Published in final edited form as:
J Am Stat Assoc. 2023 ; 118(541): 571–582. doi:10.1080/01621459.2021.1942013.

Sparse Identification and Estimation of Large-Scale Vector AutoRegressive Moving Averages

Ines Wilms[a,*], **Sumanta Basu**[b,*], **Jacob Bien**[c,†], **David S. Matteson**[b]

[a]Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands

[b]Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA

[c]Data Sciences and Operations, University of Southern California, Los Angeles, CA, USA

Abstract

The Vector AutoRegressive Moving Average (VARMA) model is fundamental to the theory of multivariate time series; however, identifiability issues have led practitioners to abandon it in favor of the simpler but more restrictive Vector AutoRegressive (VAR) model. We narrow this gap with a new optimization-based approach to VARMA identification built upon the principle of parsimony. Among all equivalent data-generating models, we use convex optimization to seek the parameterization that is simplest in a certain sense. A user-specified strongly convex penalty is used to measure model simplicity, and that same penalty is then used to define an estimator that can be efficiently computed. We establish consistency of our estimators in a double-asymptotic regime. Our non-asymptotic error bound analysis accommodates both model specification and parameter estimation steps, a feature that is crucial for studying large-scale VARMA algorithms. Our analysis also provides new results on penalized estimation of infinite-order VAR, and elastic net regression under a singular covariance structure of regressors, which may be of independent interest. We illustrate the advantage of our method over VAR alternatives on three real data examples.

Keywords

Identifiability; Forecasting; Multivariate Time Series; Sparse Estimation; VARMA

1 Introduction

Learning regulatory dynamics and forecasting are two canonical problems in the analysis of multivariate time series, with widespread applications in economics, signal processing and biostatistics amongst others. In recent years, there has been increasing focus in networks or graphical models of time series to describe how a multivariate time series' components interact with each other. Vector AutoRegressions (VAR) estimated using parsimony-inducing regularization (penalties or priors) have become a popular alternative [8, 14, 29, 18] to factor modeling of high-dimensional time series, e.g., [7]. In the classical time

[†]Corresponding author. jbien@usc.edu, http://faculty.marshall.usc.edu/Jacob-Bien/ (J. Bien).
[*]Equal Contribution.

series and signal processing literatures, Vector AutoRegressive Moving Average (VARMA) models are known to provide a more parsimonious description of a linear time invariant system than VAR. However, in practice, their use has been limited due to identification and estimation issues. The goal of this work is to overcome these challenges by theoretically and empirically investigating the large-scale VARMA as a competitive alternative to the VAR.

In a VARMA$_d(p, q)$ model, a stationary d-dimensional mean-zero vector time series y_t is modeled as a function of its own p past values and q lagged error terms. More precisely,

$$y_t = \sum_{\ell=1}^{p} \Phi_\ell y_{t-\ell} + \sum_{m=1}^{q} \Theta_m a_{t-m} + a_t, \qquad (1)$$

where $\left\{\Phi_\ell \in \mathbb{R}^{d \times d}\right\}_{\ell=1}^{p}$ are autoregressive parameter matrices, $\left\{\Theta_m \in \mathbb{R}^{d \times d}\right\}_{m=1}^{q}$ are moving average parameter matrices, and a_t denotes a d-dimensional mean-zero white noise vector time series with $d \times d$ nonsingular contemporaneous covariance matrix Σ_a. The primary focus of this work is to consider VARMA models where d is moderate or large. A VAR is a special case of the VARMA without moving average coefficients ($\Theta_m = \mathbf{0}_{d\times d}$, for $m = 1,\ldots,q$).

Although VARs are more intensively investigated (e.g., [13, 37] for computational contributions; [10, 30, 51, 9] for theoretical contributions, and [36, 23] for applications), several reasons exist for preferring the more general VARMA class. Unlike VAR, the class of VARMA is closed under marginaliztion and linear transformation [33]. In macroeconomics, VARMA is popular for its close link with linearized dynamic stochastic general equilibrium (DSGE) models [28, 21]. A parsimonious finite order VARMA can capture the dynamics of a potentially infinite-order VAR, leading to improved estimation and forecasting accuracy. Empirically, VARMAs have been shown to outperform VARs in terms of estimation and forecasting accuracy [28, 4]. Our empirical analysis also demonstrates such improvements (see Section 5). Importantly, we see that VARMA achieves this improved forecast accuracy using a more parsimonious description of the data than VAR.

Despite its advantages over VAR, VARMA has not been very popular among practitioners due to its computational and theoretical challenges in model identification and specification. The model (1) is not identifiable in general (see Section 2.1), i.e. there can be different combinations of AR and MA matrices $\{\Theta_\ell\}$ and $\{\Theta_m\}$ that lead to the same data generating process. The problem of *model identification* refers to finding a “simple” element in this equivalence set $\mathscr{E}$ of all such AR-MA matrices (see Section 2 for formal definition), usually by specifying a number of restrictions on model parameters. The problem of *model specification* refers to finding these restrictions along with the model orders p, q in a data-driven fashion.

Arguably the most popular identification procedure is the *Echelon form identification* [25, 40, 12], which amounts to selecting a basis for the row space of a block Hankel matrix (see Section 4 of [15]). Specifying an Echelon form involves selecting *Kronecker orders* (related to indices of rows that form the above basis) from a $O((p+q)^d)$-dimensional set, by comparing an equally large number of models. Data-driven strategies, involving a series of canonical correlation tests, or regressions based on model selection criteria (e.g., AIC, BIC, information theoretic criterion) were proposed [2, 3, 39]. However, all of these methods are computationally intensive and lack a formal asymptotic theory that combines specification and estimation. Assuming d is fixed, [39] proved asymptotic theory for the specification step. Then, assuming Kronecker orders are known, consistency of parameter estimation was established. This procedure has been tested only on very small d, and finite sample performances are not clear (Section 3.4, [34]).

Other popular identification and specification methods include scalar component models [44, 6, 5] and final equations form [53, 24, 48]. While these and other existing identification procedures [25, 40, 12] require different sets of assumptions—sometimes more relaxed ones than we will consider—on the structure of the process, they inherently face the same limitations for large-scale models. The uncertainty and error in the data-driven specification stage is not accounted for in the analysis of the model parameter estimation stage.

These computational and theoretical challenges of aggregating the model selection and parameter estimation are akin to the variable selection challenges in linear regression, where shrinkage methods (e.g., ridge, lasso, elastic net) have been successfully used in combining selection and parameter estimation. A key advantage of these approaches is that they allow formal asymptotic analysis of the complete specification-plus-estimation procedure.

In this work, we show that these convex optimization based techniques of regularization and dimension reduction, by now ubiquitous in the field of high-dimensional statistics, provide new perspectives and solutions to large-scale VARMA identification and estimation problems with several attractive properties.

I. Automatic identification of parsimonious VARMA models.

We show that by devising a suitable convex penalty, we can identify a parsimonious element in the equivalence class $\mathscr{E}$ in an intuitive yet objective fashion (Section 2). More formally, we can define the class of AR-MA matrices with minimum ℓ_1-norm as a partially identified class of “sparse” VARMA models $\mathscr{R}\mathscr{E} = \text{argmin}_{(\Phi, \Theta) \in \mathscr{E}}\{\sum_{\ell = 1}^{p} \|\Phi_\ell\|_1 + \sum_{m = 1}^{q} \|\Theta_m\|_1\}$. We could also use a modified, strongly convex penalty $\text{argmin}_{(\Phi, \Theta) \in \mathscr{E}}\{\sum_{\ell = 1}^{p}(\|\Phi_\ell\|_1 + \alpha\|\Phi_\ell\|_F^2) + \sum_{m = 1}^{q}(\|\Theta_m\|_1 + \alpha\|\Theta_m\|_F^2)\}$ with a very small $\alpha \approx 0$ to identify a parsimonious element in $\mathscr{R}\mathscr{E}$, viz. the *unique* AR-MA matrices with minimum Frobenius norm (Proposition 2.1).

II. Computationally efficient estimation of VARMA models.

Our identification strategy explicitly links the search for a unique, parsimonious model throughout the identification, specification and estimation stages. The same penalty used in our identification is used as a regularizer to define a natural VARMA estimator corresponding to this identified target (Section 3). We show on real and simulated data examples (Section 5 and Appendix G) that such parsimonious VARMA models lead to important gains in forecast accuracy compared to parsimoniously estimated VARs. An implementation of our fully-automated VARMA identification and estimation procedure is available in the R package bigtime [49].

III. Non-asymptotic theory for sparse VARMA.

We also provide a non-asymptotic theoretical analysis of our proposed sparse VARMA estimator (Section 4). Our analysis explicitly captures the complexity of model selection, and does not assume the identification restrictions are known *a priori* as in existing asymptotic analysis of VARMA [16, 19]. While to the best of our knowledge, consistency of VARMA estimators has been studied only in the low-dimensional, fixed d asymptotic regime [28, 20], our error bound analysis shows consistent estimation is possible in a double-asymptotic regime $d, T \to \infty$. We provide two main results on consistency (Proposition 4.1). Our first result in the spirit of *partial identification* [35, 43] states that under suitable sparsity assumptions our algorithm provides a parsimonious VARMA estimator (small ℓ_1-norm) whose distance from the equivalence class $\mathscr{E}$ asymptotically vanishes as long as log $d / T \to 0$. Our second result on *point identification* states that our estimator converges in probability to our identified target in $\mathscr{E}$ as long as d^4 log $d / T \to 0$.

2 Identification of the VARMA

We revisit the VARMA identification problem in Section 2.1, then introduce an optimization-based, parsimonious identification strategy for VARMA in Sections 2.2 and 2.3.

