Structured Regularization for Large Vector Autoregressions with Exogenous Variables

William B. Nicholson*, David S. Matteson†, and Jacob Bien‡

June 7, 2015

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 forecasting. One of the major shortcomings of the VAR that has hindered 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, forecasting using VARs 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 than their smaller counterparts. Existing methods that allow for the estimation of large VARs either tend to require *ad hoc* subjective specifications or are computationally infeasible. Moreover, as global economies become more intricately intertwined, there has been substantial interest in incorporating the impact of stochastic, unmodeled *exogenous* variables. Vector autoregression with exogenous variables (VARX) is a straightforward extension of the VAR that allows for the inclusion of unmodeled variables, but it also faces dimensionality challenges.

We adapt several prominent scalar regression regularization techniques to a vector time series context to greatly reduce the parameter space of VAR and VARX models. We formulate convex optimization procedures that are amenable to efficient solutions for the time-ordered high-dimensional problems we aim to solve. Through this framework, we propose a structured family of models and provide implementations which allow for both the efficient estimation and accurate forecasting of high-dimensional VAR and VARX models. We also highlight an extension which allows for shrinking toward reference models. We demonstrate the efficacy of our approaches in simulated data examples as well as both low and high-dimensional macroeconomic applications.

1 Introduction

The practice of macroeconomic forecasting was spearheaded by Klein and Goldberger [1955], whose eponymous simultaneous equation system jointly forecasted the behavior of 15 annual macroeconomic indicators, including con-

*PhD Candidate, Department of Statistical Science, Cornell University, 301 Malott Hall, Ithaca, NY 14853 (E-mail: wbn8@cornell.edu; Webpage: `http://www.wbnicholson.com`)

†Assistant Professor, Department of Statistical Science and Department of Social Statistics, Cornell University, 1196 Comstock Hall, Ithaca, NY 14853, (E-mail: matteson@cornell.edu; Webpage: `https://courses.cit.cornell.edu/~dm484/`)

‡Assistant Professor, Department of Biological Statistics and Computational Biology and Department of Statistical Science, Cornell University, 1178 Comstock Hall, Ithaca, NY 14853 (E-mail: jbien@cornell.edu; Webpage: `http://faculty.bscb.cornell.edu/~bien/`)

sumer expenditure, interest rates, and corporate profits. The parameterization and identification restrictions of these models were heavily influenced 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 as to the overall state of the economy and to influence policy decisions (Welfe [2013]). As the Klein-Goldberger model and its extensions were primarily motivated by Keynesian economic theory, the collapse of the Bretton Woods monetary system and severe oil price shocks led to widespread forecasting failure in the 1970s (Diebold [1998]). At this time, the vector autoregression (VAR), popularized by Sims [1980], emerged as an atheoretical forecasting technique underpinned by statistical methodology and not subject to the ebb and flow of contemporary macroeconomic theory.

In many applications, a VAR's forecasts can be improved by incorporating unmodeled exogenous variables, which are determined outside of the VAR. Examples of exogenous variables are context-dependent and can include leading indicators and weather-related characteristics. If a small open economy is being examined, global macroeconomic variables, such as world oil prices, can be considered exogenous. Such models are most commonly referred to as "VARX" in the econometrics literature, but in other fields they can go by "transfer function" or "distributed lag."

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] use a structural VARX model to analyze the effect of monetary policy shocks in Canada. The VARX is also amenable under scenarios in which forecasts are only desired from a subset of the included series in a VAR, as its corresponding VARX has a reduced parameterization. VARX models have received considerable attention not just in economics, but also marketing (Nijs et al. [2007]), political science (Wood [2009]), and real estate (Brooks and Tsolacos [2000]).

Unfortunately, dimensionality issues have limited the utility of the VARX. In the conventional VAR context, most applications are limited to no more than 6 series (cf. Bernanke et al. [2005]), forcing the practitioner to specify *a priori* a reduced subset of series to include. The VARX faces similar restrictions. As outlined in Penm et al. [1993], lag order, the maximum number of lagged observations to include, may differ between modeled and exogenous series. Hence, in order to select an optimal 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 modeled (which we will refer to as "endogenous" throughout the paper) and exogenous series. Moreover, unlike the conventional VAR, zero constraints (restrictions forcing certain model parameters to zero) are generally expected.

Shortly after the VAR's inception, efforts were made to reduce its parameterization. Early attempts, such as 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 reasonably modeled by a univariate random walk via shrinking

estimated models toward univariate unit root processes.

The Bayesian VAR with a Minnesota Prior was shown by Robertson and Tallman [1999] to produce forecasts superior to 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 it requires the specification of several hyperparameters.

Modern Bayesian VAR extensions originally proposed in Kadiyala and Karlsson [1997] and compiled by Koop [2011] impose restrictions on the parameter space while allowing for more general covariance specifications and 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] extends 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 estimation 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 proposed by Hsu et al. [2008] and further explored in Song and Bickel [2011] and Davis et al. [2012]. Theoretical properties were investigated by Kock and Callot [2013] and by Basu and Michailidis [2013]. The Lasso-VAR has several advantages over the Bayesian VAR as it is more computationally tractable in high dimensions, performs variable selection, and can readily compute multi-step forecasts and their associated prediction intervals.

As it is often viewed as an economic rather than statistical problem, reducing the parameter space of the VARX model has received considerably less attention. Ocampo and Rodríguez [2012] extend the aforementioned Bayesian VAR estimation methods to the VARX context. George et al. [2008] apply stochastic search variable selection to the VARX framework which provides a data-driven method to determine zero restrictions, but their approach is not scalable to high dimensions. Chiuso and Pillonetto [2010] propose estimating a VARX model with Lasso and Group Lasso penalties but do not elaborate on potential group structures.

This paper seeks to bridge the considerable gap between the regularization and macroeconomic modeling communities. We develop a coherent framework for penalized VARX estimation while incorporating the unique structure of the VARX model in a computationally efficient manner. Our methods; the Lag Group Lasso VARX, Own/Other Group Lasso VARX, Sparse Lag Group Lasso VARX, Sparse Own/Other Group Lasso VARX, Lasso VARX, and Endogenous-First VARX, extend the Lasso and its structured counterparts to take 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. These models offer great flexibility in capturing the true underlying dynamics of an economic system while imposing minimal assumptions on the parameter space. Moreover, unlike previous approaches, due to our adaptation of convex optimization algorithms to

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

Section 2 describes the notation used throughout the paper and introduces our structured regularization methodology. Section 3 provides our implementation details and presents both macroeconometric applications and a simulation study, Section 4 details the "Minnesota Lasso:" an extension that allows for the incorporation of mild nonstationarity by shrinking toward a vector random walk, and Section 5 contains our conclusion. The Appendix elaborates upon our solution strategies and algorithms.

2 Methodology

2.0 The VARX Model

A k-dimensional multivariate time series $\{\mathbf{y}_t\}_{t=1}^T$ and 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

$$\mathbf{y}_t = \boldsymbol{\nu} + \sum_{\ell=1}^{p} \boldsymbol{B}^{(\ell)} \mathbf{y}_{t-\ell} + \sum_{j=1}^{s} \boldsymbol{\theta}^{(j)} \mathbf{x}_{t-j} + \mathbf{u}_t \quad \text{for} \quad t = 1, \ldots, T, \tag{2.1}$$

in which $\boldsymbol{\nu}$ denotes a k-dimensional constant intercept vector, $\boldsymbol{B}^{(\ell)}$ represents a $k \times k$ endogenous coefficient matrix at lag $\ell = 1, \ldots, p$, $\boldsymbol{\theta}^{(j)}$ 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 Equation (2.1) with $\boldsymbol{\theta}^{(j)} = \mathbf{0}$ for $j = 1, \ldots, s$.

Note that throughout this paper, we use the terms endogenous and exogenous simply to denote modeled and unmodeled series, respectively. We do not make any assumptions regarding the underlying economic dynamics between series.

In a low dimensional setting, in which the number included series and maximum lag order $kp+ms$ is substantially less than the length of the series, T, the VARX can be estimated by multivariate least squares, in which $\boldsymbol{\nu}, \boldsymbol{B} = [\boldsymbol{B}^{(1)}, \ldots, \boldsymbol{B}^{(p)}]$, and $\boldsymbol{\theta} = [\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(s)}]$) are chosen as

$$\min_{\boldsymbol{\nu}, \boldsymbol{B}, \boldsymbol{\theta}} \sum_{t=1}^{T} \|\mathbf{y}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{p} \boldsymbol{B}^{(\ell)} \mathbf{y}_{t-\ell} - \sum_{j=1}^{s} \boldsymbol{\theta}^{(j)} \mathbf{x}_{t-j}\|_F^2, \tag{2.2}$$

in which $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$ denotes the Frobenius norm of a matrix A. In the absence of regularization, the $\text{VARX}_{k,m}(p,s)$ requires the estimation of $k(1+kp+ms)$ parameters. In the following section, we will apply several convex penalties to Equation (2.2) which aid in reducing the parameter space of the VARX by imposing sparsity in $\boldsymbol{B}$ and $\boldsymbol{\theta}$.

2.1 Structured Penalties for VARX Modeling

In this section, we introduce a general penalized framework for VARX models. We consider structured objectives of the form

$$\min_{\boldsymbol{B},\boldsymbol{\theta},\boldsymbol{\nu}} \sum_{t=1}^{T} \|\mathbf{y}_t - \boldsymbol{\nu} - \sum_{\ell=1}^{p} \boldsymbol{B}^{(\ell)}\mathbf{y}_{t-\ell} - \sum_{j=1}^{s} \boldsymbol{\theta}^{(j)}\mathbf{x}_{t-j}\|_F^2 + \lambda\Big(\mathcal{P}_y(\boldsymbol{B}) + \mathcal{P}_x(\boldsymbol{\theta})\Big), \tag{2.3}$$

in which $\lambda \geq 0$ is a penalty parameter selected according to a procedure that is discussed in Section 3.1, $\mathcal{P}_y(\boldsymbol{B})$ represents the group penalty structure on endogenous coefficients, and $\mathcal{P}_x(\boldsymbol{\theta})$ represents the exogenous group penalty structure. Table 1 details the penalty structures considered in this paper. In the following section, we will discuss each penalty structures in greater detail.

Equations (2.4)-(2.5) adapt the univariate *Group Lasso* penalty proposed by Yuan and Lin [2006] to the VARX setting. Our choices of $\mathcal{P}_x$ and $\mathcal{P}_y$ create structured sparsity based on pre-specified groupings, which are designed to incorporate the intrinsic structure of the VARX. Our methods have a substantial advantage over popular Bayesian approaches in that they will both shrink least squares estimates toward zero as well as 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 overparameterized. Our methods allow for feasible estimation even under scenarios in which the number of regression parameters is close to or exceeds the number of scalar observations kT.

The Group Lasso penalty function was explored in the VAR context by Song and Bickel [2011] who consider several structured groupings with a particular emphasis on creating a distinction between a series's own lags and those of another series. Theoretical properties of the use of a Group Lasso penalty in the VAR setting were further explored by Basu et al. [2012].

A disadvantage of the Group Lasso VARX 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 a multi-dimensional gridsearch to select penalty parameters. Equations (2.6)-(2.7) instead implement the *Sparse Group Lasso* penalty (Simon et al. [2013]), which extends the Group Lasso to allow within-group sparsity.

