Low-Rank and Sparse Modeling of High-dimensional Vector Autoregressions

Sumanta Basu$^1$ and George Michailidis$^2$

$^1$University of California, Berkeley and $^2$University of Florida

This draft: November 29, 2015

Abstract

Network modeling of high-dimensional time series in presence of unobserved latent variables is an important problem in macroeconomics and finance. In macroeconomic policy making and forecasting, it is often impossible to observe and incorporate all the relevant series in the analysis. Failure to include these variables often results in spurious connectivity among the observed time series in structural analyses, which may have serious policy implications. In order to accurately estimate a network of Granger causal interactions after accounting for latent effects, we introduce a novel approach of low-rank and sparse vector autoregression (VAR). We argue that in presence of a few latent pervasive factors, the transition matrix of a misspecified VAR model among the observed series can be approximated as the sum of a low-rank and a sparse component. Exploiting this connection, we consider estimating low-rank plus sparse VAR models using a combination of nuclear norm and lasso penalties. We establish non-asymptotic upper bounds on the estimation error rates of the low-rank and the sparse components and demonstrate the advantage of the proposed methodology over ordinary and sparse VAR estimates via numerical experiments.
1 Introduction

The problem of learning networks of interactions among the components of a high-dimensional time series arises routinely in many areas of economic and biological sciences. Examples include macroeconomic policy making and forecasting, assessing connectivity among financial firms and reconstructing gene-regulatory interactions from time-course gene expression data. Vector autoregressive (VAR) models provide a convenient framework to accomplish this. In recent years, high-dimensional VAR models have gained a lot of interest in statistics and econometrics community [Basu and Michailidis, 2015].

An important challenge in autoregressive modeling of multivariate time series stems from the fact that failure to include relevant variables in the model can introduce spurious correlations among the individual time series, resulting in incorrect estimation of the edge set of the underlying Granger causal network. This is also a major critique against causal interpretation of Granger-causality. This problem in VAR modeling is well-known in the economics literature. For instance, Christiano et al. [1999] argue that a positive response of prices to monetary tightening in the post-war US economy, commonly known as the “price puzzle”, is an artefact of not including forward looking variables in the model [Banbura et al., 2010]. The high-dimensional VAR framework with sparsity based regularizers like lasso resolves this problem to a certain extent by allowing many variables in the model. However, in many macroeconomic applications it is not possible to observe all the relevant variables driving the market economy. A popular strategy is factor modeling, where the key idea is that there are a few latent factors driving the major co-movements of many time series [Stock and Watson, 2005]. Indeed, empirical evidence suggests that the co-movement of many macroeconomic time series in the US economy can be explained by a small number of unobserved factors extracted from the data.

Failure to account for the presence of unobserved common factors can negatively impact high-dimensional sparse VAR modeling in two ways. First, the correlation among the time series that is driven by underlying factors introduces spurious connectivities among the
observed time series. Second, even if the true Granger causal network is sparse, failure to account for hidden factors can result in a non-sparse VAR representation of the process and lasso estimates become inaccurate (cf Examples 1 and 2 below).

In this work, we propose to deal with this problem using a low-rank and sparse modeling strategy. Low-rank approximation and low-rank+sparse decomposition of Hankel matrices, which represent the input-output structure of a linear time invariant system (LTI), have appeared in the literature [Fazel et al., 2003]. A low-rank representation of the Hankel matrix corresponds to a system of small order or dimension and a sparse Hankel matrix represents sparse input-output system [Chandrasekaran et al., 2011]. In the context of high-dimensional stationary time series, we show that a low-rank or a sparse+low-rank structure in the transition matrix arises naturally, if the components of the observed process are affected by some latent factors. We demonstrate this using two examples.

Example 1. We consider a $p$-dimensional stationary process $\{X_t\}$ with the entire dynamics driven by a $r$-dimensional ($r \ll p$) unobserved process of factors $\{F_t\}$, which itself follows a $VAR(1)$ process

$$X_t = \Lambda F_t + \xi_t, \quad \xi_t \sim N(0, \Sigma_\xi), \quad \text{Cov}(\xi_t, \xi_s) = 0 \text{ if } t \neq s$$

$$F_t = HF_{t-1} + \eta_t, \quad \eta_t \sim N(0, \Sigma_\eta), \quad \text{Cov}(\eta_t, \eta_s) = 0, \quad \text{if } t \neq s$$

This is a simple example of a static factor model used in economics. We assume the matrix of factor loadings $\Lambda_{p \times r}$ has full column rank $r$, so that it has a left inverse $\Lambda^-$ satisfying $\Lambda^- \Lambda = I_r$. It then readily follows that

$$X_t = \Lambda \left[ HF_{t-1} + \eta_t \right] + \xi_t$$

$$= \Lambda \left[ H\Lambda^- (X_{t-1} - \xi_{t-1}) + \eta_t \right] + \xi_t$$

$$= \Lambda H\Lambda^- X_{t-1} + \left[ \Lambda\eta_t + \xi_t - \Lambda H\Lambda^- \xi_{t-1} \right]$$

$$= LX_{t-1} + \epsilon_t$$
where \( L = \Lambda H \Lambda^{-1} \) has rank at most \( r \) and the new error process \( \epsilon^t = \Lambda \eta^t + \xi^t - L \xi^{t-1} \) has a MA(1) component.

