	0.600	1.000
LASSO ( $\lambda = 100$ )	2.34	1.00	0.600	1.000
WALD ( $\alpha = .05$ )	0.0347	0.15	0.035	1.000
WALD ( $\alpha = .10$ )	0.0340	0.28	0.062	1.000
WALD ( $\alpha = .20$ )	0.0337	0.49	0.120	1.000
$T = 800$	Time	FWER	FDR	TDR
LASSO ( $\lambda = 1$ )	2.24	1.00	0.600	1.000
LASSO ( $\lambda = 10$ )	2.11	1.00	0.600	1.000
LASSO ( $\lambda = 100$ )	2.13	1.00	0.600	1.000
WALD ( $\alpha = .05$ )	0.0504	0.06	0.015	1.000
WALD ( $\alpha = .10$ )	0.0552	0.24	0.068	1.000
WALD ( $\alpha = .20$ )	0.0523	0.54	0.150	1.000

**Table 3.** Simulation results for 100-dimensional VAR(1) process. Sample sizes are  $T = 40, 60, 80$ . The WALD method is assessed by Time (in seconds), FWER, FDR, and TDR.

$T = 40$	Time	FWER	FDR	TDR
WALD ( $\alpha = .05$ )	0.3005	0.5500	0.1754	0.9330
WALD ( $\alpha = .10$ )	0.3191	0.8200	0.2844	0.9387
WALD ( $\alpha = .20$ )	0.3114	0.9600	0.4321	0.9475
$T = 60$	Time	FWER	FDR	TDR
WALD ( $\alpha = .05$ )	0.3030	0.5600	0.1356	0.9450
WALD ( $\alpha = .10$ )	0.3089	0.7100	0.2099	0.9484
WALD ( $\alpha = .20$ )	0.3275	0.8800	0.3593	0.9601
$T = 80$	Time	FWER	FDR	TDR
WALD ( $\alpha = .05$ )	0.3130	0.5200	0.1195	0.9491
WALD ( $\alpha = .10$ )	0.3339	0.7000	0.1709	0.9569
WALD ( $\alpha = .20$ )	0.3394	0.9400	0.3237	0.9632

VAR-LASSO, and hence the procedure was initialized at  $\Phi^{(1)} = 0$ . It happened that the VAR-LASSO procedure for this data process required several minutes of run-time for each simulation, and still exhibited poor performance; we have omitted its presentation in our summary in Table 3. We find that for our method the computation time is still quite feasible, and TDR is quite good even in the case of  $T = 40$ . The FWER is rather high even for  $\alpha = .05$ , but FDR does drop as sample size increases from  $T = 40$  to  $T = 60$ , and is probably acceptable for  $\alpha = .05$ .

We also comment that which variables are included in our model depends on the size of VAR coefficients; however, even variables with a coefficient of zero (for predicting a core variable) can have a non-zero coefficient when we restrict to a sub-model. This is an interesting phenomenon: a variable can only be truly identified as spurious once certain other variables have been introduced. For this reason, the refining step corresponding to backwards deletion is especially important.

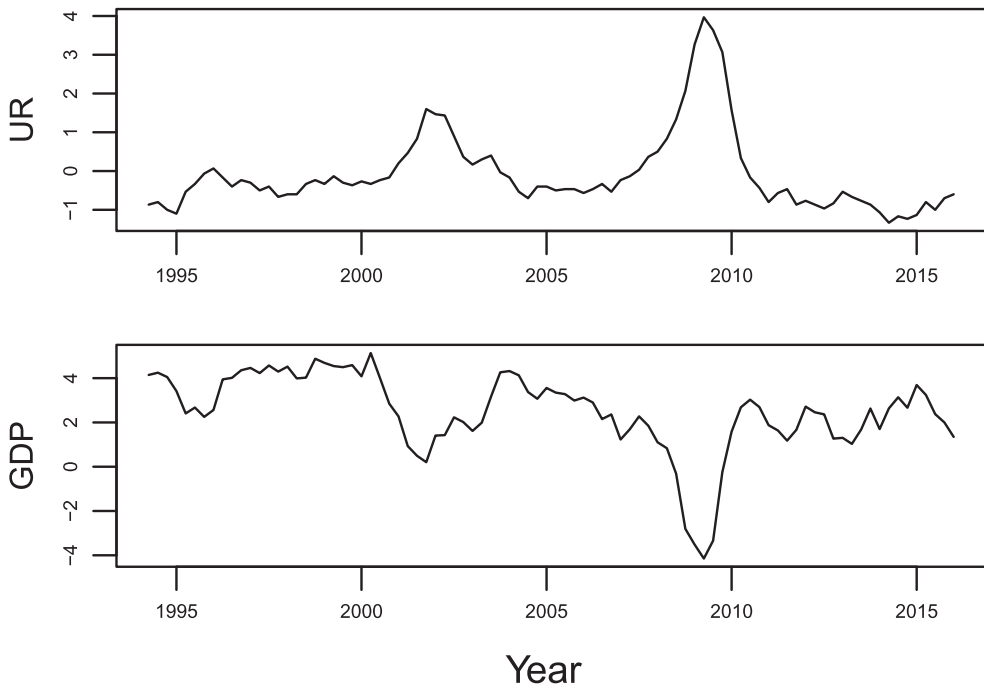
### 3. Application to QWI predictors

Real GDP growth and UR represent two of the most important measures of economic health and performance; they are widely monitored by policy-makers, government officials and economists. In our application the choice of variable types – here relating to job flows – to include is guided by knowledge about economic linkages. Note that this contrasts with a ‘throw-it-all-in’ approach where the researcher indiscriminately includes as much information as possible, with little consideration of which kinds of supporting variables are likely to have a substantive connection with the targets.