The Lasso, Equation (2.8), considers no structure, but in certain scenarios may have computational advantages as compared to more complex approaches. The Endogenous-First Group Lasso incorporates a nested penalty structure that, within a lag, prioritizes endogenous coefficients before their exogenous counterparts.

Table 1: VARX Penalty Functions. Note that $\boldsymbol{B}_{\text{on}}^{(\ell)}$ denotes the diagonal elements of coefficient matrix $\boldsymbol{B}^{(\ell)}$ whereas $\boldsymbol{B}_{\text{off}}^{(\ell)}$ denotes the off-diagonal elements.

Name	$\mathcal{P}_y(\boldsymbol{B})$	$\mathcal{P}_x(\boldsymbol{\theta})$
(2.4) Lag	$\sqrt{k^2}\sum_{\ell=1}^{p}\|\boldsymbol{B}^{(\ell)}\|_F$	$\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\theta}_{\cdot,i}^{(j)}\|_F$
(2.5) Own/Other	$\sqrt{k}\sum_{\ell=1}^{p}\|\|\boldsymbol{B}_{\text{on}}^{(\ell)}\|\|_F + \sqrt{k(k-1)}\sum_{\ell=1}^{p}\|\|\boldsymbol{B}_{\text{off}}^{(\ell)}\|\|_F$	$\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\theta}_{\cdot,i}^{(j)}\|_F$
(2.6) Sparse Lag	$(1-\alpha)\sqrt{k^2}\sum_{\ell=1}^{p}\|\boldsymbol{B}^{(\ell)}\|_F + \alpha\|\boldsymbol{B}\|_1$	$(1-\alpha)\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\theta}_{\cdot,i}^{(j)}\|_F + \alpha\|\boldsymbol{\theta}\|_1$
(2.7) Sparse Own/Other	$(1-\alpha)\big(\sqrt{k}\sum_{\ell=1}^{p}\|\|\boldsymbol{B}_{\text{on}}^{(\ell)}\|\|_F + \sqrt{k(k-1)}\sum_{\ell=1}^{p}\|\|\boldsymbol{B}_{\text{off}}^{(\ell)}\|\|_F\big) + \alpha\|\boldsymbol{B}\|_1$	$(1-\alpha)\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\theta}_{\cdot,i}^{(j)}\|_F + \alpha\|\boldsymbol{\theta}\|_1$
(2.8) No Grouping	$\|\boldsymbol{B}\|_1$	$\|\boldsymbol{\theta}\|_1$
(2.9) Endogenous-First	$\mathcal{P}_{x,y}(\boldsymbol{B},\boldsymbol{\theta}) = \sum_{\ell=1}^{p}\sum_{i=1}^{k}\left(\|[\boldsymbol{B}_{j}^{(\ell)}, \boldsymbol{\theta}_{i,\cdot}^{(\ell)}]\|_F + \|\boldsymbol{\theta}_{i,\cdot}^{(\ell)}\|_F\right)$	

Group Lasso VARX

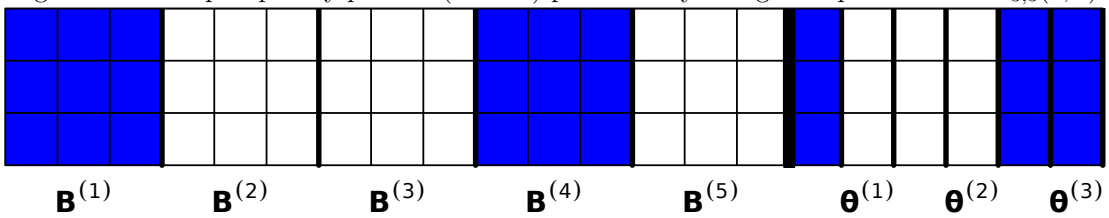
We first present the *Lag* Group Lasso VARX (2.4), in which the endogenous coefficients are grouped according to their lag matrix $\boldsymbol{B}^{(\ell)}$ for $\ell = 1, \ldots, p$, and each exogenous series is partitioned into a separate group. This structured grouping is expressed as

$$\mathcal{P}_y(\boldsymbol{B}) = \sqrt{k^2}\sum_{\ell=1}^{p}\|\boldsymbol{B}^{(\ell)}\|_F, \tag{2.10}$$

$$\mathcal{P}_x(\boldsymbol{\theta}) = \sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\theta}_{\cdot,i}^{(j)}\|_F. \tag{2.11}$$

Note that since groups differ in cardinality, it is required to weight the penalty to avoid regularization favoring larger groups. This structure implies that for each endogenous series, a coefficient at lag ℓ is either nonzero for all series or zero for all. Similarly, the relationship between an exogenous and endogenous series at lag i 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 shaded in Figure 1.

Figure 1: Example sparsity pattern (shaded) produced by a Lag Group Lasso-VARX$_{3,3}(5,3)$



In comparison to Bayesian regularization methods, such as the stochastic search variable selection approach proposed by George et al. [2008], solving the Group Lasso VARX is tractable even in high dimensions (we have considered as high as $k = 40, m = 128, T = 195$). We are able to extend the efficient Group Lasso solution method proposed by Qin et al. [2010], who utilize a block coordinate descent procedure and transform each "one group" subproblem to a trust-region framework. These subproblems can then be solved efficiently via univariate

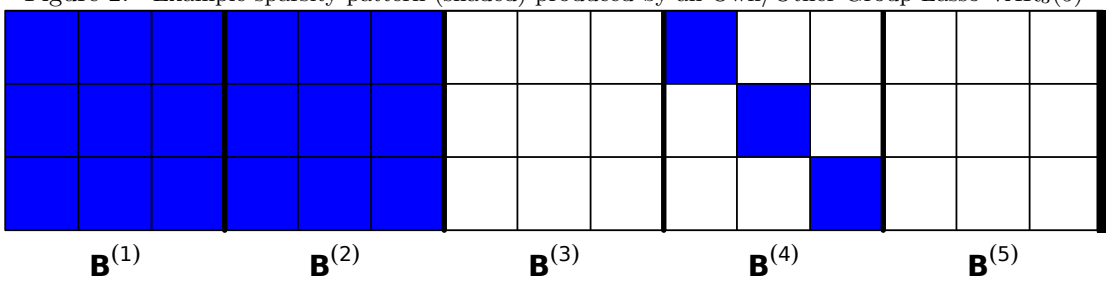
optimization. Details of this procedure are provided Section A.3.2 of the Appendix.

The Lag Group Lasso structure is advantageous for applications in which all endogenous component series tend to exhibit comparable dynamics, while allowing more flexibility for exogenous structures. It also can serve as a powerful tool for lag selection. However, in many settings, it may not be appropriate to give equal consideration to every entry in a coefficient matrix. Diagonal entries of $\boldsymbol{B}$, which represent a series' own lags, are in many applications more likely to be nonzero than are off-diagonal entries, which represent cross dependence with other components. The *Own/Other* Group Lasso VARX (2.5) allows for the partitioning of each lag matrix in $\boldsymbol{B}$ into separate groups by assigning the endogenous penalty structure

$$\mathcal{P}_y(\boldsymbol{B}) = \sqrt{k}\sum_{\ell=1}^{p}||\boldsymbol{B}_{\text{on}}^{(\ell)}||_F + \sqrt{k(k-1)}\sum_{\ell=1}^{p}||\boldsymbol{B}_{\text{off}}^{(\ell)}||_F. \tag{2.12}$$

An example of this sparsity pattern is shown in Figure 2. The modifications required to utilize the Own/Other structure are detailed in Section A.3.3 in the Appendix.

Figure 2: Example sparsity pattern (shaded) produced by an Own/Other Group Lasso-VAR$_3$(5)



Sparse Group Lasso VARX

For certain applications, a Group Lasso 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 substantially increases computation time. Moreover, it is inefficient to include an entire group if only one coefficient is nonzero. The *Sparse Group Lasso*, proposed by Simon et al. [2013] allows for within-group sparsity through a convex combination of Lasso and Group Lasso penalties. The *Sparse Lag Group Lasso* VARX (2.6) results in a penalty of the form

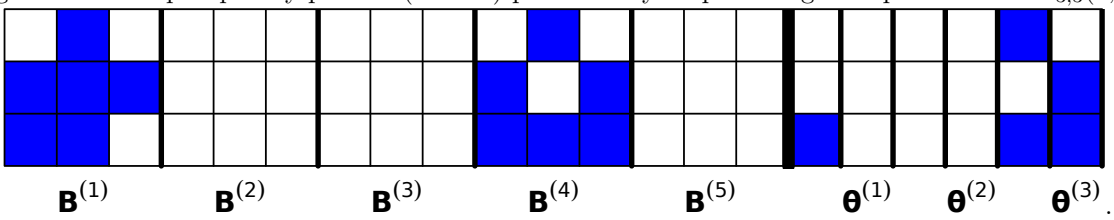
$$\mathcal{P}_y(\boldsymbol{B}) = (1-\alpha)\big(\sqrt{k^2}\sum_{\ell=1}^{p}\|\boldsymbol{B}^{(\ell)}\|_F\big) + \alpha\|\boldsymbol{B}\|_1, \tag{2.13}$$

$$\mathcal{P}_x(\boldsymbol{\theta}) = (1-\alpha)\big(\sqrt{k}\sum_{j=1}^{s}\sum_{i=1}^{m}\|\boldsymbol{\theta}_{\cdot,i}^{(j)}\|_F\big) + \alpha\|\boldsymbol{\theta}\|_1, \tag{2.14}$$

in which $0 \leq \alpha \leq 1$ is an additional penalty parameter that controls within-group sparsity. Without prior knowledge, we weight according to relative group sizes and set α to $\frac{1}{k+1}$, though α could be estimated by cross-validation. The

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

Figure 3: Example sparsity pattern (shaded) produced by a Sparse Lag Group Lasso-VARX$_{3,3}(5,3)$



Since the inclusion of within-group sparsity does not create a separable objective function, conventional solution methods, such as coordinate descent, are no longer applicable. Following Simon et al. [2013], our solution to the Sparse Group Lasso VARX makes use of proximal gradient descent. The details of this approach and our implementation are provided in Section A.3.4 of the Appendix. The Sparse Group Lasso VARX can also be extended to alternative groupings. Consequently, we also offer the Sparse "Own/Other" Group Lasso VARX (2.7) as an estimation procedure.

Lasso VARX

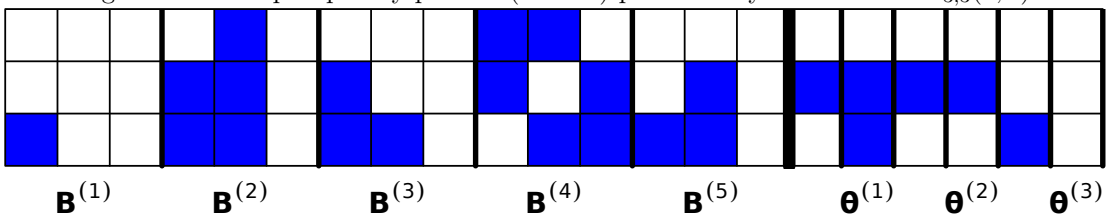
The Lasso VARX (2.8), proposed by Chiuso and Pillonetto [2010], incorporates no structure and can be viewed as a special case of the Sparse Group Lasso VARX in which $\alpha = 1$, resulting in penalties of the form

$$\mathcal{P}_y(\boldsymbol{B}) = \|\boldsymbol{B}\|_1, \tag{2.15}$$

$$\mathcal{P}_x(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1. \tag{2.16}$$

The L_1 penalty will induce sparsity in the coefficient matrices $\boldsymbol{B}$ and $\boldsymbol{\theta}$ by zeroing individual entries. An example sparsity pattern is depicted in Figure 4.

