Non-Asymptotic Guarantees for Reliable Identification of Granger Causality via the LASSO

Proloy Das, *Student Member, IEEE,* and Behtash Babadi, *Member, IEEE*

***Abstract*—Granger causality is among the widely used data-driven approaches for causal analysis of time series data with applications in various areas including economics, molecular biology, and neuroscience. Two of the main challenges of this methodology are: 1) over-fitting as a result of limited data duration, and 2) correlated process noise as a confounding factor, both leading to errors in identifying the causal influences. Sparse estimation via the LASSO has successfully addressed these challenges for parameter estimation. However, the classical statistical tests for Granger causality resort to asymptotic analysis of ordinary least squares, which require long data duration to be useful and are not immune to confounding effects. In this work, we address this disconnect by introducing a LASSO-based statistic and studying its non-asymptotic properties under the assumption that the true models admit sparse autoregressive representations. We establish fundamental limits for reliable identification of Granger causal influences using the proposed LASSO-based statistic. We further characterize the false positive error probability and test power of a simple thresholding rule for identifying Granger causal effects and provide two methods to set the threshold in a data-driven fashion. We present simulation studies and application to real data to compare the performance of our proposed method to ordinary least squares and existing LASSO-based methods in detecting Granger causal influences, which corroborate our theoretical results.**

***Index Terms*—Granger Causality, LASSO, autoregressive models, non-asymptotic analysis.**

I. Introduction

RELIABLE identification of causal influences is one of the central challenges in time series analysis, with implications for various domains such as economics [1], neuroscience [2]–[4] and computational biology [5], [6]. Granger causal (GC) characterization of time series is among the widely used methods in this regard. This framework was pioneered by Granger [7], with subsequent key generalizations provided by Geweke [8], [9]. The notion of GC influence pertains to assessing the improvements in predicting the future samples of one time series by incorporating the past samples of another one.

While causality, as the relationship between cause and effect, is a philosophically well-defined concept [10], it eludes a universal definition in empirical sciences and engineering. Granger causality is one of many definitions used in time series models (see [11], [12] for other notions), with an explicit data-driven form that admits statistical testing. The stochastic nature of the time series model, i.e., the uncertainty is the central feature of GC definition. Unlike regression analysis that merely reflects correlational association, Granger's notion of causality probes if two time series are *temporally related* [13], in terms of the precedence established by the direction of time flow, i.e., if one time series precedes, and thus forecasts, another. In principle, given two time series x_t and y_t, one asserts that y_t has a GC influence on x_t when the posterior conditional densities $p(x_t|x_{t-1}, x_{t-2}, \cdots, y_{t-1}, y_{t-2}, \cdots)$ and $p(x_t|x_{t-1}, x_{t-2}, \cdots)$ differ significantly. However, estimating these posterior densities from the observed data is a difficult task in general, and requires additional modeling assumptions. A popular set of such assumptions pertains to parametric multivariate autoregressive (MVAR) models along with certain distributional specifications (e.g., zero-mean Gaussian process noise). In these models, the aforementioned posterior densities can be fully characterized by the estimates of parameters and prediction error variances. As such, the classical notion of Granger causality focuses on the prediction error variances: one first aims at predicting x_{t+1} by a linear combination of the joint past observations $\{x_t, \cdots x_0\}$, $\{y_t, \cdots y_0\}$ (i.e., the *full* model), followed by repeating this task by excluding the past observations of y_t (i.e., the *reduced* model). If the prediction error variance in the former case is significantly smaller than the latter, we say that y_t has a GC influence on x_t (See Section II-A for a formal definition of GC).

Conventionally, the optimal linear predictors in the mean square error sense, are obtained by the ordinary least squares (OLS), and the model orders are determined by the AIC [14] or BIC [15] procedures. Then, the GC measure is defined as the logarithmic ratio of the two prediction error variances, and its statistical significance is assessed based on the corresponding asymptotic distributions [16]–[18]. While the aforementioned procedure is relatively simple to carry out, it faces two key challenges. First, in order to obtain reliable MVAR parameter estimates via OLS, a relatively long observation horizon is required. In datasets with small sample size (e.g., gene expression data [19]), the regression models typically over-fit the observed data, causing both parameter estimation and model order selection to break down [4], [20]. In addition, AIC/BIC may restrict the order of the MVAR in a way that the resulting model fails to capture the complex and long-range dynamics of the underlying couplings [2], [21]. Secondly, correlated process noise arising from latent processes, may lead to misidentification of GC influences, which is often

P. Das is with the Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, 02114 USA (e-mail: pdas6@mgh.harvard.edu).

B. Babadi is with the Department of Electrical and Computer Engineering and the Institute for Systems Research, University of Maryland, College Park, MD, 20742 USA (e-mail: behtash@umd.edu).

This work was supported in part by the National Science Foundation Awards No. CCF1552946 and ECCS1807216 and the National Institutes of Health Award No. 1U19NS107464 (Corresponding author: Proloy Das).

referred to as the confounding effect [22].

These challenges have been successfully addressed in the context of regularized MVAR estimation [23]–[31]. In particular, the theory of sparse estimation via the LASSO [32]–[36] provides a principled methodology for simultaneous parameter estimation and model selection in high dimensional MVAR models [24]–[29], [37]–[39]. In addition, the Oracle property of the LASSO in presence of correlated noise ensures robust recovery of the set of MVAR parameters arising from the direct GC influences while discarding any spurious couplings due to correlated process noise, thus alleviating the confounding effect. The LASSO and its variants have already been utilized in existing work to identify graphical GC influences based on the recovered sparsity patterns [40]–[47]. These methods construct the GC graph based on the estimated model parameters, either directly [40] or by appropriate thresholding [45], [46] to control false positive errors. This idea has even been extended to time series models that account for nonlinear dynamics using structured multilayer perceptrons or recurrent neural networks [48]. Another related class of results uses debiasing techniques in order to construct confidence intervals and thereby identify the significant GC interactions [49]–[55]. There is, however, an evident disconnect between these LASSO-based approaches and the classical OLS-based GC inference: while the LASSO-based approaches aim at identifying the GC effects based on consistent *parameter estimates* in the non-asymptotic regime, the classical GC methodology relies on the comparison of the *prediction errors* across the *full* and *reduced* models by resorting to asymptotic distributions.

In this paper, we address the aforementioned disconnect between the currently available LASSO-based approaches and the classical OLS-based GC inference by unifying these two approaches via introducing a new LASSO-based GC statistic that resembles the classical GC measure, and by leveraging the consistency properties of the LASSO to characterize the non-asymptotic properties of said GC statistic. In particular, we consider a canonical bivariate autoregressive (BVAR) process with correlated process noise. We then propose a likelihood-based scaled F-statistic as the relevant GC statistic, which we call the LGC statistic, and study its non-asymptotic properties under both the presence and absence of a GC influence. Our analysis reveals that the well-known sufficient conditions of the LASSO for stable BVAR estimation are also sufficient for accurate detection of the GC influences, if the strength of the GC effect satisfies additional conditions. We also show that the conditions on the strength of the GC effect pose a fundamental limit for GC identification using LGC. Furthermore, we characterize the false positive error probability and test power of a simple thresholding scheme for identification of GC influences and provide two methods to set the threshold in a data-driven fashion.

We present extensive simulation studies to compare the performance of the proposed LGC-based method with the classical OLS-based and two of the existing LASSO-based approaches in detecting GC influences. These studies demonstrate the validity of our theoretical claims, explore the key underlying trade-offs, and evaluate our contributions in the context of existing work. We also present an application to experimentally-recorded neural data from general anesthesia to assess the GC influence of the local field potential (LFP) on spiking activity. Our results based on LGC analysis corroborate existing hypotheses on the GC influence of LFP in mediating local spiking activity, whereas these effects are concealed by the classical GC analysis due to significant overfitting.

In summary, our main contribution is to extend the theoretical results of the LASSO to the classical characterization of GC influences, to identify the key trade-offs in terms of sampling requirements and strength of the GC effects that result in reliable GC identification, to characterize the false positive error probability and test power of GC detection by thresholding LGC, and to provide data-driven methods to set the test threshold in practice.

The rest of this paper is organized as follows: Section II provides background and our problem formulation. Our main theoretical and methodological contributions are given in Section III. Section IV presents application to simulated and experimentally-recorded neural data, followed by our concluding remarks in Section V.

II. Background and Problem Formulation

Throughout this work, we use regular and boldface lowercase letters to denote scalars and vectors, respectively. Matrices are denoted by boldface uppercase letters. The transpose and Hermitian of a matrix $\mathbf{M}$ are denoted by $\mathbf{M}^\top$ and $\mathbf{M}^H$, respectively. The ℓ_α-norm of a vector $\mathbf{v} \in \mathbb{R}^n$ is denoted by $\|\mathbf{v}\|_\alpha = \left(\sum_{i=1}^n |v_i|^\alpha\right)^{1/\alpha}$ for $\alpha \geq 1$. In addition, we use the notation $\|\mathbf{v}\|_0 = \sum_{i=1}^n \mathbb{1}\left[v_i \neq 0\right]$ to denote the number of non-zero elements of the vector $\mathbf{v}$, and drop the subscript of the ℓ_2-norm of $\mathbf{v}$ and denote it by $\|\mathbf{v}\|$. For a matrix $\mathbf{M}$, the matrix norm induced by the ℓ_α-norm is defined as $\|\mathbf{M}\|_\alpha := \sup_{\mathbf{x}\neq\mathbf{0}} \|\mathbf{M}\mathbf{x}\|_\alpha / \|\mathbf{x}\|_\alpha$. We denote the minimum and maximum eigenvalues of a $\mathbf{M}$ by $\Lambda_{\min}(\mathbf{M})$ and $\Lambda_{\max}(\mathbf{M})$, respectively.

A. Granger Causality in a Canonical BVAR Regression Model

Consider finite-duration observations from two time series x_t and y_t, given by $\{x_t, y_t\}_{t=-p+1}^n$, where n is the sample size and p is the order. The BVAR(p) model can be expressed as:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \mathbf{A}_1 \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \mathbf{A}_2 \begin{bmatrix} x_{t-2} \\ y_{t-2} \end{bmatrix} + \cdots + \mathbf{A}_p \begin{bmatrix} x_{t-p} \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_t \\ \epsilon'_t \end{bmatrix}, \quad (1)$$

with $\mathbf{A}_i \in \mathbb{R}^{2\times 2}$ for $i \in \{1, 2, \cdots, p\}$ denoting the BVAR parameters and $[\epsilon_t, \epsilon'_t]^\top$ denoting the process noise with known distribution. It is commonly assumed that $[\epsilon_t, \epsilon'_t]^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\epsilon)$. Using this BVAR($p$) model and considering $\{x_t, y_t\}_{t=-p+1}^0$ as the initial condition, one can form a prediction model of x_t as follows:

$$\mathbf{x} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (2)$$

where the response $\mathbf{x}$, regressors $\mathbf{X}$, and residuals $\boldsymbol{\epsilon}$ are defined as in (3).

The regression coefficients $\boldsymbol{\theta}$ consist of $2p$ parameters: $\{\theta_i\}_{i=1}^p$, representing the autoregression coefficients obtained

from $(\mathbf{A}_i)_{1,1}$, $i = 1, 2, \cdots, p$ and $\{\theta_i\}_{i=p+1}^{2p}$ representing the cross-regression coefficients obtained from $(\mathbf{A}_i)_{1,2}$, $i = 1, 2, \cdots, p$. Hereafter, we denote the true coefficients by $\boldsymbol{\theta}^* \in \mathbb{R}^{2p}$. Also, for a generic coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^{2p}$, the corresponding auto- and cross-regression components are denoted by $\boldsymbol{\theta}_{(1)} \in \mathbb{R}^p$ and $\boldsymbol{\theta}_{(2)} \in \mathbb{R}^p$, respectively, i.e., $\boldsymbol{\theta} =: [\boldsymbol{\theta}_{(1)}; \boldsymbol{\theta}_{(2)}]$.

In the context of this bivariate model, let $\mathcal{U}_t$ be the sigma-field generated by all random variables up to and including time $t-1$, and $(\mathcal{U}\backslash y)_t$ be the sigma-field generated by all random variables up to and including time $t-1$, except $\{y_{t-1}, y_{t-2}, \cdots\}$. In addition, let the variance of a random variable, Z conditional on these sigma-fields be $\sigma^2(Z|\mathcal{U}_t)$ and $\sigma^2(Z|(\mathcal{U}\backslash y)_t)$. With this notation, Granger Causality (GC) is formally defined for Gaussian BVAR models as follows:

Definition 1 (Granger Causality [7])**.** We say y_t has a Granger causal link to x_t, if $\sigma^2(x_t|\mathcal{U}_t) < \sigma^2(x_t|(\mathcal{U}\backslash y)_t)$, i.e., we are better able to predict x_t using all available data than excluding y_t.

While Definition 1 applies to any time t, in practice the GC is assessed within blocks of the time-series data, e.g., for $0 \leq t \leq n$, since estimating the prediction variance reliably for all time points and using a single trial of the time-series data is challenging. A convenient framework to test the GC influence of y_t on x_t is to pose it as hypothesis testing, with the null hypothesis $H_{y\mapsto x,0} : \boldsymbol{\theta}^*_{(2)} = \mathbf{0}$. To this end, one considers the following BVAR(p) models:

$$\text{Full Model: } \mathbf{x} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{4a}$$

$$\text{Reduced Model: } \mathbf{x} = \mathbf{X}\widetilde{\boldsymbol{\theta}} + \widetilde{\boldsymbol{\epsilon}}, \text{ with } \widetilde{\boldsymbol{\theta}}_{(2)} = \mathbf{0}. \tag{4b}$$

In words, in the *full* model, all columns of $\mathbf{X}$ are used to estimate $\mathbf{x}$, but in the *reduced* model, only the first p columns are used. The conventional GC measure [7]–[9] is then defined as the logarithmic ratio of the residual variances: $\mathcal{F}_{y\mapsto x} := \log\left(\text{var}(\widetilde{\boldsymbol{\epsilon}})/\text{var}(\boldsymbol{\epsilon})\right)$. Note that when the residuals are Gaussian, $\mathcal{F}_{y\mapsto x}$ is the log-likelihood ratio statistic. Given that the *reduced* model is nested within the *full* model, we have $\mathcal{F}_{y\mapsto x} \geq 0$.

In order to compute $\mathcal{F}_{y\mapsto x}$ from the time series data, empirical residual variances are used based on OLS parameter estimates under both models [1]:

$$\widehat{\boldsymbol{\theta}}_{\text{OLS}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \ \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}\right\|^2, \tag{5a}$$

$$\widehat{\widetilde{\boldsymbol{\theta}}}_{\text{OLS}} = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}_{(2)}=\mathbf{0}}{\text{argmin}} \ \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}\right\|^2. \tag{5b}$$

The estimated $\mathcal{F}_{y\mapsto x}$ is a random variable over $\mathbb{R}_{\geq 0}$, and typically has a non-degenerate distribution. Thus, a non-zero $\mathcal{F}_{y\mapsto x}$ does not necessarily imply a GC influence. To control for false discoveries, the well-established results on the asymptotic normality of maximum likelihood estimators can be utilized: under mild assumptions, $n\mathcal{F}_{y\mapsto x}$ converges in distribution to a chi-square χ^2_p with degree p. In addition, under a sequence of local alternatives $H^n_{y\mapsto x,1} : \boldsymbol{\theta}^*_{(2)} = \boldsymbol{\delta}/\sqrt{n}$, for some constant vector $\boldsymbol{\delta}$, $n\mathcal{F}_{y\mapsto x}$ converges in distribution to a non-central chi-square $\chi^2_p(\nu)$ with degree p and non-centrality $\nu > 0$ [17], [18]. These asymptotic results lead to a simple thresholding strategy: rejecting the null hypothesis if $\mathcal{F}_{y\mapsto x}$ exceeds a fixed threshold. A key consideration in this framework is choosing the model order p. To this end, criteria such as the AIC [14] and BIC [15] are widely used to strike a balance between the variance accounted for and the number of coefficients to be estimated.

While the foregoing procedure works well in practice for large sample sizes, its performance sharply degrades as the sample size decreases. This performance degradation has two main reasons:

1) The regression models become under-determined and result in poor estimates of the parameters [28], and
2) The conventional model selection criteria fail to capture possible long-range temporal coupling of the underlying processes [27].

As a result, the classical GC measure is highly susceptible to over-fitting. In addition, when the process noise elements ϵ_t and ϵ'_t are highly correlated, the OLS estimates incur additional error in capturing the true BVAR parameters, and hence result in mis-detection of the GC influences. While some existing nonparametric methods aim at entirely bypassing MVAR estimation by utilizing spectral matrix factorization [56] or multivariate embedding [57] for system identification, they are similarly prone to the adverse effects of small sample size. It is noteworthy that there also exists a slew of partial correlation-based nonparametric methods that employ conditional independence tests for GC detection [58]–[60], thus avoiding time series modeling assumptions altogether.

B. LASSO-based GC Inference in the High-Dimensional Setting

In the so-called high-dimensional setting, where the model dimension becomes comparable to or even exceeds the sample size [34], regularization schemes are employed to guard against over-fitting. These schemes include Tikhonov regularization [61], [62], ℓ_1-regularization or the LASSO [32], [33], [35], [36], smoothly clipped absolute deviation [63], [64], Elastic-Net [65], and their variants, and have particularly proven useful in MVAR estimation [23]–[26], [30], [31]. Among these techniques, the LASSO has been widely used and studied in the high-dimensional sparse MVAR setting,

$$\mathbf{x} = \begin{bmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_1 \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} x_{n-1} & \cdots & x_{n-p} & y_{n-1} & \cdots & y_{n-p} \\ x_{n-2} & \cdots & x_{n-p-1} & y_{n-2} & \cdots & y_{n-p-1} \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ x_0 & \cdots & x_{-p+1} & y_0 & \cdots & y_{-p+1} \end{bmatrix}, \ \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_n \\ \epsilon_{n-1} \\ \vdots \\ \epsilon_1 \end{bmatrix}. \tag{3}$$

under fairly general assumptions [24]–[26]. By augmenting the least squares error loss with the ℓ_1-norm of the parameters, the LASSO simultaneously guards against over-fitting and provides automatic model selection [36]–[39], [66], [67], under the hypothesis that the true parameters are sparse. In the context of MVAR estimation, assuming that the time series data admit a sparse MVAR representation, the LASSO estimates enjoy tight bounds on the estimation and prediction errors under suitable sample size requirements, even for models with correlated noise [28], [29].

By leveraging the foregoing properties, the LASSO and its variants have been utilized in existing work to identify GC influences in a graphical fashion [40]–[46]. Fujita et al. [41] used sparse autoregressive models to identify the structure of gene regulatory networks; Arnold et al. [40] provided a formal treatment of LASSO-based Granger causality detection using sparse parameter estimates. Lozano et al. [42] and Shimamura et al. [47] used Group LASSO and Elastic-Net penalties, respectively, to reduce the number of false positives while maintaining a high true positive rate in network inference. Finally, Shojaie and Michailidis [43] introduced a truncating LASSO penalty in order to identify the correct order of the VAR model and thereby enhance the reliability of GC discovery. To deal with time series data with extreme events, i.e., exhibiting heavy-tailed distributions, Liu et al. [44] utilized the theory of extreme value modeling to modify the distributional assumptions.

