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

Ines Wilms[a]*, Sumanta Basu[b]*, Jacob Bien[c]†, and 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

*Equal Contribution.

†Corresponding author. E-mail and URLs: jbien@usc.edu, http://faculty.marshall.usc.edu/Jacob-Bien/ (J. Bien), i.wilms@maastrichtuniversity.nl, https://sites.google.com/view/iwilms (I. Wilms), sumbose@cornell.edu, http://faculty.bscb.cornell.edu/~basu/ (S. Basu), matteson@cornell.edu, http://www.stat.cornell.edu/~matteson/ (D.S. Matteson).

# 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, 16, 33, 20] to factor modeling of high-dimensional time series, e.g., [7]. In the classical time 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,

$$y_t = \sum_{\ell=1}^{p} \Phi_\ell y_{t-\ell} + \sum_{m=1}^{q} \Theta_m a_{t-m} + a_t, \tag{1.1}$$

where $\{\Phi_\ell \in \mathbb{R}^{d\times d}\}_{\ell=1}^{p}$ are autoregressive parameter matrices, $\{\Theta_m \in \mathbb{R}^{d\times d}\}_{m=1}^{q}$ 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 $\Sigma_a$. 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., [15, 42] for computational contributions; [10, 34, 56, 9] for theoretical contributions, and [41, 25] 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 [38]. In macroeconomics, VARMA is popular for its close link with linearized dynamic stochastic general equilibrium (DSGE) models [32, 23]. 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 [32, 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 (1.1) is not identifiable in general (see Section 2.1), i.e. there can be different combinations of AR and MA matrices $\{\Phi_\ell\}$ 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 $\mathcal{E}$ 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* [28, 45, 14], which amounts to selecting a basis for the row space of a block Hankel matrix (see Section 4 of [17]). Specifying an Echelon form involves selecting *Kronecker orders* (related to indices of rows that form the above basis) from a $O\left((p+q)^d\right)$-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, 44]. However, all of these methods are computationally intensive and lack a formal asymptotic theory that combines specification and estimation. Assuming $d$ is fixed, [44] 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, [39]).

Other popular identification and specification methods include scalar component models [49, 6, 5] and final equations form [58, 27, 53]. While these and other existing identification procedures [28, 45, 14] 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 $\mathcal{E}$ in an intuitive yet objective fashion (Section 2). More formally, we can define the class of AR-MA matrices with minimum $\ell_1$-norm as a partially identified class of "sparse" VARMA models $\mathcal{RE} = \text{argmin}_{(\Phi,\Theta)\in\mathcal{E}}\{\sum_{\ell=1}^{p}\|\Phi_\ell\|_1 + \sum_{m=1}^{q}\|\Theta_m\|_1\}$. We could also use a modified, strongly convex penalty $\text{argmin}_{(\Phi,\Theta)\in\mathcal{E}}\{\sum_{\ell=1}^{p}(\|\Phi_\ell\|_1+\alpha\|\Phi_\ell\|_F^2) + \sum_{m=1}^{q}(\|\Theta_m\|_1+\alpha\|\Theta_m\|_F^2)\}$ with a very small $\alpha \approx 0$ to identify a parsimonious element in $\mathcal{RE}$, 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` [54].

**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 [18, 21]. While

to the best of our knowledge, consistency of VARMA estimators has been studied only in the low-dimensional, fixed $d$ asymptotic regime [32, 22], 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* [40, 48] states that under suitable sparsity assumptions our algorithm provides a parsimonious VARMA estimator (small $\ell_1$-norm) whose distance from the equivalence class $\mathcal{E}$ 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 $\mathcal{E}$ 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 (1.1) 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

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

with the lag operator $L^\ell$ defined as $L^\ell y_t = y_{t-\ell}$. We assume the model is stable and invertible meaning respectively that $\det\{\Phi(z)\} \neq 0$ and $\det\{\Theta(z)\} \neq 0$ 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 $\det\{\Pi(z)\} \neq 0$ for all $|z| \leq 1$. The $\Pi$-matrices can be computed recursively from the AR matrices $\{\Phi_\ell\}$ and MA matrices $\{\Theta_m\}$ (e.g., [12], 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,

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

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 $\mathcal{E}_{p,q}(\Pi(L))$ 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* $\Sigma_a$*. 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,* $(\Phi, \Theta) \in \mathcal{E}_{p,q}(\Pi(L))$ *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* $\rho_{zy} = \mathbb{E}[z_t y_t^\top]$ *and* $\Sigma_z = \mathbb{E}[z_t z_t^\top]$*. That is,*

$$\mathcal{E}_{p,q}(\Pi(L)) = \{(\Phi, \Theta) : \rho_{zy} = \Sigma_z\beta\}\,. \tag{2.1}$$

A proof of this proposition is provided in Appendix A.1. Note that both $\rho_{zy}$ and $\Sigma_z$ can be expressed as functions of $\Pi$ and $\Sigma_a$ alone (i.e. they do not depend on $\Theta$ and $\Phi$), and hence are uniquely defined for the underlying process $y_t$. While the AR($\infty$) representation given by $\Pi$ in Proposition 2.1 is unique, it allows an equivalent characterization in terms of many ($\Phi$, $\Theta$) 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 $\rho_{zy}$ and $\Sigma_z$ 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, $\mathcal{P}_{\text{AR}}(\Phi)$ and $\mathcal{P}_{\text{MA}}(\Theta)$. Our identification results apply equally well to any convex function: one may consider, amongst others, the $\ell_1$-norm, the $\ell_2$-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

$$\mathcal{RE}_{p,q}(\Pi(L)) = \underset{\Phi,\Theta}{\operatorname{argmin}} \{\mathcal{P}_{\text{AR}}(\Phi) + \mathcal{P}_{\text{MA}}(\Theta) \text{ s.t. } \Phi(L) = \Theta(L)\Pi(L)\}. \tag{2.2}$$

This regularized equivalence class is a subclass of the equivalence class $\mathcal{E}_{p,q}(\Pi(L))$, containing the regularized VARMA representations. If the objective function in (2.2) 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 $\ell_1$-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

$$(\Phi^{(\alpha)}, \Theta^{(\alpha)}) = \underset{\Phi,\Theta}{\operatorname{argmin}}\{\mathcal{P}_{\text{AR}}(\Phi)+\mathcal{P}_{\text{MA}}(\Theta)+\frac{\alpha}{2}\|\Phi\|_F^2+\frac{\alpha}{2}\|\Theta\|_F^2 \text{ s.t. } \Phi(L) = \Theta(L)\Pi(L)\}. \tag{2.3}$$

Problem (2.3) 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

$$(\Phi^{(0)}, \Theta^{(0)}) = \lim_{\alpha \to 0^+} (\Phi^{(\alpha)}, \Theta^{(\alpha)}). \tag{2.4}$$

The following proposition, proved in Appendix A.2, establishes that $(\Phi^{(0)}, \Theta^{(0)})$ is in the regularized equivalence class $\mathcal{RE}_{p,q}(\Pi(L))$ 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 $\ell_2$-norm over all lasso solutions (see [50], Lemma 7).

**Proposition 2.2.** *The limit in (2.4) exists and is the unique pair in the set $\mathcal{RE}_{p,q}(\Pi(L))$ whose Frobenius norm squared is smallest:*

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

## 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 (2.2) 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 $\mathcal{P}_{\text{AR}}(\Phi) = \|\Phi\|_0$ and $\mathcal{P}_{\text{MA}}(\Theta) = \|\Theta\|_0$. However, since the $\ell_0$-penalty is non-convex, a unique solution cannot be guaranteed. One can construct examples in which there exist multiple equivalent, sparsest VARMAs, see [51] and Appendix A.3.1. Strong convexity in (2.3) is key to guaranteeing uniqueness of $(\Phi^{(\alpha)}, \Theta^{(\alpha)})$. For sparsity, we may therefore add to the $\ell_2$-norm in (2.3) the $\ell_1$-norm $\mathcal{P}_{\text{AR}}(\Phi) = \|\Phi\|_1$ and $\mathcal{P}_{\text{MA}}(\Theta) = \|\Theta\|_1$ as a sparsity-inducing convex heuristic.

While our theory will focus on the $\ell_1$-norm, in the empirical sections we also investigate a time-series specific alternative penalty, the hierarchical lag (hereafter "HLag") penalty [43, 55]: $\mathcal{P}_{\text{AR}}(\Phi) = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{\ell=1}^{p}\|\Phi_{(\ell:p),ij}\|$, and $\mathcal{P}_{\text{MA}}(\Theta) = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{m=1}^{q}\|\Theta_{(m:q),ij}\|$, 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., [57]) 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 $\Phi_\ell$ and $\Theta_\ell$ 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 VARMA$_d(p,q)$ of Equation (1.1) cannot be directly estimated as it contains the unobservable (latent) lagged errors, we proceed in two phases, in the spirit of [46, 21], 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 (1.1) has a pure VAR($\infty$) representation if it is invertible (Section 2.1). We therefore approximate the errors $a_t$ by the residuals of a VAR($\widetilde{p}$) given by

$$y_t = \sum_{\tau=1}^{\widetilde{p}} \Pi_\tau y_{t-\tau} + \varepsilon_t, \tag{3.1}$$

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

Estimating the VAR($\widetilde{p}$) of Equation (3.1) is challenging since $\widetilde{p}$ needs to be sufficiently large such that the residuals $\widehat{\varepsilon}_t$ accurately approximate the errors $a_t$. Since, a large number of parameters ($\widetilde{p}d^2$), relative to the time series length $T$, needs to be estimated, we use regularized estimation. For ease of notation, first rewrite model (3.1) in compact matrix notation $Y = \Pi Z + E$, where $Y = [y_{\widetilde{p}+1} \dots y_T] \in \mathbb{R}^{d\times(T-\widetilde{p})}$, $Z = [z_{\widetilde{p}+1} \dots z_T] \in \mathbb{R}^{d\widetilde{p}\times(T-\widetilde{p})}$, with $z_t = [y_{t-1}^\top \dots y_{t-\widetilde{p}}^\top]^\top \in \mathbb{R}^{(d\widetilde{p}\times 1)}$, $E = [\varepsilon_{\widetilde{p}+1} \dots \varepsilon_T] \in \mathbb{R}^{d\times(T-\widetilde{p})}$, and $\Pi = [\Pi_1 \dots \Pi_{\widetilde{p}}] \in \mathbb{R}^{d\times d\widetilde{p}}$. The regularized autoregressive estimates $\widehat{\Pi}$ are obtained as

$$\widehat{\Pi} = \underset{\Pi}{\operatorname{argmin}} \left\{ \frac{1}{2}\|Y - \Pi Z\|_F^2 + \lambda_\Pi \mathcal{P}(\Pi) \right\}, \tag{3.2}$$

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 ($\ell_1$-norm or HLag penalty). The penalty parameter $\lambda_\Pi > 0$ then regulates the degree of sparsity in $\widehat{\Pi}$: the larger $\lambda_\Pi$, the sparser $\widehat{\Pi}$. Problem (3.2) can be efficiently solved using Algorithm 1 in

[43].

## 3.2 Phase-II: Estimating the VARMA

We continue with the approximated lagged errors $\widehat{\varepsilon}_{t-1}, \ldots, \widehat{\varepsilon}_{t-q}$ instead of the true errors $a_{t-1}, \ldots, a_{t-q}$ in Equation (1.1). The resulting model

$$y_t = \sum_{\ell=1}^{p} \Phi_\ell y_{t-\ell} + \sum_{m=1}^{q} \Theta_m \widehat{\varepsilon}_{t-m} + u_t, \tag{3.3}$$

is a regression of $y_t$ on $y_{t-1}, \ldots, y_{t-p}, \widehat{\varepsilon}_{t-1}, \ldots, \widehat{\varepsilon}_{t-q}$ 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 (3.3) 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 $z_t = [y_{t-1}^\top \ldots y_{t-\hat{p}}^\top]^\top \in \mathbb{R}^{(d\hat{p} \times 1)}$, $X = [x_{\bar{o}+1} \ldots x_T] \in \mathbb{R}^{d\hat{q} \times (T-\bar{o})}$ with $x_t = [\widehat{\varepsilon}_{t-1}^\top \ldots \widehat{\varepsilon}_{t-\hat{q}}^\top]^\top \in \mathbb{R}^{(d\hat{q} \times 1)}$, 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:

$$(\widehat{\Phi}^{(\alpha)}, \widehat{\Theta}^{(\alpha)}) = \underset{\Phi,\Theta}{\operatorname{argmin}} \{ \frac{1}{2} \|Y - \Phi Z - \Theta X\|_F^2 + \lambda_\Phi \mathcal{P}_{\mathrm{AR}}(\Phi) + \lambda_\Theta \mathcal{P}_{\mathrm{MA}}(\Theta) + \frac{\alpha}{2} (\lambda_\Phi \|\Phi\|_F^2 + \lambda_\Theta \|\Theta\|_F^2) \}, \tag{3.4}$$

where $\lambda_\Phi, \lambda_\Theta > 0$ are two penalty parameters. By adding the regularizers $\mathcal{P}_{\mathrm{AR}}(\Phi)$ and $\mathcal{P}_{\mathrm{MA}}(\Theta)$ 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 (2.3). Optimization problem (3.4) 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 $\widetilde{p}, \widehat{p}, \widehat{q}$; (ii) the penalty parameters $\lambda_\Pi, \lambda_\Phi, \lambda_\Theta$; and (iii) the parameter $\alpha$ 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 $\widetilde{p}, \widehat{p}$, and $\widehat{q}$.** We take $\widetilde{p} = \lfloor 1.5\sqrt{T} \rfloor$ and $\widehat{p} = \widehat{q} = \lfloor 0.75\sqrt{T} \rfloor$. Our theoretical analysis suggests that $\widetilde{p} \asymp T^{\frac{1}{2}-\epsilon}$ (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 $\ell_1$-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 $\lambda_\Pi, \lambda_\Phi$ and $\lambda_\Theta$.** We select the penalty parameters using cross-validation. Below, we describe the selection of $\lambda_\Pi$ 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 [24], 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 $\widehat{y}_{t+h}^{(\lambda)}$ obtained with penalty parameter $\lambda$. We select the value of $\lambda_\Pi$ that gives the most regularized model whose Mean Squared Forecast Error

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

is within one standard error (see [29]; Chapter 7) of the minimal MSFE. In simulations, we take $h = 1$; in the forecast applications, we also consider other forecast horizons.

**The parameter $\alpha$.** We will sometimes refer to Equation (3.4) as an "elastic net" problem, although, unlike $\lambda_\Phi$ and $\lambda_\Theta$, the parameter $\alpha$ 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 $\alpha$ indeed produces sparse VARMA estimates close to the unique $(\Phi^{(0)}, \Theta^{(0)})$ pair defined in Equation (2.4). For $\alpha = 0$, 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 $\alpha$ since the target $(\Phi^{(0)}, \Theta^{(0)})$ corresponds to the pair with minimum Frobenius norm among all minimum-$\ell_1$ 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 $\Sigma_z$ is potentially singular. It is not clear whether a restricted eigenvalue (RE) assumption, commonly used in high-dimensional regression [37], 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 $\ell_1$-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 [37].

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 [47] 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, $\ell_1$ and $\ell_\infty$ 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 $\mathcal{A}(L) = \sum_{\ell \geq 0} A_\ell L^\ell$, we define $|||\mathcal{A}||| := \max_{\theta \in [-\pi, \pi]} \|\mathcal{A}(e^{i\theta})\|$, and use $\mathcal{A}_{[k]}(L)$ and $\mathcal{A}_{-[k]}(L)$ 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 $\|\mathcal{A}\|_{2,1}$ 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 $\Gamma_x(h) = \mathrm{Cov}(x_t, x_{t+h}) = \mathbb{E}[x_t x_{t+h}^\top]$, $h \in \mathbb{Z}$, we define the spectral density function $f_x(\theta) := \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\Gamma_x(\ell)e^{-i\ell\theta}$, $\theta \in [-\pi, \pi]$. 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 (1.1) with $\Lambda_{\min}(\Sigma_a) > 0$, it is known that $f_y$ is non-singular on $[-\pi, \pi]$ and there exist two model dependent quantities $\bar{C} > 0$ and $\bar{\rho} \in [0, 1)$ such that $\|\Pi_\tau\| \leq \bar{C}\,\bar{\rho}^\tau$, for all integers $\tau \geq 1$ [22]. This implies for any $\tilde{p} \geq 1$, we have $\|\Pi_{-[\tilde{p}]}\|_{2,1} \leq \bar{C}\bar{\rho}^{\tilde{p}}/(1-\bar{\rho})$. The quantities $|||f_y|||$, $\left|\left|\left|f_y^{-1}\right|\right|\right|$ 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 $\Sigma$. For some $\alpha \geq 0$, $y, \varepsilon \in \mathbb{R}^N$ and $X \in \mathbb{R}^{N\times D}$, consider the linear regression model $y = X\beta^{*(\alpha)} + \varepsilon$ with identified target*

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

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

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

*for some $n$ and $M$, where $M \geq \|\beta^{*(\alpha)}\|_1$. Then for any choice of $\lambda \geq 2\left\|X^\top\varepsilon/n\right\|_\infty$ and*

$q_n \geq \left\|X^\top X/n - \Sigma\right\|_\infty$, *the following holds:*

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

$$\textit{(b) Partially-Identified Estimation: } \min_{\beta:\Sigma\beta=\rho} \|\hat{\beta}^{(\alpha)} - \beta\|^2 \leq \frac{4q_n M^2 + \lambda\left[2M + \alpha M^2/2\right]}{\Lambda^+_{\min}(\Sigma)},$$

*where* $\Lambda^+_{\min}(\Sigma)$ *is the smallest non-zero eigenvalue of* $\Sigma$.
*In addition, define the constrained version of the estimator*

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

*Then, for any* $r_n \geq \frac{1}{n}\left\|X^\top\varepsilon\right\|_\infty$, *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*

$$\textit{(c) Point-Identified Estimation: } \left\|\hat{\beta}^{(\alpha)}_{[C]} - \beta^{*(\alpha)}\right\|^2 \leq 2v_n + 2(\sqrt{D}/\alpha + M)v_n^{1/2},$$

*where* $v_n := \frac{4Mr_n + 2s_n + 4M^2 q_n}{\Lambda^+_{\min}(\Sigma)}$.

The VARMA estimator from Phase-II can be expressed in the above regression format (see Equation (4.5)) 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, $\lambda, q_n, r_n$ 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 $\ell_1$-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 $\ell_1$-norm. This error bound also increases linearly with the inverse of $\alpha$, 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 [59], 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 $\hat{\varepsilon}_t$ for use in Phase-II. As a by-product, we also provide estimation error bounds for VAR($\infty$) 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:

$$y_t = \sum_{\tau=1}^{\tilde{p}} \Pi_\tau y_{t-\tau} + \varepsilon_t, \quad \text{where} \quad \varepsilon_t = \left( a_t + \sum_{\tau=\tilde{p}+1}^{\infty} \Pi_\tau y_{t-\tau} \right). \tag{4.1}$$

The autoregressive design takes the form $\mathcal{Y}_{T\times d} = \mathcal{X}_{T\times d\tilde{p}} B_{d\tilde{p}\times d} + E_{T\times d}$, where $\mathcal{Y} = [y_T : y_{T-1} : \ldots : y_1]^\top$, $\mathcal{X} = ((y_{T-i-j+1}))_{1\le i\le T, 1\le j\le \tilde{p}}$, $B = [\Pi_1 : \ldots : \Pi_{\tilde{p}}]^\top$ and $E = [\varepsilon_T : \varepsilon_{T-1} : \ldots : \varepsilon_1]^\top$. Vectorizing this regression design with $T$ samples and $d^2\widetilde{p}$ parameters, we have $Y = Z\beta^* + \mathrm{vec}(E)$, where $Y = \mathrm{vec}(\mathcal{Y})$, $Z = I \otimes \mathcal{X}$, and $\beta^* = \mathrm{vec}(B)$. In Phase-I, we consider a lasso estimator

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^{d^2\tilde{p}}}{\mathrm{argmin}} \frac{1}{T} \|Y - Z\beta\|^2 + \lambda \|\beta\|_1, \tag{4.2}$$

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

Our next proposition provides upper bounds on the approximation error of $a_t$ by $\hat{\varepsilon}_t$ for a random realization of $(T + \widetilde{p})$ data points from the VARMA model (1.1). A complete proof is given in Appendix D.

