Causal Network Inference via Group Sparse Regularization

Andrew Bolstad, *Member, IEEE,* Barry Van Veen, *Fellow, IEEE,* and Robert Nowak, *Fellow, IEEE*

Abstract

This paper addresses the problem of inferring sparse causal networks modeled by multivariate auto-regressive (MAR) processes. Conditions are derived under which the Group Lasso (gLasso) procedure consistently estimates sparse network structure. The key condition involves a "false connection score" ψ. In particular, we show that consistent recovery is possible even when the number of observations of the network is far less than the number of parameters describing the network, provided that $\psi < 1$. The false connection score is also demonstrated to be a useful metric of recovery in non-asymptotic regimes. The conditions suggest a modified gLasso procedure which tends to improve the false connection score and reduce the chances of reversing the direction of causal influence. Computational experiments and a real network based electrocorticogram (ECoG) simulation study demonstrate the effectiveness of the approach.

I. Introduction

The problem of inferring networks of causal relationships arises in biology, sociology, cognitive science and engineering. Specifically, suppose that we are able to observe the dynamical behaviors of N individual components of a system and that some, but not necessarily all, of the components may be causally influencing each other. We will refer to such a system as a causal network. To emphasize the network-centric viewpoint, we will use the terms node and network, instead of component and system, respectively. Causal network inference is the process of identifying the significant causal influences by observing the time-series at the nodes. For example, in electrocorticography (ECoG) the electrical signals in the brain are recorded directly and a goal is to identify the direction of information flow from one brain region to another.

One common tool for modeling causal influences is the multivariate autoregressive (MAR) model [1]–[3]. MAR models assume that the current measurement at a given node is a linear combination of the previous p measurements at all N nodes, plus an innovation noise:

$$\mathbf{x}(t) = \sum_{r=1}^{p} \mathbf{A}_r \mathbf{x}(t-r) + \mathbf{u}(t) \tag{1}$$

where $\mathbf{x}(t) = \begin{bmatrix} x_1(t) & x_2(t) & \dots & x_N(t) \end{bmatrix}^T$ is a vector of signal measurements across all N nodes at time t, matrices $\mathbf{A}_r = \{a_{i,j}(r)\}$ contain autoregressive coefficients describing the influence of node j on node i at a delay of r time samples, and $\mathbf{u}(t) = \begin{bmatrix} u_1(t) & u_2(t) & \dots & u_N(t) \end{bmatrix}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ is innovation noise. The MAR model is especially conducive to the assessment of Granger Causality, where time series x_j is said to Granger-cause x_i if knowledge of the past of x_j improves the prediction of x_i compared to using only the past of x_i [4].

The MAR model in Eq. (1) allows for the possibility of a fully connected network in which every node causally influences every other node. This flexibility is somewhat unrealistic and leads to practical challenges. In many networks each node is directly influenced by only a small subset of other nodes. The MAR model is overparameterized in such cases. This leads to serious practical problems. It may be impossible to reliably infer the network from noisy, finite-length time-series because of the large number of unknown coefficients in overparameterized models. We define the Sparse MAR Time-series (SMART) model to have the same form as Eq. (1) but include an extra parameter $\mathcal{S}_{\text{active}}$ denoting the index pairs of non-zero causal influences to eliminate overparameterization. For example, if node j influences node i, then $(i,j) \in \mathcal{S}_{\text{active}}$, otherwise $(i,j) \notin \mathcal{S}_{\text{active}}$ and $a_{i,j}(r) = 0$ for all time indices r. The SMART model for node i is given by:

$$x_i(t) = u_i(t) + \sum_{j:(i,j)\in\mathcal{S}_{\text{active}}} \sum_{r=1}^{p} a_{i,j}(r) x_j(t-r) \tag{2}$$

Applying Eq. (2) to each node $i = 1, 2, \dots, N$ in turn gives the SMART model for the whole network.

If the cardinality of the active set, denoted $|\mathcal{S}_{\text{active}}|$, is equal to N^2, then the SMART model is equivalent to the MAR model. We are primarily interested in networks for which $|\mathcal{S}_{\text{active}}| \leq mN$, for some constant $m > 1$. In such cases, the main inference challenge is reliably identifying the set $\mathcal{S}_{\text{active}}$, since once this is done the task of estimating the SMART coefficients is a simple

This work supported in part by the NIBIB under NIH awards EB005473, EB009749 and by the AFOSR award FA9550-09-1-0140.

This work is sponsored by the department of the Air Force under contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

and classical problem. In general, the amount of data required to reliably estimate SMART coefficients decreases as $|\mathcal{S}_{\text{active}}|$ decreases.

Identifying $\mathcal{S}_{\text{active}}$ is a subset selection problem. Simple subset selection problems can be solved using the well-known Lasso procedure. The Lasso mixes an ℓ_2 norm on the residual error with an ℓ_1 norm penalty on the regression coefficients favoring a solution in which most coefficients are zero [5]. However, ordinary Lasso does not capture the group structure of sparse connections in the SMART model. The Group Lasso (gLasso) procedure was first proposed by [6] in a general setting to promote group-structured sparsity patterns. gLasso penalties have recently been proposed for source localization in magneto-/electroencephalography (M/EEG) [7]–[12], as well as for identifying interaction patterns in the human brain [13] and in gene regulatory networks [14]. In both [13] and [14] the gLasso is effectively applied to SMART model estimation by penalizing the sum of ℓ_2 norms of the coefficients of each network link (ℓ_1 norm of ℓ_2 norms). We study estimation consistency of this technique which we term the SMART gLasso or SG.

Our main contribution is a novel characterization of the special conditions needed for consistency of the SG. These conditions are described in Section III. Existing gLasso consistency results do not apply to the temporal structure in the SMART model. The SG consistency conditions are similar in spirit to the standard "incoherence" conditions encountered in the analysis of Lasso and its variants [15], but are fundamentally different because of the autoregressive structure of our model. We define the "false connection score" and show that it yields a condition for consistent estimation of the underlying SMART sparsity. If this score is below one, then the network connectivity pattern can be recovered with high probability in the limit as the size of the network and the number of samples tends to infinity (although the number of samples can grow much slower than the network size). Conversely, if this score is above one, than an estimate that identifies all the correct connections will also include at least one false positive with high probability.