The aforementioned methods translate the non-zero values of estimated parameters (in a group-wise sense) to GC either directly or after appropriate pruning [45] (See [46] for a through overview). Alternatively, de-biasing techniques have been introduced for constructing confidence intervals over the estimated parameters and thereby identifying the significant GC effects [49]–[55] to maintain high number of true positives while reducing false discoveries.

C. Unifying the Classical OLS-based and LASSO-based Approaches

Comparing the classical OLS-based GC and the recent LASSO-based approaches to GC inference reveals an evident disconnect: the latter approach directly utilizes the *estimated parameters* from a *single model* to identify the GC influence with non-asymptotic performance guarantees, while the former is based on comparing the *prediction error* performance of *two different models* by resorting to asymptotic distributions for statistical testing.

Our main objective here is address this disconnect by unifying these two approaches. To this end, we first replace OLS estimation in (5b) by its LASSO counterpart:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \frac{1}{n} \left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}\right\|^2 + \lambda_n \|\boldsymbol{\theta}\|_1 \tag{6a}$$

$$\widehat{\widetilde{\boldsymbol{\theta}}} = \underset{\boldsymbol{\theta}:\boldsymbol{\theta}_{(2)}=\mathbf{0}}{\operatorname{argmin}} \ \frac{1}{n} \left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}\right\|^2 + \lambda_n \|\boldsymbol{\theta}\|_1, \tag{6b}$$

where λ_n denotes the regularization parameter. Let

$$\ell(\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}) := \frac{1}{n} \left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}\right\|^2 \ \text{ with } \boldsymbol{\theta} = [\boldsymbol{\theta}_{(1)}; \boldsymbol{\theta}_{(2)}].$$

By similarly grouping the solutions of (6) as $\widehat{\boldsymbol{\theta}} = [\widehat{\boldsymbol{\theta}}_{(1)}; \widehat{\boldsymbol{\theta}}_{(2)}]$ and $\widehat{\widetilde{\boldsymbol{\theta}}} = [\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}; \mathbf{0}]$, we then propose to use the following statistic:

$$\mathcal{T}_{y\mapsto x} := \frac{\ell\Big(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\Big)}{\ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\Big)} - 1 = \frac{\ell\Big(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\Big) - \ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\Big)}{\ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\Big)}, \tag{7}$$

akin to a scaled likelihood-based version of the F-statistic, which we call the LASSO-based GC (LGC) statistic. Note that the LGC statistic can be related to the conventional GC statistic as $\mathcal{T}_{y\mapsto x} = \exp(\mathcal{F}_{y\mapsto x}) - 1$, when $\lambda_n = 0$. Therefore, it is expected for $\mathcal{T}_{y\mapsto x}$ to be near 0 under the null hypothesis. One advantage of using this statistic is that a simple thresholding strategy, similar to that used for the classical GC statistic, can be used to reject the null hypothesis $H_{y\mapsto x,0} : \boldsymbol{\theta}^*_{(2)} = \mathbf{0}$. In the next section, we will characterize the non-asymptotic properties of the LGC statistic and seek conditions that allow us to distinguish between the null (i.e., absence of a GC effect) and a suitably defined alternative (i.e., presence of a GC effect) hypothesis.

To evaluate the benefits of unifying these two approaches, i.e., using the robust *estimation* performance of the LASSO in comparing the *prediction error* performance of the full and reduced models, we will present a simulation study in Section IV-A to compare LGC-based GC identification with two existing LASSO-based approaches, namely, the confidence interval (CI)-based LASSO [50], [53], and truncating LASSO (TLASSO). Our simulation results show that when the strength of the GC link is relatively weak, the CI-based LASSO method results in a high false positive rate and TLASSO exhibits low true positive rate; LGC, however, strikes a reasonable balance between true and false positive discovery, in accordance with our main objective.

III. Theoretical Results

Our main theoretical contribution in this section is to characterize $\mathcal{T}_{y\mapsto x}$ under both the null and a suitably chosen alternative hypothesis, and establish sufficient conditions that guarantee distinguishing these hypotheses with high probability. Furthermore, we show that some of these conditions are indeed necessary and thus pose a fundamental limit for the separation of null and alternative hypotheses. We then analyze the false positive error probability corresponding to the aforementioned thresholding strategy, under slightly weakened sufficient conditions. The latter result can be used to obtain suitable thresholds in practice, as we will discuss in Section III-F and demonstrate in Section IV. Before presenting the main results, we state our key assumptions and discuss their implications in the following section.

A. Key Assumptions and Their Implications

We adopt the following assumption on the BVAR process from [28]:

Assumption 1. The $\{x_t, y_t\}_{t=-p+1}^{n}$ is a part of a realization of zero-mean bivariate process that admits a stable and invertible

BVAR(p) representation, with a zero-mean i.i.d. Gaussian process noise with positive definite covariance $\mathbf{\Sigma}_\epsilon$. Further, the initial condition of the process is such that the samples under consideration attain the stationary distribution.

To elaborate on the implications of Assumption 1 for the second order statistics of the BVAR process, let $\mathbf{\Gamma}(l)$ be the auto-covariance matrix of the BVAR process at lag l and

$$\mathbf{F}(\omega) := \frac{1}{2\pi} \sum_{l=-\infty}^{\infty} \mathbf{\Gamma}(l) \exp(-il\omega)$$

be its spectral density matrix. It can be shown that the spectral density exists, if $\sum_{l=0}^{\infty} \|\mathbf{\Gamma}(l)\|_2^2 < \infty$. Furthermore, if $\sum_{l=0}^{\infty} \|\mathbf{\Gamma}(l)\|_2 < \infty$, the spectral density is bounded and continuous, so that the essential supremum is indeed achieved [28].

For the BVAR(p) process in (1), the matrix valued characteristic polynomial is defined as $\mathbf{A}(z) := \mathbf{I} - \sum_{j=1}^{p} \mathbf{A}_j z^j$. Then, the following consequences of Assumption 1 provide a simple characterization of the spectral density matrix [28]:

1) The process noise covariance matrix $\mathbf{\Sigma}_\epsilon$ is positive definite with bounded eigen-values, i.e.,
$$0 < \Lambda_{\min}(\mathbf{\Sigma}_\epsilon) \leq \Lambda_{\max}(\mathbf{\Sigma}_\epsilon) < \infty.$$
2) The BVAR process is stable and invertible, i.e., $\det(\mathbf{A}(z)) \neq 0$ on or inside the unit circle, $\{z \in \mathbb{C} : |z| \leq 1\}$.

Under these two conditions, the spectral density matrix exists, and its maximum eigen-value is bounded almost everywhere on $[-\pi, \pi]$, i.e.,

$$\mathcal{M}(\mathbf{F}) := \operatorname*{ess\,sup}_{\omega \in [-\pi,\pi]} \Lambda_{\max}(\mathbf{F}(\omega)) < \infty.$$

Furthermore, it is bounded and continuous, and admits the representation:

$$\mathbf{F}(\omega) = \frac{1}{2\pi} \mathbf{A}^{-1}(\exp(-i\omega)) \mathbf{\Sigma}_\epsilon \mathbf{A}^{-H}(\exp(-i\omega)).$$

Additionally, consider the infimum of the spectral density over unit circle:

$$m(\mathbf{F}) := \operatorname*{ess\,inf}_{\omega \in [-\pi,\pi]} \Lambda_{\min}(\mathbf{F}(\omega)). \tag{8}$$

Then, the following useful bounds hold for the BVAR(p) process in (1) [28]:

$$\mathcal{M}(\mathbf{F}) \leq \frac{1}{2\pi} \frac{\Lambda_{\max}(\mathbf{\Sigma}_\epsilon)}{\mu_{\min}(\mathbf{A})}, \qquad m(\mathbf{F}) \geq \frac{1}{2\pi} \frac{\Lambda_{\min}(\mathbf{\Sigma}_\epsilon)}{\mu_{\max}(\mathbf{A})}, \tag{9}$$

where $\mu_{\max}(\mathbf{A}) := \max_{|z|=1} \Lambda_{\max}\left(\mathbf{A}^H(z)\mathbf{A}(z)\right)$ and $\mu_{\min}(\mathbf{A}) := \min_{|z|=1} \Lambda_{\min}\left(\mathbf{A}^H(z)\mathbf{A}(z)\right)$.

We note that the characteristic polynomial $\mathbf{A}(z)$ encodes the temporal dependencies of the process, whereas $\mathbf{\Sigma}_\epsilon$ captures the correlation between the process noise components, possibly due to latent processes. Expressing the error bounds in our theoretical analysis in terms of $\mu_{\max}(\mathbf{A}), \mu_{\min}(\mathbf{A}), \Lambda_{\max}(\mathbf{\Sigma}_\epsilon), \Lambda_{\min}(\mathbf{\Sigma}_\epsilon)$, instead of $\mathcal{M}(\mathbf{F})$ and $m(\mathbf{F})$, helps to distinguish the contributions of these two sources of BVAR dependencies [28] (See Appendix A).

We also consider the $2p$-dimensional alternative BVAR(1) representation of the 2-dimensional BVAR(p) process: $\breve{\mathbf{X}}_t = \breve{\mathbf{A}}_1 \breve{\mathbf{X}}_{t-1} + \breve{\epsilon}$, where $\mathbf{X}_t$ is the first the row of $\mathbf{X}$ in (3) organized as a column vector, and $\breve{\mathbf{A}}_1$ and $\breve{\epsilon}$ are constructed by the corresponding augmentation of $\mathbf{A}_i$'s and ϵ_t's, respectively. The process $\breve{\mathbf{X}}_t$ has a characteristic polynomial, $\breve{\mathbf{A}}(z) := \mathbf{I} - \breve{\mathbf{A}}_1 z$ and is stable if and only if the original process is stable [68].

The remaining component of our key assumptions is the following sparsity assumption, frequently arising in high-dimensional regime, specially in LASSO literature.

Assumption 2. The regression coefficients in (2), $\boldsymbol{\theta}^*$ is k-sparse, i.e., $\|\boldsymbol{\theta}^*\|_0 = k$.

The implication of this assumption for the *full* model is fairly standard, under both the null and alternative hypotheses. However, for the *reduced* model under the alternative hypothesis, where only the autoregressive parameters are unspecified and the cross-regression parameters are enforced to be $\mathbf{0}$, we need to define a suitable surrogate "true" model whose parameters can be used to quantify the estimation error and thereby establish concentration bounds (See Proposition A.4). Let the columns of $\mathbf{X}$ corresponding to $\boldsymbol{\theta}_{(i)}$ be denoted by $\mathbf{X}_{(i)}$, for $i = 1, 2$. Using the fact that the error residuals of the optimal linear estimator, in the mean square error sense, are uncorrelated with the columns of the design matrix [69, pp. 386]), we define the surrogate "true" autoregression coefficients in the *reduced* model as:

$$\widetilde{\boldsymbol{\theta}}^*_{(1)} := \boldsymbol{\theta}^*_{(1)} + \mathbf{C}_{11}^{-1} \mathbf{C}_{12} \boldsymbol{\theta}^*_{(2)}, \tag{10}$$

where

$$\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top \mathbf{X}\right] = \begin{bmatrix} \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(1)}^\top \mathbf{X}_{(1)}\right] \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(1)}^\top \mathbf{X}_{(2)}\right] \\ \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(2)}^\top \mathbf{X}_{(1)}\right] \mathbb{E}\left[\frac{1}{n}\mathbf{X}_{(2)}^\top \mathbf{X}_{(2)}\right] \end{bmatrix} =: \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} =: \mathbf{C}. \tag{11}$$

Note that even though the MVAR coefficients under the alternative hypothesis are k-sparse, the surrogate "true" autoregression coefficients under the *reduced* model may not be. To deal with this issue, we follow the treatment of [35] in analyzing the LASSO under weakly sparse or compressible parameters, and further impose a norm condition on $\boldsymbol{\theta}^*_{(2)}$ (See the statement of Theorem 1) to restrict the alternative hypothesis. The latter ensures that the *full* and *reduced* models are distinguishable under the alternative hypothesis.

Finally, we use the following definition throughout the rest of the paper:

Definition 2. We say that a threshold t *correctly distinguishes* the null and alternative hypotheses, if the ranges of the statistic $\mathcal{T}$ under the null and alternative hypotheses are respectively subsets of $\mathcal{T} < t$ and $\mathcal{T} > t$.

B. Main Theoretical Results

Our main theorem is stated as follows:

Theorem 1 (Main Theorem)**.** *Suppose that the key assumptions 1 and 2 hold. Then, for the proposed LGC statistic $\mathcal{T}_{y\mapsto x}$ evaluated at the BVAR(p) parameter estimates from the solutions of* (6) *with a regularization parameter $\lambda_n = 4\mathscr{A}\sqrt{\log(2p)/n}$, there exists a threshold that correctly distinguishes the null and the local alternative hypothesis $H^n_{y\mapsto x,1} : \|\boldsymbol{\theta}^*_{(2)}\|_2^2 \geq \mathscr{B}k\log(2p)/n$ with probability at least $1 - K_1\exp(-n\bar{c}) - K_2/p^{\bar{d}}$, if $n \geq \max\{\mathscr{C}'', \mathscr{D}''k\}\log(2p)$, with $\mathscr{A}$, $\mathscr{B}$, $\mathscr{C}''$, $\mathscr{D}''$, $K_1, K_2, \bar{c}$, and $\bar{d} > 0$ denoting constants that are explicitly given in the proof.*

Proof. The proof has three main steps. First, we bound the deviation of the empirical quantities $\ell(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)})$ (*full* model) and $\ell(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0})$ (*reduced* model) with respect to their counterparts evaluated at the true parameters. After invoking suitable concentration results for $\ell(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)})$ and $\ell(\widehat{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0})$, we can then lower bound $\mathcal{T}_{y\mapsto x}$ under the alternative hypothesis and upper bound it under the null hypothesis. Secondly, we seek conditions under which the bounds do not coincide, which further restricts the alternative hypothesis. The last step of the proof establishes that these deviation and concentration results indeed hold with high probability.

Step 1. We first assume the deviation bounds:

$$\left|\ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\Big) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big)\right| \leq \Delta_F, \tag{12}$$

$$\left|\ell\Big(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\Big) - \ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right)\right| \leq \Delta_R, \tag{13}$$

and the concentration inequalities:

$$\left|\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - (\boldsymbol{\Sigma}_\epsilon)_{1,1}\right| \leq \Delta_N \tag{14}$$

$$\left|\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - D\right| \leq \Delta_D \tag{15}$$

hold for some non-negative quantities, Δ_F, Δ_R, Δ_N, Δ_D and $D := {\boldsymbol{\theta}^*_{(2)}}^\top(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\boldsymbol{\theta}^*_{(2)}$. Using the bounds in (12) and (13), we get:

$$\frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \Delta_R}{\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) + \Delta_F} \leq \frac{\ell\Big(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\Big)}{\ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\Big)} \leq \frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) + \Delta_R}{\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - \Delta_F}, \tag{16}$$

which gives the following lower and upper bounds on $\mathcal{T}_{y\mapsto x}$:

$$\begin{aligned}\frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - \Delta_R - \Delta_F}{\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) + \Delta_F} &\leq \mathcal{T}_{y\mapsto x} \\ &\leq \frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) + \Delta_R + \Delta_F}{\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - \Delta_F}.\end{aligned} \tag{17}$$

Now, under the null hypothesis $H_{y\mapsto x,0} : \boldsymbol{\theta}^*_{(2)} = \mathbf{0}$, we have $\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) = \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big)$ and thus Δ_R in Proposition A.2 can be chosen equal to Δ_F in Proposition A.1. This implies:

$$\mathcal{T}_{y\mapsto x} \leq \frac{\Delta_R + \Delta_F}{\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - \Delta_F} \leq \frac{2\Delta_F}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} - \Delta_N - \Delta_F}, \tag{18}$$

with application of (14).

On the other hand, under a general alternative hypothesis $H_{y\mapsto x,0} : \boldsymbol{\theta}^*_{(2)} \neq \mathbf{0}$, (15) can be used to show that:

$$\begin{aligned}\mathcal{T}_{y\mapsto x} &\geq \frac{\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\right) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - \Delta_R - \Delta_F}{\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) + \Delta_F} \\ &\geq \frac{D - (\Delta_D + \Delta_R + \Delta_F)}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} + \Delta_N + \Delta_F}.\end{aligned} \tag{19}$$

From the upper and lower bounds in (18) and (19), it is possible to choose a threshold to correctly distinguish the two hypotheses, if:

$$\frac{D - (\Delta_D + \Delta_R + \Delta_F)}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} + \Delta_N + \Delta_F} > \frac{2\Delta_F}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} - \Delta_N - \Delta_F}, \tag{20}$$

which after rearrangement translates to:

$$D > \Delta_D + \Delta_R + \Delta_F\left(1 + 2\frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1} + \Delta_N + \Delta_F}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} - (\Delta_N + \Delta_F)}\right). \tag{21}$$

Next, we choose $\Delta_N = (\boldsymbol{\Sigma}_\epsilon)_{1,1}/4$. Then, assuming $\Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1}/4$, the bound of (21) further simplifies to

$$D > \Delta_D + \Delta_R + 7\Delta_F. \tag{22}$$

Step 2. Next, we assume that the following conditions hold:

Condition 1 (Restricted eigenvalue (RE) condition)**.** The symmetric matrix $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^\top\mathbf{X}/n \in \mathbb{R}^{2p\times 2p}$ satisfies restricted eigenvalue condition with curvature $\alpha > 0$ and tolerance $\tau \geq 0$, i.e., $\widehat{\boldsymbol{\Sigma}} \sim \text{RE}(\alpha, \tau)$:

$$\boldsymbol{\phi}^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\phi} \geq \alpha\|\boldsymbol{\phi}\|_2^2 - \tau\|\boldsymbol{\phi}\|_1^2, \;\; \forall \;\; \boldsymbol{\phi} \in \mathbb{R}^{2p},$$

with

$$\tau := \frac{m-1}{m}\frac{\alpha}{32k}$$

for some constant $m > 1$.

Condition 2 (Deviation condition)**.** There exist deterministic functions $\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)$, $\mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)$ such that

$$\left\|\frac{1}{n}\mathbf{X}^\top(\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*)\right\|_\infty \leq \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}},$$

$$\left\|\frac{1}{n}\mathbf{X}_{(1)}^\top\left(\mathbf{x} - \mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)\right\|_\infty \leq \mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}}.$$

Under these conditions, we can use the expressions for the non-negative quantities, Δ_F, Δ_R and Δ_D derived in Propositions A.1 and A.2 (Appendix A) and Lemma C.3 (Appendix C), respectively, to obtain the following bound on the right hand side of (22):

$$\begin{aligned}\Delta_D + \Delta_R + 7\Delta_F \leq\; &\mathscr{a}\sqrt{\frac{\log(2p)}{n}}\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2 \\ &+ \mathscr{b}\sqrt{\frac{k\log(2p)}{n}}\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2 + \mathscr{c}\frac{k\log(2p)}{n},\end{aligned}$$

where

$$\mathscr{a} = \frac{\alpha}{27}\left(\left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\|_2^2 + 1\right),$$

$$\mathscr{b} = \mathscr{A}\left[\left(32\sqrt{2m}+73\right)\left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\|+1\right],$$
$$\mathscr{c} = \frac{16\mathscr{A}^2}{\alpha/m}\left(\frac{168}{m+1}+20\right),$$

provided the LASSO problems are solved with the choice of $\lambda_n = 4\mathscr{A}\sqrt{\log(2p)/n}$, for $\mathscr{A}$ satisfying (See Propositions B.2 and B.3 in Appendix B):

$$\mathscr{A} \geq \max\left\{\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon), \mathbb{Q}'(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\right\}.$$

Also, note that the assumption $\Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1}/4$ requires:

$$\frac{24}{m+1}\frac{k\lambda_n^2}{\alpha/m} \leq \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{4}. \tag{23}$$

and imposes an upper bound on λ_n. The implications of this upper bound are further discussed in *Remark 1* at the end of this section.