Figure 4: Example sparsity pattern (shaded) produced by a Lasso-VARX$_{3,3}(5,2)$



A major advantage of the Lasso VARX over its structured counterparts is its computational tractability. Our solution approach involves the use of coordinate descent, popularized by Friedman et al. [2010]. Coordinate descent consists of partitioning the Lasso VARX into scalar subproblems for each $[\boldsymbol{B}, \boldsymbol{\theta}]_{ij}$, solving component-wise, then updating until convergence. This approach is computationally efficient since, in the Lasso VARX context, each

subproblem has a closed-form solution. Tseng [2001] establishes that global convergence arises from solving individual subproblems in the coordinate descent framework. Our solution strategy is detailed Section A.3.1 of the Appendix.

2.1.1 Endogenous-First Structure

We have previously only considered structures that assign endogenous and exogenous variables to separate groups. In this section, we consider a nested structure that can take into account the relative importance between endogenous and exogenous series.

In certain scenarios, there may exist an *a priori* importance ranking among endogenous and exogenous variables. For example, the endogenous variables could represent economic indicators of interest in a small open economy, and the exogenous variables global macroeconomic indicators. 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.

We can consider such a structure by utilizing a *hierarchical group lasso* penalty proposed by Jenatton et al. [2011]. The Endogenous-First Group Lasso VARX (2.9) takes the form

$$\mathcal{P}_{x,y}(\boldsymbol{B},\boldsymbol{\theta}) = \sum_{\ell=1}^{p}\sum_{j=1}^{k} \|[\boldsymbol{B}_j^{(\ell)}, \boldsymbol{\theta}_{j,\cdot}^{(\ell)}]\|_F + \|\boldsymbol{\theta}_{j,\cdot}^{(\ell)}\|_F. \tag{2.17}$$

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 (2.17) decouples across rows, allowing for separate nested structures across each endogenous series. This sparsity pattern is depicted in Figure 5.

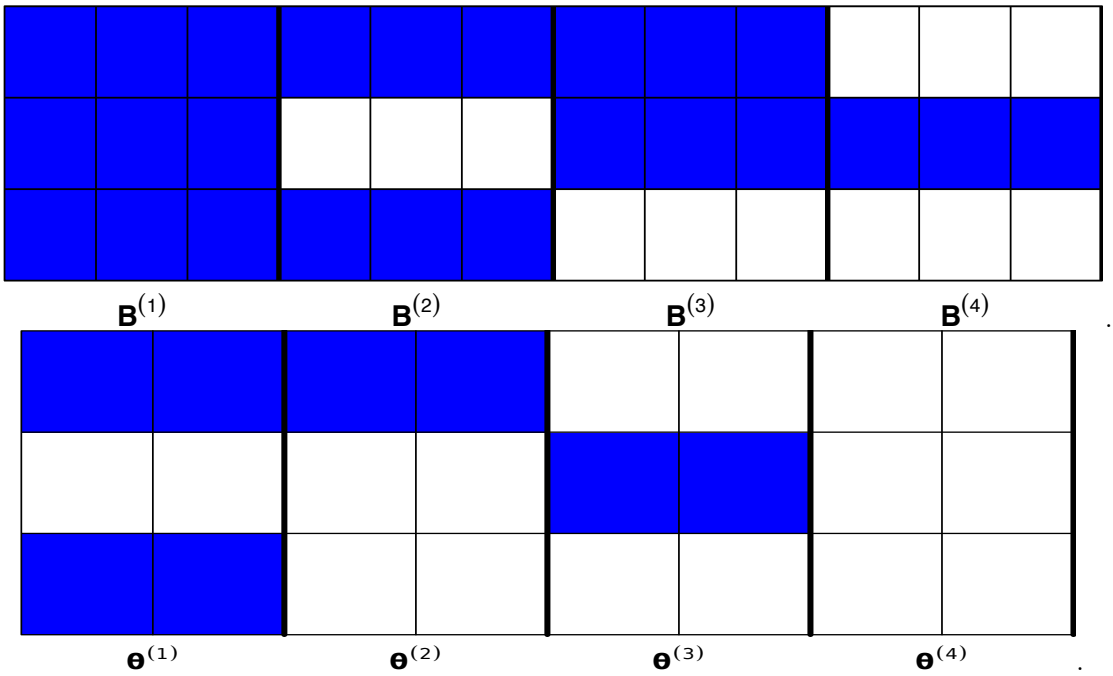


Figure 5: Example Sparsity Pattern Generated by an Endogenous-First VARX with $k = 3, p = 4, m = 2, s = 4$. Notice that a row in $\boldsymbol{\theta}^{(i)}$ can only be nonzero if the corresponding row in $\boldsymbol{B}^{(i)}$ is also nonzero.

Most Group Lasso solution methods, such as block coordinate descent, take advantage of the separability of groups to improve computational performance. Although the nested structure is not directly separable, based on the methodology of Jenatton et al. [2011] its dual can be solved in one pass of block coordinate descent. Details of the solution approach are provided in Section A.3.5 of the Appendix.

3 High Dimensional Macroeconometrics

In this section, we start by evaluating our regularization procedures in two macroeconomic data applications: one high-dimensional and one low-dimensional. In our first application, we consider applying our procedures on the widely used set of US macroeconomic indicators originally constructed by Stock and Watson [2005]. Our second example considers forecasting a small set of Canadian macroeconomic indicators and incorporating the previous indicators as exogenous series.

In addition, we examine the performance of our procedures on several simulated high-dimensional time series conforming to different sparsity patterns, with one constructed to be advantageous for each proposed structure. Section 3.1 outlines 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 the results of our simulation scenarios.

3.1 Practical Implementation

Penalty Parameter Selection

The regularization parameter, λ, is not known in practice and is typically estimated via cross-validation. In this section, we detail 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 in which all components of $[\boldsymbol{B}, \boldsymbol{\theta}]$ will be zero, and decreases in log-linear increments. This value differs for each procedure and can be inferred by their respective algorithms. The starting values are summarized in Table 9 located in Section A.4 of the Appendix. The number of gridpoints, n, as well as the depth of the grid are left to user input.

Due to time-dependence, our problem is not well-suited to traditional j-fold cross-validation. Instead, following Banbura et al. [2009], we propose choosing the optimal penalty parameter by minimizing one-step-ahead mean-square forecast error (MSFE). 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_1 and forecast $\hat{\mathbf{y}}_{T_1+1}^{\lambda_i}$ for $i = 1, \ldots, n$. We then sequentially add one observation at a time and repeat this process until we reach time T_2. This procedure is illustrated in Figure 6.

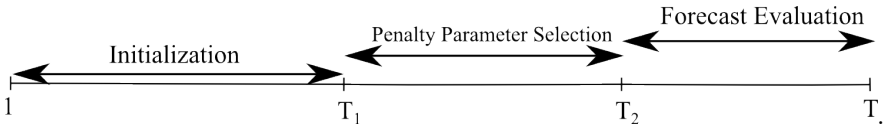


Figure 6: Illustration of Rolling Cross-Validation

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

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

Finally, from time T_2 to T, we evaluate the one-step-ahead forecast accuracy of $\hat{\lambda}$. If desired, additional forecast horizons or criterion functions can be substituted. 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 substantially decreases computation time. The penalty parameter selection procedure is expressed in Algorithm 2 in the Appendix.

3.2 Methods for Comparison

A conventional VAR model selection approach in a low-dimensional setting involves fitting a $\text{VAR}_k(\ell)$ by least squares for $0 \leq \ell \leq p$ and selecting a lag order based on an information criterion, such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC). Per Lütkepohl [2005], the AIC and BIC of a $\text{VAR}_k(\ell)$ are defined as

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

in which $\widehat{\Sigma}_u^\ell$ is the residual sample covariance matrix obtained from the estimated $\text{VAR}_k(\ell)$, and $|\Sigma|$ represents the determinant of Σ. The estimated lag order $\hat{p}$ is then chosen as the minimizer of AIC or BIC. AIC penalizes each model coefficient uniformly by a factor of 2 whereas BIC scales penalties according to series length. Hence, when T is large, BIC will tend to select more parsimonious models than will AIC.

We utilize AIC and BIC to select lag order and to ensure numerically-stable results, we construct our least squares estimates using a slight variation of the approach developed by Neumaier and Schneider [2001] that adds a small (on the order of $\epsilon_{\text{machine}}$) ridge penalty.

We additionally compare our methods against two naive approaches that provide insight with regard to the level of temporal dependence in the data. We first consider the unconditional *sample mean*, which will make one-step-ahead forecasts at time $t+1$ based upon the average of all observed data up to time t: $\hat{\mathbf{y}}_{t+1} = \frac{1}{t}\sum_{i=1}^{t} \mathbf{y}_i$. Scenarios in which the sample mean forecasts well relative to more sophisticated procedures imply weak temporal dependence.

Second, we consider the vector *random walk* model, which makes one-step-ahead forecasts based upon the last

observed realization of the series, i.e. $\hat{\mathbf{y}}_{t+1} = \mathbf{y}_t$. Superior performance of the vector random walk indicates a strong degree of temporal dependence, which is often observed in macroeconomic data.

3.3 Macroeconometric Applications

We evaluated our methods on the widely used large macroeconomic dataset created by Stock and Watson [2005] and later improved by Koop [2011]. The dataset consists of 168 quarterly US macroeconomic indicators containing information about various aspects of the economy, including income, industrial production, employment, stock prices, interest rates, exchange rates, etc. The data ranges from Quarter 1 of 1959 to Quarter 4 of 2007 ($T = 195$). Per Koop [2011], the series can be partitioned into several groups; we consider the following three:

- *Small*: 3 variables (Federal Funds Rate, Consumer Price Index, Gross Domestic Product growth rate): Core group, typically used in simple Dynamic Stochastic Generalized Equilibrium models ($k = 3$),
- *Medium*: Adds to small 17 additional variables containing aggregated economic information (e.g. consumption, labor, housing, exchange rates) ($k = 20$),
- *Medium-Large*: Adds to medium 20 additional aggregate variables ($k = 40$),

For a detailed description of each set of variables, consult Koop [2011]. As Banbura et al. [2009] found that the greatest improvements in forecast performance occurred with the *medium* VAR, that will be our focus. We will attempt to forecast the *medium* set of indicators while using the remaining variables in the *medium-large* model as exogenous predictors. Before estimation, each series is transformed to stationarity according to the transformation codes provided by Stock and Watson [2005] and standardized by subtracting the sample mean and dividing by the sample standard deviation. Quarter 3 of 1977 to Quarter 3 of 1992 is used for penalty parameter selection while Quarter 4 of 1992 to Quarter 4 of 2007 is used for forecast evaluation. Our results are summarized in Table 2.