We also propose a variant of the SG in Section II which does not penalize self-connections (i.e., each node is free to influence itself). We call this variant Self-Connected SMART gLasso (SCSG) and show that it typically results in a lower false connection score for SMART models. We provide some example networks as well as their false connection scores for the SMART gLasso and SCSG approaches in Sec. V. We demonstrate the effectiveness of our results by simulating a variety of networks in Sec. VI. We also apply our results to a realistic brain network in Sec. VII by simulating the sparse connectivity pattern observed in the macaque brain.

II. Graph Inference with Lasso-Type Procedures

In this section we introduce the Lasso, gLasso, SG, and SCSG, and discuss previous consistency results.

A. Lasso and gLasso

Tibshirani first proposed the Least Absolute Shrinkage and Selection Operator (Lasso) in 1996 to "retain the good features of both subset selection and ridge regression" [5]. Although originally stated as an ℓ_1 norm constrained least squares optimization, the Lasso can also be stated as an unconstrained mixed-norm minimization. We consider the unconstrained problem throughout:

$$\hat{\mathbf{a}}^{Lasso} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \tag{3}$$

Here it is assumed that measured length n vector $\mathbf{y}$ is the result of a sparse linear combination of columns of $\mathbf{X}$; i.e. $\mathbf{y} = \mathbf{X}\mathbf{a}$ for sparse vector $\mathbf{a}$. The first term of (3) penalizes solutions which do not fit the measured data well, while the second term favors solution which are sparse. Yuan and Lin [6] introduced the Group Lasso (gLasso) extension to Tibshirani's Lasso in 2006. While the Lasso penalizes the ℓ_1 norm of the coefficient vector, the gLasso divides the coefficient vector into predetermined sub-vectors and penalizes the sum of the ℓ_2 norms of the sub-vectors; i.e., the ℓ_1 norm of ℓ_2 norms:

$$\hat{\mathbf{a}}^{gLasso} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{n}\left\|\mathbf{y} - \mathbf{X}\begin{bmatrix}\boldsymbol{\alpha}_1\\ \vdots \\ \boldsymbol{\alpha}_N\end{bmatrix}\right\|_2^2 + \lambda\sum_{i=1}^{N}\|\boldsymbol{\alpha}_i\|_2 \tag{4}$$

Such a penalty is beneficial when each group of coefficients is believed to be either all zero or all non-zero, and the solution contains only a small number of nonzero coefficient groups, e.g., [7]–[12].

Solving the SMART model subset selection problem with the gLasso leads to the SG estimate:

$$\widehat{\mathbf{a}}_i^{SG} = \arg\min_{\mathbf{a}_i} \frac{1}{n}\|\mathbf{y}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \lambda\sum_{j=1}^{N}\|\mathbf{a}_{i,j}\|_2 \tag{5}$$

where we define:

$$
\begin{aligned}
\mathbf{y}_i &= \begin{bmatrix} x_i(t) & x_i(t-1) & \dots & x_i(t-n+1) \end{bmatrix}^T \\
\mathbf{X}_i &= \begin{bmatrix} x_i(t-1) & \dots & x_i(t-p) \\ x_i(t-2) & \dots & x_i(t-p-1) \\ \vdots & \ddots & \vdots \\ x_i(t-n) & \dots & x_i(t-p-n+1) \end{bmatrix} \\
\mathbf{X} &= \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \dots & \mathbf{X}_N \end{bmatrix} \\
\mathbf{a}_{i,j} &= \begin{bmatrix} a_{i,j}(1) & a_{i,j}(2) & \dots & a_{i,j}(p) \end{bmatrix}^T \\
\mathbf{a}_i &= \begin{bmatrix} \mathbf{a}_{i,1} & \mathbf{a}_{i,2} & \dots & \mathbf{a}_{i,N} \end{bmatrix}^T
\end{aligned}
$$

The SCSG removes the penalty for self-connections, that is, each node's own past values are allowed to predict its current value without a penalty:

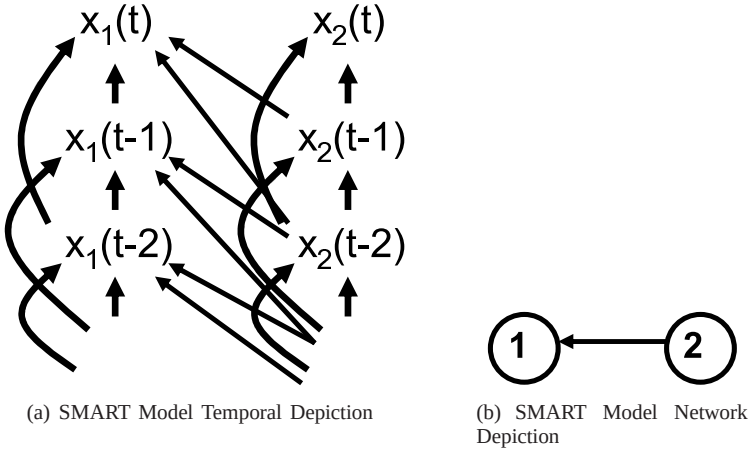
$$\hat{\mathbf{a}}_i^{SCSG} = \arg\min_{\mathbf{a}_i} \frac{1}{n} \|\mathbf{y}_i - \mathbf{X}\mathbf{a}_i\|_2^2 + \lambda \sum_{j \neq i} \|\mathbf{a}_{i,j}\|_2 \tag{6}$$

This represents the expectation of sparse connectivity between nodes.

The gLasso optimization falls into a class of well-studied convex optimization problems. Many algorithms have been proposed for solving this sort of problem (see [16] for a description and comparison of several approaches). Greedy procedures, such as group orthogonal matching pursuit, have been proposed as well [17]. The choice of optimization algorithm is not an important concern in this paper; rather the main contribution of this paper is to characterize the behavior and consistency of the solution of Eqs. (5) and (6).