2.1 Identification Problem

Consider the VARMA$_d(p, q)$ of Equation (1) with fixed autoregressive order p and moving average order q. The model can be written using compact lag operators as $\Phi(L)y_t = \Theta(L)a_t$, where the AR and MA operators are respectively given by

$$\Phi(L) = I - \Phi_1 L - \Phi_2 L^2 - \ldots - \Phi_p L^p \text{ and } \Theta(L) = I + \Theta_1 L + \Theta_2 L^2 + \ldots + \Theta_q L^q,$$

with the lag operator L^ℓ defined as $L^\ell y_t = y_{t-\ell}$. We assume the model is stable and invertible meaning respectively that det{$\Phi(z)$} ≠ 0 and det{$\Theta(z)$} ≠ 0 for all $|z| \leq 1 (z \in \mathbb{C})$. The process $\{y_t\}$ then has an infinite-order VAR representation $\Pi(L)y_t = a_t$, where $\Pi(L) = \Theta^{-1}(L)\Phi(L) = I - \Pi_1 L - \Pi_2 L^2 - \cdots$, with det{$\Pi(z)$}≠ 0 for all $|z| \leq 1$. The Π-matrices can be computed recursively from the AR matrices $\{\Phi_\ell\}$ and MA matrices $\{\Theta_m\}$ (e.g., [11], Chapter 11). The VARMA is uniquely defined in terms of the operator $\Pi(L)$, but not in terms of the AR and MA operators $\Phi(L)$ and $\Theta(L)$, in general. That is, for a given $\Pi(L)$, p, and q, one can define an equivalence class of AR and MA matrix pairs,

$$\mathscr{E}_{p,q}(\Pi(L)) = \{(\Phi, \Theta) : \Phi(L) = \Theta(L)\Pi(L)\},$$

where $\Phi = [\Phi_1 \cdots \Phi_p]$ and $\Theta = [\Theta_1 \cdots \Theta_q]$. This class can, in general, consist of more than one such pair, implying that further identification restrictions on the AR and MA matrices are needed for meaningful estimation.

In order to connect identification to estimation, we first provide an alternate characterization of the equivalence class $\mathscr{E}_{p,q}(\Pi(L))$ in terms of a Yule-Walker type equation.

Proposition 2.1—(Yule-Walker type equation for VARMA). *Consider a white noise process* $\{a_t\}_{t\in\mathbb{Z}}$ *with mean zero and variance* Σ_a. *For a stable, invertible linear filter* $\Pi(L)$ *that allows a* VARMA$_d$(p, q) *representation* $\Pi(L) = \Theta^{-1}(L)\Phi(L)$, *consider the process* $y_t = \Pi^{-1}(L)a_t$ *and define* $z_t = \left[y_{t-1}^\top:\cdots:y_{t-p}^\top:a_{t-1}^\top:\cdots:a_{t-q}^\top\right]^\top$. Then, $(\Phi,\Theta)\in\mathscr{E}_{p,q}(\Pi(L))$ *if and only if* $\beta_{d(p+q)\times d} := [\Phi_1:\ldots:\Phi_p:\Theta_1:\ldots:\Theta_q]^\top$ *is a solution to the system of equations* $\rho_{zy} = \Sigma_z\beta$, *where* $\rho_{zy} = \mathbb{E}\left[z_t y_t^\top\right]$ *and* $\Sigma_z = \mathbb{E}\left[z_t z_t^\top\right]$. *That is,*

$$\mathscr{E}_{p,q}(\Pi(L)) = \left\{(\Phi,\Theta):\rho_{zy} = \Sigma_z\beta\right\}. \tag{2}$$

A proof of this proposition is provided in Appendix A.1. Note that both ρ_{zy} and Σ_z can be expressed as functions of Π and Σ_a alone (i.e. they do not depend on Θ and Φ), and hence are uniquely defined for the underlying process y_t. While the AR(∞) representation given by Π in Proposition 2.1 is unique, it allows an equivalent characterization in terms of many (Φ, Θ) combinations. Each of these combinations is a solution to the (potentially) underdetermined system of equations in Proposition 2.1.

A key consequence of this proposition is that our identification target can be defined by optimizing over the solution set of this Yule-Walker type equation. Further, we can use sample analogues of ρ_{zy} and Σ_z in our estimation step to search for this target in a data-driven fashion.

2.2 Optimization-based Identification

We rely on strongly convex optimization to establish identification for VARMA models. Among all feasible AR and MA matrix pairs, we look for the one that gives the most parsimonious VARMA representation. We measure parsimony through a pair of convex regularizers, $\mathscr{P}_{\mathrm{AR}}(\Phi)$and $\mathscr{P}_{\mathrm{MA}}(\Theta)$. Our identification results apply equally well to any convex function: one may consider, amongst others, the ℓ_1-norm, the ℓ_2-norm, the nuclear norm, and combinations thereof. Our methodology also allows for a different choice of convex function for the AR and MA matrices if prior knowledge would allow a more informed modeling approach. This might be particularly useful in economics, for instance, where one may be interested in a parsimonious AR structure for interpretability, but can allow for a non-sparse MA polynomial to increase forecast accuracy.

We now define the *regularized* equivalence class of VARMA representations as

$$\mathscr{R}\mathscr{E}_{p,q}(\Pi(L)) = \underset{\Phi,\Theta}{\operatorname{argmin}}\left\{\mathscr{P}_{\mathrm{AR}}(\Phi) + \mathscr{P}_{\mathrm{MA}}(\Theta)\ \text{s.t.}\ \Phi(L) = \Theta(L)\Pi(L)\right\}. \tag{3}$$

This regularized equivalence class is a subclass of the equivalence class $\mathscr{E}_{p,q}(\Pi(L))$, containing the regularized VARMA representations. If the objective function in (3) is strongly convex, then the regularized equivalence class consists of one unique AR-MA matrix pair, in which case identification is established. However, for the ℓ_1-norm, for instance, the objective function is convex but not strongly convex. Hence, to ensure identification for this case, we add two extra terms to the objective function and consider

$$(\Phi^{(\alpha)}, \Theta^{(\alpha)}) = \underset{\Phi,\Theta}{\operatorname{argmin}}\{\mathscr{P}_{\mathrm{AR}}(\Phi) + \mathscr{P}_{\mathrm{MA}}(\Theta) + \frac{\alpha}{2}\|\Phi\|_F^2 + \frac{\alpha}{2}\|\Theta\|_F^2 \text{ s.t. } \Phi(L) = \Theta(L)\Pi(L)\}\,. \quad (4)$$

Problem (4) is strongly convex and thus has a *unique* solution pair $(\Phi^{(\alpha)}, \Theta^{(\alpha)})$ for each $\alpha > 0$. For any stable, invertible VARMA, we then define its unique regularized representation in terms of the AR-MA matrices as

$$(\Phi^{(0)}, \Theta^{(0)}) = \lim_{\alpha \to 0^+} (\Phi^{(\alpha)}, \Theta^{(\alpha)})\,. \quad (5)$$

The following proposition, proved in Appendix A.2, establishes that $(\Phi^{(0)}, \Theta^{(0)})$ is in the regularized equivalence class $\mathscr{RE}_{p,q}(\Pi(L))$ and furthermore is the *unique* pair of autoregressive and moving average matrices in this set having the smallest Frobenius norm. This result is similar to a result in the regression context, which states that the LARS-lasso solution has the minimum ℓ_2-norm over all lasso solutions (see [45], Lemma 7).

Proposition 2.2.—The limit in (5) exists and is the unique pair in the set $\mathscr{RE}_{p,q}(\Pi(L))$ whose Frobenius norm squared is smallest:

$$(\Phi^{(0)}, \Theta^{(0)}) = \underset{\Phi,\Theta}{\operatorname{argmin}}\left\{\|\Phi\|_F^2 + \|\Theta\|_F^2 \text{ s.t. } (\Phi, \Theta) \in \mathscr{RE}_{p,q}(\Pi(L))\right\}.$$

2.3 Sparse Identification

While our identification results apply equally well to any convex function, we give special attention to sparsity-inducing convex regularizers. In this case, the regularized equivalence class in (3) is a sparse equivalence class, meaning that, in general, we would expect many of the elements of the AR and/or MA matrices to be exactly equal to zero.

To guarantee the sparsest VARMA representation, one might consider taking $\mathscr{P}_{\mathrm{AR}}(\Phi) = \|\Phi\|_0$ and $\mathscr{P}_{\mathrm{MA}}(\Theta) = \|\Theta\|_0$. However, since the ℓ_0-penalty is non-convex, a unique solution cannot be guaranteed. One can construct examples in which there exist multiple equivalent, sparsest VARMAs, see [46] and Appendix A.3.1. Strong convexity in (4) is key to guaranteeing uniqueness of $(\Phi^{(\alpha)}, \Theta^{(\alpha)})$. For sparsity, we may therefore add to the ℓ_2-norm in (4) the ℓ_1-norm $\mathscr{P}_{\mathrm{AR}}(\Phi) = \|\Phi\|_1$ and $\mathscr{P}_{\mathrm{MA}}(\Theta) = \|\Theta\|_1$ as a sparsity-inducing convex heuristic.

While our theory will focus on the ℓ_1-norm, in the empirical sections we also investigate a time-series specific alternative penalty, the hierarchical lag (hereafter "HLag") penalty [38, 50]: $\mathscr{P}_{\mathrm{AR}}(\Phi) = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{\ell=1}^{p}\|\Phi_{(\ell:p),ij}\|$, and $\mathscr{P}_{\mathrm{MA}}(\Theta) = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{m=1}^{q}\|\Theta_{(m:q),ij}\|$, with $\Phi_{(\ell:p),ij} = [\Phi_{\ell,ij}\ldots\Phi_{p,ij}] \in \mathbb{R}^{(p-\ell+1)}$ and $\Theta_{(m:q),ij} = [\Theta_{m,ij}\ldots\Theta_{q,ij}] \in \mathbb{R}^{(q-m+1)}$. This penalty involves a lag-based hierarchical group lasso penalty (e.g., [52]) on the AR (or MA) parameters. It allows for automatic lag selection by forcing lower lags of a time series in one of the VARMA equations to be selected before its higher order lags and is thus built on the intuition of encouraging increased sparsity in Φ_ℓ and Θ_ℓ as the lag increases.

3 Sparse Estimation of the VARMA

We estimate and determine the degree of parsimony of VARMA parameters by the use of convex regularizers. Since the $\text{VARMA}_d(p, q)$ of Equation (1) cannot be directly estimated as it contains the unobservable (latent) lagged errors, we proceed in two phases, in the spirit of [41, 19], and references therein. In Phase-I, we approximate these unobservable errors. In Phase-II, we estimate the VARMA with the approximated lagged errors.