Next, we use the following lower bound on D: $D \geq \widetilde{\Lambda}_{\min}\|\boldsymbol{\theta}^*_{(2)}\|_2^2$, with $\widetilde{\Lambda}_{\min} := \Lambda_{\min}\left(\mathbf{C}_{22}-\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)$, which gives the following sufficient condition for inequality (22) to hold:

$$\widetilde{\Lambda}_{\min}\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2 \geq \mathscr{a}\sqrt{\frac{\log(2p)}{n}}\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2^2 + \mathscr{b}\sqrt{\frac{k\log(2p)}{n}}\left\|\boldsymbol{\theta}^*_{(2)}\right\|_2 + \mathscr{c}\frac{k\log(2p)}{n}. \tag{24}$$

By further requiring $n \geq \mathscr{C}'\log(2p)$, where

$$\mathscr{C}' = \left(\frac{2\alpha\left(\left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\|_2^2+1\right)}{27\widetilde{\Lambda}_{\min}}\right)^2,$$

we have $\widetilde{\Lambda}_{\min} - \mathscr{a}\sqrt{\log(2p)/n} \geq \widetilde{\Lambda}_{\min}/2$. The latter combined with (24), and an application of Lemma C.5 (Appendix C) gives the sufficient condition: $\|\boldsymbol{\theta}^*_{(2)}\|_2^2 \geq \mathscr{B}k\log(2p)/n$ for unambiguous discrimination of the LGC statistic under the null and the *local* alternative hypothesis $H^n_{y\mapsto x,0}: \|\boldsymbol{\theta}^*_{(2)}\|_2^2 \geq \mathscr{B}k\log(2p)/n$, as long as $n \geq \max\{\mathscr{C}', \mathscr{D}'k\}\log(2p)$, where

$$\mathscr{B} := \left(\frac{4\mathscr{b}^2}{\widetilde{\Lambda}_{\min}^2}+\frac{4\mathscr{c}}{\widetilde{\Lambda}_{\min}}\right), \text{ and } \mathscr{D}' := \frac{1536m}{m+1}\frac{\mathscr{A}^2}{\alpha(\boldsymbol{\Sigma}_\epsilon)_{1,1}}.$$

Note that the condition $n \geq \mathscr{D}'k\log(2p)$ ensures the upper bound (23) on λ_n.

Step 3. It only remains to provide a lower bound on the probability of the event that the proposed LGC statistic under the null and *local* alternative hypotheses are correctly distinguishable by a threshold. Proposition B.1 (Appendix B) establishes that Condition 1 holds with probability at least

$$1 - c_1\exp(-c_2 n\min\{\zeta^{-2}, 1\}), \tag{25}$$

if $n \geq C_0\max\{\zeta^2, 1\}k\log(2p)$, for some constants C_0, c_1, c_2, and ζ (>0). Also, Propositions B.2 and B.3 (Appendix B) establish that Condition 2 holds with probability at least

$$1 - \frac{d_1}{(2p)^{d_2}} - \frac{d_1'}{(2p)^{d_2'}}, \tag{26}$$

if $n \geq \max\{D_0, D_0'\}\log(2p)$, for some constants d_1, d_1', d_2, d_2', D_0, and D_0' (>0).

From Lemma C.2 (Appendix C), the first deviation bound (12) holds with probability $1-2\exp(-n/128)$ with the choice $\Delta_N = (\boldsymbol{\Sigma}_\epsilon)_{1,1}/4$. Lastly, from Lemma C.3 (Appendix C), the deviation bound (12) holds with probability $1-c_3\exp(-c_4 n\min\{\zeta^{-2}\log(2p)/n, 1\})$, under Condition 2.

Combining the two steps, the claim of the theorem holds with probability at least

$$1 - 2\exp\left(-\frac{n}{128}\right) - c_1\exp(-c_2 n\min\{\zeta^{-2}, 1\}) - c_3\exp\left(-c_4 n\min\{\zeta^{-2}\max\{D_0, D_0^{'}\}, 1\}\right) - \frac{d_1}{(2p)^{d_2}} - \frac{d_1'}{(2p)^{d_2'}}, \tag{27}$$

if $n \geq \max\{\mathscr{C}'', \mathscr{D}''k\}\log(2p)$, where $\mathscr{D}'' := \max\{\mathscr{D}', C_0\max\{\zeta^2, 1\}\}$ and $\mathscr{C}'' = \max\{\mathscr{C}', D_0, D_0'\}$. Finally, this probability can be lower bounded by

$$1 - K_1\exp(-n\bar{c}) - \frac{K_2}{p^{\bar{d}}},$$

where

$$\bar{c} = \min\left\{\frac{1}{128}, c_2, c_4, c_2\zeta^{-2}, c_4\zeta^{-2}\max\{D_0, D_0^{'}\}\right\},$$
$$\bar{d} = \min\left\{d_2, d_2^{'}\right\},$$
$$K_1 = 2 + c_1 + c_3, \qquad K_2 = \frac{d_1}{2^{d_2}} + \frac{d_1'}{2^{d_2'}}.$$

This concludes the proof of the main theorem. □

From the proof of Theorem 1, it is apparent that large enough sample size n is sufficient for the LASSO estimates under both the full and reduced models to be consistent given the choice of λ_n, which in turn allows us to control the prediction errors. Also, large enough $\|\boldsymbol{\theta}^*_{(2)}\|_2$ is sufficient to distinguish the null and alternative hypotheses.

In order to establish the fundamental limit for separating the null and alternative hypotheses, however, we need to inspect the necessity of these conditions. From the proof of Theorem 1, for any threshold t, there exists a constant $\mathscr{A}'$, such that if $\lambda_n < 4\mathscr{A}'\sqrt{\log(2p)/n}$, then $\mathcal{T}_{y\mapsto x} > t$ with high probability, which results in a type I error. Thus, the choice of λ_n in Theorem 1 is necessary to control false positives (See Remark 1 in Section III-F).

In what follows, we focus on the necessity of the lower bound on $\|\boldsymbol{\theta}^*_{(2)}\|_2$ and the condition on n for detecting a Granger causal link. To prove this necessity result, we need a lower bound on the performance of the LASSO under the alternative hypothesis. To obtain a lower bound, we take the approach of [27], [70] by constructing a family of BVAR processes with k-sparse parameters $\boldsymbol{\theta}_{(2)}$ for which the LASSO makes significant errors under the alternative hypothesis. The remaining key element of the proof is the Fano's inequality, which provides the required lower bound (See Lemma C.4):

Theorem 2 (Necessary Conditions)**.** *Suppose that the key assumptions 1 and 2 hold. Then, there exists a constant $\widetilde{\mathscr{B}}$ for which the proposed LGC statistic $\mathcal{T}_{y\mapsto x}$ in Theorem 1 fails to reject the null hypothesis for any threshold t under the alternative hypothesis $H^n_{y\mapsto x,1}$: $\|\boldsymbol{\theta}^*_{(2)}\|_2^2 \leq \widetilde{\mathscr{B}}k\log p/n$, with*

probability at least $1/2$*, if* $n < \widetilde{\mathscr{D}}' k \log p$*. The constants* $\widetilde{\mathscr{B}}$ *and* $\widetilde{\mathscr{D}}'$ *are explicitly given in the proof.*

Proof. Consider a class $\mathcal{Z}$ of BVAR processes with a fixed 1-sparse $\boldsymbol{\theta}_{(1)}$ and k-sparse $\boldsymbol{\theta}_{(2)}$ over any subset $\mathcal{K} \subset \{1, 2, \cdots, p\}$ satisfying $|\mathcal{K}| = k$, with elements given by

$$\left(\boldsymbol{\theta}_{(2)}\right)_\ell = \pm e^{-b}, \forall \ell \in \mathcal{K}. \tag{28}$$

where b is a constant to be chosen. We also assume, without loss of generality, that $\left(\boldsymbol{\theta}_{(1)}\right)_1 = e^{-b}$ or 0. Note that the vector $\boldsymbol{\theta} = [\boldsymbol{\theta}_{(1)}, \boldsymbol{\theta}_{(2)}]$ is $(k+1)$-sparse, but the following arguments can be repeated by redefining k as $k-1$, pertaining to k-sparse parameter vectors. We also add the vector of all zeros to $\mathcal{Z}$. For a fixed $\mathcal{K}$, we have $2^{k+1}+1$ such parameters forming a subfamily $\mathcal{Z}_\mathcal{K}$. Consider the maximal collection of $\binom{p}{k}$ subsets $\mathcal{K}$ for which any two subsets differ in at least $k/4$ indices. The size of this collection can be identified by $A(p, \frac{k}{4}, k)$ in coding theory, where $A(n, d, w)$ represents the maximum size of a binary code of length n with minimum distance d and constant weight w [71]. It can be shown that

$$A(p, \tfrac{k}{4}, k) \geq \frac{p^{\frac{7}{8}k-1}}{k!},$$

for large enough p [72, Theorem 6]. Also, by the Gilbert-Varshamov bound [71], there exists a subfamily $\mathcal{Z}^\star_\mathcal{K} \subset \mathcal{Z}_\mathcal{K}$, of cardinality $|\mathcal{Z}^\star_\mathcal{K}| \geq 2^{\lfloor k/8 \rfloor + 1} + 1$, such that any two distinct $\boldsymbol{\theta}^i, \boldsymbol{\theta}^j \in \mathcal{Z}^\star_\mathcal{K}$ differ in at least $k/16$ elements. Thus for $\boldsymbol{\theta}^i, \boldsymbol{\theta}^j \in \mathcal{Z}^\star := \bigcup_\mathcal{K} \mathcal{Z}^\star_\mathcal{K}$, we have

$$\left\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^j\right\|_2 \geq \frac{1}{4}\sqrt{k}e^{-b} =: h, \tag{29}$$

and $|\mathcal{Z}^\star| \geq \frac{p^{\frac{7}{8}k-1}}{k!} 2^{\lfloor k/8 \rfloor + 1} + 1$. For an arbitrary estimate $\widehat{\boldsymbol{\theta}} =: [\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}]$, consider the testing problem between the $\frac{p^{\frac{7}{8}k-1}}{k!} 2^{\lfloor k/8 \rfloor + 1} + 1$ hypotheses $H_j : \boldsymbol{\theta}^*_{(2)} = \boldsymbol{\theta}^j_{(2)} \in \mathcal{Z}^\star$, using the minimum distance decoding strategy. By construction, if $k > 16$, we have:

$$\sup_{\mathcal{Z}^\star} \mathbb{P}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \geq \frac{h}{2}\right] = \sup_j \mathbb{P}\left[\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^j \middle| H_j\right], \tag{30}$$

given that $\|\boldsymbol{\theta}^i_{(1)}\|_2 = e^{-b}$ or 0. Let $f_{\boldsymbol{\theta}^j}$ denote joint probability distribution of $\{x_k\}_{k=1}^n$ conditioned on $\{x_k\}_{k=-p+1}^0$ and $\{y_k\}_{k=-p+1}^n$ under the hypothesis H_j. Fano's inequality given in Lemma C.4 provides the following lower bound on the probability in (30):

$$\sup_j \mathbb{P}\left[\widehat{\boldsymbol{\theta}} \neq \boldsymbol{\theta}^j | H_j\right] \geq 1 - \frac{\xi + \log 2}{\log_2(|\mathcal{Z}^\star| - 1)}, \tag{31}$$

where ξ is an upper bound on the Kullback-Leibler divergence between $f_{\boldsymbol{\theta}^i}$ and $f_{\boldsymbol{\theta}^j}$ for any $i \neq j$, i.e., $\mathcal{D}_{\mathsf{KL}}\left(f_{\boldsymbol{\theta}^i} \middle\| f_{\boldsymbol{\theta}^j}\right) \leq \xi, \forall i \neq j$. Given Gaussian innovations in the BVAR process, for $i \neq j$, we have

$$\begin{aligned}
\mathcal{D}_{\mathsf{KL}}\left(f_{\boldsymbol{\theta}^i} \middle\| f_{\boldsymbol{\theta}^j}\right) &\leq \sup_{i \neq j} \mathbb{E}\left[\log \frac{f_{\boldsymbol{\theta}^i}}{f_{\boldsymbol{\theta}^j}} \middle| H_i\right] \\
&= \sup_{i \neq j} \mathbb{E}\left[-\frac{1}{2(\Sigma_\epsilon)_{1,1}}\left(\left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^i\right\|^2 \right.\right. \\
&\qquad \left.\left. - \left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^j\right\|^2\right) \middle| H_i\right] \\
&\leq \sup_{i \neq j} \frac{1}{2(\Sigma_\epsilon)_{1,1}} \mathbb{E}\left[\left\|\mathbf{X}\left(\boldsymbol{\theta}^i - \boldsymbol{\theta}^j\right)\right\|^2 \middle| H_i\right] \\
&= \frac{n}{2(\Sigma_\epsilon)_{1,1}} \sup_{i \neq j} \left(\boldsymbol{\theta}^i - \boldsymbol{\theta}^j\right)^\top \mathbf{C}\left(\boldsymbol{\theta}^i - \boldsymbol{\theta}^j\right) \\
&\leq \frac{n\Lambda_{\max}(\mathbf{C})}{2(\Sigma_\epsilon)_{1,1}} \sup_{i \neq j} \left\|\boldsymbol{\theta}^i - \boldsymbol{\theta}^j\right\|_2^2 \\
&\leq \frac{2n(k+1)\Lambda_{\max}(\mathbf{C})e^{-2b}}{(\Sigma_\epsilon)_{1,1}} =: \xi.
\end{aligned} \tag{32}$$

Using Lemma C.4, Eqs. (29), (30) and (32) yield:

$$\sup_{\mathcal{Z}^\star} \mathbb{P}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \geq \frac{h}{2}\right] \geq 1 - \frac{2\left(\frac{2n(k+1)\Lambda_{\max}(\mathbf{C})e^{-2b}}{(\Sigma_\epsilon)_{1,1}} + \log 2\right)}{(k+1)\log p}. \tag{33}$$

for p large enough so that $\log p \geq \frac{8\log k - \log 2}{3 - \frac{12}{k}}$. Assuming $(k+1)\log p \geq 8\log 2$ and choosing $b \geq \frac{1}{2}\log\left(\frac{16\Lambda_{\max}(\mathbf{C})}{(\Sigma_\epsilon)_{1,1}}\right) + \frac{1}{2}\log\left(n/\log p\right)$, we get:

$$\sup_{\mathcal{Z}^\star} \mathbb{P}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \geq \frac{h}{2}\right] \geq \frac{1}{2}, \tag{34}$$

Thus, there exist a choice in $\mathcal{Z}^*$ for which the error of an arbitrary estimator is larger than $h/2$ with probability at least $1/2$. By this choice of $\boldsymbol{\theta}^*$, the condition on b gives $\|\boldsymbol{\theta}^*_{(2)}\|_2 \leq \|\boldsymbol{\theta}^*\|_2 \leq \widetilde{\mathscr{B}} k \log p / n$, where $\widetilde{\mathscr{B}} = \frac{(\Sigma_\epsilon)_{1,1}}{16\Lambda_{\max}(\mathbf{C})}$.

Next, under the alternative hypothesis, for a given threshold t, the null hypothesis is not rejected if

$$\frac{\ell\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}\right)}{\ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right)} \leq 1 + t, \tag{35}$$

Using Lemma C.1, Lemma C.3, and Proposition A.2, the numerator on the left hand side of (35) can be upper bounded by $(\boldsymbol{\Sigma}_\epsilon)_{1,1} + D + \Delta_D + \Delta_R + \Delta_N$. Thus, (35) holds if

$$\ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right) \geq \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1} + D + \Delta_D + \Delta_R + \Delta_N}{(1+t)}. \tag{36}$$

For large enough n and p, the numerator in (36) can be further upper bounded by $5(\boldsymbol{\Sigma})_{1,1}$ with high probability. Next, we need to lower bound the left hand side of (36). We have $\ell\left(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}\right) = \frac{1}{n}\|\mathbf{X}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|_2^2 \geq \frac{\alpha(m+1)}{2m}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$, given the RE condition and the assumption on τ in Proposition A.1. Under the condition that $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \geq \frac{h}{2}$, which holds with probability at least $1/2$, if

$$\frac{\alpha(m+1)}{8m} h^2 > \frac{5(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{(1+t)}, \tag{37}$$

then the inequality (36) holds, which implies that the null hypothesis is not rejected. Replacing h from (29) into the inequality (37) gives the condition $n < \widetilde{\mathscr{D}}' k \log p$, where $\widetilde{\mathscr{D}}' := \frac{\alpha(m+1)(1+t)}{10240 m \Lambda_{\max}(\mathbf{C})}$. □

It is noteworthy that while the result of Theorem 2 imposes necessary conditions for reliable GC detection via LGC, it is a weak converse result as it does not quantify necessary conditions that hold for *all* test statistics.

C. False Positive Error Analysis

Theorem 1 provides sufficient conditions that implicitly guarantee high sensitivity and specificity in detecting GC influences using the LGC statistic. To make these implications more explicit, by slightly weakening the sufficient condition on the sample size, $n \geq \mathscr{D}'' k \log(2p)$ in Theorem 1, we arrive at the following corollary to Theorem 1 that upper bounds the false positive error probability:

Corollary 1 (False Positive Error Probability)**.** *Suppose that assumptions 1 and 2 as well as conditions 1 and 2 in the proof of Theorem 1 hold. Then, for some arbitrary $t_0 > 0$, thresholding the proposed LGC statistic $\mathcal{T}_{y\mapsto x}$ at a level $\mathfrak{t} > 0$ for rejecting the null hypothesis results in a false positive error probability of at most $2\exp\left(-n\Big/8\left(1+\gamma t_0\sqrt{\log(2p)/n}\right)^2\right)$ with $\gamma := (\mathfrak{t}+2)/\mathfrak{t}$, if*

$$n \geq 2\max\{\widetilde{\mathscr{D}}^2 k^2/t_0^2, 2\widetilde{\mathscr{D}}\gamma k\}\log(2p),$$

for some constant $\widetilde{\mathscr{D}}$ that is explicitly given in proof.