B. Graphical Model Identification

Lasso-like algorithms have found application in high dimensional graphical model identification. The seminal work in this area was done by Meinshausen and Bühlmann [18] who consider estimating the structure of sparse Gaussian graphical models by identifying the nonzero entries of the inverse covariance matrix. They consider an undirected graph where each vertex represents a variable and edges represent conditional dependence between two variables given all other variables. Conditionally independent variables do not share an edge and correspond to a zero entry in the inverse covariance matrix. Identifying the edge set, or nonzero entries in the inverse covariance matrix, is achieved by writing independent samples of one variable as a sparse, but unknown linear combination of the corresponding samples of the other variables, then using the Lasso. Meinshausen and Bühlmann [18] show that this procedure consistently identifies the edge set even when the number of variables (vertices) grows faster than the number of samples. Ravikumar, et al., [19] propose an alternative Lasso like approach to the same problem by maximizing the ℓ_1 norm penalized log-likelihood function. In this case the first term of Eq. (3) is replaced with an inner product and log-determinant of the covariance matrix. The graphical lasso technique solves this type of problem efficiently for very large problems [20].



(a) SMART Model Temporal Depiction

(b) SMART Model Network Depiction

Fig. 1. Two graphical depictions of a two node, second order SMART model. (a) Explicit time dependence structure. (b) Shorthand depiction of (a) suppressing time and self-connections.

The SMART model is a graphical model involving causal relationships and consequently, an element of time. The resulting model is a directed graph, and each node can be represented by multiple vertices: one for the current value, and potentially

infinitely many for past values at that node as shown in Fig. 1(a). To ease visualization, we suppress time dependence and illustrate causal influence with a single arrow linking one vertex per node as shown in Fig. 1(b). Here we have not shown self-connections. Nodes which have a causal influence are termed "parent nodes" (node 2 in Fig. 1) and the nodes they influence "child nodes" (node 1 in Fig. 1). Given that graphs representing MAR models are directed, the existing analyses by Ravikumar, et al., [19] and Meinshausen and Bühlmann [18] are insufficient. The additional notions of causality and a temporal element place the SMART model in the realm of graphical Granger models [14], [21].

C. Existing Lasso and gLasso Consistency Results

There are many existing results on consistency of the Lasso (e.g., [18], [22]) and extensions of these to the gLasso or closely related problems (e.g., [17], [23]–[30]). An important concept in all these results is mutual incoherence, the maximum absolute inner product between two columns of $\mathbf{X}$. Mutual incoherence is extended to grouped variables by using the maximum singular value of $\mathbf{X}_i^T\mathbf{X}_j$ in place of the vector inner product. Analyzing mutual coherence in the SMART model setting is challenging due to the strong statistical dependence between columns of $\mathbf{X}$. Both Lasso and gLasso have recently been successfully applied to SMART networks (e.g. [13], [14], [31], [32]), but consistency was not considered. In independent work, the consistency of first-order AR models (a special case of the general problem considered here) is investigated in [33]. We identify novel incoherence conditions tailored specifically to the SMART model, and show how the network structure of the model affects these conditions. Thus these incoherence conditions provide unique insight into the capabilities and limitations of SG model identification.

III. Asymptotic Consistency of SMART gLasso

In this section we provide sufficient conditions for the asymptotic consistency of the SG estimate assuming the data are generated by a SMART model. Our general approach is similar to the style of argument used in the analysis of gLasso consistency [30] and other graph inference methods based on sparse regression [18]. An important distinction in SG is the MAR structure of the design matrix $\mathbf{X}$.

Let $\mathcal{S}_i = \{j \in \{1,\dots,N\} : (i,j) \in \mathcal{S}_{\text{active}}\}, i = 1,\dots,N$ indicate the subset of nodes that causally influence node i. Define $\mathbf{X}_{\mathcal{S}_i}$ and $\mathbf{X}_{\mathcal{S}_i^C}$ to be submatrices of $\mathbf{X}$ composed of the matrices $\mathbf{X}_j$, $j \in \mathcal{S}_i$ and $\mathbf{X}_j$, $j \notin \mathcal{S}_i$, respectively. An oracle that knows $\mathcal{S}_i$ does not need to solve the subset selection problem but only a regression problem with design matrix $\mathbf{X}_{\mathcal{S}_i}$ and parameters $\mathbf{a}_{i,j}$, $j \in \mathcal{S}_i$.

Our main result makes use of a regression problem with the same design matrix. Consider a node j with $j \notin \mathcal{S}_i$. The optimal linear predictor of $\mathbf{X}_j$ given $\mathbf{X}_{\mathcal{S}_i}$ is $\sum_{k\in\mathcal{S}_i}\mathbf{X}_k\mathbf{\Psi}_{j,k}$ where the $\mathbf{\Psi}_{j,k}$ minimize $\mathbb{E}[\|\mathbf{X}_j - \sum_{k\in\mathcal{S}_i}\mathbf{X}_k\mathbf{\Psi}_{j,k}\|_F^2]$. If we stack $\{\mathbf{\Psi}_{j,k}\}_{k\in\mathcal{S}_i}$ to form a matrix $\mathbf{\Psi}_{j,\mathcal{S}_i}$, then we can write $\sum_{k\in\mathcal{S}_i}\mathbf{X}_k\mathbf{\Psi}_{j,k} = \mathbf{X}_{\mathcal{S}_i}^T\mathbf{\Psi}_{j,\mathcal{S}_i}$. Using standard matrix calculus it is not difficult to verify that

$$\mathbf{\Psi}_{j,\mathcal{S}_i} = \mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i}^{-1}\mathbb{E}[\mathbf{X}_{\mathcal{S}_i}^T\mathbf{X}_j]$$

where the covariance matrix

$$\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i} = \mathbb{E}[\mathbf{X}_{\mathcal{S}_i}^T\mathbf{X}_{\mathcal{S}_i}].$$