3.1 Phase-I: Approximating the unobservable errors

The VARMA of Equation (1) has a pure VAR(∞) representation if it is invertible (Section 2.1). We therefore approximate the errors a_t by the residuals of a $\text{VAR}(\tilde{p})$ given by

$$y_t = \sum_{\tau = 1}^{\tilde{p}} \prod_\tau y_{t - \tau} + \varepsilon_t, \tag{6}$$

for $(\tilde{p} + 1) \leq t \leq T$, with $\tilde{p}$ a finite number, $\left\{\Pi_\tau \in \mathbb{R}^{d \times d}\right\}_{\tau = 1}^{\tilde{p}}$ the AR parameter matrices, and ε_t a vector error series. Denote the estimates by $\widehat{\Pi}_\tau$ and residuals by $\varepsilon_t = y_t - \sum_{\tau = 1}^{\tilde{p}} \widehat{\Pi}_\tau y_{t - \tau}$.

Estimating the $\text{VAR}(\tilde{p})$ of Equation (6) is challenging since $\tilde{p}$ needs to be sufficiently large such that the residuals ε_t accurately approximate the errors a_t. Since, a large number of parameters $\left(\tilde{p}d^2\right)$, relative to the time series length T, needs to be estimated, we use regularized estimation. For ease of notation, first rewrite model (6) in compact matrix notation $Y = \Pi Z + E$, where $Y = [y_{\tilde{p} + 1} \ldots y_T] \in \mathbb{R}^{d \times (T - \tilde{p})}$, $Z = [z_{\tilde{p} + 1} \ldots z_T] \in \mathbb{R}^{d\tilde{p} \times (T - \tilde{p})}$, with $z_t = \left[y_{t - 1}^\top \ldots y_{t - \tilde{p}}^\top\right]^\top \in \mathbb{R}^{(d\tilde{p} \times 1)}$, $E = [\varepsilon_{\tilde{p} + 1} \ldots \varepsilon_T] \in \mathbb{R}^{d \times (T - \tilde{p})}$, and $\Pi = [\Pi_1 \ldots \Pi_{\tilde{p}}] \in \mathbb{R}^{d \times d\tilde{p}}$. The regularized autoregressive estimates $\widehat{\Pi}$ are obtained as

$$\Pi = \underset{\Pi}{\operatorname{argmin}}\left\{\frac{1}{2}\left\|Y - \Pi Z\right\|_F^2 + \lambda_\Pi \mathcal{P}(\Pi)\right\}, \tag{7}$$

where we use the squared Frobenius norm as loss function and $\mathcal{P}(\Pi)$ is any convex regularizer. In our simulations and applications, we focus on sparsity-inducing regularizers (ℓ_1-norm or HLag penalty). The penalty parameter $\lambda_\Pi > 0$ then regulates the degree of sparsity in Π: the larger λ_Π, the sparser Π. Problem (7) can be efficiently solved using Algorithm 1 in [38].

3.2 Phase-II: Estimating the VARMA

We continue with the approximated lagged errors $\varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ instead of the true errors $a_{t-1}, \ldots, a_{t-q}$ in Equation (1). The resulting model

$$y_t = \sum_{\ell=1}^{p} \Phi_\ell y_{t-\ell} + \sum_{m=1}^{q} \Theta_m \varepsilon_{t-m} + u_t, \tag{8}$$

is a regression of y_t on $y_{t-1},\ldots,y_{t-p}$, $\varepsilon_{t-1},\ldots,\varepsilon_{t-q}$ with vector error series u_t. To tackle the VARMA overparameterization problem and establish identification simultaneously with estimation, we again use regularization.

Rewrite the lagged regression (8) in compact matrix notation $Y = \Phi Z + \Theta X + U$, where $Y = [y_{\bar{o}+1}\ldots y_T] \in \mathbb{R}^{d\times(T-\bar{o})}$, $Z = [z_{\bar{o}+1}\ldots z_T] \in \mathbb{R}^{d\hat{p}\times(T-\bar{o})}$, with $z_t = \left[y_{t-1}^\top \ldots y_{t-\hat{p}}^\top\right]^\top \in \mathbb{R}^{(d\hat{p}\times 1)}$, $X = [x_{\bar{o}+1}\ldots x_T] \in \mathbb{R}^{d\hat{q}\times(T-\bar{o})}$ with $x_t = \left[\varepsilon_{t-1}^\top\ldots\varepsilon_{t-\hat{q}}^\top\right]^\top \in \mathbb{R}^{(d\hat{q}\times 1)}$, with $\bar{o} = \max(\hat{p},\hat{q})$, for specified order $\hat{p}$, $\hat{q}$, $U = [u_{\bar{o}+1}\ldots u_T] \in \mathbb{R}^{d\times(T-\bar{o})}$, $\Phi = [\Phi_1\ldots\Phi_{\hat{p}}] \in \mathbb{R}^{d\times d\hat{p}}$, and $\Theta = [\Theta_1\ldots\Theta_{\hat{q}}] \in \mathbb{R}^{d\times d\hat{q}}$. The regularized VARMA estimates are obtained as:

$$\begin{aligned}(\Phi^{(\alpha)},\Theta^{(\alpha)}) = \underset{\Phi,\Theta}{\operatorname{argmin}}\{\frac{1}{2}\|Y-\Phi Z-\Theta X\|_F^2 + \lambda_\Phi \mathscr{P}_{\mathrm{AR}}(\Phi) + \lambda_\Theta \mathscr{P}_{\mathrm{MA}}(\Theta)\\ + \frac{\alpha}{2}(\lambda_\Phi\|\Phi\|_F^2 + \lambda_\Theta\|\Theta\|_F^2)\},\end{aligned} \tag{9}$$

where λ_Φ, λ_Θ, > 0 are two penalty parameters. By adding the regularizers $\mathscr{P}_{\mathrm{AR}}(\Phi)$ and $\mathscr{P}_{\mathrm{AR}}(\Phi)$ to the objective function, estimation of large-scale VARMAs is feasible. The addition of the squared Frobenius norms makes the problem strongly convex, ensuring a unique solution in the same way as was done in the identification scheme (4). Optimization problem (9) can be solved via the proximal gradient algorithm in Appendix F. We investigate the forecast accuracy of the proposed VARMA on simulated data in Appendix G.

3.3 Choosing Tuning Parameters

The estimation procedure involves three sets of user-defined choices: (i) the maximum lag orders $\tilde{p}$, $\hat{p}$, $\hat{q}$; (ii) the penalty parameters λ_Π, λ_Φ, λ_Θ; and (iii) the parameter α to ensure uniqueness. We choose these in either a data-driven or computationally inexpensive manner. Below we motivate our choices and address implications of misspecification.

The maximal lag orders $\tilde{p}$, $\hat{p}$, and $\hat{q}$.—We take $\tilde{p} = \lfloor 1.5\sqrt{T}\rfloor$ and $\hat{p} = \hat{q} = \lfloor 0.75\sqrt{T}\rfloor$. Our theoretical analysis suggests that $\tilde{p} \asymp T^{\frac{1}{2}-\epsilon}$ (Proposition 4.2), and for larger d, overselecting AR/MA orders only affects the estimation and prediction performance at a rate of log d (Proposition 4.4). To simplify practical implementation, we therefore set these values at a slightly larger order $O(\sqrt{T})$.

We perform a simulation study (Appendix G.4) to investigate misspecification of the maximal lag orders. We find that, in general, overselecting is less severe than underselecting. The price to pay for overselection is smaller for the HLag penalty than for the ℓ_1-penalty since the former performs automatic lag selection. As such, it can reduce the effective maximal order of each series in each equation of the VAR (Phase-I) and VARMA (Phase-II).

The penalty parameters λ_{Π}, λ_{Φ}, and λ_{Θ}.—We select the penalty parameters using cross-validation. Below, we describe the selection of λ_{Π} in Phase-I; in Phase-II, we proceed similarly but using a two-dimensional grid search for the penalty parameters $(\lambda_{\Phi}, \lambda_{\Theta})$.

Following [22], we use a grid of ten penalty parameters starting from $\lambda_{\Pi,\max}$, an estimate of the smallest value for which all parameters are zero, and then decreasing in log linear increments. We then use the following time series cross-validation approach: For each time point $t = S, \ldots, T - h$, with $S = \lfloor 0.9 \cdot T \rfloor$ and forecast horizon h, we estimate the model and obtain parameter estimates. This results in ten different parameter estimates, one for each value of the penalty parameter in the grid. From these estimates, we compute h-step ahead forecasts $y_{t+h}^{(\lambda)}$ obtained with penalty parameter λ. We select the value of λ_{Π} that gives the most regularized model whose Mean Squared Forecast Error

$$\mathrm{MSFE}_h^{(\lambda)} = \frac{1}{T - h - S + 1} \sum_{t = S}^{T - h} \frac{1}{d} \| y_{t+h} - y_{t+h}^{(\lambda)} \|^2,$$

is within one standard error (see [26]; Chapter 7) of the minimal MSFE. In simulations, we take $h = 1$; in the forecast applications, we also consider other forecast horizons.

The parameter α.—We will sometimes refer to Equation (9) as an “elastic net” problem, although, unlike λ_{Φ} and λ_{Θ}, the parameter α is not treated as a statistical tuning parameter; rather, as a small positive value simply used to ensure uniqueness. Our simulation study in Appendix A.3.2 reveals that the addition of a small non-zero α indeed produces sparse VARMA estimates close to the unique $(\Phi^{(0)}, \Theta^{(0)})$ pair defined in Equation (5). For α = 0, we still retrieve sparse VARMA estimates that are close to *an* element in the sparse equivalence class. The resulting estimates are typically sparser (i.e. they have fewer non-zero components) than the estimates obtained with a small non-zero α since the target $(\Phi^{(0)}, \Theta^{(0)})$ corresponds to the pair with minimum Frobenius norm among all minimum-ℓ_1 VARMA representations. Since our main objectives are to produce VARMA estimates that are close to the sparse equivalent class and have good out-of-sample forecast performance, we prefer to work with the sparser estimates and thus take $\alpha = 0$ in practice, as we have done in our forecast applications (Section 5) and simulations (Appendix G).