**Example 2.** Consider the same process \( \{X^t\} \), but assume now that its dynamics is governed by two sources: an underlying process of factors \( \{F^t\} \) as before and the interaction among its components, as captured by a VAR(1) process with a sparse transition matrix \( S \)

\[
X^t = \Lambda F^t + S X^{t-1} + \xi^t, \quad S \text{ sparse} \tag{3}
\]

\[
F^t = H F^{t-1} + \eta^t \tag{4}
\]

A similar calculation shows

\[
X^t = \Lambda H \Lambda^{-1} [X^{t-1} - S X^{t-2} - \xi^{t-1}] + S X^{t-1} + \xi^t + \Lambda \eta^t
\]

\[
= (L + S) X^{t-1} - L S X^{t-2} + \epsilon^t
\]

\[
\approx (L + S) X^{t-1} + \epsilon^t, \text{ assuming the second order effects in } L S \text{ are small}
\]

Motivated by the above connections, we propose to model the process \( \{X^t\} \) as a stable VAR(1) process with the transition matrix having a low-rank and a sparse component. Using a combination of nuclear norm and \( \ell_1 \) penalties, we show that a penalized least squares method can provide accurate estimates of the individual sparse and low-rank components, hence recover the network of Granger causal interactions after accounting for the latent effects. Under suitable regularity conditions on the degree of identifiability of the individual components, we derive non-asymptotic upper bounds on the estimation and in-sample prediction errors and establish consistency under high-dimensional scaling.

The rest of the paper is organized as follows. In section 2, we review the existing works on factor models and low-rank models in the literature. In section 3, we introduce the modeling framework and describe our estimation procedure. Section 4 is devoted to the theoretical properties of the proposed estimates. In section 5, we conduct numerical experiments to
compare the performance of the proposed method with ordinary least squares (OLS) and lasso.

**Notations.** We reserve the symbol $\| \cdot \|$ to denote the $\ell_2$-norm of a vector and the spectral norm of a matrix. The symbol $\| A \|_*$ is used to denote the nuclear norm, i.e., sum of the singular values, of a matrix. $A^*$ denotes the conjugate transpose of a matrix $A$. For any matrix $A$, we use the notations $\| A \|_0$ to denote $\text{card}(\text{vec}(A))$, $\| A \|_1$ to denote $\| \text{vec}(A) \|_1$ and $\| A \|_{\text{max}}$ to denote $\| \text{vec}(A) \|_{\infty}$. Throughout the chapter, $\Lambda_{\text{max}}(\cdot)$, $\Lambda_{\text{min}}(\cdot)$ are used to denote the maximum and minimum eigenvalues of a symmetric or Hermitian matrix. For any integer $p \geq 1$, we use $\mathbb{S}^{p-1}$ to denote the unit ball $\{ v \in \mathbb{R}^p : \| v \| = 1 \}$. We use $\{ e_1, e_2, \ldots \}$ generically to denote unit vectors in $\mathbb{R}^p$, when $p$ is clear from the context. Throughout the chapter, we write $A \gtrsim B$ if there exists an absolute constant $c$, independent of the model parameters, such that $A \geq cB$.

## 2 Related Work

Factor models have a long history in the statistics and econometrics literature as a popular technique for dimension reduction. Bai and Ng [2008] provide a comprehensive review of the theoretical and empirical work on factor models. In a typical static or dynamic factor model, the intertemporal dynamics among the observed series is entirely driven by the dynamics of the latent factors and it does not accommodate additional Granger causal effects among the observed series in the model.

The problem that we consider in this work, however, is considerably different in nature. We are interested in learning both the effect of the latent variables on the system and the Granger causal estimates or interactions among the system components, after accounting for the effects of latent factors.

Bernanke et al. [2005]; Bai et al. [2014]; Stock and Watson [2005] considered a similar problem using factor augmented vector autoregressive (FAVAR) models. The authors pro-
posed to model the joint process \([F^t], (X^t)\) as a vector autoregression, with the restriction that there is no effect from \(\{X^t\}\) to \(\{F^t\}\). FAVAR allows additional intertemporal dynamics of each of the observed series on top of the common factors, but does not allow Granger causal effects among the observed series. In the context of Example 2 in Section 1, FAVAR only allows diagonal \(S\).