Recall the following variables: N, the number of nodes in the network; m, the maximum number of parent nodes; p, the SMART model order; and n, the number of observations. The main result concerning the consistency of SMART gLasso is

Theorem 1: Let C_{power}, C_{con}, C_{min}, C_{max}, and C_{fcs} be non-negative constants. Assume entries in $\mathbf{y}_i$ and the corresponding row of each $\mathbf{X}_j$ matrix come from independent realizations of the SMART model. Assume the following conditions hold:

1) **Scaling:** N, m, and p are $\mathcal{O}(n^c)$, while λ is $\Theta(n^{-c})$ for different $c > 0$ with $m\lambda^2 = o(1)$ and $\frac{p}{n\lambda^2} = o(1)$.
2) **Signal Power:**
$$\max_{i\in\{1,\dots,N\}}\sigma_i^2 = \mathbb{E}[x_i^2(t)] \le C_{power} < \infty$$
3) **Connection Strength:** $\min_{(i,j)\in\mathcal{S}_{active}}\|\mathbf{a}_{i,j}\|_2 \ge C_{con} > 0$
4) **Minimum Power:** $\max_i\|\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i}^{-1}\|_2 \le C_{min}^{-1} < \infty$
5) **Maximum Cross Correlation:**
$$\max_i\|\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i^C}\|_2 \le C_{max} < \infty$$
where
$$\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i^C} = \mathbb{E}[\mathbf{X}_{\mathcal{S}_i}^T\mathbf{X}_{\mathcal{S}_i^C}]$$
6) **False Connection Score:** For all $(i,j) \in \mathcal{S}_{\text{active}}^C$
$$\psi_{j\to i}^{FC} := \left\|\sum_{k\in\mathcal{S}_i}\mathbf{\Psi}_{j,k}^T\frac{\mathbf{a}_{i,k}}{\|\mathbf{a}_{i,k}\|_2}\right\|_2 \le C_{fcs} < 1 \tag{7}$$

Then for all n sufficiently large, the set of links identified by SG satisfies $\hat{\mathcal{S}} = \mathcal{S}_{\text{active}}$ with probability greater than $1-\exp(-\Theta(n))$; i.e., zero and nonzero links identified by SG agree with those of the underlying true model.

Proof: The proof is presented in Appendix A. ■

Note we have used the following notation: $f(n) = \mathcal{O}(g(n))$ implies $|f(n)| \leq k|g(n)|$ for some $k > 0$ and large n, $f(n) = \Theta(g(n))$ implies $k_1|g(n)| \leq |f(n)| \leq k_2|g(n)|$ for some positive constants k_1 and k_2 and large n, and $f(n) = o(g(n))$ implies $|f(n)| \leq k|g(n)|$ for all $k > 0$ and large n.

Assumption 1 specifies how network parameters grow as a function of the number of observations n. It may be possible to allow some or all of the constants C_{power}, C_{con}, C_{min}, C_{max}, and C_{fcs} to depend on n, but for the purposes of this paper we will take these to be constants. The number of nodes in the network N can grow at any polynomial rate, including both faster or slower than the number of observations n, or remain fixed. Assumptions 2–5 are rather mild. They are used to show that there will be no false negatives for sufficiently small λ. In practice, signals are often normalized to have equal power across nodes, which automatically achieves 2, though only this weaker assumption is necessary here. The effect of normalization on the other assumptions, particularly 6, is an interesting open question. Assumption 4 essentially says that each time sample in the active set contains some independent information. Assumption 5 ensures that any influence due to the nodes in $\mathcal{S}_i$ cannot be easily generated using nodes in $\mathcal{S}_i^c$ instead.

Assumption 6 is the most restrictive and most informative. In the proof of the theorem, Assumption 6 is used to show that the probability of declaring a nonzero connection when none exists (i.e. a false connection or false alarm) goes to zero for large n. In order to understand the implications of the assumption, we point out a more restrictive, but less complicated alternative: $\sum_{k\in\mathcal{S}_i} \|\mathbf{\Psi}_{j,k}\|_2 \leq C_{fcs} < 1$. If this inequality holds, Assumption 6 follows from simple norm bounds. The inequality also suggests the following interpretation of Assumption 6. Nodes that do not directly drive the node of interest (i.e., nodes in $\mathcal{S}_i^C$) cannot be easily predicted from nodes that are directly driving the node of interest. In Section V we provide example networks that do and do not satisfy Assumption 6 to gain insight into the nature of which networks can be recovered. We show next that Assumption 6 is necessary for a large class of networks, including those of fixed size.

Theorem 2: Suppose Assumptions 2–5 of Theorem 1 hold, but $\psi_{j\to i}^{FC} \geq 1 + c$ for some pair (i, j) and constant $c > 0$. Suppose also that $m^2 p < n$ for large n. Then with probability exceeding $1 - \exp(-\Theta(n))$, the connections recovered by SG will not be the true connections.

Proof: A proof is given in Appendix B. ■

Theorem 2 suggests that the false connection score is extremely important in sparse network recovery, especially in finite parameter networks, which are discussed below in Sec. IV-A.

The SCSG (6) assumes that each node is driven by its own past. The conditions of Theorem 1, with minor modification, still govern the ability to recover the correct connectivity pattern using SCSG:

Corollary 1: Suppose Assumptions 1–5 of Theorem 1 hold for all l. In place of Assumption 6, assume:

$$\widetilde{\psi}_{j\to i}^{FC} = \left\| \sum_{k\in\mathcal{S}_i, k\neq i} \mathbf{\Psi}_{j,k}^T \frac{\mathbf{a}_{i,k}}{\|\mathbf{a}_{i,k}\|_2} \right\|_2 \leq C_{fcs} < 1. \tag{8}$$

Then with probability exceeding $1 - \exp(-\Theta(n))$, the connections recovered by SCSG (6) will be the true connections.

Proof: See Appendix C. ■

As we will show in the next section, $\widetilde{\psi}_{j\to i}^{FC}$ is typically lower than $\psi_{j\to i}^{FC}$, though cancellation between the self-connection term and other terms in the sum of (7) is possible.