4 Theoretical Properties

We establish consistency of our VARMA estimator with the lasso penalty in Phase-I and elastic net penalty in Phase-II under a double asymptotic regime where dimension d grows with the sample size. Our Phase-II estimator is essentially an elastic net regression, but introduces additional complexities compared to i.i.d. or stochastic regression that need to be dealt with in the asymptotic analysis. The rows of the design matrix consist of consecutive observations from an *approximate* version of the time series $z_t = [y_{t-1}^{\top} : \ldots : y_{t-p}^{\top} : a_{t-1}^{\top} : \ldots : a_{t-q}^{\top}]^{\top}$, with a_t approximated by Phase-I residuals $\hat{\varepsilon}_t$. The error term in the regression involves $\hat{\varepsilon}_t$ which do not have an analytically tractable distribution. In addition, since $\Phi(L)y_t = \Theta(L)a_t$, the population covariance matrix of the predictors Σ_z is potentially singular. It is not clear whether a restricted eigenvalue (RE) assumption, commonly used in high-dimensional regression [32], holds in Phase-II regression.

We start by establishing in Section 4.1 deterministic upper bounds on the estimation error of a generic elastic net regression under some sufficient conditions. A crucial step to verify these sufficient conditions is to derive upper bounds to control the approximation error of a_t by $\hat{\varepsilon}_t$ in Phase-I. We do this in Section 4.2. Finally, in Section 4.3 we show that these sufficient conditions for Phase-II elastic net regression are satisfied with high probability for random realizations from the VARMA model, and present estimation error bounds. To maintain analytical tractability when tackling the VARMA specific complexities, we consider two modifications in Phase-II. First, we use $\hat{y}_t := y_t - \hat{\varepsilon}_t$, the fitted values from Phase-I, instead of y_t, as response in Phase-II. The analysis can be modified in a straightforward fashion to use y_t as response, although the resulting upper bounds become larger. Second, we consider a constrained version of the penalized Phase-II estimator with an additional side constraint on the ℓ_1-norm of the regression coefficient. Equivalence of the constrained and penalized versions follows from duality of the convex programs. The additional side constraint on the regression coefficient is easy to implement in practice [1], and has been used for technical convenience in earlier literature on high-dimensional statistics [32].

We assume Gaussianity in our analysis, primarily to apply some concentration inequalities for Gaussian processes in our non-asymptotic error bound analysis. The results can be extended to non-Gaussian VARMA using recent concentration bounds for non-Gaussian linear processes [42] with potentially slower convergence rate for processes with heavier tails than Gaussian, although the technical exposition becomes more cumbersome.

Notation.

We denote the sets of integers, real, and complex numbers by $\mathbb{Z}$, $\mathbb{R}$, and $\mathbb{C}$, respectively. We use $\|.\|$ to denote the Euclidean norm of a vector and the operator norm of a matrix. We reserve $\|.\|_0$, $\|.\|_1$ and $\|.\|_\infty$ to denote the number of nonzero elements, ℓ_1 and ℓ_∞ norms of a vector or the vectorized version of a matrix, respectively, and $\|.\|_F$ to denote the Frobenius norm of a matrix. For a matrix-valued, possibly infinite-order lag polynomial $\mathscr{A}(L) = \sum_{\ell \geq 0} A_\ell L^\ell$, we define $|||\mathscr{A}||| := \max_{\theta \in [-\pi, \pi]} \|\mathscr{A}(e^{i\theta})\|$, and use $\mathscr{A}_{[k]}(L)$ and $\mathscr{A}_{-[k]}(L)$ to denote the truncated version $\sum_{\ell = 0}^{k} A_\ell L^\ell$ and the tail series $\sum_{\ell > k} A_\ell L^\ell$, respectively. We also use $\|\mathscr{A}\|_{2,1}$ to denote the sum of the operator norms of its coefficients, $\sum_{\ell \geq 0} \|A_\ell\|$. More generally, for any complex matrix-valued function $f(\theta)$ of frequencies $\theta \in [-\pi, \pi]$ to $\mathbb{C}^{p \times p}$, we define $|||f||| := max_{\theta \in [-\pi, \pi]} \| f(\theta) \|$. In our theoretical analyses, we use c_i, $i = 0,1,2,\ldots$, to denote universal positive constants whose values do not rely on the model dimensions and parameters. For two model dependent positive quantities A and B, we also use $A \gtrsim B$ to mean that for any universal constant $c > 0$, we have $A \geq cB$ for sufficiently large sample size. Finally, $A \asymp B$ means $A \gtrsim B$ and $A \lesssim B$.

Remark 4.1 (Measures of Dependence).—We adopt the spectral density based measures of dependence introduced in [10] to capture the role of temporal dependence in our non-asymptotic error bounds. For a d-dimensional centered stationary time series $\{x_t\}_{t \in \mathbb{Z}}$ with autocovariance function $\Gamma_x(h) = \text{Cov}(x_t, x_{t+h}) = \mathbb{E}\left[x_t x_{t+h}^\top\right]$, $h \in \mathbb{Z}$, we define the spectral density function $f_x(\theta) := \frac{1}{2\pi}\sum_{\ell = -\infty}^{\infty} \Gamma_x(\ell) e^{-i\ell\theta}$, $\theta \in [-\boldsymbol{\pi}, \boldsymbol{\pi}]$. The quantity $|||f_X|||$ is taken as a measure of temporal and cross-sectional dependence in the time series $\{x_t\}$. For a stable, invertible VARMA process y_t in (1) with $\Lambda_{\min}(\Sigma_a) > 0$, it is known that f_y is non-singular on $[-\boldsymbol{\pi}, \boldsymbol{\pi}]$ and there exist two model dependent quantities $\bar{C} > 0$ and $\bar{\rho} \in [0, 1)$ such that $\|\Pi_\tau\| \le \bar{C}\bar{\rho}^\tau$, for all integers $\boldsymbol{\tau} \ge 1$ [20]. This implies for any $\tilde{p} \ge 1$, we have $\|\Pi_{-[\tilde{p}]}\|_{2,1} \le \bar{C}\bar{\rho}^{\tilde{p}}/(1 - \bar{\rho})$. The quantities $|||f_y|||$, $|||f_y^{-1}|||$ and $\|\Pi_{-[\tilde{p}]}\|_{2,1}$ appear in our error bounds, and capture the effects of temporal dependence on the convergence rates.

4.1 Elastic Net with Singular Gram Matrix

Consider an elastic net penalized regression problem where the population covariance matrix of the predictors is singular. The problem is non-identifiable in the sense that there is no "true" coefficient vector. Rather, the elastic net penalty itself is used to specify an identified target among all equivalent data-generating models. The following proposition provides deterministic upper bounds on estimation and in-sample prediction errors under some sufficient conditions. The proof is in Appendix C.

Proposition 4.1.—*Let* $\Sigma \in \mathbb{R}^{D \times D}$ *be a non-negative definite matrix with* $\Lambda_{\min}(\Sigma) = 0$ *and let* $\rho \in \mathbb{R}^D$ *be in the column space of* Σ. *For some* $\alpha \ge 0$, $y, \mathscr{E} \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N \times D}$, *consider the linear regression model* $y = X_\beta^{*(\alpha)} + \varepsilon$ *with identified target*

$$\beta^{*(\alpha)} := \underset{\beta}{\text{argmin}}\{\mathscr{P}_\alpha(\beta) \text{s.t. } \Sigma\beta = \rho\},$$

where $\mathscr{P}_\alpha(\beta) := \|\beta\|_1 + (\alpha/2)\|\beta\|^2$, *and define the estimator*

$$\hat{\beta}^{(\alpha)} := \underset{\beta : \|\beta\|_1 \le M}{\text{argmin}} \frac{1}{n}\|y - X\beta\|^2 + \lambda\mathscr{P}_\alpha(\beta),$$

for some n and M, where $M \ge \|\beta^{*(\alpha)}\|_1$. *Then for any choice of* $\lambda \ge 2\|X^\top \varepsilon / n\|_\infty$ *and* $q_n \ge \|X^\top X / n - \Sigma\|_\infty$, *the following holds:*

a. *In - Sample Prediction*: $\frac{1}{n}\|X\hat{\beta}^{(\alpha)} - X\beta^{*(\alpha)}\|^2 \le \lambda\left[2M + \alpha M^2/2\right]$,

b. *Partially - Identified Estimation*: $\min_{\beta : \Sigma\beta = p} \|\hat{\beta}^{(\alpha)} - \beta\|^2 \le \frac{4q_n M^2 + \lambda\left[2M + \alpha M^2/2\right]}{\Lambda_{\min^+}(\Sigma)}$,

where $\Lambda_{\min^+}(\Sigma)$ *is the smallest non-zero eigenvalue of* Σ.

In addition, define the constrained version of the estimator

$$\hat{\beta}_{[C]}^{(\alpha)} := \underset{\beta}{\text{argmin}}\left\{\mathscr{P}_\alpha(\beta) \text{s.t. } \frac{1}{n}\|y - X\beta\|^2 \le A_n,\ \|\beta\|_1 \le M\right\}.$$

Then, for any $r_n \geq \frac{1}{n}\|X^\top \varepsilon\|_\infty$, *and* $S_n \geq \left|\frac{1}{n}\|\varepsilon\|^2 - \sigma^2\right|$, $A_n = \sigma^2 + s_n$ *and* $M \geq \|\beta^{*(\alpha)}\|_1$, *we have*

c. *Point - Identified Estimation*: $\left\|\widehat{\beta}_{[C]}^{(\alpha)} - \beta^{*(\alpha)}\right\|^2 \leq 2\upsilon_n + 2(\sqrt{D}/\alpha + M)\upsilon_n^{1/2}$, *where* $\upsilon_n := \frac{4Mr_n + 2s_n + 4M^2 q_n}{\Lambda_{\min^+}(\Sigma)}$.

The VARMA estimator from Phase-II can be expressed in the above regression format (see Equation (14)) with $n = T - q$, $N = nd$, $\Sigma = \Sigma_Z$ and $D = d^2(p + q)$. We will show that modulo some terms capturing the effect of temporal dependence, λ, q_n, r_n can be chosen in the order of at most $O(\sqrt{\log D/n})$ with high probability.

Under this setting, part (a) will imply in-sample prediction consistency in the high-dimensional regime $\log D / n \to 0$ as long as the identification target $\beta^{*(\alpha)}$ is *weakly sparse*, i.e. its ℓ_1-norm grows sufficiently slowly. Consequently, our VARMA forecasts will asymptotically converge to the optimal forecasts.

