International Journal of Forecasting 33 (2017) 627–651

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

VARX-L: Structured regularization for large vector autoregressions with exogenous variables

William B. Nicholson [a,*], David S. Matteson [b,c], Jacob Bien [d,e]

[a] Department of Statistical Science, Cornell University, 301 Malott Hall, Ithaca, NY 14853, United States
[b] Department of Statistical Science, Cornell University, 1196 Comstock Hall, Ithaca, NY 14853, United States
[c] Department of Social Statistics, Cornell University, 1196 Comstock Hall, Ithaca, NY 14853, United States
[d] Department of Biological Statistics and Computational Biology, Cornell University, 1178 Comstock Hall, Ithaca, NY 14853, United States
[e] Department of Statistical Science, Cornell University, 1178 Comstock Hall, Ithaca, NY 14853, United States

ARTICLE INFO

Keywords:
Big data
Forecasting
Group lasso
Macroeconometrics
Time series

ABSTRACT

The vector autoregression (VAR) has long proven to be an effective method for modeling the joint dynamics of macroeconomic time series, as well as for forecasting. One major shortcoming of the VAR that has limited its applicability is its heavy parameterization: the parameter space grows quadratically with the number of series included, quickly exhausting the available degrees of freedom. Consequently, using VARs for forecasting is intractable for low-frequency, high-dimensional macroeconomic data. However, empirical evidence suggests that VARs that incorporate more component series tend to result in more accurate forecasts. Most conventional methods that allow for the estimation of large VARs either require *ad hoc* subjective specifications or are computationally infeasible. Moreover, as global economies become more intricately intertwined, there has been a substantial interest in incorporating the impact of stochastic, unmodeled *exogenous* variables. Vector autoregression with exogenous variables (VARX) extends the VAR to allow for the inclusion of unmodeled variables, but faces similar dimensionality challenges.

This paper introduces the VARX-L framework, a structured family of VARX models, and provides a methodology that allows for both efficient estimation and accurate forecasting in high-dimensional analysis. VARX-L adapts several prominent scalar regression regularization techniques to a vector time series context, which greatly reduces the parameter space of VAR and VARX models. We also highlight a compelling extension that allows for shrinking toward reference models, such as a vector random walk. We demonstrate the efficacy of VARX-L in both low- and high-dimensional macroeconomic forecasting applications and simulated data examples. Our methodology is easy to reproduce in a publicly available R package.

1. Introduction

The practice of macroeconomic forecasting was spearheaded by Klein and Goldberger (1955), whose eponymous simultaneous equation system forecast the behaviors of 15 annual macroeconomic indicators jointly, including

* Corresponding author.
E-mail addresses: wbn8@cornell.edu (W.B. Nicholson), matteson@cornell.edu (D.S. Matteson), jbien@cornell.edu (J. Bien).
URLs: http://www.wbnicholson.com (W.B. Nicholson), http://www.stat.cornell.edu/~matteson/ (D.S. Matteson), http://faculty.bscb.cornell.edu/~bien/ (J. Bien).

http://dx.doi.org/10.1016/j.ijforecast.2017.01.003
0169-2070/

consumer expenditures, interest rates, and corporate profits. The parameterization and identification restrictions of these models were influenced heavily by Keynesian economic theory. As computing power increased, such models became larger and began to utilize higher frequency data. Forecasts and simulations from these models were commonly used to inform government policymakers regarding the overall state of the economy and to influence policy decisions (Welfe, 2013). As the Klein-Goldberger model and its extensions were motivated primarily by Keynesian economic theory, the collapse of the Bretton Woods monetary system and severe oil price shocks led to widespread forecasting failures in the 1970s (Diebold, 1998). At this time, the vector autoregression (VAR), popularized by Sims (1980), emerged as an atheoretical forecasting technique that was underpinned by statistical methodology and not subject to the ebb and flow of contemporary macroeconomic theory.

Unfortunately, the flexibility of the VAR can create modeling complications. In the absence of prior information, the VAR assumes that every series interacts linearly with both its own past values and those of every other series included. Such a model is known as an *unrestricted* VAR. As most economic series are low-frequency (monthly, quarterly, or annual) there is rarely enough data available to allow accurate forecasts using large unrestricted VARs. Such models are overparameterized, provide inaccurate forecasts, and are very sensitive to changes in economic variables. Consequently, the VAR's parameter space must be reduced in such applications, either in a data-driven manner or based upon the modeler's knowledge of the underlying economic system. This model selection process has been described as "blending data and personal beliefs according to a subjective, undocumented procedure that others cannot duplicate" (Todd, 1990, p. 18).

Despite their overparameterization, large VARs can be preferable to their smaller counterparts in many applications, as small models can exclude potentially relevant variables. Ideally, a variable should always be included in the model unless one has prior knowledge that it is irrelevant. For example, as was described by Lütkepohl (2014), modeling the Taylor rule (Taylor, 1993) requires an estimate of the "output gap" between the real gross domestic product and the potential output. The output gap is difficult to measure and can include many explanatory variables that encompass disaggregated economic measurements. Moreover, recent work by Hendry and Hubrich (2011) and Ibarra (2012) showed that incorporating disaggregated series improves the forecasts of macroeconomic aggregates such as the consumer price index. Hence, a large vector autoregression with a coherent variable selection procedure is required in order to utilize all relevant economic information properly in these scenarios.

Shortly after the inception of the VAR, efforts were made to develop a systematic approach to the reduction of its parameterization. Early attempts, such as that of Litterman (1979), centered upon a Bayesian approach underpinned by contemporary macroeconomic theory. In applying a Bayesian VAR with a Gaussian prior (analogous to ridge regression), specific priors were formulated based upon stylized facts regarding US macroeconomic data. For example, the popular *Minnesota prior* incorporates the prevailing belief that macroeconomic variables can be modeled reasonably by means of a univariate random walk, by shrinking the estimated models toward univariate unit root processes.

The Bayesian VAR with a Minnesota prior was shown by Robertson and Tallman (1999) to produce forecasts that were superior to those from the conventional VAR, univariate models, and traditional simultaneous equation models. However, this approach is very restrictive; in particular, it assumes that all series are contemporaneously uncorrelated, and requires the specification of several hyperparameters. Moreover, the Minnesota prior cannot accommodate large VARs itself. As was pointed out by Banbura, Giannone, and Reichlin (2009), when constructing a 40-variable system, in addition to the Minnesota prior, Litterman (1986b) also imposed strict economically-motivated restrictions in order to limit the number of variables in each equation.

The modern Bayesian VAR extensions that were originally proposed by Kadiyala and Karlsson (1997) and compiled by Koop (2013) show that empirical regularization methods alone allow for the accurate forecasting of large VARs. Such procedures impose data-driven restrictions on the parameter space while allowing for more general covariance specifications and the estimation of hyperparameters via empirical Bayes or Markov chain Monte Carlo methods. These approaches are computationally expensive, and multi-step forecasts are nonlinear and must be obtained by additional simulation. Using a conjugate Gaussian-Wishart prior, Banbura et al. (2009) extended the Minnesota prior to a high-dimensional setting with a closed-form posterior distribution. Their technique uses a single hyperparameter, which is estimated by cross-validation. However, their specification does not perform variable selection, and their penalty parameter selection procedure is more natural within a frequentist framework.

More recent attempts to reduce the parameter space of VARs have incorporated the *lasso* (Tibshirani, 1996), a least squares variable selection technique. These approaches include the *lasso-VAR* that was proposed by Hsu, Hung, and Chang (2008) and explored further by Davis, Zang, and Zheng (2016), Li and Chen (2014), and Song and Bickel (2011). Its theoretical properties were investigated by Basu and Michailidis (2015) and Kock and Callot (2015). Gefang (2012) considered a Bayesian implementation of the elastic net, an extension of the lasso proposed by Zou and Hastie (2005) that accounts for highly correlated covariates. However, their implementation is not computationally tractable and little forecasting improvement relative to existing methods is observed. The lasso-VAR has several advantages over Bayesian approaches, as it is more computationally efficient in high dimensions, performs variable selection, and can readily compute multi-step forecasts and their associated prediction intervals.

In many applications, a VAR's forecasts can be improved through the incorporation of unmodeled exogenous variables, which are determined outside the VAR. Examples of exogenous variables are context-dependent and range from leading indicators to weather-related measurements. In many scenarios, global macroeconomic variables, such

as world oil prices, may be considered exogenous. Such models are most commonly referred to as "VARX" in the econometrics literature, but are also known as "transfer function" or "distributed lag" models.

VARX has become an especially popular approach in the modeling of small open economies, as they are generally sensitive to a wide variety of global macroeconomic variables which evolve independently of their internal indicators. For example, Cushman and Zha (1997) used a structural VARX model to analyze the effects of monetary policy shocks in Canada. The VARX is also amenable under scenarios in which forecasts are desired from only a subset of the series included in a VAR, as its corresponding VARX has a reduced parameterization by construction. VARX models have received considerable attention not only in economics, but also in marketing (Nijs, Srinivasan, & Pauwels, 2007), political science (Wood, 2009), and real estate (Brooks & Tsolacos, 2000).

Unfortunately, dimensionality issues have limited the utility of the VARX. As a result of the aforementioned overparameterization concerns, most applications are limited to no more than six series in the conventional unrestricted VAR context (Bernanke, Boivin, & Eliasz, 2005), forcing the practitioner to specify *a priori* a reduced subset of series to include. The VARX faces similar restrictions. As was outlined by Penm, Penm, and Terrell (1993), the lag order (the maximum number of lagged observations to include) may differ between modeled and unmodeled series. Hence, in order to select a VARX model using standard information criterion minimization based methods, one must fit all subset models up to the predetermined maximal lag order for both the series of forecasting interest (which we refer to as *endogenous* throughout this paper) and the exogenous series. Moreover, unlike the conventional VAR, zero constraints (restrictions fixing certain model parameters to zero) are generally expected.

As reducing the parameter space of the VARX model is often viewed as an economic rather than a statistical problem, it has received considerably less attention. Ocampo and Rodríguez (2012) extended the aforementioned Bayesian VAR estimation methods to the VARX context. George, Sun, and Ni (2008) applied stochastic search variable selection to the VARX framework; this provides a data-driven method for determining zero restrictions, but is not scalable to high dimensions. Chiuso and Pillonetto (2010) proposed the estimation of a VARX model with lasso and group lasso penalties, but did not elaborate on potential group structures.

This paper seeks to bridge the considerable gap between the regularization and macroeconomic forecasting communities. We develop the VARX-L framework, which allows for high-dimensional penalized VARX estimation while incorporating the unique structure of the VARX model in a computationally efficient manner. We implement this framework by developing substantial modifications to existing lasso and group lasso solution algorithms, which were designed primarily for univariate regression applications with no time dependence.

We extend the lasso and its structured counterparts so as to impose structured sparsity on the VARX, taking into account characteristics such as lag coefficient matrices, the delineation between a component's own lags and those of another component, and a potential nested structure between endogenous and exogenous variables. Our methods offer a considerable degree of flexibility in capturing the underlying dynamics of an economic system while imposing minimal assumptions on the parameter space.

Moreover, unlike previous approaches, our adaptation of convex optimization algorithms to a multivariate time series setting means that our models are well-suited for the simultaneous forecasting of high-dimensional low-frequency macroeconomic time series. In particular, our models allow for prediction in scenarios where the number of component series and exogenous variables included is close to or exceeds the length of the series. Our procedures, which avoid the use of subjective or complex hyperparameters, are publicly available in our R package `BigVAR` and can be easily applied by practitioners.

Section 2 describes the notation used throughout the paper and introduces our structured regularization methodology. Section 3 provides details of our implementation and presents three macroeconomic forecasting applications. Section 4 describes the "Minnesota VARX-L", an extension that allows for the incorporation of unit root non-stationarity by shrinking toward a vector random walk. Section 5 presents a simulation study, and Section 6 contains our conclusion. The Appendix details the solution strategies and algorithms that comprise the VARX-L class of models.

2. Methodology

A k-dimensional multivariate time series $\{\mathbf{y}_t\}_{t=1}^T$ and a m-dimensional exogenous multivariate time series $\{\mathbf{x}_t\}_{t=1}^T$ follow a vector autoregression with exogenous variables of order (p, s), denoted $\text{VARX}_{k,m}(p, s)$, if the following linear relationship holds (conditional on initialization):

$$\mathbf{y}_t = \boldsymbol{\nu} + \sum_{\ell=1}^{p} \boldsymbol{\Phi}^{(\ell)} \mathbf{y}_{t-\ell} + \sum_{j=1}^{s} \boldsymbol{\beta}^{(j)} \mathbf{x}_{t-j} + \mathbf{u}_t \quad \text{for } t = 1, \ldots, T, \tag{1}$$

where $\boldsymbol{\nu}$ denotes a k-dimensional constant intercept vector, $\boldsymbol{\Phi}^{(\ell)}$ represents a $k \times k$ endogenous coefficient matrix at lag $\ell = 1, \ldots, p$, $\boldsymbol{\beta}^{(j)}$ represents a $k \times m$ exogenous coefficient matrix at lag $j = 1, \ldots, s$, and $\mathbf{u}_t$ denotes a k-dimensional white noise vector that is independent and identically distributed with mean zero and nonsingular covariance matrix $\boldsymbol{\Sigma}_u$. A VAR, which is a special case of the VARX, can be represented by Eq. (1) with $\boldsymbol{\beta}^{(j)} = \mathbf{0}$ for $j = 1, \ldots, s$.

In a low-dimensional setting, in which the number of predictors included is substantially smaller than the length of the series, T, the VARX model can be fitted using multivariate least squares, with $\boldsymbol{\nu}, \boldsymbol{\Phi} = [\boldsymbol{\Phi}^{(1)}, \ldots, \boldsymbol{\Phi}^{(p)}]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(s)}]$ estimated as

$$\underset{\boldsymbol{\nu}, \boldsymbol{\Phi}, \boldsymbol{\beta}}{\text{argmin}} \quad \sum_{t=1}^{T} \left\| \mathbf{y}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{p} \boldsymbol{\Phi}^{(\ell)} \mathbf{y}_{t-\ell} - \sum_{j=1}^{s} \boldsymbol{\beta}^{(j)} \mathbf{x}_{t-j} \right\|_F^2, \tag{2}$$

where $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ denotes the Frobenius norm of a matrix A, which reduces to the L_2 norm when A is a vector. In the absence of regularization, the VARX$_{k,m}(p, s)$ requires the estimation of $k(1+kp+ms)$ regression parameters. The following section will apply several convex penalties to Eq. (2) which help to reduce the parameter space of the VARX by imposing sparsity on $\mathbf{\Phi}$ and $\boldsymbol{\beta}$.

2.1. VARX-L: structured penalties for VARX modeling

This section introduces VARX-L, a general penalized multivariate regression framework for large VARX models. We consider structured objectives of the form

$$\min_{\boldsymbol{\nu},\mathbf{\Phi},\boldsymbol{\beta}} \sum_{t=1}^{T} \left\| \mathbf{y}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{p} \mathbf{\Phi}^{(\ell)} \mathbf{y}_{t-\ell} - \sum_{j=1}^{s} \boldsymbol{\beta}^{(j)} \mathbf{x}_{t-j} \right\|_F^2 + \lambda \Big(\mathcal{P}_y(\mathbf{\Phi}) + \mathcal{P}_x(\boldsymbol{\beta}) \Big), \tag{3}$$

where $\lambda \geq 0$ is a penalty parameter which is selected in a sequential, *rolling* manner according to a procedure that is discussed in Section 3.1; $\mathcal{P}_y(\mathbf{\Phi})$ denotes a penalty function on endogenous coefficients; and $\mathcal{P}_x(\boldsymbol{\beta})$ denotes a penalty function on exogenous coefficients.

Table 1 details the penalty structures proposed in this paper, with all but the last having this separable structure. The following section will discuss each penalty structure in detail. Note that since we utilize a single penalty parameter for all model coefficients, it is required that all series included be on the same scale; hence, we assume that the series are standardized to each have a zero mean and unit variance, prior to estimation.

Eqs. (4)–(5) adapt the *group lasso* penalty proposed by Yuan and Lin (2006) to the VARX setting. The group lasso partitions the parameter space into groups of related variables which are shrunk toward zero. Within a group, all variables are either set identically to zero or are nonzero. Our choices of $\mathcal{P}_y$ and $\mathcal{P}_x$ create structured sparsity based on pre-specified groupings, which are designed to incorporate the intrinsic lagged structure of the VARX. The proposed "lag group" methods have a substantial advantage over popular Bayesian approaches, in that they both shrink least squares estimates toward zero and perform variable selection in a computationally efficient manner.

Sparsity in the coefficient matrix is desirable when k and m are large because the conventional VARX is grossly overparameterized. As was stated by Litterman (1984), it is widely believed in macroeconomic forecasting that small bits of relevant information exist throughout the data, and economic theory is not informative with regard to the manner in which this information is scattered. The proposed VARX-L framework provides a systematic approach in order to filter as much information as possible, assigning an appropriate weight to each bit.

The group lasso penalty function was explored in the VAR context by Song and Bickel (2011), who considered several structured groupings, with a particular emphasis on creating a distinction between a series' own lags and those of another series. The theoretical properties of the use of a group lasso penalty in the VAR setting were explored further by Basu, Shojaie, and Michailidis (2015).

One feature of the lag group VARX-L is that it does not impose sparsity within a group. Song and Bickel (2011) attempt to circumvent this constraint by including several additional lasso penalties, but such an approach requires the use of a multi-dimensional gridsearch for selecting the penalty parameters. The penalties for the proposed sparse lag group VARX-L and sparse own/other group VARX-L, listed in Eqs. (6)–(7), instead implement the *sparse group lasso* (Simon, Friedman, Hastie, & Tibshirani, 2013), which extends the group lasso to allow within-group sparsity. The sparse group lasso can be viewed analogously to the elastic net (Zou & Hastie, 2005), but extended to structured penalties.