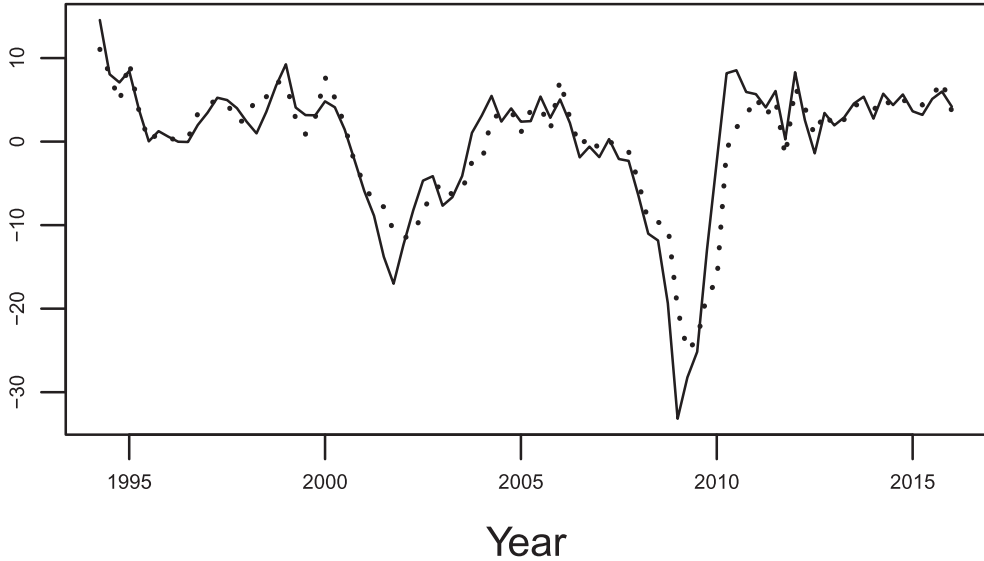
Here, we focus on the workforce indicators for various industry groups; this gives a large but tractable dataset where a statistical treatment is required. Industry-level data are instrumental in capturing the effects of labor market re-allocation, a hypothesized economic mechanism, and such data involve highly detailed and idiosyncratic elements in an expansive set of candidate predictors. It would not be practical or feasible to determine the relative importance of each such variable beforehand. (Even if that were possible, the assessment of dynamic lead/lag relationships across sectors and the quantitative measurement of effects requires a statistical approach for the time series dataset.)

In particular, data on quarterly real GDP is taken from the Bureau of Economic Analysis (BEA) and is converted to an annualized rate of growth (i.e. annual, or year-on-year, differences of log-transformed GDP), as shown in Figure 1. The quarterly Unemployment Rate (UR) is taken from the Bureau of Labor Statistics, and is converted to annual or year-on-year differences. The auxiliary variables are 114 time series of Quarterly Workforce Indicators (QWI), drawn from the LEHD database of the U.S. Census Bureau. The QWI time series used here represent six kinds of indicators recorded for 19 major industry groups, as detailed in Appendix C. Each of these 114 series is converted to an annual rate, in the same way as GDP, i.e. we take 100 times the annual difference of the logged data. All time series are considered for the sample period 1993.Q1 through 2017.Q2 to match the available history for the supporting variables. Considering the annualized or year-on-year changes allows us to focus directly on the dynamics and direction of the movements in these core variables. From a statistical perspective many of the series required differencing to yield stationary time series amenable to a stable VAR framework.

The LEHD data represent labor market information based on certain administrative sources (such as Unemployment Insurance data and the Quarterly Census of Employment and Wages) outside the scope of more traditional indicators, and these industry-level measures are able to capture concepts like employment reallocation; see [1] for details. While two of the indicator types, Earnings and Employment, are ‘classic’ measures, the remaining four types – Hires, Separations (abbreviated as ‘Seps’), Job Generations or Creations (‘JC’), and Job Destructions (‘JD’) – represent less studied labor flows measures at the individual (Hires and Seps) and firm (JC and JD) level.

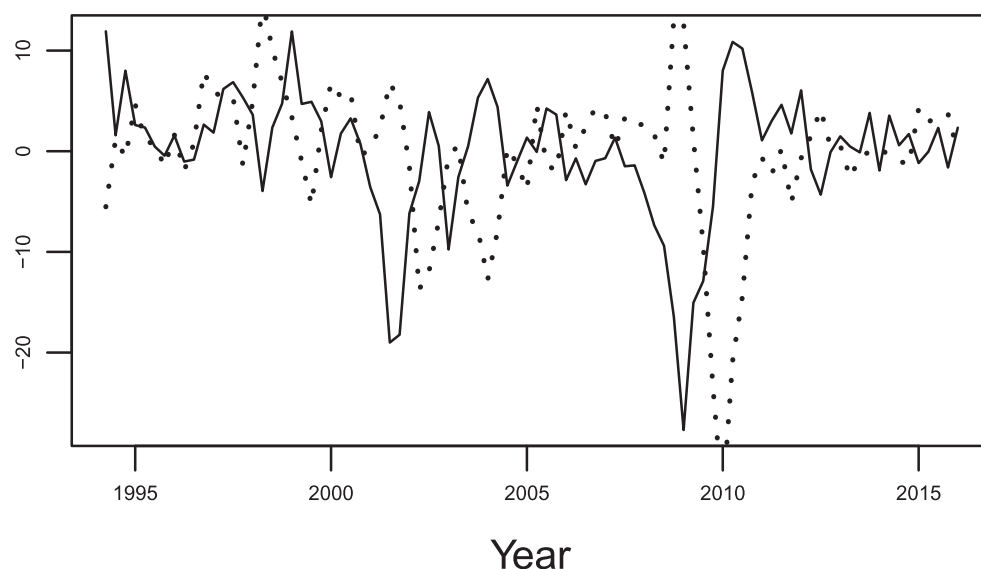


**Figure 1.** UR (top) and GDP (bottom) as annualized growth rates.



**Figure 2.** Aggregate Hires (solid) and Seps (dotted) as annualized growth rates.

Figure 2 shows aggregate (over 19 industry groups) Hires and Seps over the full sample period. Both measures dip around the recession of 2001–2002 and plunge during the Great Recession. Recoveries to modest growth rates are evident during economic expansions;