Proof. First, we note that under Condition 1 and Condition 2 there exist some real numbers $s, t > 0$ such that

$$\left|\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - (\boldsymbol{\Sigma}_\epsilon)_{1,1}\right| \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1}/s, \tag{38a}$$

$$\Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1} t/s. \tag{38b}$$

This allows us to set a problem independent threshold, since the upper bound on $\mathcal{T}_{y\mapsto x}$ under the null hypothesis given in (18) simplifies to:

$$\mathcal{T}_{y\mapsto x} \leq \frac{2t/s}{1-(1+t)/s}. \tag{39}$$

Now, given any threshold $\mathfrak{t} > 0$, we can solve for s in terms of t and $\mathfrak{t}$ as:

$$s = 1 + \frac{2+\mathfrak{t}}{t} t = 1 + \gamma t. \tag{40}$$

To ensure $\Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1} t/s$, we need the following to hold:

$$\frac{(m+1)\alpha}{42m}\frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{16\mathscr{A}^2}\frac{n}{k\log(2p)} - \frac{1}{t} - \gamma \geq 0. \tag{41}$$

On the other hand, using the expression for s from (40) and invoking Lemma C.2 (Appendix C) yield the following statement that can be used to bound the false positive error probability:

$$\mathbb{P}\left[\left|\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - (\boldsymbol{\Sigma}_\epsilon)_{1,1}\right| \geq \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{s}\right] \leq 2\exp\left(-\frac{n}{8\left(1+\gamma t\right)^2}\right). \tag{42}$$

With a choice of $t = t_0\sqrt{\log(2p)/n}$ for any $t_0 > 0$, applying Lemma C.5 (Appendix C) on (41) then gives the sampling requirement of

$$n \geq \left(\widetilde{\mathscr{D}}/t_0\right)^2 k^2 \log(2p) + 2\widetilde{\mathscr{D}}\gamma k \log(2p),$$

and the false positive error probability given in the corollary, where

$$\widetilde{\mathscr{D}} = \frac{42m}{(m+1)\alpha}\frac{16\mathscr{A}^2}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}.$$

This concludes the proof of the corollary.

Note that by further choosing $t_0 \leq \sqrt{\widetilde{\mathscr{D}}k/2\gamma}$, one can simplify the sampling requirement and the corresponding upper bound on false positive error probability to $n \geq 2\left(\widetilde{\mathscr{D}}/t_0\right)^2 k^2\log(2p)$ and $2\exp\left(-n/\left(1+\sqrt{8}\right)^2\right)$, respectively. The later inequality follows from the fact $n \geq 4\widetilde{\mathscr{D}}k\gamma\log(2p)$ which is also guaranteed by choice of t_0. □

D. Power Analysis

Intuitively, detecting a GC effect arising from a small cross-regression coefficient $\boldsymbol{\theta}^*_{(2)}$ is challenging, and often requires a long observation horizon to be identified. Theorem 1 quantifies this intuition via a lower bound on the norm of the cross-regression coefficients in terms of the spectral properties of the process (via $\mathscr{B}$), sparsity k, sample size n and model order p. In particular, as $\|\boldsymbol{\theta}^*_{(2)}\|_2 \to 0$, a scaling of $n = \mathcal{O}(k\log(2p)/\|\boldsymbol{\theta}^*_{(2)}\|_2^2)$ maintains the sensitivity/specificity of $\mathcal{T}_{y\mapsto x}$ with high probability. This lower bound on $\|\boldsymbol{\theta}^*_{(2)}\|_2$ exhibits the same scaling as that in the thresholding procedure of [45] (i.e., the scaling of the LASSO estimation error), as well as the classical scaling of [17], [18] (up to logarithmic factors), and we thus believe is not significantly improvable.

In Corollary 1, we formalized the foregoing argument in terms of the test specificity. In order to similarly quantify the test sensitivity, we consider the effect size $\mathfrak{r} := \frac{1}{k}\|\boldsymbol{\theta}^*_{(2)}\|^2$ pertaining to the cross-regression coefficients. The power analysis for detecting a GC effect using LGC is given in the following corollary to Theorem 1:

Corollary 2 (Power Analysis)**.** *Suppose that assumptions 1 and 2 as well as conditions 1 and 2 in the proof of Theorem 1 hold. Then, for an effect size $\mathfrak{r} := \frac{1}{k}\|\boldsymbol{\theta}^*_{(2)}\|^2$ and an arbitrary constant $0 < \psi < 1$, thresholding the proposed LGC statistic $\mathcal{T}_{y\mapsto x}$ at a level $\mathfrak{t} > 0$ gives a test power of at least $1 - 2\exp\left(-\frac{n}{128}\right) - 2\exp(-cn\min\{\zeta^{-2}\log(2p)/n, 1\})$, if $\mathfrak{r} \geq \left(\frac{\mathfrak{r}_0}{\psi^2} + \frac{\mathfrak{r}_1+\mathfrak{r}_2 t}{\psi}\right)\frac{\log(2p)}{n}$ and*

$$n \geq \max\left\{\frac{\mathscr{C}''}{(1-\psi)^2}, \mathscr{D}' k\right\}\log(2p),$$

for the same constant $\mathscr{D}'$ in Theorem 1 and some constants $\mathfrak{r}_0$, $\mathfrak{r}_1$, $\mathfrak{r}_2$, and $\mathscr{C}''$ that are explicitly given in proof.

Proof. Starting from (19), we seek sufficient conditions to ensure

$$t < \frac{D - (\Delta_D + \Delta_R + \Delta_F)}{(\boldsymbol{\Sigma}_\epsilon)_{1,1} + \Delta_N + \Delta_F} \tag{43}$$

with high probability for a given threshold $\mathfrak{t}$. Recall that:

$$\Delta_F := \frac{24}{m+1}\frac{k\lambda_n^2}{\alpha/m},$$

$$\Delta_R := 20\frac{k\lambda_n^2}{\alpha/m} + \left(8\sqrt{2m}+18\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1,$$

$$D := {\boldsymbol{\theta}^*_{(2)}}^\top (\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\boldsymbol{\theta}^*_{(2)},$$

and

$$\Delta_D := \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}} \left\| \left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)} \right] \right\|_1 + \frac{\alpha}{27} \left\| \left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)} \right] \right\|_2^2,$$

with a choice of $\lambda_n = 4\mathscr{A}\sqrt{\log(2p)/n}$. By rearranging (43), we equivalently seek conditions to ensure:

$$D > \Delta_D + \Delta_R + (1+t)\Delta_F + t\Delta_N + t(\boldsymbol{\Sigma}_\epsilon)_{1,1}. \quad (44)$$

As in the proof of Theorem 1, we have:

$$\Delta_D + \Delta_R \leq \mathscr{a}\sqrt{\frac{\log(2p)}{n}} \left\| \boldsymbol{\theta}^*_{(2)} \right\|_2^2 + \mathscr{b}\sqrt{\frac{k\log(2p)}{n}} \left\| \boldsymbol{\theta}^*_{(2)} \right\|_2 + \mathscr{c}'\frac{k\log(2p)}{n},$$

where

$$\mathscr{a} = \frac{\alpha}{27}\left(\left\| \mathbf{C}_{11}^{-1}\mathbf{C}_{12} \right\|_2^2 + 1 \right),$$
$$\mathscr{b} = \mathscr{A}\left[\left(32\sqrt{2m} + 73 \right) \left\| \mathbf{C}_{11}^{-1}\mathbf{C}_{12} \right\| + 1 \right],$$
$$\mathscr{c}' = \frac{320 m \mathscr{A}^2}{\alpha},$$

and $D \geq \widetilde{\Lambda}_{\min}\|\boldsymbol{\theta}^*_{(2)}\|_2^2$, with $\widetilde{\Lambda}_{\min} := \Lambda_{\min}\left(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)$. Choosing $\Delta_N = (\boldsymbol{\Sigma}_\epsilon)_{1,1}/4$ and assuming that the upper bound (23) on λ_n is met with equality, we have $\Delta_F = (\boldsymbol{\Sigma}_\epsilon)_{1,1}/4$ and hence (44) holds if:

$$\widetilde{\Lambda}_{\min} \left\| \boldsymbol{\theta}^*_{(2)} \right\|_2^2 \geq \mathscr{a}\sqrt{\frac{\log(2p)}{n}} \left\| \boldsymbol{\theta}^*_{(2)} \right\|_2^2 + \mathscr{b}\sqrt{\frac{k\log(2p)}{n}} \left\| \boldsymbol{\theta}^*_{(2)} \right\|_2 + \mathscr{c}''\frac{k\log(2p)}{n}, \quad (45)$$

where $\mathscr{c}'' := \frac{16 m \mathscr{A}^2}{\alpha}\left(\frac{24(6t+1)}{m+1} + 20 \right)$. By an application of Lemma C.5, the inequality in (45) holds if

$$\left\| \boldsymbol{\theta}^*_{(2)} \right\|_2^2 \geq \left(\frac{\mathscr{b}^2}{\left(\widetilde{\Lambda}_{\min} - \mathscr{a}\sqrt{\frac{\log(2p)}{n}} \right)^2} + \frac{2\mathscr{c}''}{\left(\widetilde{\Lambda}_{\min} - \mathscr{a}\sqrt{\frac{\log(2p)}{n}} \right)} \right) \times \frac{k\log(2p)}{n}. \quad (46)$$

Given that $\|\boldsymbol{\theta}^*_{(2)}\|_2^2 = k\mathscr{r}$ and using the condition $n \geq \frac{\mathscr{C}''}{(1-\psi)^2}\log(2p)$ with $\mathscr{C}'' := \mathscr{a}^2/\widetilde{\Lambda}^2_{\min}$, the following bound on the effect size $\mathscr{r}$ ensures the inequality in (46):

$$\mathscr{r} \geq \left(\frac{\mathscr{r}_0}{\psi^2} + \frac{\mathscr{r}_1 + \mathscr{r}_2 t}{\psi} \right) \frac{\log(2p)}{n},$$

where

$$\mathscr{r}_0 := \frac{\mathscr{b}^2}{\widetilde{\Lambda}^2_{\min}},$$
$$\mathscr{r}_1 := \frac{32 m \mathscr{A}^2}{\alpha\widetilde{\Lambda}_{\min}}\left(\frac{24}{m+1} + 20 \right),$$
$$\mathscr{r}_2 := \frac{4608 m \mathscr{A}^2}{\alpha(m+1)\widetilde{\Lambda}_{\min}}.$$

Finally, assuming that Conditions 1 and 2 hold, the deviation bound for the full model with $\Delta_N = (\boldsymbol{\Sigma}_\epsilon)_{1,1}/4$ holds with probability at least $1 - 2\exp(-n/128)$ and the deviation bound under the reduced model holds with probability at least $1 - 2\exp(-cn\min\{\zeta^{-2}\log(2p)/n, 1\})$, with the constant ζ in Condition 2. Combining the two, the test power is at least

$$1 - 2\exp\left(-\frac{n}{128}\right) - 2\exp(-cn\min\{\zeta^{-2}\log(2p)/n, 1\}),$$

This concludes the proof of the corollary. □

E. Data-driven Choice of the Threshold for Hypothesis Testing

While Theorem 1 establishes the existence of a threshold t that can separate the null and alternative hypotheses, a key practical factor in the utility of LGC is to be able to set this threshold in a data-driven fashion. A careful inspection of the proofs of Theorem 1 and Corollaries 1 and 2 indeed allows us to estimate these thresholds in practice. Here, we consider two practical methods to choose the test threshold in a data-driven fashion:

Method 1: As in most practical applications of LASSO, we assume that n is sufficiently large and satisfies the sufficient lower bound of Corollary 1 for a nominal choice of the free parameter t_0. Then, by setting $t_0 = 1$, we have the following proposition for the first method to choose the test threshold and the corresponding type I and II error characterization:

Proposition 1. *Suppose that assumptions 1 and 2 as well as conditions 1 and 2 in the proof of Theorem 1 hold. Given a target false positive probability π_{F}, the threshold t can be chosen as:*

$$t = 2 \Bigg/ \left(\frac{n}{\sqrt{8\log\left(2/\pi_{\mathsf{F}}\right)\log(2p)}} - \sqrt{\frac{n}{\log(2p)}} - 1 \right). \quad (47)$$

Furthermore, if the effect size is large enough such that $\mathscr{r} \geq \left(\frac{\mathscr{r}_0}{\psi^2} + \frac{\mathscr{r}_1 + \mathscr{r}_2 t}{\psi} \right) \frac{\log(2p)}{n}$, the type II error probability β satisfies:

$$\beta \leq 2\exp\left(-\frac{n}{128}\right) + 2\exp(-cn\min\{\zeta^{-2}\log(2p)/n, 1\}). \quad (48)$$

Proof. This result is a specialization of Corollaries 1 and 2 to the specific choice of the threshold given in (47). □

We used this method of setting the threshold in our simulation study in Section IV-A (Fig. 2, dashed traces) as well as in our real data validation in Section IV-C.

Method 2: If the assumption of Method 1, that n is large enough to satisfy the sufficient lower bound in Corollary 1, is not valid, we need to choose the free parameter t_0 to be compatible with the lower bound on n in Theorem 1. To this end, we need to estimate three key parameters in a data-driven fashion: the sparsity level k, the noise variance $(\boldsymbol{\Sigma}_\epsilon)_{1,1}$, and the curvature parameter α in the RE condition. We hereafter assume that τ in the RE condition is small enough so that $m = 2$. We also assume that λ_n is empirically chosen via cross-validation. Also, let $\mathscr{M}(\mathbf{X})$ denote the mutual coherence of the matrix $\mathbf{X}$. We have the following proposition that

provides the second method to choose the threshold t and the corresponding type I and II error analysis:

Proposition 2. *Suppose that assumptions 1 and 2 hold and the sufficient conditions of Theorem 1 are met, but with $\mathscr{B}$ replaced by another constant $\mathscr{B}'$ (given in the proof). Given a target false positive probability π_{F}, let π_0 be such that $\pi_{\mathsf{F}} \leq \pi_0 + K_1' \exp(-n\bar{c}') + \frac{K_2'}{p^{\bar{d}'}}$, for some known constants $K_1', \bar{c}', K_2'$, and $\bar{d}'$. Then, the threshold t can be chosen as:*

$$t = 2\Big/\left(\frac{1}{t_0}\left(\frac{n}{\sqrt{8\log(2/\pi_0)\log(2p)}} - \sqrt{\frac{n}{\log(2p)}}\right) - 1\right), \tag{49}$$

where

$$t_0 = \frac{28\widehat{k}\lambda_n^2}{\kappa\widehat{\alpha}(\widehat{\boldsymbol{\Sigma}}_\epsilon)_{1,1}}\sqrt{\frac{\log(2p)}{8\log(2/\pi_0)}}, \tag{50}$$

with

- $\widehat{k} := \left\|\widehat{\boldsymbol{\theta}}^\delta\right\|_0$, *where $\widehat{\boldsymbol{\theta}}^\delta$ is the estimate $\widehat{\boldsymbol{\theta}}$ thresholded at a level $\delta = \mathcal{O}(\lambda_n/\alpha)$,*
- $(\widehat{\boldsymbol{\Sigma}}_\epsilon)_{1,1} = \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2$,
- $\widehat{\alpha} := \frac{4}{3}\min_{0\leq i\leq 2p}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{i,i}\left(1 - \mathscr{M}(\mathbf{X})\frac{\widehat{k}}{2\kappa}\right)$, *where $\mathscr{M}(\mathbf{X})$ is the mutual coherence of $\mathbf{X}$,*

and κ is a constant given in the proof. Furthermore, if the effect size is large enough such that it satisfies the condition in Corollary 2, the type II error probability β satisfies:

$$\beta \leq K_1'' \exp(-n\bar{c}'') + \frac{K_2''}{p^{\bar{d}''}}, \tag{51}$$

where $K_1'', \bar{c}'', K_2''$, and $\bar{d}''$ are constants explicitly given in the proof.

Proof. Given a choice of t_0, choosing the threshold t as:

$$t = 2\Big/\left(\frac{1}{t_0}\left(\frac{n}{\sqrt{8\log(2/\pi_0)\log(2p)}} - \sqrt{\frac{n}{\log(2p)}}\right) - 1\right), \tag{52}$$

ensures that the false positive error expression in Corollary 1 is bounded by π_0. On the other hand, for (41) to hold, we need to have

$$\gamma\widetilde{\mathscr{D}}k\frac{\log(2p)}{n} \leq 1 - \frac{\widetilde{\mathscr{D}}k}{t_0}\sqrt{\frac{\log(2p)}{n}}. \tag{53}$$

We choose t_0 to satisfy the foregoing bound with equality, which along with the threshold in (52) leads to:

$$t_0 = k\widetilde{\mathscr{D}}\sqrt{\frac{\log(2p)}{8\log(2/\pi_0)}} = k\frac{42m}{(m+1)\alpha}\frac{\lambda_n^2}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}\sqrt{\frac{\log(2p)}{8\log(2/\pi_0)}}. \tag{54}$$

For this choice of t_0, one needs to empirically estimate the sparsity level k, the noise variance $(\boldsymbol{\Sigma}_\epsilon)_{1,1}$, and the curvature parameter α in the RE condition. For k, we use the estimator $\widehat{k} = \left\|\widehat{\boldsymbol{\theta}}^\delta\right\|_0$, where $\widehat{\boldsymbol{\theta}}^\delta$ is the estimate $\widehat{\boldsymbol{\theta}}$ thresholded at a level $\delta > 0$. It can be shown that if δ is chosen as $\mathcal{O}(\lambda_n/\alpha)$, then $\widehat{k} \geq (1-\mathscr{C})k$ for some constant $\mathscr{C}$, given Conditions 1 and 2 [73, Theorem 7.3].

Next, we estimate $(\boldsymbol{\Sigma}_\epsilon)_{1,1}$ as

$$(\widehat{\boldsymbol{\Sigma}}_\epsilon)_{1,1} = \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}\widehat{\boldsymbol{\theta}}\right\|^2, \tag{55}$$

which with probability at least $1 - 2\exp(-n/32) - c_1\exp(-c_2 n\min\{\zeta^{-2}, 1\}) - \frac{d_1}{(2p)^{d_2}}$, satisfies

$$|(\widehat{\boldsymbol{\Sigma}}_\epsilon)_{1,1} - (\boldsymbol{\Sigma}_\epsilon)_{1,1}| \leq \frac{(\boldsymbol{\Sigma}_\epsilon)_{1,1}}{2} + \Delta_F, \tag{56}$$

where $\Delta_F = \frac{24}{m+1}\frac{k\lambda_n^2}{\alpha/m}$.

Estimating the curvature parameter α, however, is not trivial. We thus replace it with a lower bound using the mutual coherence of the design matrix, $\mathscr{M}(\mathbf{X})$, following a result from [74] that holds with high probability:

$$\alpha \geq \frac{2m}{m+1}\min_{0\leq i\leq 2p}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)_{i,i}\Big(1 - \mathscr{M}(\mathbf{X})k\Big). \tag{57}$$

Recalling the sufficient condition $\|\boldsymbol{\theta}^*_{(2)}\|_2^2 \geq \mathscr{B}k\log(2p)/n$ in the statement of Theorem 1, where $\mathscr{B}$ linearly depends on $\frac{m}{(m+1)\alpha}$, we weaken the lower bound on $\|\boldsymbol{\theta}^*_{(2)}\|_2^2$ as $\mathscr{B}'k\log(2p)/n$, where $\mathscr{B}'$ is the same as $\mathscr{B}$ with α replaced by the foregoing lower bound. The estimate $\widehat{\alpha}$ is obtained by using the lower bound with an estimate of $\frac{\widehat{k}}{1-\mathscr{C}}$ for k. Note that this lower bound on α may be too conservative in practice and may result in reducing the test power.