The penalty for the basic VARX-L adapts the lasso in Eq. (8); it considers no structure, or can be viewed as a group lasso penalty that assigns each coefficient to a singleton group. This most basic penalty has computational advantages over more complex approaches in very high-dimensional scenarios. Finally, the penalty for the proposed endogenous-first VARX-L, Eq. (9), incorporates a nested penalty structure such that, within a lag, endogenous coefficients are prioritized before their exogenous counterparts. Since this penalty structure is not separable in the manner of Eq. (3), its penalty function is denoted as $\mathcal{P}_{y,x}$.

2.1.1. Lag group VARX-L

We first present the *lag* group VARX-L (Eq. (4)), in which the endogenous coefficients are grouped according to their lagged coefficient matrix $\mathbf{\Phi}^{(\ell)}$ for $\ell = 1, \ldots, p$, and each exogenous component series is partitioned into its own group at every lag. This structured grouping is expressed as

$$\mathcal{P}_y(\mathbf{\Phi}) = \sqrt{k^2} \sum_{\ell=1}^{p} \|\mathbf{\Phi}^{(\ell)}\|_F,$$

$$\mathcal{P}_x(\boldsymbol{\beta}) = \sqrt{k} \sum_{j=1}^{s} \sum_{i=1}^{m} \|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F.$$

Note that it is necessary to weight the penalty in order to avoid regularization favoring larger groups because the cardinalities of the endogenous and exogenous groups differ. This structure implies that a coefficient matrix at lag ℓ is either entirely nonzero or entirely zero for each endogenous series. Similarly, the relationship between an exogenous and an endogenous series at lag j will either be nonzero for all endogenous series or identically zero. A potential sparsity pattern generated by this structure (with $k = 3$, $p = 5$, $m = 2$, and $s = 3$) is shown in Fig. 1, with the active (i.e., nonzero) elements shaded.

Estimating the lag group VARX-L is tractable even in high dimensions relative to Bayesian regularization methods, such as stochastic search variable selection (George et al., 2008). We are able to extend the efficient group lasso solution method proposed by Qin, Scheinberg, and Goldfarb (2010), who utilize a block coordinate descent procedure and transform each "one group" subproblem to a

Table 1
The proposed VARX-L penalty functions.

Group name	$\mathcal{P}_y(\boldsymbol{\Phi})$	$\mathcal{P}_x(\boldsymbol{\beta})$
(4) Lag	$\sqrt{k^2}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}^{(\ell)}\|_F$	$\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F$
(5) Own/other	$\sqrt{k}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}_{\text{on}}^{(\ell)}\|_F+\sqrt{k(k-1)}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}_{\text{off}}^{(\ell)}\|_F$	$\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F$
(6) Sparse lag	$(1-\alpha)\sqrt{k^2}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}^{(\ell)}\|_F+\alpha\|\boldsymbol{\Phi}\|_1$	$(1-\alpha)\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F+\alpha\|\boldsymbol{\beta}\|_1$
(7) Sparse own/other	$(1-\alpha)\left(\sqrt{k}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}_{\text{on}}^{(\ell)}\|_F+\sqrt{k(k-1)}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}_{\text{off}}^{(\ell)}\|_F\right)+\alpha\|\boldsymbol{\Phi}\|_1$	$(1-\alpha)\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F+\alpha\|\boldsymbol{\beta}\|_1$
(8) Basic	$\|\boldsymbol{\Phi}\|_1$	$\|\boldsymbol{\beta}\|_1$
(9) Endogenous-first	$\mathcal{P}_{y,x}(\boldsymbol{\Phi},\boldsymbol{\beta})=\sum_{\ell=1}^{p}\sum_{j=1}^{k}\left(\|[\boldsymbol{\Phi}_{j,\cdot}^{(\ell)},\boldsymbol{\beta}_{j,\cdot}^{(\ell)}]\|_F+\|\boldsymbol{\beta}_{j,\cdot}^{(\ell)}\|_F\right)$	

Note that $\boldsymbol{\Phi}_{\text{on}}^{(\ell)}$ and $\boldsymbol{\Phi}_{\text{off}}^{(\ell)}$ denote the diagonal and off-diagonal elements of the coefficient matrix $\boldsymbol{\Phi}^{(\ell)}$, respectively.

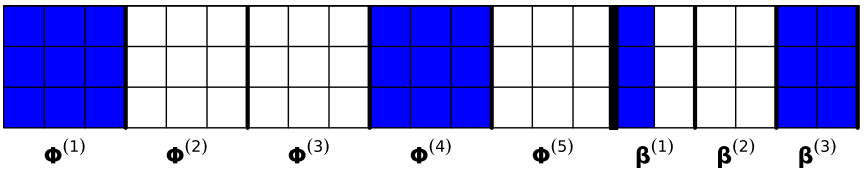

Fig. 1. Example sparsity pattern (active elements shaded) produced by a lag group VARX-L$_{3,2}$(5, 3).

trust-region framework. These subproblems can then be solved efficiently via univariate optimization. Details of our algorithm are provided in Appendix A.3.2.

The lag group structure is advantageous for applications in which all series tend to exhibit comparable dynamics, such as forecasting the disaggregate subcomponents of a composite index. It can also serve as a powerful tool for lag selection. However, giving equal consideration to every entry in a coefficient matrix may be inappropriate in many settings. In many applications, the diagonal entries of each $\boldsymbol{\Phi}^{(\ell)}$, which represent regressions on a series' own lags, are more likely to be nonzero than the off-diagonal entries, which represent lagged cross-dependence with other components. The *own/other* group VARX-L (Eq. (5)) allows each lag matrix $\boldsymbol{\Phi}^{(\ell)}$ to be partitioned into separate groups by assigning the endogenous penalty structure

$$\mathcal{P}_y(\boldsymbol{\Phi})=\sqrt{k}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}_{\text{on}}^{(\ell)}\|_F+\sqrt{k(k-1)}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}_{\text{off}}^{(\ell)}\|_F,$$

where $\boldsymbol{\Phi}_{\text{on}}^{(\ell)}$ denotes the diagonal elements of $\boldsymbol{\Phi}^{(\ell)}$ and $\boldsymbol{\Phi}_{\text{off}}^{(\ell)}$ denotes its off-diagonal entries.

An example of this sparsity pattern is shown in Fig. 2. The computational modifications required to utilize the own/other structure are detailed in Appendix A.3.3. This delineation between own and other lags is often incorporated in macroeconomic forecasting. As was detailed by Litterman (1986a), the traditional Minnesota prior operates under the assumption that a series' own past values account for most of its variation, meaning that they are shrunk by a smaller factor than the realizations of other series. The strong forecasting performances of the VARX-L procedures that utilize the own/other structure in Section 3.3 provide further justification for Litterman's beliefs.

In addition to the own/other and lag group structures, one could also consider a variety of application-dependent partitions. This idea was explored briefly by Song and Bickel (2011), who posit potentially segmenting financial or economic series based upon industry or sector characteristics, in addition to a lag-based grouping. However, since many applications lack *a priori* information about the potential relationships among the series included, we prefer to structure our groupings around the intrinsic structure of the VARX.

2.1.2. Sparse lag group VARX-L

For certain applications, a group penalty might be too restrictive. If a group is active, all coefficients in the group will be nonzero, and including a large number of groups increases the computation time substantially. Moreover, it is inefficient to include an entire group if, for example, only one coefficient is truly nonzero. The *sparse group lasso*, proposed by Simon et al. (2013), allows for within-group sparsity through a convex combination of the lasso and group lasso penalties. The sparse lag group VARX-L in Eq. (6) results in a penalty of the form

$$\mathcal{P}_y(\boldsymbol{\Phi})=(1-\alpha)\left(\sqrt{k^2}\sum_{\ell=1}^{p}\|\boldsymbol{\Phi}^{(\ell)}\|_F\right)+\alpha\|\boldsymbol{\Phi}\|_1,$$

$$\mathcal{P}_x(\boldsymbol{\beta})=(1-\alpha)\left(\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\beta}_{\cdot,i}^{(j)}\|_F\right)+\alpha\|\boldsymbol{\beta}\|_1,$$

where $0\leq\alpha\leq 1$ is an additional penalty parameter that controls the within-group sparsity. If we have no prior knowledge of the important predictors, we weight according to the relative group sizes and set $\alpha=\frac{1}{k+1}$, though α could also be estimated by cross-validation.

The inclusion of the L_1 norm allows for within-group sparsity, meaning that individual coefficients within a group can be set to zero even if the group is considered active. An example sparsity pattern is depicted in Fig. 3.

Since the inclusion of within-group sparsity does not create a separable objective function, conventional group lasso solution methods, such as coordinate descent, are no longer applicable. Following Simon et al. (2013), our

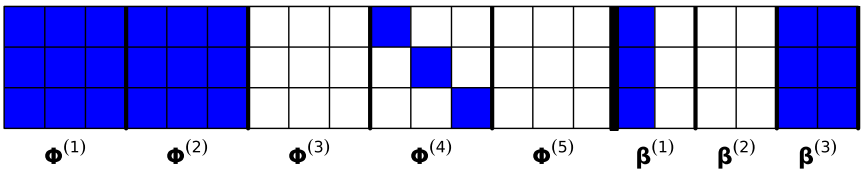

Fig. 2. Example sparsity pattern (active elements shaded) produced by an own/other group VARX-$L_{3,2}(5, 3)$.

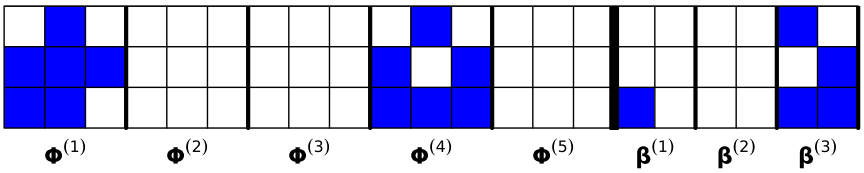

Fig. 3. Example sparsity pattern (active elements shaded) produced by a sparse lag group VARX-$L_{3,2}(5, 3)$.

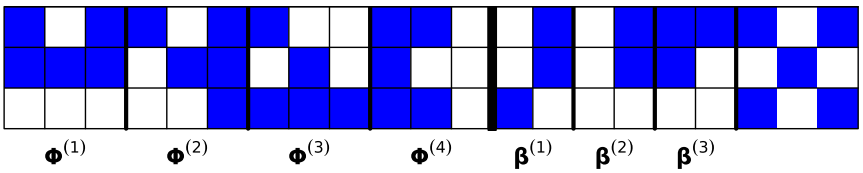

Fig. 4. Example sparsity pattern (active elements shaded) produced by a basic VARX-$L_{3,2}(5, 3)$.

estimation algorithm for the sparse lag group VARX-L makes use of proximal gradient descent. The details of this approach and our implementation are provided in Appendix A.3.4. This penalty can also be extended to alternative groupings. Consequently, we also consider the *sparse own/other group* VARX-L in Eq. (7) as an estimation procedure.

2.1.3. Basic VARX-L

The basic VARX-L in Eq. (8), proposed by Chiuso and Pillonetto (2010), incorporates no structure and can be viewed as a special case of the sparse lag group VARX-L in which $\alpha = 1$, resulting in penalties of the form

$$\mathcal{P}_y(\boldsymbol{\Phi}) = \|\boldsymbol{\Phi}\|_1, \qquad \mathcal{P}_x(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1.$$

The L_1 penalty induces sparsity in the coefficient matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\beta}$ by zeroing individual entries. An example sparsity pattern is depicted in Fig. 4.

One major advantage of the basic VARX-L over its structured counterparts is its computational tractability. Our solution approach involves the use of coordinate descent, which was popularized for lasso problems by Friedman, Hastie, and Tibshirani (2010). Coordinate descent consists of partitioning the basic VARX-L into subproblems for each scalar element $[\boldsymbol{\Phi}, \boldsymbol{\beta}]_{ij}$, solving component-wise, and updating until convergence. This approach is computationally efficient because, in the basic VARX-L context, each subproblem has a closed-form solution. Tseng (2001) established that global convergence arises from individual subproblems being solved in the coordinate descent framework. Our solution strategy and algorithm are detailed in Appendix A.3.1.

2.2. An endogenous-first active set

Previously, we have considered only structures that assign endogenous and exogenous variables to separate groups. This section considers a nested structure that can take into account the relative importance of endogenous and exogenous predictor series.

There may be an *a priori* importance ranking among endogenous and exogenous variables in certain scenarios. For example, the endogenous variables could represent economic indicators that are of interest in a small open economy, with global macroeconomic indicators serving as exogenous variables. In such a scenario, it may be desirable for exogenous variables to enter into a forecasting equation only if endogenous variables are also present at a given lag ℓ. We can consider such a structure by utilizing a *hierarchical group lasso* penalty (see e.g. Jenatton, Mairal, Obozinski, & Bach, 2011). The endogenous-first VARX-L penalty function in Eq. (9) takes the form

$$\mathcal{P}_{y,x}(\boldsymbol{\Phi}, \boldsymbol{\beta}) = \sum_{\ell=1}^{p} \sum_{j=1}^{k} \left(\|[\boldsymbol{\Phi}_{j,\cdot}^{(\ell)}, \boldsymbol{\beta}_{j,\cdot}^{(\ell)}]\|_F + \|\boldsymbol{\beta}_{j,\cdot}^{(\ell)}\|_F \right). \tag{10}$$

Under this structure, at a given lag, exogenous variables can enter the model only after the endogenous variables at the same lag. Note that this structure requires that $s \leq p$. It should also be noted that Eq. (10) decouples across rows, allowing for a separate nested structures across each endogenous series. This sparsity pattern is depicted in Fig. 5.

Most group lasso solution methods, such as block coordinate descent, take advantage of the separability of groups to improve the computational performance.

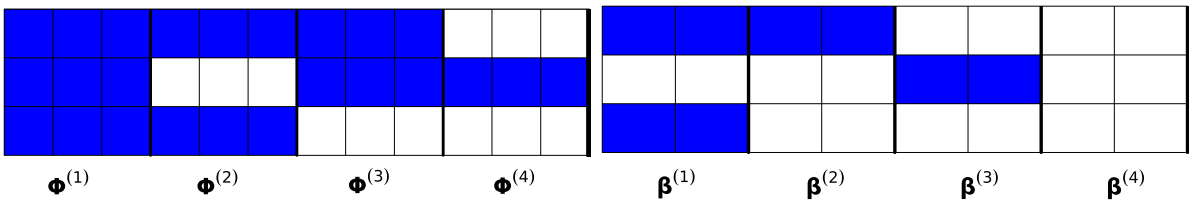

Fig. 5. Example sparsity pattern (active elements shaded) generated by an endogenous-first VARX-L$_{3,2}$(4, 4). Note that a row in $\boldsymbol{\beta}^{(\ell)}$ can only be nonzero if the corresponding row in $\boldsymbol{\Phi}^{(\ell)}$ is also nonzero.

Although the nested structure is not directly separable, its dual can be solved in one pass of block coordinate descent based on the methodology of Jenatton et al. (2011). Details of the solution approach and our algorithm are provided in Appendix A.3.5. A distinct hierarchical group penalty, with respect to coefficient lags, is considered in Nicholson, Bien, and Matteson (2014).

3. High-dimensional macroeconometrics

This section evaluates our regularization procedures in three macroeconomic data applications: two high-dimensional and one low-dimensional. Our first two applications involve applying the proposed VARX-L procedures to the widely-used set of US macroeconomic indicators originally constructed by Stock and Watson (2005). Our second example involves forecasting a small set of Canadian macroeconomic indicators, incorporating the previous US dataset as an exogenous series. Section 3.1 outlines the practical implementation of our penalty parameter selection procedure, Section 3.2 describes the benchmarks that we compare our models against, Section 3.3 details our macroeconomic applications, and Section 3.4 provides an expanded comparison of the relative forecasting performances of the competing models using the model confidence sets framework developed by Hansen, Lunde, and Nason (2011).

3.1. Practical implementation

The regularization parameter, λ, is not known in practice, and is typically chosen via cross-validation. This section details our strategy for selecting λ. Following Friedman et al. (2010), we select from a grid of potential penalty parameters that starts with the smallest value for which all components of $[\boldsymbol{\Phi}, \boldsymbol{\beta}]$ will be zero, then decreases in log-linear increments. This value differs for each procedure and can be inferred from the respective algorithm. The starting values are summarized in Table 17, located in Appendix A.5. Both the number of gridpoints, N, and the depth of the grid are left to user input. A deep grid and a large number of gridpoints result in increased computational costs, and often do not improve the forecasting performance. We have found that a grid depth of $\frac{1}{25}\lambda_{\max}$ and 10 gridpoints achieve an adequate forecast performance in most scenarios.

Due to the time-dependence, our problem is not well-suited to traditional K-fold cross-validation. Instead, following Banbura et al. (2009), we propose that the optimal penalty parameter be chosen by minimizing the h-step-ahead mean square forecast error (MSFE), in which $h = 1, 2, \ldots,$ denotes the desired forecast horizon. We divide the data into three periods: one for initialization, one for training, and one for forecast evaluation. Define time indices $T_1 = \left\lfloor \frac{T}{3} \right\rfloor, T_2 = \left\lfloor \frac{2T}{3} \right\rfloor$.

We start our validation process by fitting a model using all data up to time $t = T_1$ and forecasting $\hat{\mathbf{y}}^{\lambda_i}_{T_1+h}$ for $i = 1, \ldots, N$. We then add one observation at a time sequentially and repeat this process until time $t = T_2 - h$. This procedure is illustrated in Fig. 6.

We select $\hat{\lambda}$ as the minimizer of

$$\text{MSFE}(\lambda_i) = \frac{1}{(T_2 - T_1 - h + 1)} \sum_{t=T_1}^{T_2-h} \|\hat{\mathbf{y}}^{\lambda_i}_{t+h} - \mathbf{y}_{t+h}\|_F^2. \quad (11)$$

Finally, we evaluate the h-step-ahead forecast accuracy of $\hat{\lambda}$ from time origins $t = T_2 - h$ to $t = T - h$. Additional criterion functions can also be substituted if desired, but MSFE is the most natural criterion, given our use of the least squares objective function. Rather than parallelizing the cross-validation procedure, our approach uses the result from the previous period as an initialization or "warm start", which decreases the computation time substantially. The penalty parameter selection procedure is presented in Algorithm 2 in the Appendix.