**Figure 3.** Aggregate JC (solid) and JD (dotted) as annualized growth rates.

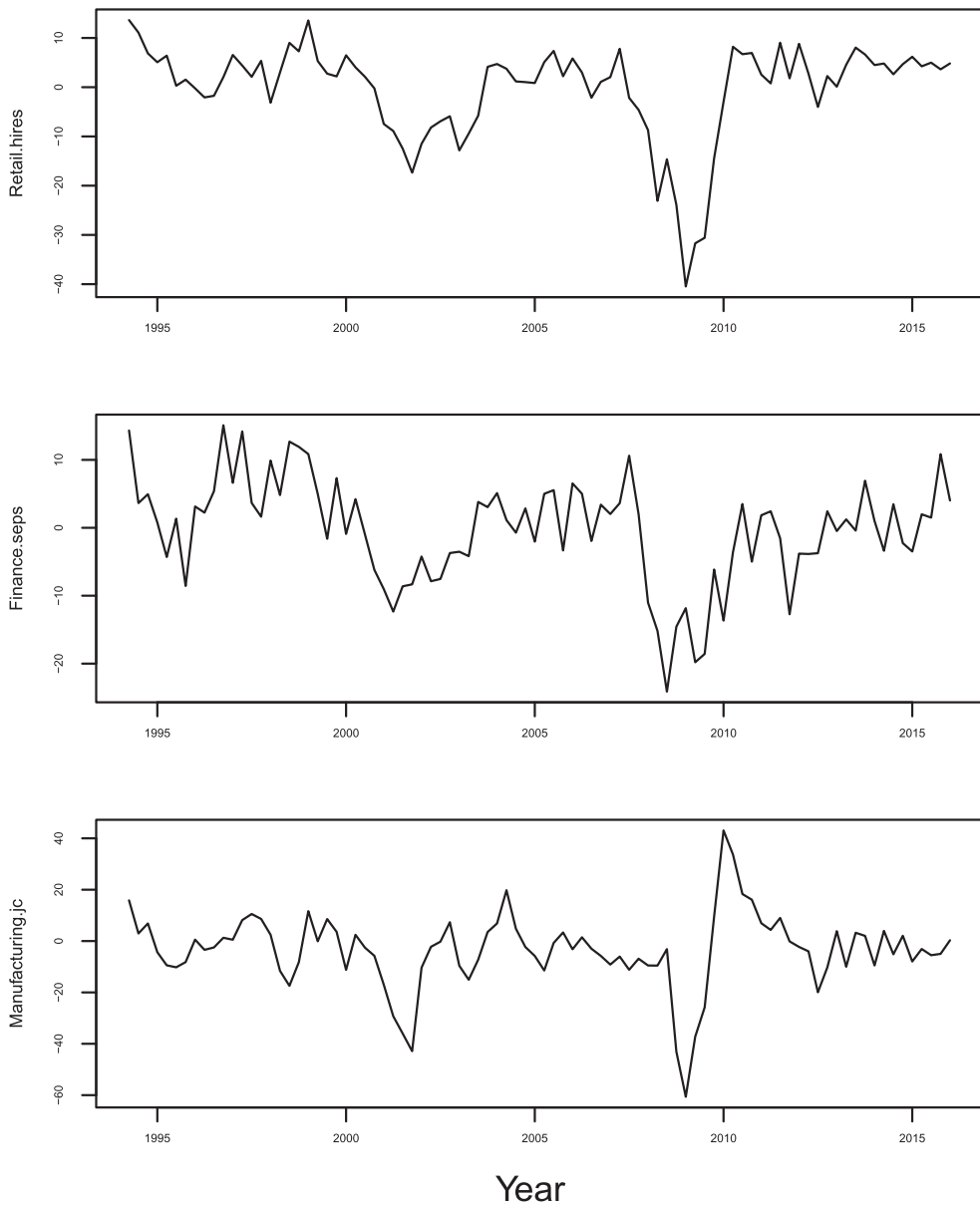
the recessions are marked by a more noteworthy negativity in Seps than in Hires, which coincides with the reduction in GDP and UR evident in Figure 1.

Figure 3 displays aggregate JC versus JD in log-annual differences; the plot of firm-side quantities appears rather different from the pattern in Figure 2. JC tends to increase during expansion and decrease modestly during recessions, whereas JD increases during recessions; it is clear that during the Great Recession, JD has the primary influence on employment changes through its precipitous decline. Though the raw series (not displayed) could perhaps be regarded as stationary, the transformation of log-annual differencing is still needed to remove seasonality, and it is also useful to have comparable data inputs for the machine-learning algorithm.

For both GDP and UR in turn, we take the first fifteen years (1993Q1 to 2007Q4) as our estimation or ‘training’ sample, and fit a VAR(4) model (see Appendix C of the supplement for more details on the model selection) using the threshold  $\alpha = .001$ , resulting in the selection of Retail.hires, Finance.seps, and Manufacturing.jc as auxiliary indicators (see Figure 4). This value of  $\alpha$  is very stringent in its inclusion of auxiliary variables; by increasing  $\alpha$ , much larger models can be entertained. The size of the model strikes a good balance between allowing for too many supporting series (hence hindering the implementation of the chosen VAR model and diluting the effects of the major auxiliary variables) and too few (thus not allowing the forecast errors enough room for improvement). We also examined the VAR-LASSO approach discussed in Section 2.3, but for various values of the tuning parameter the method selects either all or none of the auxiliaries (just as in the simulations). Henceforth we focus on the new methodology.

Model adequacy diagnostics gave satisfactory results (see Appendix C of the supplement for details on diagnostics). The coefficient t-statistics are given in Tables 4 and 5; recall that the constrained estimation discussed in Section 2.2 enforces a zero wherever the





**Figure 4.** Auxiliary QWI series, as annualized growth rate: Retail.hires, Finance.seps, and Manufacturing.jc.

t-statistic is small (and re-estimating all the coefficients subject to this constraint) so long as the Whittle likelihood does not significantly change. Note that some remaining t-statistics appear to be insignificant, but the standard critical values do not apply due to interdependence – replacing these by zeros was found to yield a significantly worse likelihood, and hence they should be left as is. (Appendix C of the supplement contains additional results, including the rest of the VAR coefficient matrices based upon the constrained Yule-Walker

**Table 4.** VAR coefficient t-statistics for variables impacting GDP, by lag. Entries marked 0 correspond to a zero imposed by constrained Yule-Walker estimation.