Table 2: Out of sample MSFE of one-step ahead forecasts of 20 macroeconomic indicators with 20 exogenous predictors $p = 4, s = 4$, $T = 195$

Model/ VARX Penalty Structure	MSFE	Standard Error
Lasso	12.248	1.924
Lag Group Lasso	12.998	1.903
Own/Other Group Lasso	11.517	1.689
Sparse Lag Group Lasso	12.184	1.912
Sparse Own/Other Group Lasso	11.780	1.670
Endogenous-First VARX	12.521	2.008
VAR with lag selected by AIC	22.814	2.614
VAR with lag selected by BIC	13.236	1.798
Sample Mean	15.135	2.279
Random Walk	30.196	5.396

Our procedures substantially outperform the benchmarks, with the Own/Other Group Lasso and Sparse Own/Other Group Lasso VARX achieving the best performance, providing evidence that making the distinction between a series'

own lags and those of other series can improve forecasts in practice.

Canadian Macroeconomic Data Application

We also consider a low-dimensional application in which we forecast Canadian series using US macroeconomic variables 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, Racette and Raynauld [1992] and Cushman and Zha [1997] show 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 4 Canadian macroeconomic series using our previously defined *medium* dataset as exogenous predictors. The added series are

1. M1 (a measure of the liquid components of money supply)
2. Industrial Production
3. GDP (relative to 2000)
4. Canada/US exchange rate

Quarter 1 of 1977 to Quarter 1 of 1992 is used for penalty parameter selection while Quarter 2 of 1992 to Quarter 4 of 2007 is used for forecast evaluation. In addition to the standard measures, we also compare against our procedures in the VAR framework, in which the exogenous predictors are ignored. Our results are summarized in Table 3.

Table 3: Out of sample MSFE of one-step ahead VARX forecasts of 4 Canadian macroeconomic indicators with 20 exogenous predictors $p = 4, s = 4$ and VAR forecasts of 4 Canadian macroeconomic indicators, $p = 4$, $T = 195$

VARX Penalty Structure	MSFE	Standard Error
Lasso	2.996	0.428
Lag Group Lasso	2.988	0.424
Own/Other Group Lasso	2.995	0.427
Sparse Lag Group Lasso	2.959	0.443
Sparse Own/Other Group Lasso	2.984	0.433
Endogenous-First VARX	3.033	0.443
VAR Penalty Structure	MSFE	S.E
Lasso	3.027	0.468
Lag Group Lasso	3.075	0.455
Own/Other Group Lasso	3.303	0.468
Sparse Lag Group Lasso	3.042	0.466
Sparse Own/Other Group Lasso	3.037	0.465
VAR with lag selected by AIC	3.341	0.496
VAR with lag selected by BIC	3.201	0.433
Sample Mean	3.052	0.395
Random Walk	4.545	0.446

Even at this low dimension, despite a weak temporal dependence as evidenced by the relatively strong performance of the sample mean, we find that all of our methods substantially outperform the AIC and BIC benchmarks, with the Sparse Lag Group Lasso VARX achieving superior performance. Moreover, we find that our methods are able to

effectively leverage relevant information from the exogenous predictors, as every VARX procedure outperforms its corresponding VAR procedure.

3.4 Simulation Scenarios

In this section, we consider evaluating the forecasting performance of our procedures on several simulated high-dimensional time series. All simulations operate on a $\text{VARX}_{10,10}(4,4)$ of length $T = 100$, and each simulation is repeated 100 times. The choice of 4 was selected for p and s because it represents one year of dependence for quarterly series, which is a common frequency of macroeconomic data. The middle third of the data is used for penalty parameter selection while the last third is used for forecast evaluation. Under each scenario, Σ_u is distributed according to a multivariate normal distribution with mean $\mathbf{0}_k$ and covariance $(0.01) \times \boldsymbol{I}_{10}$. We do not include an intercept in any simulation scenarios.

Creating Simulation Structures

In order to simulate from a $\text{VARX}_{10,10}(4,4)$, we start by constructing a $\text{VAR}_{20}(4,4)$. Denoting the first 10 series as $\mathbf{y}_t$ and the second 10 as $\mathbf{x}_t$, we simulate according to the relationship

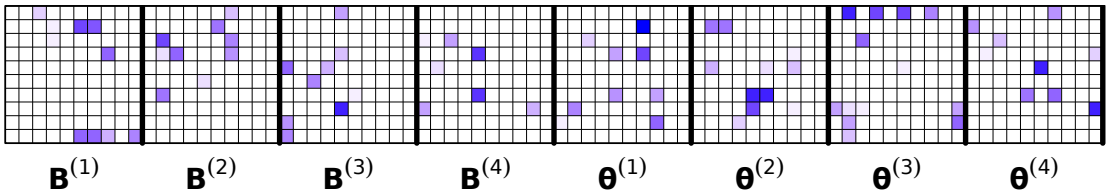
$$\begin{pmatrix} \mathbf{y}_t \\ \mathbf{x}_t \end{pmatrix} = \begin{pmatrix} \boldsymbol{B}^{(1)} & \boldsymbol{\theta}^{(1)} \\ \mathbf{0} & \Gamma \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-1} \\ \mathbf{x}_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{B}^{(2)} & \boldsymbol{\theta}^{(2)} \\ \mathbf{0} & \Gamma \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-2} \\ \mathbf{x}_{t-2} \end{pmatrix} + \begin{pmatrix} \boldsymbol{B}^{(3)} & \boldsymbol{\theta}^{(3)} \\ \mathbf{0} & \Gamma \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-3} \\ \mathbf{x}_{t-3} \end{pmatrix} + \begin{pmatrix} \boldsymbol{B}^{(4)} & \boldsymbol{\theta}^{(4)} \\ \mathbf{0} & \Gamma \end{pmatrix} \begin{pmatrix} \mathbf{y}_{t-4} \\ \mathbf{x}_{t-4} \end{pmatrix} + \mathbf{u}_t,$$

in which Γ is an unspecified matrix denoting the dependence structure of the exogenous series $\mathbf{x}_t$, and $\mathbf{u}_t \overset{\text{iid}}{\sim} N(\mathbf{0}, .01 \times \boldsymbol{I}_{20})$. After ensuring that the resulting VAR is stationary, in the following scenarios, we solely forecast $\mathbf{y}_t$.

Scenario 1: Unstructured Sparsity

We first consider a scenario in which the sparsity is completely random. Under such a design, we should expect superior performance from the Lasso VARX, which assumes no structure. This sparsity pattern is depicted in Figure 7 and the results are summarized in Table 4.

Figure 7: Sparsity Pattern Scenario 1: Unstructured Sparsity. Darker shading represents coefficients that are larger in magnitude.



In this scenario, as expected, we find that the Lasso VARX achieves the best performance. Nonetheless, all structured approaches outperform the AIC and BIC benchmarks.

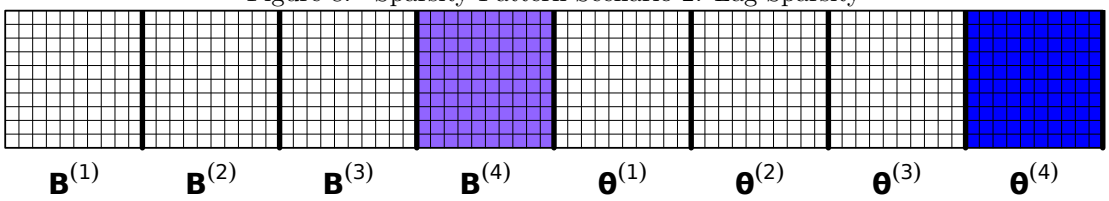
Table 4: Out-of-sample MSFE of one-step ahead forecasts after 100 simulations: Scenario 1

Model/VARX Penalty Structure	MSFE	Standard Error
Lasso	0.1482	0.0177
Lag Group Lasso	0.2371	0.0136
Own/Other Group Lasso	0.2109	0.0108
Sparse Lag Group Lasso	0.2075	0.0132
Sparse Own/Other Group Lasso	0.2170	0.0111
Endogenous-First VARX	0.3477	0.0235
VAR with lag selected by AIC	0.4848	0.0741
VAR with lag selected by BIC	0.5874	0.0109
Sample Mean	2.8743	0.1637
Random Walk	2.1139	0.0812

Scenario 2: Lag Sparsity

We next consider a scenario in which lags $\boldsymbol{B}^{(4)}$ and $\boldsymbol{\theta}^{(4)}$ are dense with coefficients of the same magnitude, and all other coefficients are set to zero. Under such a design, we should expect superior performance from the structured approaches that partition all coefficients within a lag to the same group, such as the Lag Group Lasso VARX and the Endogenous-First VARX. This sparsity pattern is depicted in Figure 8, and the results are summarized in Table

Figure 8: Sparsity Pattern Scenario 2: Lag Sparsity



5.

Table 5: Out-of-sample MSFE of one-step ahead forecasts after 100 simulations: Scenario 2

Model/VARX Penalty Structure	MSFE	Standard Error
Lasso	0.1239	0.0059
Lag Group Lasso	0.1169	0.0052
Own/Other Group Lasso	0.1180	0.0053
Sparse Lag Group Lasso	0.1142	0.0054
Sparse Own/Other Group Lasso	0.1182	0.0054
Endogenous-First VARX	0.1164	0.0052
VAR with lag selected by AIC	0.2272	0.0031
VAR with lag selected by BIC	0.2367	0.0107
Sample Mean	0.2228	0.0097
Random Walk	0.5129	0.0257

Under Scenario 2, we find that the Sparse Lag Group Lasso, which conforms to this grouping but also allows for within-group sparsity, achieves the best forecasts, though there is not a substantial difference across all structured approaches. Note that since AIC and BIC select from models of sequentially increasing lag order, they cannot

accommodate this sparsity pattern and perform even worse than the sample mean

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. Only matrices $\boldsymbol{B}^{(1)}$, $\boldsymbol{B}^{(4)}, \boldsymbol{\theta}^{(1)}$, and $\boldsymbol{\theta}^{(4)}$ contain nonzero coefficients. Additionally, similarly to Scenario 1, sparsity within each lag is unstructured. In such a scenario, we should expect procedures that allow for within-group sparsity, such as the Sparse Lag Group Lasso and Lasso VARX to achieve the best forecast performance. This sparsity pattern is depicted in Figure 9 and the results are summarized in Table 6.

Figure 9: Sparsity Pattern Scenario 3: Structured Lagwise, Unstructured within Lag

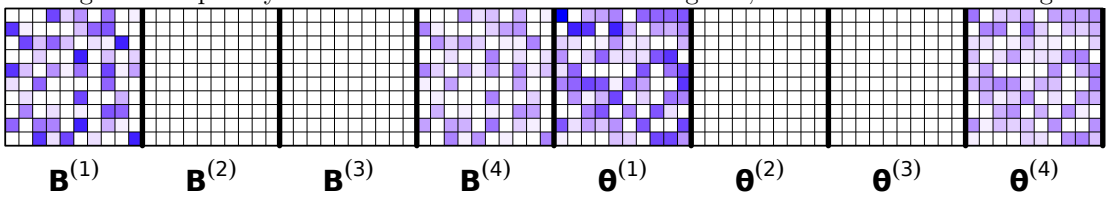


Table 6: Out-of-sample MSFE of one-step ahead forecasts after 100 simulations: Scenario 3

Model/VARX Penalty Structure	MSFE	Standard Error
Lasso	0.1634	0.0018
Lag Group Lasso	0.1630	0.0018
Own/Other Group Lasso	0.1625	0.0018
Sparse Lag Group Lasso	0.1676	0.0018
Sparse Own/Other Group Lasso	0.1641	0.0017
Endogenous-First VARX	0.1724	0.0018
VAR with lag selected by AIC	0.3514	0.0042
VAR with lag selected by BIC	0.2814	0.0037
Sample Mean	0.5984	0.0135
Random Walk	1.0439	0.0259

Under this scenario, the Lag Group Lasso and Own/Other Group Lasso VARX achieve the best performance, though the difference in forecasting performance between most of our regularized methods is negligible.