In addition, since the process of factors is unobserved, the modeling strategy amounts to iteratively estimating the factors and using them in the VAR model. The method relies on consistent estimation of the number of factors and testing the restrictions imposed by the factor structure. Our approach of modeling the transition matrix as a combination of sparse and low-rank component does not require estimating the number of factors or the factor process separately and provides a framework for jointly estimating the common effects of the market and interactions among the system components. Further, the theory presented in the subsequent section can be easily generalized for approximately sparse and low-rank matrices and is applicable when the factor structure is not exactly, but only approximately, correct.

Low rank approximation of a given matrix is a popular technique of dimension reduction in many areas of science and engineering [Fazel, 2002], including matrix completion problems, principal component analysis and factor analysis. In recent years, decomposing a given matrix into sparse and low-rank component has gained considerable interest, with applications in video surveillance [Candès et al., 2011], neuroimaging and recommender systems. Finding the best low-rank plus sparse representation of an observed matrix via rank constrained optimization is computationally expensive due to the nonconvex nature of the problem. A tractable alternative commonly used in practice is the convex relaxation

\[
\min_{(L, S) : \ L + S = A} \|L\|_* + \gamma \|S\|_1, \ \gamma > 0
\]

where the nuclear/trace norm (sum of singular values of a matrix) serves as a surrogate for
the rank constraint and the $\ell_1$ norm serves as a surrogate for the sparsity constraint. Several algorithms for solving the above optimization problem have been proposed in the literature, including semidefinite programming [Chandrasekaran et al., 2011] and alternating direction method of multipliers [Yuan and Yang, 2009].

In many noisy settings such as ours, the matrix $A$ is not observed and needs to be estimated from data. An example closely related to our problem is the problem of Gaussian graphical model selection in the presence of latent variables from independent samples [Chandrasekaran et al., 2012]. Some other applications in factor analysis and multi-task regression have been covered in Agarwal et al. [2012]. To the best of our knowledge, the properties of these estimators have not been studied in the context of time series and dependent data.

3 Model and Estimation Procedure

In this section, we formally introduce the low-rank plus sparse VAR models and describe the estimation procedure using a combination of nuclear norm and lasso penalties. Here we focus only on VAR(1) models and leave extensions to higher-order VAR models for future work.

3.1 Model Description

Consider a VAR(1) model where the transition matrix has a low-rank and a sparse component

$$X_t = AX_{t-1} + \epsilon_t, \epsilon_t \sim N(0, \Sigma_\epsilon)$$

$$A = L^0 + S^0, \text{rank}(L^0) = r, \|S^0\|_0 = s, r \ll p, s \ll p^2$$

In this model, the matrix $L^0$ captures the effects of latent variables and $S^0$ encodes the dynamics among the individual time series, after accounting for the latent effects. The goal is to estimate $S^0$ and $L^0$ with high accuracy using small to moderate sample sizes. We restrict
our analysis to only models with serially uncorrelated errors. A general model with serially correlated error structure, although more well-suited for the examples described in Section 1, poses significant technical challenges due to endogeneity (correlation between predictors and the noise in the regression) and we intend to pursue it as a separate problem.

**Stability.** As shown in Basu and Michailidis [2015], stability of a VAR process is important for the estimation and prediction accuracy of regularized VAR estimates. If a VAR model (6) is stable, it has a spectral density

\[
f_X(\theta) = \frac{1}{2\pi} \left( A^{-1}(e^{-i\theta}) \right) \Sigma_{\epsilon} \left( A^{-1}(e^{-i\theta}) \right)^*, \quad \theta \in [-\pi, \pi],
\]

where \( A(z) := I_p - Az \). For stable VAR processes, \( \det(A(z)) \neq 0 \) on the unit circle of the complex plane \( \{z \in \mathbb{C} : |z| = 1\} \).

Following Basu and Michailidis [2015], we define the following quantities, which capture the effect of dependence on the accuracy of the regularized estimates:

\[
\mathcal{M}(f_X) = \text{ess sup}_{\theta \in [-\pi, \pi]} \Lambda_{\text{max}}(f_X(\theta)), \quad m(f_X) = \text{ess sup}_{\theta \in [-\pi, \pi]} \Lambda_{\text{min}}(f_X(\theta)),
\]

\[
\mu_{\text{max}}(A) = \max_{|z|=1} \Lambda_{\text{max}}(A^*(z)A(z)), \quad \mu_{\text{min}}(A) = \min_{|z|=1} \Lambda_{\text{min}}(A^*(z)A(z)),
\]

where \( \Lambda_{\text{max}}(.) \) and \( \Lambda_{\text{min}}(.) \) denote the maximum and minimum eigenvalues of a symmetric or Hermitian matrix, respectively.