**Proposition 4.2.** *Consider any solution $\hat{\beta}$ of (4.2) using a random realization of $\{y_t\}_{t=1-\tilde{p}}^{T}$ from the VARMA model (1.1). Choose $\tilde{p} \asymp T^{\frac{1}{2}-\epsilon}$ for some $\epsilon \in (0, 1/2)$, and $\lambda \geq \lambda_0$, where*

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

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

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \|\hat{\varepsilon}_t - \varepsilon_t\|^2 &\leq \Delta_\varepsilon^2 := 2\lambda \sum_{\tau=1}^{\tilde{p}} \|\Pi_\tau\|_1 , \\
\max_{1\leq j\leq d} \frac{1}{T}\sum_{t=1}^{T} (\hat{\varepsilon}_{tj} - a_{tj})^2 &\leq \Delta_a^2 := 4 \max\left\{ \Delta_\varepsilon^2, 4\pi \left\| \Pi_{-[\tilde{p}]} \right\|_{2,1}^2 |||f_y||| \right\}, \\
\frac{1}{T}\sum_{t=1}^{T} \|\hat{\varepsilon}_t - a_t\|^2 &\leq 4 \max\left\{ \Delta_\varepsilon^2, 4\pi d \left\| \Pi_{-[\tilde{p}]} \right\|_{2,1}^2 |||f_y||| \right\}.
\end{aligned}$$

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

**Remark 4.3** (Convergence Rate & Truncation Bias)**.** The error bounds $\Delta_\varepsilon^2$ and $\Delta_a^2$ scale with $\lambda_0$, which has two terms. The first term decays polynomially with $T$. The second term $\left\| \Pi_{-[\tilde{p}]} \right\|_{2,1}$ captures the *truncation bias* arising from using a VAR($\widetilde{p}$) approximation to a VAR($\infty$) process. When $\widetilde{p} \asymp T^{\frac{1}{2}-\epsilon}$, this term decays exponentially with $T^{\frac{1}{2}-\epsilon}$ since

$$\left\| \Pi_{-[\tilde{p}]} \right\|_{2,1} \leq \frac{\bar{C}}{1-\bar{\rho}} \bar{\rho}^{\tilde{p}} = \frac{\bar{C}}{1-\bar{\rho}} \exp\left[ -T^{\frac{1}{2}-\epsilon} \log(1/\bar{\rho}) \right], \tag{4.3}$$

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 $\tilde{p}$, Slow & Fast Rates, and RE Condition)**.** As long as $\widetilde{p}$ increases polynomially fast with $T$, the truncation bias vanishes as $T \to \infty$ and the approximation errors $\Delta_\varepsilon$ and $\Delta_a$ decay with $T$ at a rate $O(\sqrt{\log d/T})$. However, under sparsity of $\Pi$ 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 $\widetilde{p}$ matches with low-dimensional VARMA analysis presented in [22].

## 4.3 Prediction and Estimation Error in Phase-II

For simplicity of exposition, we assume that $p$ and $q$ are known and $\widetilde{p} > p + q$. 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:

$$y_t = \sum_{\ell=1}^{p} \Phi_\ell y_{t-\ell} + \sum_{m=0}^{q} \Theta_m \hat{\varepsilon}_{t-m} + u_t, \qquad \text{for } t = 1, 2, \ldots, n, \qquad n = T - q, \tag{4.4}$$

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

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

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

$$\underbrace{\mathrm{vec}(\mathcal{Y})}_{Y} = \underbrace{\left(I \otimes \tilde{\mathcal{Z}}\right)}_{\tilde{Z}} \underbrace{\mathrm{vec}(B)}_{\beta^*} + \underbrace{\mathrm{vec}(\mathcal{U})}_{U}. \tag{4.5}$$

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 $\lambda$) that hold with high probability for a random realization of $(T+\tilde{p})$ consecutive observations from the VARMA process. To this end, note that $\left\| \left(I \otimes \tilde{\mathcal{Z}}\right)^\top \left(I \otimes \tilde{\mathcal{Z}}\right)/n - I \otimes \Sigma_z \right\|_\infty = \left\| \tilde{\mathcal{Z}}^\top \tilde{\mathcal{Z}}/n - \Sigma_z \right\|_\infty$.

In Section 4.2, we have discussed how the approximation errors $\Delta_a$, $\Delta_\varepsilon$ 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 (4.5) with design matrix $I \otimes \tilde{\mathcal{Z}}$ and error vector vec$(\mathcal{U})$. Set $\sigma_j^2 = e_j^\top \mathrm{Var}\left(\Theta(L)\Pi_{-[\tilde{p}]}(L)y_t\right) e_j$, for $j = 1, \ldots, d$. Then there exist universal constants $c_i > 0$ such that the event*

$$\mathcal{E} := \left\{ \left\| \tilde{\mathcal{Z}}^\top \tilde{\mathcal{Z}}/n - \Sigma_z \right\|_\infty \le q_n, \frac{1}{n} \left\| \tilde{\mathcal{Z}}^\top \mathcal{U} \right\|_\infty \le r_n, \left| \frac{1}{n} \|vec(\mathcal{U})\|^2 - \sum_{j=1}^{d} \sigma_j^2 \right| \le s_n \right\} \tag{4.6}$$

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

$$\begin{aligned}
q_n &= \varphi_{q,1}\sqrt{\frac{\log\, d^2(p+q)}{n}} + \varphi_{q,2}\left(\Delta_a + \Delta_a^2\right), \\
r_n &= \varphi_{r,1}\sqrt{\frac{\log\, d^2(p+q)}{n}} + \varphi_{r,2}\left(\Delta_\varepsilon + \Delta_\varepsilon^2 + \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}\right), \\
s_n/d &= \varphi_{s,1}\sqrt{\frac{\log\, d^2(p+q)}{n}} + \varphi_{s,2}\left(\Delta_\varepsilon + \Delta_\varepsilon^2\right),
\end{aligned}$$

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

$$\begin{aligned}
\varphi_{q,1} &= 2\pi |||f_y||| \left(p + q \left|||\Pi_{[\tilde{p}]}|||^2\right)^2, \\
\varphi_{q,2} &= \max\{2q, 2\sqrt{2\pi q} |||f_y|||^{1/2} \left(p + q \left|||\Pi_{[\tilde{p}]}|||^2\right)^{1/2}\}, \\
\varphi_{s,1} &= 2\pi |||\Theta||| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2 |||f_y|||, \\
\varphi_{s,2} &= \max\left\{2\|\Theta\|_{2,1}^2, 4\sqrt{2\pi} |||\Theta|||^{1/2} \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1} |||f_y|||^{1/2} \|\Theta\|_{2,1}\right\}, \\
\varphi_{r,1} &= c_1 |||f_y||| A \max\left\{1, |||\Theta|||^2 \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2, \left|||\Pi_{[\tilde{p}]}|||^2\right\}, \right. \\
\varphi_{r,2} &= c_2 |||f_y||| \|\Theta\|_{2,1} \max\{1, \left\|\Pi_{[\tilde{p}]}\right\|_{2,1}\}.
\end{aligned}$$

Using Proposition 2.1, the identification target in (2.3) with an elastic net penalty becomes

$$(\Phi^{(\alpha)}, \Theta^{(\alpha)}) = \underset{\Phi,\Theta}{\operatorname{argmin}} \left\{ \|[\Phi : \Theta]\|_1 + \frac{\alpha}{2} \|[\Phi : \Theta]\|_F^2 \text{ s.t. } \operatorname{vec}(\rho_{zy}) = (I \otimes \Sigma_z)\operatorname{vec}(\beta) \right\}, \quad (4.7)$$

where $\rho_{zy}, \Sigma_z$ and $\beta$ are as defined in Proposition 2.1. We consider the penalized and constrained versions of the estimator

$$\begin{aligned}
\operatorname{vec}\left([\hat{\Phi}^{(\alpha)} : \hat{\Theta}^{(\alpha)}]^\top\right) &= \underset{\|\beta\|_1 \le M}{\operatorname{argmin}} \frac{1}{n} \left\|\operatorname{vec}(\mathcal{Y}) - (I \otimes \tilde{\mathcal{Z}})\beta\right\|^2 + \lambda \mathcal{P}_\alpha(\beta) \\
\operatorname{vec}\left([\hat{\Phi}_{[C]}^{(\alpha)} : \hat{\Theta}_{[C]}^{(\alpha)}]^\top\right) &= \underset{\|\beta\|_1 \le M}{\operatorname{argmin}} \left\{ \mathcal{P}_\alpha(\beta) \text{ s.t. } \frac{1}{n} \left\|\operatorname{vec}(\mathcal{Y}) - (I \otimes \tilde{\mathcal{Z}})\beta\right\|^2 \le A_n \right\}.
\end{aligned}$$

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* (1.1)*, and let* $n = T - q$ *denote the sample size in Phase-II. Denote* $K_y := \max\{|||f_y|||, \|\Pi\|_{2,1}, \left\|\Theta^{(\alpha)}\right\|_{2,1}\}$.

*(a)* <u>*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* $\tilde{y}_t = \sum_{\ell=1}^p \widehat{\Phi}_\ell y_{t-\ell} + \sum_{m=1}^q \widehat{\Theta}_m \hat{\varepsilon}_{t-m}$ *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* $M \geq \|\Phi^{(\alpha)}\|_1 + \|\Theta^{(\alpha)}\|_1$ *for some* $\alpha \geq 0$,

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

*(b)* Partially-identified Estimation*: With the same choice of* $\lambda$*,* $M$ *and* $\alpha$ *in (a), the penalized estimator is partially identified and satisfies*

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

*(c)* Point-identified Estimation*: 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*

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

Part (a) of this proposition ensures that as long as the identification target is parsimonious in the sense of small $\ell_1$-norm and the penalty parameter is chosen appropriately, the VARMA forecasts converge to the optimal forecasts (which uses any element from the equivalence class $\mathcal{E}_{p,q}(\Pi)$) in the asymptotic regime $\log d/n \to 0$. The truncation bias term $\|\Pi_{-[\tilde{p}]}\|_{2,1}$ and the approximation error from Phase-I $\Delta_\varepsilon$ 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 $\mathcal{E}_{p,q}(\Pi)$, so searching for the correct identification target within this equivalence class still has a complexity of the order of $d^2$. The tuning parameter $\alpha$ also affects the convergence rate, since this captures

the degree of curvature of the term $\mathcal{P}_\alpha(.)$ in the loss function. However, taking a sequence of $\alpha_n$ 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 [35] (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 $\widehat{L}_{\widehat{\Theta}}$ of the estimated VARMA whose elements are $\widehat{L}_{\widehat{\Theta},ij} = \max\{m : \widehat{\Theta}_{m,ij} \neq 0\}$, where $\widehat{L}_{\widehat{\Theta},ij} = 0$ if $\widehat{\Theta}_{m,ij} = 0$ for all $m = 1\ldots, \hat{q}$. This lag

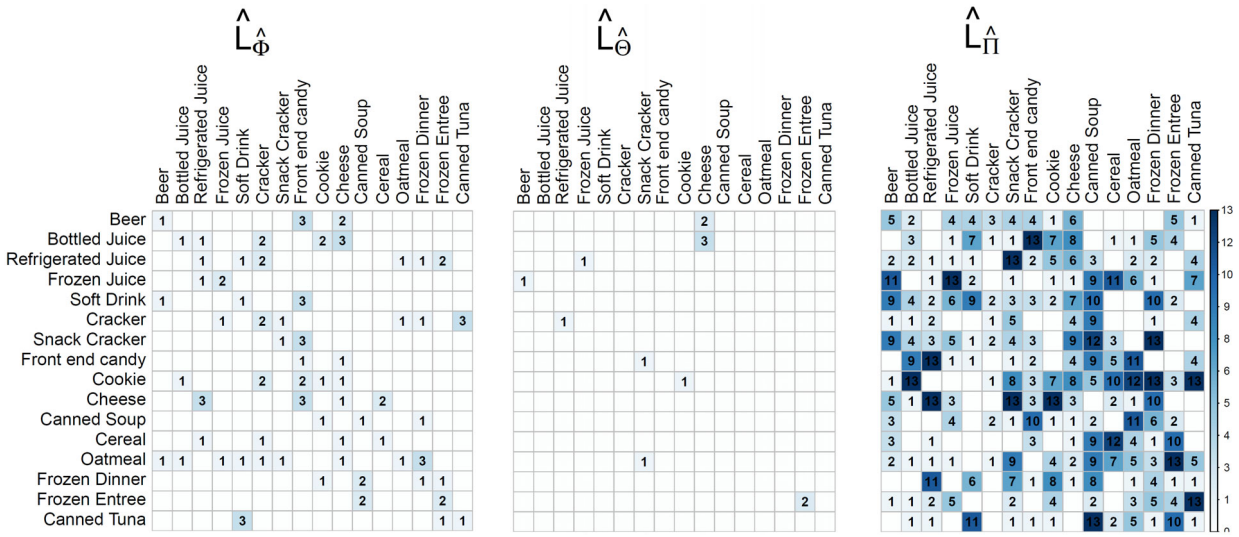


Figure 1: Demand data set: AR-lag matrix (left) and MA-lag matrix (middle) of the estimated VARMA, and AR-lag matrix of the estimated VAR (right).

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 $\widehat{L}_{\widehat{\Phi}}$ of the estimated VARMA and the autoregressive lag matrix $\widehat{L}_{\widehat{\Pi}}$ 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

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 Demand Data | | | Monthly Volatility Data | | | Quaterly 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) |

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 (3.5). 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 ([19]).

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 [18]. 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 [31, 52], and are left for future research.

## 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.

# Supplement to "Sparse Identification and Estimation of Large-Scale Vector AutoRegressive Moving Averages"

We present the proofs of sparse identification in Section A. Proofs of key technical ingredients required for Phase-I and II analyses are in Section B, along with some additional lemmas to control the error due to using $\hat{\varepsilon}_t$ instead of $\varepsilon_t$ in Phase-II. Sections C, D and E contain results for error bound analysis in elastic net, Phase-I and II, respectively. Section F contains details of Phase-I and II algorithms. Section G presents the results on several numerical experiments.

*Notations.* 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, $\ell_1$ and $\ell_\infty$ norms of a vector or the vectorized version of a matrix, respectively, and $\|.\|_F$ to denote the Frobenius norm of a matrix. The symbol $\mathbb{S}^{d-1}$ is used to denote the vectors $v \in \mathbb{R}^d$ with $\|v\| = 1$. We use $\Lambda_{\max}(.)$ and $\Lambda_{\min}(.)$ to denote the maximum and minimum eigenvalues of a (symmetric or Hermitian) matrix. We use $|.|$ to denote the absolute value of a real number or complex number. We use $V^*$ to denote the conjugate transpose of a complex matrix, vector or scalar $V$. For a matrix-valued, possibly infinite-order lag polynomial $\mathcal{A}(L) = \sum_{\ell \geq 0} A_\ell L^\ell$, we define $|||\mathcal{A}||| := \max_{\theta \in [-\pi,\pi]} \|\mathcal{A}(e^{i\theta})\|$, and use $\mathcal{A}_{[k]}(L)$ and $\mathcal{A}_{-[k]}(L)$ 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 $\|\mathcal{A}\|_{2,1}$ 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 $f$ of frequencies from $[-\pi, \pi]$ 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. Their values are allowed to change from equation to equation. For example, we will use $c_0$ instead of $2c_0, c_0 + 2$ etc. within a proof to keep the notations simple. 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 $A \lesssim B$.

*Measures of Dependence.* We adopt the spectral density based measures of dependence introduced in [10] to conduct our non-asymptotic analysis. For a $d$-dimensional centered stationary Gaussian time series $\{x_t\}_{t\in\mathbb{Z}}$ with autocovariance function $\Gamma_x(h) = \mathrm{Cov}(x_t, x_{t+h}) = \mathbb{E}[x_t x_{t+h}^\top]$, $h \in \mathbb{Z}$, we assume the spectral density function $f_x(\theta) := \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\Gamma_x(\ell)e^{-i\ell\theta}$, $\theta \in [-\pi,\pi]$, exists, is non-singular a.e. on $[-\pi,\pi]$, and $|||f_x||| < \infty$. The quantity $|||f_x|||$ is taken as a measure of temporal and cross-sectional dependence in the time series $\{x_t\}$. We say that the time series $x_t$ is *stable* if $|||f_x||| < \infty$. More generally, for any pair of $d$-dimensional centered, stable time series $\{x_t\}$ and $\{y_t\}$, the cross-spectral density is defined as $f_{x,y}(\theta) = \frac{1}{2\pi}\sum_{\ell=-\infty}^{\infty}\Gamma_{x,y}(\ell)e^{-i\ell\theta}$, where $\Gamma_{x,y}(h) = \mathrm{Cov}(x_t, y_{t+h})$, for $h \in \mathbb{Z}$. If the joint process $w_t = [x_t^\top, y_t^\top]^\top$ is stable, i.e. it satisfies $|||f_w||| < \infty$, it follows that $|||f_{x,y}|||^2 \le |||f_x|||\,|||f_y|||$.

For a stable, invertible VARMA process $y_t$ in (1.1) with $\Lambda_{\min}(\Sigma_a) > 0$, it is known that $f_y$ is non-singular on $[-\pi,\pi]$ 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 $\tau \ge 1$. 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 $|||f_y|||$, $\left|\left|\left|f_y^{-1}\right|\right|\right|$ and $\|\Pi_{-[\tilde{p}]}\|_{2,1}$ appear in our error bounds, and captures the effects of temporal dependence on the convergence rates.

# A Proofs for Sparse Identification

## A.1 Yule-Walker type Equations for VARMA

*Proof of Proposition 2.1.* Define

$$\tilde{\mathcal{E}}_{p,q}(\Pi(L)) := \{(\Phi,\Theta) : \rho_{zy} = \Sigma_z\beta\}\,.$$

We first show that $\mathcal{E}_{p,q}(\Pi(L)) \subseteq \tilde{\mathcal{E}}_{p,q}(\Pi(L))$. To this end, note that any $(\Phi,\Theta) \in \mathcal{E}_{p,q}(\Pi(L))$ satisfies $y_t = \beta^\top z_t + a_t$. Therefore, $\mathbb{E}\left[y_t z_t^\top\right] = \beta^\top\mathbb{E}\left[z_t z_t^\top\right] + \mathbb{E}\left[a_t z_t^\top\right]$. Since $\mathbb{E}\left[a_t z_t^\top\right] = 0$, this implies $\rho_{zy}^\top = \beta^\top\Sigma_z$.

Next we show that $\tilde{\mathcal{E}}_{p,q}(\Pi(L)) \subseteq \mathcal{E}_{p,q}(\Pi(L))$. To this end, note that the set $\mathcal{E}_{p,q}(\Pi(L))$ can be characterized precisely as the set of matrix AR and MA parameters $\Phi$ and $\Theta$ which satisfy almost surely (a.s.)

$$y_t = \beta^\top z_t + a_t, \quad t \in \mathbb{Z}, \tag{A.1}$$

for a process $y_t = \Pi^{-1}(L)a_t$, where $a_t \overset{i.i.d.}{\sim} (0, \Sigma_a)$ is a white noise process.

Now, consider a solution of the Yule-Walker type linear systems of equation $\beta \in \tilde{\mathcal{E}}_{p,q}(\Pi(L))$. Since $\mathcal{E}_{p,q}(\Pi(L)) \subseteq \tilde{\mathcal{E}}_{p,q}(\Pi(L))$ and $\mathcal{E}_{p,q}(\Pi(L)) \neq \phi$, this solution takes the form $\beta = \beta^* + \delta$, where $\beta^* = \left[\Phi_1^* : \ldots : \Phi_p^* : \Theta_1^* : \ldots : \Theta_q^*\right]^\top \in \mathcal{E}_{p,q}(\Pi(L))$ is a particular solution of the linear systems, and $\delta = [\delta_{11} : \ldots : \delta_{1p} : \delta_{21} : \ldots : \delta_{2q}]^\top$ satisfies $\Sigma_z \delta = \mathbf{0}_{d(p+q)\times d}$.

This implies $\delta^\top \Sigma_z \delta = \mathbf{0}_{d\times d}$, i.e. $var(\delta^\top z_t) = \mathbf{0}_{d\times d}$. In other words, $\delta^\top z_t$ is almost surely a constant. Since $\mathbb{E}[z_t] = 0$, we conclude that $\delta^\top z_t = 0$ a.s.