Scenario 4: Sparse and Diagonally Dominant

Our final scenario consists consists of a diagonally-dominant sparsity structure, in which all diagonal elements in $\boldsymbol{B}^{(4)}$ are equal in magnitude and substantially larger than the off-diagonal components. In a manner similar to Scenario 2, the coefficients in $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(4)}$ are identical in magnitude. $\boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)}$, $\boldsymbol{B}^{(3)}, \boldsymbol{\theta}^{(2)}$, and $\boldsymbol{\theta}^{(3)}$ are set identically to zero. Under this setting, one would expect top performance from the Own/Other Group Lasso VARX. The sparsity pattern is depicted in Figure 10 and the simulation results are summarized in Table 7.

Under Scenario 4, as expected, the Own/Other and Sparse Own/Other Group Lasso VARX achieve superior fore-

Figure 10: Sparsity Pattern Scenario 4: Sparse and Diagonally Dominant

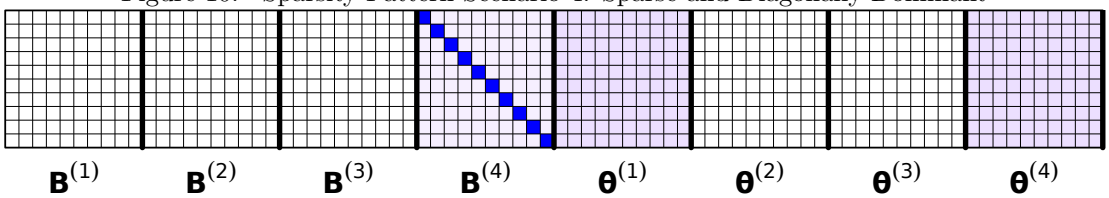


Table 7: Out-of-sample MSFE of one-step ahead forecasts after 100 simulations: Scenario 4

Model/VARX Penalty Structure	MSFE	Standard Error
Lasso	0.1425	0.00649
Lag Group Lasso	0.1551	0.00718
Own/Other Group Lasso	0.1182	0.00517
Sparse Lag Group Lasso	0.1331	0.00011
Sparse Own/Other Group Lasso	0.1190	0.00520
Endogenous-First VARX	0.1659	0.00746
VAR with lag selected by AIC	0.2433	0.00252
VAR with lag selected by BIC	0.2445	0.00342
Sample Mean	0.2295	0.00372
Random Walk	0.4610	0.00981

casts. The other structured approaches that cannot accommodate this sparsity pattern still substantially outperform the information criterion benchmarks.

Discussion

All of our procedures are fairly robust to sparsity patterns not conforming to their true group structures. In each scenario, every method substantially outperforms all benchmarks. Scenario 1 is the only case in which the structured approaches perform poorly relative to the Lasso VARX. We expect such an unstructured sparsity pattern to occur only rarely in data applications.

4 Extending the VAR to incorporate mild nonstationarity

In this section, we outline a possible extension that allows for shrinking toward reference models, such as a vector random walk, that can account for mild non-stationarity, which is ubiquitous in macroeconomic data.

The "Minnesota Lasso"

Our proposed algorithms can easily be modified to shrink toward a known constant matrix. Shrinking toward a constant matrices $\mathbf{C}_y \in \mathbf{R}^{k\times kp}, \mathbf{C}_x \in \mathbf{R}^{k\times ms}$ results in a penalty of the form

$$\min_{\boldsymbol{B},\boldsymbol{\theta},\boldsymbol{\nu}} \frac{1}{2}\sum_{t=1}^{T} \|\mathbf{y}_t - \boldsymbol{\nu} - \boldsymbol{B}\mathbf{Y}_{t-1} - \boldsymbol{\theta}\mathbf{X}_{t-1}\|_F^2 + \lambda\Big(\mathcal{P}_y(\boldsymbol{B}-\mathbf{C}_y) + \mathcal{P}_x(\boldsymbol{\theta}-\mathbf{C}_x)\Big).$$

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

Let $[\boldsymbol{B}, \boldsymbol{\theta}]^{\lambda}(\sum_{t=1}^{T} \mathbf{y}_t, \mathbf{C}_y, \mathbf{C}_x)$ denote a solution to this problem. Now, by a change of variables $\widetilde{\boldsymbol{B}} = \boldsymbol{B} - \mathbf{C}_y, \widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \mathbf{C}_x$, we obtain the equivalent problem

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

or

$$\min_{\boldsymbol{B},\boldsymbol{\theta},\boldsymbol{\nu}} \frac{1}{2} \sum_{t=1}^{T} \|\mathbf{y}_t - \boldsymbol{\nu} - \mathbf{C}_y\mathbf{Y}_{t-1} - \boldsymbol{B}\mathbf{Y}_{t-1} - \mathbf{C}_x\mathbf{X}_{t-1} - \boldsymbol{\theta}\mathbf{X}_{t-1}\|_F^2 + \lambda\Big(\mathcal{P}_y(\widetilde{\boldsymbol{B}}) + \mathcal{P}_x(\widetilde{\boldsymbol{\theta}})\Big).$$

Thus, the solution to this transformed problem is given by $\widehat{\boldsymbol{B}}^{\lambda}(\mathbf{y}_t - \mathbf{C}_y\mathbf{Y}_{t-1}, 0)$ and $\widehat{\boldsymbol{\theta}}^{\lambda}(\mathbf{y}_t - \mathbf{C}_x\mathbf{X}_{t-1}, 0)$. Transforming back to the original variable (i.e., from $[\widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{\theta}}]$ to $[\boldsymbol{B}, \boldsymbol{\theta}]$), we see that

$$\widehat{\boldsymbol{B}}^{\lambda}(\mathbf{y}_t, \mathbf{C}_y) = \mathbf{C}_y + \widehat{\boldsymbol{B}}^{\lambda}(\mathbf{y}_t - \mathbf{C}_y\mathbf{Y}_{t-1}, 0), \tag{4.1}$$

$$\widehat{\boldsymbol{\theta}}^{\lambda}(\mathbf{y}_t, \mathbf{C}_x) = \mathbf{C}_x + \widehat{\boldsymbol{\theta}}^{\lambda}(\mathbf{y}_t - \mathbf{C}_x\mathbf{X}_{t-1}, 0). \tag{4.2}$$

As an example, with $\mathbf{C}_y = (\mathbf{I}_k, \mathbf{0}_{k\times k}, \ldots, \mathbf{0}_{k\times k}), \mathbf{C}_x = \mathbf{0}_{k\times ms}$, we could implement a variant of the Minnesota Prior, in which we shrink toward a random walk. This approach could be of use in economic applications as it is widely believed that many macroeconomic time series can be well approximated by a random walk (Litterman [1979]). In order to test this procedure, we follow the methodology of Banbura et al. [2009], who also utilize the data from Stock and Watson [2005], but eschew stationarity transformations and work directly with the non-stationary series. We again apply our estimation procedures on the aforementioned *medium* set of series, but choose not to perform any stationarity transformations and instead shrink toward a vector random walk. Our results are summarized in Table 8. We find that in this non-stationary setting, exogenous series are less helpful in forecasting; hence in this section, we fit VAR models, setting $m = s = 0$.

Table 8: Out of sample MSFE of one-step ahead forecasts on 20 nonstationary macroeconomic indicators which shrink toward a vector random walk.

	$p = 4$		$p = 13$	
Model/Penalty	MSFE	S.E.	MSFE	S.E.
Lasso	2.322	1.515	2.038	1.207
Lag Group Lasso	2.278	1.537	2.428	1.657
Own/Other Group Lasso	2.074	1.302	2.445	1.649
Sparse Lag Group Lasso	2.156	1.417	2.421	1.643
Sparse Own/Other Group Lasso	1.914	1.171	2.421	1.638
VAR with lag selected by AIC	11.001	5.838	11.136	6.071
VAR with lag selected by BIC	11.001	5.838	11.136	6.071
Sample Mean	31.30	3.709	29.12	1.657
Random Walk	2.513	1.766	2.626	1.850

For both maximal lag orders, all of our approaches outperform the random walk benchmark, though performance starts to degrade as the lag order increases.

5 Conclusion

The structured regularization framework is quite flexible in that it can accommodate a variety of potential dynamic structures. Each of the proposed methods consistently outperforms benchmark procedures. Forecast performance for all methods appears to be robust across multiple sparsity structures. Moreover, upon examining actual macroeconomic data, structured approaches tend to outperform their unstructured counterparts.

Our work has considerable room for extensions. Primarily, our procedures require a coherent maximal lag selection mechanism. The currently accepted 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 incorporate an additional penalty parameter that grows as the lag order increases, as in Song and Bickel [2011], but this approach would require a multi-dimensional gridsearch as well as specifying a functional form for the lag penalty. Nicholson et al. [2014] addresses this concern in the VAR context by extending the hierarchical group lasso penalty to explicitly incorporate lag order selection.

An `R` package containing our algorithms and validation procedures, `BigVAR`, is available on Github.

References

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, pages 19–53, 2011.

Marta Banbura, Domenico Giannone, and Lucrezia Reichlin. Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92, 2009.

Sumanta Basu and George Michailidis. Estimation in high-dimensional vector autoregressive models. *arXiv preprint arXiv:1311.4175*, 2013.

Sumanta Basu, Ali Shojaie, and George Michailidis. Network granger causality with inherent grouping structure. *arXiv preprint arXiv:1210.3711*, 2012.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Stephen R Becker, Emmanuel J Candès, and Michael C Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

Ben S Bernanke, Jean Boivin, and Piotr Eliasz. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (favar) approach. *The Quarterly Journal of Economics*, 120(1):387–422, 2005.

Chris Brooks and Sotiris Tsolacos. Forecasting models of retail rents. *Environment and Planning A*, 32(10):1825–1840, 2000.

Alessandro Chiuso and Gianluigi Pillonetto. Nonparametric sparse estimators for identification of large scale linear systems. In *Decision and Control (CDC), 2010 49th IEEE Conference on*, pages 2942–2947. IEEE, 2010.

David O Cushman and Tao Zha. Identifying monetary policy in a small open economy under flexible exchange rates. *Journal of Monetary economics*, 39(3):433–448, 1997.

Richard A. Davis, Pengfei Zang, and Tian Zheng. Sparse vector autoregressive modeling. 2012. journal: arXiv preprint arXiv:1207.0520.

Francis X Diebold. The past, present, and future of macroeconomic forecasting. *The Journal of Economic Perspectives*, 12(2):175–192, 1998.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Milton Friedman. 'comment:' a test of an econometric model for the united states, 1921-1947 by carl christ. In *Conference on business cycles*, pages 35–130. NBER, 1951.