3.1.1. Multi-step predictions

The VARX-L framework can accommodate multi-step-ahead forecasting easily, by modifying our solution algorithms. Essentially, direct h-step forecasts are computed by solving the standard VARX-L objective Eq. (3), leaving a gap of h observations.

$$\min_{\boldsymbol{\nu},\boldsymbol{\Phi},\boldsymbol{\beta}} \sum_{t=h}^{T} \Bigg\| \mathbf{y}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{p} \boldsymbol{\Phi}^{(\ell)} \mathbf{y}_{t-h-\ell+1} - \sum_{j=1}^{s} \boldsymbol{\beta}^{(j)} \mathbf{x}_{t-h-j+1} \Bigg\|_F^2 + \lambda \Big(\mathcal{P}_y(\boldsymbol{\Phi}) + \mathcal{P}_x(\boldsymbol{\beta}) \Big).$$

As per Clark and McCracken (2013), the direct h-step-ahead forecast can be calculated as

$$\hat{\mathbf{y}}_{t+h} = \hat{\boldsymbol{\nu}} + \hat{\boldsymbol{\Phi}}^{(1)} \mathbf{y}_t + \cdots + \hat{\boldsymbol{\Phi}}^{(p)} \mathbf{y}_{t-p+1} + \hat{\boldsymbol{\beta}}^{(1)} \mathbf{x}_t + \cdots + \hat{\boldsymbol{\beta}}^{(s)} \mathbf{x}_{t-s+1}.$$

The *iterative* approach, in which multi-step forecasts are computed recursively as 1-step-ahead forecasts using predicted values, is another popular technique for computing long-horizon forecasts. However, this approach cannot be

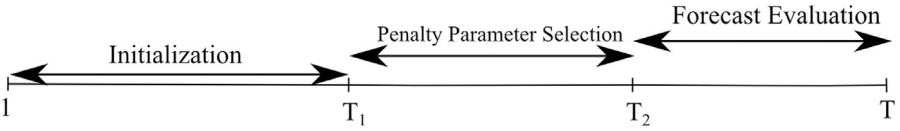

Fig. 6. Illustration of rolling cross-validation.

extended to the VARX setting directly because we do not forecast the exogenous series.

If iterative multi-step predictions are desired, one could instead fit the full VAR_{k+m}. However, as was shown by Marcellino, Stock, and Watson (2006), direct forecasts are more robust to model misspecification, making them more appropriate in high-dimensional settings. As our macroeconomic applications consider quarterly data, Section 3.3 computes both 1- and 4-step-ahead forecasts.

3.1.2. Selecting a structure

The VARX-L framework offers many different choices of structured penalties that are suitable for a wide range of applications. Under scenarios in which little is known about the potential dynamic dependence of the series included, the basic VARX-L makes no underlying structural assumptions.

The lag group and sparse lag group VARX-L structures are most appropriate when the endogenous series are closely related; for example, if the series of interest comprise unemployment rates segmented by state or census region. The own/other group and sparse own/other group VARX-L structures are most appropriate for macroeconomic applications in which a series' own lags are thought to have a substantially different temporal dependence from those of "other" series. As an example, one could consider a disparate group of series that are examined traditionally in small-scale macroeconometric forecasting applications: the US federal funds rate, the GDP growth rate, and the consumer price index. The endogenous-first structure is best suited to applications in which the forecasting effectiveness of the exogenous series is unknown.

One potential diagnostic tool involves fitting the sparse lag group VARX-L with both λ and α selected according to rolling cross-validation. Selecting a value of α that is close to zero is evidence of strong groupwise sparsity, while a value that is close to one indicates unstructured sparsity, and values in the middle provide evidence for some combination of the two.

In practice, since the computational time required to apply all procedures is manageable, we suggest fitting several VARX-L structures and selecting the approach that achieves the best out-of-sample forecasting performance.

3.2. Methods for comparison

A conventional VARX model selection approach in a low-dimensional setting involves fitting a $\text{VARX}_{k,m}(\ell, j)$ using least squares for $0 \leq \ell \leq p$, $0 \leq j \leq s$, and selecting lag orders for both the endogenous and exogenous series based on an information criterion, such as Akaike's information criterion (AIC) or the Bayesian information criterion (BIC). As per Lütkepohl (2005), the AIC and BIC of a $\text{VARX}_{k,m}(\ell, j)$ are defined as

$$\text{AIC}(\ell, j) = \log|\widehat{\Sigma}_u^{\ell,j}| + \frac{2\big(k(k\ell + mj)\big)}{T},$$

$$\text{BIC}(\ell, j) = \log|\widehat{\Sigma}_u^{\ell,j}| + \frac{\log(T)\big(k(k\ell + mj)\big)}{T},$$

where $\widehat{\Sigma}_u^{\ell,j}$ is the residual sample covariance matrix obtained from the estimated $\text{VARX}_{k,m}(\ell, j)$, and $|\Sigma|$ represents the determinant of Σ. The selected lag orders (ℓ, j) are then chosen as the values that minimize the AIC or BIC. The AIC penalizes each model coefficient uniformly by a factor of two whereas the BIC scales penalties relative to the series length. Hence, the BIC will tend to select more parsimonious models than the AIC when T is large.

We compare our methods with least squares model selection procedures that utilize the AIC and BIC for selecting lag orders. Since we are considering high-dimensional applications in which $\widehat{\Sigma}_u$ could be ill-conditioned, we construct our least squares estimates using a variation of the approach developed by Neumaier and Schneider (2001). This procedure constructs the least squares estimates using a QR decomposition, which obviates the need for explicit matrix inversion. In addition, following a heuristic proposed by Hansen (1998), we impose a ridge penalty: $\big((k \cdot \ell + m \cdot j)^2 + (k \cdot \ell + m \cdot j) + 1\big)\epsilon_{\text{machine}}$, scaled by the column norms of the lagged series $\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-p}, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-s}$, where $\epsilon_{\text{machine}}$ denotes machine precision. This penalty ensures that the determinant of $\widehat{\Sigma}_u$ is well-defined, but without noticeably impacting the calculation of the degrees of freedom. However, in very high-dimensional settings where $k\big(kp + ms\big) > kT$, the covariance Σ_u is not well-defined, meaning that information criterion based methods are not appropriate. Details of our information criterion minimization routine are provided by Nicholson, Matteson, and Bien (2016).

In addition, we compare our methods with two naïve approaches that provide insights in regard to the level of temporal dependence in the data. We first consider the unconditional *sample mean*, which will make h-step-ahead forecasts at time $t + h$ based on the average of all data observed up until time t: $\hat{\mathbf{y}}_{t+h} = \frac{1}{t}\sum_{i=1}^{t}\mathbf{y}_i$. A strong relative performance of the sample mean indicates that the sophisticated models of interest have low predictive power relative to a white noise process, which could occur for a number of reasons, including weak temporal dependence or dependence of a nonlinear nature.

Second, we consider the vector *random walk* model, which makes h-step-ahead forecasts based upon the most recent realization of the series, i.e., $\hat{\mathbf{y}}_{t+h} = \mathbf{y}_t$. A superior performance of the vector random walk indicates a high persistence or a strong degree of temporal dependence, as is often observed in macroeconomic data.

Finally, we compare our methods with two more sophisticated approaches. First, we consider the popular Bayesian VAR with the modified Minnesota prior proposed by Banbura et al. (2009) (henceforth BGR). Their approach is very similar to ridge regression, in that it shrinks the least squares coefficients toward zero, with the degree of regularization being determined by a single penalty parameter. This parameter is chosen using rolling cross-validation, as described in Section 3.1. Following Banbura et al. (2009), we fit a $\text{VAR}_{k+m}(p)$ instead of a $\text{VARX}_{k,m}(p, s)$, and select the regularization parameter as that which minimizes the h-step-ahead MSFE across the k endogenous series. This allows BGR's method to make forecasts by utilizing information from both the endogenous and exogenous series, and creates a direct comparison with our VARX-L framework.

BGR's approach modifies the Minnesota prior to make it computationally tractable in high dimensions, but does not return sparse solutions. The superior performance of the VARX-L methods relative to BGR's approach provides evidence on the importance of imposing sparsity in order to obtain accurate forecasts. Details of our implementation of BGR's procedure are provided in Appendix A.4.

In addition, we also provide a comparison with a factor model (Anderson, 1984; Stock & Watson, 2006). In a manner similar to principal components analysis, a factor model attempts to find a low-rank structure in the data that adequately accounts for a high percentage of the variation across the series. Our implementation forecasts using a reduced-rank structure that accounts for at least 95% of the total variance in the series included. For implementation details, consult (Ensor, 2013). We expect the factor model to perform well for scenarios in which a small subset of common factors drive the underlying dynamics of the larger cross-section of series modeled.

3.3. Macroeconometric applications

We evaluate our methods on the large and widely utilized macroeconomic dataset created by Stock and Watson (2005) and later amended by Koop (2013). The dataset consists of 168 quarterly US macroeconomic indicators that contain information about various aspects of the economy, including income, industrial production, employment, stock prices, interest rates, exchange rates, etc. It starts in quarter 2 of 1959 and finishes in quarter 3 of 2007 ($T = 195$). As per Koop (2013), the series can be categorized into several levels, of which we consider:

- *Small* ($k = 3$): three variables (federal funds rate, consumer price index, gross domestic product growth rate). This is the core group, which is typically used in simple dynamic stochastic generalized equilibrium models.
- *Medium* ($k = 20$): small plus 17 additional variables containing aggregated economic information (e.g., consumption, labor, housing, exchange rates).
- *Medium–large* ($k = 40$): medium plus 20 additional aggregate variables.
- *Large* ($k = 168$): medium–large plus 128 additional variables, consisting primarily of the components that make up the aggregated information.

For a detailed description of each set of variables, consult (Koop, 2013). As Banbura et al. (2009) found that the greatest improvements in forecast performance occurred with the *medium* VAR, we will initially attempt to forecast the *medium* set of indicators ($k = 20$) while using the additional variables from the *medium–large* category as exogenous predictors ($m = 20$).

Prior to estimation, each series is transformed to stationarity according to the specifications provided by Stock and Watson (2005) and standardized by subtracting the sample mean and dividing by the sample standard deviation. The period from quarter 2 of 1976 to quarter 3 of 1992 is used for penalty parameter selection while the period from quarter 4 of 1992 to quarter 3 of 2007 is used for forecast evaluation. The MSFE (relative to the sample mean) over the evaluation period for each model is reported in Table 2. In addition, Table 3 records the average *sparsity ratio* for each VARX-L model. The sparsity ratio is the average proportion of model coefficients that are set to zero across the evaluation period. A sparsity ratio of one implies that all coefficients are set to zero, whereas a least squares fit that includes all coefficients has a sparsity ratio of zero.

Most of our VARX-L procedures outperform the benchmarks substantially at both forecast horizons, with the own/other group VARX-L and sparse own/other group VARX-L achieving the best performances. This illustrates that making the distinction between a series' own lags and those of other series can provide improved forecasts in macroeconomic applications. The relatively poor performance of the lag group VARX-L suggests that a lag-based grouping may be too restrictive for such a disparate group of series, and hence, not appropriate for this application.

The imposition of sparsity appears to be crucial, as BGR's Bayesian VAR performs worse than all of the VARX-L procedures at both horizons except for the lag group VARX-L at $h = 4$. It performs very similarly to the least squares VARX with lags selected by the BIC at $h = 1$, but slightly better at $h = 4$. The factor model also performs very poorly at both forecast horizons, indicating that a low-rank structure may not be appropriate for this application.

The VARX with lags selected by the AIC is outperformed substantially by the sample mean at both horizons, whereas the VARX with lags selected by the BIC outperforms it slightly at $h = 1$, but is outperformed at $h = 4$. Since the AIC imposes a weaker penalty for higher lag orders than the BIC, it has a tendency to construct overparameterized models, whereas the BIC has a tendency to underfit, thus missing out on potential dynamic relationships that the VARX-L procedures are able to capture. Since neither approach imposes variable selection, these models tend to result in very noisy multi-step-ahead forecasts.

We observe that the basic VARX-L imposes the most sparsity at both horizons, as expected, since it is not bound to any group structure. The sparse own/other group VARX-L, which has a very flexible group structure as well as the ability to impose within-group sparsity, returns the second most sparse model across both horizons. The

Table 2
One- and four-step-ahead MSFEs of $k = 20$ macroeconomic indicators (relative to the sample mean) with $m = 20$ exogenous predictors $p = 4, s = 4$.

Model/VARX-L penalty structure	One-step-ahead out-of-sample relative MSFE	Four-step-ahead out-of-sample relative MSFE
Basic	0.8064	0.9672
Lag group	0.8747	0.9798
Own/other group	0.7773	0.9582
Sparse lag group	0.8206	0.9702
Sparse own/other group	0.7823	0.9590
Endogenous-first	0.8531	0.9748
VARX with lags selected by the AIC	5.0223	7.8363
VARX with lags selected by the BIC	0.9455	1.1603
BGR's Bayesian VAR	0.9414	0.9765
Factor model	0.9904	0.9819
Sample mean	1.0000	1.0000
Random walk	1.9909	1.8706

Table 3
Average sparsity ratios (proportions of least squares coefficients set to zero) of VARX-L models for forecasting $k = 20$ macroeconomic indicators with $m = 20$ exogenous predictors $p = 4, s = 4, k^2p + kms = 3200$.

VARX-L penalty structure	One-step-ahead average sparsity ratio	Four-step-ahead average sparsity ratio
Basic	0.9090	0.9636
Lag group	0.2259	0.6315
Own/other group	0.3049	0.7030
Sparse lag group	0.6743	0.7095
Sparse own/other group	0.7118	0.8853
Endogenous-first	0.1411	0.6068

lag group VARX-L, the own/other group VARX-L and the endogenous-first VARX-L, all of which lack the ability to impose within-group sparsity, return relatively less sparse solutions at $h = 1$, but impose much more sparsity at $h = 4$, which could reflect the decreasing predictive power of the data for forecasting at longer horizons.

As an additional application to showcase the tractability of our methods on high-dimensional problems, we attempt to forecast the *medium–large* ($k = 40$) set of indicators, using the remaining variables in the *large* ($m = 128$) set as exogenous predictors. The forecast performance is recorded in Table 4, and the average sparsity ratios of the VARX-L procedures in Table 5.

In this application, we observe that the own/other group VARX-L achieves the best forecasting performance at $h = 1$, followed by the basic VARX-L. At $h = 4$, the sparse lag group VARX-L achieves the best performance. As in the previous application, all of the VARX-L procedures outperform the benchmark models at $h = 1$. However, BGR's Bayesian VAR achieves the second-best forecasting performance at $h = 4$.

Most of our VARX-L methods impose considerably more sparsity in this larger application than in Table 3, which is to be expected, as many of the exogenous series included are probably either irrelevant or redundant for forecasting purposes. Strangely, the endogenous-first VARX-L imposes less sparsity than in the previous application, but this could be due to its unique nested penalty structure, which forces endogenous coefficients to be active along with their exogenous counterparts. It is possible that the relatively poor performances of several VARX-L models at $h = 4$ are the result of not imposing enough sparsity. For example, the sparse own/other group VARX-L imposes less sparsity at $h = 4$ than at $h = 1$.

3.3.1. Canadian macroeconomic data application

We next consider a low-dimensional application in which we forecast Canadian indicators using US macroeconomic series as exogenous predictors. As a small, relatively open economy, Canada's macroeconomic indicators have been shown to be very sensitive to their US counterparts. In particular, Cushman and Zha (1997) and Racette and Raynauld (1992) demonstrate that the US gross domestic product and federal funds rate are very influential in modeling Canada's analogous monetary policy proxy variables. Taking this into consideration, we forecast $k = 4$ Canadian macroeconomic series using our previously defined *medium* dataset as exogenous predictors ($m = 20$). The endogenous series are Canadian M1 (a measure of the liquid components of money supply), Canadian industrial production, Canadian GDP (relative to 2000), and the Canada/US exchange rate.

The Canadian series cover the period from quarter 3 of 1960 to quarter 3 of 2007. The period from quarter 3 of 1977 to quarter 2 of 1993 is used for penalty parameter selection, while the period from quarter 3 of 1993 to quarter 3 of 2007 is used for forecast evaluation ($T = 191$). In addition to the standard benchmarks, we also compare against our procedures in the VAR-L framework, in which the exogenous predictors are ignored. Our results are summarized in Table 6 and Table 7.

Even in this low dimension, we still find that all of our models outperform the AIC and BIC benchmarks substantially across both forecast horizons, with the sparse

Table 4
One- and four-step-ahead MSFEs of $k = 40$ macroeconomic indicators (relative to the sample mean) with $m = 128$ exogenous predictors $p = 4, s = 4$.

Model/VARX-L penalty structure	One-step-ahead out-of-sample relative MSFE	Four-step-ahead out-of-sample relative MSFE
Basic	0.7810	0.9824
Lag group	0.8550	0.9756
Own/other group	0.7587	0.9680
Sparse lag group	0.7899	0.9575
Sparse own/other group	0.7917	0.9952
Endogenous-first	0.8278	0.9657
VARX with lags selected by the AIC	1.9789	2.5105
VARX with lags selected by the BIC	0.9999	0.9999
BGR's Bayesian VAR	0.9062	0.9646
Factor model	0.9758	1.0166
Sample mean	1.0000	1.0000
Random walk	1.6763	1.6763

Table 5
Average sparsity ratios (proportions of least squares coefficients set to zero) of VARX-L models for forecasting $k = 40$ macroeconomic indicators with $m = 128$ exogenous predictors $p = 4, s = 4, k^2p + kms = 26{,}880$.

VARX-L penalty structure	One-step-ahead average sparsity ratio	Four-step-ahead average sparsity ratio
Basic	0.9683	0.9862
Lag group	0.6300	0.9584
Own/other group	0.8144	0.9257
Sparse lag group	0.8577	0.8711
Sparse own/other group	0.9912	0.9825
Endogenous-first	0.0702	0.1697

Table 6
One- and four-step-ahead MSFEs (relative to the sample mean) for the VARX forecasts of $k = 4$ Canadian macroeconomic indicators with $m = 20$ exogenous predictors $p = 4, s = 4$, and the VAR forecasts of four Canadian macroeconomic indicators, $p = 4$.