Part (b) will ensure that the Euclidean distance of our VARMA estimator from the set of data-generating vectors $\{\beta : \Sigma_Z \beta = \rho_{ZY}\}$ converges to zero in the asymptotic regime $\log D / n \to 0$, assuming weak sparsity of $\beta^{*(\alpha)}$. The rate of convergence also relies on the curvature of the population loss captured by $\Lambda_{\min^+}(\Sigma)$.

Error bound for the point identification part (c) will imply that with an appropriate choice of s_n, consistent estimation of our identification target is possible in the double-asymptotic regime $D^2 \log(D) / n \to 0$, as long as $\beta^{*(\alpha)}$ is weakly sparse in the sense of small ℓ_1-norm. This error bound also increases linearly with the inverse of α, the parameter capturing curvature of the penalty function $\mathcal{P}_\alpha(\beta)$.

Remark 4.2.—We focus on prediction and estimation instead of model selection consistency for two reasons. First, model selection consistency in penalized regression holds only under incoherence or irrepresentable conditions [54], which are stringent even for i.i.d. data, and are not known to hold with high probability for multivariate stationary time series data. Second, since we work with *an equivalence class of models* potentially having different sparsity patterns, it is not obvious how to define sparsity of a true model, in general. However, we have conducted a simulation experiment (Appendix A.3.2) to assess model selection properties of our estimator in finite samples, which shows promising results.

4.2 Approximation Error in Phase-I

Our main interest in this section is in approximating the errors a_t by the Phase-I residuals $\widehat{\mathscr{E}}_t$ for use in Phase-II. As a by-product, we also provide estimation error bounds for VAR(∞) coefficients (see Proposition D.1).

Suppose we re-index data in the form $(y_{-(\tilde{p}-1)}, y_{-(\tilde{p}-2)}, \ldots, y_{-1}, y_0, y_1, \ldots, y_T)$. In Phase-I, we regress y_t on its most recent $\tilde{p}$ lags:

$$y_t = \sum_{\tau = 1}^{\tilde{p}} \Pi_\tau y_{t-\tau} + \varepsilon_t, \text{ where } \varepsilon_t = \left(a_t + \sum_{\tau = \tilde{p}+1}^{\infty} \Pi_\tau y_{t-\tau} \right). \tag{10}$$

The autoregressive design takes the form $\mathscr{Y}_{T \times d} = \mathscr{X}_{T \times d\tilde{p}} \boldsymbol{B}_{d\tilde{p} \times d} + E_{T \times d}$, where $\mathscr{Y} = [y_T : y_{T-1} : \ldots : y_1]^\top$, $\mathscr{X} = ((y_{T-i-j+1}))_{1 \le i \le T, 1 \le j \le \tilde{p}}$, $\boldsymbol{B} = [\Pi_1 : \ldots : \Pi_{\tilde{p}}]^\top$ and $E = [\boldsymbol{\varepsilon}_T : \boldsymbol{\varepsilon}_{T-1} : \ldots : \boldsymbol{\varepsilon}_1]^\top$. Vectorizing this regression design with T samples and $d^2\tilde{p}$ parameters, we have $Y = Z\beta^* + \text{vec}(E)$, where $Y = \text{vec}(\mathscr{Y})$, $Z = I \otimes \mathscr{X}$, and $\beta^* = \text{vec}(B)$. In Phase-I, we consider a lasso estimator

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{d^2\tilde{p}}}{\text{argmin}} \frac{1}{T} \| Y - Z\beta \|^2 + \lambda \| \beta \|_1, \tag{11}$$

where $\hat{\beta} = \text{vec}(\hat{B})$ and $\hat{B} = [\hat{\Pi}_1 : \ldots : \hat{\Pi}_{\tilde{p}}]^\top$. We denote the residuals of the Phase-I regression as $\hat{\varepsilon}_t = y_t - \sum_{\tau = 1}^{\tilde{p}} \hat{\Pi}_\tau y_{t-\tau}$.

Our next proposition provides upper bounds on the approximation error of a_t by $\hat{\varepsilon}_t$ for a random realization of $(T + \tilde{p})$ data points from the VARMA model (1). A complete proof is given in Appendix D.

Proposition 4.2.—*Consider any solution* $\hat{\beta}$ *of (11) using a random realization of* $\{y_t\}_{t = 1 - \tilde{p}}^{T}$ *from the VARMA model (1). Choose* $\tilde{p} \asymp T^{\frac{1}{2} - \epsilon}$ *for some* $\epsilon \in (0, 1/2)$, and $\lambda \ge \lambda_0$, *where*

$$\lambda_0 := 2\pi |||f_y||| \left[3A \max\left\{ |||\Pi_{[\tilde{p}]}|||^2, 1 \right\} \sqrt{\log\left(d^2\tilde{p}\right)/T} + \|\Pi_{-[\tilde{p}]}\|_{2,1} \right], \text{ for some } A > 1.$$

Then, for $T \gtrsim \log d^2\tilde{p}$, *there exist universal constants* $c_i > 0$ *such that with probability at least* $1 - c_0 \exp\left[-\left(c_1 A^2 - 2 \right) \log d^2\tilde{p} \right]$,

$$\frac{1}{T} \sum_{t = 1}^{T} \|\hat{\varepsilon}_t - \varepsilon_t\|^2 \le \Delta_\varepsilon^2 := 2\lambda \sum_{\tau = 1}^{\tilde{p}} \|\Pi_\tau\|_1,$$

$$\max_{1 \le j \le d} \frac{1}{T} \sum_{t = 1}^{T} (\hat{\varepsilon}_{tj} - a_{tj})^2 \le \Delta_a^2 := 4 \max\left\{ \Delta_\varepsilon^2, 4\pi \|\Pi_{-[\tilde{p}]}\|_{2,1}^2 |||f_y||| \right\},$$

$$\frac{1}{T} \sum_{t = 1}^{T} \|\hat{\varepsilon}_t - a_t\|^2 \le 4 \max\left\{ \Delta_\varepsilon^2, 4\pi d \|\Pi_{-[\tilde{p}]}\|_{2,1}^2 |||f_y||| \right\}.$$

If, in addition, $\{\Pi_1, \ldots, \Pi_{\tilde{p}}\}$ *are sparse so that* $k := \sum_{\tau = 1}^{\tilde{p}} \|\Pi_\tau\|_0 \lesssim T$, *then for any choice of* $\lambda \ge 2\lambda_0$ *and* $T \gtrsim \max\{\tilde{p}^2 |||f_y|||^2 |||f_y^{-1}|||^2, 1\} k (\log d + \log \tilde{p})$, *we can use a potentially tighter upper bound* $\Delta_\varepsilon^2 := (128/\pi) |||f_y^{-1}||| k \lambda^2$.

Remark 4.3 (Convergence Rate & Truncation Bias).—The error bounds Δ_ε^2 and Δ_a^2 scale with λ_0, which has two terms. The first term decays polynomially with T. The second term $\|\Pi_{-[\tilde{p}]}\|_{2,1}$ captures the *truncation bias* arising from using a VAR($\tilde{p}$) approximation to a VAR(∞) process. When $\tilde{p} \asymp T^{\frac{1}{2}-\epsilon}$, this term decays exponentially with $T^{\frac{1}{2}-\epsilon}$ since

$$\|\Pi_{-[\tilde{p}]}\|_{2,1} \le \frac{\bar{C}}{1-\bar{\rho}}\bar{\rho}^{\tilde{p}} = \frac{\bar{C}}{1-\bar{\rho}}\exp\left[-T^{\frac{1}{2}-\epsilon}\log(1/\bar{\rho})\right], \tag{12}$$

where $\bar{C}$, $\bar{\rho}$ are as defined in Remark 4.1. This bias also appears in our Phase-II analysis.

Remark 4.4 (Choice of [INEQ-START).—$\tilde{p}$, Slow & Fast Rates, and RE Condition] As long as $\tilde{p}$ increases polynomially fast with T, the truncation bias vanishes as $T \to \infty$ and the approximation errors Δ_ε and Δ_a decay with T at a rate $O(\sqrt{\log d/T})$. However, under sparsity of Π and choosing $\tilde{p} \asymp T^{1/2-\epsilon}$, a suitable Restricted Eigenvalue (RE) condition holds with high probability (see Appendix D for details), and these approximation errors decay at a faster rate $O(\log d / T)$. The choice of $(1/2 - \epsilon)$ in the exponent ensures that $T \gtrsim \tilde{p}^2$ holds asymptotically. This choice of $\tilde{p}$ matches with low-dimensional VARMA analysis presented in [20].

4.3 Prediction and Estimation Error in Phase-II

For simplicity of exposition, we assume that p and q are known and $\tilde{p} > p + q$. It will be evident from our analysis that similar conclusions hold as long as we replace these with any upper bounds of p and q. Without loss of generality, we also assume that the Phase-II regressions are run with the following re-indexing of observations:

$$y_t = \sum_{\ell=1}^{p} \Phi_\ell y_{t-\ell} + \sum_{m=0}^{q} \Theta_m \hat{\varepsilon}_{t-m} + u_t, \quad \text{for } t = 1, 2, \ldots, n, \ n = T - q, \tag{13}$$

where $u_t = \Theta(L)(a_t - \hat{\varepsilon}_t)$, and $\Theta_0 = I$. As mentioned earlier, we consider a variant of the Phase-II regression where the fitted values from Phase-I, $\hat{y}_t = y_t - \hat{\varepsilon}_t$, are used as response instead of y_t. The autoregressive moving average design then takes the form

$$\underbrace{\begin{bmatrix} \hat{y}_n^\top \\ \hat{y}_{n-1}^\top \\ \vdots \\ \hat{y}_1^\top \end{bmatrix}}_{\mathscr{Y}_{n\times d}} = \underbrace{\begin{bmatrix} y_{n-1}^\top & \cdots & y_{n-p}^\top & \hat{\varepsilon}_{n-1}^\top & \cdots & \hat{\varepsilon}_{n-q}^\top \\ y_{n-2}^\top & \cdots & y_{n-1-p}^\top & \hat{\varepsilon}_{n-2}^\top & \cdots & \hat{\varepsilon}_{n-1-q}^\top \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_0^\top & \cdots & y_{1-p}^\top & \hat{\varepsilon}_0^\top & \cdots & \hat{\varepsilon}_{1-q}^\top \end{bmatrix}}_{\mathscr{X}_{n\times d(p+q)}} \underbrace{\begin{bmatrix} \Phi^\top \\ \Theta^\top \end{bmatrix}}_{B_{d(p+q)\times d}} + \underbrace{\begin{bmatrix} u_n^\top \\ \vdots \\ u_1^\top \end{bmatrix}}_{\mathscr{U}_{n\times d}},$$

where $\Phi = [\Phi_1 : \ldots : \Phi_p]$, and $\Theta = [\Theta_1 : \ldots : \Theta_q]$. Vectorizing the above regression problem with n samples and $d^2(p+q)$ parameters, we have

$$\underbrace{\mathrm{vec}(\mathscr{Y})}_{Y} = \underbrace{(I \otimes \mathscr{X})}_{Z}\underbrace{\mathrm{vec}(B)}_{\beta^*} + \underbrace{\mathrm{vec}(\mathscr{U})}_{U}. \tag{14}$$