We next set $m = 2$. For large enough n and p, we also have $\Delta_F \leq (\boldsymbol{\Sigma}_\epsilon)_{1,1}/2$, so that the estimate of t_0, namely $\widehat{t_0}$, using the empirical estimates of k and $(\boldsymbol{\Sigma}_\epsilon)_{1,1}$ satisfies

$$\frac{\widehat{t_0}}{t_0} \leq \frac{1-\mathscr{C}}{2} := \kappa. \tag{58}$$

Thus, the false positive error can be bounded as

$$\begin{aligned}\pi_{\mathsf{F}} \leq{}& \pi_0 + 2\exp(-n/32) + c_1\exp(-c_2 n\min\{\zeta^{-2}, 1\}) \\ &+ c_3\exp\left(-c_4 n\min\{\zeta^{-2}\max\{D_0, D_0^{'}\}, 1\}\right) + \frac{d_1}{(2p)^{d_2}}.\end{aligned} \tag{59}$$

Letting

$$\bar{c}' = \min\left\{\frac{1}{32}, c_2, c_4, c_2\zeta^{-2}, c_4\zeta^{-2}\max\{D_0, D_0^{'}\}\right\},$$
$$\bar{d}' = d_2, \quad K_1' = 2 + c_1 + c_3, \quad K_2' = \frac{d_1}{2^{d_2}},$$

provides the assumed bound on π_F.

To bound the type II error, we use the result of Corollary 2, but need to ensure that the deviation condition for the reduced model also holds. Therefore, we have:

$$\begin{aligned}\beta \leq{}& 2\exp(-n/128) + 2\exp(-n/32) \\ &+ 2\exp(-cn\min\{\zeta^{-2}\log(2p)/n, 1\}) \\ &+ c_3\exp\left(-c_4 n\min\{\zeta^{-2}\max\{D_0, D_0^{'}\}, 1\}\right) \\ &+ \frac{d_1}{(2p)^{d_2}} + \frac{d_1'}{(2p)^{d_2'}}.\end{aligned} \tag{60}$$

Letting

$$\bar{c}'' = \min\left\{\frac{1}{128}, c, c_4, c\zeta^{-2}, c_4\zeta^{-2}\max\{D_0, D_0^{'}\}\right\},$$

$$\bar{d}'' = \min\{d_2, d_2'\}, \quad K_1'' = 6 + c_3, \quad K_2'' = \frac{d_1}{2^{d_2}} + \frac{d_1'}{2^{d_2'}},$$

establishes the bound on β. This concludes the proof of the proposition. □

While the choice of $t_0 = 1$ in Method 1 is quite convenient and results in a threshold independent of the details of the BVAR process, it may be too conservative in practice and may result in low test power if n is not sufficiently large. Method 2, however, selects t_0 based on empirical estimates of process-dependent parameters, and is thus expected to provide a higher test power. We will further evaluate the type I and II error for these two methods for setting the test threshold using simulation studies in Section IV-B and compare their performance.

F. Discussion of the Results

To discuss the implications of these results, several remarks are in order:

Remark 1. Unlike the conventional estimation error results of the LASSO that specify a lower bound on the regularization parameter λ_n, Theorem 1 prescribes a fixed choice of λ_n for both the *full* and *reduced* estimation problems in (6). This is due to an interesting phenomenon revealed by our analysis: while conventional analyses of LASSO focus on the estimation performance of a single model and thus provide a lower bound on λ_n, in our framework we have two competing models (i.e., *full* and *reduced*) which need to be distinguishable under the null and alternative hypotheses in order to reliably detect the GC influences. The latter imposes an *upper* bound on λ_n. As such, there is a suitable interval for choosing λ_n that results in both consistent estimation and discrimination of the two models. We note that we have presented our theoretical results under the assumption that a single λ_n within the aforementioned interval is used for both models, for the simplicity of analysis. The strategy of using the 'same λ_n for both full and reduced models' is also appealing from a practical perspective, since the empirical methods for choosing λ_n generally involve solving the problems for multiple values of λ_n to evaluate a selection criterion and to pick the *optimal* λ_n. This process is typically computationally intensive and performing it only for the full model saves processing time. For example, the user may select the *optimal* λ_n via k-fold cross-validation for the full model which leads to smallest out of sample error across different folds (i.e., best predictive performance), and then use the resulting value of λ_n for both the full and reduced models to estimate the VAR parameters and prediction error variance using the whole data, thus avoiding extra computational costs of cross-validation.

Remark 2. Corollary 1 bounds the false positive error probability, i.e., type I error rate, for a simple thresholding scheme for detecting GC influences from $\mathcal{T}_{y\mapsto x}$, under a slightly weakened sufficient condition on n, i.e., $n = \mathcal{O}(k^2 \log(2p))$ instead of $n = \mathcal{O}(k \log(2p))$. This non-asymptotic result provides a principled guideline for choosing a threshold that controls the false positive error rate, as shown in Section III-E. As such, this result extends the conventional statistical testing framework based on the asymptotics of log-likelihood ratio statistic using OLS to the non-asymptotic setting using the LASSO. This result can further be utilized in the assignment of p-values (i.e., the false positive error probability when the observed LGC statistic is used as the threshold), as is commonly done in the conventional OLS-based setting based on the asymptotic χ^2 distribution under the null hypothesis.

Remark 3. We have presented our results for a BVAR model in order to parallel the classical GC analysis. Our results can be extended to the general MVAR setting by using the *conditional* notion of Geweke [9] in a natural fashion, given that Condition 1 and Condition 2 readily generalize to this setting. To elaborate on this point, consider a case where the BVAR(p) model in (1) is augmented by $(d-2)$ other time series to result in a general d-dimensional MVAR(p) setting. Following the formulation of Section II-C, the conditional LGC statistic can be defined as:

$$\mathcal{T}_{y\mapsto x} := \frac{\ell\Big(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}, \mathbf{0}, \widehat{\widetilde{\boldsymbol{\theta}}}_{(3)}, \cdots, \widehat{\widetilde{\boldsymbol{\theta}}}_{(d)}\Big)}{\ell\Big(\widehat{\boldsymbol{\theta}}_{(1)}, \widehat{\boldsymbol{\theta}}_{(2)}, \widehat{\boldsymbol{\theta}}_{(3)}, \cdots, \widehat{\boldsymbol{\theta}}_{(d)}\Big)} - 1, \tag{61}$$

where $\widehat{\widetilde{\boldsymbol{\theta}}}_{(i)}$ and $\widehat{\boldsymbol{\theta}}_{(i)}$ denote the MVAR parameters from the i^{th} process to the first time series, under the *reduced* and *full* models, respectively. Then, using a similar procedure as in the proof of Theorem 1 and Corollary 1, the results can be extended to this setting, by replacing the various occurrences of $2p$ by dp and by adopting slightly different constants.

Remark 4. The constants in the proofs of Theorems 1 and 2 and Corollaries 1 and 2 solely depend on the joint spectrum of processes x_t, y_t as well as some absolute constants. As an illustrative example, by assuming $\boldsymbol{\Sigma}_\epsilon = 0.01\mathbf{I}$, $\mu_{\max}(\mathbf{A}) = 0.9$, $\mu_{\min}(\mathbf{A}) = \mu_{\min}(\breve{\mathbf{A}}) = 0.01$, $\breve{\Lambda}_{\min} = 0.7$, $\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\|_2 = 0.5$, $\|[\mathbf{C}_{11}^{-1}\mathbf{C}_{12}; \mathbf{I}]\boldsymbol{\theta}^*_{(2)}\|_2 = 1.5$, $m = 8$, $d_0 = 5.2 \times 10^{-4}$, $d_0' = 4.2 \times 10^{-4}$, $D_0 = 100$, $C_0 = 10^{-6}$, and $c = 0.02$, the key constants in Theorem 1 take the following numerical values: $\mathscr{A} = 10^{-3}$, $\mathscr{B} = 5.136$, $\mathscr{C}' = 7.41 \times 10^{-4}$, $\mathscr{D} = 24.38$, $K_1 = 6$, $K_2 = 6$, $\bar{c} = 2.06 \times 10^{-4}$ and $d = 1$. These translate to $\lambda_n = 10^{-3}\sqrt{\log(2p)/n}$, a requirement of $n > \max\{100, 24.38k\}\log(2p)$, local alternative hypotheses satisfying $\|\boldsymbol{\theta}^*_{(2)}\|_2^2 > 5.136k\log(2p)/n$, and failure probability $< 6\exp\big(-2.06 \times 10^{-6}n\big) + 6/p$. Similarly, for Corollary 1, we get $\widetilde{\mathscr{D}} = 10.67$, translating to a sample size requirement of $n > 2\max\big\{28.46k^2, 396k\big\}\log(2p)$ (with $t_0 = 2$ and $t = 0.114$). The potentially large numerical values of some of these constants suggest that the non-asymptotic advantage may come with large values of n and p.

IV. Application to Simulated and Experimentally-Recorded Data

In this section, we examine our theoretical results through application to simulated and real data, and by comparing the performance of classical OLS-based GC and the proposed LGC statistic in detecting GC influences. We use the fast implementation in [75] to solve the LASSO problems. Unless otherwise stated, the regularization parameter λ_n is chosen via five-fold cross-validation performed over the *full* model, with the same λ_n used for the *reduced* model.

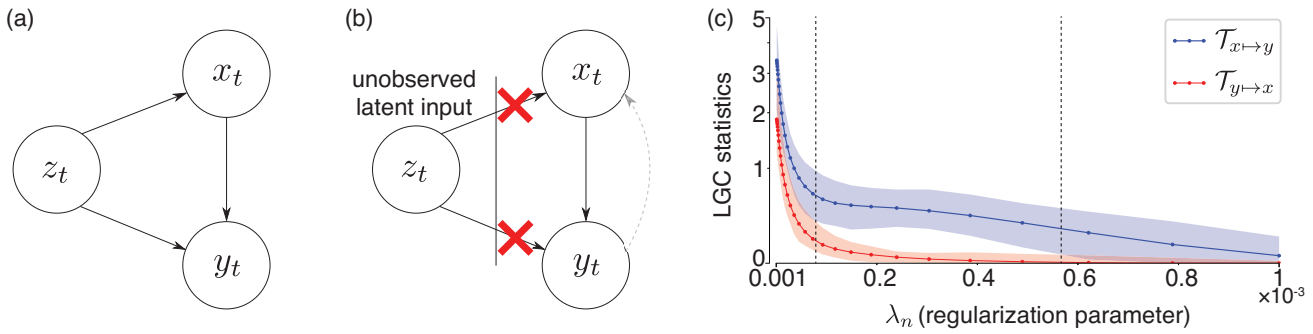

Fig. 1. Simulation Results. (a) Ground truth GC pattern. (b) Estimation setup, in which z_t is latent and thus introduces a spurious GC link (dashed gray arrow) from y_t to x_t. (c) Effect of λ_n on the LGC statistics for $n = 250, p = 100$. The LGC statistics $\mathcal{T}_{y\mapsto x}$ (red) and $\mathcal{T}_{x\mapsto y}$ (blue) are separable for a suitable range of λ_n, marked by the dashed vertical lines (colored hulls show the range of LGC over 30 realizations).

A. Simulation Studies

We simulated three time series x_t, y_t, z_t according to the sparse MVAR(11) model:

$$\begin{aligned} x_t =& -0.67x_{t-1} + 0.2x_{t-5} - 0.1x_{t-11} + 0.05z_{t-3} + \nu_{1,t}, \\ y_t =& -0.62y_{t-1} + 0.1y_{t-5} - 0.2y_{t-11} - 0.1x_{t-2} - 0.1x_{t-3}, \\ &+ 0.5x_{t-11} - 0.001z_{t-4} - 0.004z_{t-5} + \sqrt{0.6}\nu_{2,t} \\ z_t =& -0.9025z_{t-2} + \nu_{3,t}. \end{aligned}$$

where $\nu_{i,t} \sim \mathcal{N}(0,1)$, i.i.d. for $i = 1, 2, 3$. In this model, x_t has a direct GC influence on y_t, but there is no GC influence from y_t to x_t. The *latent* process z_t, however, influences both x_t and y_t (Fig. 1(a)). As such, the correlated process noise components ϵ_t and ϵ'_t in (1) are modeled as $0.05z_{t-3}+\nu_{1,t}$ and $-0.001z_{t-4} - 0.004z_{t-5} + \sqrt{0.6}\nu_{2,t}$, respectively. As shown in Fig. 1(b), removing z_t from the analysis indeed induces a false (i.e., indirect) GC influence from y_t to x_t. We performed two sets of numerical experiments to evaluate the effects of λ_n, n, and p on the identification of GC influences between x_t and y_t based on $\mathcal{T}_{y\mapsto x}$ and $\mathcal{T}_{x\mapsto y}$:

1) Evaluating the Effect of λ_n: Fig. 1(c) shows the LGC statistics $\mathcal{T}_{y\mapsto x}$ (red) and $\mathcal{T}_{x\mapsto y}$ (blue) for $n = 250$ and $p = 100$, obtained by varying λ_n in the interval $[10^{-6}, 10^{-3}]$ uniformly in the log-scale. The dotted lines and colored hulls represent the average and range of the values, respectively, over 30 realizations. As discussed in *Remark 1*, there is an evident range of λ_n that provides a meaningful separation between $\mathcal{T}_{y\mapsto x}$ and $\mathcal{T}_{x\mapsto y}$, which is marked by the dashed vertical lines in Fig. 1(c).

2) Evaluating the Effect of the Sample Size n: We fixed a model order of $p = 100$ and varied n uniformly in the interval $[100, 1000]$. Fig. 2(a) and (b) show the resulting LGC statistics $\mathcal{T}_{y\mapsto x}$ and $\mathcal{T}_{x\mapsto y}$ corresponding to the LASSO and OLS ($\lambda_n = 0$), respectively. In Fig. 2(a), we also plotted the threshold t corresponding to a false positive probability of 0.01, according to Method 1 in Section III-E (dashed line). As n grows larger, the ranges of the two LGC statistics are saliently separated. The proposed thresholding rule of Corollary 1 is also able to correctly identify the true GC effects for $n \geq 250$.

The OLS results shown in Fig. 2(b), however, require much larger values of n to be stable, whereas the LGC statistic provided by the LASSO (Fig. 2(a)) are stable even for $n < 2p$. In addition, OLS requires $n \geq 400$ for the ranges of the GC measures to be distinguishable.

3) Evaluating the Effect of the Model Order p: Finally, we fixed $n = 300$ and varied p in the interval $[10, 300]$ uniformly in the log-scale. Fig. 2(c) and (d) show the corresponding GC measures for the LASSO and OLS, respectively, along with the threshold t corresponding to a false positive probability of 0.01. For $p \ll n$, the LASSO and OLS exhibit similar performance. But, the OLS-based GC measures become unstable for $p \approx n$, whereas those of the LASSO remain stable throughout. The LGC statistics also remain saliently separable over a wider range of p for the LASSO, as compared to their OLS counterparts.

4) Comparison with Existing LASSO-based Methods: We used the same setting as in Section IV-A to compare the performance of LGC with two existing LASSO-based methods, namely the confidence interval (CI)-based LASSO and truncating LASSO (TLASSO) GC detection. For the CI-based LASSO, we used de-biasing via node-wise regression [50] and the confidence interval construction technique of [53]: if the confidence interval of at least one of the cross-regression coefficients does not include zero (after Bonferroni correction), we identify it as a GC link. We used the Bonferroni correction for its simplicity and common usage in the applications of GC [76], [77]. It is however possible to use other multiple comparison correction schemes such as the Benjamini-Hochberg procedure [78]. The TLASSO, as a graphical LASSO-based method, uses the truncating LASSO penalty to automatically determine the order of the VAR models, and thus performs model simplification by reducing the number of covariates to control false discoveries [43].

Fig. 3 shows the receiver operating characteristic (ROC) plot for varying sample sizes, $n \in \{200, 350, 500, 750, 1000, 1250, 1500\}$, computed from 200 realization of the same process in Section IV-A. The marker sizes are proportional to n, for visual convenience. We fixed the BVAR order p at 50 for LGC and CI-based LASSO, while TLASSO automatically determined the BVAR order ($p \leq 50$). For LGC, we considered a threshold for false positive error rate of 0.1 based on Method 1 in Section III-E, and accordingly set a confidence level of 90% for the CI-based LASSO and a false negative rate of 0.1 for TLASSO. The TLASSO is the most conservative of the three methods in terms of controlling the false positive error,

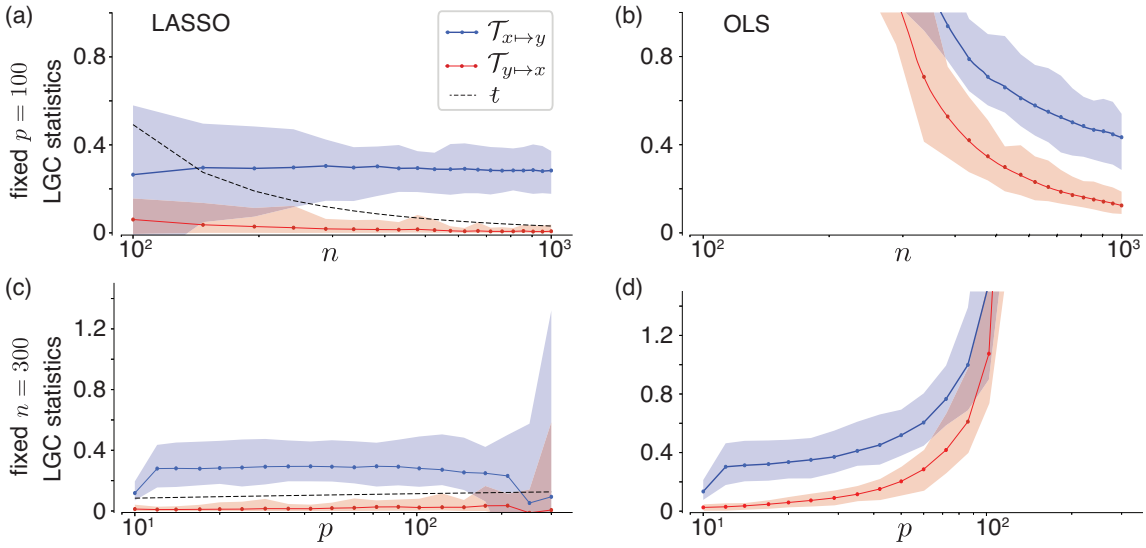

Fig. 2. Simulation Results (continued). LASSO-based (left) and OLS-based ($\lambda_n = 0$, right) LGC statistics $\mathcal{T}_{y\mapsto x}$ (red) and $\mathcal{T}_{x\mapsto y}$ (blue) obtained by varying n for fixed $p = 100$ (top panels (a) and (b)) and varying the model order p for fixed $n = 300$ (bottom panels (c) and (d)). The dashed lines in panels (a) and (c) show the threshold t at a false positive error level of 0.01 (colored hulls show the range LGC over 30 realizations).

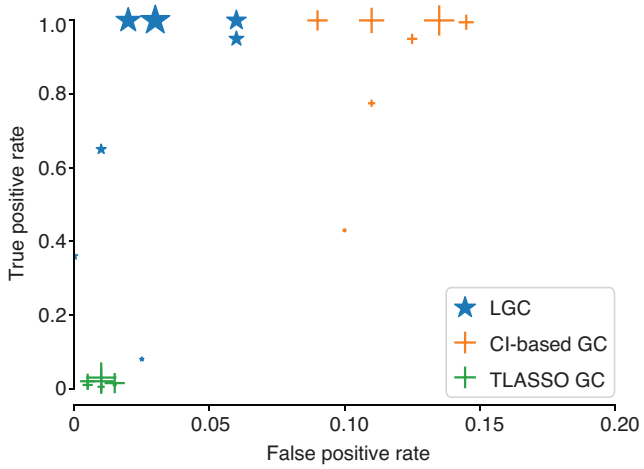

Fig. 3. Comparing the performance of LGC, CI-based LASSO, and TLASSO for GC identification. True positive and false positive error rates are computed from 200 realization of the BVAR process in Section IV-A with $p = 50$ by varying $n \in \{200, 350, 500, 750, 1000, 1250, 1500\}$. The marker sizes are proportional to n.

which comes at the cost of low test power. On the other hand, the CI-based LASSO exhibits a high test power, but at cost of increasing false positive errors. LGC, however, adheres to the middle ground between these two extremes by striking a reasonable balance between true and false positive error rates. Note that this simulation study was intentionally designed to evaluate the performance of LGC in presence of weak GC influences. As expected, when the GC influence gets stronger, all three methods exhibit similar performance.