Model/VARX-L penalty structure	One-step-ahead out-of-sample RMSFE	Four-step-ahead out-of-sample RMSFE
Basic	0.8406	0.9187
Lag group	0.8357	0.9285
Own/other group	0.8376	0.9143
Sparse lag group	0.8274	0.9129
Sparse own/other group	0.8327	0.9181
Endogenous-first	0.8454	0.9593
VARX with lags selected by the AIC	1.3680	1.7739
VARX with lags selected by the BIC	0.8785	1.0941
BGR's Bayesian VAR (with exogenous series)	1.0058	0.9748
Model/VAR-L penalty structure	**One-step-ahead out-of-sample RMSFE**	**Four-step-ahead out-of-sample RMSFE**
Basic	0.8465	0.9645
Lag group	0.8575	0.9965
Own/other group	0.8491	0.9604
Sparse lag group	0.8506	0.9623
Sparse own/other group	0.8493	0.9655
VAR with lag selected by the AIC	0.9190	1.1365
VAR with lag selected by the BIC	0.8785	1.0941
BGR's Bayesian VAR (without exogenous series)	1.0066	0.9891
Factor model	0.9033	1.0152
Sample mean	1.0000	1.0000
Random walk	1.3388	1.7180

lag group VARX-L achieving superior performances at both horizons. This low-dimensional example is better suited to lag-based groupings than our previous application was. Consequently, the relative forecasting performances of

Table 7
Average sparsity ratios (proportions of least squares coefficients set to zero) of VARX-L models in forecasting $k = 4$ macroeconomic indicators with $m = 20$ exogenous predictors $p = 4$, $s = 4$, $k^2p + kms = 384$, and VAR-L models lacking exogenous series $p = 4$, $k^2p = 64$.

Model/VARX-L penalty structure	One-step-ahead average sparsity ratio	Four-step-ahead average sparsity ratio
Basic	0.7882	0.9091
Lag group	0.5953	0.7733
Own/other group	0.5909	0.8125
Sparse lag group	0.7282	0.7979
Sparse own/other group	0.6764	0.8554
Endogenous-first	0.0005	0.0236
Model/VAR-L penalty structure	**One-step-ahead average sparsity ratio**	**Four-step-ahead average sparsity ratio**
Basic	0.6151	0.7958
Lag group	0.0000	0.8403
Own/other group	0.0002	0.6104
Sparse lag group	0.2114	0.3114
Sparse own/other group	0.2494	0.7000

the lag group VARX-L and the sparse lag group VARX-L improve substantially.

In addition, we find that our methods are able to use relevant information from the exogenous predictors effectively, as every VARX-L procedure achieves a better out-of-sample performance than its corresponding VAR-L. Conversely, the information criterion based VARX approaches fail to outperform their VAR counterparts. At $h = 1$ and $h = 4$, BIC produces identical forecast errors in both the VAR and VARX settings, indicating that it never selects any exogenous series.

BGR's Bayesian VAR performs poorly in this scenario, achieving a forecast performance similar to that of the sample mean across both horizons, with or without exogenous series, outperforming only the lag group VAR-L at $h = 4$. Its poor performances in both settings suggest that it is desirable to impose sparsity even in low-dimensional applications. The factor model also performs poorly, indicating that a low-rank structure may not be appropriate for this small-scale macroeconomic forecasting application.

We observe that less sparsity is imposed in the VAR framework than in the VARX. At $h = 1$, the lag group VAR-L imposes no sparsity and the own/other group VAR-L imposes only a negligible amount. This relative lack of sparsity is likely to be due to the fact that the VAR-L models already operate from a reduced model space relative to their VARX-L counterparts. However, considerably more sparsity is imposed at $h = 4$ under both frameworks. Across all models, we find that the level of sparsity generally increases with the number of potential coefficients, as well as with longer forecast horizons.

3.4. Evaluating model performance with model confidence sets

In addition to simply evaluating model performances based on relative MSFEs, we also consider applying the model confidence sets framework proposed by Hansen et al. (2011), which conducts a sequence of pairwise hypothesis tests in order to construct a set of "superior models", within which the null hypothesis of *equal predictive ability* cannot be rejected.

This procedure starts by computing the sum of squared forecast errors (SSFE) for each candidate model over the evaluation period ($T_2 + 1$ to T). The SSFE for model i at time t is defined as $\text{SSFE}_{i,t} = \|\mathbf{y}_t - \hat{\mathbf{y}}_t^{(i)}\|_F^2$. After constructing the SSFE, the MCS procedure computes the pairwise loss differential at each time t for each pair of models $i < j$: $d_{i,j,t} = \text{SSFE}_{i,t} - \text{SSFE}_{j,t}$.

Relative model performance is assessed according to the hypothesis of equal predictive ability, which tests whether the pairwise loss averaged over time differs from zero across all model combinations. We evaluate this hypothesis by constructing the test statistic

$$v_{i,j} = \frac{\bar{d}_{ij}}{\sqrt{\widehat{\text{var}}(\bar{d}_{ij})}}, \tag{12}$$

in which $\bar{d}_{ij} = \frac{1}{T-T_2}\sum_{T_2+1}^{T} d_{i,j,t}$, and $\widehat{\text{var}}(\bar{d}_{ij})$ is estimated according to a block bootstrap procedure. The asymptotic distribution of this test statistic is non-standard, and therefore is also estimated using a block bootstrap in a manner similar to the variance.

The model confidence sets algorithm is initialized by setting M equal to all candidate models and testing for equal predictive ability iteratively using the aggregate test statistic $V_{R,M} = \max_{i,j\in M} |v_{ij}|$. If equal predictive ability is rejected at a given confidence level $1 - \alpha$, the worst performing model (i.e., the model with the largest average pairwise loss differential) is removed from M and the procedure is repeated on the reduced subset of models. The algorithm terminates once equal predictive ability cannot be rejected. For more details on the MCS methodology, consult Bernardi and Catania (2014) and Hansen et al. (2011).

We implement this procedure using the `R` package `MCS` (Bernardi & Catania, 2014). Following the package's default settings, we choose $\alpha = 0.15$ and perform 5000 bootstrap replications. Our resulting sets of equal predictive ability are displayed in Table 8.

Table 8
Model sets M of equal predictive ability ($\alpha = 0.15$). Within each set of models, we cannot reject the null hypothesis of equal predictive ability, though they achieve superior forecasting performances relative to all excluded models.

Application	One-step M	Four-step M
Stock and Watson (2005) US macroeconomic data ($k = 20, m = 20$)	Own/other VARX-L Sparse own/other VARX-L	Own/other VARX-L Sparse own/other VARX-L
Stock and Watson (2005) US macroeconomic data ($k = 40, m = 128$)	Own/other VARX-L	Sparse lag group VARX-L Own/other VARX-L Endogenous-first VARX-L BGR's Bayesian VAR
Canadian macroeconomic data ($k = 4, m = 20$)	Sparse own/other VARX-L Sparse lag group VARX-L	Sparse own/other VARX-L Basic VARX-L Sparse lag group VARX-L Own/other VARX-L

We find that the MCS procedure is able to distinguish not only between the forecasting performances of our VARX-L models and those of the benchmarks, but also between those within the VARX-L class of models. In the Canadian macroeconomic data application, no VAR-L models are included at either forecast horizon, indicating that the exogenous series have predictive power. In addition, we also find that either the sparse own/other group VARX-L or the own/other group VARX-L is in the MCS in every application. This provides further evidence supporting the use of a group structure that distinguishes between a series' own lags and those of other series in macroeconomic applications. The only application in which a competitor's model is included in the MCS is the large US macroeconomic application ($k = 40, m = 128$) at $h = 4$, though several VARX-L models are also included. The relatively poor performance of the VARX-L models in this scenario suggests that long-horizon forecasts in large models deserve additional scrutiny.

4. Extending the VARX-L to unit-root non-stationarity

In some scenarios, it may not be appropriate to shrink every coefficient toward zero. In traditional time series analysis, economic series that exhibit persistence are transformed to stationarity. However, this framework has several drawbacks. This process can be labor-intensive and subjective unless pre-established transformation guidelines are available. Second, as was stated by Kennedy (2003), stationarity transformations destroy information about the long-run relationships of economic variables. Ideally, it would be preferable to work with the untransformed series directly, in order to forecast effectively using all available information. This section outlines a possible extension that allows for shrinkage toward reference models, such as a vector random walk, that can account for mild non-stationarity, which is ubiquitous in macroeconomic data.

4.1. The "Minnesota" VARX-L

The proposed VARX-L models can be modified easily to shrink coefficients toward a known constant matrix. Shrinking toward constant matrices $\mathbf{C}_y \in \mathbf{R}^{k\times kp}, \mathbf{C}_x \in \mathbf{R}^{k\times ms}$ results in a slightly modified objective of the form

$$\min_{\boldsymbol{\nu},\boldsymbol{\Phi},\boldsymbol{\beta}} \sum_{t=1}^{T} \|\mathbf{y}_t - \boldsymbol{\nu} - \boldsymbol{\Phi}\mathbf{Y}_{t-1} - \boldsymbol{\beta}\mathbf{X}_{t-1}\|_F^2 + \lambda\left(\mathcal{P}_y(\boldsymbol{\Phi} - \mathbf{C}_y) + \mathcal{P}_x(\boldsymbol{\beta} - \mathbf{C}_x)\right), \tag{13}$$

in which $\mathbf{Y}_t = [\mathbf{y}_t^\top, \ldots, \mathbf{y}_{t-p}^\top]$ and $\mathbf{X}_t = [\mathbf{x}_t^\top, \ldots, \mathbf{x}_{t-s}^\top]$.

Let $[\boldsymbol{\Phi}, \boldsymbol{\beta}]^\lambda(\mathbf{C}_y, \mathbf{C}_x)$ denote a solution to this problem. Now, by a change of variables $\widetilde{\boldsymbol{\Phi}} = \boldsymbol{\Phi} - \mathbf{C}_y$ and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} - \mathbf{C}_x$, we obtain the equivalent problem

$$\min_{\boldsymbol{\nu},\widetilde{\boldsymbol{\Phi}},\widetilde{\boldsymbol{\beta}}} \sum_{t=1}^{T} \|\mathbf{y}_t - \boldsymbol{\nu} - \mathbf{C}_y\mathbf{Y}_{t-1} - \widetilde{\boldsymbol{\Phi}}\mathbf{Y}_{t-1} - \mathbf{C}_x\mathbf{X}_{t-1} - \widetilde{\boldsymbol{\beta}}\mathbf{X}_{t-1}\|_F^2 + \lambda\left(\mathcal{P}_y(\widetilde{\boldsymbol{\Phi}}) + \mathcal{P}_x(\widetilde{\boldsymbol{\beta}})\right),$$

which can be expressed as

$$\min_{\boldsymbol{\nu},\widetilde{\boldsymbol{\Phi}},\widetilde{\boldsymbol{\beta}}} \sum_{t=1}^{T} \|\widetilde{\mathbf{y}}_t - \boldsymbol{\nu} - \widetilde{\boldsymbol{\Phi}}\mathbf{Y}_{t-1} - \widetilde{\boldsymbol{\beta}}\mathbf{X}_{t-1}\|_F^2 + \lambda\left(\mathcal{P}_y(\widetilde{\boldsymbol{\Phi}}) + \mathcal{P}_x(\widetilde{\boldsymbol{\beta}})\right),$$

where $\widetilde{\mathbf{y}}_t = \mathbf{y}_t - \mathbf{C}_y\mathbf{Y}_{t-1} - \mathbf{C}_x\mathbf{X}_{t-1}$. We can view the solution to this transformed problem as $[\widetilde{\boldsymbol{\Phi}}, \widetilde{\boldsymbol{\beta}}]^\lambda(\mathbf{0}, \mathbf{0})$ operating on $\widetilde{\mathbf{y}}_t$. Hence, transforming back to the setting of Eq. (13), we find that

$$[\boldsymbol{\Phi}, \boldsymbol{\beta}]^\lambda(\mathbf{C}_y, \mathbf{C}_x) = [\mathbf{C}_y, \mathbf{C}_x] + [\widetilde{\boldsymbol{\Phi}}, \widetilde{\boldsymbol{\beta}}]^\lambda(\mathbf{0}_{k\times kp}, \mathbf{0}_{k\times ms}).$$

As an example, consider $\mathbf{C}_y = [\mathbf{I}_k, \mathbf{0}_{k\times k}, \ldots, \mathbf{0}_{k\times k}]$, $\mathbf{C}_x = \mathbf{0}_{k\times ms}$, which implements a variant of the Minnesota prior, shrinking the VARX-L model toward a vector random walk. We refer to this extension as the "Minnesota" VARX-L. It could be very useful in economic applications, as it is believed widely that many persistent macroeconomic time series can be approximated well by a random walk (Litterman, 1979).

We validate this alternative approach by following the methodology of Banbura et al. (2009), who also utilize the data from Stock and Watson (2005), but eschew stationarity transformations and work with the untransformed

Table 9
One- and four-step-ahead MSFEs (relative to a random walk) for $k = 20$ non-stationary macroeconomic indicators with $m = 20$ exogenous predictors which shrink toward a vector random walk.

Model/Minnesota VARX-L penalty structure	One-step-ahead out-of-sample RMSFE	Four-step-ahead out-of-sample RMSFE
Basic	0.8173	0.9460
Lag group	0.9450	0.9590
Own/other group	0.8155	0.9520
Sparse lag group	0.9858	0.9702
Sparse own/other group	0.8808	0.9550
Endogenous-first	0.9746	0.9518
VARX with lag selected by the AIC	1.2764	1.1896
VARX with lag selected by the BIC	1.2764	1.1896
BGR's Bayesian VAR	1.3475	1.0083
Factor model	4.7979	1.5968
Sample mean	11.304	5.7747
Random walk	1.0000	1.0000

Table 10
Average sparsity ratios (proportions of least squares coefficients set to zero) of VARX-L models in forecasting $k = 20$ non-stationary macroeconomic indicators with $m = 20$ exogenous predictors $p = 4, s = 4, k^2p + kms = 3200$.

VARX-L penalty structure	One-step-ahead average sparsity ratio	Four-step-ahead average sparsity ratio
Basic	0.9746	0.9652
Lag group	0.7650	0.7595
Own/other group	0.7602	0.7533
Sparse lag group	0.8484	0.8842
Sparse own/other group	0.8004	0.8099
Endogenous-first	0.7594	0.7875

series directly. We again apply our VARX-L forecasting procedures by forecasting the aforementioned *medium* set of ($k = 20$) series using the remaining 20 variables in the *medium large* set as exogenous predictors, but choose not to perform any stationarity transformations, instead shrinking toward a vector random walk.

One advantage of not applying stationarity transformations is that it allows us to utilize more of our data. The data used in Section 3.3 extend to quarter 4 of 2008, but one series, non-borrowed depository institutional reserves (FMRNBA), becomes negative in early 2008, due in part to changes in both monetary policy and Federal Reserve accounting (Ip, 2008). The stationarity transformation guidelines provided by Stock and Watson (2005) for this series propose taking the first difference of logs, which is obviously not appropriate for negative values.

The period from quarter 3 of 1976 to quarter 2 of 1993 is used for penalty parameter selection, while the period from quarter 3 of 1993 to quarter 4 of 2008 is used for forecast evaluation ($T = 200$). In this application, we also shrink BGR's Bayesian VAR toward a random walk. Our results are summarized in Table 9, and the average sparsity ratios are recorded in Table 10.

We find that each of these Minnesota VARX-L procedures outperforms the random walk at both forecast horizons, with the own/other group Minnesota VARX-L achieving the best out-of-sample performance at $h = 1$ and the basic VARX-L performing the best at $h = 4$.

We observe that the choice of structure affects the forecasting performance substantially under this scenario. Lag-based groupings, such as the lag group, sparse lag group, and endogenous-first perform relatively poorly at $h = 1$ but improve slightly relative to other methods at $h = 4$; however, they still outperform the naïve methods at both horizons. This reduction in their relative performance is likely to be due to their inability to distinguish between the diagonal random walk component and the coefficients on other lags in the lag one coefficient matrix $\mathbf{\Phi}^{(1)}$.

AIC and BIC are not well suited to a non-stationarity setting, and hence are completely uninformative, selecting lag orders of $p = 1$ and $s = 0$ at every point in time across both horizons. Despite the imposition of a random walk prior, BGR's procedure produces forecasts that are inferior to those of both the VARX-L procedures and the naïve random walk. The factor model is also poorly suited to this application and is outperformed substantially by the random walk.

A considerable amount of sparsity is imposed across forecast horizons, regardless of the structure, with the basic VARX-L returning the most sparse models. This indicates that a greater level of sparsity may be more appropriate when shrinking toward a reference model as opposed to shrinking identically toward zero.

5. Simulation scenarios

This section evaluates the forecasting performances of our procedures on several simulated multivariate time series conforming to different sparsity patterns, with one constructed to be advantageous for each proposed

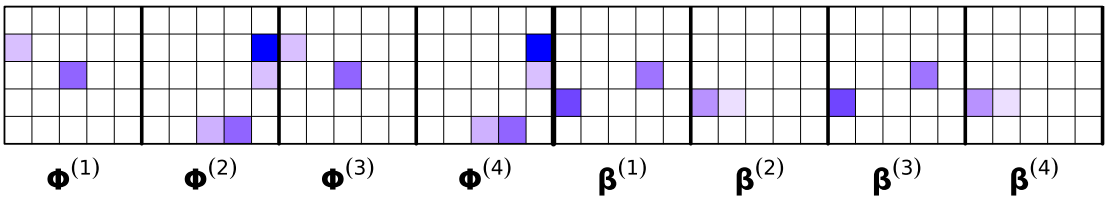

Fig. 7. Sparsity pattern scenario 1: Unstructured sparsity. Darker shading represents coefficients that are larger in magnitude.