In order to apply Proposition 4.1 on this regression problem with $N = nd$ and $D = d^2(p+q)$, we first provide suitable choices of q_n, s_n and r_n (same as choice of λ) that hold with high probability for a random realization of $(T + \tilde{p})$ consecutive observations from the VARMA process. To this end, note that

$$\left\|(I \otimes \mathscr{Z})^{\top}\left(I \otimes \mathscr{Z}\right)/n - I \otimes \Sigma_z\right\|_{\infty} = \left\|\mathscr{Z}^{\top}\mathscr{Z}/n - \Sigma_z\right\|_{\infty}.$$

In Section 4.2, we have discussed how the approximation errors Δ_a, Δ_ε and the truncation bias term $\|\Pi_{-[\tilde{p}]}\|_{2,1}$ decay with the sample size. In this proposition, we show that q_n, r_n and s_n / d can be chosen to be a linear combination of the above terms and $\sqrt{\log d^2(p+q)/n}$, where the coefficients of this linear combination depend on model parameters and capture the role of temporal dependence in these convergence rates.

Proposition 4.3.—*Consider the Phase-II regression (14) with design matrix $I \otimes \mathscr{Z}$ and error vector* $\mathrm{vec}(\mathscr{U})$. *Set $\sigma_j^2 = e_j^{\top}\mathrm{Var}(\Theta(L)\Pi_{-[\tilde{p}]}(L)y_t)e_j$, for $j = 1,\ldots,d$. Then there exist universal constants $c_i > 0$ such that the event*

$$\mathscr{E} := \left\{\left\|\mathscr{Z}^{\top}\mathscr{Z}/n - \Sigma_z\right\|_{\infty} \le q_n, \frac{1}{n}\left\|\mathscr{Z}^{\top}\mathscr{U}\right\|_{\infty} \le r_n, \left|\frac{1}{n}\|\mathrm{vec}(\mathscr{U})\|^2 - \sum_{j=1}^{d}\sigma_j^2\right| \le s_n\right\} \qquad (15)$$

holds with probability at least $1 - c_0 \exp(-(c_1 A^2 - 2)\log d^2(p+q)]$, *where*

$$q_n = \varphi_{q,1}\sqrt{\frac{\log d^2(p+q)}{n}} + \varphi_{q,2}\left(\Delta_a + \Delta_a^2\right),$$

$$r_n = \varphi_{r,1}\sqrt{\frac{\log d^2(p+q)}{n}} + \varphi_{r,2}\left(\Delta_\varepsilon + \Delta_\varepsilon^2 + \|\Pi_{-[\tilde{p}]}\|_{2,1}\right),$$

$$s_n/d = \varphi_{s,1}\sqrt{\frac{\log d^2(p+q)}{n}} + \varphi_{s,2}\left(\Delta_\varepsilon + \Delta_\varepsilon^2\right),$$

and $\varphi_{\mathrm{q},1}$, $\varphi_{\mathrm{q},2}$, $\varphi_{\mathrm{r},1}$, $\varphi_{\mathrm{r},2}$, $\varphi_{\mathrm{s},1}$, $\varphi_{\mathrm{s},2}$, *are functions of the model parameters*

$$\varphi_{q,1} = 2\pi|||f_y|||\left(p + q|||\Pi_{[\hat{p}]}|||^2\right)^2,$$

$$\varphi_{q,2} = \max\left\{2q, 2\sqrt{2\pi q}|||f_y|||^{1/2}\left(p + q|||\Pi_{[\hat{p}]}|||^2\right)^{1/2}\right\},$$

$$\varphi_{s,1} = 2\pi|||\Theta|||\,\|\Pi_{-[\hat{p}]}\|_{2,1}^2|||f_y|||,$$

$$\varphi_{s,2} = \max\left\{2\|\Theta\|_{2,1}^2, 4\sqrt{2\pi}|||\Theta|||^{1/2}\|\Pi_{-[\hat{p}]}\|_{2,1}|||f_y|||^{1/2}\left\|\Theta\right\|_{2,1}\right\},$$

$$\varphi_{r,1} = c_1|||f_y|||A\max\left\{1, |||\Theta|||^2\|\Pi_{-[\hat{p}]}\|_{2,1}^2, |||\Pi_{[\hat{p}]}|||^2\right\},$$

$$\varphi_{r,2} = c_2|||f_y|||\,\|\Theta\|_{2,1}\max\{1, \|\Pi_{[\tilde{p}]}\|_{2,1}\}.$$

Using Proposition 2.1, the identification target in (4) with an elastic net penalty becomes

$$(\Phi^{(\alpha)}, \Theta^{(\alpha)}) = \underset{\Phi,\Theta}{\mathrm{argmin}}\{\|[\Phi:\Theta]\|_1 + \frac{\alpha}{2}\|[\Phi:\Theta]\|_F^2 \text{ s.t. } \mathrm{vec}\left(\rho_{zy}\right) = \left(I \otimes \Sigma_z\right)\mathrm{vec}(\beta)\}, \qquad (16)$$

where ρ_{zy}, Σ_z and β are as defined in Proposition 2.1. We consider the penalized and constrained versions of the estimator

$$\operatorname{vec}\left([\widehat{\Phi}^{(\alpha)}:\widehat{\Theta}^{(\alpha)}]^{\top}\right) = \underset{\|\beta\|_1 \le M}{\operatorname{argmin}} \frac{1}{n}\left\|\operatorname{vec}(\mathscr{Y}) - (I \otimes \mathscr{Z})\beta\right\|^2 + \lambda\mathscr{P}_{\alpha}(\beta)$$

$$\operatorname{vec}\left([\widehat{\Phi}_{[C]}^{(\alpha)}:\widehat{\Theta}_{[C]}^{(\alpha)}]^{\top}\right) = \underset{\|\beta\|_1 \le M}{\operatorname{argmin}} \left\{\mathscr{P}_{\alpha}(\beta) \text{ s.t. } \frac{1}{n}\left\|\operatorname{vec}(\mathscr{Y}) - (I \otimes \mathscr{Z})\beta\right\|^2 \le A_n\right\}.$$

A direct application of Proposition 4.1 with the choices of $q_n r_{n,}$ s_n in Proposition 4.3 then leads to the following upper bounds on the prediction and estimation error of the penalized and constrained versions of our two-phase VARMA estimator.

Proposition 4.4 (VARMA Estimation and Prediction Errors).—*Consider a random realization of* $T + \tilde{p}$ *consecutive observations* $\{y_1, \ldots, y_{T+\tilde{p}}\}$ *from a stable, invertible Gaussian VARMA model (1), and let* $n = T - q$ *denote the sample size in Phase-II. Denote* $K_y: = \max\{|||f_y|||, \left\|\Pi\right\|_{2,1}, \|\Theta^{(\alpha)}\|_{2,1}\}$.

<u>(a) Forecast Error:</u> *Let* $y_t^* = \sum_{\ell=1}^{p} \Phi_{\ell} y_{t-\ell} + \sum_{m=1}^{q} \Theta_m a_{t-m}$ *and* $\tilde{y}_t = \sum_{\ell=1}^{p} \widehat{\Phi}_{\ell} y_{t-\ell} + \sum_{m=1}^{q} \widehat{\Theta}_m \widehat{\varepsilon}_{t-m}$ *denote the optimal and the penalized VARMA forecasts respectively. Then, for a choice of* $\lambda \asymp K_y^3 \max\left\{\sqrt{\log d^2(p+q)/n}, \Delta_{\varepsilon}\right\}$, *and* $M \ge \|\Phi^{(\alpha)}\|_1 + \|\Theta^{(\alpha)}\|_1$ *for some* $\alpha \ge 0$,

$$\frac{1}{n}\sum_{t=1}^{n} \|\tilde{y}_t - y_t^*\|^2 = O_{\mathbb{P}}\left(K_y^3 M^2 \max\left\{\sqrt{\frac{\log d^2(p+q)}{n}}, \|\Pi_{-[\tilde{p}]}\|_{2,1}, \Delta_{\varepsilon}\right\}\right).$$

<u>(b) Partially-identified Estimation:</u> With the same choice of λ, M and α in (a), the penalized estimator is partially identified and satisfies

$$\min_{(\Phi,\Theta) \in \varepsilon_{p,q}(\Pi(L))} \left\|\left(\widehat{\Phi}^{(\alpha)}, \widehat{\Theta}^{(\alpha)}\right) - (\Phi, \Theta)\right\|_F^2 = O_{\mathbb{P}}\left(\frac{K_y^3 M^2}{\Lambda_{\min^+}(\Gamma_z(0))} \max\left\{\sqrt{\frac{\log d^2(p+q)}{n}}, \|\Pi_{-[\tilde{p}]}\|_{2,1}, \Delta_{\varepsilon}\right\}\right).$$

<u>(c) Point-identified Estimation:</u> *For a choice of* $A_n \asymp K_y^3 \|\Pi_{-[\tilde{p}]}\|_{2,1}^2 \max\{d\sqrt{\log d^2(p+q)/n}, \Delta_{\varepsilon}\}$ *and any* $\alpha > 0$, *the constrained version of the estimator is point* identified and satisfies

$$\left\|\left(\widehat{\Phi}_{[C]}^{(\alpha)}, \widehat{\Theta}_{[C]}^{(\alpha)}\right) - \left(\Phi^{(\alpha)}, \Theta^{(\alpha)}\right)\right\|_F^2 = O_{\mathbb{P}}\left(\frac{K_y^3 M^2}{\alpha\sqrt{\Lambda_{\min^+}(\Gamma_z(0))}} \max\left\{d^3\sqrt{\frac{\log d^2(p+q)}{n}}, \|\Pi_{-[\tilde{p}]}\|_{2,1}, \Delta_{\varepsilon}\right\}^{1/2}\right).$$

Part (a) of this proposition ensures that as long as the identification target is parsimonious in the sense of small ℓ_1-norm and the penalty parameter is chosen appropriately, the VARMA forecasts converge to the optimal forecasts (which uses any element from the equivalence class $\mathscr{E}_{p,q}(\Pi)$) in the asymptotic regime $\log d / n \rightarrow 0$. The truncation bias term $\|\Pi_{-[\tilde{p}]}\|_{2,1}$ and the approximation error from Phase-I Δ_{ε} also converges to zero in this asymptotic regime, as shown in Section 4.2. The convergence rates are further affected by the strength of temporal dependence in the VARMA process, as captured by the term K_y.