As shown in Basu and Michailidis [2015], \( \mathcal{M}(f_X) \) and \( m(f_X) \) together capture the narrowness of the spectral density of the underlying stochastic process. Processes with stronger temporal and cross-sectional dependence have narrower spectra and slower convergence rates of regularized estimates. For VAR models, \( \mathcal{M}(f_X) \) and \( m(f_X) \) are related to \( \mu_{\text{max}}(A) \) and \( \mu_{\text{min}}(A) \) as follows:

\[
m(f_X) \geq \frac{1}{2\pi} \frac{\Lambda_{\text{min}}(\Sigma_{\epsilon})}{\mu_{\text{max}}(A)}, \quad \mathcal{M}(f_X) \leq \frac{1}{2\pi} \frac{\Lambda_{\text{max}}(\Sigma_{\epsilon})}{\mu_{\text{min}}(A)} \quad (10)
\]

Proposition 2.2 in Basu and Michailidis [2015] provides a lower bound on \( \mu_{\text{min}}(A) \). For
the special structure of the models considered here, we can get an improved upper bound on \( \mu_{\text{max}}(A) \), as shown in the following lemma:

**Lemma 3.1.** For a stable \( \text{VAR}(1) \) model of the class (6), we have

\[
\mu_{\text{max}}(A) \leq [1 + l + \left( v_{\text{in}} + v_{\text{out}} \right)/2]^2 \tag{11}
\]

where \( l \) is the largest singular value of \( L^0 \), \( v_{\text{in}} = \max_{1 \leq j \leq p} |S_{0j}^0| \) and \( v_{\text{out}} = \max_{1 \leq i \leq p} |S_{ij}^0| \).

**Proof.** \( \|A(z)\| = \|I - (L^0 + S^0)z\| \leq \|I\| + \|L^0\| + \|S^0\| \) for any \( z \in \mathbb{C} \) with \( |z| = 1 \). The result follows from the fact that \( \mu_{\text{max}}(A) = \max_{|z|=1} \|A(z)\|^2 \). \( \square \)

### 3.2 Estimation Procedure

Based on the data \( \{X^0, \ldots, X^T\} \) generated according to the model (6), we form the autoregressive design

\[
\begin{bmatrix}
(X^T)' \\
\vdots \\
(X^1)' \\
y'
\end{bmatrix} = \begin{bmatrix}
(X^{T-1})' \\
\vdots \\
(X^0) \\
x'
\end{bmatrix} A' + \begin{bmatrix}
(\epsilon^T)' \\
\vdots \\
(\epsilon^1)' \\
E'
\end{bmatrix} \tag{12}
\]

This is a linear regression problem with \( N = T \) samples and \( q = p^2 \) variables. The goal is to estimate \( L^0 \) and \( S^0 \) with high accuracy when \( N \ll p^2 \).

There is an inherent identifiability issue in the estimation of (6). Suppose the low-rank component \( L^0 \) itself is \( s \)-sparse and the sparse component \( S^0 \) is of rank \( r \). In that scenario, we cannot hope for any method to estimate \( L^0 \) and \( S^0 \) separately without imposing any further constraints. So, a minimal condition for low-rank and sparse recovery is that the low rank part should not be too sparse and the sparse part should not be low-rank.

This issue has been addressed in the literature by several authors [Chandrasekaran et al., 2011; Candès et al., 2011]. By and large, all the authors propose to ensure the above
identifiability under some form of incoherence type condition. These conditions serve as sufficient conditions for exact recovery of the low rank and the sparse component by solving the convex program (5). In a recent paper, Agarwal et al. [2012] showed that in a noisy setting where exact recovery of the two components is impossible, it is still possible to achieve good approximation under comparatively mild assumption. In particular, they formulated a general measure of the radius of nonidentifiability of the problem and established a non-asymptotic upper bound on the approximation error

\[ \| \hat{L} - L^0 \|_F^2 + \| \hat{S} - S^0 \|_F^2 \]  

which depend on this radius. The key idea is to allow for sparse and low-rank matrices in the model, but controlling for the error introduced. We refer the readers to the above paper for a more detailed discussion on this notion of non-identifiability. The low-rank and sparse decomposition problem under restrictions on the radius of nonidentifiability takes the form

\[ (\hat{L}', \hat{S}') = \arg\min_{L,S \in \mathbb{R}^{p \times p}, \|S\|_{\text{max}} \leq \alpha/p} \frac{1}{2} \| Y - X(L + S) \|_F^2 + \lambda_N \|L\|_* + \mu_N \|S\|_1 \]  

where \( \lambda_N, \mu_N \) are non-negative tuning parameters controlling the regularization of sparse and low-rank part. The parameter \( \alpha \) controls for degree of non-identifiable matrices allowed in the model class.