Now, consider any centered linear process $y_t = \Pi^{-1}(L)a_t$, as mentioned above. Then, since $\beta^* \in \mathcal{E}_{p,q}(\Pi(L))$, for any $t \in \mathbb{Z}$ we have

$$y_t = \Phi_1^* y_{t-1} + \ldots + \Phi_p^* y_{t-p} + \Theta_1^* a_{t-1} + \ldots + \Theta_q^* a_{t-q} + a_t.$$

Also, since, $\delta^\top z_t = 0$ a.s., we have

$$\begin{aligned} y_t = {} & (\Phi_1^* + \delta_{11})y_{t-1} + (\Phi_2^* + \delta_{12})y_{t-2} + \ldots + (\Phi_p^* + \delta_{1p})y_{t-p} \\ & + (\Theta_1^* + \delta_{21})a_{t-1} + \ldots + (\Theta_q^* + \delta_{2q})a_{t-q} + a_t \text{ a.s.} \end{aligned}$$

It follows from (A.1) that $\beta \in \mathcal{E}_{p,q}(\Pi(L))$, proving $\tilde{\mathcal{E}}_{p,q}(\Pi(L)) \subseteq \mathcal{E}_{p,q}(\Pi(L))$.

□

## A.2 Optimization-based Identification

Consider the convex minimization problem

$$\mathcal{C}^* = \arg\min_{x\in\mathcal{L}} f(x)$$

where $f : \mathbb{R}^n \to [0, \infty)$ is a convex function and $\mathcal{L} \subseteq \mathbb{R}^n$ is an affine space. We assume that $\mathcal{C}^*$ is non-empty (i.e. the minimum is attained) and let

$$x^* = \arg\min_{x\in\mathcal{L}} \|x\|^2 \text{ s.t. } x \in \mathcal{C}^*,$$

which is unique since this is a strongly convex problem.

Defining $f(x,\alpha) = f(x) + \frac{\alpha}{2}\|x\|^2$, we see that $f(\cdot,\alpha)$ is $\alpha$-strongly convex for each $\alpha > 0$ and therefore there is a unique minimizer

$$x_\alpha := \arg\min_{x\in\mathcal{L}} f(x,\alpha).$$

**Proposition A.1.** *The sequence of minimizers of $f(\cdot,\alpha)$ converge, as $\alpha \to 0^+$, to the unique minimizer of $f(\cdot)$ that has smallest $\ell_2$-norm:* $\lim_{\alpha\to 0^+} x_\alpha = x^*$.

*Proof.* We begin with a lemma.

**Lemma 1.**

$$\lim_{\alpha\to 0^+} \frac{f(x_\alpha,\alpha) - f(x^*,\alpha)}{\alpha} = 0.$$

*Proof.* By definition of $x^*$,

$$\|x^*\|^2 = \min_{x\in\mathcal{L}} \|x\|^2 \text{ s.t. } f(x) \le f^*,$$

where $f^* = \min_{x\in\mathcal{L}} f(x)$. This can be equivalently expressed (see, e.g., [11]) as

$$\|x^*\|^2 = \min_{x\in\mathcal{L}} \sup_{\lambda\ge 0} L(x;\lambda)$$

where

$$L(x;\lambda) = \|x\|^2 + \lambda(f(x) - f^*) = \lambda[f(x, 2/\lambda) - f^*].$$

By strong duality (Slater's condition holds since $\mathcal{C}^* \neq \emptyset$), we can interchange the "min" and the "sup":

$$\|x^*\|^2 = \sup_{\lambda\ge 0} g(\lambda),$$

where $g(\lambda) = \min_{x\in\mathcal{L}} L(x;\lambda)$. Now, for $\bar{\lambda} > \lambda \geq 0$,

$$\begin{aligned} g(\bar{\lambda}) &= \min_{x\in\mathcal{L}} L(x,\bar{\lambda}) \\ &= \min_{x\in\mathcal{L}} \left\{ L(x,\lambda) + (\bar{\lambda}-\lambda)[f(x) - f^*] \right\} \\ &\geq \min_{x\in\mathcal{L}} L(x,\lambda) + (\bar{\lambda}-\lambda) \min_{x\in\mathcal{L}} [f(x) - f^*] \\ &\geq g(\lambda). \end{aligned}$$

Thus, $g$ is a non-decreasing function, and

$$\lim_{\lambda\to\infty} g(\lambda) = \sup_{\lambda\geq 0} g(\lambda) = \|x^*\|^2.$$

Now,

$$\|x^*\|^2 = \lim_{\lambda\to\infty} g(\lambda) = \lim_{\lambda\to\infty} \left\{ \lambda[f(x_{2/\lambda}, 2/\lambda) - f^*] \right\} = \lim_{\alpha\to 0^+} (2/\alpha)[f(x_\alpha,\alpha) - f^*]$$

or, subtracting $\|x^*\|^2$ from both sides,

$$0 = \lim_{\alpha\to 0^+} (2/\alpha)[f(x_\alpha,\alpha) - f(x^*,\alpha)].$$

□

By $\alpha$-strong convexity of $f(\cdot,\alpha)$,

$$f(y,\alpha) \geq f(x_\alpha,\alpha) + \frac{\alpha}{2}\|x_\alpha - y\|^2 \tag{A.2}$$

for any $y \in \mathcal{L}$.

Applying this with $y = x^*$ gives

$$f(x^*,\alpha) \geq f(x_\alpha,\alpha) + \frac{\alpha}{2}\|x_\alpha - x^*\|^2$$

or

$$\|x_\alpha - x^*\|^2 \leq (2/\alpha)[f(x^*,\alpha) - f(x_\alpha,\alpha)].$$

Taking the limit of both sides, Lemma 1 gives

$$\lim_{\alpha \to 0} \|x_\alpha - x^*\|^2 \leq 0.$$

Thus,

$$\lim_{\alpha \to 0} x_\alpha = x^*.$$

□

The above results are now easily applied to prove Proposition 2.2.

*Proof of Proposition 2.2.* Denote $x = (\Phi, \Theta)$. Consider the convex function $f(x) = \mathcal{P}_{\mathrm{AR}}(\Phi) + \mathcal{P}_{\mathrm{MA}}(\Theta)$, and the affine space $\mathcal{L}$ in which $\Phi(L) = \Theta(L)\Pi(L)$ holds. It follows from Proposition A.1 that $\lim_{\alpha \to 0^+}(\Phi^{(\alpha)}, \Theta^{(\alpha)}) = (\Phi^{(0)}, \Theta^{(0)})$. □

## A.3 Identification for Multiple, Sparsest VARMA Representations

### A.3.1 A Toy Example

In Section 2.3, we refer to multiple equivalent, sparsest VARMA representations, as, for instance, discussed in Section 4.5.2 of [51]. As an example, we consider the VAR(1) and VMA(1) models

$$y_t = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} y_{t-1} + a_t \Leftrightarrow y_t = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} a_{t-1} + a_t.$$

In this section, we establish our unique identification target for this example.

Following the VARMA($p$,$q$) notation of our paper we write $\Phi(L)y_t = \Theta(L)a_t$, where the AR and MA operators are respectively given by

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

with the lag operator $L^\ell$ defined as $L^\ell y_t = y_{t-\ell}$. For the VMA(1) example with MA-coefficient

matrix equal to

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

we equivalently have $\Phi(L) = (I - 0L) = I$ and $\Theta(L) = (I + AL)$ in the VARMA(1,1) formulation. Further, since $\det\{\Phi(z)\} \neq 0$ and $\det\{\Theta(z)\} \neq 0$ for all $|z| \leq 1$ ($z \in \mathbb{C}$), this model is stable and invertible, and the process $\{y_t\}$ then has an infinite-order VAR representation $\Pi(L)y_t = a_t$, where $\Pi(L) = \Theta^{-1}(L)\Phi(L) = I - \Pi_1 L - \Pi_2 L^2 - \cdots$, which in this case simplifies to $\Pi(L) = (I + AL)^{-1}I = I - AL$. We then recognize this model is equivalent to a VAR(1) model with AR-coefficient matrix $A$, or we equivalently have $\widetilde{\Phi}(L) = (I - AL)$ and $\widetilde{\Theta}(L) = (I + 0L) = I$, in its VARMA(1,1) formulation.

In our paper, we therefore note that both models, as defined by their AR and MA coefficient matrix pairs $(\Phi, \Theta)$: $(I, A)$ and $(A, I)$, respectively, are in the same VARMA(1,1) equivalence class $\mathcal{E}_{1,1}$ with respect to $\Pi(L) = (I - AL)$. This is defined for the general VARMA(p,q) model as $\mathcal{E}_{p,q}(\Pi(L)) = \{(\Phi, \Theta) : \Phi(L) = \Theta(L)\Pi(L)\}$, and in this case we specifically have

$$\mathcal{E}_{1,1}(I - AL) = \{(\overline{\Phi}, \overline{\Theta}) : (I - \overline{\Phi}L) = (I + \overline{\Theta}L)(I - AL)\}.$$

For the equivalence relation of $\mathcal{E}_{1,1}$ to hold for the given $A$, any matrix pair $(\overline{\Phi}, \overline{\Theta})$ in the set must also be of the form

$$\overline{\Phi} = \overline{\Phi}(a) = \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix} \quad \textit{and} \quad \overline{\Theta} = \overline{\Theta}(b) = \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix},$$

for $a, b \in \mathbb{R}$ such that, $a + b = 1$, and so there are many solutions, not just the two identified above.

We now turn to considering this problem from the proposed optimization-based identification perspective (Section 2.2) by using strongly convex optimization to establish identification. Among all feasible AR and MA matrix pairs $(\overline{\Phi}, \overline{\Theta})$, we look for the one that gives the most parsimonious representation of the VARMA. Specifically, we measure parsimony through a pair of convex regularizers, $\mathcal{P}_{\text{AR}}(\overline{\Phi})$ and $\mathcal{P}_{\text{MA}}(\overline{\Theta})$. Our identification results apply

equally well to any convex function; one can consider, amongst others, the $\ell_1$-norm, the $\ell_2$-norm, the nuclear norm, and convex combinations thereof.

To be concrete for this particular example, let us specifically consider using the $\ell_1$-norm: $\mathcal{P}_{\mathrm{AR}}(\overline{\Phi}) = \|\overline{\Phi}\|_1$ and $\mathcal{P}_{\mathrm{MA}}(\overline{\Theta}) = \|\overline{\Theta}\|_1$. Then for any fixed $\alpha > 0$, a uniquely identified solution $(\overline{\Phi}^{(\alpha)}, \overline{\Theta}^{(\alpha)})$ is

$$\underset{\overline{\Phi},\overline{\Theta}}{\operatorname{argmin}} \left\{ \|\overline{\Phi}\|_1 + \|\overline{\Theta}\|_1 + \frac{\alpha}{2}\|\overline{\Phi}\|_F^2 + \frac{\alpha}{2}\|\overline{\Theta}\|_F^2 \text{ s.t. } (I - \overline{\Phi}L) = (I + \overline{\Theta}L)(I - AL) \right\}$$

(see general case, Equation (2.3)), and this is equivalent to

$$\underset{\overline{\Phi}(a),\overline{\Theta}(b)}{\operatorname{argmin}} \left\{ \left\| \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix} \right\|_1 + \left\| \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \right\|_1 + \frac{\alpha}{2} \left\| \begin{pmatrix} 0 & a \\ 0 & 0 \end{pmatrix} \right\|_F^2 + \frac{\alpha}{2} \left\| \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \right\|_F^2 \text{ s.t. } a + b = 1 \right\},$$

or more simply,

$$\underset{\overline{\Phi}(a),\overline{\Theta}(b)}{\operatorname{argmin}} \left\{ |a| + |b| + \frac{\alpha}{2}|a|^2 + \frac{\alpha}{2}|b|^2 \text{ s.t. } a + b = 1 \right\},$$

This optimization problem is strongly convex and thus has a unique solution pair $(\overline{\Phi}^{(\alpha)}, \overline{\Theta}^{(\alpha)})$ for each value of $\alpha > 0$. We further define our final (unique) optimization-based identified VARMA representation as

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

a result which is proved (in the general case) in Proposition 2.2 to be the *unique* pair of autoregressive and moving average matrices in the 'regularized equivalent' class having smallest Frobenius norm, i.e., the regularized equivalent (sub-) class of $\mathcal{E}_{1,1}(I - AL)$ in this example is defined as $\mathcal{RE}_{1,1}(I - AL) =$

$$\underset{\overline{\Phi},\overline{\Theta}}{\operatorname{argmin}} \left\{ \|\overline{\Phi}\|_1 + \|\overline{\Theta}\|_1 \text{ s.t. } (I - \overline{\Phi}L) = (I + \overline{\Theta}L)(I - AL) \right\} = \underset{\overline{\Phi}(a),\overline{\Theta}(b)}{\operatorname{argmin}} \left\{ |a| + |b| \text{ s.t. } a + b = 1 \right\}$$

(which has many solutions, i.e., $b = 1 - a, a \in [0, 1]$). Then our final unique solution for this

specific problem is

$$(\overline{\Phi}^{(0)}, \overline{\Theta}^{(0)}) = \underset{\overline{\Phi}(a),\overline{\Theta}(b)}{\operatorname{argmin}} \left\{ |a|^2 + |b|^2 \text{ s.t. } (\overline{\Phi}(a), \overline{\Theta}(b)) \in \mathcal{RE}_{1,1}(I - AL) \right\}$$
$$= \underset{\overline{\Phi}(a),\overline{\Theta}(b)}{\operatorname{argmin}} \left\{ |a|^2 + |b|^2 \text{ s.t. } b = 1 - a, a \in [0, 1] \right\}.$$

This has the unique solution $a = b = 0.5$, or

$$\overline{\Phi} = \overline{\Theta} = \begin{pmatrix} 0 & 0.5 \\ 0 & 0 \end{pmatrix},$$

and we can further confirm by hand this solution is in $\mathcal{E}_{1,1}(I - AL)$, since

$$\left(I + \begin{pmatrix} 0 & 0.5 \\ 0 & 0 \end{pmatrix} L\right)^{-1} \left(I - \begin{pmatrix} 0 & 0.5 \\ 0 & 0 \end{pmatrix} L\right) = \left(I - \begin{pmatrix} 0 & 0.5 \\ 0 & 0 \end{pmatrix} L\right)^{2} = \left(I - \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} L\right) = (I - AL).$$

Finally, we note that although the proposed unique VARMA(1,1) solution above does not have as few non-zero parameters as either the pure VMA(1) or VAR(1) model (in which there was just one), in finding this solution via optimization with constraints there was still only one free parameter, and therefore the same overall model complexity in this regard. Furthermore, this is only the unique solution derived under the $\ell_1$-norm choice of regularization, and we reiterate that the flexible framework that we propose also allows any (user specified) convex function for regularization-based identification, including the $\ell_1$-norm, the $\ell_2$-norm, the nuclear norm, and convex combinations thereof.

### A.3.2 Simulation

We further illustrate sparse identification with a small simulation study. Figure A1 (panel a) shows a $\text{VARMA}_{d=8}(1, 1)$ model

$$\Phi_{\text{dense}} = \begin{bmatrix} \mathbf{0.2} & \mathbf{0.05} \\ \mathbf{0} & \mathbf{0.1} \end{bmatrix}, \text{ and } \Theta_{\text{dense}} = \begin{bmatrix} \mathbf{0} & \mathbf{-0.25} \\ \mathbf{0} & \mathbf{-0.1} \end{bmatrix},$$

with the dense $(\Phi, \Theta)$ having 80 nonzero entries. However, this VARMA model can be alternatively expressed in terms of

$$\Phi_{\text{sparse}} = \begin{bmatrix} \mathbf{0.2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \text{ and } \Theta_{\text{sparse}} = \begin{bmatrix} \mathbf{0} & \mathbf{-0.2} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

a sparse $(\Phi, \Theta)$ having only 32 nonzero entries (panel b); or

$$\Phi^{(0)} = \begin{bmatrix} \mathbf{0.1} & \mathbf{-0.1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \text{ and } \Theta^{(0)} = \begin{bmatrix} \mathbf{0.1} & \mathbf{-0.1} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

having only 64 nonzero entries (panel c).

Note that there are multiple equivalent, minimum-$\ell_1$ VARMA representations. Two of these are visualized in panels (b) and (c) but others exist such as the pair where the AR and MA matrices of the "sparse" design are swapped. All have minimal $\ell_1$-norm (i.e. $||\Phi||_1 = ||\Theta||_1 = 3.2$). Panel (c) displays the unique pair $(\Phi^{(0)}, \Theta^{(0)})$, defined in Equation (2.4), as the one having minimal $\ell_2$-norm (i.e. $||\Phi^{(0)}||_F^2 = ||\Theta^{(0)}||_F^2 = 0.32$). When choosing the $\ell_1$-norm as the convex regularizer, our optimization-based identification strategy would favor the sparser VARMA representations over the denser one since the former have a smaller $\ell_1$-norm (i.e. $\ell_1$-norm for the dense design is $||\Phi||_1 = ||\Theta||_1 = 5.6$).

To illustrate the link between our identification and estimation stages, consider the following simulation experiment: We take $\Sigma_a = I_d$ and generate time series of length $T = 1000$ (after 200 burn-in observations) from the *dense* VARMA (Figure A1, panel a). We then use our sparse VARMA procedure with $\ell_1$-norm as convex regularizer and take $p = q = 1$ to obtain the AR and MA parameter estimates. The number of simulations is $N = 500$.

First, we estimate the VARMA with $\alpha = 0$; the corresponding estimates are visualized in Figure A1 panel (d), for an illustrative simulation run. The results are very stable from one simulation run to another. Although we generate the time series from the dense DGP, our procedure encourages identification and estimation of sparser models and thus returns sparser estimates. Since there are infinitely many equivalent "true" $(\Phi, \Theta)$ pairs, we are not interested in comparing the estimates to the dense $(\Phi, \Theta)$ pair used to originally generate the

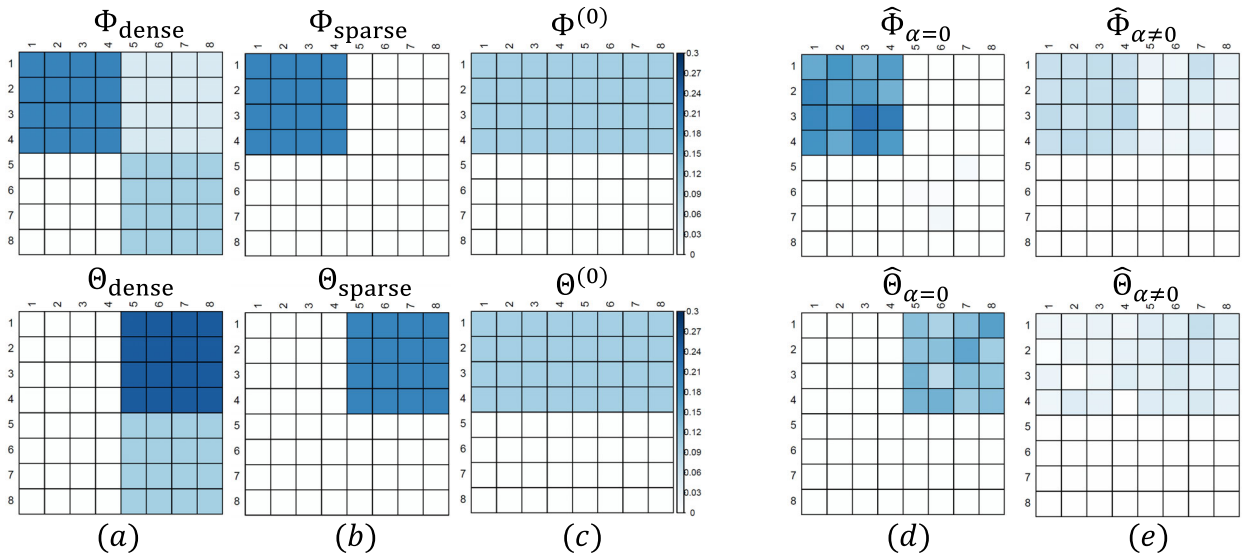


Figure A1: True AR and MA matrices from three equivalent VARMA representations: (a) a dense, (b) a sparse and (c) the target VARMA. The estimates obtained with our sparse VARMA estimation procedure for $\alpha = 0$ are displayed in panel (d); for a small $\alpha \neq 0$ in panel (e). Darker shading of cells indicate parameters that are larger in (absolute) magnitude.

data. Instead, we aim to produce estimates that are close to the sparse equivalence class. In almost all simulation runs (496 out of 500), Matthews Correlation Coefficient (MCC) between the sparse $(\Phi, \Theta)$ (Figure A1, panel b) and the estimates equals one; thereby providing perfect recovery. By taking $\alpha = 0$, our simulation results thus show that our estimates are very close to one of the elements in the sparse equivalance class, which is in line with our theoretical result on partially identified estimation.