Edward I George, Dongchu Sun, and Shawn Ni. Bayesian stochastic search for var model restrictions. *Journal of Econometrics*, 142(1):553–580, 2008.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

N. J. Hsu, H. L. Hung, and Y. M. Chang. Subset selection for vector autoregressive processes using lasso. 52(7): 3645–3657, 2008. journal: Computational Statistics & Data Analysis.

Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*, 12:2297–2334, 2011.

K Kadiyala and Sune Karlsson. Numerical methods for estimation and inference in bayesian var-models. *Journal of Applied Econometrics*, 12(2):99–132, 1997.

Lawrence Robert Klein and Arthur S Goldberger. An econometric model of the united states, 1929-1952, 1955.

Anders Bredahl Kock and Laurent AF Callot. Oracle inequalities for high dimensional vector autoregressions. *arXiv preprint arXiv:1311.0811*, 2013.

Gary Koop. Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics*, 2011.

Robert B. Litterman. Techniques of forecasting using vector autoregressions. Working papers, Federal Reserve Bank of Minneapolis, 1979.

Helmut Lütkepohl. New introduction to multiple time series analysis. 2005.

Arnold Neumaier and Tapio Schneider. Estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1):27–57, 2001.

William B Nicholson, Jacob Bien, and David S Matteson. Hierarchical vector autoregression. *arXiv preprint arXiv:1412.5250*, 2014.

Vincent R Nijs, Shuba Srinivasan, and Koen Pauwels. Retail-price drivers and retailer profits. *Marketing Science*, 26(4):473–487, 2007.

Jorge. Nocedal and Stephen J Wright. *Numerical optimization*, volume 2. Springer New York, 1999.

Sergio Ocampo and Norberto Rodríguez. An introductory review of a structural var-x estimation and applications. *Revista Colombiana de Estadística*, 35(3):479–508, 2012.

Jack HW Penm, Jammie H Penm, and RD Terrell. The recursive fitting of subset varx models. *Journal of Time Series Analysis*, 14(6):603–619, 1993.

Zhiwei Qin, Katya Scheinberg, and Donald Goldfarb. Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation*, pages 1–27, 2010.

Daniel Racette and Jacques Raynauld. Canadian monetary policy: will the checklist approach ever get us to price stability? *Canadian Journal of Economics*, pages 819–838, 1992.

John C Robertson and Ellis William Tallman. Improving forecasts of the federal funds rate in a policy model. Technical report, Federal Reserve Bank of Atlanta, 1999.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

Christopher A Sims. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48, 1980.

Song Song and Peter Bickel. Large vector auto regressions. 2011. journal: arXiv preprint arXiv:1106.3915.

James H Stock and Mark W Watson. An empirical comparison of methods for forecasting using many predictors. *Manuscript, Princeton University*, 2005.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494, 2001.

Władysław Welfe. Macroeconometric models of the united states and canada. In *Macroeconometric Models*, pages 15–46. Springer, 2013.

B. Dan Wood. Presidential saber rattling and the economy. *American Journal of Political Science*, 53(3):695–709, 2009. ISSN 1540-5907. doi: 10.1111/j.1540-5907.2009.00395.x. URL `http://dx.doi.org/10.1111/j.1540-5907.2009.00395.x`.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

A 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

$$
\begin{array}{llll}
\boldsymbol{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T] & (k \times T); & & \\
\boldsymbol{Z}_t = [\mathbf{y}_t^\top, \ldots, \mathbf{y}_{t-p}^\top, \mathbf{x}_t^\top, \ldots, \mathbf{x}_{t-s}^\top]^\top & [(kp + ms) \times 1]; & \boldsymbol{Z} = [\boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{T-1}] & [(kp + ms) \times T]; \\
\boldsymbol{B} = [\boldsymbol{B}^{(1)}, \boldsymbol{B}^{(2)}, \ldots, \boldsymbol{B}^{(p)}] & (k \times kp); & \boldsymbol{\theta} = [\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(s)}] & [k \times ms]; \\
\boldsymbol{\Phi} = [\boldsymbol{B}, \boldsymbol{\theta}] & [k \times (kp + ms)]; & & \\
\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_T] & (k \times T); & \mathbf{1} = [1, \ldots, 1]^\top & (T \times 1).
\end{array}
$$

Equation (2.1) then becomes

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

and the least squares procedure (2.2) can be expressed as minimizing $\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{\nu}\mathbf{1}^\top - \boldsymbol{\Phi}\boldsymbol{Z}\|_F^2$ over $\boldsymbol{\nu}$ and $\boldsymbol{\Phi}$, where $\|\mathbf{A}\|_F$ denotes the Frobenius norm of the matrix $\mathbf{A}$.

A.2 Intercept Term

Using compact matrix notation, we can express the unpenalized portion of (2.3) as

$$
\begin{aligned}
f(\boldsymbol{\Phi}, \boldsymbol{\nu}) &= \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{\nu}\mathbf{1}^\top - \boldsymbol{\Phi}\boldsymbol{Z}||_F^2, && \text{(A.1)} \\
&= \frac{1}{2}\sum_{kt}(Y_{kt} - \nu_k - \boldsymbol{\Phi}Z_{\cdot t})^2. && \text{(A.2)}
\end{aligned}
$$

In regularization problems, the intercept $\hat{\boldsymbol{\nu}}$ is not typically shrunk and can be derived separately from $\widehat{\boldsymbol{\Phi}} = \operatorname{argmin} f$. We can find $\hat{\boldsymbol{\nu}}$ by calculating the gradient of (A.1) with respect to $\boldsymbol{\nu}$

$$
\begin{aligned}
0 = \nabla_{\boldsymbol{\nu}} f(\boldsymbol{\Phi}, \boldsymbol{\nu}) &= (\boldsymbol{Y} - \hat{\boldsymbol{\nu}}\mathbf{1}^\top - \widehat{\boldsymbol{\Phi}}\boldsymbol{Z})\mathbf{1}, \\
\implies \hat{\boldsymbol{\nu}}_j(\lambda) &= \overline{Y}_{k\cdot} - \widehat{\boldsymbol{\Phi}}\overline{Z}_{k\cdot},
\end{aligned}
$$

in which $\overline{Y}_{k\cdot} = \frac{1}{T}\sum_t Y_{kt}$, and $\overline{Z}_{k\cdot} = \frac{1}{T}\sum_t Z_{kt}$. This provides some insight into the scaling, as we can rewrite (A.1) as

$$\min_{\boldsymbol{\Phi}} \frac{1}{2}||\boldsymbol{Y} - (\overline{Y} - \boldsymbol{\Phi}\overline{Z})\mathbf{1}^\top - \boldsymbol{\Phi}\boldsymbol{Z}||_F^2, \tag{A.3}$$

$$= \min_{\boldsymbol{\Phi}} \frac{1}{2}||(\boldsymbol{Y} - \overline{Y}\mathbf{1}^\top) - \boldsymbol{\Phi}(\boldsymbol{Z} - \overline{Z}\mathbf{1}^\top)||_F^2, \tag{A.4}$$

in which $\overline{\mathbf{Y}}$ is a $k \times 1$ vector of row means and $\overline{\mathbf{Z}}$ is a $(kp+ms) \times 1$ vector of row means. Note that since the Laplacian with respect to $\boldsymbol{\nu}$ is -T, $\hat{\boldsymbol{\nu}}$ is a maximum.

A.3 Solution Strategies

In the following sections, assume that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are centered as in Equation (A.4).

A.3.1 Lasso VARX

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

$$\min_{\boldsymbol{\Phi}_{jr}} \frac{1}{2}(\boldsymbol{Y}_{jt} - \sum_{\ell \neq r} \boldsymbol{\Phi}_{j\ell}\boldsymbol{Z}_{\ell t} - \boldsymbol{\Phi}_{jr}\boldsymbol{Z}_{jt})^2 + \lambda|\boldsymbol{\Phi}_{jr}|. \tag{A.5}$$

Let $\boldsymbol{R}_t = \boldsymbol{Y}_{jt} - \sum_{\ell \neq r} \boldsymbol{\Phi}_{j\ell}\boldsymbol{Z}_{\ell t}$ denote the partial residual. Then, we can rewrite Equation (A.5) as

$$g_{jr}(\boldsymbol{\Phi}) = \min_{\boldsymbol{\Phi}_{jr}} \frac{1}{2}(\boldsymbol{R}_t - \boldsymbol{\Phi}_{jr}\boldsymbol{Z}_{jt})^2 + \lambda|\boldsymbol{\Phi}_{jr}|$$

$$= \min_{\boldsymbol{\Phi}_{jr}} \frac{1}{2}(\sum_t \mathbf{R}_t^2 - \boldsymbol{\Phi}_{jr}^2\boldsymbol{Z}_{jt}^2 - 2\mathbf{R}_t\boldsymbol{Z}_{jt}\boldsymbol{\Phi}_{jr}) + \lambda|\boldsymbol{\Phi}_{jr}|.$$

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

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

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

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

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

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

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

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

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

A.3.2 Lag Group Lasso VARX

Rather than vectorizing the Lag Group Lasso VARX and solving the corresponding univariate least squares problem, if the groups are proper submatrices we can exploit the matrix structure for considerable computational gains. Without loss of generality, we will consider the "one lag" problem for $\boldsymbol{B}^{(q)}$ (the problem for $\boldsymbol{\theta}^{(q)}$ is analogous).

$$\frac{1}{2}\|\boldsymbol{R}_{-q} - \boldsymbol{B}^{(q)}\boldsymbol{Z}_q\|_F^2 + \lambda\|\boldsymbol{B}^{(q)}\|_F, \tag{A.6}$$

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

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

The subgradient with respect to $\boldsymbol{B}^{(q)}$ then is

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

where ω is defined as

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

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

$$\begin{aligned}
&\frac{\hat{\boldsymbol{B}}^{(q)}\boldsymbol{Z}_q\boldsymbol{Z}_q^\top - \boldsymbol{R}_{-q}\boldsymbol{Z}_q^\top}{\lambda} \in \{U : \|U\|_F \leq 1\}, \\
&\iff \|\hat{\boldsymbol{B}}^{(q)}\boldsymbol{Z}_q\boldsymbol{Z}_q^\top - \boldsymbol{R}_{-q}\boldsymbol{Z}_q^\top\|_F \leq \lambda, \\
&\iff \|\boldsymbol{R}_{-q}\boldsymbol{Z}_q^\top\|_F \leq \lambda, \\
&\iff \hat{\boldsymbol{B}}^{(q)} = \mathbf{0}.
\end{aligned}$$

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

$$\begin{aligned}
&\boldsymbol{B}^{(q)}\boldsymbol{Z}_q\boldsymbol{Z}_q^{\top} - \boldsymbol{R}_{-q}\boldsymbol{Z}_q^{\top} + \lambda(\frac{\boldsymbol{B}^{(q)}}{\|\boldsymbol{B}^{(q)}\|_F}) = 0,\\
&\boldsymbol{B}^{(q)}\boldsymbol{Z}_q\boldsymbol{Z}_q^{\top} + \lambda(\frac{\boldsymbol{B}^{(q)}}{\|\boldsymbol{B}^{(q)}\|_F}) = \boldsymbol{R}_{-q}\boldsymbol{Z}_q^{\top},\\
&\boldsymbol{B}^{(q)}\left(\boldsymbol{Z}_q\boldsymbol{Z}_q^{\top} + \frac{\lambda}{\|\boldsymbol{B}^{(q)}\|_F}\boldsymbol{I}_k\right) = \boldsymbol{R}_{-q}\boldsymbol{Z}_q^{\top}. \qquad \text{(A.7)}
\end{aligned}$$