4 Theoretical Properties

In this section, we derive non-asymptotic upper bounds on the estimation errors of the low-rank and sparse components of the transition matrix. The main result shows that consistent estimation is possible with a sample size of the order \( N \sim p \mathcal{M}^2(f_X)/m^2(f_X) \), as long as the process \( \{X_t\} \) is stable, stationary and the radius of nonidentifiability, as measured by \( \|S^0\|_{\text{max}} \) is small in an appropriate sense.
We build upon the results of Agarwal et al. [2012] for fixed $X$ and $E$. In particular, it follows from Corollary 1 of the above paper that for a single realization of $\{X^0, \ldots, X^T\}$, for any $\alpha \geq \|S^0\|_{\text{max}}$, if $\gamma_N := \Lambda_{\text{min}}(X'X) > 0$, then any solution $(\hat{L}, \hat{S})$ of the convex program (14) with

$$\lambda_N \geq 4\|X'E\|, \quad \mu_N \geq 4\|X'E\|_{\text{max}} + \frac{4\gamma_N \alpha}{p}$$

(15)

satisfies, for some universal positive constants $c_i > 0$,

$$\|\hat{L} - L^0\|_F^2 + \|\hat{S} - S^0\|_F^2 \leq c_1 \frac{\lambda_N^2}{\gamma_N^2} r + c_2 \frac{\mu_N^2}{\gamma_N^2} s$$

(16)

In order to obtain meaningful results in the context of our problem, we need upper bounds on $\|X'E\|$ and $\|X'E\|_{\text{max}}$ and a lower bound on $\Lambda_{\text{min}}(X'X)$ that hold with high probability. In the context of time series where all the entries of the matrix $X'$ are dependent on each other, it is a non-trivial task to establish such deviation bounds. The main technical contribution of this chapter is to derive these deviation bounds, which lead to meaningful analysis in the context of VAR. The results rely on the measure of stability defined in (9) and an analysis of the joint spectrum of $\{X^{t-1}\}$ and $\{e^t\}$.

**Proposition 1.** Consider a random realization of $\{X^0, \ldots, X^T\}$ generated according to a stable VAR(1) process (6) and form the autoregressive design (12). Define

$$\phi(A, \Sigma_e) = \Lambda_{\text{max}}(\Sigma_e) \left[ 1 + \frac{1 + \mu_{\text{max}}(A)}{\mu_{\text{min}}(A)} \right]$$

Then there exist universal positive constants $c_i > 0$ such that

1. for $N \gtrsim p$,

$$\mathbb{P}\left[ \|X'E/N\| > c_0 \phi(A, \Sigma_e) \sqrt{p/N} \right] \leq c_1 \exp\left[ -c_2 \log p \right]$$

and for any $N \gtrsim \log p$,

$$\mathbb{P}\left[ \|X'E/N\|_{\text{max}} > c_0 \phi(A, \Sigma_e) \sqrt{\log p/N} \right] \leq c_1 \exp\left[ -c_2 \log p \right]$$
2. For \( N \gtrsim p \mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X) \),

\[
P \left[ \Lambda_{\min}(\mathbf{X}'\mathbf{X}/N) > \frac{\Lambda_{\min}(\Sigma_{\epsilon})}{2\mu_{\max}(\mathbf{A})} \right] \leq c_1 \exp\left[-c_2 \log p\right]
\]

Using the above deviation bounds in the non-asymptotic error (16), we obtain the final result for approximate recovery of the low-rank and the sparse components using nuclear and \( \ell_1 \) norm relaxation, as shown next.

**Proposition 2.** Consider the setup of Proposition 1. There exist universal positive constants \( c_i > 0 \) such that for \( N \gtrsim p \mathcal{M}^2(f_X)/\mathfrak{m}^2(f_X) \), for any \( S^0 \) with \( \|S^0\|_{\max} \leq \alpha \), any solution \((\hat{L}, \hat{S})\) of the program (14) satisfies, with probability at least \( 1 - c_1 \exp[-c_2 \log p] \),

\[
\|\hat{S} - S^0\|_F^2 + \|\hat{L} - L^0\|_F^2 \leq \frac{c_0 \phi^2(A, \Sigma_{\epsilon}) \mu_{\max}(\mathbf{A})}{\Lambda_{\min}(\Sigma_{\epsilon})} \frac{(rp + s \log p)}{N} + \frac{32 \Lambda_{\min}^2(\Sigma_{\epsilon}) \alpha^2}{\mu_{\max}^2(\mathbf{A})} \frac{s}{p^2} \tag{17}
\]

*Remarks.* The error bound presented in the above proposition consists of two key terms. The first term is the error of estimation emanating from randomness in the data and limited sample capacity. For a given model, this error goes to zero as the sample size increases. The second term represents the error due to the unidentifiability of the problem. This is more fundamental to the structure of the true low-rank and sparse components, depends only on the model parameters and does not change with sample size.