Next, we estimate the VARMA with a non-zero but small $\alpha$ (we take $\alpha = 10^{-2}$, thus small relative to the selected values for $\lambda_\Phi \approx 10^2$ and $\lambda_\Theta \approx 10$). By adding the $\ell_2$-norm to the objective function, we expect to produce estimates that are closer to the unique $(\Phi^{(0)}, \Theta^{(0)})$. This expectation is confirmed by our results, as can be seen from the corresponding estimates, visualized in Figure A1 panel (e). The average (over the simulation runs) MCC between the target (Figure A1, panel c) and the estimates is 0.97 with a standard error smaller than 0.001.

Since there exist multiple equivalent, sparsest VARMA representations with different support, we do not focus on model selection consistency in the paper but instead on forecasting. For forecasting purposes, we are interested in obtaining a parsimonious VARMA represen-

tation with good out-of-sample performance. For this reason, we prefer to use $\alpha = 0$ in the simulation study and forecast applications since our numerical experiments showed that this generally produces a sparser (i.e. with fewer non-zero coefficients) estimated VARMA compared to the estimates obtained when taking $\alpha$ non-zero but small.

# B Key Technical Ingredients

Our first technical ingredient provides a deviation bound (in element-wise maximum norm) for the product of two random matrices, whose rows consist of consecutive observations from two time series that are outputs of a linear filter applied on the same stationary Gaussian time series. In the analysis of both Phase-I and Phase-II, we use this result to control upper bounds on inner products of columns of the design matrix and the error matrix. This proposition generalizes a similar concentration bound in [10] for uncorrelated time series.

**Proposition B.1.** *Let $\{y_t\}_{t\in\mathbb{Z}}$ be a $d$-dimensional stable, Gaussian, centered time series with spectral density $f_y$. Consider two time series $X_t = \mathcal{A}(L)y_t$ and $Y_t = \mathcal{B}(L)y_t$, whose $d \times d$ matrix-valued lag polynomials $\mathcal{A}(L)$ and $\mathcal{B}(L)$ satisfy $\|\mathcal{A}\|_{2,1} < \infty$, $\|\mathcal{B}\|_{2,1} < \infty$. Let $\mathcal{X} = [X_T : X_{T-1} : \cdots : X_1]^\top$ and $\mathcal{Y} = [Y_T : Y_{T-1} : \cdots : Y_1]^\top$ be two data matrices, each containing in its rows $T$ consecutive observations from the time series $\{X_t\}$ and $\{Y_t\}$, respectively. Then there exists a universal constant $c > 0$ such that for any $\eta > 0$ and any $u, v \in \mathbb{S}^{d-1}$, we have*

$$\mathbb{P}\left[\left|u^\top\left(\mathcal{X}^\top\mathcal{Y}/T - \Gamma_{X,Y}(0)\right)v\right| > 6\pi|||f_y|||\max\left\{|||\mathcal{A}|||^2, |||\mathcal{B}|||^2\right\}\eta\right] \tag{B.1}$$

*is at most $6\exp[-cT\min\{\eta, \eta^2\}]$.*
*In addition, if $T \gtrsim \log d$, then for any $A > 0$, the following upper bound holds with probability at least $1 - 6\exp\left[-2(cA^2 - 1)\log d\right]$:*

$$\left\|\mathcal{X}^\top\mathcal{Y}/T\right\|_\infty \leq 2\pi|||f_y|||\left[3A\ \max\left\{|||\mathcal{A}|||^2, |||\mathcal{B}|||^2\right\}\sqrt{2\log d/T} + |||\mathcal{A}|||\ \|\mathcal{B}\|_{2,1}\right].$$

**Remark B.1.** The two terms in the above bound can be viewed as the variance and bias terms. The first term provides a bound on the deviation of $\mathcal{X}^\top\mathcal{Y}/T$ around its expectation in element-wise maximum norm. This bound scales with the dimension $d$ at a rate $\sqrt{\log d/T}$

similar to the case of i.i.d. random variables. In addition, the terms $|||f_y|||$, $|||\mathcal{A}|||$ and $|||\mathcal{B}|||$ capture the effect of temporal dependence on the convergence rates. The second term provides a bound on the bias, i.e. the population covariance between the time series $X_t$ and $Y_t$. This Hölder-type bound involves the operator norms of the spectral density of $y_t$ (across frequencies), and the linear filters $\mathcal{A}(L)$ and $\mathcal{B}(L)$ applied on $y_t$. The bound on bias can be potentially improved using additional structures of the linear filters (see remark after proof below).

*Proof of Proposition B.1.* In order to obtain a high probability concentration bound, we first state a generalized version of Proposition 2.4(b) in [10], allowing for correlation between the two time series. The proof follows along the same line, only replacing $(2/n)\sum_{t=1}^{n} w^t z^t$ with $(2/n)\sum_{t=1}^{n} w^t z^t - Cov(z^t, w^t)$ in the left hand side of the first equation in their proof.

Let $\{X_t\}_{t\in\mathbb{Z}}$ and $\{Y_t\}_{t\in\mathbb{Z}}$ be two $d$-dimensional stationary Gaussian centered time series, with autocovariance function $\Gamma_{X,Y}(h) = cov(X_t, Y_{t+h}) = \mathbb{E}[X_t Y_{t+h}^\top]$ and cross-spectral density $f_{X,Y}$. Assume the process $Z_t = [X_t^\top : Y_t^\top]^\top$ is stable so that it has bounded cross-spectrum $|||f_{X,Y}||| < \infty$. Let $\mathcal{X}$ and $\mathcal{Y}$ be $T \times d$ data matrices, with rows corresponding to consecutive observations from the time series $\{X_t\}$ and $\{Y_t\}$, respectively. Then, for any $u$, $v \in \mathbb{R}^d$ with $\|u\| \le 1$, $\|v\| \le 1$, and any $\eta > 0$, we have

$$\mathbb{P}\left[\left|u^\top(\mathcal{X}^\top\mathcal{Y}/T - \Gamma_{X,Y}(0))v\right| > 2\pi\left[|||f_{X,Y}||| + |||f_X||| + |||f_Y|||\right]\eta\right] \tag{B.2}$$

is at most $6\exp[-cT\min\{\eta, \eta^2\}]$ for some universal constant $c > 0$.

Next, we use the fact that $|||f_{X,Y}|||^2$ is at most $|||f_X|||\,|||f_Y|||$, so that $|||f_{X,Y}||| + |||f_X||| + |||f_Y|||$ is at most $3\max\{|||f_X|||, |||f_Y|||\}$.

By definition of $X_t$ and $Y_t$, the spectral densities take the form

$$\begin{aligned} f_X(\theta) &= \mathcal{A}(e^{i\theta}) f_y(\theta) \mathcal{A}^*(e^{i\theta}), \\ f_Y(\theta) &= \mathcal{B}(e^{i\theta}) f_y(\theta) \mathcal{B}^*(e^{i\theta}), \\ f_{X,Y}(\theta) &= \mathcal{A}(e^{i\theta}) f_y(\theta) \mathcal{B}^*(e^{i\theta}). \end{aligned}$$

This implies $|||f_X||| \le |||\mathcal{A}|||^2 |||f_y|||$, $|||f_Y||| \le |||\mathcal{B}|||^2 |||f_y|||$ and $|||f_{X,Y}||| \le |||\mathcal{A}|||\,|||\mathcal{B}|||\,|||f_y||| < \infty$, so

that the above concentration bound can be applied. Plugging in these upper bounds into the above concentration inequality, we prove the first part of our proposition.

In order to prove the second part, we set $\eta = A\sqrt{(\log d^2)/T}$ and take union bound of the event in (B.2) over $d^2$ choices of $u, v \in \{e_1, \ldots, e_d\}$, the set of canonical unit vectors in $\mathbb{R}^d$. Since $T \gtrsim \log d$, we have $\min\{\eta, \eta^2\} = \eta^2$ so that the above inequality implies

$$\mathbb{P}\left[\left\|\mathcal{X}^\top \mathcal{Y}/T\right\|_\infty > \left\|\Gamma_{X,Y}(0)\right\|_\infty + 6\pi A |||f_y||| \max\left\{|||\mathcal{A}|||^2, |||\mathcal{B}|||^2\right\} \sqrt{2\log d/T}\right]$$

is at most $6d^2 \exp[-cA^2 \log d^2] = 6\exp\left[-(cA^2 - 1)\log d^2\right]$.

Next, in order to get an upper bound on $\|\Gamma_{X,Y}(0)\|_\infty$, note that

$$\begin{aligned}
\Gamma_{X,Y}(0) &= \mathrm{Cov}(\mathcal{A}(L)y_t, \mathcal{B}(L)y_t) \\
&= \sum_{\ell\geq 0}\sum_{m\geq 0} A_\ell \Gamma(\ell - m) B_m^\top \\
&= \int_{-\pi}^{\pi} \sum_{\ell\geq 0}\sum_{m\geq 0} e^{i(\ell-m)\theta} A_\ell f(\theta) B_m^\top d\theta \\
&= \sum_{m\geq 0}\left[\int_{-\pi}^{\pi}\left(\sum_{\ell\geq 0} A_\ell e^{i\ell\theta}\right) f(\theta) e^{-im\theta} d\theta\right] B_m^\top.
\end{aligned}$$

Therefore,

$$\|\Gamma_{X,Y}(0)\|_\infty \leq \|\Gamma_{X,Y}(0)\| \leq 2\pi |||\mathcal{A}||| |||f_y||| \|\mathcal{B}\|_{2,1}.$$

□

**Remark B.2.** Note that the bound on $\|\Gamma_{X,Y}(0)\|$ may be improved using information on the dependence between $X_t$ and $Y_t$. For instance, if we consider $X_t = y_{t-\ell}$ and $Y_t = y_t$, then we can expect that $\Gamma_{X,Y}(0)$, the covariance between $X_t$ and $Y_t$, will decay with larger $\ell$, but our bound does not. A tighter bound on $\|\Gamma_{X,Y}(0)\|$ can potentially be obtained using special structures of $X_t$ and $Y_t$, as in our proof of Proposition D.2.

Our second key technical ingredient will be used to provide an upper bound on the operator norm of the spectral density of a time series of the form $z_t$ in Proposition 2.1 in terms of the spectral density of $y_t$ and the linear filter used to generate $a_t$ from $y_t$. We use this to provide a finite-sample upper bound on the deviation of the sample Gram matrix in

the Phase-II regression from its population analogue.

**Proposition B.2.** *Consider a $d$-dimensional centered stable process $\{y_t\}$, and a $d\times d$ matrix-valued lag polynomial $\mathcal{C}(L)$ with finite $\|\mathcal{C}\|_{2,1}$. Then the spectral density of the $d(p+q)$-dimensional derived process*

$$z_t = \left[y_{t-1}^\top, y_{t-2}^\top, \ldots, y_{t-p}^\top, \mathcal{C}(L)y_{t-1}^\top, \ldots, \mathcal{C}(L)y_{t-q}^\top\right]^\top$$

*satisfies* $|||f_z||| \leq \left(p+q|||\mathcal{C}|||^2\right) |||f_y|||$.

*Proof of Proposition B.2.* Let $\mathcal{C}(L) = \sum_{\ell\geq 0} C_\ell L^\ell$ be a potentially infinite order $d\times d$ matrix-valued lag polynomial. The autocovariance function of the process $\{z_t\}$ takes the form

$$\Gamma_z(h) = \mathrm{Cov}(z_t, z_{t+h}) = \mathrm{Cov}\left(\begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-p} \\ \mathcal{C}(L)y_{t-1} \\ \vdots \\ \mathcal{C}(L)y_{t-q} \end{bmatrix}, \begin{bmatrix} y_{t-1+h} \\ \vdots \\ y_{t-p+h} \\ \mathcal{C}(L)y_{t-1+h} \\ \vdots \\ \mathcal{C}(L)y_{t-q+h} \end{bmatrix}\right).$$

The $d(p+q)\times d(p+q)$ matrix on the right can be partitioned into four blocks. Since $\|\Gamma_y(h)\| \leq 2\pi|||f_y||| < \infty$ for all $h\in\mathbb{Z}$, and $\|\mathcal{C}\|_{2,1} = \sum_{\ell\geq 0}\|C_\ell\| < \infty$, using dominated convergence theorem we can express the four blocks as follows.

1. *Block (1,1), size $dp\times dp$*: consists of $p^2$ submatrices of size $d\times d$ each, the $(r,s)^{th}$ submatrix given by $\mathrm{Cov}\left(y_{t-r}, y_{t-s+h}\right) = \Gamma_y(r-s+h)$, for $1\leq r,s,\leq p$;

2. *Block (1,2), size $dp\times dq$*: consists of $pq$ submatrices of size $d\times d$ each, the $(r,s)^{th}$ submatrix given by $\mathrm{Cov}\left(y_{t-r}, \sum_{\ell\geq 0} C_\ell y_{t-s+h-\ell}\right) = \sum_{\ell\geq 0}\Gamma_y(r-s+h-\ell)C_\ell^\top$, for $1\leq r\leq p$, $1\leq s\leq q$;

3. *Block (2,1), size $dq\times dp$*: consists of $pq$ submatrices of size $d\times d$ each, the $(r,s)^{th}$ submatrix given by $\mathrm{Cov}\left(\sum_{\ell\geq 0} C_\ell y_{t-r-\ell}, y_{t-s+h}\right) = \sum_{\ell\geq 0} C_\ell\Gamma_y(r-s+h+\ell)$, for $1\leq r\leq q$, $1\leq s\leq p$;

4. *Block (2,2), size $dq \times dq$*: consists of $q^2$ submatrices of size $d \times d$ each, the $(r,s)^{th}$ submatrix given by $\text{Cov}\left(\sum_{\ell \geq 0} C_\ell y_{t-r-\ell}, \sum_{\ell' \geq 0} C_{\ell'} y_{t-s+h-\ell'}\right) = \sum_{\ell,\ell' \geq 0} C_\ell \Gamma_y(h+r-s+\ell-\ell')C_{\ell'}$, for $1 \leq r, s \leq q$.

Similarly, the spectral density $f_z(\theta) = (1/2\pi)\sum_{h=-\infty}^{\infty} \Gamma_z(h)e^{-ih\theta}$, for any $\theta \in [-\pi, \pi]$ can be partitioned into four blocks as follows:

Block (1,1): the $(r,s)^{th}$ submatrix, for $1 \leq r \leq p$, $1 \leq s \leq q$, is given by

$$\frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Gamma_y(h+r-s)e^{-ih\theta} = e^{i(r-s)\theta} f_y(\theta)$$

Block (1,2): the $(r,s)^{th}$ submatrix, for $1 \leq r \leq p$, $1 \leq s \leq q$, is given by

$$\begin{aligned}
& \frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\sum_{\ell \geq 0} \Gamma_y(r-s+h-\ell)C_\ell^\top e^{-ih\theta} \\
= & \sum_{\ell \geq 0}\left[\frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Gamma_y(r-s+h-\ell)e^{-i(r-s+h-\ell)\theta}\right] C_\ell^\top e^{i(r-s-\ell)\theta} \\
= & f_y(\theta)\mathcal{C}^*(e^{i\theta})e^{i(r-s)\theta}.
\end{aligned}$$

Block (2,1): the $(r,s)^{th}$ submatrix, for $1 \leq r \leq q$, $1 \leq s \leq p$ is given by

$$\begin{aligned}
& \frac{1}{2\pi}\sum_{\ell \geq 0} C_\ell \Gamma_y(r-s+h+\ell)e^{-ih\theta} \\
= & \sum_{\ell \geq 0}\left[\frac{1}{2\pi}\sum_{h=-\infty}^{\infty} \Gamma_y(r-s+h+\ell)e^{-i(h+r-s+\ell)\theta}\right] e^{i(r-s+\ell)\theta} \\
= & e^{i(r-s)\theta}\left(\sum_{\ell \geq 0} C_\ell e^{i\ell\theta}\right) f_y(\theta) = e^{i(r-s)\theta}\mathcal{C}(e^{i\theta})f_y(\theta).
\end{aligned}$$

*Block (2,2)*: the $(r,s)^{th}$ submatrix, for $1 \le r \le q$, $1 \le s \le q$ is given by

$$\begin{aligned}
& \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \sum_{\ell,\ell' \ge 0} C_\ell \Gamma_y(h+r-s+\ell-\ell') C_{\ell'}^\top e^{-ih\theta} \\
= & \sum_{\ell,\ell' \ge 0} C_\ell \left( \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \Gamma_y(h+r-s+\ell-\ell') e^{-i(h+r-s+\ell-\ell')\theta} \right) C_{\ell'}^\top e^{i(r-s+\ell-\ell')\theta} \\
= & e^{i(r-s)\theta} \left( \sum_{\ell \ge 0} C_\ell e^{i\ell\theta} \right) f_y(\theta) \left( \sum_{\ell' \ge 0} C_{\ell'}^\top e^{-i\ell'\theta} \right) = e^{i(r-s)\theta} \mathcal{C}(e^{i\theta}) f_y(\theta) \mathcal{C}^*(e^{i\theta}).
\end{aligned}$$

Let $v_p$ and $v_q$ denote the vectors $[e^{i\theta}, \ldots, e^{ip\theta}]^\top$ and $[e^{i\theta}, \ldots, e^{iq\theta}]^\top$ respectively. Then the four blocks of $f_z(\theta)$ can be expressed as $\left(v_p v_p^*\right) \otimes f_y(\theta)$, $\left(v_p v_q^*\right) \otimes \left(f_y(\theta)\mathcal{C}^*(e^{i\theta})\right)$, $\left(v_q v_p^*\right) \otimes \left(\mathcal{C}(e^{i\theta}) f_y(\theta)\right)$ and $\left(v_q v_q^*\right) \otimes \left(\mathcal{C}(e^{i\theta}) f_y(\theta) \mathcal{C}^*(e^{i\theta})\right)$ respectively. Since $\|v_p\| = \sqrt{v_p^* v_p} = \sqrt{p}$, and $\|v_q\| = \sqrt{q}$, and $\|A \otimes B\| = \|A\|\|B\|$, by the norm compression inequality we obtain

$$\begin{aligned}
|||f_z||| &\le \left\| \begin{bmatrix} p|||f_y||| & \sqrt{pq}|||f_y|||||\mathcal{C}||| \\ \sqrt{pq}|||f_y|||||\mathcal{C}||| & q|||f_y|||||\mathcal{C}|||^2 \end{bmatrix} \right\| \\
&= |||f_y||| \left(p + q|||\mathcal{C}|||^2\right).
\end{aligned}$$

□

**Lemma B.1** (Controlling $\hat{\varepsilon}_t - \varepsilon_t$)**.** Let $\{x_t\}$ be a $d$-dimensional, centered, stable Gaussian time series, and let $z_t = \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t)$, where $\mathcal{B}(L)$ is a finite order lag polynomial of degree $q$ and $\{\hat{\varepsilon}_t - \varepsilon_t\}$, $t = 1, , \ldots, n+q$ is a sequence of $d$-dimensional random vectors satisfying $\sum_{t=1}^{n+q} \|\hat{\varepsilon}_t - \varepsilon_t\|^2/(n+q) \le \Delta_\varepsilon^2$ on an event $\mathcal{E}$ such that $\mathbb{P}(\mathcal{E}) \ge 1 - c_0 \exp[-(c_1 A^2 - 1)\log d^2\tilde{p}]$. Also, let $\{w_t\}_{t=1-j}^{n-j}$ be a sequence of random vectors given by $w_t = \hat{\varepsilon}_{t-j} - \varepsilon_{t-j}$, for some $j \in \{1, \ldots, q\}$. Consider data matrices $\mathcal{X}, \mathcal{Z}$ and $\mathcal{W}$ containing $n$ consecutive observations from the time series $x_t, z_t$ and $w_t$ respectively, and assume $n \gtrsim \log(d^2\tilde{p})$. Then there exist constants $c_i > 0$ such that for any two unit vectors $u, v \in \mathbb{S}^{d-1}$, each of the following statements holds with probability at least $1 - c_0 \exp[-(c_1 A^2 - 1)\log d^2\tilde{p}]$:

(i) $\left|u^\top \left(\mathcal{X}^\top \mathcal{Z}/n\right) v\right| \le \left[2\pi|||f_x||| \left(1 + A\sqrt{\log d^2\tilde{p}\,/n}\right)\right]^{1/2} \sqrt{(1+q/n)}\Delta_\varepsilon \|\mathcal{B}\|_{2,1}$;

(ii) $\left|u^\top \left(\mathcal{W}^\top \mathcal{Z}/n\right) v\right| \le (1+q/n)\,\Delta_\varepsilon^2 \|\mathcal{B}\|_{2,1}$.