Lag	1	2	3	4
UR	0.094	0	0	0
GDP	8.052	−1.173	−1.267	0
Retail.hires	1.318	0	−3.426	1.751
Finance.seps	0	2.289	1.354	0
Manufacturing.jc	−1.100	0	1.254	−1.209

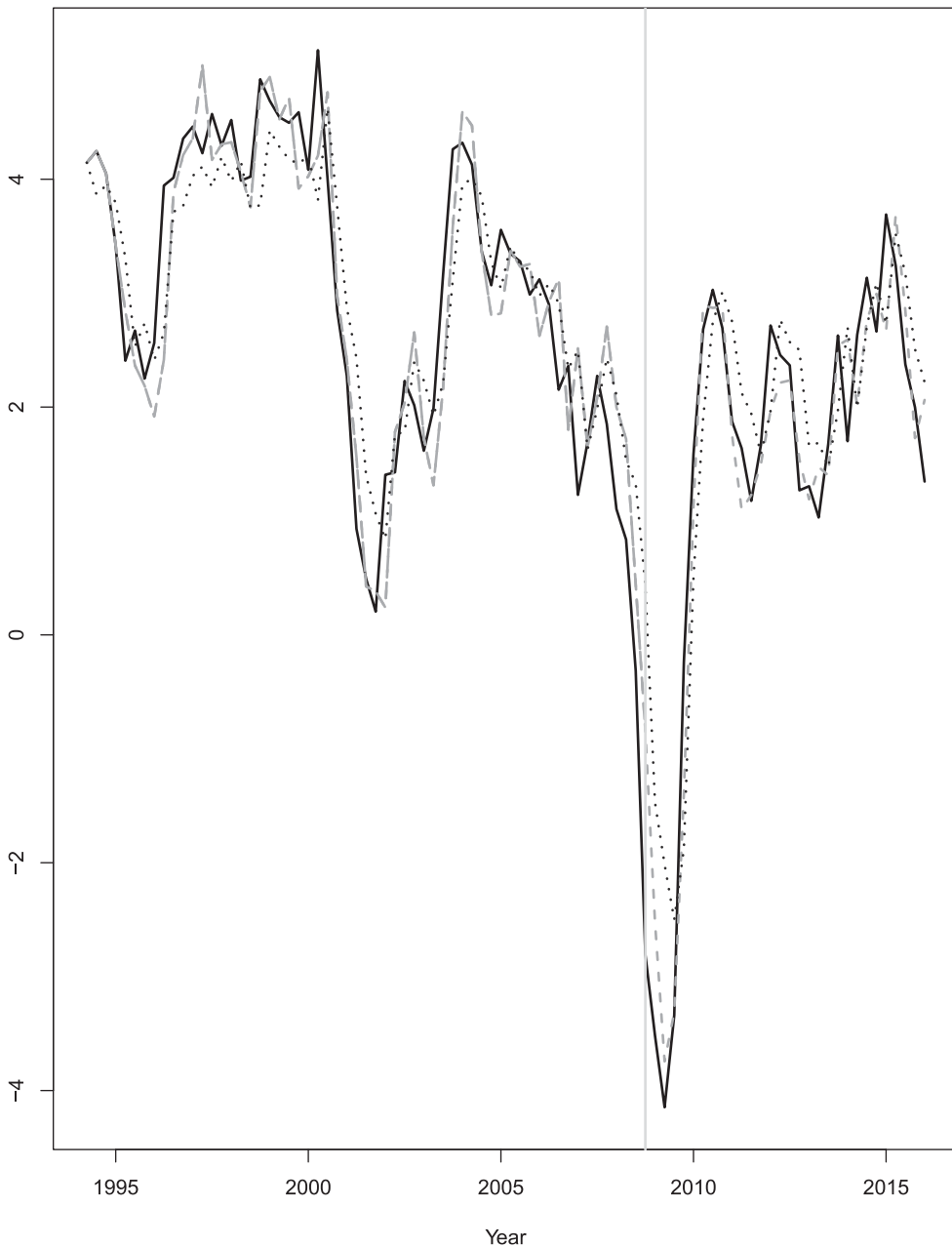
**Table 5.** VAR coefficient t-statistics for variables impacting UR, by lag. Entries marked 0 correspond to a zero imposed by constrained Yule-Walker estimation.

Lag	1	2	3	4
UR	5.580	0	0	−0.067
GDP	−2.565	0	2.852	0
Retail.hires	0	0	0	0
Finance.seps	0	−1.381	0	0
Manufacturing.jc	−1.721	0	0	0

fits, and the corresponding t-statistics, as well as unconstrained estimation results. Also, the structural VAR for the final five series is determined, and used to generate an impulse response plot.)

In order to assess our results, we compare out-of-sample VAR(4) forecasts for each core series to a best univariate AR competitor, which is found by fitting an AR( $r$ ) model to each core series, where  $r$  is chosen according to the Akaike Information Criterion (AIC). If the VAR( $p$ ) model is correctly specified and the auxiliary variables indeed Granger cause the core variable, then the innovation variance of the AR( $r$ ) should be greater than the corresponding entry of the error covariance matrix of  $\epsilon_t$ , whenever  $r \leq p$ . Plots of the forecasts are given for GDP (Figure 5) and UR (Figure 6). It is clear that the in-sample (grey long-dash) and out-of-sample (grey dashed) forecasts from the VAR(4) more effectively track the movements of each respective core series over the out-of-sample period, as compared to the dotted grey line (both in-sample and out-of-sample) that corresponds to AR( $r$ ) forecasts.