Table 11
Out-of-sample MSFEs of one-step-ahead forecasts after 100 simulations: Scenario 1.

Model/VARX-L penalty structure	MSFE	MSFE relative to sample mean
Basic	0.0645 (0.0012)	0.0454
Lag group	0.0755 (0.0010)	0.0532
Own/other group	0.0734 (0.0010)	0.0517
Sparse lag group	0.0724 (0.0009)	0.0510
Sparse own/other group	0.0699 (0.0009)	0.0492
Endogenous-first	0.0779 (0.0010)	0.0549
VARX with lags selected by the AIC	0.1040 (0.0017)	0.0733
VARX with lags selected by the BIC	0.1183 (0.0032)	0.0833
BGR's Bayesian VAR	0.3675 (0.0124)	0.2590
Sample mean	1.4187 (0.0681)	1.0000
Random walk	0.8416 (0.0272)	0.5932

Note: Standard errors are shown in parentheses.

structure. Note that we omit the factor model from this section since it is designed to forecast well in low-rank regimes rather than sparse regimes, and is therefore anticipated to perform poorly.

Our objective is to quantify the relative performances under both matched and unmatched model sparsity and penalty function structures. All simulations operate on a $\text{VARX}_{5,5}(4, 4)$ of length $T = 100$, and each simulation is repeated 100 times. The choice of $p = s = 4$ was selected because it represents one year of dependence for quarterly series, which is a common frequency of macroeconomic data. The middle third of the dataset is used for penalty parameter selection, while the last third is used for forecast evaluation. Under the first five scenarios, Σ_u is distributed according to a multivariate normal distribution with mean $\mathbf{0}_5$ and covariance $(0.01) \times \boldsymbol{I}_5$; the sixth scenario utilizes a more general specification. We do not include intercepts in any of the simulation scenarios. The coefficient matrix from each simulation scenario was designed to ensure that a stationary process would be generated. This procedure is elaborated upon by Nicholson et al. (2016).

We simulate from a $\text{VARX}_{5,5}(4, 4)$ by first constructing a $\text{VAR}_{10}(4)$. Denoting the first five series by $\mathbf{y}_t$ and the second five by $\mathbf{x}_t$, we simulate according to the unidirectional relationship

$$\begin{pmatrix}\mathbf{y}_t\\ \mathbf{x}_t\end{pmatrix} = \sum_{\ell=1}^{4}\begin{pmatrix}\boldsymbol{\Phi}^{(\ell)} & \boldsymbol{\beta}^{(\ell)}\\ \mathbf{0} & \boldsymbol{\Gamma}^{(\ell)}\end{pmatrix}\begin{pmatrix}\mathbf{y}_{t-\ell}\\ \mathbf{x}_{t-\ell}\end{pmatrix} + \mathbf{u}_t,$$

where $\boldsymbol{\Gamma}^{\ell} \in \mathbb{R}^{m\times m}$ denotes the dependence structure of the exogenous series $\mathbf{x}_t$ (which follows the same sparsity pattern as $\boldsymbol{\Phi}^{\ell}$), and $\mathbf{u}_t \overset{\text{i.i.d.}}{\sim} N(\mathbf{0}, \Sigma_u)$. The residual covariance Σ_u is set to $0.01 \times I_{10}$ for the first five scenarios, while the covariance used in Scenario 6 is depicted in Fig. 12.

5.1. Scenario 1: unstructured sparsity

We begin by considering a scenario in which the sparsity is completely random; our sparsity pattern was generated in such a manner that each coefficient was given an equal probability (10%) of being active, resulting in a coefficient matrix in which roughly 90% of the coefficients are zero. Under such a design, we would expect a superior performance from the basic VARX-L, which assumes no group structure. We do not expect such a structure to be a common occurrence in macroeconomic applications, but it may be present in other application areas, such as internet traffic, in which the series included can vary substantially and probably will not exhibit any group structure. This sparsity pattern is depicted in Fig. 7, and the results are summarized in Table 11.

In this scenario, as expected, we find that the basic VARX-L achieves the best performance. Of the structured methods, the sparse own/other VARX-L performs the best, as it partially accommodates this sparsity pattern. Also as expected, the other approaches, which impose a structure that is not present in the data, suffer from degraded forecasts, but all of the structured approaches outperform the AIC and BIC benchmarks substantially. BGR's Bayesian VAR, which cannot perform either variable or lag order selection, achieves a substantially worse forecast performance than either of the information criterion based methods.

5.2. Scenario 2: lag group sparsity

We next consider a scenario in which $\boldsymbol{\Phi}^{(4)}$ and $\boldsymbol{\beta}^{(4)}$ are dense, with coefficients of the same magnitude, and all of the other coefficients are set to zero. Such a sparsity pattern may be found in disaggregated macroeconomic

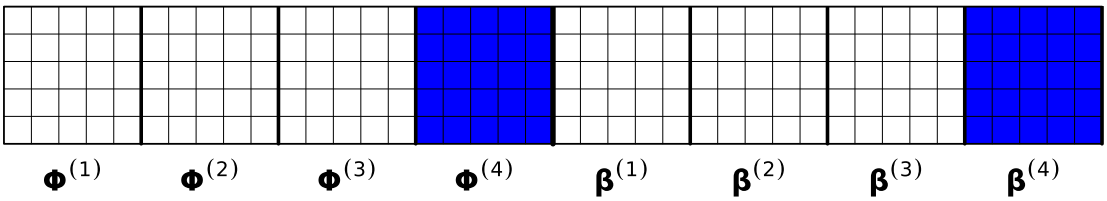

Fig. 8. Sparsity pattern scenario 2: Lag group sparsity.

Table 12
Out-of-sample MSFEs of one-step-ahead forecasts after 100 simulations: Scenario 2.

Model/VARX-L penalty structure	MSFE	MSFE relative to sample mean
Basic	0.0786 (0.0012)	0.1397
Lag group	0.0709 (0.0011)	0.1260
Own/other group	0.0713 (0.0011)	0.1268
Sparse lag group	0.0739 (0.0012)	0.1314
Sparse own/other group	0.0742 (0.0011)	0.1319
Endogenous-first	0.0720 (0.0011)	0.1280
VARX with lags selected by the AIC	1.0084 (0.0273)	1.7933
VARX with lags selected by the BIC	0.9927 (0.0282)	1.7654
BGR's Bayesian VAR	0.5769 (0.0146)	1.0259
Sample mean	0.5623 (0.0123)	1.0000
Random walk	1.1279 (0.0322)	2.0058

Note: Standard errors are shown in parentheses.

series, such as agricultural price indices, which follow a purely seasonal autoregressive relationship and exhibit a substantial degree of cross-dependence. Under such a design, we should expect a superior performance from the lag group VARX-L, which partitions all coefficients within a lag into the same group. This sparsity pattern is depicted in Fig. 8, and the results are summarized in Table 12.

As expected, we find that the lag group VARX-L achieves the best performance, and all of the structured approaches outperform the basic VARX-L. Under this scenario, all of the VARX-L procedures offer substantial improvements over the benchmarks. This is probably a result of their ability to effectively take advantage of the strong signal from the exogenous predictors. Note that although the AIC and BIC benchmarks utilize this exogenous information, they are restricted to selecting from models of sequentially increasing lag orders; hence, they cannot accommodate this sparsity pattern and probably overfit the data, resulting in a performance that is comparable to that of a random walk. BGR's Bayesian VAR improves upon the information criterion based benchmarks, but performs substantially worse than any of the VARX-L methods because it cannot perform variable selection.

5.3. Scenario 3: structured lagwise sparsity, unstructured within-lag

Our third scenario can be thought of as a hybrid of Scenarios 1 and 2. As in Scenario 2, certain coefficient matrices are set identically to zero, with only the matrices $\mathbf{\Phi}^{(1)}$, $\mathbf{\Phi}^{(4)}$, $\boldsymbol{\beta}^{(1)}$, and $\boldsymbol{\beta}^{(4)}$ containing nonzero coefficients. In addition, the sparsity within each lag is unstructured, in a manner similar to Scenario 1. This scenario can be viewed as a less restrictive and more realistic version of the structure presented in Scenario 2, as it allows the degree of cross-dependence to vary across components. In such a scenario, we would expect procedures that allow for within-group sparsity, such as the sparse lag group VARX-L and the basic VARX-L, to achieve the best forecast performances. This sparsity pattern is depicted in Fig. 9 and the results are summarized in Table 13.

Under this scenario, the basic VARX-L achieves the best performance, followed closely by the sparse lag group VARX-L. Unlike Scenario 2, the information criterion based benchmarks are able to capture a portion of the true underlying structure in both the endogenous and exogenous series, since this structure exhibits dependence in the first lag; thus, they outperform the naïve benchmarks substantially. However, since they cannot account for within-lag sparsity, they are still outperformed considerably by all of the VARX-L methods. As in Scenario 1, BGR's Bayesian VAR performs very poorly, since it cannot perform either variable or lag order selection.

5.4. Scenario 4: sparse and diagonally dominant

Our final scenario consists of a diagonally-dominant sparsity structure, in which all of the diagonal elements of $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(4)}$ are equal in magnitude, whereas all of the off-diagonal endogenous coefficients are set to zero. As in scenario 2, the coefficients in $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(4)}$ are identical in magnitude. This structure incorporates the belief posited by Litterman (1986a) that macroeconomic series' own lags are more informative in forecasting applications than those of other series. Under this setting, one would expect superior performances from the own/other group VARX-L.

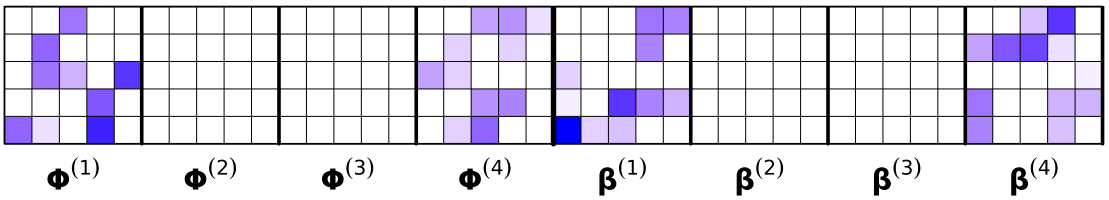

Fig. 9. Sparsity pattern scenario 3: Structured lagwise, unstructured within lag.

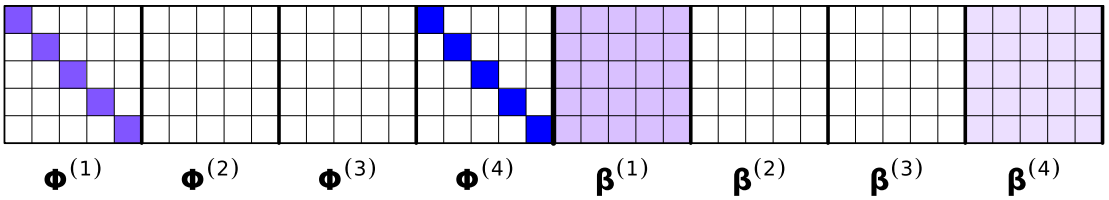

Fig. 10. Sparsity pattern scenario 4: Sparse and diagonally dominant.

Table 13
Out-of-sample MSFEs of one-step-ahead forecasts after 100 simulations: Scenario 3.

Model/VARX-L penalty structure	MSFE	MSFE relative to sample mean
Basic	0.0665 (0.0008)	0.1258
Lag group	0.0696 (0.0008)	0.1317
Own/other group	0.0699 (0.0009)	0.1322
Sparse lag group	0.0677 (0.0008)	0.1281
Sparse own/other group	0.0683 (0.0008)	0.1293
Endogenous-first	0.0711 (0.0009)	0.1345
VARX with lags selected by the AIC	0.1300 (0.0019)	0.2458
VARX with lags selected by the BIC	0.2501 (0.0061)	0.4730
BGR's Bayesian VAR	0.7568 (0.0515)	1.4314
Sample mean	0.5287 (0.0275)	1.0000
Random walk	1.3000 (0.0731)	2.4588

Note: Standard errors are shown in parentheses.

Table 14
Out-of-sample MSFEs of one-step-ahead forecasts after 100 simulations: Scenario 4.

Model/VARX-L penalty structure	MSFE	MSFE relative to sample mean
Basic	0.0669 (0.0008)	0.0406
Lag group	0.0720 (0.0008)	0.0437
Own/other group	0.0626 (0.0008)	0.0380
Sparse lag group	0.0729 (0.0011)	0.0442
Sparse own/other group	0.0625 (0.0008)	0.0379
Endogenous-first	0.0725 (0.0011)	0.0440
VARX with lags selected by the AIC	0.1043 (0.0015)	0.0633
VARX with lags selected by the BIC	0.1044 (0.0015)	0.0634
BGR's Bayesian VAR	0.7741 (0.0394)	0.4702
Sample mean	1.6460 (0.0902)	1.0000
Random walk	0.7512 (0.0390)	0.4563

Note: Standard errors are shown in parentheses.

The sparsity pattern is depicted in Fig. 10 and the simulation results are summarized in Table 14.

Under Scenario 4, as expected, the own/other group and sparse own/other group VARX-L achieve superior forecasts. Since the magnitudes of the coefficients within a lag matrix vary substantially, structures that utilize lag-based groupings, such as the lag group and the endogenous-first VARX-L, are unable to capture this discrepancy, and thus perform relatively poorly. However, they still outperform the benchmark procedures substantially. Again, we find that VARX methods with lags selected by the AIC and BIC perform very poorly, as they are restricted to selecting from sequentially increasing lag orders and cannot account for within-lag sparsity. BGR's Bayesian VAR also performs poorly for similar reasons.

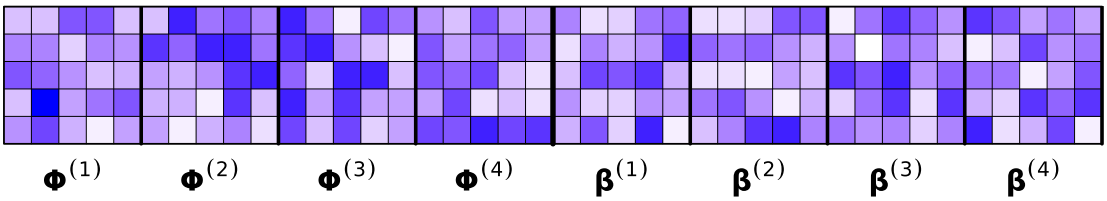

Fig. 11. Magnitudes of coefficients scenario 5: No sparsity.

5.5. Scenario 5: no within-lag sparsity

Next, we consider a scenario in which the entire 5×40 coefficient matrix is dense. Although the coefficient matrix is not sparse, as in the other scenarios, the information criterion based methods can impose sparsity by truncating the maximum lag orders $\hat{p}$ and $\hat{s}$. The relative magnitudes of the coefficient matrix are depicted in Fig. 11. This scenario gives an advantage to procedures that do not impose any within-lag sparsity: the information criterion based methods and BGR's Bayesian VAR. The results from this scenario are depicted in Table 15.

We observe that all of the VARX-L methods outperform the information criterion benchmarks and BGR's Bayesian VAR substantially, in spite of a lack of sparsity in the data generating process. The lag group VARX-L achieves the best performance, which is to be expected. Since the lag group VARX-L has fewer groups than the own/other structures and does not impose within-group sparsity, it has a tendency to employ ridge-like penalization, as opposed to setting an entire group to zero. Within the class of VARX-L models, the basic VARX-L performs the worst. This suggests that structured groupings are more robust in applications in which the true model is not sparse.

5.6. Scenario 6: non-diagonal covariance

The previous five simulation scenarios impose a diagonal structure on Σ_u, the contemporaneous covariance matrix. Such scenarios occur only rarely in practice. This scenario pairs the sparsity pattern from Scenario 1 with the covariance structure depicted in Fig. 12 in order to examine the robustness of the VARX-L procedures in the presence of a non-scaled identity covariance matrix.

Note that, unlike either BGR's Bayesian VAR or the least squares VARX with lags selected by the AIC or BIC, the VARX-L procedures do not incorporate Σ_u explicitly. Hence, one would expect the benchmark procedures to have a slight advantage in this scenario.

In this scenario we observe (see Table 16) that the basic VARX-L, which conforms to the true sparsity pattern, achieves the best performance, followed by the sparse own/other group VARX-L, despite the more complex error structure. As in every other scenario, all of the VARX-L models outperform the benchmarks substantially. BGR's Bayesian VAR performs substantially worse than the information criterion based methods, even though it incorporates the covariance of $\mathbf{u}_t$ explicitly; its poor performance is likely to be the result of incorporating an unreliable covariance estimate. For an expanded discussion and simulation study involving non-diagonal covariance structures, as well as a procedure for incorporating the covariance when fitting VARX-L models, consult Nicholson et al. (2016).

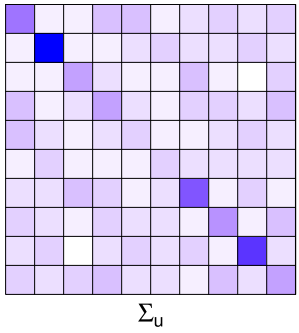

Fig. 12. Covariance matrix: Simulation scenario 6.

Overall, all of the VARX-L models proposed are fairly robust to sparsity patterns that do not conform to their true group structures. In each scenario, every method outperforms all of the benchmark procedures substantially. Scenario 1 is the only case in which the structured approaches perform poorly relative to the basic VARX-L. We expect such unstructured sparsity patterns to occur only rarely in macroeconomic applications.

6. Conclusion

We have shown that the proposed VARX-L structured regularization framework is highly amenable to the VARX setting, in that it can reduce its parameter space while still incorporating useful information from both endogenous and exogenous predictors. VARX-L models scale well with the dimension of the data and are quite flexible at accommodating a wide variety of potential dynamic structures. Each of the proposed methods consistently outperforms the benchmark procedures, both in simulations and in macroeconomic forecasting applications. The forecast performances of all models appear to be robust across multiple sparsity structures and forecast horizons. Moreover, an examination of actual macroeconomic data indicates that structured VARX-L models tend to outperform the basic VARX-L.