*Proof.* In order to prove (i), note that

$$\begin{aligned} u^\top \left(\mathcal{X}^\top \mathcal{Z}/n\right) v &= \frac{1}{n}\sum_{t=1}^{n} \left(u^\top x_t\right) \left[v^\top \sum_{k=0}^{q} B_k \left(\hat{\varepsilon}_{t-k} - \varepsilon_{t-k}\right)\right] \\ &= \sum_{k=0}^{q} \frac{1}{n}\sum_{t=1}^{n} \left(u^\top x_t\right) \left(v^\top B_k(\hat{\varepsilon}_{t-k} - \varepsilon_{t-k})\right) \\ &\leq \sum_{k=0}^{q} \left[\frac{1}{n}\sum_{t=1}^{n} \left(u^\top x_t\right)^2\right]^{1/2} \left[\frac{1}{n}\sum_{t=1}^{n} \left(v^\top B_k(\hat{\varepsilon}_{t-k} - \varepsilon_{t-k})\right)^2\right]^{1/2}. \end{aligned}$$

Using the concentration inequality in Proposition 2.4 and the upper bound on the spectral norm of population covariance matrix in Proposition 2.3 of [10], square of the first term in each summand is at most $2\pi|||f_x|||(1 + A\sqrt{\log d^2\tilde{p}/n})$ with probability at least $1 - c_0 \exp\left[-(c_1 A^2 - 1)\log d^2\tilde{p}\right]$. Also, using the Cauchy-Schwarz inequality, square of the second term in the $k^{th}$ summand above satisfies, on the event $\mathcal{E}$,

$$\frac{1}{n}\sum_{t=1}^{n}\left[\left(v^\top B_k(\hat{\varepsilon}_{t-k} - \varepsilon_{t-k})\right)^2\right] \leq \frac{1}{n}\sum_{t=1}^{n} \|B_k\|^2 \|\hat{\varepsilon}_{t-k} - \varepsilon_{t-k}\|^2 = \|B_k\|^2(1 + q/n)\Delta_\varepsilon^2.$$

Together, this implies $|u^\top(\mathcal{X}^\top\mathcal{Z}/n)v|$ is upper bounded by $2\pi|||f_x|||\sqrt{1+q/n}\,(1 + A\sqrt{\log d^2\tilde{p}/n})^{1/2}\left(\sum_{k=0}^{q}\|B_k\|\right)\Delta_\varepsilon$ with the specified probability.

In order to prove (ii), note that

$$\begin{aligned} u^\top \left(\mathcal{W}^\top \mathcal{Z}/n\right) v &= \frac{1}{n}\sum_{t=1}^{n} \left(u^\top(\hat{\varepsilon}_{t-j} - \varepsilon_{t-j})\right) \left(v^\top \sum_{k=0}^{q} B_k(\hat{\varepsilon}_{t-k} - \varepsilon_{t-k})\right) \\ &\leq \sum_{k=0}^{q} \left[\frac{1}{n}\sum_{t=1}^{n} \left(u^\top(\hat{\varepsilon}_{t-j} - \varepsilon_{t-j})\right)^2\right]^{1/2} \left[\frac{1}{n}\sum_{t=1}^{n} \left(v^\top B_k(\hat{\varepsilon}_{t-k} - \varepsilon_{t-k})\right)^2\right]^{1/2}. \end{aligned}$$

Using the argument above, we can check that on the event $\mathcal{E}$, the square of the first term in each summand is at most $(1 + q/n)\Delta_\varepsilon^2$ and the square of the second term in the $k^{th}$ summand is at most $(1 + q/n)\|B_k\|^2\Delta_\varepsilon^2$. Putting things together, the right hand side of the above inequality is bounded above by $(1 + q/n)\left(\sum_{k=0}^{q}\|B_k\|\right)\Delta_\varepsilon^2$. □

**Lemma B.2.** Consider $\hat{\varepsilon}_t$ and $\Delta_\varepsilon$ as in Lemma B.1, and $\Delta_a$ as defined in (D.5). Let $\tilde{\mathcal{Z}}$ be a data matrix consisting of $n$ consecutive observations from the time series

$[y_{t-1}^\top, y_{t-2}^\top, \ldots, y_{t-p}^\top, \hat{\varepsilon}_{t-1}^\top, \hat{\varepsilon}_{t-2}^\top, \ldots, \hat{\varepsilon}_{t-q}^\top]^\top$, and $\mathcal{Z}$ a data matrix for $\{z_t\} = [y_{t-1}^\top, y_{t-2}^\top, \ldots, y_{t-p}^\top, a_{t-1}^\top, a_{t-2}^\top, \ldots, a_{t-q}^\top]^\top$. Assume $n \gtrsim \log(d^2(p+q))$ and $\tilde{p} \geq p+q$. Then there exist universal constants $c_i > 0$ such that for any $u, v \in \mathbb{S}^{d(p+q)-1}$, with probability at least $1 - c_0 \exp[-(c_1 A^2 - 1)\log d^2(p+q)]$, the following holds:

$$\begin{aligned}\left|u^\top \left(\tilde{\mathcal{Z}}^\top \tilde{\mathcal{Z}}/n - \Gamma_z(0)\right) v\right| &\leq 2\pi |||f_z|||(1 + A\sqrt{\log d^2(p+q)/n}) + q(1+q/n)\Delta_a^2 \\ &\quad + 2\left[2\pi |||f_z|||(1 + A\sqrt{\log d^2(p+q)/n}) q(1+q/n)\right]^{1/2} \Delta_a.\end{aligned}$$

*Proof.* We begin by re-writing $\tilde{z}_t$ as $z_t + w_t$, where $w_t = \left[0^\top, \ldots, 0^\top, (\hat{\varepsilon}_{t-1} - a_{t-1})^\top, \ldots, (\hat{\varepsilon}_{t-q} - a_{t-q})^\top\right]^\top$. Then the following decomposition holds:

$$\begin{aligned}\left|u^\top \left(\tilde{\mathcal{Z}}^\top \tilde{\mathcal{Z}}/n - \Gamma_z(0)\right) v\right| &\leq \left|u^\top \left(\mathcal{Z}^\top \mathcal{Z}/n - \Gamma_z(0)\right) v\right| + \left|u^\top \left(\mathcal{Z}^\top \mathcal{W}/n\right) v\right| \\ &+ \left|v^\top \left(\mathcal{Z}^\top \mathcal{W}/n\right) u\right| + \left|u^\top \left(\mathcal{W}^\top \mathcal{W}/n\right) v\right|.\end{aligned}$$

Using Proposition 2.4 in [10], we can obtain a high probability upper bound on the first term on the right hand side in terms of $|||f_z|||$. In order to control the second and third terms, assume $v = [v_1^\top, v_2^\top, \ldots, v_{p+q}^\top]^\top$, where each $v_j \in \mathbb{R}^d$, and note that

$$\begin{aligned}&\left|u^\top \left(\mathcal{Z}^\top \mathcal{W}/n\right) v\right| = \left|\frac{1}{n}\sum_{t=1}^n (u^\top z_t)(v^\top w_t)\right| \\ \leq\ &\left[\frac{1}{n}\sum_{t=1}^n (u^\top z_t)^2\right]^{1/2} \left[\frac{1}{n}\sum_{t=1}^n (v^\top w_t)^2\right]^{1/2} \\ \leq\ &\left[2\pi |||f_z|||(1 + A\sqrt{\log d(p+q)/n})\right]^{1/2} \left[\sum_{k=0}^q \frac{1}{n}\sum_{t=1}^n \left(v_{p+k}^\top (\hat{\varepsilon}_{t-k} - a_{t-k})\right)^2\right]^{1/2} \\ \leq\ &\left[2\pi |||f_z|||(1 + A\sqrt{\log d(p+q)/n})\right]^{1/2} \left[\sum_{k=0}^q (1+q/n)\|v_{p+k}\|^2 \Delta_a^2\right]^{1/2}.\end{aligned}$$

The result follows by using the fact that on the event $\mathcal{E}$, square of the second term in the above product is upper bounded by $q(1+q/n)\Delta_a^2$, and noting that $u^\top(\mathcal{W}^\top \mathcal{W}/n)v = \frac{1}{n}\sum_{t=1}^n (u^\top w_t)(v^\top w_t) \leq q(1+q/n)\Delta_a^2$. □

# C Proof of Proposition 4.1 (Elastic Net)

*Proof of Proposition 4.1.* Set $\hat{\beta} \leftarrow \hat{\beta}^{(\alpha)}$, and define $\beta_P^* := \mathcal{P}_{\mathcal{S}}(\hat{\beta})$, the projection of $\hat{\beta}$ onto the affine space $\mathcal{S} := \{\beta : \Sigma\beta = \rho\}$. Note that $\beta^* \in \mathcal{S}$. Set $v = \hat{\beta} - \beta^*$, $v_1 = \hat{\beta} - \beta_P^*$, $v_2 = \beta_P^* - \beta^*$. Then $v_2 \in \mathcal{N}(\Sigma)$, $v_1 \perp \mathcal{N}(\Sigma)$, and $\|v\|^2 = \|v_1\|^2 + \|v_2\|^2$. Consider

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^{\bar{d}}}{\operatorname{argmin}} \frac{1}{n}\|Y - X\beta\|^2 + \lambda\left(\|\beta\|_1 + \frac{\alpha}{2}\|\beta\|^2\right) \text{ subject to } \|\beta\|_1 \le M$$

where $M \ge \|\beta^*\|_1$.

Start with the basic inequality

$$\frac{1}{n}\left\|Y - X\hat{\beta}\right\|^2 + \lambda\left(\|\hat{\beta}\|_1 + \frac{\alpha}{2}\|\hat{\beta}\|^2\right) \le \frac{1}{n}\|Y - X\beta^*\|^2 + \lambda\left(\|\beta^*\|_1 + \frac{\alpha}{2}\|\beta^*\|^2\right)$$

This implies

$$\frac{1}{n}\|Xv\|^2 - \frac{2}{n}v^\top X^\top \varepsilon \le \lambda\left[(\|\beta^*\|_1 - \|\beta^* + v\|_1) + \frac{\alpha}{2}\left(\|\beta^*\|^2 - \|\beta^* + v\|^2\right)\right]$$

Since $\|X^\top\varepsilon/n\|_\infty \le \lambda/2$, moving the second term to the right we get

$$v^\top\left(X^\top X/n\right)v \le \lambda\|v\|_1 + \lambda\left[(\|\beta^*\|_1 - \|\beta^* + v\|_1) + \frac{\alpha}{2}\left(\|\beta^*\|^2 - \|\beta^* + v\|^2\right)\right],$$

which in turn implies, by triangle inequality,

$$v^\top\left(X^\top X/n\right)v \le \lambda\left[2\|\beta^*\|_1 + \frac{\alpha}{2}\|\beta^*\|^2\right] \le \lambda\left[2M + \alpha M^2/2\right].$$

This implies

$$\begin{aligned} v^\top\Sigma v &= v^\top\left(\Sigma - X^\top X/n\right)v + v^\top\left(X^\top X/n\right)v \\ &\le \left\|\Sigma - X^\top X/n\right\|_\infty \|v\|_1^2 + \lambda\left[2M + \alpha M^2/2\right] \\ &\le 4q_n M^2 + \lambda\left[2M + \alpha M^2/2\right], \text{ since } \|v\|_1 \le \|\hat{\beta}\|_1 + \|\beta^*\|_1 \le 2M. \end{aligned}$$

By the orthogonal decomposition $v = v_1 + v_2$, we have

$$v^\top \Sigma v = v_1^\top \Sigma v_1 \geq \Lambda_{\min}^+(\Sigma) \|v_1\|^2.$$

Combining the above two inequalities,

$$\|v_1\|^2 = \|\hat{\beta} - \beta_P^*\|^2 \leq \frac{4q_n M^2 + \lambda\left[2M + \alpha M^2/2\right]}{\Lambda_{\min}^+(\Sigma)}.$$

We restate the point identification result of part (c) in the form of a complete proposition C.1. □

**Proposition C.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 $\Sigma$. Consider the linear regression model $y_{N\times 1} = X_{N\times D}\beta_{D\times 1}^{*(\alpha)} + \varepsilon_{N\times 1}$ with identified target*

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

*where $\mathcal{P}_\alpha(\beta) := \|\beta\|_1 + (\alpha/2)\|\beta\|^2$, and let*

$$\hat{\beta}^{(\alpha)} := \underset{\beta}{\operatorname{argmin}} \left\{\mathcal{P}_\alpha(\beta) \text{ s.t. } \frac{1}{n}\|y - X\beta\|^2 \leq A_n, \quad \|\beta\|_1 \leq M\right\}$$

*be the estimator. On the event*

$$\mathcal{E} := \left\{\left\|X^\top X/n - \Sigma\right\|_\infty \leq q_n, \frac{1}{n}\left\|X^\top \varepsilon\right\|_\infty \leq r_n, \left|\frac{1}{n}\|\varepsilon\|^2 - \sigma^2\right| \leq s_n\right\}$$

*and choosing $A_n = \sigma^2 + s_n$ and $M \geq \|\beta^{*(\alpha)}\|_1$, the following holds:*

$$\left\|\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)}\right\|^2 \leq 2v_n + 2(\sqrt{D}/\alpha + M)v_n^{1/2},$$

*where $v_n := \frac{4Mr_n + 2s_n + 4M^2 q_n}{\Lambda_{\min}^+(\Sigma)}$ and $\Lambda_{\min}^+(\Sigma)$ is the smallest non-zero eigenvalue of $\Sigma$.*

*Proof of Proposition C.1.* The estimator can be written as

$$\hat{\beta}^{(\alpha)} := \underset{\beta}{\operatorname{argmin}} \left\{ \mathcal{P}_\alpha(\beta) \text{ s.t. } \beta \in \mathcal{A}_n, \quad \|\beta\|_1 \leq M \right\}, \tag{C.1}$$

where $\mathcal{A}_n = \{\beta : \frac{1}{n}\|y - X\beta\|^2 \leq A_n\}$. Our proof consists of a series of lemmas. We begin by relating the estimator's constraint set to the equivalence class of parameters that could have generated the data.

**Lemma C.1.** If $A_n \geq \sigma^2 + s_n$, then $\beta^{*(\alpha)} \in \mathcal{A}_n$ on the event $\mathcal{E}$.

*Proof.* By the triangle inequality,

$$\frac{1}{n}\|y - X\beta^{*(\alpha)}\|^2 = \frac{1}{n}\|\varepsilon\|^2 \leq \sigma^2 + s_n.$$

□

Our next lemma is a result about our estimator's in-sample prediction performance.

**Lemma C.2.** If we choose $A_n = \sigma^2 + s_n$, then on the event $\mathcal{E}$,

$$\frac{1}{n}\|X(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 4Mr_n + 2s_n.$$

*Proof.* We rewrite the inequality $\frac{1}{n}\|y - X\hat{\beta}^{(\alpha)}\|^2 \leq A_n$ as

$$\|X(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq n(A_n - \frac{1}{n}\|\varepsilon\|^2) + 2\varepsilon^\top X(\beta^{*(\alpha)} - \hat{\beta}^{(\alpha)}).$$

Our choice of $A_n$ means that

$$A_n - \frac{1}{n}\|\varepsilon\|^2 = \sigma^2 + s_n - \frac{1}{n}\|\varepsilon\|^2 \leq \left|\sigma^2 - \frac{1}{n}\|\varepsilon\|^2\right| + s_n.$$

On the event $\mathcal{E}$, we know that this is bounded by $2s_n$. Thus,

$$\|X(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 2ns_n + 2\varepsilon^\top X(\beta^{*(\alpha)} - \hat{\beta}^{(\alpha)}).$$

Furthermore,

$$\|X(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 2ns_n + 2\|X^\top \varepsilon\|_\infty \cdot \|\beta^{*(\alpha)} - \hat{\beta}^{(\alpha)}\|_1.$$

Dividing both sides by $n$ and recalling the definition of $r_n$ (through the event $\mathcal{E}$) gives

$$\frac{1}{n}\|X(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 2s_n + 2r_n\|\beta^{*(\alpha)} - \hat{\beta}^{(\alpha)}\|_1.$$

The triangle inequality and recalling that both vectors are bounded by $M$ in $\ell_1$ norm gives

$$\frac{1}{n}\|X(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 4Mr_n + 2s_n.$$

□

Our next lemma extends this prediction result from $X$ to $\Sigma^{1/2}$.

**Lemma C.3.** If we choose $A_n = \sigma^2 + s_n$, then on the event $\mathcal{E}$,

$$\|\Sigma^{1/2}(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 4Mr_n + 2s_n + 4M^2 q_n.$$

*Proof.* Writing

$$\|\Sigma^{1/2}(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 = \frac{1}{n}\|X(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 + (\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})^\top(\Sigma - \frac{1}{n}X^T X)(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)}),$$

we apply Lemma C.2 to get

$$\|\Sigma^{1/2}(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 4Mr_n + 2s_n + (\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})^\top(\Sigma - \frac{1}{n}X^T X)(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)}).$$

Now, for a matrix $A$, $v^\top A v = \sum_{ij} v_i A_{ij} v_j \leq \|A\|_\infty \sum_{ij} |v_i||v_j| = \|A\|_\infty \|v\|_1^2$, and recalling the definition of $q_n$ (through the event $\mathcal{E}$) we have

$$\|\Sigma^{1/2}(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \leq 4Mr_n + 2s_n + q_n\|\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)}\|_1^2.$$

The result follows by the triangle inequality and that both vectors are bounded by $M$ in $\ell_1$

norm. □

At this point, we move from prediction bounds to estimation bounds. Our next step is to translate the previous result to a statement about our estimator not being too far from the set of possible parameters that generated our data, that is the affine space $\{\beta : \Sigma\beta = \rho\}$.

**Lemma C.4.** Let $\hat{\beta}_P^{(\alpha)}$ denote the projection of $\hat{\beta}^{(\alpha)}$ onto the affine subspace $\{\beta : \Sigma\beta = \rho\}$:

$$\hat{\beta}_P^{(\alpha)} := \arg\min_{\beta} \left\{ \|\hat{\beta}^{(\alpha)} - \beta\|^2 \text{ s.t. } \Sigma\beta = \rho \right\}.$$

If we choose $A_n = \sigma^2 + s_n$, then on the event $\mathcal{E}$,

$$\|\hat{\beta}^{(\alpha)} - \hat{\beta}_P^{(\alpha)}\|^2 \leq v_n,$$

where

$$v_n := \frac{4Mr_n + 2s_n + 4M^2 q_n}{\Lambda_{\min}^+(\Sigma)}$$

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

*Proof.* The distance of $\hat{\beta}^{(\alpha)}$ to the affine space is given by

$$\begin{aligned}
\|\hat{\beta}^{(\alpha)} - \hat{\beta}_P^{(\alpha)}\|^2 &= \min_{\beta} \left\{ \|\hat{\beta}^{(\alpha)} - \beta\|^2 \text{ s.t. } \Sigma\beta = \rho \right\} \\
&= \min_{\delta} \left\{ \|\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)} - \delta\|^2 \text{ s.t. } \Sigma\delta = 0 \right\} \\
&= \|\Sigma\Sigma^+(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2
\end{aligned}$$

where in the second equality we use that $\Sigma\beta^{*(\alpha)} = \rho$ and in the third equality we use that the row space and null space are orthogonal complements and therefore the residual after projecting onto the null space is equivalent to the projection onto the row space of $\Sigma$ (and

here the row space and column space are identical). Now, $\Sigma\Sigma^+ = (\Sigma^{1/2})^+\Sigma^{1/2}$ and so

$$\begin{aligned}\|\hat{\beta}^{(\alpha)} - \hat{\beta}_P^{(\alpha)}\|^2 &= \|(\Sigma^{1/2})^+\Sigma^{1/2}(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \\ &\leq \|(\Sigma^{1/2})^+\|^2\|\Sigma^{1/2}(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2 \\ &\leq \|\Sigma^{1/2}(\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)})\|^2/\Lambda^+_{\min}(\Sigma).\end{aligned}$$

The result follows from the previous lemma. □

At this point, we have bounded the distance between our estimator and the identified target in the direction orthogonal to the affine space. The remainder of the proof of the proposition is aimed at bounding the distance along the affine space. To do so, we make use of the strong convexity of the objective function $\mathcal{P}_\alpha$.