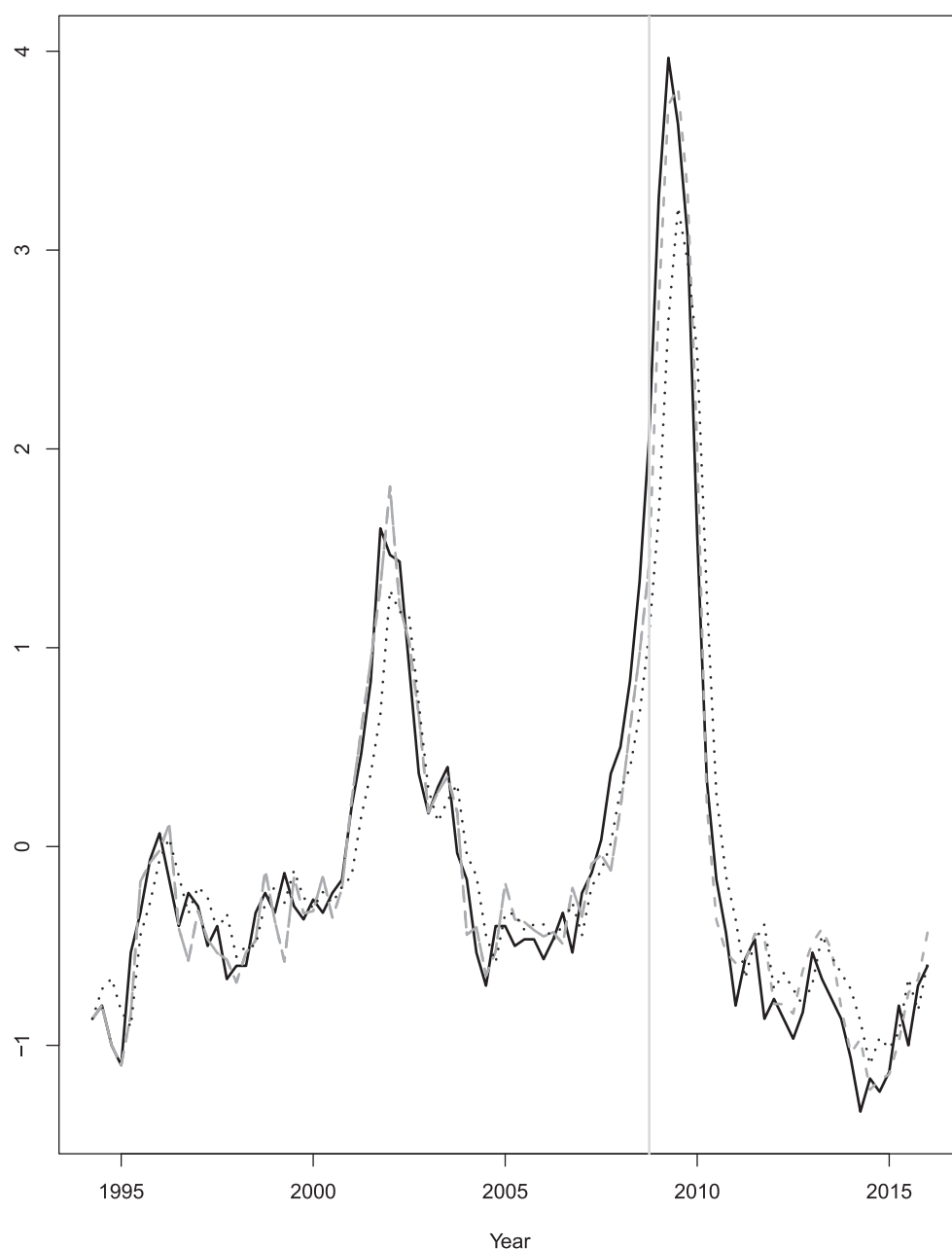
We empirically compare the forecast error for both methods by taking a running sum of squared forecast errors (omitting the first five years, so that a sufficient number of forecasts are present in the cumulation) over the in-sample and out-of-sample period, and plot the cumulative ratio of VAR( $p$ ) to AR( $r$ ), for both GDP (Figure 7) and UR (Figure 8). If this ratio is less than one, then the VAR( $p$ ) model is favored over the univariate model; we observe that the VAR model appears to have superior performance, for both core series, over both the in-sample and out-of-sample periods. In order to determine whether the superiority of the VAR results is likely due to sampling variability, we also compute Diebold-Marriano (DM) test statistics (with a Bartlett taper and bandwidth equal to half the sample size) for these forecast comparisons; see [15]. These t-statistics – which are just computed over the out-of-sample period – were 1.847 ( $p$ -value .038) and 1.395 ( $p$ -value



**Figure 5.** In-sample (grey long-dash) and out-of-sample (grey dashed) forecasts of GDP (annualized growth rate) from the fitted VAR(4) model, with univariate AR forecasts (dark grey, dotted). Solid grey line denotes the boundary between the in-sample and out-of-sample period.