The error in estimation again consists of two terms - the second term \((rp + s \log p)/N\) consists of the dimensionality parameters and matches the parametric convergence rates for independent observations. The effect of dependence in the data is captured through the first part of the term: \( \frac{c_0 \phi^2(A, \Sigma_{\epsilon}) \mu_{\max}^2(\mathbf{A})}{\Lambda_{\min}^2(\Sigma_{\epsilon})} \). As discussed in Basu and Michailidis [2015], this term is larger when the spectral density is more spiky, indicating a stronger temporal and cross-sectional dependence in the data.
Table 1: Estimation Error $\|\hat{A} - A\|_F / \|A\|_F$ of OLS, lasso and low-rank+sparse estimates of a VAR(1) model $X^t = AX^{t-1} + \epsilon^t$. The transition matrix $A = L + S$ has a low rank component $L$ of rank 2 and a sparse component $S$ with 2−3% non-zero entries.

5 Numerical Experiments

In this section we conduct numerical experiments to assess the performance of low rank and sparse modeling in VAR analysis and compare it with the performances of ordinary least squares (OLS) and lasso estimates.

We consider three different VAR(1) models with $p = 30, 50$ and 100 variables. For each of these models, we generate $N = 50, 100, 200, 300$ and 500 observations from a Gaussian VAR(1) process $X^t = AX^{t-1} + \epsilon^t$, where $A = L + S$ can be decomposed into a low-rank matrix $L$ of rank 2 and a sparse matrix $S$ with $2−3\%$ non-zero entries. We rescale the entries of $A$ to ensure stability of the process (the spectral radius is set to $\rho(A) = 0.7$) and rescale the error variance so that $SNR = 2$. We compare the estimation and in-sample prediction error of the different estimates using two performance metrics:

1. Estimation Error: $\|\hat{A} - A\|_F / \|A\|_F$

2. In-sample Prediction Error: $\|\hat{Y} - Y\|_F^2 / \|Y\|_F^2$

The tuning parameters for lasso and low-rank plus sparse estimates are chosen according to Equation (15). We report median and IQR of the performance metrics from 50 iterations of the above experiments.
Table 2: In-sample prediction error $\|\hat{Y} - Y\|_F^2 / \|Y\|_F^2$ of OLS, lasso and low-rank+sparse estimates of a VAR(1) model $X_t = AX_{t-1} + \epsilon_t$. The transition matrix $A = L + S$ has a low rank component $L$ of rank 2 and a sparse component $S$ with $2 - 3\%$ non-zero entries.

<table>
<thead>
<tr>
<th>Prediction Error</th>
<th>OLS</th>
<th>lasso</th>
<th>low-rank+sparse</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p=30, N=50$</td>
<td>3.50(0.38)</td>
<td>1.00(0.02)</td>
<td>0.96(0.02)</td>
</tr>
<tr>
<td>$p=30, N=100$</td>
<td>1.53(0.08)</td>
<td>0.99(0.01)</td>
<td>0.96(0.02)</td>
</tr>
<tr>
<td>$p=30, N=200$</td>
<td>1.15(0.04)</td>
<td>0.99(0.01)</td>
<td>0.96(0.02)</td>
</tr>
<tr>
<td>$p=30, N=300$</td>
<td>1.08(0.02)</td>
<td>0.98(0.01)</td>
<td>0.95(0.02)</td>
</tr>
<tr>
<td>$p=30, N=500$</td>
<td>1.03(0.02)</td>
<td>0.98(0.01)</td>
<td>0.95(0.01)</td>
</tr>
<tr>
<td>$p=50, N=50$</td>
<td>-</td>
<td>1.01(0.02)</td>
<td>0.96(0.02)</td>
</tr>
<tr>
<td>$p=50, N=100$</td>
<td>2.51(0.16)</td>
<td>1.00(0.01)</td>
<td>0.97(0.01)</td>
</tr>
<tr>
<td>$p=50, N=200$</td>
<td>1.39(0.04)</td>
<td>1.00(0.00)</td>
<td>0.97(0.01)</td>
</tr>
<tr>
<td>$p=50, N=300$</td>
<td>1.21(0.02)</td>
<td>0.99(0.00)</td>
<td>0.97(0.01)</td>
</tr>
<tr>
<td>$p=50, N=500$</td>
<td>1.10(0.01)</td>
<td>0.99(0.01)</td>
<td>0.97(0.01)</td>
</tr>
<tr>
<td>$p=100, N=50$</td>
<td>-</td>
<td>1.02(0.01)</td>
<td>0.98(0.01)</td>
</tr>
<tr>
<td>$p=100, N=100$</td>
<td>-</td>
<td>1.01(0.01)</td>
<td>0.98(0.01)</td>
</tr>
<tr>
<td>$p=100, N=200$</td>
<td>2.50(0.05)</td>
<td>1.00(0.00)</td>
<td>0.99(0.01)</td>
</tr>
<tr>
<td>$p=100, N=300$</td>
<td>1.66(0.02)</td>
<td>1.00(0.00)</td>
<td>0.98(0.01)</td>
</tr>
<tr>
<td>$p=100, N=500$</td>
<td>1.29(0.01)</td>
<td>1.00(0.00)</td>
<td>0.98(0.00)</td>
</tr>
</tbody>
</table>