**Lemma C.5.** Under the same setup and conditions as the previous lemma,

$$\mathcal{P}_\alpha(\hat{\beta}_P^{(\alpha)}) - \mathcal{P}_\alpha(\hat{\beta}^{(\alpha)}) \leq (\sqrt{D} + \alpha M)v_n^{1/2} + \frac{\alpha}{2}v_n,$$

where $v_n$ is defined in Lemma C.4.

*Proof.* By the triangle inequality, $\|\hat{\beta}_P^{(\alpha)}\|_1 - \|\hat{\beta}^{(\alpha)}\|_1 \leq \|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|_1$ and $\|\hat{\beta}_P^{(\alpha)}\| \leq \|\hat{\beta}^{(\alpha)}\| + \|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|$. Squaring this second inequality gives

$$\begin{aligned}\|\hat{\beta}_P^{(\alpha)}\|^2 &\leq \|\hat{\beta}^{(\alpha)}\|^2 + \|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|^2 + 2\|\hat{\beta}^{(\alpha)}\| \cdot \|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\| \\ &\leq \|\hat{\beta}^{(\alpha)}\|^2 + \|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|^2 + 2M\|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|\end{aligned}$$

Thus,

$$\begin{aligned}\mathcal{P}_\alpha(\hat{\beta}_P^{(\alpha)}) - \mathcal{P}_\alpha(\hat{\beta}^{(\alpha)}) &\leq \|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|_1 + \frac{\alpha}{2}\left(\|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|^2 + 2M\|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|\right) \\ &\leq (\sqrt{D} + \alpha M)\|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\| + \frac{\alpha}{2}\|\hat{\beta}_P^{(\alpha)} - \hat{\beta}^{(\alpha)}\|^2.\end{aligned}$$

The result follows from the previous lemma. □

**Lemma C.6.** Let $\hat{\beta}_P^{(\alpha)}$ be the projection of $\hat{\beta}^{(\alpha)}$ onto $\{\beta : \Sigma\beta = \rho\}$. If we choose $A_n = \sigma^2 + s_n$ and $M \geq \|\beta^{*(\alpha)}\|_1$, then on the event $\mathcal{E}$,

$$\|\hat{\beta}_P^{(\alpha)} - \beta^{*(\alpha)}\|^2 \leq 2(\sqrt{D}/\alpha + M)v_n^{1/2} + v_n,$$

where $v_n := \frac{4Mr_n + 2s_n + 4M^2 q_n}{\Lambda_{\min}^+(\Sigma)}$.

*Proof.* By $\alpha$-strong convexity of $\mathcal{P}_\alpha$ and the definition of $\beta^{*(\alpha)}$, we have that for any $\gamma$ such that $\Sigma\gamma = \rho$,

$$\mathcal{P}_\alpha(\gamma) \geq \mathcal{P}_\alpha(\beta^{*(\alpha)}) + \frac{\alpha}{2}\|\gamma - \beta^{*(\alpha)}\|^2$$

Substituting $\hat{\beta}_P^{(\alpha)}$ for $\gamma$ and rearranging terms gives

$$\|\hat{\beta}_P^{(\alpha)} - \beta^{*(\alpha)}\|^2 \leq (2/\alpha)\left[\mathcal{P}_\alpha(\hat{\beta}_P^{(\alpha)}) - \mathcal{P}_\alpha(\beta^{*(\alpha)})\right].$$

By Lemma C.1, $\beta^{*(\alpha)} \in \mathcal{A}_n$ and by assumption $\|\beta^{*(\alpha)}\|_1 \leq M$, thus $\beta^{*(\alpha)}$ is feasible for (C.1), meaning that $\mathcal{P}_\alpha(\hat{\beta}^{(\alpha)}) \leq \mathcal{P}_\alpha(\beta^{*(\alpha)})$. Thus,

$$\|\hat{\beta}_P^{(\alpha)} - \beta^{*(\alpha)}\|^2 \leq (2/\alpha)\left[\mathcal{P}_\alpha(\hat{\beta}_P^{(\alpha)}) - \mathcal{P}_\alpha(\hat{\beta}^{(\alpha)})\right].$$

We apply the previous lemma to the right-hand side to conclude the proof. □

The results of Lemma C.4 and Lemma C.6 can now be combined to give the desired estimation result:

$$\begin{aligned}
\|\hat{\beta}^{(\alpha)} - \beta^{*(\alpha)}\|^2 &= \|\hat{\beta}^{(\alpha)} - \hat{\beta}_P^{(\alpha)}\|^2 + \|\hat{\beta}_P^{(\alpha)} - \beta^{*(\alpha)}\|^2 \\
&\leq v_n + 2(\sqrt{D}/\alpha + M)v_n^{1/2} + v_n \\
&\leq 2v_n + 2(\sqrt{D}/\alpha + M)v_n^{1/2}.
\end{aligned}$$

This establishes the proposition. □

# D Proof of Proposition 4.2 (Phase-I Analysis)

We divide the proof of Proposition 4.2 in four steps. First, in Proposition D.1, we provide deterministic upper bounds on the estimation errors $(\widehat{\Pi}-\Pi)$ and approximation errors around the *regression* residuals $(\hat{\varepsilon}_t-\varepsilon_t)$ for a given realization of $(T+\tilde{p})$ consecutive observations from the VARMA process, under some sufficient conditions. Then we show in Propositions D.2 and D.3 that for a random realization from the VARMA process, these conditions are satisfied with high probability when the sample size is sufficiently large. Finally, we provide upper bound on the approximation errors around the *true* VARMA errors $(\hat{\varepsilon}_t - a_t)$ in Proposition D.4.

We start with the deterministic upper bound on the deviation of the estimated residuals $\hat{\varepsilon}_t$ around $\varepsilon_t$ without making any assumption on the design matrix $Z$. This is essentially a so-called "slow rate" bound, as appears in the lasso regression literature [26]. Then we provide a tighter upper bound on the above deviation, and an upper bound on the deviation of $\{\widehat{\Pi}_\tau\}_{\tau=1}^{\tilde{p}}$ around $\{\Pi_\tau\}_{\tau=1}^{\tilde{p}}$, under a restricted eigenvalue (RE) condition [37, 10]:

**Assumption D.1** (Restricted Eigenvalue, RE)**.** *A symmetric matrix $G_{r\times r}$ satisfies the restricted eigenvalue (RE) condition with curvature $\gamma > 0$ and tolerance $\delta > 0$ if*

$$v^\top G v \geq \gamma\|v\|^2 - \delta\|v\|_1^2, \qquad \textit{for all} \quad v \in \mathbb{R}^r. \tag{D.1}$$

These upper bounds involve the curvature and tolerance parameters $\gamma$, $\delta$ as well as the quantity $\|Z^\top\mathcal{E}/T\|_\infty$, and do not relate directly to the VARMA parameters. Propositions D.2 and D.3 then provide insight into how these quantities depend on VARMA parameters, when we have a random realization from a stable, invertible VARMA model (1.1).

**Proposition D.1.** *Consider any solution $\hat{\beta}$ of (4.2) using a given realization of $\{y_t\}_{t=1-\tilde{p}}^T$ from the VARMA model (1.1), and set $\mathcal{E} = vec(E)$. Then, for any choice of the penalty parameter $\lambda \geq 2\left\|Z^\top\mathcal{E}/T\right\|_\infty$, we have*

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

*Further, assume* $\{\Pi_1, \ldots, \Pi_{\tilde{p}}\}$ *are sparse so that* $k := \sum_{\tau=1}^{\tilde{p}} \|\Pi_\tau\|_0$*, and the sample Gram matrix* $Z^\top Z/T$ *satisfies* $RE(\gamma, \delta)$ *of Assumption D.1 for some model dependent quantities* $\gamma > 0, \delta > 0$ *such that* $k\delta \leq \gamma/32$*. Then for any choice of* $\lambda \geq 4\left\|Z^\top \mathcal{E}/T\right\|_\infty$*, we have the following upper bounds*

$$\sum_{\tau=1}^{\tilde{p}} \left\|\widehat{\Pi}_\tau - \Pi_\tau\right\|_1 \leq 64k\lambda/\gamma, \left[\sum_{\tau=1}^{\tilde{p}} \left\|\widehat{\Pi}_\tau - \Pi_\tau\right\|_F^2\right]^{1/2} \leq 16\sqrt{k}\lambda/\gamma,$$
$$\frac{1}{T}\sum_{t=1}^{T} \|\hat{\varepsilon}_t - \varepsilon_t\|^2 \leq 128k\lambda^2/\gamma =: \Delta_\varepsilon^2. \tag{D.3}$$

*Proof of Proposition D.1.* Since $\hat{\beta}$ is a minimizer of (4.2), we have

$$\frac{1}{T}\left\|Y - Z\hat{\beta}\right\|^2 + \lambda\left\|\hat{\beta}\right\|_1 \leq \frac{1}{T}\|Y - Z\beta^*\|^2 + \lambda\|\beta^*\|_1.$$

Let $v = \hat{\beta} - \beta^*$ denote the error vector. Substituting $Y = Z\beta^* + \mathcal{E}$ in the above, we obtain

$$\frac{1}{T}\|\mathcal{E} - Zv\|^2 + \lambda\|\beta^* + v\|_1 \leq \frac{1}{T}\|\mathcal{E}\|^2 + \lambda\|\beta^*\|_1.$$

Moving some terms to the right hand side of the inequality, we get

$$v^\top\left(Z^\top Z/T\right)v \leq 2v^\top\left(Z^\top\mathcal{E}/T\right) + \lambda\left(\|\beta^*\|_1 - \|\beta^* + v\|_1\right). \tag{D.4}$$

Since $\lambda \geq 2\|Z^\top\mathcal{E}/T\|_\infty$, and the first term on the right is at most $2\|v\|_1\|Z^\top\mathcal{E}/T\|_\infty$, we have

$$v^\top\left(Z^\top Z/T\right)v \leq \lambda\left(\|v\|_1 + \|\beta^*\|_1 - \|\beta^* + v\|_1\right) \leq 2\lambda\|\beta^*\|_1 = 2\lambda\sum_{\tau=1}^{\tilde{p}}\|\Pi_\tau\|_1.$$

Then (D.2) follows from the fact that

$$v^\top\left(Z^\top Z/T\right)v = \frac{1}{T}\left\|\mathcal{X}\widehat{B} - \mathcal{X}B\right\|_F^2 = \frac{1}{T}\sum_{t=1}^{T}\|\hat{\varepsilon}_t - \varepsilon_t\|^2.$$

Next, suppose $J$ denotes the support of $\beta^*$, i.e. $J = \left\{j \in \{1, \ldots, d^2\tilde{p}\} : \beta_j^* \neq 0\right\}$. By our

assumption, $|J| \leq k$. Inequality (D.4), together with our choice of $\lambda$, then leads to

$$\begin{aligned} 0 \leq v^\top \left(Z^\top Z/T\right) v &\leq \frac{\lambda}{2}\left(\|v_J\|_1 + \|v_{J^c}\|_1\right) + \lambda\left(\|\beta_J^*\|_1 - \|\beta_J^* + v_J\|_1 - \|v_{J^c}\|_1\right) \\ &\leq \frac{\lambda}{2}\left(\|v_J\|_1 + \|v_{J^c}\|_1\right) + \lambda\left(\|v_J\|_1 - \|v_{J^c}\|_1\right) \\ &\leq \frac{3\lambda}{2}\|v_J\|_1 - \frac{\lambda}{2}\|v_{J^c}\|_1 \leq 2\lambda\|v_J\|_1 \leq 2\lambda\|v\|_1. \end{aligned}$$

Since $\lambda > 0$, the first inequality on the last line ensures $\|v_{J^c}\|_1 \leq 3\|v_J\|_1$, so that $\|v\|_1 \leq 4\|v_J\|_1 \leq 4\sqrt{k}\|v\|$. Using the RE condition (D.1) and the upper bound on $k\delta$, we have

$$v^\top \left(Z^\top Z/T\right) v \geq \gamma\|v\|^2 - \delta\|v\|_1^2 \geq (\gamma - 16k\delta)\|v\|^2 \geq \frac{\gamma}{2}\|v\|^2.$$

Combining these upper and lower bounds on $v^\top \left(Z^\top Z/T\right) v$, we obtain the final inequalities as follows:

$$\begin{aligned} & \gamma\|v\|^2/2 \leq v^\top \left(Z^\top Z/T\right) v \leq 8\lambda\sqrt{k}\|v\| \\ \Rightarrow \quad & \|v\| \leq 16\lambda\sqrt{k}/\gamma, \end{aligned}$$

and consequently $\|v\|_1 \leq 4\sqrt{k}\|v\| \leq 64k\lambda/\gamma$.

Together with $v^\top \left(Z^\top Z/T\right) v \leq 2\lambda\|v\|_1$, we obtain the final in-sample prediction error bound $128k\lambda^2/\gamma$. □

Our next proposition provides a non-asymptotic upper bound on $\|Z^\top \mathcal{E}/T\|_\infty$ which holds with high probability for large $d, \tilde{p}$. If $\lambda$ is chosen as the same order of this bound, Proposition D.1 then shows how the upper bounds of estimation and approximation errors vary with model parameters.

**Proposition D.2** (Deviation Condition: Phase-I)**.** *If $\{y_{-(\tilde{p}-1)}, \ldots, y_T\}$ is a random realization from a stable, invertible VARMA model (1.1), then there exist universal constants $c_i > 0$ such that for any $A > 1$, with probability at least $1 - c_0 \exp[-(c_1 A^2 - 1)\log d^2\tilde{p}]$,*

$$\|Z^\top \mathcal{E}/T\|_\infty \leq 2\pi |||f_y||| \left[3A \max\left\{|||\Pi_{[\tilde{p}]}|||^2, 1\right\}\sqrt{\log(d^2\tilde{p})/T} + \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}\right].$$

*Proof of Proposition D.2.* Note that $\|\tilde{Z}^\top \mathcal{E}/T\|_\infty = \|\mathcal{X}^\top E/T\|_\infty = \max_{1\le h\le \tilde{p}} \|\mathcal{X}_{(h)}^\top E/T\|_\infty$, where $\mathcal{X}_{(h)} = [(y_{T-h}) : \ldots : (y_{1-h})]^\top$.

Define $X_t = y_{t-h} = L^h y_t$ and $Y_t = \varepsilon_t = a_t + \sum_{\tau=\tilde{p}+1}^{\infty} \Pi_\tau y_{t-\tau} = \Pi_{[\tilde{p}]}(L) y_t$. The first term in our upper bound follows from (B.1) in Proposition B.1, by using $X_t = L^h y_t$, $Y_t = \varepsilon_t = \Pi_{[\tilde{p}]}(L) y_t$ and $\eta = A\sqrt{\log d^2\tilde{p}/T}$. To obtain the second term, i.e. the bound on the bias term $\Gamma_{X,Y}(0)$, we use the representation $Y_t = \varepsilon_t = a_t + \sum_{t=\tilde{p}+1}^{\infty} \Pi_\tau y_{t-\tau}$ as follows:

$$\Gamma_{X,Y}(0) = \mathrm{Cov}\left(y_{t-h}, a_t + \sum_{\tau=\tilde{p}+1}^{\infty} \Pi_\tau y_{t-\tau}\right) = \sum_{\tau=\tilde{p}+1}^{\infty} \Gamma_y(h-\tau)\Pi_\tau^\top.$$

First, note that the entries of $\Gamma_{X,Y}(0)$ are upper bounded as follows:

$$\|\Gamma_{X,Y}(0)\|_\infty \le \|\Gamma_{X,Y}(0)\| \le \left(\max_{h\in\mathbb{Z}} \|\Gamma_y(h)\|\right) \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1} \le 2\pi \|\!|\!| f_y |\!|\!| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}$$

The last inequality holds since for any $h \in \mathbb{Z}$, $\Gamma_y(h) = \int_{-\pi}^{\pi} e^{ih\theta} f_y(\theta) d\theta$.

□

The next proposition investigates sample size requirements for the RE condition to hold with high probability, and also provides insight into how the tolerance and curvature parameters depend on the VARMA model parameters.

**Proposition D.3** (Verifying Restricted Eigenvalue Condition)**.** *Consider a random realization of $(T+\tilde{p})$ data points $\{y_{-(\tilde{p}-1)}, \ldots, y_T\}$ from a stable, invertible VARMA model (1.1) with $\Lambda_{\min}(\Sigma_a) > 0$. Then there exist universal constants $c_i > 0$ such that for $T \gtrsim \max\{\omega^2, 1\} k(\log d + \log\tilde{p})$, the matrix $Z^\top Z/T$ satisfies RE($\gamma, \delta$) with probability at least $1 - c_1 \exp(-c_2 T min\{\omega^{-2}, 1\})$, where*

$$\gamma = \pi / \|\!|\!| f_y^{-1} |\!|\!|, \; \omega = c_3 \tilde{p} \|\!|\!| f_y |\!|\!| \|\!|\!| f_y^{-1} |\!|\!|, \; \delta = \gamma \max\{\omega^2, 1\} \log(d\tilde{p})/T.$$

*Proof of Proposition D.3.* The proof follows along the same line of arguments as in Proposition 4.2 of [10], where the restricted eigenvalue condition was verified for processes $\{y_t\}$ generated according to a *finite-order* VAR process. In particular, rows of the design matrix were generated from the process $\tilde{y}_t = [y_t^\top, \ldots, y_{t-\tilde{p}+1}^\top]^\top$ allowing a VAR(1) representation

with closed form expressions of spectral density and autocovariance. In the present context, $\{\tilde{y}_t\}$ does not have a VAR representation. However, a close inspection of the proof in [10] shows that it is sufficient to derive a lower bound on $\Lambda_{\min}(\Gamma_{\tilde{y}}(0))$ and an upper bound on $|||f_{\tilde{y}}|||$, and the rest of the argument follows. Next, we derive these two bounds for the process $\{\tilde{y}_t\}$.

First we consider $\Lambda_{\min}(\Gamma_{\tilde{y}}(0))$. Note that $\Gamma_{\tilde{y}}(0)$ can be viewed as the variance-covariance of a vectorized data matrix containing $\tilde{p}$ consecutive observations from the process $y_t$. Hence, using Proposition 2.3 and Equation (2.6) of [10], we can show that

$$\Lambda_{\min}\left(\Gamma_{\tilde{y}}(0)\right) \geq \min_{\theta\in[-\pi,\pi]} 2\pi\Lambda_{\min}\left(f_y(\theta)\right) = 2\pi\left|\left|\left|f_y^{-1}\right|\right|\right|^{-1}.$$

The upper bound on $|||f_{\tilde{y}}|||$ follows from Proposition B.2, by setting $\mathcal{C}(L) = 0$, which implies $|||f_{\tilde{y}}||| \leq \tilde{p}|||f_y|||$. □

**Proposition D.4.** *Consider the Phase-I regression residuals $\hat{\varepsilon}_t$ in Proposition D.1. Assume $(1/T)\sum_{t=1}^{T}\|\hat{\varepsilon}_t - \varepsilon_t\|^2 \leq \Delta_\varepsilon^2$ with probability at least $1 - c_0\exp[-(c_1A^2-1)\log d^2\tilde{p}]$ for some universal constants $c_i > 0$, and $T \gtrsim \log(d^2\tilde{p})$. Then there exist $c_i > 0$ such that*

*(a) For any $v \in \mathbb{S}^{d-1}$, with probability at least $1 - c_0\exp[-(c_1A^2-1)\log(d^2\tilde{p})]$,*

$$\frac{1}{T}\sum_{t=1}^{T}\left(v^\top(\hat{\varepsilon}_t - a_t)\right)^2 \leq 4\max\left\{\Delta_\varepsilon^2, 4\pi\left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2|||f_y|||\right\} =: \Delta_a^2. \tag{D.5}$$

*(b) In particular, with probability at least $1 - c_0\exp[-(c_1A^2-2)\log(d^2\tilde{p})]$,*

$$\max_{1\leq j\leq d}\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\varepsilon}_{tj} - a_{tj}\right)^2 \leq \Delta_a^2. \tag{D.6}$$

*(c) With probability at least $1 - c_0\exp[-(c_1A^2-2)\log(d^2\tilde{p})]$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\hat{\varepsilon}_t - a_t\|^2 \leq 4\max\left\{\Delta_\varepsilon^2, 4\pi d\left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2|||f_y|||\right\}. \tag{D.7}$$

*Proof.* We use the decomposition $\hat{\varepsilon}_t - a_t = (\hat{\varepsilon}_t - \varepsilon_t) + (\varepsilon_t - a_t)$ and analyze the sum of squares for the two parts separately. In particular, note that

$$\frac{1}{T}\sum_{t=1}^{T}\left(v^\top(\hat{\varepsilon}_t - a_t)\right)^2 \le 4\max\left\{\frac{1}{T}\sum_{t=1}^{T}\left(v^\top(\hat{\varepsilon}_t - \varepsilon_t)\right)^2, \frac{1}{T}\sum_{t=1}^{T}\left(v^\top(\varepsilon_t - a_t)\right)^2\right\}.$$

By assumption, the first part is at most $\Delta_\varepsilon^2$ with probability at least $1 - c_0\exp\left[-(c_1A^2 - 1)\log d^2\tilde{p}\right]$. To work with the second part, note that $\varepsilon_t - a_t = \Pi_{-[\tilde{p}]}(L)y_t =: w_t$, say. The spectral density of $w_t$ satisfies $|||f_w||| \le \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}|||f_y|||$. Using Propositions 2.3 and 2.4 of [10], we obtain the following upper bound for any $\eta > 0$,

$$\mathbb{P}\left[v^\top\left(\frac{1}{T}w_t w_t^\top\right)v > 2\pi|||f_w|||(1+\eta)\right] \le 2\exp\left[-c_0 T\min\left\{\eta, \eta^2\right\}\right].$$

Setting $\eta = (c_1A^2 - 1)\log d^2\tilde{p}/T$ (note that $\eta < 1$ when $T \gtrsim \log d^2\tilde{p}$), we conclude that the second term is at most $4\pi|||f_w|||$ with probability at least $1 - 2\exp\left[-(c_1A^2 - 1)\log d^2\tilde{p}\right]$.