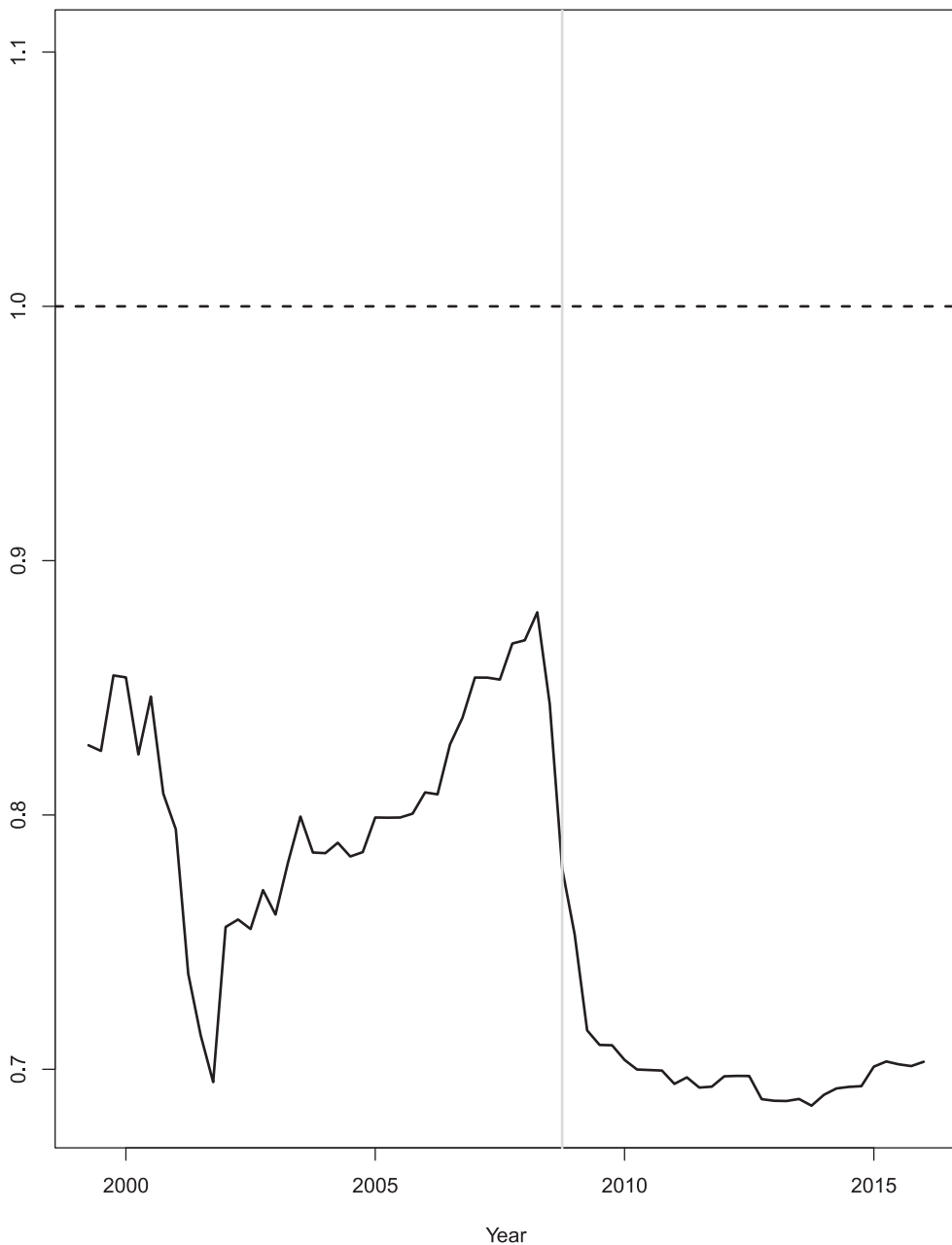
.087) respectively for GDP and UR, indicating a significant (at 10% level) superiority of VAR over AR in out-of-sample forecasting.

We have experimented with alterations, such as different thresholds, and have observed that increasing the number of selected variables can potentially increase or decrease



**Figure 6.** In-sample (grey long-dash) and out-of-sample (grey dashed) forecasts of UR from the fitted VAR(4) model, with univariate AR forecasts (dark grey, dotted). Solid grey line denotes the boundary between the in-sample and out-of-sample period.

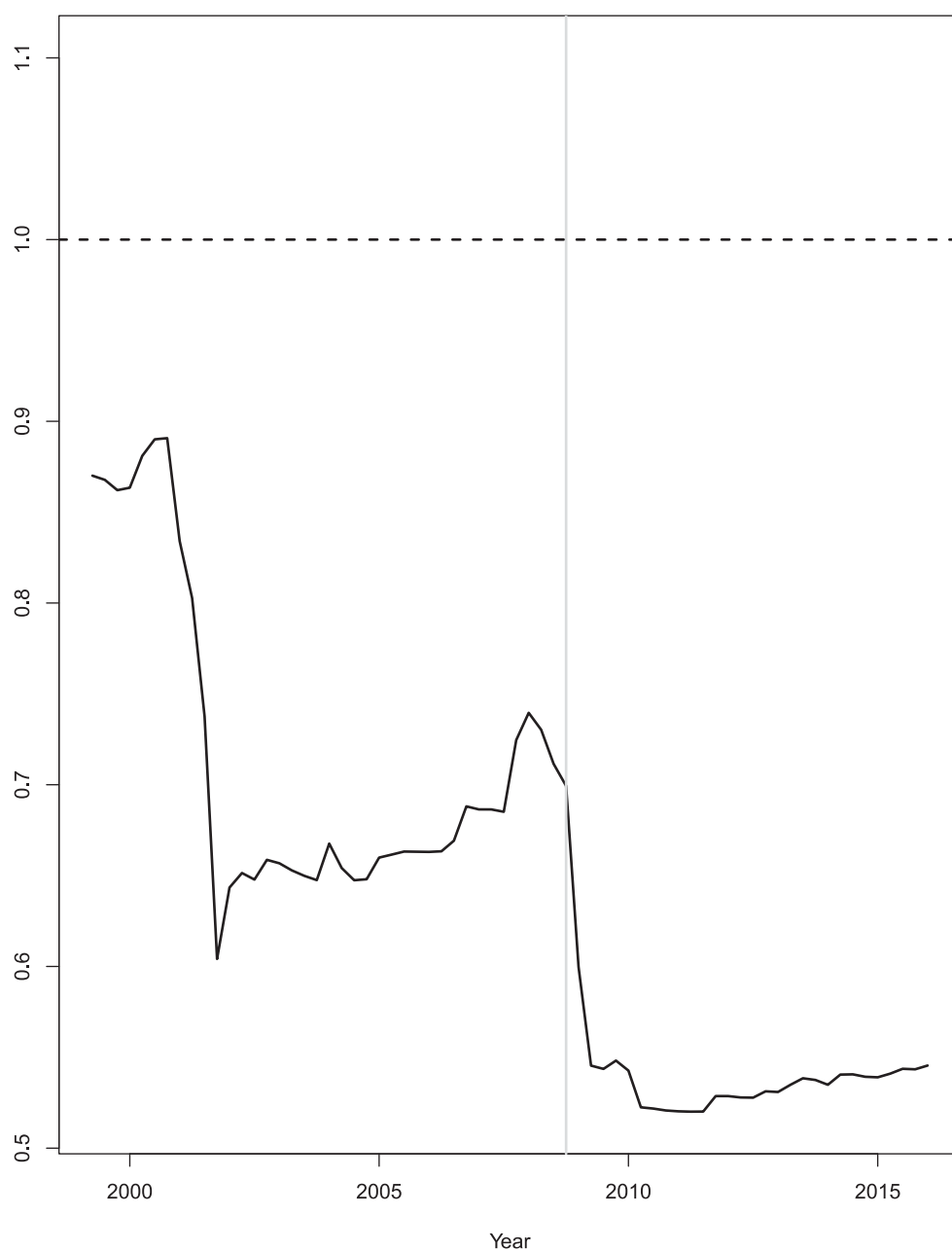
performance, as assessed via the DM test. Altering sparsity in the coefficient estimation can impact the quality of fit, as assessed through residual serial correlation, and enforcing too many zeros decreases forecast performance. Also, altering the sample span can lead to the selection of completely different auxiliary variables (cf. Figure 2 of [33]), but the resulting VAR still has superior out-of-sample performance.