The estimation errors are reported in Table 1. In all the three settings, we find that the low-rank plus sparse VAR estimates outperform the estimates using ordinary least-squares and lasso. Among the three estimates, OLS has the worst estimation error with a high IQR, whereas the two regularized estimates produce lower estimation error with low IQR. For $p = 30$ and $N = 50$, the median estimation error of OLS is 7.20 with an IQR of 1.16, whereas lasso has an estimation error of 1.17 with an IQR of 0.09. The sparse plus low rank estimate has the lowest estimation error of 0.96 with an IQR of 0.07. The estimation errors of all three methods decrease with increase in sample sizes. We also report the error in estimating separately the low rank and the sparse components in the last column of Table 1.

The in-sample prediction errors of the three estimation methods are reported in Table 2. As in the case with estimation error, we see that low-rank plus sparse VAR estimates outperform OLS and lasso estimates in terms of prediction error in nearly all the settings. The prediction error of OLS for $p = 30$ and $N = 50$ is 3.50 with an IQR of 0.38, which indicates that the OLS prediction errors are 3.50 times larger than the errors from fitting a white noise model to the data (i.e., assuming $A = 0$). This effect of overfitting is lower in the lasso regularized estimates, where the prediction error (1.00) is of the same order of the white noise model with an IQR of 0.02. By accounting for a latent low-rank structure of
Figure 1: Estimated Granger causal networks using lasso and low-rank plus sparse VAR estimates. The top panel displays the true transition matrix $A$, its low-rank component $L$ and the structure of its sparse component $S$. The bottom panel displays the structure of the Granger causal networks estimated by lasso ($\hat{A}_{\text{lasso}}$), the low-rank plus sparse modeling strategy ($\hat{S}$) and the estimated low-rank component ($\hat{L}$).

the transition matrix, the low-rank plus sparse estimates produce a lower prediction error of 0.95 with an IQR of 0.03. The results of the other settings are qualitatively similar.

In addition to its improved estimation and prediction performance, the low-rank plus sparse modeling strategy help recover the underlying Granger causal network after accounting for the latent structure. In Figure 1, we demonstrate this using a VAR(1) model with $p = 50$ and $N = 500$. The top panel displays the true transition matrix $A$, its low-rank component $L$ and the structure of its sparse component $S$. The bottom panel displays the structure of the Granger causal networks estimated by lasso ($\hat{A}_{\text{lasso}}$), the low-rank plus sparse modeling strategy ($\hat{S}$) and the estimated low-rank component ($\hat{L}$). As predicted by the theory, we see that the lasso estimate of the Granger causal network, $\hat{A}_{\text{lasso}}$, selects many false positives due to its failure to account for the latent structure. On the other hand, the low-rank plus sparse estimate $\hat{S}$ provides a sparser estimate of the network with significantly less false positives.

It is interesting to note that the estimation performance of the regularized estimates in
low-rank plus sparse VAR models is worse than the performance of lasso in sparse VAR models of similar dimension [Basu and Michailidis, 2015], even for the same sample sizes. This is in line with the error bounds presented in Proposition 2. The estimation error in low-rank plus sparse models is of the order of $O(rp + s\log p)/N$ while the error of lasso in sparse VAR models scales at a faster rate of $O(s\log p/N)$. This can also be viewed in the factor model examples of Section 1. Using the notation of (1) and (3), a $s$-sparse VAR requires estimating $s$ parameters in $S$ while the presence of $r$ factors introduces an additional $rp$ parameters in the loading matrix $\Lambda$.

A Deviation Bounds

Proof of Proposition 1. 1. We want to find upper bounds on $\|X'E/N\|\max$ and $\|X'E/N\|$ that hold with high probability. Note that such an upper bound for $\|X'E/N\|\max$ has already been derived in Basu and Michailidis [2015]. Here we adopt a different technique that takes a unified approach to provide upper bounds on both quantities. To this end, note that the two norms have the following representations

$$\frac{1}{N}\|X'E\| = \sup_{u,v \in S^{p-1}} \frac{1}{N} u'X'Ev, \quad \frac{1}{N}\|X'E\|\max = \sup_{u,v \in \{e_1, \ldots, e_p\}} \frac{1}{N} u'X'Ev$$

For any given $u, v \in S^{p-1}$, we first provide a bound on $u'(X'E/N)v$.