The second inequality (D.6) follows by taking an union bound over the choices $v = e_1, \ldots, e_d$, the unit vectors, and multiplying the tail probability by $d^2\tilde{p} \ge d$. The third inequality (D.7) follows by adding up these $d$ terms corresponding to the $d$ unit vectors. □

*Proof of Proposition 4.2.* The slow rate bounds follow from Propositions D.1, D.2 and D.4. To establish the fast rate, note that by Proposition D.3, the RE condition with $\gamma = \pi/\left|||f_y^{-1}|||\right.$, $\omega = c_3\tilde{p}|||f_y|||\left|||f_y^{-1}|||\right.$ and $\delta = \gamma\max\{\omega^2, 1\}\log(d\tilde{p})/T$ holds with probability at least $1 - c_1\exp\left[-c_2 k\log(d\tilde{p})\right]$, for $T \gtrsim \max\{\omega^2, 1\}k(\log d\tilde{p})$. Since $k(\log d\tilde{p}) \ge \log(d^2\tilde{p})$ for $k \ge 2$, the event where both RE and deviation condition of Proposition D.2 hold has probability at least $1 - c_0\exp\left[-(c_1A^2 - 1)\log(d^2\tilde{p})\right]$ for some universal constants $c_i > 0$. These choices also ensure $k\delta/\gamma = \max\{\omega^2, 1\}\log(d\tilde{p})/T \le 1/32$ for large enough $T$. Plugging in the value of $\gamma$ in the final inequality of Proposition D.1 leads to the tighter upper bound $\Delta_\varepsilon^2 = 128k\lambda^2/\gamma = (128/\pi)\left|||f_y^{-1}|||\right.k\lambda^2$. □

# E Propositions and Proofs for Phase-II Analysis

Before presenting the proof of 4.3, in Proposition E.1 we provide a high probability upper bound on $\left\|\tilde{\mathcal{Z}}^\top \mathcal{U}/n\right\|_\infty$, which is required for the choices of both $\lambda$ in the penalized version and $r_n$ in the constrained version. Deriving upper bounds for the other quantities $q_n$, $s_n$ follow similar arguments, and an outline is provided in the proof of Proposition 4.3.

**Proposition E.1** (Deviation Bound: Phase-II)**.** *There exist universal constants $c_i > 0$ such that if $n \gtrsim \log d^2(p+q)$, then for any $A > 0$ the following holds with probability at least $1 - c_0 \exp[-(c_1 A^2 - 2) \log d^2(p+q)]$:*

$$\left\|\tilde{\mathcal{Z}}^\top \mathcal{U}/n\right\|_\infty \leq \varphi_1 \sqrt{\frac{\log d^2(p+q)}{n}} + \varphi_2 \cdot \left(\Delta_\varepsilon + \Delta_\varepsilon^2 + \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}\right),$$

*where*

$$\begin{aligned}
\varphi_1 &= c_1 |||f_y||| A \max\left\{1, |||\Theta|||^2 \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2, \left|||\Pi_{[\tilde{p}]}|||^2\right\}, \\
\varphi_2 &= c_2 |||f_y||| \, \|\Theta\|_{2,1} \max\{1, \left\|\Pi_{[\tilde{p}]}\right\|_{2,1}\}.
\end{aligned}$$

*Proof of Proposition E.1.* Recall $n = T - q$ is the number of observations in the Phase-II regression. The element-wise maximum norm can be expressed as

$$\left\|\mathcal{Z}^\top \mathcal{U}/n\right\|_\infty = \max_{\substack{1 \leq \ell \leq p \\ 1 \leq m \leq q}} \max\left\{\left\|\mathcal{Y}_{(\ell)}^\top \mathcal{U}/n\right\|_\infty, \left\|\hat{E}_{(m)}^\top \mathcal{U}/n\right\|_\infty\right\},$$

where $\mathcal{Y}_{(\ell)} = [y_{n-\ell} : \ldots : y_{1-\ell}]^\top$ is a data matrix with $n$ consecutive observations from the process $\{y_t\}$, $\hat{E}_{(m)} = [\hat{\varepsilon}_{n-m} : \ldots : \hat{\varepsilon}_{1-m}]^\top$ is a data matrix with $n$ consecutive observations from the process $\{\hat{\varepsilon}_t\}$, and $\mathcal{U}$ is a data matrix with $n$ consecutive observations from the

process $\{u_t\}$. Also, the process $\{u_t\}$ can be alternately expressed as

$$\begin{aligned} u_t &= \Phi(L)y_t - \Theta(L)\hat{\varepsilon}_t \\ &= \Theta(L)(a_t - \hat{\varepsilon}_t) \\ &= \Theta(L)(a_t - \varepsilon_t) - \Theta(L)(\hat{\varepsilon}_t - \varepsilon_t) \\ &= \Theta(L)\left(\Pi(L) - \Pi_{[\tilde{p}]}(L)\right) y_t - \Theta(L)(\hat{\varepsilon}_t - \varepsilon_t) \\ &= \mathcal{A}(L)y_t + \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t), \text{ say.} \end{aligned}$$

The lag polynomial $\mathcal{A}(L) = \Theta(L)\Pi_{-[\tilde{p}]}(L)$ satisfies $|||\mathcal{A}||| \leq |||\Theta|||||\Pi_{-[\tilde{p}]}||_{2,1}$, and $\mathcal{B}(L) = -\Theta(L)$ is a finite order lag polynomial.

Now note that each term $\mathcal{Y}_{(\ell)}^\top \mathcal{U}/n$ can be expressed in the form of a sample covariance matrix $\widehat{\mathrm{Cov}}(L^\ell y_t, u_t) := \sum_{t=1}^n y_{t-\ell}u_t^\top/n$. With this notation, we can decompose this into two terms and apply deviation bounds from Proposition B.1 and Lemma B.1 on each term separately. To be precise, for any $\ell$, $1 \leq \ell \leq p$, we have

$$\widehat{\mathrm{Cov}}(y_{t-\ell}, u_t) = \widehat{\mathrm{Cov}}(L^\ell y_t, \mathcal{A}(L)y_t) + \widehat{\mathrm{Cov}}(L^\ell y_t, \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t)).$$

Similarly, for any $m$, $1 \leq m \leq q$, we can decompose $\widehat{\mathrm{Cov}}(\hat{\varepsilon}_{t-m}, u_t)$ into four parts as

$$\begin{aligned} &\widehat{\mathrm{Cov}}(\hat{\varepsilon}_{t-m} - \varepsilon_{t-m}, \mathcal{A}(L)y_t) + \widehat{\mathrm{Cov}}(\hat{\varepsilon}_{t-m} - \varepsilon_{t-m}, \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t)) \\ &+\widehat{\mathrm{Cov}}(\Pi_{[\tilde{p}]}(L)L^m y_t, \mathcal{A}(L)y_t) + \widehat{\mathrm{Cov}}(\Pi_{[\tilde{p}]}(L)L^m y_t, \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t)). \end{aligned}$$

Using bounds from Proposition B.1 and Lemma B.1 then implies that there are universal constants $c_i > 0$ such that each of the following events hold with probability at least $1 -$

$c_0 d^2 \exp[-(c_1 A^2 - 1)\log d^2(p+q)]$ as long as $n > q$, $\tilde{p} \geq p+q$ and $n \gtrsim \log d^2(p+q)$:

$$\begin{aligned}
\left\|\widehat{\mathrm{Cov}}(L^\ell y_t, \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t))\right\|_\infty &\leq 2\sqrt{2\pi}|||f_y|||^{1/2}\Delta_\varepsilon \|\Theta\|_{2,1} \\
\left\|\widehat{\mathrm{Cov}}(L^\ell y_t, \mathcal{A}(L) y_t)\right\|_\infty &\leq 2\pi|||f_y||| \Big[|||\Theta||| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1} + \\
&\qquad 3A\max\left\{1, |||\Theta|||^2 \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2\right\}\sqrt{\log d^2(p+q)/n}\Big] \\
\left\|\widehat{\mathrm{Cov}}(\hat{\varepsilon}_{t-m} - \varepsilon_{t-m}, \mathcal{A}(L) y_t)\right\|_\infty &\leq 2\sqrt{2\pi}|||f_y|||^{1/2}|||\Theta||| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}\Delta_\varepsilon \\
\left\|\widehat{\mathrm{Cov}}(\hat{\varepsilon}_{t-m} - \varepsilon_{t-m}, \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t))\right\|_\infty &\leq 2\|\Theta\|_{2,1}\Delta_\varepsilon^2 \\
\left\|\widehat{\mathrm{Cov}}(\Pi_{[\tilde{p}]}(L)L^m y_t, \mathcal{B}(L)(\hat{\varepsilon}_t - \varepsilon_t))\right\|_\infty &\leq 2\sqrt{2\pi}|||f_y|||^{1/2}\left|\left|\left|\Pi_{[\tilde{p}]}\right|\right|\right|^{1/2}\Delta_\varepsilon \|\Theta\|_{2,1} \\
\left\|\widehat{\mathrm{Cov}}(\Pi_{[\tilde{p}]}(L)L^m y_t, \mathcal{A}(L) y_t)\right\|_\infty &\leq 2\pi|||f_y||| \Big[|||\Theta||| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1} \left|\left|\left|\Pi_{[\tilde{p}]}\right|\right|\right| + \\
&\qquad 3A\max\left\{\left|\left|\left|\Pi_{[\tilde{p}]}\right|\right|\right|^2, |||\Theta|||^2 \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2\right\}\sqrt{\log d^2(p+q)/n}\Big].
\end{aligned}$$

Summing up the six terms above and taking a union bound over $1 \leq \ell \leq p$, $1 \leq m \leq q$, we obtain the final upper bound. □

*Proof of Proposition 4.3.* We start by deriving a suitable choice of $s_n$. To this end, note that for each $j$, $1 \leq j \leq d$, we have $\widehat{\mathrm{Var}}(u_{tj})$ can be expressed as

$$\begin{aligned}
e_j^\top \widehat{\mathrm{Cov}}(u_t, u_t) e_j \;=\; & e_j^\top \widehat{\mathrm{Cov}}\left(\Theta(L)\Pi_{-[\tilde{p}]}(L) y_t, \Theta(L)\Pi_{-[\tilde{p}]}(L) y_t\right) e_j \\
& -2 e_j^\top \widehat{\mathrm{Cov}}\left(\Theta(L)\Pi_{-[\tilde{p}]}(L) y_t, \Theta(L)(\hat{\varepsilon}_t - \varepsilon_t)\right) e_j \\
& + e_j^\top \widehat{\mathrm{Cov}}\left(\Theta(L)(\hat{\varepsilon}_t - \varepsilon_t), \Theta(L)(\hat{\varepsilon}_t - \varepsilon_t)\right) e_j.
\end{aligned}$$

We then use upper bounds on the individual terms using the deviation bounds provided in our technical ingredients.

In particular, set $w_t := \Theta(L)\Pi_{-[\tilde{p}]}(L) y_t$. Then $|||f_w||| \leq |||\Theta||| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2 |||f_y|||$. Setting $\sigma_j^2 = e_j^\top \Gamma_w(0) e_j$, Proposition 2.4 of [10] implies, with probability at least $1 - c_1 \exp\left[-(c_2 A^2 - 1)\log d^2(p+q)\right]$, the following holds:

$$\left| e_j^\top \widehat{\mathrm{Cov}}(w_t, w_t) e_j - \sigma_j^2 \right| \leq 2\pi|||f_w|||A\sqrt{\log d^2(p+q)/n}.$$

The second term in the above expansion, $e_j^\top \widehat{\mathrm{Cov}}\left(w_t, \Theta(L)(\hat{\varepsilon}_t - \varepsilon_t)\right) e_j$, can be bounded in absolute value (use Lemma B.1 and note that $n > q$, $n \gtrsim \log d^2(p+q)$) by the following:

$$2\sqrt{2\pi}|||f_w|||^{1/2}\Delta_\varepsilon \left\|\Theta\right\|_{2,1}.$$

The last term in the above expansion, $e_j^\top \widehat{\mathrm{Cov}}\left(\Theta(L)(\hat{\varepsilon}_t - \varepsilon_t), \Theta(L)(\hat{\varepsilon}_t - \varepsilon_t)\right) e_j$, can be bounded in absolute value (see proof of Lemma B.1(ii)) by the following:

$$2\left\|\Theta\right\|_{2,1}^2 \Delta_\varepsilon^2.$$

Combining these, we obtain the following choice of $s_n$ (with $\sigma^2$ as $\sum_{j=1}^d \sigma_j^2$):

$$\begin{aligned} s_n \;=\; & 2\pi|||\Theta||| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2 |||f_y|||A\, d\, \sqrt{\log d^2(p+q)/n} + \\ & 4\sqrt{2\pi|||\Theta||| \left\|\Pi_{-[\tilde{p}]}\right\|_{2,1}^2 |||f_y|||}\, d\, \Delta_\varepsilon \left\|\Theta\right\|_{2,1} + 2d\, \left\|\Theta\right\|_{2,1}^2 \Delta_\varepsilon^2. \end{aligned}$$

The choice of $q_n$ follows from Lemma B.2, with a union bound over $d^2(p+q)^2$ choices of $u, v$ as canonical unit vectors in $\mathbb{R}^{d(p+q)}$. In particular, we have

$$q_n = 2\pi|||f_z|||A\sqrt{\log d^2(p+q)/n} + 2q\Delta_a^2 + 2\sqrt{2\pi|||f_z|||q}\Delta_a.$$

The choice of $r_n$ follows directly from the Proposition E.1. □

# F Implementation of the Sparse VARMA Procedure

**Phase-II Proximal Gradient Algorithm.** The objective function in (3.4) is separable over the $d$ rows of $\Phi, \Theta$ and can thus be solved in parallel by solving the "one-row" subproblems, see e.g., [43]. Denote the $i^{th}$ row of $Y$ by $Y_{i\cdot} = \mathbb{R}^{1\times(T-\bar{o})}$, the $i^{th}$ row of $\Phi$ by $\Phi_{i\cdot} \in \mathbb{R}^{1\times dp}$ and the $i^{th}$ row of $\Theta$ by $\Theta_{i\cdot} \in \mathbb{R}^{1\times dq}$. The Proximal Gradient Algorithm for the one-row subproblems is given in Algorithm 1.

**Choice of convex regularizers.** As indicated in Section 3, we focus on the $\ell_1$-norm

**Algorithm 1** Proximal Gradient Algorithm to solve Phase-II

**Input** $Y_{i\cdot}$, $Z$, $X$, $p$, $q$, $\Phi_{i\cdot}[0]$, $\Theta_{i\cdot}[0]$, $\lambda_\Phi$, $\lambda_\Theta$, $\alpha, \mathcal{P}_{\text{AR}}(\Phi), \mathcal{P}_{\text{MA}}(\Theta)$, $\epsilon$

**Initialization** Set

- $\Phi_{i\cdot}[2] \leftarrow \Phi_{i\cdot}[1] \leftarrow \Phi_{i\cdot}[0]$
- $\Theta_{i\cdot}[2] \leftarrow \Theta_{i\cdot}[1] \leftarrow \Theta_{i\cdot}[0]$
- step size $s = 1/\sigma_1(A)^2$, with $\sigma_1(A)$ the largest singular value of the matrix $A = \binom{Z}{X}$

**Iteration** For $r = 3, 4, \ldots$

- $\widehat{\phi} \leftarrow \Phi_{i\cdot}[r-1] + \dfrac{r-2}{r+1}\left(\Phi_{i\cdot}[r-1] - \Phi_{i\cdot}[r-2]\right)$
- $\Phi_{i\cdot}[r] \leftarrow \dfrac{1}{1+\alpha\cdot\lambda_\Phi} \cdot \text{Prox}_{s\lambda_\Phi P_i^{(\Phi)}}\left(\widehat{\phi} - s\nabla_\Phi \mathcal{L}_i(\widehat{\phi})\right)$,
  where
  - $\nabla_\Phi \mathcal{L}_i(\widehat{\phi}) = -(Y_{i\cdot} - \widehat{\phi}Z - \Theta_{i\cdot}[r-1]X)Z^\top$,
  - $\text{Prox}_{s\lambda_\Phi P_i^{(\Phi)}}(\cdot)$ the proximal operator of the function $s\lambda_\Phi P_i^{(\Phi)}(\cdot)$ where $\mathcal{P}_{\text{AR}}(\Phi) = \sum_i P_i^{(\Phi)}(\Phi_{i\cdot})$.
- $\widehat{\theta} \leftarrow \Theta_{i\cdot}[r-1] + \dfrac{r-2}{r+1}\left(\Theta_{i\cdot}[r-1] - \Theta_{i\cdot}[r-2]\right)$
- $\Theta_{i\cdot}[r] \leftarrow \dfrac{1}{1+\alpha\cdot\lambda_\Theta} \cdot \text{Prox}_{s\lambda_\Theta P_i^{(\Theta)}}\left(\widehat{\theta} - s\nabla_\Theta \mathcal{L}_i(\widehat{\theta})\right)$,
  where
  - $\nabla_\Theta \mathcal{L}_i(\widehat{\theta}) = -(Y_{i\cdot} - \Phi_{i\cdot}[r]Z - \widehat{\theta}X)X^\top$,
  - $\text{Prox}_{s\lambda_\Theta P_i^{(\Theta)}}(\cdot)$ the proximal operator of the function $s\lambda_\Theta P_i^{(\Theta)}(\cdot)$ where $\mathcal{P}_{\text{MA}}(\Theta) = \sum_i P_i^{(\Theta)}(\Theta_{i\cdot})$.

**Convergence** Iterate until $||\Phi_{i\cdot}[r] - \Phi_{i\cdot}[r-1]||_\infty \leq \epsilon$ and $||\Theta_{i\cdot}[r] - \Theta_{i\cdot}[r-1]||_\infty \leq \epsilon$

**Output** $\widehat{\Phi}_{i\cdot} \leftarrow \Phi_{i\cdot}[r]; \widehat{\Theta}_{i\cdot} \leftarrow \Theta_{i\cdot}[r]$

---

and HLag penalty as choices of convex regularizers. For the $\ell_1$-norm,

$$P_i^{(\Phi)}(\Phi_i) = \sum_{j=1}^{d}\sum_{\ell=1}^{p}|\Phi_{\ell,ij}| \text{ and } P_i^{(\Phi)}(\Theta_i) = \sum_{j=1}^{d}\sum_{m=1}^{q}|\Theta_{m,ij}|.$$

For the HLag penalty,

$$P_i^{(\Phi)}(\Phi_i) = \sum_{j=1}^{d}\sum_{\ell=1}^{p}||\Phi_{(\ell:p),ij}|| \text{ and } P_i^{(\Theta)}(\Theta_i) = \sum_{j=1}^{d}\sum_{m=1}^{q}||\Theta_{(m:q),ij}||.$$

# G Simulation Study

We investigate the performance of the proposed VARMA estimator through a simulation study. We generate data from a VARMA$_d(p, q)$ with time series length $T = 100$. To ensure identification, we take $\Phi_\ell$, $1 \leq \ell \leq p$, diagonal matrices and set each diagonal element of $\Phi_\ell$ equal to $0.4/\ell$. For the autoregressive order, we take $p = 4$. For the error covariance matrix, we take $\Sigma_a = I_d$. To reduce the influence of initial conditions on the data generating processes, the first 200 observations were discarded as burn-in for each simulation run.