**Figure 7.** Ratio of cumulative in-sample and out-of-sample squared forecast errors of GDP (annualized growth rate), for fitted VAR(4) model versus fitted univariate AR model. Solid grey line denotes the boundary between the in-sample and out-of-sample period.

#### 4. Conclusions

This paper introduces a pragmatic, methodologically sound set of tools for analysis of sizable VARs, when the practitioner wants to determine which of a large collection of auxiliary



**Figure 8.** Ratio of cumulative in-sample and out-of-sample squared forecast errors of UR, for fitted VAR(4) model versus fitted univariate AR model. Solid grey line denotes the boundary between the in-sample and out-of-sample period.

variables are most useful for forecasting a small set of core variables of interest. The efficacy of this strategy is borne out by our application to forecasting annual growth rates of GDP and changes in unemployment with various Workforce Indicator series. In this case study, the application of the method is very fast from a computational standpoint; for example,

the auxiliary variables are identified at each stage and the rolling forecasts are calculated, performed over all stages within a few seconds.

The key mathematical proposition provides the theoretical basis and indicates a procedure for reducing prediction errors for any given combination of core and auxiliary series. As a consequence, it suffices to analyze simple sub-collections of variables and sift the results according to computed likelihoods (which can be achieved very rapidly in practice). Also, a finding from our simulations is that a straightforward implementation of LASSO for variable selection in a VAR context does not work well, and is extremely expensive, although further research into better penalization structures is warranted.

To re-iterate our framework, with regard to other approaches available in the forecasting literature, our method simultaneously addresses three key considerations that commonly arise in actual applications:

- (i) *Dimension Reduction*: Dynamic factor analysis and random projections produce reductions in the parameter count by expressing the observation vector in terms of a smaller number of processes. Our approach reduces dimension by screening out auxiliary variables. This enables variable preservation, as noted in (iii).
- (ii) *Variable Targeting*: This allows one to concentrate on predictive power for the main time series of interest. With most VAR treatments – whether with LASSO, Dynamic Factors, or other model types – variables are considered in a generic or symmetric fashion (i.e. having equal importance with regards to forecast performance), so that predictive accuracy for the main series may be sacrificed to attain greater accuracy for series of only minor importance. Our approach achieves the aim of variable targeting without the great computational expense of a full LASSO or Bayesian VAR.
- (iii) *Variable Preservation*: Our approach retains, or preserves, the dynamic representations of the original variables, which is discarded when using techniques such as dynamic factor analysis and random projections. Also, our method produces a single model involving core and auxiliary variables, which is lost in model averaging approaches.

We use the proposed framework in an application to modeling real GDP and unemployment on the basis of a large set of indicators spanning various major sectors of the economy and different measures of labor market condition. The resulting base VAR containing the core pair along with selected auxiliaries is highly compact, and allows for convenient estimation and rapid sequential computation of predictions as the out-of-sample period is updated on a rolling basis. Empirically, there are substantial improvements in forecasting national GDP and unemployment with these auxiliaries compared to a univariate approach. In particular, the corresponding out-of-sample forecasts lead to collective forecast errors that are notably smaller (with the differences in error series approaching conventional levels of significance).

Here, in an economic application, our focus has been to handle large dimensional data with an eye on specific variables of interest. We have shown how to determine an optimal set of predictive indicators based on an empirical methodology using a VAR model, which satisfies the goals of attaining an efficient approach that keeps the computational burden in check while also satisfying the three major objectives of dimension reduction, variable targeting, and variable preservation.



The use of VARs as a general statistical framework continues to expand into diverse datasets and subject areas, and situations with Big Data have become increasingly more common. Our methodology helps assess whether the extensive information contained in a given large dataset actually improves our predictions of key indicators, and how to most efficiently use the numerous possibilities for supporting variables in analyzing trends and making forecasts of prime interest.

## Disclosure statement

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the U.S. Census Bureau.