In addition, part (b) ensures that the distance of our penalized estimator from the equivalence class also asymptotically vanishes in this high-dimensional regime. Further, the convergence rates are affected by the minimum positive eigenvalue of the variance-covariance matrix of the process z_t, which captures the curvature of the loss function.

Part (c) shows that our constrained estimator converges in probability to our identification target, but in a low-dimensional regime $d^3\sqrt{\log d}/n \to 0$. This slow rate is a consequence of the fact that we did not assume sparsity on the entire equivalence class $\mathscr{E}_{p,q}(\Pi)$, so searching for the correct identification target within this equivalence class still has a complexity of the order of d^2. The tuning parameter α also affects the convergence rate, since this captures the degree of curvature of the term $\mathscr{P}_\alpha(\,.\,)$ in the loss function. However, taking a sequence of α_n that converges to 0 at a rate slower than $d^3\sqrt{\log d^2(p+q)/n}$, we can still guarantee consistent estimation of the target $(\Phi^{(0)}, \Theta^{(0)})$ with the minimum Frobenius norm.

5 Forecast Applications

We present three forecast applications:

(i) Demand forecasting.

Weekly sales data (in dollars) are collected for $d = 16$ product categories of Dominick's Finer Foods from January 1993 to July 1994 ($T = 76$). Data are taken from https://research.chicagobooth.edu/kilts/marketing-databases/dominicks. To ensure stationarity, we take each series in log differences and consider sales growth. Augmented Dickey-Fuller tests help support that the sales growth series are stationary.

(ii) Volatility forecasting.

We collect monthly realized variances for $d = 17$ stock market indices, from January 2009 to December 2016 ($T = 96$). Realized variances, computed from five minute returns, are obtained from http://realized.oxford-man.ox.ac.uk/data/download and log-transformed following standard practice. Augmented Dickey-Fuller tests help support that the log-realized variances are stationary.

(iii) Macro-economic forecasting.

We consider $d = 168$ quarterly macro-economic series of length $T = 60$ ending in 2008, Quarter 4. Data are taken from the Journal of Applied Econometrics Data Archive, a full list of the series is available in [31] (Data Appendix), along with the transformations to make them approximately stationary.

In all considered cases, the number of time series d is large relative to the time series length T. First, we discuss the model parsimony of the estimated VARMA and VAR with HLag penalties. Secondly, we compare their forecast accuracy for different forecast horizons.

5.1 Model Parsimony

Since the sparse VARMA and VAR estimators with HLag penalties both perform automatic lag selection, they give information on the effective maximum AR and MA orders. Consider the $d \times d$ moving average lag matrix $\widehat{L}_{\Theta}$ of the estimated VARMA whose elements are $\widehat{L}_{\Theta,ij} = \max\{m{:}\Theta_{m,ij} \neq 0\}$, where $\widehat{L}_{\Theta,ij} = 0$ if $\Theta_{m,ij} = 0$ for all $m = 1\ldots,\widehat{q}$. This lag matrix shows the maximal MA lag for each series j in each equation i of the corresponding estimated VARMA. If entry ij is zero, this means that all lagged MA coefficients of time series j on time series i are estimated as zero. If entry ij is, for instance, three, this means that the third lagged moving average term of series j on series i is estimated as non-zero, but the forth and higher as all zero. Similarly, one can construct the autoregressive lag matrix $\widehat{L}_{\Phi}$ of the estimated VARMA and the autoregressive lag matrix $\widehat{L}_{\Pi}$ of the estimated VAR.

Figure 1 shows the lag matrices of the estimated VARMA and VAR on the demand data. Similar findings are obtained for the other data sets and therefore omitted. The MA lag matrix of the VARMA (middle panel) is very sparse: 247 out of 256 entries are equal to zero. By adding just few MA terms to the model, serial correlation in the error terms is captured. As a result, a more parsimonious VARMA model is obtained: 107 out of the 3,072 (around 3%) estimated VARMA parameters are non-zero. In contrast, 877 out of the 3,328 (around 25%) estimated VAR parameters are non-zero. We find the more parsimonious VARMA to often give more accurate forecasts than the VAR, as discussed next.

5.2 Forecast Accuracy

We compare the forecast accuracy of VARMA to VAR through an expanding window forecast exercise. Let h be the forecast horizon. At each time point $t = S,\ldots,T - h$, we sparsely estimate the VARMA and VAR. We take S such that forecasts are computed for the last 25% of observations. We estimate the model on the standardized series and obtain h-step-ahead forecasts and corresponding forecast errors $e_{i,t+h}^{(i)} = y_{i,t+h} - \widehat{y}_{i,t+h}$ for each series $1 \leq i \leq d$. The overall forecast performance is measured by computing the Mean Squared Forecast Error for a particular forecast horizon h, as in Equation (10). For the weekly marketing data set, we take $h = 1, 8, 13$. For the monthly volatility data set, we take $h = 1, 6, 12$. For the quarterly macro-economic data set, we take $h = 1, 4, 8$. To assess the difference in forecast performance between VARMA and VAR, we use a Diebold-Mariano (DM-) test ([17]).

The MSFEs on the three data sets are given in Table 1. Across all considered data sets and horizons, VARMA gives either a significantly lower MSFE than the VAR estimator (in 5 out of 9 cases at the 5% level, in 1 case at the 10% level) or performs equally well (in 3 out of 9 cases). The gain in forecast accuracy over VAR is typically the largest for the longest forecast horizons. VARMA not only gives a lower MSFE averaged over the considered time points, but it also attains the lowest MSFE for the large majority of time points. For the demand data at horizon $h = 13$, for instance, it outperforms VAR for all time points except two. The sparse VARMA method is thus a valuable addition to the forecaster’s toolbox for large-scale multivariate time series models. It exploits the serial correlation between the error terms and, as a consequence, often gives more parsimonious forecast models with competitive or better forecast accuracy than a sparse VAR.

6 Conclusion

We present sparse identification and estimation for VARMA models. Our estimator, available in the R package bigtime, is naturally aligned with our identified target through the use of sparsity-inducing convex regularizers and can be computed efficiently even for large-scale VARMAs. Under a double-asymptotic regime where both $d, T \to \infty$, we prove consistency of our two-step sparse VARMA estimation for stable, invertible Gaussian VARMA processes. Simulation and real data analyses show that our sparse VARMA model can produce better forecasts compared to sparse VAR by fitting more parsimonious models.

There are several questions we did not address. Our two-stage procedure can be generalized to an iterative method, as in [16]. However, developing a double-asymptotic theory for such an iterative method is complex and left for future research. The convergence rates of our point-identified Phase-II estimator can be potentially sharpened under restricted eigenvalue assumptions. Identifying a class of sparse VARMAs for which such assumptions hold with high probability is an interesting theoretical question. Inference of model parameters can be pursued by adopting debiasing approaches [27, 47], and are left for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the editor and reviewers for their thorough review and highly appreciate their comments which substantially improved the quality of the manuscript. The authors wish to thank Profs. Christophe Croux, George Michailidis, Suhasini Subba Rao and Ruey S. Tsay for stimulating discussions and helpful comments. IW was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 832671. SB was supported by NSF award DMS-1812128 and NIH awards 1R01GM135926–01 and 1R21NS120227–01. JB was supported by an NSF CAREER award (DMS-1748166). DSM was supported by NSF (1455172, 1934985, 1940124, 1940276), Xerox PARC, the Cornell University Atkinson Center for a Sustainable Future (AVF-2017), USAID, and the Cornell University Institute of Biotechnology & NYSTAR.

References

[1]. Agarwal A; Negahban S and Wainwright MJ (2010), "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in Advances in Neural Information Processing Systems, pp. 37–45.

[2]. Akaike H (1974), "A new look at the statistical model identification," IEEE transactions on automatic control, 19, 716–723.

[3]. — (1976), "Canonical correlation analysis of time series and the use of an information criterion," in Mathematics in Science and Engineering, Elsevier, vol. 126, pp. 27–96.

[4]. Anthanasopoulos G and Vahid F (2008), "VARMA versus VAR macroeconomic forecasting," Journal of Business & Economic Statistics, 26, 237–252.

[5]. Athanasopoulos G; Poskitt DS and Vahid F (2012), "Two canonical VARMA forms: Scalar component models vis-a-vis the echelon form," Econometric Reviews, 31, 1.

[6]. Athanasopoulos G and Vahid F (2008), "A complete VARMA modelling methodology based on scalar components," Journal of Time Series Analysis, 29, 533–554.

[7]. Bai J and Ng S (2008), "Large dimensional factor analysis," Foundations and Trends[textregistered] in Econometrics, 3, 89–163.

[8]. Banbura M; Giannone D and Reichlin L (2010), “Large Bayesian vector auto regressions,” Journal of Applied Econometrics, 25(1), 71–92.

[9]. Basu S; Li X and Michailidis G (2019), “Low rank and structured modeling of high-dimensional vector autoregressions,” IEEE Transactions on Signal Processing, 67, 1207–1222.

[10]. Basu S and Michailidis G (2015), “Regularized estimation in sparse high-dimensional time series models,” The Annals of Statistics, 43(4), 1535–1567.