Now, since $\boldsymbol{Z}_q\boldsymbol{Z}_q^{\top} \succ 0$ and $\lambda > 0, \boldsymbol{Z}_q\boldsymbol{Z}_q^{\top} + \frac{\lambda}{\|\boldsymbol{B}^{(q)}\|_F}\boldsymbol{I}_k \succ 0$, it is possible to create a trust region subproblem which coincides with Equation (A.6). However, we need to transform $\mathbf{R}_{-q}\boldsymbol{Z}_q$ into a scalar. Define

$$\begin{aligned}
&\boldsymbol{r}_q = \text{vec}(\mathbf{R}_{-q}\boldsymbol{Z}_q),\\
&\mathbf{X}_q = \boldsymbol{Z}_q\boldsymbol{Z}_q^{\top} \otimes \mathbf{I}_k,\\
&\boldsymbol{b}_q = \text{vec}(\boldsymbol{B}^{(q)}).
\end{aligned}$$

Hence, we can rewrite Equation (A.7) as

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

Applying the same transformation to the original subproblem, there exists a $\Delta > 0$ in which the optimal solution to the trust-region subproblem corresponds to the optimal solution of Equation A.6

$$\begin{aligned}
\min \frac{1}{2}\boldsymbol{b}_q^{\top}\boldsymbol{X}_q\boldsymbol{X}_q^{\top}\boldsymbol{b}_q + \boldsymbol{r}_q^{\top}\boldsymbol{b}_q,\\
\text{s.t.}\|\boldsymbol{b}_q\|_F \leq \Delta,
\end{aligned}$$

in which Δ corresponds to the trust-region radius. By the Karush-Kuhn-Tucker (KKT) conditions, we must have that: $\lambda(\Delta - \|\boldsymbol{b}_q^*\|_F) = 0$, which implies that $\|\boldsymbol{b}_q^*\|_F = \Delta$. Then, applying Theorem 4.1 of Nocedal and Wright [1999], we can conclude that

$$\boldsymbol{b}_q^* = -\left(\boldsymbol{X}_q + \frac{\lambda}{\Delta}\boldsymbol{I}_{k^2}\right)^{-1}\boldsymbol{r}_q. \qquad \text{(A.8)}$$

These transformations allow for the use of the methodology described in Qin et al. [2010]. Equation (A.8) can also be expressed as $\boldsymbol{b}_q^* = \Delta y_q(\Delta)$, where

$$y_q(\Delta) = -\left(\Delta\boldsymbol{X}_q + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{r}_q,$$

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 efficiently compute $\|y_q(\Delta)\|_F^2$ via an eigen-decomposition of $\boldsymbol{X}_q$

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

in which $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$ represent the respective eigenvectors and eigenvalues of $\boldsymbol{X}_q$. Finally, we can determine the optimal Δ by applying Newton's method to find the root of

$$\phi(\Delta) = 1 - \frac{1}{\|y_j(\Delta)\|_F}. \tag{A.9}$$

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

A.3.3 Own/Other Group Lasso VARX

Since, in this scenario, the groups are not proper submatrices, Equation (2.12) must be transformed into a least squares problem. In order to do so, we define the following

$$\begin{aligned}
r_{-qq} &= \text{vec}(\mathbf{R}_{-qq}),\\
b_{qq} &= \text{vec}(\text{diag}(\boldsymbol{B}^{(q)})),\\
\boldsymbol{M}_{qq} &= (\boldsymbol{Z}^\top \otimes I_k)_{qq}.
\end{aligned}$$

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

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

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

$$\frac{\partial}{\partial b_{qq}} = \boldsymbol{M}_{qq}^\top \boldsymbol{M}_{qq} b_{qq} + \boldsymbol{M}_{qq}^\top r + \lambda\omega(b),$$

where ω is defined as

$$\omega(s) \in \begin{cases} \{\frac{s}{\|s\|_F}\} & s \neq 0 \\ \{u : \|u\|_F \leq 1\} & s = 0. \end{cases}$$

Thus, we can apply a slightly adapted version of Algorithm 3.

A.3.4 Sparse Lag Group Lasso VARX

As with the Lag Group Lasso, we will consider the one-block subproblem for lag $\boldsymbol{B}^{(q)}$

$$\min_{\boldsymbol{B}^{(q)}} \frac{1}{2k}\|\mathbf{R}_{-q} - \boldsymbol{B}^{(q)}\boldsymbol{Z}_q\|_F^2 + (1-\alpha)\lambda\|\boldsymbol{B}^{(q)}\|_F + \alpha\lambda\|\boldsymbol{B}^{(q)}\|_1. \tag{A.10}$$

Since the inclusion of within-group sparsity does not allow for separability, coordinate descent based procedures are no longer appropriate, therefore, following Simon et al. [2013] our solution to the Sparse Group Lasso VARX utilizes gradient descent methods. We express Equation (A.10) 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 only makes use of first-order information around its current estimate $\boldsymbol{B}_0$ (borrowing from Simon et al. [2013], for notational ease, let $\boldsymbol{B} \equiv \boldsymbol{B}^{(q)}$, $\ell(\boldsymbol{B})$ represent the unpenalized loss function, and $\mathcal{P}(\boldsymbol{B})$ represent the penalty term). Then, we can express the linearization as

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

in which h represents the step size. Removing terms independent of $\boldsymbol{B}$, our objective function becomes

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

Then, generalizing the arguments outlined by Simon et al. [2013], we can infer that the optimal $\boldsymbol{B}$ can be expressed as

$$U(\boldsymbol{B}) = \left(1 - \frac{h(1-\alpha)\lambda}{\|ST(\boldsymbol{B}_0 - h(\boldsymbol{B}_0\boldsymbol{Z}_q - \boldsymbol{R}_{-q})\boldsymbol{Z}_q^\top, h\alpha\lambda)\|_F}\right)_+ ST(\boldsymbol{B}_0 - h(\boldsymbol{B}_0\boldsymbol{Z}_q - \boldsymbol{R}_{-q})\boldsymbol{Z}_{-q}^\top, h\alpha\lambda).$$

As in Simon et al. [2013], we apply a Nesterov accelerated update. At step m, we update according to

$$\hat{\boldsymbol{B}}[j] \leftarrow \hat{\boldsymbol{B}}[j-1] + \frac{j}{j+3}(U(\boldsymbol{B}) - \hat{\boldsymbol{B}}[j-1]), \tag{A.11}$$

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

The calculation of the step size h can be problematic. Ideally, the step size should be as large as possible, as it leads to faster convergence, but if the step size is too large, the algorithm may diverge. The conventional method for determining step size, described in Simon et al. [2013] and Beck and Teboulle [2009], is to decrease h until

$$\ell(\widehat{\boldsymbol{B}}, h) \leq \ell(\boldsymbol{B}) + \text{vec}(\nabla_q)^\top \text{vec}(\Delta_{l,h}) + \frac{1}{2h}||\Delta_{l,h}||_F^2. \tag{A.12}$$

However, as noted in section 5.3 of Becker et al. [2011], Equation (A.12) has severe cancellation errors when $\ell(\widehat{\boldsymbol{B}}, h) \approx \ell(\boldsymbol{B}, h)$. They posit a more conservative approach, iterating until

$$\ell(\widehat{\boldsymbol{B}}, h) \leq \frac{1}{2hk}||\Delta_{l,h}||_F^2. \tag{A.13}$$

They recommend a hybrid approach: choosing Equation (A.12) when $\ell(\boldsymbol{B}, t) - \ell(\widehat{\boldsymbol{B}}, t) \geq \gamma \ell(\widehat{\boldsymbol{B}}, t)$, for some small $\gamma > 0$ and choosing Equation (A.13) otherwise.

Unfortunately, we have found even this hybrid approach to be unstable. This could be due to the use of a Nesterov-style accelerated update which, per Bach et al. [2011], can result in the algorithm not decreasing at each step, causing the above specifications to diverge. We instead analytically derive 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 $\boldsymbol{B}^{(q)}$. We have that

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

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

As only the maximum eigenvalue is required, the power method is much more computationally efficient than a

computation of the entire eigensystem. 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 time point in the cross-validation and evaluation stages.

The inner loop of of the Sparse Group-Lasso VAR procedure is detailed in Algorithm (4). An outline of the algorithm is below:

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

In a manner similar to Algorithm 3, an "active-set" approach is used to minimize computation time.

Upon performing transformations as in the Own/Other Group Lasso VARX scenario, the Sparse Own/Other Group Lasso VARX follows almost the exact same procedure as its Lag counterpart.

A.3.5 Endogenous-First VARX

The Endogenous-First Group Lasso VARX is of the form

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

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

$$\min_{\boldsymbol{\Phi}_i} \frac{1}{2}\|\boldsymbol{Y}_i - \boldsymbol{\Phi}_i\boldsymbol{Z}|_F^2 + \lambda \sum_{\ell=1}^{p}\sum_{i=1}^{k}\left(\|[\boldsymbol{B}_j^{(\ell)}, \boldsymbol{\theta}_{i,\cdot}^{(\ell)}]\|_F + \|\boldsymbol{\theta}_{i,\cdot}^{(\ell)}\|_F\right). \tag{A.14}$$

In a manner similar to the Sparse Group Lasso VARX, the Endogenous-First VARX is solved via proximal gradient descent. For ease of notation, let $\mathcal{P}(\boldsymbol{B}, \boldsymbol{\theta})$ represent the nested penalty. The update step for the Endogenous-First VARX (at step j) can be expressed as

$$\boldsymbol{\Phi}_i[j] = \text{Prox}_{h\lambda,\mathcal{P}(\boldsymbol{B},\boldsymbol{\theta})}(\boldsymbol{\Phi}_i[j-1] - h\nabla\ell(\boldsymbol{\Phi}_i)), \tag{A.15}$$

in which h denotes step size and $\ell(\boldsymbol{\Phi})$ denotes the unpenalized loss function. Note that $\nabla\ell(\boldsymbol{\Phi}_i) = -(\boldsymbol{Y}_i - \boldsymbol{\Phi}_i\boldsymbol{Z})\boldsymbol{Z}^\top$. Similar to the Sparse Group Lasso setting, a fixed step size is used; h is set to the inverse of the square root of the largest singular value of $\boldsymbol{Z}$. To speed convergence, as in the Sparse Group Lasso update step (A.11), we apply a similar Nesterov-style accelerated update:

$$\hat{\boldsymbol{\Phi}} \leftarrow \hat{\boldsymbol{\Phi}}[j-1] + \frac{j-2}{j+1}(\hat{\boldsymbol{\Phi}}[j-1] - \hat{\boldsymbol{\Phi}}[j-2]),$$

Thus, (A.15) becomes

$$\boldsymbol{\Phi}_i[j] = \text{Prox}_{h_j\lambda,\mathcal{P}(\boldsymbol{B},\boldsymbol{\theta})}(\hat{\boldsymbol{\Phi}} - h\nabla\ell(\boldsymbol{\Phi}_i)), \tag{A.16}$$

Definition A.1 (Jenatton et al. [2011])**.** *The proximal operator associated with the Endogenous-First VARX can be expressed as*

$$Prox_{h_j\lambda,\mathcal{P}(\boldsymbol{B},\boldsymbol{\theta})} \arg\min_{v\in\mathbf{R}^{kp+ms}} \big\{\frac{1}{2}\|u-v\|_F^2 + h_j\lambda\mathcal{P}(v)\big\} \tag{A.17}$$

in other words, the proximal operator will map a vector $u \in \mathbf{R}^{kp+ms}$ to the unique solution of (A.17)*.*

Jenatton et al. [2011] observed that the dual of (A.17) can be solved with one pass of block coordinate descent. Moreover, the block updates are extremely simple and available in closed-form. Algorithm 5 details the prox function within one lag ℓ. Note that the Endogenous-First VARX consists of p separate nested structures in each series. Thus, solving (A.14) essentially amounts to calling Algorithm 5 p times at each update step. For more details about the implementation of nested penalty structures, consult Nicholson et al. [2014].