B. Comparing the Data-Driven Methods for Setting the Test Threshold

In the foregoing simulation studies, we used Method 1 in Section III-E to set the test threshold due to its convenient form. In order to compare the two data-driven methods for setting the test threshold given in Section III-E, we consider a slightly modified version of the simulation setting in Section IV-A:

$$\begin{aligned} x_t =& -0.1x_{t-1} + 0.2x_{t-5} - 0.1x_{t-11} + 0.05z_{t-3} + \nu_{1,t}, \\ y_t =& -0.1y_{t-1} - a \times 0.1x_{t-2} + a \times 0.5x_{t-11} - 0.001z_{t-4} \\ & - 0.004z_{t-5} + \sqrt{0.6}\nu_{2,t}, \\ z_t =& -0.9025z_{t-2} + \nu_{3,t}. \end{aligned}$$

In this modified setting, the sparsity level is reduced to $k = 3$, and the leading VAR coefficients are decreased, so that the estimate of the curvature parameter α using the mutual coherence is reliable for small values of n and p. The multiplier $0 \leq a \leq 1$ is chosen to control the effect size of the GC link. Similar to the previous case, we consider $p = 50$, and use a range of $n \in \{100, 150, 200, 250, 300, 350\}$ to closely examine the role of the effect size in the performance of the two methods. To speed up simulations, we tuned λ_n for $n = 100$ via 2-fold cross-validation, and used the scaling $\sqrt{100/n}$ to obtain the subsequent values of λ_n.

Fig. 4(a) shows the ROC plot for the two methods for setting the test thresholds given in Section III-E. The marker sizes are proportional to n. As it can be observed from the figure, Method 1 (blue stars) is expectedly more conservative and while it maintains negligible false positive rates, it requires larger values of n for reliable detection of the GC link. Method 2 (purple stars), however, is more sensitive, but for $n \geq 150$ achieves both high test power (more than 0.9) and a false positive error less than the target level of 0.1.

Fig. 4(b) shows the performance of both methods with respect to the effect size. The line widths are proportional to n. Consistent with Fig. 4(a), Method 1 is more conservative and while consistently achieves a negligible false alarm rate, it requires a larger value of n to detect GC links with small

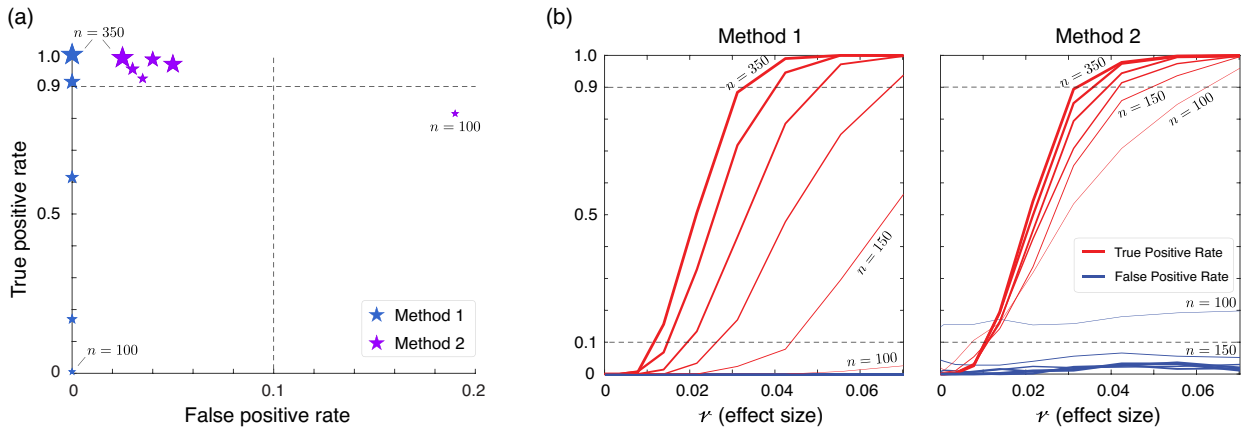

Fig. 4. Comparing the performance of the two methods of Section III-E for selecting the threshold in a data-driven fashion. (a) True positive and false positive rates are computed from 200 realization of the modified BVAR process with $a = 0.75$, $p = 50$, and varying $n \in \{100, 150, 200, 250, 300, 350\}$, for a target false positive probability of 0.1. The results for Methods 1 and 2 given in Section III-E are shown by blue and purple stars, respectively. The marker sizes are proportional to n. (b) True positive (red) and false positive (blue) rates as a function of the effect size r, by varying $a \in [0, 1]$. Line widths are proportional to n.

effect size, as indicated by the horizontal dashed line at 0.9. Method 2, however, achieves a false alarm rate below the target level (horizontal dashed line at 0.1) for $n \geq 150$, while reliably capturing GC links of small effect size. As expected, the performance of both methods become comparable as n gets larger.

C. Application to Experimentally-Recorded Neural Data from General Anesthesia

Finally, we present an application to simultaneous local field potentials (LFPs) and an ensemble of single-unit recordings from the temporal cortex in a human subject under Propofol-induced general anesthesia (Data from [79]). The LFP signal is the electrical field potential measured at the cortical surface and represents mesoscale dynamics of brain activity with both cortical and sub-cortical (e.g., thalamic) origins. Single-unit spike recordings, on the other hand, represent the neuronal scale cortical dynamics.

Brain states under anesthesia and sleep are associated with the emergence of periodic and profound suppression of neuronal spiking activity that is strongly phase-locked to the peaks of the LFP slow oscillations [79]–[82]. Specifically, by comparing the average LFP signals triggered at the trough of the slow oscillations under no-spike and many-spike conditions, [80] argues that neuronal spiking activity may have a GC influence in high-amplitude peaks of the slow oscillations manifested in the LFP. Here, we examine the role of neuronal spiking activity in mediating the LFP slow oscillations by assessing the GC influences between them.

We use a time duration of 51.2 s during anesthesia, corresponding to $n = 1280$ samples (sampling frequency of 25 Hz). The ensemble spiking activity is represented by its peristimulus time histogram (PSTH) (i.e., ensemble average over 23 units). Fig. 5(a) shows the LFP (green) and PSTH (orange) signals used in the analysis. We use a model order of $p = 100$, corresponding to a history length of 4 s, to ensure that slow oscillations ($\sim$0.25 Hz to 0.5 Hz) can be captured by the BVAR model. Fig. 5(b) shows the estimated BVAR coefficients by the LASSO (top) and OLS (bottom). A visual comparison of the two sets of coefficients suggests that OLS has likely over-fitted the data. The corresponding LGC statistics $\mathcal{T}_{\mathsf{LFP}\mapsto\mathsf{PSTH}}$ and $\mathcal{T}_{\mathsf{PSTH}\mapsto\mathsf{LFP}}$ for both methods are reported in Table I. The numbers in parentheses show the p-values, i.e., the false positive error probability, when the observed GC statistics are used as thresholds. For the LGC statistics, the p-values are computed via Corollary 1 (with a choice of $t_0 = 0.25$), while in the conventional OLS setting we used the χ^2 distribution against the threshold $\mathcal{F}_{x\mapsto y} = \log(1 + \mathcal{T}_{x\mapsto y})$.

For a false positive error probability of 0.01, the LGC-based test clearly detects the GC effect $\mathsf{PSTH} \mapsto \mathsf{LFP}$ (boldface number) as significant, and discards the GC effect $\mathsf{LFP} \mapsto \mathsf{PSTH}$. The conventional χ^2 test applied to the classical GC statistics $\mathcal{F}_{\mathsf{LFP}\mapsto\mathsf{PSTH}}$ and $\mathcal{F}_{\mathsf{PSTH}\mapsto\mathsf{LFP}}$, however, fails to detect any GC influence, even at a significance level as high as 0.05. The outcome of the LASSO-based LGC analysis is therefore consistent with the aforementioned hypothesis in [80] on the GC influence of spiking activity in mediating the LFP dynamics.

TABLE I
OBTAINED LGC STATISTICS AND p-VALUES

	LASSO	OLS
$\mathcal{T}_{\mathsf{LFP}\mapsto\mathsf{PSTH}}$	0.0055 (0.1079)	0.1001 (0.1832)
$\mathcal{T}_{\mathsf{PSTH}\mapsto\mathsf{LFP}}$	**0.0096** (**0.0015**)	0.1046 (0.1135)

V. CONCLUDING REMARKS

In this work, we proposed a GC statistic based on the LASSO parameter estimates, namely the LGC statistic, in order to identify GC influences in a canonical sparse BVAR model with correlated process noise. By analyzing the non-asymptotic properties of LGC statistic, we established that

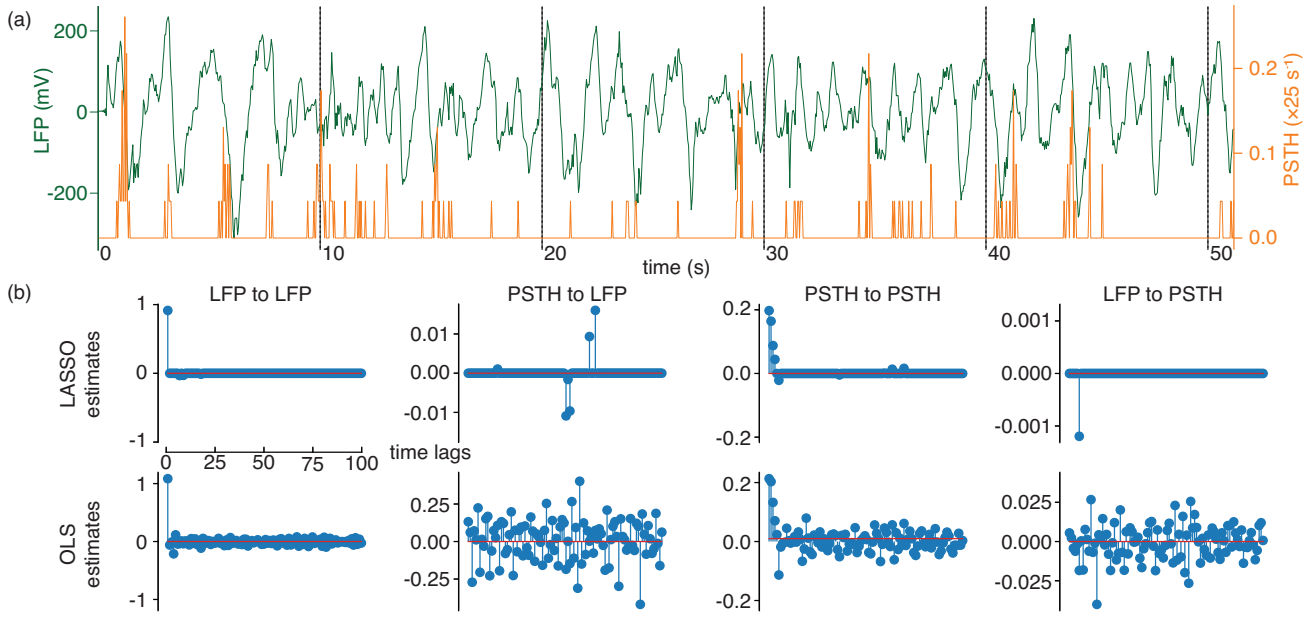

Fig. 5. Analysis of neural data from general anesthesia. (a) LFP (green) and PSTH (orange) traces for a time window of duration 51.2 s. (b) BVAR parameter estimates corresponding to the *full* models for LASSO (top) and OLS (bottom).

the well-known sufficient conditions for the consistency of LASSO also suffice for accurate identification of GC influences, if the strength of the GC effect is large enough. We also established the necessity of the constraint on the strength of the GC effect for reliable GC identification via LGC. We also analyzed the false positive error performance and test power of a simple thresholding rule for detecting GC influences and provided data-driven methods to set the test threshold in practice. We validated our theoretical claims through application to simulated and experimentally-recorded neural data from general anesthesia. In particular, we showed that the proposed LGC statistic is able to identify a GC effect from spiking activity to LFP slow oscillations under anesthesia, whereas the conventional OLS-based GC analysis does not detect this effect. Our contribution compared to existing literature is to provide a simple statistic inspired by the classical log-likelihood ratio statistic used for GC analysis, which can be directly computed from the LASSO estimates without the need to resort to de-biasing procedures or asymptotic results for testing. Future work includes extending our results to autoregressive generalized linear models with time-varying parameters and obtaining necessary conditions that hold for any test statistic.

Appendix A
Prediction Error Analysis of the Full and Reduced Models

In this section, we establish the deviation bounds of (12) and (13) under Conditions 1 and 2. Note that both the *full* and *reduced* models share the same RE condition (Condition 1), since the *reduced* model is nested within the *full* model. However, the deviation conditions required by Condition 2 are different for the two models. For the *full* model, we require:

$$\left\|\frac{1}{n}\mathbf{X}^\top(\mathbf{x}-\mathbf{X}\boldsymbol{\theta}^*)\right\|_\infty \leq \mathbb{Q}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}}, \tag{62}$$

for some deterministic function $\mathbb{Q}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)$. In the *reduced* model, however, we require

$$\left\|\frac{1}{n}\mathbf{X}_{(1)}^\top\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)\right\|_\infty \leq \mathbb{Q}'(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}} \tag{63}$$

for another deterministic function $\mathbb{Q}'(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)$. These conditions guarantee consistent and stable recovery of the autoregressive parameters in the LASSO problems of (6), and hence lead to suitable deviation results for both the *full* and *reduced* models.

Proposition A.1 (Deviation Result for the *Full* Model)**.** Suppose $\widehat{\boldsymbol{\Sigma}} \sim \mathrm{RE}(\alpha,\tau)$, with τ satisfying $32k\tau/\alpha = (m-1)/m$ for some $m > 1$ and $(\mathbf{X},\mathbf{x})$ satisfying the deviation bound (62). Then for any $\lambda_n \geq 4\mathbb{Q}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)\sqrt{\log(2p)/n}$, the solution to the *full* model in (6) satisfies:

$$\left|\ell\left(\widehat{\boldsymbol{\theta}}_{(1)},\widehat{\boldsymbol{\theta}}_{(2)}\right)-\ell\left(\boldsymbol{\theta}^*_{(1)},\boldsymbol{\theta}^*_{(2)}\right)\right| \leq \frac{24}{m+1}\frac{k\lambda_n^2}{\alpha/m} =: \Delta_F. \tag{64}$$

Proof. For the *full* model, we have:

$$\ell\left(\widehat{\boldsymbol{\theta}}_{(1)},\widehat{\boldsymbol{\theta}}_{(2)}\right)-\ell\left(\boldsymbol{\theta}^*_{(1)},\boldsymbol{\theta}^*_{(2)}\right) = \frac{1}{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\mathbf{X}^\top\mathbf{X}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) - \frac{2}{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\mathbf{X}^\top(\mathbf{x}-\mathbf{X}\boldsymbol{\theta}^*). \tag{65}$$

To obtain bounds on the left hand side of (65), we bound the terms on the right hand side individually. The bound on the first term follows from (70) on the consistency of the

LASSO under the RE and deviation bound assumptions (See Proposition A.3),

$$\frac{1}{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) \leq \frac{18m}{m+1}\frac{k\lambda_n^2}{\alpha},$$

while the second terms can be bounded using (69) as

$$\begin{aligned}\left|\frac{1}{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top \mathbf{X}^\top(\mathbf{x}-\mathbf{X}\boldsymbol{\theta}^*)\right| &\leq \left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1 \left\|\frac{1}{n}\mathbf{X}^\top(\mathbf{x}-\mathbf{X}\boldsymbol{\theta}^*)\right\|_\infty \\ &\leq \frac{3m}{m+1}\frac{k\lambda_n^2}{\alpha},\end{aligned}$$

where we have used the choice of λ_n and (62) to conclude:

$$\left\|\mathbf{X}^\top(\mathbf{x}-\mathbf{X}\boldsymbol{\theta}^*)\right\|_\infty \leq \frac{\lambda_n}{4}.$$

Combining these two bounds via the triangle inequality concludes the proof. □

Proposition A.2 (Deviation Result for the *Reduced* Model)**.** Suppose $\widehat{\boldsymbol{\Sigma}} \sim \mathrm{RE}(\alpha,\tau)$, with τ satisfying $32k\tau/\alpha = (m-1)/m$ for some $m > 1$ and $(\mathbf{X}_{(1)},\mathbf{x})$ satisfying the deviation bound (63). Let J denote the support of $\boldsymbol{\theta}^*_{(1)}$, with its complement denoted by J^c. Then, for any $\lambda_n \geq 4\mathbb{Q}'(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)\sqrt{\log(2p)/n}$, the solution to *reduced* model in (6) satisfies:

$$\begin{aligned}\left|\ell\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)},\mathbf{0}\right)-\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)},\mathbf{0}\right)\right| &\leq 20\frac{k\lambda_n^2}{\alpha/m} \\ &\quad+\left(8\sqrt{2m}+18\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1 \\ &=: \Delta_R, \end{aligned} \tag{66}$$

Proof. In a similar fashion to Proposition A.1, by invoking the consistency results of LASSO for the *reduced* model under the RE and deviation bound assumptions (See Proposition A.4) and noting (63) we have:

$$\begin{aligned}\left|\ell\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)},\mathbf{0}\right)-\ell\left(\widetilde{\boldsymbol{\theta}}^*_{(1)},\mathbf{0}\right)\right| &\leq 12\frac{k\lambda_n^2}{\alpha/m} \\ &\quad+\left(8\left(\sqrt{2m}+1\right)+2\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1 \\ &\quad+16\sqrt{\frac{\lambda_n^3 k}{\alpha/m}}\sqrt{\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1}. \end{aligned} \tag{67}$$

Simplifying the last term using Arithmetic Mean-Geometric Mean inequality proves the claim. □

To establish the consistency results used in Proposition A.1 and Proposition A.2, we first state a result adapted from [28] on the prediction error of LASSO under the *full* model:

Proposition A.3 (Prediction Error for the *Full* Model)**.** Suppose $\widehat{\boldsymbol{\Sigma}} \sim \mathrm{RE}(\alpha,\tau)$, with τ satisfying $32k\tau/\alpha = (m-1)/m$ for some $m > 1$ and $(\mathbf{X},\mathbf{x})$ satisfying the deviation bound (62). Then for any $\lambda_n \geq 4\mathbb{Q}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}}$, the solution to the *full* model in (6) satisfies:

$$\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_2 \leq \frac{3m}{m+1}\frac{\sqrt{k}\lambda_n}{\alpha}, \tag{68}$$

$$\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|_1 \leq \frac{12m}{m+1}\frac{k\lambda_n}{\alpha}, \tag{69}$$

$$(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)^\top\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*) \leq \frac{18m}{m+1}\frac{k\lambda_n^2}{\alpha}. \tag{70}$$

Proof. The proof closely follows that of [28, Proposition 4.1], and is thus omitted for brevity. □

In what follows, we show that the particular choice of τ satisfying $32k\tau/\alpha = (m-1)/m$ for some $m > 1$ will simplify the prediction error analysis of the *reduced* model. As for the *reduced* model, the main technical difficulty in establishing prediction error bounds stems from the fact that $\widetilde{\boldsymbol{\theta}}^*_{(1)}$ is no longer k–sparse. We address this issue in the following proposition:

Proposition A.4 (Prediction Error for the *Reduced* Model)**.** Suppose $\widehat{\boldsymbol{\Sigma}} \sim \mathrm{RE}(\alpha,\tau)$, with τ satisfying the relation in Proposition A.3 and $(\mathbf{X}_{(1)},\mathbf{x})$ satisfying the deviation bound (63). Let J denote the support of $\boldsymbol{\theta}^*_{(1)}$, with its complement denoted by J^c. Then, for any $\lambda_n \geq 4\mathbb{Q}'(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)\sqrt{\log(2p)/n}$, the solution to *reduced* model in (6) satisfies the inequalities (71), (72) and (73).