[11]. Brockwell PJ and Davis RA (1991), Time series: Theory and methods, Springer Series in Statistics.

[12]. Chan JCC; Eisenstat E and Koop G (2016), “Large Bayesian VARMAs,” Journal of Econometrics, 192(2), 374–390.

[13]. Davis R; Zang P and Zheng T (2016), “Sparse vector autoregressive modeling,” Journal of Computational and Graphical Statistics, 25(4), 1077–1096.

[14]. De Mol C; Giannone D and Reichlin L (2008), “Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?” Journal of Econometrics, 146, 318–328.

[15]. Deistler M (1985), “General structure and parametrization of ARMA and state-space systems and its relation to statistical problems,” Handbook of statistics, 5, 257–277.

[16]. Dias GF and Kapetanios G (2018), “Estimation and forecasting in vector autoregressive moving average models for rich datasets,” Journal of Econometrics, 202, 72–91.

[17]. Diebold F and Mariano R (1995), “Comparing predictive accuracy,” Journal of Business and Economic Statistics, 13, 253–263.

[18]. Diebold FX and Y ilmaz K (2014), “On the network topology of variance decompositions: Measuring the connectedness of financial firms,” Journal of Econometrics, 182, 119–134.

[19]. Dufour J and Jouini T (2014), “Asymptotic distributions for quasi-efficient estimators in echelon VARMA models,” Computational Statistics & Data Analysis, 73, 69–86.

[20]. Dufour J-M and Jouini T (2005), “Asymptotic distribution of a simple linear estimator for VARMA models in echelon form,” in Statistical modeling and analysis for complex data problems, Springer, pp. 209–240.

[21]. Fernández-Villaverde J; Rubio-Ramírez JF; Sargent TJ and Watson MW (2007), “ABCs (and Ds) of understanding VARs,” American Economic Review, 97, 1021–1026.

[22]. Friedman J; Hastie T and Tibshirani R (2010), “Regularization paths for generalized linear models via coordinate descent,” Journal of Statistical Software, 33(1), 1–22. [PubMed: 20808728]

[23]. Gelper S; Wilms I and Croux C (2016), “Identifying demand effects in a large network of product categories,” Journal of Retailing, 92(1), 25–39.

[24]. Hannan EJ (1976), “The Identification and Parameterization of ARMAX and State Space Forms,” Econometrica, 44, 713–723.

[25]. Hannan EJ and Kavalieris L (1984), “Multivariate linear time series models,” Advances in Applied Probability, 16, 492–561.

[26]. Hastie T; Tibshirani R and Friedman J (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.

[27]. Javanmard A and Montanari A (2014), “Confidence intervals and hypothesis testing for high-dimensional regression,” Journal of Machine Learning Research, 15, 2869–2909.

[28]. Kascha C (2012), “A comparison of estimation methods for vector Autoregressive Moving-Average Models,” Econometric Reviews, 31, 297–324.

[29]. Kilian L and Lütkepohl H (2017), Structural vector autoregressive analysis, Cambridge University Press, chap. 16: Structural VAR Analysis in a Data-Rich Environment.

[30]. Kock AB and Callot (2015), “Oracle inequalities for high dimensional vector autoregressions,” Journal of Econometrics, 186, 325–344.

[31]. Koop GM (2013), “Forecasting with medium and large Bayesian VARs,” Journal of Applied Econometrics, 28(2), 177–203.

[32]. Loh P-L and Wainwright MJ (2012), “High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity.” The Annals of Statistics, 40, 1637–1664.

[33]. Lütkepohl H (2005), New introduction to multiple time series analysis, Springer-Verlag: Berlin-Germany.

[34]. — (2006), “Forecasting with VARMA models,” Handbook of economic forecasting, 1, 287–325.

[35]. Manski CF (2010), “Partial identification in econometrics,” in Microeconometrics, Springer, pp. 178–188.

[36]. Matteson DS and Tsay RS (2011), “Dynamic orthogonal components for multivariate time series,” Journal of the American Statistical Association, 106(496), 1450–1463.

[37]. Nicholson W; Matteson DS and Bien J (2017), “VARX-L: Structured regularization for large vector autoregressions with exogenous variables,” International Journal of Forecasting, 33(3), 627–651.

[38]. Nicholson WB; Wilms I; Bien J and Matteson DS (2020), “High dimensional forecasting via interpretable vector autoregression,” Journal of Machine Learning Research, 21, 1–52. [PubMed: 34305477]

[39]. Poskitt DS (1992), “Identification of Echelon canonical forms for vector linear processes using least squares,” The Annals of Statistics, 20, 195–215.

[40]. — (2016), “Vector autoregressive moving average identification for macroeconomic modeling: A new methodology,” Journal of Econometrics, 192, 468–484.

[41]. Spliid H (1983), “A fast estimation method for the vector autoregressive moving average model with exogenous variables,” Journal of the American Statistical Association, 78(384), 843–849.

[42]. Sun Y; Li Y; Kuceyeski A and Basu S (2018), “Large spectral density matrix estimation by thresholding,” arXiv preprint arXiv:1812.00532.

[43]. Tamer E (2010), “Partial identification in econometrics,” Annu. Rev. Econ, 2, 167–195.

[44]. Tiao GC and Tsay RS (1989), “Model specification in multivariate time series,” Journal of the Royal Statistical Society Series B, 51, 157–213.

[45]. Tibshirani RJ (2013), “The lasso problem and uniqueness,” Electronic Journal of statistics, 7, 1456–1490.

[46]. Tsay RS (2014), Multivariate Time Series Analysis: With R and Financial Applications, Wiley.

[47]. van de Geer S; Bühlmann P; Ritov Y and Dezeure R (2014), “On asymptotically optimal confidence regions and tests for high-dimensional models, ” The Annals of Statistics, 42, 1166–1202.

[48]. Wallis KF (1977), “Multiple time series analysis and the final form of econometric models,” Econometrica, 45, 1481–1497.

[49]. Wilms I; Basu S; Bien J and Matteson DS (2017), bigtime: Sparse Estimation of Large Time Series Models, R package version 0.1.0. https://CRAN.R-project.org/package=bigtime.

[50]. — (2017), “Interpretable vector autoregressions with exogenous time series,” NIPS 2017 Symposium on Interpretable Machine Learning, arXiv:1711.03623.

[51]. Wu W-B and Wu YN (2016), “Performance bounds for parameter estimates of high-dimensional linear models with correlated errors,” Electronic Journal of Statistics, 10, 352–379.

[52]. Yan X and Bien J (2017), “Hierarchical sparse modeling: A choice of two group lasso formulations,” Statistical Science, 32, 531–560.

[53]. Zellner A and Palm F (1974), “Time series analysis and simultaneous equation econometric model,” Journal of Econometrics, 2, 17–54.

[54]. Zhao P and Yu B (2006), “On model selection consistency of Lasso,” Journal of Machine Learning Research, 7, 2541–2563.

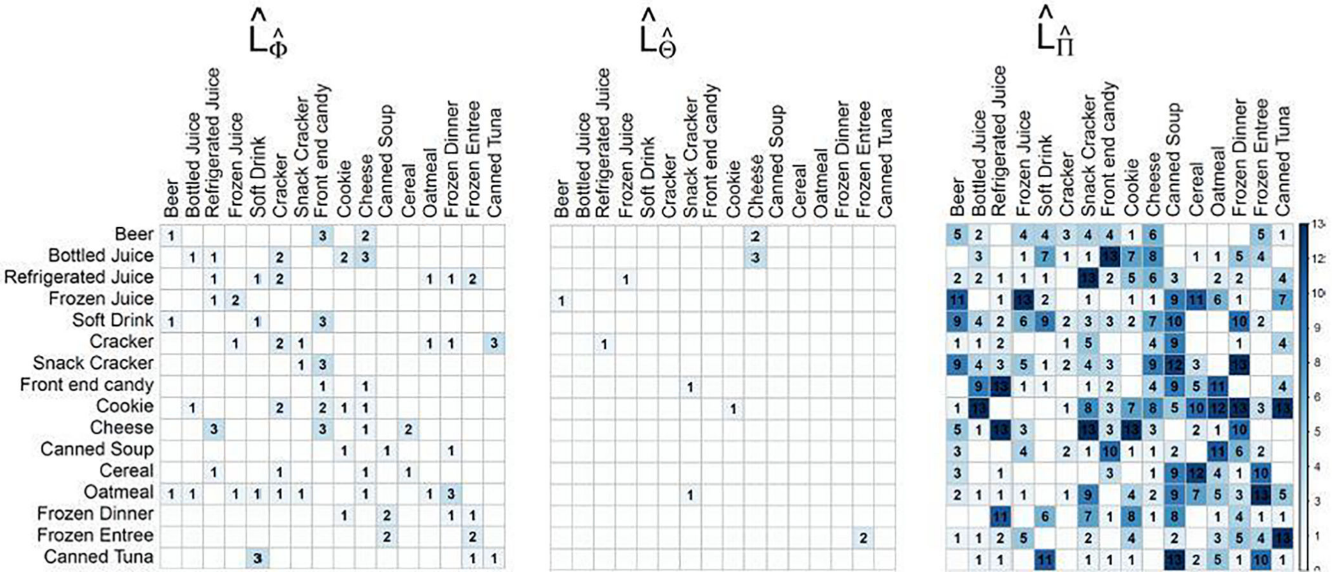

Fig. 1.

Demand data set: AR-lag matrix (left) and MA-lag matrix (middle) of the estimated VARMA, and AR-lag matrix of the estimated VAR (right).

Table 1

Mean Squared Forecast Errors at different forecast horizons for the two estimators on the three data sets. *P*-values of the Diebold-Mariano tests are given in parentheses.

Estimator	**Weekly**			**Monthly**			**Quaterly**		
	Demand Data			**Volatility Data**			**Macro-economic Data**		
	$h = 1$	$h = 8$	$h = 13$	$h = 1$	$h = 6$	$h = 12$	$h = 1$	$h = 4$	$h = 8$
VARMA	0.473	0.578	0.550	0.781	1.080	1.065	0.974	1.152	1.281
VAR	0.499 (0.141)	0.703 (0.041)	0.715 (<0.001)	0.728 (0.142)	1.209 (0.050)	1.429 (0.007)	0.977 (0.412)	1.170 (0.080)	1.401 (0.003)