Using Proposition 2.3 of Basu and Michailidis [2015], we obtain

$$\mathbb{P}\left(\|u'(X'E/N)v\| > 2\pi\eta\phi(A, \Sigma_c)\right) \leq 6 \exp\left[-cN \min\{\eta, \eta^2\}\right] \quad (18)$$

for any $u, v \in S^{p-1}$ and any $\eta > 0$.

To derive the deviation bound on $\|X'E/N\|\max$, we simply take a union bound over

16
the \( p^2 \) possible choices of \( u, v \in \{e_1, e_2, \ldots, e_p\} \). This leads to

\[
P[\|\mathcal{X}'E/N\|_{\text{max}} > 2\pi\eta\phi(A, \Sigma_\epsilon)] \leq 6 \exp[-cN \min\{\eta, \eta^2\} + 2 \log p]
\]

Since \( N \gtrsim p \), we can set \( \eta = \sqrt{(2 + c_1)\log p/cN} \) so that \( \eta < 1 \) (i.e., \( \eta^2 < \eta \)) will be satisfied for large enough \( N \). This implies that

\[
P[\|\mathcal{X}'E/N\|_{\text{max}} > c_0\phi(A, \Sigma_\epsilon)] \leq c_1 \exp[-c_2 \log p]
\]

for some universal constants \( c_i > 0 \).

To derive the deviation bound on the spectral norm, we discretize the unit ball \( S^{p-1} \) using an \( \epsilon \)-net \( \mathcal{N} \) of cardinality at most \( (1 + 2/\epsilon)^p \). An argument along the line of Supplementary Lemma F.2 in Basu and Michailidis [2015] then shows that for a small enough \( \epsilon > 0 \),

\[
\sup_{u, v \in S^{p-1}} |u'(\mathcal{X}'E/N)v| \leq K \sup_{u, v \in \mathcal{N}} |u'(\mathcal{X}'E/N)v|
\]

for some constant \( K > 1 \), possibly dependent on \( \epsilon \). As before, taking a union bound over the \( (1 + 2/\epsilon)^{2p} \) choices of \( u \) and \( v \), we get

\[
P[\|\mathcal{X}'E/N\| > 2\pi K \eta\phi(A, \Sigma_\epsilon)] \leq 6 \exp[-cN \min\{\eta, \eta^2\} + 2p \log(1 + 2/\epsilon)]
\]

Since \( N \gtrsim p \), choosing \( \eta = \sqrt{(c_1 + 2 \log(1 + 2/\epsilon))p/cN} \) ensures \( \eta < 1 \) for large enough \( N \). Setting \( \eta \) as above concludes the proof.

2. We want to obtain a lower bound on the minimum eigenvalue of \( \mathcal{X}'\mathcal{X}/N \) that holds with high probability.

Since \( \Lambda_{\min}(\mathcal{X}'\mathcal{X}/N) = \inf_{v \in S^{p-1}} v'(\mathcal{X}'\mathcal{X}/N)v \), we start with the single deviation bound
of Proposition 2.3 in Basu and Michailidis [2015],

\[ \mathbb{P} \left[ |v' (\mathcal{X}' \mathcal{X}/N - \Gamma X(0)) v| > 2 \pi \eta \mathcal{M}(f_X) \right] \leq 2 \exp \left[ -cN \min\{\eta, \eta^2\} \right] \]

for any \( v \in \mathbb{S}^{p-1} \) and \( \eta > 0 \).

The next step is to extend this single deviation bound uniformly on the set \( \mathbb{S}^{p-1} \). As in the proof of part 1, we construct a \( \epsilon \)-net of cardinality at most \( (1 + 2/\epsilon)^p \) and approximate the quadratic form using its values on the net. This yields the following deviation bound

\[ \mathbb{P} \left[ \sup_{v \in \mathbb{S}^{p-1}} |v' \left( \frac{\mathcal{X}' \mathcal{X}}{N} - \Gamma X(0) \right) v| > 2K \pi \eta \mathcal{M}(f_X) \right] \leq 2 \exp \left[ -cN \min\{\eta, \eta^2\} + p \log \left( 1 + \frac{2}{\epsilon} \right) \right] \]

for some constant \( K > 1 \). Setting \( \eta = m(f_X)/4K \pi \mathcal{M}(f_X) < 1 \) and noting that \( N \gtrsim \mathcal{M}^2(f_X)/m^2(f_X)p \), we conclude

\[ \mathbb{P} \left[ \sup_{v \in \mathbb{S}^{p-1}} |v' (\mathcal{X}' \mathcal{X}/N - \Gamma X(0)) v| > m(f_X)/2 \right] \leq c_0 \exp \left[ -c_1 \log p \right] \]

The result follows from the lower bound on \( m(f_X) \) presented in (10) and the fact that \( v' \Gamma X(0) v \geq m(f_X) \) for all \( v \in \mathbb{S}^{p-1} \).

\[ \square \]

References