We consider several settings for the moving average parameters. We take banded matrices for $\Theta_m$, $1 \leq m \leq q$ with the diagonal elements of $\Theta_m$ equal to $\theta/m$, the elements on the first lower and upper subdiagonals equal to $\theta/(10m)$, and the elements on the second lower and upper subdiagonals equal to $\theta/(100m)$. The parameter $\theta$ regulates the strength of the moving average signal. The parameter $q$ regulates the moving average order. We investigate the effect of the following features. *(i) The MA signal strength*: we vary the parameter $\theta \in \{0, 0.4, 0.6, 0.8\}$. The larger $\theta$, the stronger the moving average signal. Note that for $\theta = 0$, the true model is a VAR. *(ii) The MA order*: we vary the parameter $q \in \{4, 6, 8, 10\}$. *(iii) The number of time series*: we vary the number of time series $d \in \{5, 10, 20, 40\}$. In all considered settings, the VARMA models are invertible and stable.

**Estimators.** We compare the following estimators. (i) "VARMA$(p, q; a_t)$": the VARMA estimator of model (1.1) with an oracle providing the true errors $a_t$ and orders $p$ and $q$. (ii) "VARMA$(p, q; \widehat{\varepsilon}_t)$" the VARMA estimator of model (3.3) with approximated errors and an oracle providing the orders $p$ and $q$. (iii) "VARMA$(\widehat{p}, \widehat{q}; \widehat{\varepsilon}_t)$": the VARMA estimator of model (3.3) with approximated errors and specified orders $\widehat{p} = \widehat{q} = \lfloor 0.75\sqrt{T} \rfloor$. (iv) "VAR$(\widetilde{p})$": the VAR estimator of model (3.1) with specified order $\widetilde{p} = \lfloor 1.5\sqrt{T} \rfloor$. We use both the $\ell_1$-norm and the HLag penalty to obtain our estimates.

**Performance Measure.** We compare the performance of the estimators in terms of out-of-sample forecast accuracy. We generate time series of length $T + 1$ and use the last observation to measure forecast accuracy. We compute the Mean Squared Forecast Error

$$\text{MSFE} = \frac{1}{N} \sum_{s=1}^{N} \frac{1}{d} \| y_{T+1}^{(s)} - \widehat{y}_{T+1}^{(s)} \|^2,$$

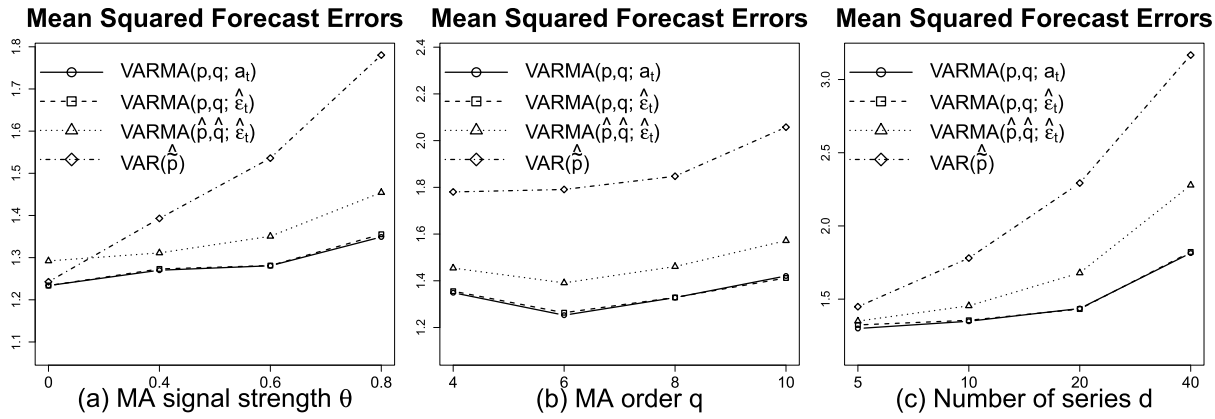


Figure A2: Mean Squared Forecast Errors (averaged over the simulation runs) of the four estimators for different values of (a) the moving average parameter $\theta$, (b) the moving average order $q$, and (c) the number of time series $d$.

where $y_t^{(s)}$ is the vector of time series at time point $t$ in the $s^{th}$ simulation run, and $\widehat{y}_t^{(s)}$ is its predicted value. The number of simulations is $N = 500$. We focus on out-of-sample forecast accuracy in the simulation study, in line with the discussion of the applications in Section 5. We did also compare the estimators in terms of the estimation accuracy of the $\Pi$-matrices; similar conclusions are obtained and available from the authors upon request.

## G.1 Effect of the Moving Average Signal Strength

Figure A2 panel (a) shows the MSFEs (averaged over the simulation runs) of the four estimators for different values of the moving average parameter $\theta$, which regulates the moving average signal strength. We report the results for the HLag penalty and $d = 10, q = 4$.

If the true model is a VARMA (i.e. $\theta \neq 0$), the VARMA estimators perform better than the VAR, as expected. The larger $\theta$, the larger the gain of VARMA over VAR. The differences in forecast accuracy between the VARMA estimators and the VAR estimator are all significant, as confirmed by paired $t$-tests (at the 5% significance level). Among the VARMA estimators, there is no significant difference between "VARMA$(p, q; a_t)$" and "VARMA$(p, q; \widehat{\varepsilon}_t)$" thus supporting the validity of the two-phase approach. The VARMA estimator with estimated errors and selected orders (i.e., "VARMA$(\widehat{p}, \widehat{q}; \widehat{\varepsilon}_t)$") performs, for all values of $\theta$, very similarly to the one with known orders. The loss in forecast accuracy of

Table A1: Mean Square Forecast Errors (averaged over the simulation runs) of the four estimators with either HLag penalty or $\ell_1$-norm and for different values of the moving average parameter $\theta$. $P$-values of a paired $t$-test are in parentheses.

| | VARMA$(p,q;a_t)$ | | VARMA$(p,q;\widehat{\varepsilon}_t)$ | | VARMA$(\widehat{p},\widehat{q};\widehat{\varepsilon}_t)$ | | VAR$(\widetilde{p})$ | |
|---|---|---|---|---|---|---|---|---|
| | HLag | $\ell_1$ | HLag | $\ell_1$ | HLag | $\ell_1$ | HLag | $\ell_1$ |
| $\theta=0$ | 1.234 | 1.263 (<0.01) | 1.234 | 1.263 (<0.01) | 1.292 | 1.334 (<0.01) | 1.243 | 1.317 (<0.01) |
| $\theta=0.4$ | 1.270 | 1.299 (0.415) | 1.273 | 1.303 (0.396) | 1.311 | 1.387 (0.040) | 1.393 | 1.558 (<0.01) |
| $\theta=0.6$ | 1.281 | 1.315 (0.360) | 1.281 | 1.321 (0.293) | 1.351 | 1.459 (<0.01) | 1.536 | 1.802 (<0.01) |
| $\theta=0.8$ | 1.349 | 1.383 (0.275) | 1.355 | 1.399 (0.170) | 1.454 | 1.582 (<0.01) | 1.780 | 2.159 (<0.01) |

not knowing the autoregressive or moving average order is limited to 5% on average.

If the true model is a VAR (i.e. $\theta = 0$), the VARMA estimators with known orders both reduce to a VAR($p$) estimator since $\theta = 0$, hence $q = 0$. They give the lowest MSFE. However, in practice, the orders of the model are not known. For unknown orders, the VARMA estimator is competitive to the VAR estimator. The VARMA estimator attains this competitveness since, in general, it returns a more parsimonious model (i.e. the estimated VARMA has more sparse AR coefficients with some sparse MA coefficients than the number of sparse coefficients in the estimated VAR).

The relative performance of the four estimators with HLag penalty are compared to the results with $\ell_1$-norm in Table A1. For the estimators with unknown maximal lag orders (i.e. VARMA$(\widehat{p},\widehat{q};\widehat{\varepsilon}_t)$ and VAR$(\widetilde{p})$), HLag outperforms the $\ell_1$-norm in all considered cases ($p$-values paired $t$-test $< 0.05$). For the estimators with known maximal lag orders, (e.g., VARMA$(p,q;a_t)$ and VARMA$(p,q;\widehat{\varepsilon}_t)$), HLag performs, overall, as good as the $\ell_1$-norm. These results are in line with the findings of [43].

## G.2 Effect of the Moving Average Order

Figure A2 panel (b) shows the MSFEs of the four estimators for different values of the moving average order $q$. We report the results for the HLag penalty and $d = 10, \theta = 0.8$. Similar conclusions are obtained with the $\ell_1$-norm and other values of $d$ and $\theta$, therefore omitted. For all values of $q$, the VARMA estimators perform significantly better than the VAR estimator. The oracle VARMA estimators perform equally good and are closely followed by

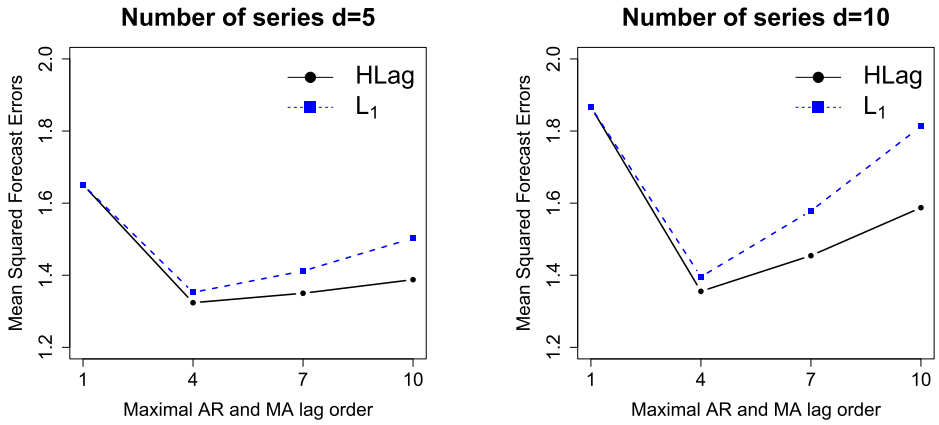


Figure A3: Mean Squared Forecast Errors (averaged over the simulation runs) of the VARMA($\widehat{p}$, $\widehat{q}$; $\widehat{\epsilon}_t$) estimator with HLag penalty (black solid line) and $\ell_1$-penalty (blue dashed line) for different values of the maximal lag orders $p$ and $q$ (horizontal axis) and number of time series $d = 5$ (left), $d = 10$ (right).

the VARMA estimator with approximated errors and unknown orders. The latter improves forecast accuracy over the VAR estimator by about 20% on average.

## G.3 Effect of the Number of Time Series

Figure A2 panel (c) shows the MSFEs for different values of the number of time series $d$. We report the results for the HLag penalty and $q = 4, \theta = 0.8$. As the number of time series increases relative to the fixed time series length $T$, it becomes more difficult to accurately estimate the model. Hence, the MSFEs of all estimators increase. For all values of $d$, the VARMA estimators attain lower values of the MSFE than the VAR estimator. All differences are significant. The loss in forecast accuracy of not knowing the AR and MA order is only 2% for $k = 5$ and remains limited to 20% for $k = 40$. The margin by which the VARMA estimator (with approximated errors and unknown orders) improves forecast accuracy over the VAR increases from around 7% for $k = 5$ to around 30% for $k = 40$.

## G.4 Implications of Misspecifying the Maximal Lag Orders

Next, we investigate the implications of misspecifying the maximal AR and MA lag orders of the VARMA. We generate data from the VARMA model with $p = q = 4, \theta = 0.8$, $d = 5, 10$

Table A2: Data-based Simulation Design: Mean Squared Forecast Errors (averaged over the simulation runs) of the four estimators with HLag penalty and different forecast horizons. Standard errors around the reported results are in parentheses.

| Forecast horizon | VARMA$(p, q; a_t)$ | VARMA$(p, q; \widehat{\varepsilon}_t)$ | VARMA$(\widehat{p}, \widehat{q}; \widehat{\varepsilon}_t)$ | VAR$(\widehat{p})$ |
|---|---|---|---|---|
| $h = 1$ | 0.764 (0.021) | 0.763 (0.022) | 0.765 (0.022) | 0.758 (0.025) |
| $h = 8$ | 0.782 (0.022) | 0.783 (0.022) | 0.785 (0.022) | 0.877 (0.026) |
| $h = 13$ | 0.784 (0.022) | 0.785 (0.022) | 0.787 (0.022) | 0.877 (0.026) |

and estimate a sparse VARMA model with maximal lag orders smaller than, equal to and larger than the true orders. Note that a maximal lag order of seven corresponds to our recommendation ($\hat{p} = \hat{q} = \lfloor 0.75\sqrt{100} \rfloor = 7$). Figure A3 shows the MSFEs for the sparse VARMA estimator with HLag penalty and $\ell_1$ penalty, different values of the maximal AR and MA lag orders (horizontal axis) and number of time series (panels).

The lowest MSFEs are attained at the true maximal lag order of four, as expected. At a maximal lag order of one, all models are misspecified and the MSFEs are the largest. Using too small maximal lag orders thus has more severe consequences than using too large maximal lag orders. Furthermore, the drop in MSFE at our recommended maximal lag orders (of seven) remains small provided that one uses an HLag penalty. Indeed, the price to pay for too large maximal lag orders is smaller for HLag than the standard $\ell_1$ penalty since HLag encourages low maximal lag orders.

## G.5 Data-based Simulation Design

As a last experiment, we consider a data-based design [30]. Similar to [13], we carry out a simulation by bootstrapping the actual demand set with $d = 16$ and $T = 76$ as discussed in Section 5 of the paper. We start from the autoregressive and moving average estimates obtained with the sparse VARMA method with HLag penalties and $\hat{p} = \hat{q} = \lfloor 0.75\sqrt{T} \rfloor = 6$ We then generate data from a VARMA$_d(\hat{p}, \hat{q})$ using a non-parametric residual bootstrap procedure (e.g., [36]) with bootstrap errors an i.i.d. sequence of discrete random variables uniformly distributed on $\{1, \ldots, T\}$.

Table A2 gives the MSFEs of the four estimators at different forecast horizons $h = 1, 8, 13$, as used in Section 5. For the VARMA estimators with known orders, we use $p = q = 3$, in line

with the largest reported values in Figure 1. First of all, note that it becomes more difficult to obtain accurate forecasts for longer horizons; the MSFEs of all estimators increases with $h$. The relative performance of VARMA compared to VAR is tied to the forecast horizon: at $h = 1$, all estimators perform equally well (i.e. there are no significant differences, as confirmed through paired $t$-tests). At longer forecast horizons, the VARMA estimators still perform equally well but statistically outperform the VAR estimator. These findings support the results from Section 5.

# References


[1] Agarwal, A.; Negahban, S. and Wainwright, M. J. (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, D. S. 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® 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] Boyd, S. and Vandenberghe, L. (2004), *Convex optimization*, Cambridge University Press.

[12] Brockwell, P. J. and Davis, R. A. (1991), *Time series: Theory and methods*, Springer Series in Statistics.

[13] Carriero, A.; Kapetanios, G. and Marcellino, M. (2012), "Forecasting government bond yields with large Bayesian vector autoregressions," *Journal of Banking & Finance*, 36, 2026–2047.

[14] Chan, J. C. C.; Eisenstat, E. and Koop, G. (2016), "Large Bayesian VARMAs," *Journal of Econometrics*, 192(2), 374–390.

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

[16] 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.

[17] 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.

[18] Dias, G. F. and Kapetanios, G. (2018), "Estimation and forecasting in vector autoregressive moving average models for rich datasets," *Journal of Econometrics*, 202, 72–91.

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

[20] Diebold, F. X. and Yılmaz, K. (2014), "On the network topology of variance decompositions: Measuring the connectedness of financial firms," *Journal of Econometrics*, 182, 119–134.

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

[22] 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.

[23] Fernández-Villaverde, J.; Rubio-Ramírez, J. F.; Sargent, T. J. and Watson, M. W. (2007), "ABCs (and Ds) of understanding VARs," *American Economic Review*, 97, 1021–1026.

[24] 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.

[25] 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.

[26] Greenshtein, E. and Ritov, Y. (2004), "Persistence in high-dimensional linear predictor selection and the virtue of overparametrization," *Bernoulli*, 10, 971–988.

[27] Hannan, E. J. (1976), "The Identification and Parameterization of ARMAX and State Space Forms," *Econometrica*, 44, 713–723.

[28] Hannan, E. J. and Kavalieris, L. (1984), "Multivariate linear time series models," *Advances in Applied Probability*, 16, 492–561.

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

[30] Ho, M. S. and Sorensen, B. E. (1996), "Finding cointegration rank in high dimensional systems using the Johansen test: An illustration using data based Monte Carlo simulations," *Review of Economics and Statistics*, 78, 726–732.

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

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

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

[34] Kock, A. B. and Callot (2015), "Oracle inequalities for high dimensional vector autoregressions," *Journal of Econometrics*, 186, 325–344.

[35] Koop, G. M. (2013), "Forecasting with medium and large Bayesian VARs," *Journal of Applied Econometrics*, 28(2), 177–203.

[36] Kreiss, J. P. and Lahiri, S. (2012), *Bootstrap methods for time series*, In: Rao, T., Rao, S. and Rao, C. (Eds.) Handbook of Statistics 30. Time Series Analysis: Methods and Applications. North Holland.

[37] Loh, P.-L. and Wainwright, M. J. (2012), "High-dimensional regression with noisy and

missing data: provable guarantees with nonconvexity." *The Annals of Statistics*, 40, 1637–1664.

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

[39] — (2006), "Forecasting with VARMA models," *Handbook of economic forecasting*, 1, 287–325.

[40] Manski, C. F. (2010), "Partial identification in econometrics," in *Microeconometrics*, Springer, pp. 178–188.

[41] Matteson, D. S. and Tsay, R. S. (2011), "Dynamic orthogonal components for multivariate time series," *Journal of the American Statistical Association*, 106(496), 1450–1463.

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

[43] Nicholson, W. B.; Wilms, I.; Bien, J. and Matteson, D. S. (2020), "High dimensional forecasting via interpretable vector autoregression," *Journal of Machine Learning Research*, 21, 1–52.

[44] Poskitt, D. S. (1992), "Identification of Echelon canonical forms for vector linear processes using least squares," *The Annals of Statistics*, 20, 195–215.

[45] — (2016), "Vector autoregressive moving average identification for macroeconomic modeling: A new methodology," *Journal of Econometrics*, 192, 468–484.

[46] 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.

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

[48] Tamer, E. (2010), "Partial identification in econometrics," *Annu. Rev. Econ.*, 2, 167–195.

[49] Tiao, G. C. and Tsay, R. S. (1989), "Model specification in multivariate time series," *Journal of the Royal Statistical Society Series B*, 51, 157–213.

[50] Tibshirani, R. J. (2013), "The lasso problem and uniqueness," *Electronic Journal of statistics*, 7, 1456–1490.

[51] Tsay, R. S. (2014), *Multivariate Time Series Analysis: With R and Financial Applications*, Wiley.

[52] 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.

[53] Wallis, K. F. (1977), "Multiple time series analysis and the final form of econometric models," *Econometrica*, 45, 1481–1497.

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

[55] — (2017), "Interpretable vector autoregressions with exogenous time series," *NIPS 2017 Symposium on Interpretable Machine Learning*, arXiv:1711.03623.

[56] Wu, W.-B. and Wu, Y. N. (2016), "Performance bounds for parameter estimates of high-dimensional linear models with correlated errors," *Electronic Journal of Statistics*, 10, 352–379.

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

[58] Zellner, A. and Palm, F. (1974), "Time series analysis and simultaneous equation econo-

metric model," *Journal of Econometrics*, 2, 17–54.

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