Our work has considerable room for extension. This paper focuses solely on forecasting applications, but our VARX-L framework could also be extended to structural analysis and policy evaluation using an approach similar to that of Furman (2014). In addition, our current implementation requires a coherent maximal lag selection mechanism. The common procedure of choosing a lag order based on the frequency of the data is problematic, in that it can lead to overfitting. One could potentially

Table 15
Out-of-sample MSFEs of one-step-ahead forecasts after 100 simulations: Scenario 5.

Model/VARX-L penalty structure	MSFE	MSFE relative to sample mean
Basic	0.0780 (0.0010)	0.6142
Lag group	0.0731 (0.0008)	0.5724
Own/other group	0.0736 (0.0009)	0.5761
Sparse lag group	0.0747 (0.0009)	0.5848
Sparse own/other group	0.0747 (0.0009)	0.5849
Endogenous-first	0.0736 (0.0009)	0.5766
VARX with lags selected by the AIC	0.1207 (0.0016)	0.9444
VARX with lags selected by the BIC	0.1157 (0.0022)	0.9053
BGR's Bayesian VAR	0.1148 (0.0018)	0.8982
Sample mean	0.1278 (0.00296)	1.0000
Random walk	0.2020 (0.00369)	1.5806

Note: Standard errors are shown in parentheses.

Table 16
Out-of-sample MSFEs of one-step-ahead forecasts after 100 simulations: Scenario 6.

Model/VARX-L penalty structure	MSFE	MSFE relative to sample mean
Basic	0.7982 (0.0113)	0.1034
Lag group	0.9089 (0.0133)	0.1177
Own/other group	0.8926 (0.0134)	0.1156
Sparse lag group	0.8833 (0.0122)	0.1144
Sparse own/other group	0.8746 (0.0127)	0.1133
Endogenous-first	0.9283 (0.0135)	0.1202
VARX with lags selected by the AIC	1.4237 (0.0232)	0.1844
VARX with lags selected by the BIC	1.0895 (0.0185)	0.1411
BGR's Bayesian VAR	2.8876 (0.0782)	0.4022
Sample mean	7.1789 (0.3739)	1.0000
Random walk	4.2543 (0.1465)	0.5511

Note: Standard errors are shown in parentheses.

incorporate an additional penalty parameter that grows as the lag order increases, as per Song and Bickel (2011), but this approach requires the use of a multi-dimensional penalty parameter selection procedure and the subjective specification of a functional form for the lag penalty.

An R package containing our algorithms and validation procedures, `BigVAR`, is available on the Comprehensive R Archive Network (CRAN).

Acknowledgments

The authors thank Gary Koop for providing his data transformation script, Marta Bańbura, Domenico Giannone, and Lucrezia Reichlin for sharing their BVAR code, and the attendees of the 2014 NBER/NSF Time Series Conference for their constructive comments, as well as two anonymous referees for helpful suggestions. This research was supported by an Amazon Web Services in Education Research Grant. DSM was supported by a Xerox PARC Faculty Research Award and National Science Foundation Grant DMS-1455172. JB was supported by National Science Foundation DMS-1405746.

Appendix

A.1. Compact matrix notation

In deriving the solution methods for our algorithms, we find it convenient to express the VARX using compact matrix notation as given in Box I: Eq. (1) then becomes

$$\boldsymbol{Y} = \boldsymbol{\nu}\boldsymbol{1}^\top + \boldsymbol{B}\boldsymbol{Z} + \mathbf{U},$$

and the least squares procedure in Eq. (2) can be expressed as minimizing $\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\nu}\boldsymbol{1}^\top - \boldsymbol{B}\boldsymbol{Z}\|_F^2$ over $\boldsymbol{\nu}$ and $\boldsymbol{B}$.

A.2. Intercept term

In regularization problems, the intercept $\hat{\boldsymbol{\nu}}$ typically is not regularized, and instead can be derived separately. Using compact matrix notation, we can express the unpenalized portion of Eq. (3) as

$$f(\boldsymbol{B}, \boldsymbol{\nu}) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\nu}\boldsymbol{1}^\top - \boldsymbol{B}\boldsymbol{Z}\|_F^2. \tag{14}$$

We can find $\hat{\boldsymbol{\nu}}$ by calculating the gradient of Eq. (14) with respect to $\boldsymbol{\nu}$

$$0 = \nabla_{\boldsymbol{\nu}} f(\boldsymbol{B}, \boldsymbol{\nu}) = (\boldsymbol{Y} - \hat{\boldsymbol{\nu}}\boldsymbol{1}^\top - \widehat{\boldsymbol{B}\boldsymbol{Z}})\boldsymbol{1},$$
$$\implies \hat{\boldsymbol{\nu}}_k(\lambda) = \overline{\boldsymbol{Y}}_{k\cdot} - \widehat{\boldsymbol{B}\overline{\boldsymbol{Z}}}_{k\cdot},$$

where $\overline{\boldsymbol{Y}}_{k\cdot} = \frac{1}{T}\sum_t \boldsymbol{Y}_{kt}$ and $\overline{\boldsymbol{Z}}_{k\cdot} = \frac{1}{T}\sum_t \boldsymbol{Z}_{kt}$. This provides some insight into the scaling, as we can rewrite Eq. (14) as

$$\min_{\boldsymbol{B}} \frac{1}{2}\|\boldsymbol{Y} - (\overline{\boldsymbol{Y}} - \boldsymbol{B}\overline{\boldsymbol{Z}})\boldsymbol{1}^\top - \boldsymbol{B}\boldsymbol{Z}\|_F^2,$$
$$= \min_{\boldsymbol{B}} \frac{1}{2}\|(\boldsymbol{Y} - \overline{\boldsymbol{Y}}\boldsymbol{1}^\top) - \boldsymbol{B}(\boldsymbol{Z} - \overline{\boldsymbol{Z}}\boldsymbol{1}^\top)\|_F^2, \tag{15}$$

$$\begin{array}{llll} \boldsymbol{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T] & (k \times T); & \boldsymbol{1} = [1, \ldots, 1]^\top & (T \times 1); \\ \boldsymbol{Z}_t = [\mathbf{y}_{t-1}^\top, \ldots, \mathbf{y}_{t-p}^\top, \mathbf{x}_{t-1}^\top, \ldots, \mathbf{x}_{t-s}^\top]^\top & [(kp + ms) \times 1]; & \boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T] & [(kp + ms) \times T]; \\ \boldsymbol{\Phi} = [\boldsymbol{\Phi}^{(1)}, \boldsymbol{\Phi}^{(2)}, \ldots, \boldsymbol{\Phi}^{(p)}] & (k \times kp); & \boldsymbol{\beta} = [\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(s)}] & [k \times ms]; \\ \boldsymbol{B} = [\boldsymbol{\Phi}, \boldsymbol{\beta}] & [k \times (kp + ms)]; & \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_T] & (k \times T). \end{array}$$

Box I.

where $\overline{\boldsymbol{Y}}$ is a $k \times 1$ vector of row means and $\overline{Z}$ is a $(kp + ms) \times 1$ vector of row means.

A.3. Solution strategies

In the following sections, assume that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are centered as in Eq. (15).

A.3.1. Basic VARX-L

Utilizing the coordinate descent framework, we can find $\widehat{\boldsymbol{B}}$ via scalar updates. To generalize to a multivariate context, we can express the one-variable update for the (j, r) entry of $\boldsymbol{B}$, $\boldsymbol{B}_{jr}$ as

$$\min_{\boldsymbol{B}_{jr}} \frac{1}{2} \sum_t \left(\boldsymbol{Y}_{jt} - \sum_{\ell \neq r} \boldsymbol{B}_{j\ell} \boldsymbol{Z}_{\ell t} - \boldsymbol{B}_{jr} \boldsymbol{Z}_{jt} \right)^2 + \lambda |\boldsymbol{B}_{jr}|. \tag{16}$$

Let $\mathbf{R}_t = \boldsymbol{Y}_{jt} - \sum_{\ell \neq r} \boldsymbol{B}_{j\ell} \boldsymbol{Z}_{\ell t}$ denote the partial residual. Then, we can rewrite Eq. (16) as

$$\begin{aligned} g_{jr}(\boldsymbol{B}) &= \min_{\boldsymbol{B}_{jr}} \frac{1}{2} \sum_t (\mathbf{R}_t - \boldsymbol{B}_{jr} \boldsymbol{Z}_{jt})^2 + \lambda |\boldsymbol{B}_{jr}| \\ &= \min_{\boldsymbol{B}_{jr}} \frac{1}{2} \sum_t (\mathbf{R}_t^2 - \boldsymbol{B}_{jr}^2 \boldsymbol{Z}_{jt}^2 - 2 \mathbf{R}_t \boldsymbol{Z}_{jt} \boldsymbol{B}_{jr}) + \lambda |\boldsymbol{B}_{jr}|. \end{aligned}$$

Now, differentiating with respect to $\boldsymbol{B}_{jr}$ gives the subgradient as

$$\partial g_{jr}(\boldsymbol{B}) = \boldsymbol{B}_{jr} \sum_t \boldsymbol{Z}_{jt}^2 - \sum_t \mathbf{R}_t \boldsymbol{Z}_{jt} + \lambda \psi(\boldsymbol{B}_{jr}),$$

where we define $\psi(\boldsymbol{B}_{jr})$ as

$$\psi \in \begin{cases} \{\operatorname{sgn}(\boldsymbol{B}_{jr})\} & \boldsymbol{B}_{jr} \neq 0 \\ [-1, 1] & \boldsymbol{B}_{jr} = 0. \end{cases}$$

For $\hat{\boldsymbol{B}}_{jr}$ to be a global minimum, $0 \in \partial g(\hat{\boldsymbol{B}}_{jr})$. After some algebra, the optimal update can be expressed as

$$\hat{\boldsymbol{B}}_{jr} \leftarrow \frac{\mathcal{ST}\left(\sum_t \mathbf{R}_t \boldsymbol{Z}_{jt}, \lambda \right)}{\sum_t \boldsymbol{Z}_{jt}^2},$$

where $\mathcal{ST}$ represents the soft-threshold operator

$$\mathcal{ST}(x, \phi) = \operatorname{sgn}(x)(|x| - \phi)_+,$$

with sgn denoting the signum function, and $(|x| - \phi)_+ = \max(|x| - \phi, 0)$. The basic VARX-L procedure is detailed in Algorithm 1.

A.3.2. Lag group VARX-L

If the groups are proper submatrices, we can achieve considerable computational gains by exploiting the matrix structure rather than vectorizing the lag group VARX-L and solving the corresponding univariate least squares problem. Without loss of generality, we will consider the "one lag" problem for $\boldsymbol{\Phi}^{(q)}$ (the problem for $\boldsymbol{\beta}^{(q)}$ is analogous):

$$\min_{\boldsymbol{\Phi}^{(q)}} \frac{1}{2} \|\mathbf{R}_{-q} - \boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q\|_F^2 + \lambda \|\boldsymbol{\Phi}^{(q)}\|_F, \tag{17}$$

where we incorporate the weighting into the penalty parameter directly for notational ease, by defining $\lambda = k\lambda$, where $\mathbf{R}_q = \boldsymbol{\Phi}^{(-q)} \boldsymbol{Z}_{-q} - \boldsymbol{Y} \in \mathbb{R}^{k \times T}$ again represents the partial residual. Taking the gradient of $\|\mathbf{R}_{-q} - \boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q\|_F^2$ with respect to $\boldsymbol{\Phi}^{(q)}$ results in

$$\begin{aligned} &\nabla_{\boldsymbol{\Phi}^{(q)}} \frac{1}{2} \|\mathbf{R}_{-q} - \boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q\|_F^2 \\ &\quad = \nabla_{\boldsymbol{\Phi}^{(q)}} \operatorname{Tr}\left((\mathbf{R}_{-q} - \boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q)(\mathbf{R}_{-q} - \boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q)^\top \right), \\ &\quad = \boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^\top - \mathbf{R}_{-q} \boldsymbol{Z}_q^\top, \end{aligned}$$

where Tr denotes the trace operator. The subgradient of Eq. (17) with respect to $\boldsymbol{\Phi}^{(q)}$ is then

$$\boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^\top - \mathbf{R}_{-q} \boldsymbol{Z}_q^\top + \lambda \omega(\boldsymbol{\Phi}^{(q)}),$$

where ω is defined as

$$\omega(\boldsymbol{\Phi}^{(q)}) = \begin{cases} \dfrac{\boldsymbol{\Phi}^{(q)}}{\|\boldsymbol{\Phi}^{(q)}\|_F} & \boldsymbol{\Phi}^{(q)} \neq 0 \\ \{U : \|U\|_F \leq 1\} & \boldsymbol{\Phi}^{(q)} = 0. \end{cases}$$

Consider the case where $\hat{\boldsymbol{\Phi}}^{(q)} = \mathbf{0}$. Then

$$\begin{aligned} &\frac{\hat{\boldsymbol{\Phi}}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^\top - \mathbf{R}_{-q} \boldsymbol{Z}_q^\top}{\lambda} \in \{U : \|U\|_F \leq 1\}, \\ &\iff \|\hat{\boldsymbol{\Phi}}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^\top - \mathbf{R}_{-q} \boldsymbol{Z}_q^\top\|_F \leq \lambda, \\ &\iff \|\mathbf{R}_{-q} \boldsymbol{Z}_q^\top\|_F \leq \lambda. \end{aligned}$$

We can conclude that $\hat{\boldsymbol{\Phi}}^{(q)} = 0 \iff \|\mathbf{R}_{-q} \boldsymbol{Z}_q^\top\|_F \leq \lambda$. Now, assuming $\hat{\boldsymbol{\Phi}}^{(q)} \neq 0$, we have that

$$\begin{aligned} &\boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^\top - \mathbf{R}_{-q} \boldsymbol{Z}_q^\top + \lambda \left(\frac{\boldsymbol{\Phi}^{(q)}}{\|\boldsymbol{\Phi}^{(q)}\|_F} \right) = 0, \\ &\boldsymbol{\Phi}^{(q)} \boldsymbol{Z}_q \boldsymbol{Z}_q^\top + \lambda \left(\frac{\boldsymbol{\Phi}^{(q)}}{\|\boldsymbol{\Phi}^{(q)}\|_F} \right) = \mathbf{R}_{-q} \boldsymbol{Z}_q^\top, \\ &\boldsymbol{\Phi}^{(q)} \left(\boldsymbol{Z}_q \boldsymbol{Z}_q^\top + \frac{\lambda}{\|\boldsymbol{\Phi}^{(q)}\|_F} \boldsymbol{I}_k \right) = \mathbf{R}_{-q} \boldsymbol{Z}_q^\top. \end{aligned} \tag{18}$$

Since $\boldsymbol{Z}_q\boldsymbol{Z}_q^\top$ is positive semidefinite and $\lambda > 0$, we can infer that $\boldsymbol{Z}_q\boldsymbol{Z}_q^\top + \frac{\lambda}{\|\boldsymbol{\Phi}^{(q)}\|_F}\boldsymbol{I}_k$ is positive definite; hence, it is possible to create a trust region subproblem that coincides with Eq. (17). However, we need to transform $\mathbf{R}_{-q}\boldsymbol{Z}_q^\top \in \mathbb{R}^{k\times k}$ into a vector. Define

$$\boldsymbol{r}_q = \text{vec}(\mathbf{R}_{-q}\boldsymbol{Z}_q^\top),$$
$$\mathbf{G}_q = \boldsymbol{Z}_q\boldsymbol{Z}_q^\top \otimes \mathbf{I}_k,$$
$$\boldsymbol{\phi}_q = \text{vec}(\boldsymbol{\Phi}^{(q)}),$$

where $\otimes$ denotes the Kronecker product. Hence, we can rewrite Eq. (18) as

$$\boldsymbol{\phi}_q^\top \left(\mathbf{G}_q + \frac{\lambda}{\|\boldsymbol{\phi}_q\|_F}\boldsymbol{I}_{k^2}\right) = \boldsymbol{r}_q.$$

Applying the same transformation to the original subproblem, we can express Eq. (17) as the trust region subproblem

$$\min \frac{1}{2}\boldsymbol{\phi}_q^\top \mathbf{G}_q\mathbf{G}_q^\top\boldsymbol{\phi}_q + \boldsymbol{r}_q^\top\boldsymbol{\phi}_q,$$
$$\text{s.t.} \|\boldsymbol{\phi}_q\|_F \le \Delta,$$

where $\Delta > 0$ is the trust-region radius that corresponds to the optimal solution of Eq. (17). These modifications allow for the use of the block coordinate descent algorithm described by Qin et al. (2010). Expanding upon their arguments, according to the Karush–Kuhn–Tucker (KKT) conditions, we have that $\lambda(\Delta - \|\boldsymbol{\phi}_q^*\|_F) = 0$, which implies that $\|\boldsymbol{\phi}_q^*\|_F = \Delta$. Then, applying Theorem 4.1 of Nocedal and Wright (1999), we can conclude that

$$\boldsymbol{\phi}_q^* = -\left(\mathbf{G}_q + \frac{\lambda}{\Delta}\boldsymbol{I}_{k^2}\right)^{-1}\boldsymbol{r}_q. \tag{19}$$

Qin et al. (2010) remarks that Eq. (19) can also be expressed as $\boldsymbol{\phi}_q^* = \Delta y_q(\Delta)$, where

$$y_q(\Delta) = -\left(\Delta\mathbf{G}_q + \lambda\boldsymbol{I}_{k^2}\right)^{-1}\boldsymbol{r}_q, \tag{20}$$

Note that, based on the KKT conditions, $\|y_q(\Delta)\|_F = 1$. Hence, the optimal Δ can be chosen to satisfy $\|y_q(\Delta)\|_F = 1$. We can compute $\|y_q(\Delta)\|_F^2$ efficiently via an eigendecomposition of $\mathbf{G}_q$. We start by rewriting Eq. (20) as

$$\begin{aligned} y_q(\Delta) &= -\left(\Delta\boldsymbol{W}V\boldsymbol{W}^\top + \lambda\boldsymbol{I}_{k^2}\right)^{-1}\boldsymbol{r}_q, \\ &= -\boldsymbol{W}(\Delta V + \lambda\boldsymbol{I}_{k^2})^{-1}\boldsymbol{W}^\top\boldsymbol{r}_q, \end{aligned}$$

where the first line follows from the spectral decomposition of a symmetric positive semidefinite matrix.
Finally, we can express $\|y_q(\Delta)\|_F^2$ as

$$\|y_q(\Delta)\|_F^2 = \sum_i \frac{(\mathbf{w}_i^\top\boldsymbol{r}_q)^2}{(\mathbf{v}_i\Delta + \lambda)^2},$$

where $\boldsymbol{w}_i$ denotes the columns of $\boldsymbol{W}$ and $\boldsymbol{v}_i$ the diagonal elements of $\boldsymbol{V}$. Qin et al. (2010) notes that we can determine the optimal Δ by using Newton's method to find the root of

$$\Omega(\Delta) = 1 - \frac{1}{\|y_q(\Delta)\|_F}. \tag{21}$$

The full lag group VARX-L procedure is detailed in Algorithm 3. Our algorithm organizes iterations around an "active set", as described by Friedman et al. (2010). This approach starts by cycling through every group but iterating only on the subset of $\boldsymbol{B}$ that are nonzero (the "active set") until convergence. If a full pass through all $\boldsymbol{B}$ does not change the active set, the algorithm has converged; otherwise, the process is repeated. This approach reduces the computation time considerably, especially for large values of λ, in which most of the model coefficients are zero.