Proof. Recall that $\widetilde{\boldsymbol{\theta}}^*_{(1)} := \boldsymbol{\theta}^*_{(1)} + \mathbf{C}_{11}^{-1}\mathbf{C}_{21}\boldsymbol{\theta}^*_{(2)}$. Let us define

$$F(\boldsymbol{\theta}_{(1)}) := \frac{1}{n}\left\|\mathbf{x}-\mathbf{X}_{(1)}\boldsymbol{\theta}_{(1)}\right\|_2^2 + \lambda_n\|\boldsymbol{\theta}_{(1)}\|_1.$$

and consider the quantity, $\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) := F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) - F\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)$. Also, let $\mathbf{v} := \boldsymbol{\theta}^*_{(1)} - \widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}$ and $\widetilde{\mathbf{v}} := \mathbf{v} + \mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}$. Then, $\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right)$ can be simplified as:

$$\begin{aligned}\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) &= F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) - F\left(\widetilde{\boldsymbol{\theta}}^*_{(1)}\right) \\ &= \frac{1}{n}\widetilde{\mathbf{v}}^\top\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\widetilde{\mathbf{v}} + \frac{2}{n}\mathbf{X}_{(1)}^\top\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)}\right) \\ &\quad+\lambda_n\left(\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right\|_1 - \left\|\widetilde{\boldsymbol{\theta}}^*_{(1)}\right\|_1\right).\end{aligned}$$

Using the fact that

$$\left\|\frac{2}{n}\mathbf{X}_{(1)}^\top\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)\right\|_\infty \leq \frac{\lambda_n}{2},$$

we get:

$$\begin{aligned}\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) &\geq \frac{1}{n}\widetilde{\mathbf{v}}^\top\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\widetilde{\mathbf{v}} - \frac{\lambda_n}{2}(\|\widetilde{\mathbf{v}}_J\|_1+\|\widetilde{\mathbf{v}}_{J^c}\|_1) \\ &\quad+\lambda_n\left(\|\widetilde{\mathbf{v}}_{J^c}\|_1-\|\widetilde{\mathbf{v}}_J\|_1-2\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right),\end{aligned}$$

where we split the error $\widetilde{\mathbf{v}}$ into components within J and components within J^c. Now, since $\widetilde{\mathbf{v}}^\top\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\widetilde{\mathbf{v}}$ is non-negative, we arrive at:

$$\begin{aligned}0 &\geq -\frac{\lambda_n}{2}\left(\|\widetilde{\mathbf{v}}_J\|_1 + \|\widetilde{\mathbf{v}}_{J^c}\|_1\right) \\ &\quad+\lambda_n\left(\|\widetilde{\mathbf{v}}_{J^c}\|_1 - \|\widetilde{\mathbf{v}}_J\|_1 - 2\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right) \\ &= -\frac{\lambda_n}{2}\left(3\|\widetilde{\mathbf{v}}_J\|_1-\|\widetilde{\mathbf{v}}_{J^c}\|_1+4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right).\end{aligned} \tag{74}$$

The rest of the treatment follows the derivation of weakly sparse or compressible models (See, for example, the derivation of the main theorem in [35]). Using the inequality (74), we can write:

$$\|\widetilde{\mathbf{v}}\|_1 = \|\widetilde{\mathbf{v}}_J\|_1+\|\widetilde{\mathbf{v}}_{J^c}\|_1 \leq 4\|\widetilde{\mathbf{v}}_J\|_1+4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1$$

$$\leq 4\sqrt{k}\|\widetilde{\mathbf{v}}_J\|_2+4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1, \tag{75}$$

where the last inequality follows from the fact $|J| \leq k$. This inequality together with the RE condition and τ satisfying the relation in Proposition A.3, allows to write:

$$\begin{aligned}\frac{1}{n}\widetilde{\mathbf{v}}^\top\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\widetilde{\mathbf{v}} &\geq \alpha\|\widetilde{\mathbf{v}}\|_2^2-\alpha\frac{m-1}{m}\|\widetilde{\mathbf{v}}\|_2^2-\frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2\\ &\geq \frac{\alpha}{m}\|\widetilde{\mathbf{v}}\|_2^2-\frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2, \end{aligned}\tag{76}$$

using the inequality $\sqrt{2(a^2+b^2)} \geq (a+b)$.

Combining (74) and (76), we then finally arrive at:

$$\begin{aligned}\Delta F\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}\right) &\geq \frac{\alpha}{m}\|\widetilde{\mathbf{v}}\|_2^2-\frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2\\ &\quad-\frac{\lambda_n}{2}\left(3\|\widetilde{\mathbf{v}}_J\|_1-\|\widetilde{\mathbf{v}}_{J^c}\|_1+4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right)\\ &\geq \frac{\alpha}{m}\|\widetilde{\mathbf{v}}\|_2^2-\frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2 &(77)\\ &\quad-\frac{\lambda_n}{2}\left(3\sqrt{k}\|\widetilde{\mathbf{v}}\|_2+4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right), &(78)\end{aligned}$$

where the last step follows from the inequalities $\|\widetilde{\mathbf{v}}_J\|_1 \leq \sqrt{k}\|\widetilde{\mathbf{v}}_J\|_2 \leq \sqrt{k}\|\widetilde{\mathbf{v}}\|_2$ and $\|\widetilde{\mathbf{v}}_{J^c}\|_1 \geq 0$.

An application of Lemma C.5 establishes that the right hand side of the inequality (78) will be positive if:

$$\begin{aligned}\|\widetilde{\mathbf{v}}\|^2 \geq \frac{9}{4}\frac{\lambda_n^2}{\alpha^2/m^2}k\\ +\frac{\lambda_n}{\alpha/m}\left(\frac{2}{\lambda_n}\frac{\alpha}{m}\frac{m-1}{k}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1^2+4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right).\end{aligned}\tag{79}$$

From the latter inequality, the first claim of the proposition follows using the fact that $\|\mathbf{a}\|_1 \geq \|\mathbf{a}\|_2$; the second claim follows from the first claim together with (75); the last claim follows from the fact that:

$$\frac{1}{n}\widetilde{\mathbf{v}}^\top\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\widetilde{\mathbf{v}} \leq \frac{\lambda_n}{2}\left(3\|\widetilde{\mathbf{v}}_J\|_1-\|\widetilde{\mathbf{v}}_{J^c}\|_1+4\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1\right),\tag{80}$$

which concludes the proof of the proposition. □

APPENDIX B
VERIFICATION OF CONDITIONS 1 AND 2

Having established Propositions A.1 and A.2, it only remains to show that Condition 1 and Condition 2 hold with high probability, under both the *full* and *reduced* models.

The following proposition establishes that the RE condition (Condition 1) holds with high probability:

Proposition B.1 (Verifying RE for $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^\top\mathbf{X}/n$)**.** Let

$$\zeta := 54\frac{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)/\mu_{\min}(\breve{\mathbf{A}})}{\Lambda_{\min}(\boldsymbol{\Sigma}_\epsilon)/\mu_{\max}(\mathbf{A})}, \quad \alpha := \frac{\Lambda_{\min}(\boldsymbol{\Sigma}_\epsilon)}{2\mu_{\max}(\mathbf{A})},$$
$$\tau := \frac{4\alpha\max\{\zeta^2,1\}}{c}\frac{\log(2p)}{n}.$$

Then, for $n \geq C_0\max\{\zeta^2,1\}k\log(2p)$, there exist constants c_1, c_2 such that

$$\mathbb{P}\left[\widehat{\boldsymbol{\Sigma}} \sim \mathrm{RE}\left(\alpha,\tau\right)\right] \geq 1-c_1\exp(-c_2 n\min\{\zeta^{-2},1\}).\tag{81}$$

Proof. The proof closely follows that of [28, proof of Proposition 4.2], and is thus omitted for brevity. □

Remark B.1. In order for τ to satisfy $32k\tau/\alpha = (m-1)/m$ for some $m > 1$, we precisely need $n \geq (128/c)(m/(m-1))\max\{\zeta^2,1\}k\log(2p)$.

Finally, the following two propositions establish that the deviation conditions (Condition 2) hold with high probability.

Proposition B.2 (Deviation Condition for the *full* Model)**.** For $n \geq D_0\log(2p)$, there exist constants d_0, d_1 and $d_2 > 0$ such that

$$\mathbb{P}\left[\left\|\frac{1}{n}\mathbf{X}^\top(\mathbf{x}-\mathbf{X}\boldsymbol{\theta}^*)\right\|_\infty \geq \mathbb{Q}(\boldsymbol{\theta}^*,\Sigma_\epsilon)\sqrt{\frac{\log(2p)}{n}}\right] \leq \frac{d_1}{(2p)^{d_2}},\tag{82}$$

where

$$\mathbb{Q}(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon) := d_0\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\left(1+\frac{1+\mu_{\max}(\mathbf{A})}{\mu_{\min}(\mathbf{A})}\right).$$

Proof. The proof follows that of [28, Proposition 4.3], and is thus omitted for brevity. □

Proposition B.3 (Deviation Condition for the *Reduced* Model)**.** For $n \geq D_0'\log(2p)$, there exist constants d_0', d_1', and $d_2' > 0$ such that

$$\mathbb{P}\left[\left\|\frac{1}{n}\mathbf{X}_{(1)}^\top\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)\right\|_\infty \geq \mathbb{Q}'(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}}\right] \leq \frac{d_1'}{(2p)^{d_2'}},\tag{83}$$

where $\mathbb{Q}'(\boldsymbol{\theta}^*,\boldsymbol{\Sigma}_\epsilon)$ is defined as in (84).

Proof. In the *reduced* model, the deviation can be expressed as:

$$\begin{aligned}&\frac{1}{n}\mathbf{X}_{(1)}^\top\left(\mathbf{x}-\mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)}-\mathbf{X}_{(2)}\mathbf{0}\right)\\ =&\frac{1}{n}\mathbf{X}_{(1)}^\top\left(-\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}+\mathbf{X}_{(2)}\boldsymbol{\theta}^*_{(2)}+\boldsymbol{\epsilon}\right)\end{aligned}$$

$$\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}-\widetilde{\boldsymbol{\theta}}^*_{(1)}\right\|_2 \leq \frac{3}{2}\frac{\lambda_n\sqrt{k}}{\alpha/m}+\sqrt{\frac{2m}{k}}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1+\sqrt{\frac{4\lambda_n}{\alpha/m}\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1},\tag{71}$$

$$\left\|\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}-\widetilde{\boldsymbol{\theta}}^*_{(1)}\right\|_1 \leq 6\frac{\lambda_n k}{\alpha/m}+4(\sqrt{2m}+1)\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1+8\sqrt{\frac{\lambda_n k}{\alpha/m}}\sqrt{\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1},\tag{72}$$

$$\frac{1}{n}\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}-\widetilde{\boldsymbol{\theta}}^*_{(1)}\right)^\top\mathbf{X}_{(1)}^\top\mathbf{X}_{(1)}\left(\widehat{\widetilde{\boldsymbol{\theta}}}_{(1)}-\widetilde{\boldsymbol{\theta}}^*_{(1)}\right) \leq 9\frac{\lambda_n^2 k}{\alpha/m}+\left(6\left(\sqrt{2m}+1\right)+2\right)\lambda_n\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1+12\sqrt{\frac{\lambda_n^3 k}{\alpha/m}}\sqrt{\left\|\widetilde{\boldsymbol{\theta}}^*_{(1)J^c}\right\|_1}.\tag{73}$$

$$= -\frac{1}{n}\mathbf{X}_{(1)}^\top \mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)} + \frac{1}{n}\mathbf{X}_{(1)}^\top \mathbf{X}_{(2)}\boldsymbol{\theta}^*_{(2)} + \frac{1}{n}\mathbf{X}_{(1)}^\top \boldsymbol{\epsilon}.$$

The last term can be bounded in a similar fashion as done for the deviation in Proposition B.2. The i^{th} component of the first two terms can be expressed as $\mathbf{e}_i^\top \frac{1}{n}\mathbf{X}^\top\mathbf{X}\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right]$, for $i = 1, 2, \cdots, p$ where $\mathbf{e}_i$'s are the standard unit bases in $\mathbb{R}^{2p}$. Invoking [28, Proposition 2.4(a)] and noting that $\mathcal{M}(\mathbf{F}) \leq \Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\big/\mu_{\min}(\breve{\mathbf{A}})$, we get (85). Using the union bound, we can then arrive at (86). Using the latter inequality, the statement of the proposition follows from the same arguments used in the proof of [28, Proposition 4.3]. □

Appendix C
Concentration Inequalities and Technical Lemmas

Lemma C.1. Given i.i.d. samples from a normal distribution, i.e., $w_t \sim \mathcal{N}(0, \sigma^2)$, the following holds:

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^{n}\frac{w_t^2}{\sigma^2} - 1\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{8}\right), \forall t \in (0,1). \tag{87}$$

Proof. Define $z_t = w_t/\sigma \sim \mathcal{N}(0,1)$, then $\sum_{t=1}^n z_t^2 \sim \chi^2(n)$. Clearly, z_t^2 is sub-exponential with parameters $(2,4)$, so is the sum $\sum_{t=1}^n z_t^2$ with parameters $(2\sqrt{n}, 4)$. The claim of the lemma then follows from standard sub-exponential tail bounds. □

Lemma C.2 (Concentration of the *Full* Model Deviation)**.** Under the *full* model, we have, for $\Delta_N < (\boldsymbol{\Sigma}_\epsilon)_{1,1}$:

$$\mathbb{P}\left[\left|\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - (\boldsymbol{\Sigma}_\epsilon)_{1,1}\right| \geq \Delta_N\right] \leq 2\exp\left(-\frac{n\Delta_N^2}{8(\boldsymbol{\Sigma}_\epsilon)_{1,1}}\right).$$

Proof. Note that $\ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) = \sum_{t=1}^n \epsilon_t^2/n$. Since $\epsilon_t \sim \mathcal{N}(0, (\boldsymbol{\Sigma}_\epsilon)_{1,1})$, using Lemma C.1 we get:

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{t=1}^{n}\frac{\epsilon_t^2}{(\boldsymbol{\Sigma}_\epsilon)_{1,1}} - 1\right| \geq t\right] \leq 2\exp\left(-\frac{nt^2}{8}\right), \forall t \in (0,1). \tag{88}$$

By letting $\Delta_N := t(\boldsymbol{\Sigma}_\epsilon)_{1,1}$, the claim of the lemma follows. □

Lemma C.3 (Concentration of the *Reduced* Model Deviation)**.** Suppose that the deviation conditions (Condition 2) hold. Then, there exist constants c_3 and $c_4 > 0$ such that

$$\left|\ell\Big(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\Big) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) - D\right| \leq \Delta_D, \tag{89}$$

with probability at least

$$1 - c_3\exp\left(-c_4 n\min\left\{\zeta^{-2}\frac{\log(2p)}{n}, 1\right\}\right),$$

$$\Delta_D := \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}}\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right]\right\|_1 + \frac{\alpha}{27}\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right]\right\|_2^2,$$

and

$$D := {\boldsymbol{\theta}^*_{(2)}}^\top(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\boldsymbol{\theta}^*_{(2)}.$$

with α and $\mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)$ defined in Proposition B.1 and Proposition B.2, respectively.

Proof. Since $\widetilde{\boldsymbol{\theta}}^*_{(1)} = \boldsymbol{\theta}^*_{(1)} + \mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}$, we have:

$$\begin{aligned}
&\ell\Big(\widetilde{\boldsymbol{\theta}}^*_{(1)}, \mathbf{0}\Big) - \ell\Big(\boldsymbol{\theta}^*_{(1)}, \boldsymbol{\theta}^*_{(2)}\Big) \\
&\quad= \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}_{(1)}\widetilde{\boldsymbol{\theta}}^*_{(1)} - \mathbf{X}_{(2)}\mathbf{0}\right\|_2^2 \\
&\qquad- \frac{1}{n}\left\|\mathbf{x} - \mathbf{X}\boldsymbol{\theta}^*\right\|_2^2 \\
&\quad= \frac{1}{n}\left\|-\mathbf{X}_{(1)}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)} + \mathbf{X}_{(2)}\boldsymbol{\theta}^*_{(2)} + \boldsymbol{\epsilon}\right\|_2^2 \\
&\qquad- \frac{1}{n}\|\boldsymbol{\epsilon}\|_2^2 \\
&\quad= \frac{2}{n}\boldsymbol{\epsilon}^\top\mathbf{X}\boldsymbol{\vartheta} + \boldsymbol{\vartheta}^\top\frac{1}{n}\mathbf{X}^\top\mathbf{X}\boldsymbol{\vartheta},
\end{aligned}$$

where $\boldsymbol{\vartheta} := \left[(-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)})^\top, {\boldsymbol{\theta}^*_{(2)}}^\top\right]^\top$. Using the Condition 2, we get:

$$\begin{aligned}
\left|\frac{1}{n}\boldsymbol{\epsilon}^\top\mathbf{X}\boldsymbol{\vartheta}\right| &\leq \left\|\frac{1}{n}\mathbf{X}^\top\epsilon\right\|_\infty \|\boldsymbol{\vartheta}\|_1 \\
&\leq \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}}\|\boldsymbol{\vartheta}\|_1 \\
&\leq \mathbb{Q}(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}_\epsilon)\sqrt{\frac{\log(2p)}{n}}\left(\left\|\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right\|_1 + 1\right)\left\|\boldsymbol{\theta}^*_{(2)}\right\|_1.
\end{aligned}$$

$$\mathbb{Q}'(\boldsymbol{\theta}^*, \Sigma_\epsilon) := d_0'\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\left(1 + \frac{1+\mu_{\max}(\mathbf{A})}{\mu_{\min}(\mathbf{A})} + \frac{3\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right]\right\|_2}{\mu_{\min}(\breve{\mathbf{A}})}\right). \tag{84}$$

$$\mathbb{P}\left[\left|\mathbf{e}_i^\top\frac{1}{n}\mathbf{X}^\top\mathbf{X}\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right]\right| \geq 3\frac{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)}{\mu_{\min}(\breve{\mathbf{A}})}\eta\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right]\right\|_2\right] \leq 6\exp[-cn\min\{\eta, \eta^2\}]. \tag{85}$$

$$\mathbb{P}\left[\left|\frac{1}{n}\mathbf{e}_i^\top\left(\mathbf{X}^\top\mathbf{X}\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right] + \mathbf{X}^\top\boldsymbol{\epsilon}\right)\right| \geq \Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)\left(1 + \frac{1+\mu_{\max}(\mathbf{A})}{\mu_{\min}(\mathbf{A})} + \frac{3\left\|\left[-\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\boldsymbol{\theta}^*_{(2)}; \boldsymbol{\theta}^*_{(2)}\right]\right\|_2}{\mu_{\min}(\breve{\mathbf{A}})}\right)\eta\right] \leq 12\exp[-cn\min\{\eta, \eta^2\}]. \tag{86}$$