IV. Network Recovery

In Section III we established conditions which guarantee high probability recovery of SMART networks asymptotically, allowing the network size to grow faster than the number of samples. Next we explore the differences between the asymptotic setting and finite sample regimes.

A. Recovery of Finite Parameter Networks

In practice, the network parameters are typically fixed, and we are interested in performance as the number of measurements n grows. The results of Theorems 1 and 2 still apply. In the finite network case, m, p, and N are fixed, so $(m^2 p)/n$ tends to zero and Assumption 1 is satisfied as long as $\lambda^2 = \mathcal{O}(n^{-c})$ with $0 < c < 1$. Also, Assumptions 2–5 are automatically satisfied as long as there is driving noise in each node. Assumption 6 is the only one that does not necessarily hold. This implies the following corollary, which follows immediately from the proof of Theorem 1.

Corollary 2: For a SMART model with fixed parameters, (5) will recover the correct network structure with probability greater than $1 - \exp(-\Theta(n))$ if $\psi_{j\to i}^{FC} < 1$ for all pairs $(i, j) \in \mathcal{S}_{\text{active}}^C$. If $\psi_{j\to i}^{FC} > 1$ for some $(i, j) \in \mathcal{S}_{\text{active}}^C$, then (5) will fail to recover the correct structure with probability exceeding $1 - \exp(-\Theta(n))$. The same result holds for (6) using $\tilde{\psi}_{j\to i}^{FC}$.

B. Recovery of Known Networks

Given Corollary 2 it is easy to check whether a given SMART model structure can be recovered via (5) or (6). Define $\mathbf{\Gamma}(\tau) = \mathbb{E}[\mathbf{x}(t)\mathbf{x}^T(t-\tau)]$, and recall $\mathbf{\Sigma}$ is the driving noise $\mathbf{u}(t)$ covariance matrix. If we define the collection of MAR coefficients $\mathbf{A}$ and $\tilde{\mathbf{\Sigma}}$ as:

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix}\mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_p \\ & \mathbf{I}_{N(p-1)} & & \mathbf{0}_{N(p-1),N}\end{bmatrix}, \\ \tilde{\mathbf{\Sigma}} &= \begin{bmatrix}\mathbf{\Sigma} & \mathbf{0}_{(p-1)N} \\ \mathbf{0}_{(p-1)N} & \mathbf{0}_{(p-1)N}\end{bmatrix},\end{aligned}$$

then $\mathbf{\Gamma}(\tau)$ can be calculated via (see e.g. [4])

$$\mathbf{\Gamma} = \mathbf{A}\mathbf{\Gamma}\mathbf{A}^T + \tilde{\mathbf{\Sigma}} \tag{9}$$

where

$$\mathbf{\Gamma} = \begin{bmatrix}\mathbf{\Gamma}(0) & \mathbf{\Gamma}(1) & \dots & \mathbf{\Gamma}(p-1) \\ \mathbf{\Gamma}(-1) & \mathbf{\Gamma}(0) & \dots & \mathbf{\Gamma}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Gamma}(1-p) & \mathbf{\Gamma}(2-p) & \dots & \mathbf{\Gamma}(0)\end{bmatrix}.$$

Using properties of Kronecker products, (9) can be solved in closed form:

$$\mathrm{vec}\left(\mathbf{\Gamma}\right) = \left(\mathbf{I} - \mathbf{A}\otimes\mathbf{A}\right)^{-1}\mathrm{vec}\left(\tilde{\mathbf{\Sigma}}\right). \tag{10}$$

Given this closed form expression for $\mathbf{\Gamma}$, matrices $\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i}$ and $\mathbf{R}_{\mathcal{S}_i,\mathcal{S}_i^C}$ are formed for each node i by selecting the appropriate entries from covariance matrix $\mathbf{\Gamma}$ and subsequently used to calculate $\mathbf{\Psi}_{j,\mathcal{S}_i}$. Given $\mathbf{\Psi}_{j,\mathcal{S}_i}$ and $\mathbf{a}_{i,k}$ for all $k \in \mathcal{S}_i$, $\psi^{FC}_{j\to i}$ or $\tilde{\psi}^{FC}_{j\to i}$ can be calculated and compared to one via Eq. (7) or (8), respectively.

C. Challenges in Realistic Networks

The theoretical basis for SMART model recovery relies on independent data samples and asymptotic probability concentration arguments. We now consider consequences of more realistic data sets.

Our analysis focuses on the dependence accross columns of $\mathbf{X}$ and the corresponding entry of $\mathbf{y}_i$ induced by the SMART model. To prove Theorems 1 and 2, we assumed each row of $\mathbf{X}$ and the corresponding entry of $\mathbf{y}_i$ to be independent from other rows. This is not true in realistic networks where each $\mathbf{X}_i$ is actually Toeplitz; however, rows of $\mathbf{X}$ and $\mathbf{y}_i$ decorrelate as the time lag between them grows ($\mathbb{E}[\mathbf{x}(t)\mathbf{x}^T(t-\tau)] \approx \mathbf{0}$). The simulations in Secs. VI and VII use correlated rows and reveal that the false alarm score has a more significant impact on performance than the row dependence. The effect of row dependence has been consider in the special case of first order (p = 1) AR models in [33], which yields a lower bound on the required number of observations.

An additional challenge – and motivation for group sparse approaches – is the limited number of data samples available. Specific connectivity patterns in a SMART model of a real network may change over time, which limits the number of samples for which the network is approximately stationary. Analysis of the performance of (5) or (6) is difficult for limited data cases (finite n); however, the asymptotic theory and the simulations presented in Section VI suggest that when $\psi^{FC}_{j\to i}$ is small, connectivity estimation is easier. Also, weak connections (for which $\|\mathbf{a}_{i,j}\|_2$ is small) are more difficult to recover with limited data. For small enough λ and large enough n, all connections will probably be recovered. When n is limited, the probability of recovering all connections, particularly weak ones, is decreased.