A.3.3. Own/other group VARX-L

In the own/other setting, Eq. (3) must be transformed into a least squares problem in order to partition each $\boldsymbol{\Phi}^{(\ell)}$ properly into separate groups for own and other lags, since the groups are not proper submatrices. We perform a least squares transformation by defining

$$r_{-qq} = \text{vec}(\mathbf{R}_{-qq}),$$
$$\boldsymbol{\phi}_{qq} = \text{vec}(\boldsymbol{\Phi}_{\text{on}}^{(q)}),$$
$$\boldsymbol{M}_{qq} = (\boldsymbol{Z}^\top \otimes I_k)_{qq}.$$

Then, the one-block subproblem for own lags (group qq) can be expressed as

$$\begin{aligned} &\min_{\boldsymbol{\phi}_{qq}} \frac{1}{2}\|\boldsymbol{M}_{qq}\boldsymbol{\phi}_{qq} + r_{-qq}\|_F^2 + \lambda\|\boldsymbol{\phi}_{qq}\|_F, \\ &\quad= \min_{\boldsymbol{\phi}_{qq}} \frac{1}{2}r_{-qq}^\top r_{-qq} + \boldsymbol{\phi}_{qq}^\top\boldsymbol{M}_{qq}^\top\boldsymbol{M}_{qq}\boldsymbol{\phi}_{qq} \\ &\qquad + r_{-qq}^\top\boldsymbol{M}_{qq}\boldsymbol{\phi}_{qq} + \lambda\|\boldsymbol{\phi}_{qq}\|_F, \\ &\quad= \min_{\boldsymbol{\phi}_{qq}} \frac{1}{2}\boldsymbol{\phi}_{qq}^\top\boldsymbol{M}_{qq}^\top\boldsymbol{M}_{qq}\boldsymbol{\phi}_{qq} + r_{-qq}^\top\boldsymbol{M}_{qq}\boldsymbol{\phi}_{qq} + \lambda\|\boldsymbol{\phi}_{qq}\|_F. \end{aligned}$$

At $\hat{\boldsymbol{\phi}}_{qq}$, we must have that $0 \in \partial f(\hat{\boldsymbol{\phi}}_{qq})$. The subgradient can be expressed as

$$\frac{\partial}{\partial\boldsymbol{\phi}_{qq}} = \boldsymbol{M}_{qq}^\top\boldsymbol{M}_{qq}\boldsymbol{\phi}_{qq} + \boldsymbol{M}_{qq}^\top r_{-qq} + \lambda\omega(\boldsymbol{\phi}_{qq}),$$

where ω is defined as

$$\omega(s) \in \begin{cases} \left\{\dfrac{s}{\|s\|_F}\right\} & s \neq 0 \\ \{u : \|u\|_F \le 1\} & s = 0. \end{cases}$$

Thus, after applying these transformations, we can apply a slightly adapted version of Algorithm 3.

A.3.4. Sparse lag group VARX-L

As with the lag group VARX-L, we will consider the one-block subproblem for lag $\boldsymbol{\Phi}^{(q)}$

$$\begin{aligned} &\min_{\boldsymbol{\Phi}^{(q)}} \frac{1}{2k}\|\mathbf{R}_{-q} - \boldsymbol{\Phi}^{(q)}\boldsymbol{Z}_q\|_F^2 \\ &\quad + (1-\alpha)\lambda\|\boldsymbol{\Phi}^{(q)}\|_F + \alpha\lambda\|\boldsymbol{\Phi}^{(q)}\|_1. \end{aligned} \tag{22}$$

Since the inclusion of within-group sparsity does not allow for separability, coordinate descent based procedures are no longer appropriate; thus, our solution to the sparse lag group VARX-L utilizes gradient descent based methods, following Simon et al. (2013). We express Eq. (22) as the sum of a generic differentiable function with a Lipschitz gradient and a non-differentiable function.

We start by linearizing the quadratic approximation of the unpenalized loss function that makes use only of first-order information around its current estimate $\mathbf{\Phi}_0$ (borrowing from Simon et al. (2013), let $\mathbf{\Phi} \equiv \mathbf{\Phi}^{(q)}$, $\ell(\mathbf{\Phi})$ represent the unpenalized loss function, and $\mathcal{P}(\mathbf{\Phi})$ represent the penalty term, for notational ease). Then, we can express the linearization as

$$\begin{aligned} M(\mathbf{\Phi}, \mathbf{\Phi}_0) &= \ell(\mathbf{\Phi}_0) + \operatorname{vec}(\mathbf{\Phi} - \mathbf{\Phi}_0)^\top \operatorname{vec}(\nabla \ell(\mathbf{\Phi}_0)) \\ &\quad + \frac{1}{2d}\|\mathbf{\Phi} - \mathbf{\Phi}_0\|_F^2 + \mathcal{P}(\mathbf{\Phi}) \\ &= \frac{1}{2k}\|\mathbf{R}_{-q} - \mathbf{\Phi}_0 \mathbf{Z}_q\|_F^2 \\ &\quad + \langle \mathbf{\Phi} - \mathbf{\Phi}_0, (\mathbf{\Phi}_0 \mathbf{Z}_q - \mathbf{R}_{-q})\mathbf{Z}_q^\top \rangle \\ &\quad + \frac{1}{2d}\|\mathbf{\Phi} - \mathbf{\Phi}_0\|_F^2 + \mathcal{P}(\mathbf{\Phi}), \end{aligned}$$

in which d represents the step size. Removing terms that are independent of $\mathbf{\Phi}$, our objective function becomes

$$\begin{aligned} &\operatorname*{argmin}_{\mathbf{\Phi}} M(\mathbf{\Phi}, \mathbf{\Phi}_0), \\ &= \operatorname*{argmin}_{\mathbf{\Phi}} \frac{1}{2d}\|\mathbf{\Phi} - \left(\mathbf{\Phi}_0 - d(\mathbf{\Phi}_0 \mathbf{Z}_q - \mathbf{R}_{-q})\mathbf{Z}_q^\top\right)\|_F^2 \\ &\quad + \mathcal{P}(\mathbf{\Phi}). \end{aligned}$$

Then, generalizing the arguments outlined by Simon et al. (2013), we can infer that the optimal update $U(\mathbf{\Phi})$ can be expressed as

$$\begin{aligned} U(\mathbf{\Phi}) = &\left(1 - \frac{d(1-\alpha)\lambda}{\|ST(\mathbf{\Phi}_0 - d(\mathbf{\Phi}_0 \mathbf{Z}_q - \mathbf{R}_{-q})\mathbf{Z}_q^\top, d\alpha\lambda)\|_F}\right)_+ \\ &\times ST(\mathbf{\Phi}_0 - d(\mathbf{\Phi}_0 \mathbf{Z}_q - \mathbf{R}_{-q})\mathbf{Z}_{-q}^\top, d\alpha\lambda). \end{aligned}$$

Following Simon et al. (2013), we apply a Nesterov accelerated update. At step j, we update according to

$$\hat{\mathbf{\Phi}}[j] \leftarrow \hat{\mathbf{\Phi}}[j-1] + \frac{j}{j+3}(U(\mathbf{\Phi}) - \hat{\mathbf{\Phi}}[j-1]), \tag{23}$$

which, as per Beck and Teboulle (2009), converges at the rate $1/j^2$, as opposed to the $1/j$ rate of the standard proximal gradient descent.

We use a constant step size according to the Lipschitz constant, H, which must satisfy

$$\|\nabla_X \ell(X) - \nabla_Y \ell(Y)\| \leq H\|X - Y\|.$$

Consider two submatrices $\mathbf{A}^{(q)}$ and $\mathbf{C}^{(q)}$. We have that

$$\begin{aligned} &\nabla_{\mathbf{A}^{(q)}} \ell(\mathbf{A}^{(q)}) = \mathbf{A}^{(q)} \mathbf{Z}_q \mathbf{Z}_q^\top - \mathbf{R}_{-q}\mathbf{Z}_q^\top, \\ &\nabla_{\mathbf{C}^{(q)}} \ell(\mathbf{C}^{(q)}) = \mathbf{C}^{(q)} \mathbf{Z}_q \mathbf{Z}_q^\top - \mathbf{R}_{-q}\mathbf{Z}_q^\top, \\ \Longrightarrow &\nabla_{\mathbf{A}^{(q)}} \ell(\mathbf{A}^{(q)}) - \nabla_{\mathbf{C}^{(q)}} \ell(\mathbf{C}^{(q)}) = (\mathbf{A}^{(q)} - \mathbf{C}^{(q)})\mathbf{Z}_q \mathbf{Z}_q^\top, \\ \Longrightarrow &\|(\mathbf{A}^{(q)} - \mathbf{C}^{(q)})\mathbf{Z}_q \mathbf{Z}_q^\top\|_2 \leq \|\mathbf{A}^{(q)} - \mathbf{C}^{(q)}\|_2 \|\mathbf{Z}_q \mathbf{Z}_q^\top\|_2. \end{aligned}$$

The last inequality follows from the sub-multiplicity of the matrix 2-norm. Therefore, we can conclude that the Lipschitz constant is $\|\mathbf{Z}_q \mathbf{Z}_q^\top\|_2 = \sigma_1(\mathbf{Z}_q \mathbf{Z}_q^\top)$ i.e., the largest singular value of $\mathbf{Z}_q \mathbf{Z}_q^\top$, which has the dimension $k \times k$ for $\mathbf{\Phi}^{(1)}, \ldots, \mathbf{\Phi}^{(p)}$ and is a scalar for exogenous groups. Since $\mathbf{Z}_q \mathbf{Z}_q^\top$ is symmetric and positive semidefinite, it is diagonalizable, and the maximum eigenvalue can be computed efficiently using the power method described by Golub and Van Loan (2012).

The power method is much more computationally efficient than a computation of the entire eigensystem because only the maximum eigenvalue is required. Moreover, we retain the corresponding eigenvector produced by this procedure to use as a "warm start" that substantially decreases the amount of time required to compute the maximal eigenvalue at each point in time during the cross-validation and forecast evaluation procedures.

In a manner similar to Algorithm 3, an "active set" approach is used to minimize the computation time. The inner loop of the sparse group VARX-L procedure is detailed in Algorithm (4). An outline of the algorithm is given below.

1. Iterate through all groups. For each group:
   (a) Check if the group is active via the condition $\|(\mathbf{\Phi}^{(q)}\mathbf{Z}_q - \mathbf{R}_{-q})\mathbf{Z}_q^\top\|_F \leq (1-\alpha)\lambda$.
   (b) If active, go to the inner loop (Algorithm 4); if not active, set group identically to zero.
   (c) Repeat until convergence.

Upon performing the least squares transformations as in the own/other group VARX-L (Appendix A.3.3), the sparse own/other group VARX-L follows almost exactly the same procedure as its lag counterpart.

A.3.5. Endogenous-first VARX-L

The endogenous-first VARX-L is of the form

$$\min_{\mathbf{\Phi}, \boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{BZ}\|_F^2 + \lambda \sum_{\ell=1}^{p} \sum_{j=1}^{k} \left(\|[\mathbf{\Phi}_j^{(\ell)}, \boldsymbol{\beta}_{j,\cdot}^{(\ell)}]\|_F + \|\boldsymbol{\beta}_{j,\cdot}^{(\ell)}\|_F\right).$$

Since the optimization problem decouples across rows, we will consider solving the *one row* subproblem (for row i)

$$\min_{\mathbf{\Phi}_i, \boldsymbol{\beta}_i} \frac{1}{2}\|\boldsymbol{Y}_i - \boldsymbol{B}_i\boldsymbol{Z}\|_F^2 + \lambda \sum_{\ell=1}^{p} \left(\|[\mathbf{\Phi}_i^{(\ell)}, \boldsymbol{\beta}_{i,\cdot}^{(\ell)}]\|_F + \|\boldsymbol{\beta}_{i,\cdot}^{(\ell)}\|_F\right). \tag{24}$$

The endogenous-first VARX-L is solved via proximal gradient descent, in a manner similar to that for the sparse group VARX-L. For ease of notation, let $\mathcal{P}(\mathbf{\Phi}, \boldsymbol{\beta})$ represent the nested penalty. The update step for the endogenous-first VARX-L (at step j) can be expressed as

$$\boldsymbol{B}_i[j] = \operatorname{Prox}_{d\lambda, \mathcal{P}(\mathbf{\Phi}, \boldsymbol{\beta})}(\boldsymbol{B}_i[j-1] - d\nabla \ell(\boldsymbol{B}_i[j-1])), \tag{25}$$

where d is the step size and $\ell(\boldsymbol{B}_i)$ denotes the unpenalized loss function. Note that $\nabla\ell(\boldsymbol{B}_i) = -(\boldsymbol{Y}_i - \boldsymbol{B}_i\boldsymbol{Z})\boldsymbol{Z}^\top$. Similarly to the sparse group VARX-L setting, a fixed step size is used: $d = \frac{1}{\sigma_1(\boldsymbol{Z}\boldsymbol{Z}^\top)}$. To speed convergence, as in the sparse group VARX-L update step (Eq. (23)), we apply a similar Nesterov-style accelerated update:

$$\hat{\boldsymbol{B}}_i \leftarrow \hat{\boldsymbol{B}}_i[j-1] + \frac{j-2}{j+1}(\hat{\boldsymbol{B}}_i[j-1] - \hat{\boldsymbol{B}}_i[j-2]).$$

Thus, Eq. (25) becomes

$$\boldsymbol{B}_i[j] = \text{Prox}_{d\lambda,\mathcal{P}(\boldsymbol{\Phi},\boldsymbol{\beta})}(\hat{\boldsymbol{B}}_i - d\nabla\ell(\hat{\boldsymbol{B}}_i)). \tag{26}$$

Jenatton et al. (2011) observed that the dual of Eq. (26) can be solved with one pass of the block coordinate descent. Moreover, the block updates are extremely simple and are available in closed form. Algorithm 5 details the prox function for the endogenous-first VARX-L. Note that it consists of p separate nested structures for each series, meaning that solving Eq. (26) essentially amounts to calling the same proximal function p times at each update step.

A.4. Banbura et al. (2009) implementation

The Bayesian VAR proposed by Banbura et al. (2009) utilizes a normal inverted Wishart prior. Defining $\phi = \text{vec}(\boldsymbol{\Phi})$, the prior has the form

$$\phi|\Omega \sim N(\phi_0, \Omega \otimes \Omega_0)$$
$$\Omega \sim iW(S_0, \alpha_0),$$

in which iW denotes the inverse Wishart distribution. This prior is implemented by adding the following dummy observations to $\boldsymbol{Y}$ and $\boldsymbol{Z}^\top$ (which we define as $\boldsymbol{X}$):

$$\boldsymbol{Y}_{d_1} = \begin{pmatrix} \text{diag}(\delta\sigma_1, \ldots, \delta\sigma_k)/\lambda \\ 0_{k\times(p-1)\times k} \\ \text{diag}(\sigma_1, \ldots, \sigma_k) \\ 0_{1\times k} \end{pmatrix}$$

$$\boldsymbol{X}_{d_1} = \begin{pmatrix} 0_{kp\times 1} & J_p \otimes \text{diag}(\sigma_1, \ldots, \sigma_k) \\ 0_{k\times 1} & 0_{k\times kp} \\ \epsilon & 0_{1\times kp} \end{pmatrix}.$$

The scale parameters for the prior variances of each series, $\sigma_1, \ldots, \sigma_k$, are estimated using univariate autoregressive models. $J_p = \text{diag}(1, 2, \ldots, p)$, and ϵ denotes the prior on the intercept and is set to a very small number (e.g., 1e−5). δ serves as an indicator for the prior belief that the series have high persistence. We set $\delta = 0$ in all of our forecasting applications except for the Minnesota VARX-L application in Section 4.

In addition to the construction above, BGR adds a prior that imposes a bound on the sum of coefficients by shrinking $\Pi = (I_k - \boldsymbol{\Phi}_1 - \cdots - \boldsymbol{\Phi}_p)$ toward zero, following Doan, Litterman, and Sims (1984). This prior is implemented by adding the additional dummy observations

$$\boldsymbol{Y}_{d_2} = \text{diag}(\delta\mu_1, \ldots, \delta\mu_k)/\tau,$$
$$\boldsymbol{X}_{d_2} = \begin{pmatrix} \boldsymbol{0}_{k\times 1} & 1_{1\times p} \otimes \text{diag}(\delta\mu_1, \ldots, \delta\mu_k)/\tau \end{pmatrix},$$

Table 17
Starting values of the penalty grid for each procedure; ρ_q denotes the number of variables in group q.