A.4 Penalty Grid Selection

Table 9: Starting values of the penalty grid for each procedure; ρ_q represents the number of variables in group q.

Structure	Starting Value of Λ_{Grid}
Lasso	$\max(\boldsymbol{Z}\boldsymbol{Y}^\top)$
Lag Group Lasso	$\max_q(\boldsymbol{Z}_q\boldsymbol{Y}^\top)$
Sparse Lag Group Lasso	$\max_q(\boldsymbol{Z}_q\boldsymbol{Y}^\top\alpha)$
Own/Other Group Lasso	$\max_q \frac{(\boldsymbol{Z}\otimes\boldsymbol{I}_k)_q\text{vec}(\boldsymbol{Y}^\top)}{\sqrt{\rho_q}}$
Sparse Own/Other Group Lasso	$\max_q \frac{(\boldsymbol{Z}\otimes\boldsymbol{I}_k)_q\text{vec}(\boldsymbol{Y}^\top)}{\sqrt{\rho_q}}\alpha$
Endogenous-First VARX	$\max_i \boldsymbol{Z}\boldsymbol{Y}_i^\top$,

A.5 Algorithms

Algorithm 1 LASSO-VARX(p,s)

Require: $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Phi}^{\text{INI}}, \lambda$
  $\boldsymbol{\Phi}^{\text{OLD}} \leftarrow \boldsymbol{\Phi}^{\text{INI}}$
  repeat
    for i in k, j in $kp + ms + 1$ **do**
      $\boldsymbol{R}_t \leftarrow \boldsymbol{Y}_{jt} - \sum_{\ell \neq j} \boldsymbol{B}_{j\ell} \boldsymbol{Z}_{\ell t}$
      $\boldsymbol{B}_{ij}^{\text{NEW}} \leftarrow \frac{\text{ST}(\sum_t \boldsymbol{R}_t \boldsymbol{Z}_{jt}, \lambda)}{\sum_t \boldsymbol{Z}_{jt}^2}$
    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 $\boldsymbol{B}^{\text{NEW}}, \hat{\boldsymbol{\nu}}$

Algorithm 2 LASSO-VARX(p,s) Cross-Validation

Require: $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Phi}^{\text{INI}}, \Lambda_{\text{grid}}$
  $\boldsymbol{Y} \leftarrow \bar{\boldsymbol{Y}} \boldsymbol{1}^\top$
  $\boldsymbol{Z} \leftarrow \bar{\boldsymbol{Z}} \boldsymbol{1}^\top$
  $\boldsymbol{\Phi}^{\text{LAST}} \leftarrow \boldsymbol{\Phi}^{\text{INI}}$

  for j in $[T_1, T_2 - 1]$ **do**
    $\boldsymbol{Y}_{\text{TRAIN}}^{(j)} \leftarrow \boldsymbol{Y}_{1:j}$
    $\boldsymbol{Z}_{\text{TRAIN}}^{(j)} \leftarrow \boldsymbol{Z}_{,1:j}$
    for i in Λ_{Grid} **do**
      $\boldsymbol{\nu}_i, \boldsymbol{\Phi}_i^{\text{NEW}} \leftarrow$ Lasso-VARX$(\boldsymbol{Y}_{\text{TRAIN}}^{(j)}, \boldsymbol{Z}_{\text{TRAIN}}^{(j)}, \boldsymbol{\Phi}_i^{\text{LAST}}, \lambda_i, \epsilon)$
      $SSFE^{(j,i)} \leftarrow \|\boldsymbol{Y}_{j+1} - [\nu_i, \boldsymbol{\Phi}_i^{\text{NEW}}] * [\boldsymbol{1}, \boldsymbol{Z}_{TRAIN}^{(j)}]\|_F^2$
      $MSFE^{(j)} \leftarrow \frac{1}{T_2 - T_1} \sum_j SSFE\{i, j\}$
      $\boldsymbol{\Phi}_i^{\text{LAST}} \leftarrow \boldsymbol{\Phi}_i^{\text{NEW}}$

    end for
  end for
  return $\lambda_{\min \text{MSFE}}$

Algorithm 3 Lag Group LASSO-VARX(p,s)

Require: $\mathbf{\Phi}_{\text{INI}}, \mathcal{G}, \boldsymbol{Y}, \boldsymbol{Z}, \mathcal{A}_{\text{INI}}, \Lambda$

Define:

$$\text{for } g = 1, \ldots, p + ms:$$
$$\mathbf{M}_g = \boldsymbol{Z}_g \boldsymbol{Z}_g^\top,$$
$$\mathbf{X}_g = \mathbf{M}_g \otimes \mathbf{I}_k.$$

for $\lambda \in \Lambda$ **do**
  $\mathbf{\Phi}_{\lambda,\mathcal{A}} \leftarrow \mathbf{\Phi}_{\lambda,\text{INI}}$,
  $\mathcal{A}_\lambda \leftarrow \mathcal{A}_{\lambda,\text{INI}}$
  repeat
    $\mathbf{\Phi}_{\lambda,\mathcal{A}} \leftarrow$ThresholdUpdate$(\mathcal{A}, \mathbf{\Phi}_{\lambda,\mathcal{A}}, \lambda)$
    $\mathbf{\Phi}_{\lambda,\mathcal{A}_{\text{FULL}}}, \mathcal{A}_\lambda \leftarrow$BlockUpdate$(\mathcal{A}_{\text{FULL}}, \mathbf{\Phi}_{\lambda,\mathcal{A}}, \lambda)$
  until $\mathbf{\Phi}_{\lambda,\mathcal{A}} = \mathbf{\Phi}_{\lambda,\mathcal{A}_{\text{FULL}}}$
  $\hat{\nu} \leftarrow \bar{\boldsymbol{Y}} - \mathbf{\Phi}_{\lambda,\mathcal{A}} \bar{\boldsymbol{Z}}$
end for
return $\hat{\nu}, \mathbf{\Phi}_\Lambda, A_\Lambda$
procedure BLOCKUPDATE$(\mathcal{G}, \mathbf{\Phi}_{\text{INI}}, \lambda)$
  $\mathbf{\Phi} \leftarrow \mathbf{\Phi}_{\text{INI}}$
  for $g \in \mathcal{G}$ **do**
    $\mathbf{R} \leftarrow \mathbf{\Phi}_{-g} \boldsymbol{Z}_{-g} - \boldsymbol{Y}$
    $\mathbf{r} \leftarrow \mathbf{R} \boldsymbol{Z}_g^\top$
    if $\|\mathbf{r}\|_F \leq \lambda$ **then**
      $\mathbf{\Phi}_g^* \leftarrow \mathbf{0}_{|g|}$
      $\mathcal{A}_g \leftarrow \emptyset$
    end if
    if $\|\mathbf{r}\|_F > \lambda$ **then**
      $\Delta \leftarrow$ the root of $\phi(\Delta)$ defined in (A.9)
      $\text{vec}(\mathbf{\Phi}_g) \leftarrow -(\mathbf{X}_g + \frac{\lambda}{\Delta}\mathbf{I})^{-1}\mathbf{r}$
      $\mathcal{A}_g \leftarrow g$
    end if
  end for
return $\mathbf{\Phi}_\lambda$, $\mathcal{A}$
end procedure
procedure THRESHOLDUPDATE$(\mathcal{A}_\lambda, \mathbf{\Phi}_{\lambda,\text{INI}}, \lambda)$
  if $\mathcal{A} = \emptyset$ **then return** $\mathbf{0}_{k \times kp+ms}$
  end if
  if $\mathcal{A} \neq \emptyset$ **then**
    $\mathbf{\Phi}_{\lambda,\text{OLD}} \leftarrow \mathbf{\Phi}_{\lambda,\text{INI}}$
    repeat
      $\mathbf{\Phi}_{\lambda,\text{NEW}}, \mathcal{A}_\lambda \leftarrow$ BlockUpdate$(\mathcal{A}_\lambda, \mathbf{\Phi}_{\lambda,\text{OLD}}, \lambda)$
      $\mathbf{\Phi}_{\lambda,\text{OLD}} \leftarrow \mathbf{\Phi}_{\lambda,\text{NEW}}$
    until Desired threshold is reached
  end if
return $\mathbf{\Phi}_{\lambda,\text{NEW}}, \mathcal{A}$
end procedure

Algorithm 4 Sparse Group LASSO-VARX(p) inner loop

Require: $\mathbf{B}_0, \boldsymbol{Z}_q, \boldsymbol{R}_{-q}, h$

$h \leftarrow \frac{1}{\sqrt{\sigma_1(\boldsymbol{Z}_q \boldsymbol{Z}_q^\top)}}$

$\mathbf{B}_0 \leftarrow \mathbf{B}^1$

repeat

$j \leftarrow 1$

$\boldsymbol{G}_q \leftarrow \frac{(\mathbf{B}^j \boldsymbol{Z}_q - \boldsymbol{R}_{-q})\boldsymbol{Z}_q^\top}{k}$

$\gamma^j \leftarrow \mathbf{B}^j \ \text{vec}(\boldsymbol{\gamma}^{(j+1)}) \leftarrow \left(1 - \frac{h(1-\alpha)\lambda}{\|ST(\mathbf{B}^j - h(\mathbf{B}^j \boldsymbol{Z}_q - \boldsymbol{R}_{-q})\boldsymbol{Z}_q^\top, h\alpha\lambda)\|_F}\right)_+ ST(\text{vec}(\boldsymbol{B}^j) - h\mathrm{G}_q, h\alpha\lambda)$

$\boldsymbol{B}^{j+1} \leftarrow \gamma^{j+1} + \frac{j}{j+3}(\gamma^{j+1} - \gamma^j)$

$j \leftarrow j + 1$

until Desired threshold is reached

Algorithm 5 Solving (A.14) at lag ℓ [Nicholson et al. [2014]]

Require: $\tilde{v}, g_1, g_2$

$r \leftarrow \tilde{v}$

for $h = 1, 2$ **do**

$r_{g_h} \leftarrow (1 - \lambda / \|r_{g_h}\|_F)_+ r_{g_h}$

end for

return r.