Although Theorem 1 indicates how λ should scale with n, selecting λ for non-asymptotic regimes can be difficult. As seen in Section VI, λ balances missed connections (Type II errors) with false positives (Type I errors). Ideally, one would select λ to achieve a specified famlywise error rate or false discovery rate; however, calculating p-values of each connection for a given λ is an open problem.

Due to the difficulty of selecting an appropriate regularization parameter, it can be beneficial to consider the family of solutions achieved by varying λ. The expectation-maximization (EM) algorithm described in [12] efficiently solves the SG or, with slight modification, SCSG problem over a range of λ, successively adding connections as λ decreases. In that work, a heuristic is used to select a single λ from the family of possible solutions [12]. In Sec. VI we use tenfold cross-validation to select the λ which performs best on held out data. Another possibility is to apply a Wald test for Granger-causality [4] successively to the last connection which enters the model and stop when a connection passes the test. A recently proposed stability selection technique combines lasso and randomized subsampling to provide subset selection with false discovery rate bounds [34]. This technique could potentially be applied to the SMART model at the expense of additional computation.

D. Normalization

Measurements from each node are often normalized to have equal power [18], [35]. We can account for normalization in any SMART model as follows. Equal power in all channels means the diagonal of $\mathbf{\Gamma}$ consists of all ones. Thus we can transform $\mathbf{\Gamma}$ to a normalized model using a diagonal matrix $\mathbf{D}^{-1/2}$ to obtain $\tilde{\mathbf{\Gamma}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{D}^{-\frac{1}{2}}$. Eq. (9) implies:

$$\begin{aligned}\tilde{\mathbf{\Gamma}} &= \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{\frac{1}{2}}\left(\mathbf{D}^{-\frac{1}{2}}\mathbf{\Gamma}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{D}^{\frac{1}{2}}\mathbf{A}^T\mathbf{D}^{-\frac{1}{2}} \\ &\quad +\mathbf{D}^{-\frac{1}{2}}\widetilde{\mathbf{\Sigma}}\mathbf{D}^{-\frac{1}{2}} \\ &= \tilde{\mathbf{A}}\tilde{\mathbf{\Gamma}}\tilde{\mathbf{A}}^T + \tilde{\mathbf{\Sigma}}^*\end{aligned}$$

where:

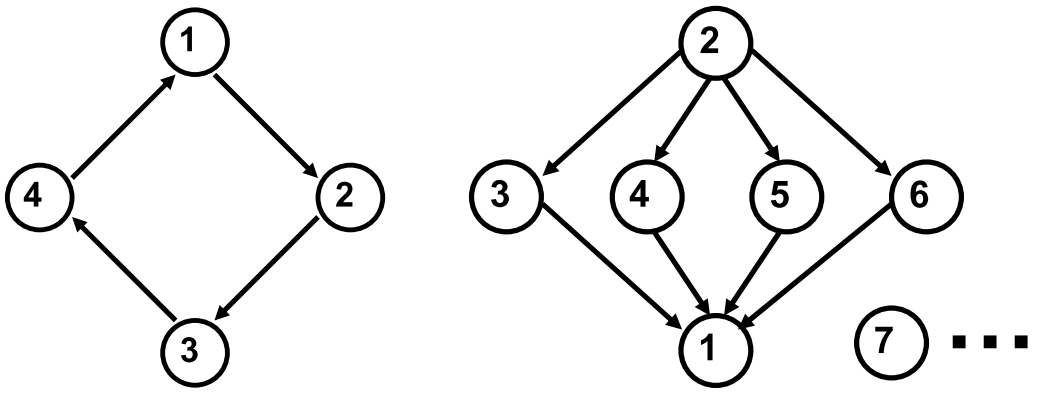
$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{D}_N \end{bmatrix}$$

Here $\mathbf{D}_i = \sigma_i^2\mathbf{I}_p$ where σ_i^2 is the power in each node before normalization.

The effect of normalization on the ability of group sparse approaches to recover network structures is complicated. We have found that normalization tends to decrease $\psi_{max} = \max_{(i,j)\in\mathcal{S}^C_{\text{active}}} \psi_{j\to i}$, indicating an improvement in asymptotic recoverability (for fixed m, p, and N at least). On the other hand, normalization clearly alters connection strength, meaning some connections may be weakened due to normalization and difficult to recover in the finite sample case.

V. Example MAR networks

The false connection scores $\psi^{FC}_{j\to i}$ and $\widetilde{\psi}^{FC}_{j\to i}$ are the key quantities that determine whether SG or SCSG will recover the connections which influence node i. We consider four example networks in this section to develop insight on the nature of identifiable topologies. Figure 2 depicts circular and parallel topologies constructed for this paper while Fig. 3 depicts networks that have been studied in previous literature (see [3], [13][1]). We compute the false connection scores for both the original network and after normalization (Sec. IV-D) to determine whether the network is identifiable as $n \to \infty$ for SG and SCSG. The maximum false connection scores for each network are listed in Table I.