## References

- [1] J. Abowd and L. Vilhuber, *National estimates of gross employment and job flows from the quarterly workforce indicators with demographic and industry detail*, J. Econom. 161 (2011), pp. 82–99.
- [2] R. Barnichon and P. Garda, *Forecasting unemployment across countries: the ins and outs*, Eur. Econ. Rev. 84 (2016), pp. 165–183.
- [3] R. Barnichon and C.J. Nekarda, *The ins and outs of forecasting unemployment: using labor force flows to forecast the labor market*, Brookings. Pap. Econ. Act. 2 (2012), pp. 83–131.
- [4] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. R. Statist. Soc. Ser. B (Methodological) 57 (1995), pp. 289–300.
- [5] F. Blasques, S.J. Koopman, M. Mallee, and Z. Zhang, *Weighted maximum likelihood estimator for mixed frequency dynamic factor models*, J. Econom. 193 (2016), pp. 405–417.
- [6] L.F. Bringmann, L.H.J.M. Lemmens, M.J.H. Huibers, D. Borsboom, and F. Tuerlinckx, *Revealing the dynamic network structure of the beck depression inventory-II*, Psychol. Med. 45 (2015), pp. 747–757.
- [7] R. Brüggemann and H. Lütkepohl, *Lag selection in subset VAR models with an application to a U.S. monetary system*, in *Econometric Studies: A Festschrift in Honour of Joachim Frohn*, R. Friedmann, L. Knüppel and H. Lütkepohl, eds., LIT Verlag, Münster, 2001, pp. 107–128.
- [8] A. Carriero, T.E. Clark, and M. Marcellino, *Common drifting volatility in large Bayesian VARs*, J. Bus. Econ. Statist. 34 (2016), pp. 375–390.
- [9] C. Chakraborty and A. Joseph, *Machine learning at central banks*, Bank of England: Staff Working Paper No 674, 2017, pp. 1–84.
- [10] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan, *Productivity and selection of human capital with machine learning*, Am. Econ. Rev. 106 (2016), pp. 124–127.
- [11] J.C. Chan, E. Eisenstat, and G. Koop, *Large Bayesian VARMA*s, J. Econom. 192 (2016), pp. 374–390.
- [12] L.J. Christiano and L. Ljungqvist, *Money does Granger-cause output in the bivariate money-output relation*, J. Monet. Econ. 22 (1988), pp. 217–235.
- [13] R.A. Davis, P. Zang, and T. Zheng, *Sparse vector autoregressive modeling*, J. Comput. Graph. Stat. 25 (2015), pp. 1077–1096.
- [14] G. Deshpande and X. Hu, *Investigating effective brain connectivity from FMRI data: past findings and current issues with reference to granger causality analysis*, Brain. Connect. 2 (2012), pp. 235–245.
- [15] F. Diebold and R. Mariano, *Comparing predictive accuracy*, J. Bus. Econ. Stat. 13 (1995), pp. 253–263.
- [16] W. Enders, T. Sandler, and G. Khusrav, *Domestic versus transnational terrorism: data, decomposition, and dynamics*, J. Peace. Res. 48 (2011), pp. 319–337.
- [17] M. Forni, M. Hallin, M. Lippi, and L. Reichlin, *The generalized dynamic factor model: identification and estimation*, Rev. Econ. Statist. 82 (2000), pp. 540–554.

- [18] B.M. Friedman and K.N. Kuttner, *Money, income, prices, and interest rates*, Am. Econ. Rev. 82 (1992), pp. 472–492.
- [19] D. Gefang, *Bayesian doubly adaptive elastic-net lasso for VAR shrinkage*, Int. J. Forecast. 30 (2014), pp. 1–11.
- [20] D. Giannone, M. Lenza, and G. Primiceri, *Prior selection for vector autoregressions*, Rev. Econ. Statist. 97 (2015), pp. 436–451.
- [21] C. Granger, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica 37 (1969), pp. 424–438.
- [22] S. Guo, Y. Wang, and Q. Yao, *High-dimensional and banded vector autoregressions*, Biometrika 103 (2016), pp. 889–903.
- [23] F. Hoffmann and T. Lemieux, *Unemployment in the great recession: a comparison of Germany, Canada, and the United States*, J. Labor. Econ. 34 (2016), pp. S95–S139.
- [24] H.R. Hyatt and J.R. Spletzer, *The shifting job tenure distribution*, Labour. Econ. 41 (2016), pp. 363–377.
- [25] H.R. Hyatt and J.R. Spletzer, *The recent decline of single quarter jobs*, Labour. Econ. 46 (2017), pp. 166–176.
- [26] L. Kilian, *Small-sample confidence intervals for impulse response functions*, Rev. Econ. Stat. 80 (1998), pp. 218–230.
- [27] G.M. Koop, *Forecasting with medium and large Bayesian VARs*, J. Appl. Econ. 28 (2013), pp. 177–203.
- [28] G.M. Koop, *Bayesian methods for empirical macroeconomics with big data*, Review of Economic Analysis 9 (2017), pp. 33–56.
- [29] G.M. Koop and D. Korobilis, *Forecasting with high dimensional panel VARs*. <https://sites.google.com/site/garykoop/research>.
- [30] G.M. Koop, D. Korobilis, and D. Pettenuzzo, *Bayesian compressed vector autoregressions*, J. Econom. 210 (2019), pp. 135–154.
- [31] H. Lütkepohl, *New Introduction to Multiple Time Series*, Springer, New York, 2007.
- [32] T.S. McElroy and D. Findley, *Fitting constrained vector autoregression models*, in *Empirical Economic and Financial Research – Theory, Methods, and Practice*, J. Beran, Y. Feng, and H. Hebbel, eds., Springer, New York, 2015, pp. 451–470.
- [33] S. Mullainathan and J. Spiess, *Machine learning: an applied econometric approach*, J. Econ. Perspectives 31 (2017), pp. 87–106.
- [34] W.B. Nicholson, D.S. Matteson, and J. Bien, *Varx-l: structured regularization for large vector autoregressions with exogenous variables*, Int. J. Forecast. 33 (2017), pp. 627–651.
- [35] J.H.W. Penm, T.J. Brailsford, and R.D. Terrell, *A robust algorithm in sequentially selecting subset time series systems using neural networks*, J. Time Ser. Anal. 21 (2000), pp. 389–412.
- [36] J.H.W. Penm and R.D. Terrell, *Multivariate subset autoregressive modelling with zero constraints for detecting overall causality*, J. Econom. 24 (1984), pp. 311–330.
- [37] C. Sims, *Macroeconomics and reality*, Econometrica 1 (1980), pp. 1–48.
- [38] S. Song and P.J. Bickel, *Large vector auto regressions*, preprint (2011). arXiv:1106.3915.
- [39] J. Stock and M. Watson, *Macroeconomic forecasting using diffusion indexes*, J. Bus. Econ. Stat. 20 (2002), pp. 147–162.
- [40] J. Stock and M. Watson, *Factor models and structural vector autoregressions in macroeconomics*, Handb. Macroecon. 8 (2016), pp. 415–525.
- [41] M. Taniguchi and Y. Kakizawa, *Asymptotic Theory of Statistical Inference for Time Series*, Springer, New York, 2000.