Furthermore, from [28, Proposition 2.4], for any $\boldsymbol{\vartheta} \in \mathbb{R}^{2p}$ and $\eta \geq 0$, there exists a constant $c > 0$ such that:

$$\mathbb{P}\left[\left|\boldsymbol{\vartheta}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}-\mathbf{C}\right)\boldsymbol{\vartheta}\right| \geq \eta\|\boldsymbol{\vartheta}\|^2\frac{\Lambda_{\max}(\boldsymbol{\Sigma}_\epsilon)}{\mu_{\min}(\breve{\mathbf{A}})}\right] \leq 2\exp[-cn\min\{\eta,\eta^2\}].$$

Next, with the choice of $\eta = \zeta^{-1}\sqrt{\log 2p/n}$, the latter concentration inequality establishes that:

$$\left|\boldsymbol{\vartheta}^\top\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\boldsymbol{\vartheta} - \boldsymbol{\vartheta}^\top\mathbf{C}\boldsymbol{\vartheta}\right| \leq \frac{\alpha}{27}\sqrt{\frac{\log(2p)}{n}}\,\|\boldsymbol{\vartheta}\|_2^2\,,$$

with probability at least $1 - 2\exp(-cn\min\{\zeta^{-2}\log(2p)/n, 1\})$. This along with the following observation concludes the proof of the lemma: $\boldsymbol{\vartheta}^\top\mathbf{C}\boldsymbol{\vartheta} = \boldsymbol{\theta}^{*\top}_{(2)}(\mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\boldsymbol{\theta}^*_{(2)}$. □

Lemma C.4 (Fano's Inequality)**.** Let $\mathcal{Z}$ be a class of densities with a subclass $\mathcal{Z}^\star$ of densities $f_{\boldsymbol{\theta}^i_{(2)}}$, parameterized by $\boldsymbol{\theta}^i_{(2)}$, for $i \in \{0, \cdots, 2^M\}$. Suppose that for any two distinct $\boldsymbol{\theta}^i_{(2)}, \boldsymbol{\theta}^j_{(2)} \in \mathcal{Z}^\star$, $\mathcal{D}_{\mathsf{KL}}\left(f_{\boldsymbol{\theta}^i_{(2)}}\middle\|f_{\boldsymbol{\theta}^j_{(2)}}\right) \leq \xi$ for some constant ξ. Let $\widehat{\boldsymbol{\theta}}$ be an estimate of the parameters. Then

$$\sup_j \mathbb{P}\left[\widehat{\boldsymbol{\theta}}_{(2)} \neq \boldsymbol{\theta}^j_{(2)}|H_j\right] \geq 1 - \frac{\xi + \log 2}{M}, \tag{90}$$

where H_j denotes the hypothesis that $\boldsymbol{\theta}_j$ is the true parameter, and induces the probability measure $\mathbb{P}[\cdot|H_j]$.

Proof. See [83, Page 323]. □

Finally, the following elementary lemma provides a useful technical tool for simplifying some of the algebraic inequalities:

Lemma C.5. The quadratic function $f(x) = ax^2 - bx - c$ with $a, b, c > 0$ is positive for all real x satisfying

$$x^2 \geq \left(\frac{b}{a}\right)^2 + 2\frac{c}{a}.$$

Proof. Clearly, $f(x)$ has only one positive root, denoted here by x_+, which satisfies $f(x) > 0$, $\forall\, x > x_+$. Using the following instance of Jensen's inequality,

$$\sqrt{1+z} \leq \frac{1}{2}z + 1 \quad \forall \quad z > 0,$$

the positive root x_+ can be upper bounded as:

$$\begin{aligned} x_+^2 &= \left(\frac{b}{2a} + \sqrt{\left(\frac{b}{2a}\right)^2 + \frac{c}{a}}\right)^2 \\ &= 2\left(\frac{b}{2a}\right)^2 + \frac{c}{a} + 2\frac{b}{2a}\sqrt{\left(\frac{b}{2a}\right)^2 + \frac{c}{a}} \\ &\leq 2\left(\frac{b}{2a}\right)^2 + \frac{c}{a} + 2\left(\frac{b}{2a}\right)^2\left(\frac{1}{2}\left(\frac{2a}{b}\right)^2\frac{c}{a} + 1\right) \\ &= \left(\frac{b}{a}\right)^2 + 2\frac{c}{a} = x_0^2 \end{aligned}$$

Then, $x^2 \geq x_0^2$ implies $ax^2 - bx - c > 0$. □

ACKNOWLEDGMENT

The authors would like to thank Patrick L. Purdon and Emery N. Brown for providing the data from [80].

REFERENCES

[1] W. H. Greene, *Econometric Analysis*, 5th ed. The address: Pearson Education, Inc., Upper Saddle River, New Jersey, 07458, 2003.

[2] S. L. Bressler and A. K. Seth, "Wiener-Granger causality: A well established methodology," *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.

[3] K. Friston, R. Moran, and A. K. Seth, "Analysing connectivity with Granger causality and dynamic causal modelling," *Curr. Opin. Neurobiol.*, vol. 23, no. 2, pp. 172–178, 2013.

[4] A. K. Seth, A. B. Barrett, and L. Barnett, "Granger Causality Analysis in Neuroscience and Neuroimaging," *J. Neurosci.*, vol. 35, no. 8, pp. 3293–3297, 2015.

[5] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach, *Systems biology in practice: concepts, implementation and application*. John Wiley & Sons, 2005.

[6] J. Feng, J. Jost, and M. Qian, *Networks: from biology to theory*. Springer, 2007, vol. 80.

[7] C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[8] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *J. Am. Stat. Assoc.*, vol. 77, no. 378, pp. 304–313, 1982.

[9] J. F. Geweke, "Measures of Conditional Linear Dependence and Feedback Between Time Series," *J. Am. Stat. Assoc.*, vol. 79, no. 388, pp. 907–915, 1984.

[10] S. Psillos, *Causation and Explanation*. Routledge, 2014.

[11] I. E. Marinescu, P. N. Lawlor, and K. P. Kording, "Quasi-experimental causality in neuroscience and behavioural research," *Nature human behaviour*, vol. 2, no. 12, pp. 891–898, 2018.

[12] J. Pearl, *Causality*. Cambridge university press, 2009.

[13] C. W. J. Granger and P. Newbold, *Forecasting Economic Time Series*. 2nd Ed., Academic Press, 1986.

[14] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. 19, no. 6, pp. 716–723, 1974.

[15] G. Schwarz, "Estimating the Dimension of a Model," *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[16] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A Granger causality measure for point process models of ensemble neural spiking activity," *PLoS Comput. Biol.*, vol. 7, no. 3, 2011.

[17] A. Wald, "Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large," *Trans. Am. Math. Soc.*, vol. 54, no. 3, pp. 426–482, 1943.

[18] R. R. Davidson and W. E. Lever, "The Limiting Distribution of the Likelihood Ratio Statistic under a Class of Local Alternatives," *Sankhy Indian J. Stat. Ser. A*, vol. 32, no. 2, pp. 209–224, 1970.

[19] C. Zou and J. Feng, "Granger causality vs. dynamic Bayesian network inference: a comparative study," *BMC Bioinformatics*, vol. 10, no. 1, p. 122, 2009.

[20] A. K. Seth, "A MATLAB toolbox for Granger causal connectivity analysis," *J. Neurosci. Methods*, vol. 186, no. 2, pp. 262–273, 2010.

[21] P. P. Mitra and B. Pesaran, "Analysis of dynamic brain imaging data," *Biophys. J.*, vol. 76, no. 2, pp. 691–708, 1999.

[22] M. T. Bahadori and Y. Liu, "An examination of practical Granger causality inference," in *Proc. 2013 SIAM Int. Conf. data Min.* SIAM, 2013, pp. 467–475.

[23] A. Goldenshluger and A. Zeevi, "Nonasymptotic Bounds for Autoregressive Time Series Modeling," *Ann. Stat.*, vol. 29, no. 2, pp. 417–444, Mar 2001.

[24] H. Wang, G. Li, and C.-L. Tsai, "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 69, no. 1, pp. 63–78, 2007.

[25] Y. Nardi and A. Rinaldo, "Autoregressive process modeling via the Lasso procedure," *J. Multivar. Anal.*, vol. 102, no. 3, pp. 528–549, 2011.

[26] F. Han and H. Liu, "Transition matrix estimation in high dimensional time series," in *Int. Conf. Mach. Learn.*, 2013, pp. 172–180.

[27] A. Kazemipour, S. Miran, P. Pal, B. Babadi, and M. Wu, "Sampling requirements for stable autoregressive estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 9, pp. 2333–2347, 2017.

[28] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Stat.*, vol. 43, no. 4, pp. 1535–1567, 2015.

[29] K. C. Wong, Z. Li, and A. Tewari, "Lasso guarantees for β-mixing heavy-tailed time series," *Ann. Stat.*, vol. 48, no. 2, pp. 1124–1142, 2020.

[30] A. Skripnikov and G. Michailidis, "Regularized joint estimation of related vector autoregressive models," *Comput. Stat. Data Anal.*, vol. 139, pp. 164–177, 2019.

[31] S. Basu, X. Li, and G. Michailidis, "Low Rank and Structured Modeling of High-Dimensional Vector Autoregressions," *IEEE Trans. Signal Process.*, vol. 67, no. 5, pp. 1207–1222, 2019.

[32] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[33] P.-L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity," in *Adv. Neural Inf. Process. Syst. 24*, 2011, pp. 2726–2734.

[34] P. Bühlmann and S. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*. Berlin: Springer Science & Business Media, 2011.

[35] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers," *Stat. Sci.*, vol. 27, no. 4, pp. 538–557, 2012.

[36] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[37] N.-J. Hsu, H.-L. Hung, and Y.-M. Chang, "Subset selection for vector autoregressive processes using Lasso," *Comput. Stat. Data Anal.*, vol. 52, no. 7, pp. 3645–3657, 2008.

[38] Y. Ren and X. Zhang, "Subset selection for vector autoregressive processes via adaptive Lasso," *Stat. Probab. Lett.*, vol. 80, no. 23, pp. 1705–1712, 2010.

[39] ——, "Model selection for vector autoregressive processes via adaptive Lasso," *Commun. Stat. Methods*, vol. 42, no. 13, pp. 2423–2436, 2013.

[40] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.*, 2007, pp. 66–75.

[41] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. C. Sogayar, and C. E. Ferreira, "Modeling gene expression regulatory networks with the sparse vector autoregressive model," *BMC Systems Biology*, vol. 1, no. 1, p. 39, Dec. 2007.

[42] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical Granger modeling for gene expression regulatory networks discovery," *Bioinformatics*, vol. 25, no. 12, pp. i110–i118, 2009.

[43] A. Shojaie and G. Michailidis, "Discovering graphical Granger causality using the truncating lasso penalty," *Bioinformatics*, vol. 26, no. 18, pp. i517—i523, 2010.

[44] Y. Liu, M. T. Bahadori, and H. Li, "Sparse-gev: Sparse latent space model for multivariate extreme value time serie modeling," in *Proc. 29 th Int. Conf. Mach. Learn. Edinburgh, Scotland, UK,*, 2012.

[45] S. Basu, A. Shojaie, and G. Michailidis, "Network Granger Causality with Inherent Grouping Structure," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 417–453, 2015.

[46] G. Michailidis and F. D'Alché-Buc, "Autoregressive models for gene regulatory network inference: Sparsity, stability and causality issues," *Math. Biosci.*, vol. 246, no. 2, pp. 326–334, 2013.

[47] T. Shimamura, S. Imoto, R. Yamaguchi, A. Fujita, M. Nagasaki, and S. Miyano, "Recursive regularization for inferring gene networks from time-course gene expression profiles," *BMC Systems Biology*, vol. 3, no. 1, p. 41, Dec. 2009.

[48] A. Tank, I. Covert, N. Foti, A. Shojaie, and E. B. Fox, "Neural granger causality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4267–4279, 2021.

[49] W. Tang, S. L. Bressler, C. M. Sylvester, G. L. Shulman, and M. Corbetta, "Measuring Granger causality between cortical regions from voxelwise fMRI BOLD signals with LASSO," *PLoS Comput. Biol.*, vol. 8, no. 5, 2012.

[50] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure, "On asymptotically optimal confidence regions and tests for high-dimensional models," *Ann. Stat.*, vol. 42, no. 3, pp. 1166–1202, 2014.

[51] S. van de Geer, "On the asymptotic variance of the debiased Lasso," *Electron. J. Stat.*, vol. 13, no. 2, pp. 2970–3008, 2019.

[52] A. Javanmard and A. Montanari, "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2869–2909, 2014.

[53] ——, "Debiasing the lasso: Optimal sample size for gaussian designs," *Ann. Stat.*, vol. 46, no. 6A, pp. 2593–2622, 2018.

[54] A. Javanmard and H. Javadi, "False discovery rate control via debiased lasso," *Electron. J. Stat.*, vol. 13, no. 1, pp. 1212–1253, 2019.

[55] C.-H. Zhang and S. S. Zhang, "Confidence intervals for low dimensional parameters in high dimensional linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 1, pp. 217–242, Jan. 2014.

[56] M. Dhamala, G. Rangarajan, and M. Ding, "Analyzing information flow in brain networks with nonparametric Granger causality," *Neuroimage*, vol. 41, no. 2, pp. 354–362, 2008.

[57] I. Vlachos and D. Kugiumtzis, "Nonuniform state-space reconstruction and coupling detection," *Phys. Rev. E*, vol. 82, no. 1, p. 16207, Jul 2010.

[58] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.

[59] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, 2019.

[60] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

[61] G. H. Golub, P. C. Hansen, and D. P. O'Leary, "Tikhonov Regularization and Total Least Squares," *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 1, pp. 185–194, 1999.

[62] F. Natterer, "Error bounds for Tikhonov regularization in Hilbert scales," *Appl. Anal.*, vol. 18, no. 1-2, pp. 29–37, 1984.

[63] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Am. Stat. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[64] H. Xie and J. Huang, "SCAD-penalized regression in high-dimensional partially linear models," *Ann. Stat.*, vol. 37, no. 2, pp. 673–696, 2009.

[65] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B (statistical Methodol.)*, vol. 67, no. 2, pp. 301–320, 2005.

[66] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.

[67] J. Ding, V. Tarokh, and Y. Yang, "Model Selection Techniques: An Overview," *IEEE Signal Process. Mag.*, vol. 35, no. 6, pp. 16–34, Nov 2018.

[68] H. Lütkepohl, *New introduction to multiple time series analysis*. Springer-Verlag, Berlin, 2005.

[69] S. M. Kay, *Fundamentals of Statistical Signal Processing*, ser. Prentice Hall Signal Processing Series. Englewood Cliffs, N.J: Prentice-Hall PTR, 1993.

[70] A. Goldenshluger and A. Zeevi, "Nonasymptotic bounds for autoregressive time series modeling," *Annals of statistics*, pp. 417–444, 2001.

[71] F. J. MacWilliams and N. J. A. Sloane, *The theory of error correcting codes*. Elsevier, 1977, vol. 16.

[72] R. L. Graham and N. Sloane, "Lower bounds for constant weight codes," *Information Theory, IEEE Transactions on*, vol. 26, no. 1, pp. 37–43, 1980.

[73] S. van de Geer, P. Bühlmann, and S. Zhou, "The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso)," *Electronic Journal of Statistics*, vol. 5, pp. 688–749, 2011.

[74] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, vol. 37, no. 4, Aug. 2009.

[75] T. Goldstein and S. Osher, "The Split Bregman Method for L1-Regularized Problems," *SIAM J. Imaging Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[76] A. P. Wu, R. Singh, and B. Berger, "Granger causal inference on DAGs identifies genomic loci regulating transcription," in *International Conference on Learning Representations*, 2021.

[77] A. Shojaie, A. Jauhiainen, M. Kallitsis, and G. Michailidis, "Inferring regulatory networks by combining perturbation screens and steady state gene expression profiles," *PloS one*, vol. 9, no. 2, p. e82393, 2014.

[78] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[79] L. D. Lewis, S. Ching, V. S. Weiner, R. A. Peterfreund, E. N. Eskandar, S. S. Cash, E. N. Brown, and P. L. Purdon, "Local cortical dynamics of burst suppression in the anaesthetized brain," *Brain*, vol. 136, no. 9, pp. 2727–2737, 2013.

[80] L. D. Lewis, V. S. Weiner, E. A. Mukamel, J. A. Donoghue, E. N. Eskandar, J. R. Madsen, W. S. Anderson, L. R. Hochberg, S. S. Cash, E. N. Brown, and P. L. Purdon, "Rapid fragmentation of neuronal networks at the onset of propofol-induced unconsciousness," *Proc. Natl. Acad. Sci.*, vol. 109, no. 49, pp. E3377–E3386, 2012.
[81] S. Chauvette, S. Crochet, M. Volgushev, and I. Timofeev, "Properties of slow oscillation during slow-wave sleep and anesthesia in cats," *J. Neurosci.*, vol. 31, no. 42, pp. 14 998–15 008, 2011.
[82] B. O. Watson, D. Levenstein, J. P. Greene, J. N. Gelinas, and G. Buzsáki, "Network Homeostasis and State Dynamics of Neocortical Sleep," *Neuron*, vol. 90, no. 4, pp. 839–852, 2016.
[83] I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: asymptotic theory*. Springer Science & Business Media, LLC, 1981.

Proloy Das (S'17) received the B.Tech. degree (Hons.) in Instrumentation Engineering (with a minor in Electronics and Electrical Communication Engineering) from Indian Institute of Technology, Kharagpur, India, in 2015 and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Maryland, College Park, MD, USA in 2019 and 2020, respectively. He is currently a Research Fellow in the Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA, USA. He is a recipient of the 2020 ECE Distinguished Dissertation Award and the 2020 Deans Doctoral Student Research Award (second place) at the A. James Clark School of Engineering and received an Honorable Mention in the 2021 Charles Caramello Distinguished Dissertation Award Competition at the Graduate School, University of Maryland, College Park. His research interests span statistical signal processing, adaptive filtering, and machine learning with primary focus on neural and biological data analysis.

Behtash Babadi (S'08–M'14) received the B.Sc. degree in Electrical Engineering from Sharif University of Technology, Tehran, Iran in 2006, and the M.Sc. and Ph.D. degrees in Engineering Sciences from Harvard University, Cambridge, MA in 2008 and 2011, respectively. From 2011 to 2013, he was a post-doctoral fellow at the Department of Brain and Cognitive Sciences at Massachusetts Institute of Technology, Cambridge, MA as well as at the Department of Anesthesia, Critical Care and Pain Medicine at Massachusetts General Hospital, Boston, MA. He is currently an Associate Professor in the Department of Electrical and Computer Engineering and the Institute for Systems Research at the University of Maryland, College Park, MD, USA. He is the recipient of a 2016 NSF CAREER Award. His research interests include statistical and adaptive signal processing, neural signal processing, compressed sensing, and systems neuroscience.