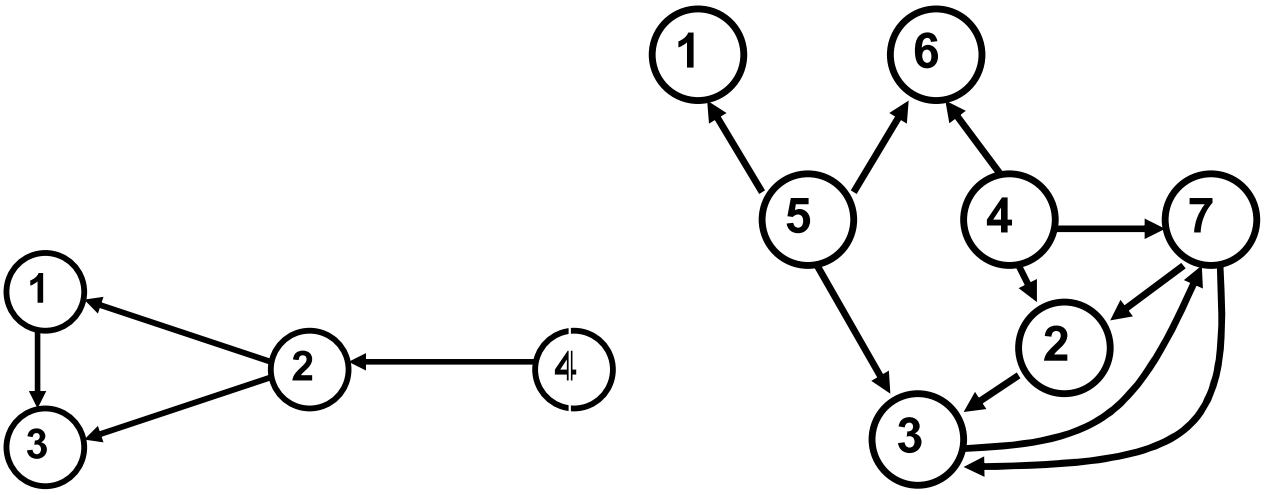
(a) Circle Network (b) Parallel Network

Fig. 2. Contrasting example MAR topologies, self-connections not shown.

Each node in the "Circle Network" shown in Fig. 2(a) is driven by it's own past as well as one other node forming the topology of a large feedback loop. We chose MAR order $p = 4$ and drew MAR coefficients from a normal distribution ($\mathcal{N}(\mathbf{0}, 0.04\mathbf{I})$). The first realization which resulted in a stable network is selected. The maximum false connection scores for this network are $\psi^{FC}_{j\to i} = 0.47$ and $\widetilde{\psi}^{FC}_{j\to i} = 0.43$. Since these are less than one, the network connectivity can be recovered (as $n \to \infty$) using both SG and SCSG.

The parallel network (Fig. 2(b)) connectivity structure and coefficients were selected deliberately to confound group sparse approaches. We chose $\mathbf{a}_{2\to 2} = \begin{bmatrix} .2 & .2 & .2 & .2 \end{bmatrix}^T$ and $\mathbf{a}_{i\to i} = \begin{bmatrix} .05 & .05 & .05 & .05 \end{bmatrix}^T$ for $i \neq 2$. All other connections shown are given by $\mathbf{a}_{i\to j} = \begin{bmatrix} .15 & .15 & .15 & .15 \end{bmatrix}^T$. This network highlights several important aspects of SCSG, so we explore it in some detail. The false connection scores for this network are summarized in Table II.

[1]In [13] the direction of causal influence is unclear. The network structure is described by a matrix of ones and zeros, but it is unclear whether a one in the $(i,j)^{th}$ position represents a connection from i to j or vice versa. We show one possibility here and note that the other possible network (not shown) has similar properties.



(a) Winterhalder Network (b) Haufe Network

Fig. 3. MAR network topologies from existing literature.

TABLE I
MAXIMUM FALSE CONNECTION SCORES.

Network	Original		Normalized	
	ψ^{FC}_{max}	$\tilde{\psi}^{FC}_{max}$	ψ^{FC}_{max}	$\tilde{\psi}^{FC}_{max}$
Circle	0.47	0.43	0.47	0.43
Parallel	1.93	1.06	1.04	1.03
Winterhalder	0.46	0.29	0.24	0.15
Haufe	0.83	0.56	0.71	0.57

No matter which approach is used, a false connection from node 2 to node 1 will be established with high probability as $n \to \infty$. This is due to the fact that there are four parallel paths connecting node 2 to node 1. Since node 2 has such a strong combined influence on node 1, group sparse approaches are likely to identify a direct link. False connections from node 1 to nodes 3–6 are also likely for large n when SG is used. On the other hand, the probability of linking 1 to 3–6 goes to zero as n increases if SCSG is used. This illustrates an important characteristic of SCSG: the asymptotic likelihood of false connections from a child to a parent tends to be reduced when self-connections are not penalized. Proving this is always true seems difficult, but we provide some rationale. The difference between $\psi^{FC}_{j\to i}$ and $\tilde{\psi}^{FC}_{j\to i}$ is the term $\mathbf{\Psi}^T_{j,i}\frac{\mathbf{a}_{i,i}}{\|\mathbf{a}_{i,i}\|_2}$, whose norm lies between the singular values of the square matrix $\mathbf{\Psi}_{j,i}$. While it is difficult to verify that vector $\mathbf{a}_{i,i}$ lines up with a strong left singular vector of $\mathbf{\Psi}_{j,i}$, we can expect that $\mathbf{\Psi}_{j,i}$ will be "large" relative to other $\mathbf{\Psi}_{j,k}$ since there is a connection from i to j.

The false connection score from node 1 to node 2 in Fig. 2(b) highlights another important (and related) feature of SCSG. The probability of falsely identifying connections to any node i which is only influenced by its own past goes to zero as n goes to ∞ since $\tilde{\psi}^{FC}_{j\to i}$ is always zero.

The parallel network example also indicates that additional, unconnected nodes (i.e., node 7) do not change the false

TABLE II
FALSE CONNECTION SCORES FOR PARALLEL NETWORK.

Connection	Original		Normalized	
	$\psi^{FC}_{i\to j}$	$\tilde{\psi}^{FC}_{i\to j}$	$\psi^{FC}_{i\to j}$	$\tilde{\psi}^{FC}_{i\to j}$
$1 \to 2$	1.41	0	0.74	0
$2 \to 1$	1.06	1.06	1.04	1.03
$1 \to 3$ $1 \to 4$ $1 \to 5$ $1 \to 6$	1.93	0.71	1.00	0.37
$3 \to 2$ $4 \to 2$ $5 \to 2$ $6 \to 2$	0.61	0	0.63	0

connection scores of connected nodes. The chance of a false connection will increase in the finite n case, but asymptotically such additional nodes do not matter since, as n grows, the estimated correlation between two unconnected nodes will go to zero.