Structure	Starting value of Λ_{Grid}
Basic	$\|\boldsymbol{Z}\boldsymbol{Y}^\top\|_\infty$
Lag group	$\max_q(\|\boldsymbol{Z}_q\boldsymbol{Y}^\top\|_F)$
Sparse lag	$\max_q\left(\|\boldsymbol{Z}_q\boldsymbol{Y}^\top\|_F\right)\alpha$
Own/other group	$\max_q\left(\|(\boldsymbol{Z}^\top \otimes \boldsymbol{I}_k)_q \text{vec}(\boldsymbol{Y})\|_F/\sqrt{\rho_q}\right)$
Sparse own/other group	$\max_q\left(\|(\boldsymbol{Z}^\top \otimes \boldsymbol{I}_k)_q \text{vec}(\boldsymbol{Y})\|_F/\sqrt{\rho_q}\right)\alpha$
Endogenous-first	$\max_i(\|\boldsymbol{Z}\boldsymbol{Y}_i^\top\|_F)$

where $\mu_1, \ldots, \mu_k$ are meant to capture the average level of each series, set according to their unconditional means, and τ denotes a loose prior which is set to 10λ. After appending the dummy observations to $\boldsymbol{Y}$ and $\boldsymbol{X}$ and creating the augmented matrices $\boldsymbol{Y}_*$ and $\boldsymbol{X}_*$, the posterior mean can be calculated in closed form as:

$$\tilde{\boldsymbol{\Phi}}^\top = (\boldsymbol{X}_*^\top\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^\top\boldsymbol{Y}_*.$$

A.5. Penalty grid selection

See Table 17.

A.6. Algorithms

Algorithm 1 Basic VARX-L$_{k,m}(p, s)$

Require: $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{B}^{\text{INI}}, \lambda$
  $\boldsymbol{B}^{\text{OLD}} \leftarrow \boldsymbol{B}^{\text{INI}}$
  repeat
    for i in [1, k], j in [1, $kp + ms$] **do**
      $\boldsymbol{R} \leftarrow \boldsymbol{Y}_{i\cdot} - \sum_{\ell\neq j}\boldsymbol{B}_{i\ell}\boldsymbol{Z}_{\ell\cdot}$
      $\boldsymbol{B}_{ij}^{\text{NEW}} \leftarrow \frac{\text{ST}(\sum_t \boldsymbol{R}\boldsymbol{Z}_{j\cdot}, \lambda)}{\sum_t \boldsymbol{Z}_{jt}^2}$
    end for
    $\boldsymbol{B}^{\text{OLD}} \leftarrow \boldsymbol{B}^{\text{NEW}}$
  until desired threshold is reached
  $\hat{\boldsymbol{\nu}} \leftarrow \bar{\boldsymbol{Y}} - \boldsymbol{B}^{\text{NEW}}\bar{\boldsymbol{Z}}$
return $\hat{\boldsymbol{\nu}}, \boldsymbol{B}^{\text{NEW}}$

Algorithm 2 Basic VARX-L(p,s) cross-validation

Require: $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{B}^{\text{INI}}, \Lambda_{\text{grid}}, h$
  $\boldsymbol{B}^{\text{LAST}} \leftarrow \boldsymbol{B}^{\text{INI}}$

  for t in $[T_1, T_2 - h]$ **do**
    $\boldsymbol{Y}_{\text{TRAIN}}^{(t)} \leftarrow \boldsymbol{Y}_{,h:(t-1)}$
    $\boldsymbol{Z}_{\text{TRAIN}}^{(t)} \leftarrow \boldsymbol{Z}_{,1:(t-h+1)}$
    for i in Λ_{Grid} **do**
      $\nu_i, \boldsymbol{B}_i^{\text{NEW}} \leftarrow$ Basic-VARX-L($\boldsymbol{Y}_{\text{TRAIN}}^{(t)}, \boldsymbol{Z}_{\text{TRAIN}}^{(t)}, \boldsymbol{B}_i^{\text{LAST}}, \lambda_i$)
      $\boldsymbol{Z}_{\text{TEST}}^{(t)} \leftarrow \boldsymbol{Z}_{,(t+1)}$
      $SSFE^{(t,i)} \leftarrow \|\boldsymbol{Y}_{t+h} - [\nu_i, \boldsymbol{B}_i^{\text{NEW}}] * [\boldsymbol{1}, \boldsymbol{Z}_{TEST}^{(t)}]\|_F^2$
      $\boldsymbol{B}_i^{\text{LAST}} \leftarrow \boldsymbol{B}_i^{\text{NEW}}$

    end for
    for i in Λ_{Grid} **do**
      $MSFE^{(i)} \leftarrow \frac{1}{T_2 - T_1 - h + 1}\sum_t SSFE^{(t,i)}$
    end for

  end for
  return $\lambda_{\hat{\imath}}$, where $\hat{\imath} = \text{argmin}_i\, MSFE^{(i)}$

Algorithm 3 Lag group VARX-L$_{k,m}(p, s)$ with active-set strategy

Require: $\boldsymbol{B}_{\text{INI}}, \mathcal{G}, \boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}_{\text{INI}}, \Lambda$
Define:
for $g = 1, \ldots, p + ms$:
$$\mathbf{G}_g = \mathbf{M}_g \otimes \mathbf{I}_k.$$
$\boldsymbol{B}_{\lambda_0,INI} \leftarrow \boldsymbol{B}_{INI}$
$\mathcal{A}_{\lambda_0,INI} \leftarrow \mathcal{A}_{INI}$
for $\lambda \in \Lambda_{\text{Grid}}$ **do**
  repeat
    $\boldsymbol{B}_{\lambda,\mathcal{A}_\lambda} \leftarrow$ThresholdUpdate$(\mathcal{A}_\lambda, \boldsymbol{B}_{\lambda,\mathcal{A}_\lambda}, \lambda)$
    $\boldsymbol{B}_{\lambda,\mathcal{A}_{\text{FULL}}}, \mathcal{A}_\lambda \leftarrow$BlockUpdate$(\mathcal{A}_{\text{FULL}}, \boldsymbol{B}_{\lambda,\mathcal{A}_\lambda}, \lambda)$
  until $\boldsymbol{B}_{\lambda,\mathcal{A}_\lambda} = \boldsymbol{B}_{\lambda,\mathcal{A}_{\text{FULL}}}$
  $\hat{v} \leftarrow \bar{\boldsymbol{Y}} - \boldsymbol{B}_{\lambda,\mathcal{A}}\bar{\boldsymbol{Z}}$
end for
return $\hat{v}, \boldsymbol{B}_\Lambda, A_\Lambda$
procedure BlockUpdate$(\mathcal{G}, \boldsymbol{B}_{\text{INI}}, \lambda)$ ▷ Makes one full pass through all groups
  $\boldsymbol{B} \leftarrow \boldsymbol{B}_{\text{INI}}$
  for $g \in \mathcal{G}$ **do**
    $\mathbf{R} \leftarrow \boldsymbol{B}_{-g}\boldsymbol{Z}_{-g} - \boldsymbol{Y}$
    $\mathbf{r} \leftarrow \mathbf{R}\boldsymbol{Z}_q^\top$
    if $\|\mathbf{r}\|_F \leq \lambda$ **then**
      $\boldsymbol{B}_g^* \leftarrow \mathbf{0}_{|g|}$
      $\mathcal{A}_g \leftarrow \emptyset$
    end if
    if $\|\mathbf{r}\|_F > \lambda$ **then**
      $\Delta \leftarrow$ the root of $\Omega(\Delta)$ defined in Eq. (21)
      $\text{vec}(\boldsymbol{B}_g) \leftarrow -(\mathbf{G}_g + \frac{\lambda}{\Delta}\mathbf{I})^{-1}\mathbf{r}$
      $\mathcal{A}_g \leftarrow g$
    end if
  end for
  return $\boldsymbol{B}_\lambda, \mathcal{A}$
end procedure
procedure ThresholdUpdate$(\mathcal{A}_\lambda, \boldsymbol{B}_{\lambda,\text{INI}}, \lambda)$ ▷ Iterates through active set until convergence
  if $\mathcal{A}_\lambda = \emptyset$ **then return** $\mathbf{0}_{k\times kp+ms}$
  end if
  if $\mathcal{A}_\lambda \neq \emptyset$ **then**
    $\boldsymbol{B}_{\lambda,\text{OLD}} \leftarrow \boldsymbol{B}_{\lambda,\text{INI}}$
    repeat
      $\boldsymbol{B}_{\lambda,\text{NEW}}, \mathcal{A}_\lambda \leftarrow$BlockUpdate$(\mathcal{A}_\lambda, \boldsymbol{B}_{\lambda,\text{OLD}}, \lambda)$
      $\boldsymbol{B}_{\lambda,\text{OLD}} \leftarrow \boldsymbol{B}_{\lambda,\text{NEW}}$
    until desired threshold is reached
  end if
return $\boldsymbol{B}_{\lambda,\text{NEW}}, \mathcal{A}$
end procedure

Algorithm 4 Sparse lag group VARX-L inner loop

Require: $\Phi_0, \boldsymbol{Z}_q, \mathbf{R}_{-q}, \alpha$
$h \leftarrow \frac{1}{\sigma_1(\boldsymbol{Z}_q)}$
$\Phi_0 \leftarrow \Phi^1$
repeat
  $j \leftarrow 1$
  $\boldsymbol{F}_q \leftarrow \frac{(\Phi^j\boldsymbol{Z}_q - \mathbf{R}_{-q})\boldsymbol{Z}_q^\top}{k}$
  $\gamma^j \leftarrow \Phi^j$
  $\text{vec}(\gamma^{(j+1)}) \leftarrow \left(1 - \frac{h(1-\alpha)\lambda}{\|ST(\Phi^j - h(\Phi^j\boldsymbol{Z}_q - \mathbf{R}_{-q})\boldsymbol{Z}_q^\top, h\alpha\lambda)\|_F}\right)_+ ST(\text{vec}(\Phi^j) - h\text{vec}(\boldsymbol{F}_q), h\alpha\lambda)$
  $\Phi^{j+1} \leftarrow \gamma^{j+1} + \frac{j}{j+3}(\gamma^{j+1} - \gamma^j)$
  $j \leftarrow j + 1$
until desired threshold is reached

Algorithm 5 Endogenous-first VARX-L proximal problem

Require: $\tilde{v}, \lambda, k, p, m, s$
for i=1,…,p **do**
  $g_1 \leftarrow [\big((i-1)\cdot k + 1\big) : \big((i-1)\cdot k + k\big)]$
  $g_2 \leftarrow [\big((i-1)\cdot m + p\cdot k\big) : \big((i-1)\cdot m + p\cdot k + m\big)]$
  $v_{\{g_1,g_2\}} \leftarrow \text{Prox}(v_{\{g_1,g_2\}}, \lambda, k)$
end for
return v.
procedure Prox(v,λ,k)
  $h_2 \leftarrow (k+1) : (k+m)$
  $h_1 \leftarrow 1 : (k+m)$
  for $j = 1, 2$ **do**
    $v_{h_j} \leftarrow (1 - \lambda/\|v_{h_j}\|_F)_+ v_{h_j}$
  end for
return v.
end procedure

References

Anderson, T. (1984). *Multivariate statistical analysis*. New York, NY: Wiley and Sons.
Banbura, M., Giannone, D., & Reichlin, L. (2009). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*, 71–92.
Basu, S., & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, *43*, 1535–1567.
Basu, S., Shojaie, A., & Michailidis, G. (2015). Network Granger causality with inherent grouping structure. *Journal of Machine Learning Research*, *16*(2), 417–453.
Beck, A., & Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, *2*, 183–202.
Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, *120*, 387–422.
Bernardi, M., & Catania, L. (2014). The model confidence set package for R. arXiv preprint arXiv:1410.8504.
Brooks, C., & Tsolacos, S. (2000). Forecasting models of retail rents. *Environment and Planning A*, *32*, 1825–1840.
Chiuso, A., & Pillonetto, G. (2010). Nonparametric sparse estimators for identification of large scale linear systems. In *2010 49th IEEE Conference on Decision and Control*, (CDC). (pp. 2942–2947). IEEE.
Clark, T. E., & McCracken, M. W. (2013). Evaluating the accuracy of forecasts from vector autoregressions. In *Advances in econometrics*: *Vol. 32. VAR Models in macroeconomics — new developments and applications: essays in honor of Christopher A. Sims* (pp. 117–168). Emerald Group Publishing Limited.
Cushman, D. O., & Zha, T. (1997). Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary Economics*, *39*, 433–448.
Davis, R. A., Zang, P., & Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, *25*, 1077–1096.
Diebold, F. X. (1998). The past, present, and future of macroeconomic forecasting. *The Journal of Economic Perspectives*, *12*, 175–192.
Doan, T., Litterman, R., & Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, *3*, 1–100.
Ensor, K. B. (2013). Time series factor models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*, 97–104.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.
Furman, Y. (2014). VAR Estimation with the Adaptive Elastic Net. Available at SSRN 2456510.
Gefang, D. (2012). Bayesian doubly adaptive elastic-net lasso for VAR shrinkage. *International Journal of Forecasting*, *30*, 1–11.
George, E. I., Sun, D., & Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics*, *142*, 553–580.
Golub, G. H., & Van Loan, C. F. (2012). *Matrix computations. Vol. 3*. JHU Press.
Hansen, P. C. (1998). *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion. Vol. 4*. Siam.
Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*, 453–497.

Hendry, D. F., & Hubrich, K. (2011). Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of Business and Economic Statistics*, *29*, 216–227.
Hsu, N. J., Hung, H. L., & Chang, Y. M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, *52*, 3645–3657.
Ibarra, R. (2012). Do disaggregated CPI data improve the accuracy of inflation forecasts? *Economic Modelling*, *29*, 1305–1313.
Ip, G. (2008). Non-borrowed reserves: False alarm. *The Wall Street Journal*, URL http://blogs.wsj.com/economics/2008/02/08/non-borrowed-reserves-false-alarm/.
Jenatton, R., Mairal, J., Obozinski, G., & Bach, F. (2011). Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*, *12*, 2297–2334.
Kadiyala, K., & Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics*, *12*, 99–132.
Kennedy, P. (2003). *A guide to econometrics*. MIT press.
Klein, L. R., & Goldberger, A. S. (1955). *An econometric model of the United States, 1929–1952*. Amsterdam: North-Holland Publishing Company.
Kock, A. B., & Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics*, *186*, 325–344.
Koop, G. (2013). Forecasting with medium and large Bayesian VARs. *Journal of Applied Econometrics*, *28*, 177–203.
Li, J., & Chen, W. (2014). Forecasting macroeconomic time series: Lasso-based approaches and their forecast combinations with dynamic factor models. *International Journal of Forecasting*, *30*, 996–1015.
Litterman, R.B. (1979). Techniques of forecasting using vector autoregressions. Working papers, Federal Reserve Bank of Minneapolis.
Litterman, R. B. (1984). Forecasting and policy analysis with Bayesian vector autoregression models. *Federal Reserve Bank of Minneapolis Quarterly Review*, *8*, 30–41.
Litterman, R. B. (1986a). Forecasting with Bayesian vector autoregressions — five years of experience. *Journal of Business and Economic Statistics*, *4*, 25–38.
Litterman, R. B. (1986b). A statistical approach to economic forecasting. *Journal of Business and Economic Statistics*, *4*, 1–4.
Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Cambridge Univ Press.
Lütkepohl, H. (2014). Structural vector autoregressive analysis in a data rich environment: A survey. DIW Berlin Discussion Paper.
Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics*, *135*, 499–526.
Neumaier, A., & Schneider, T. (2001). Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software*, *27*, 27–57.
Nicholson, W.B., Bien, J., & Matteson, D.S. (2014). High Dimensional Forecasting via Interpretable Vector Autoregression. arXiv preprint arXiv:1412.5250.
Nicholson, W.B., Matteson, D.S., & Bien, J. (2016). BigVAR Tools for modeling sparse high-dimensional multivariate time series. arXiv preprint arXiv:1702.07094.
Nijs, V. R., Srinivasan, S., & Pauwels, K. (2007). Retail-price drivers and retailer profits. *Marketing Science*, *26*, 473–487.
Nocedal, J., & Wright, S. J. (1999). *Numerical optimization. Vol. 2*. New York: Springer.
Ocampo, S., & Rodríguez, N. (2012). An introductory review of a structural VAR-X estimation and applications. *Revista Colombiana de Estadística*, *35*, 479–508.
Penm, J. H., Penm, J. H., & Terrell, R. (1993). The recursive fitting of subset VARX models. *Journal of Time Series Analysis*, *14*, 603–619.
Qin, Z., Scheinberg, K., & Goldfarb, D. (2010). Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation*, *5*, 1–27.
Racette, D., & Raynauld, J. (1992). Canadian monetary policy: will the checklist approach ever get us to price stability? *Canadian Journal of Economics*, *25*, 819–838.
Robertson, J. C., & Tallman, E. W. (1999). *Improving forecasts of the federal funds rate in a policy model. Technical Report*. Federal Reserve Bank of Atlanta.
Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*, 231–245.
Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, *48*, 1–48.
Song, S., & Bickel, P. (2011). Large vector auto regressions. arXiv preprint arXiv:1106.3915.
Stock, J.H., & Watson, M.W. (2005). An empirical comparison of methods for forecasting using many predictors. Manuscript, Princeton University.
Stock, J. H., & Watson, M. W. (2006). Forecasting with many predictors. *Handbook of Economic Forecasting*, *1*, 515–554.
Taylor, J. B. (1993). Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy. Vol. 39* (pp. 195–214). Elsevier.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*, 267–288.
Todd, R. M. (1990). Improving economic forecasting with Bayesian vector autoregression. In C. Granger (Ed.), *Modelling economic series* (pp. 214–234). Oxford: Clarendon Press.
Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, *109*, 475–494.
Welfe, W. (2013). Macroeconometric models of the United States and Canada. In *Macroeconometric models* (pp. 15–46). Springer.
Wood, B. D. (2009). Presidential saber rattling and the economy. *American Journal of Political Science*, *53*, 695–709.
Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *68*, 49–67.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*, 301–320.