The network in Fig. 3(a) (see [3]) is not only group sparse, but sparse as well; every connection but one (self-connection of node 4) consists of only one coefficient at one time lag, as shown by:

$$\begin{aligned} x_1(t) &= 0.8x_1(t-1) + 0.65x_2(t-4) + u_1(t) \\ x_2(t) &= 0.6x_2(t-1) + 0.6x_4(t-5) + u_2(t) \\ x_3(t) &= 0.5x_3(t-3) - 0.6x_1(t-1) + 0.4x_2(t-4) \\ &\quad + u_3(t) \\ x_4(t) &= 1.2x_4(t-1) - 0.7x_4(t-2) + u_4(t) \end{aligned}$$

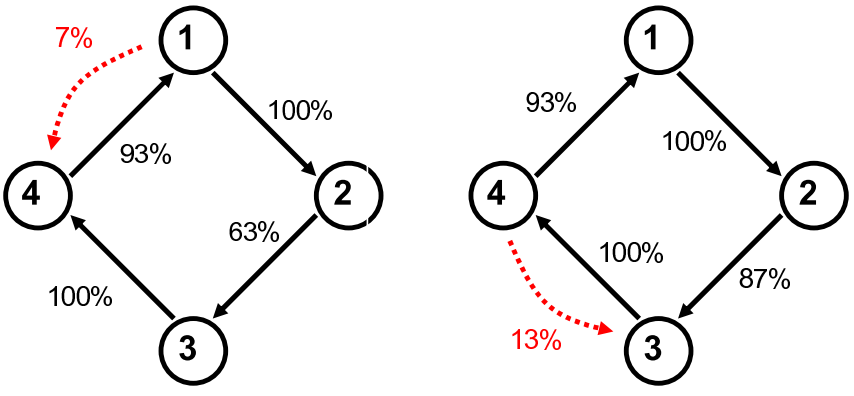
As shown in Table I, this network is recoverable by either method.

The structure of the network shown in Fig. 3(b) is taken from Fig. 1 of [13]. As in [13], we draw coefficients from a $\mathcal{N}(\mathbf{0}, 0.04\mathbf{I})$ distribution and check for stability. This network, which includes multiple paths of influence and feedback loops, can be recovered via both SG and SCSG with high probability as n increases.

VI. Simulations

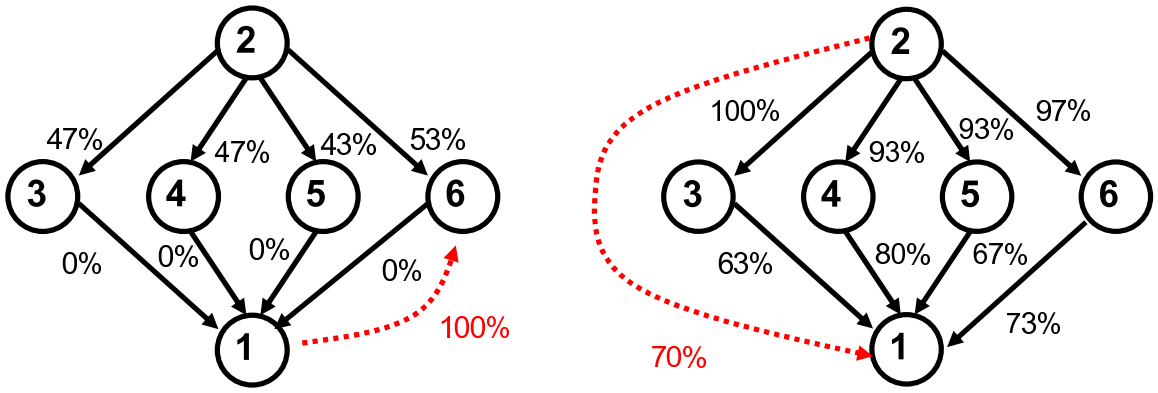
We now simulate the circle and parallel networks depicted in Fig. 2 to illustrate SG and SCSG network recovery performance with finite n. (Simulations of the Haufe and Winterhalder networks performed similarly to the circle network and are omitted for space.) Signals were simulated via (1) with the initial condition for each simulation determined from the steady state distribution and with white driving noise of equal power in each node. The expectation-maximization (EM) algorithm described in [12] is used to solve the SG and SCSG optimization problems for $\lambda \in [0.05\lambda_{max}, \lambda_{max}]$, where λ_{max} is the minimum λ such that $\hat{\mathbf{a}}_i = \mathbf{0}$. A specific λ is selected separately for each node via tenfold cross validation using prediction error on held out data. We assume the correct model order p is known. Thirty realizations of each network are generated with $n = 150$ time samples. We count the percentage of the 30 trials in which the true connections are correctly identified as well as the percentage of trials in which nonexistent connections are incorrectly identified.

The results for SG and SCSG applied to the circle network are illustrated graphically in Fig. 4. The true connections are identified in most of the cases for the circle network. The strength of the four connections are given by $\|\mathbf{a}_{2,1}\|_2 = 0.46$, $\|\mathbf{a}_{3,2}\|_2 = 0.30$, $\|\mathbf{a}_{4,3}\|_2 = 0.37$, and $\|\mathbf{a}_{1,4}\|_2 = 0.28$. The two true connections that are most often missed are the weakest connections of the four ($2 \to 3$ and $4 \to 1$). The SCSG approach identifies the connection from $2 \to 3$ considerably more often, however. The most common false connection with SG was from node 1 to node 4 and occurred in only 2 of 30 trials, while a false connection from node 4 to node 3 was identified in 4 of 30 trials using SCSG. Qualitatively similar results are obtained for $n = 50$ and $n = 100$ with the performance improving for most connections as the number of samples increases. A noticeable improvement in ability to identify true connections results as the number of samples increases from $n = 50$ to $n = 150$.



(a) Circle Network (b) Circle Network

Fig. 4. Inferring the circle network using SG and SCSG with cross validation from $n = 150$ time samples. Black lines and numbers illustrate true connections and the percentage of 30 trials in which they are correctly identified. Red dotted lines and text identify the most common false connection and percentage of occurrence over 30 trials.



(a) Parallel Network (b) Parallel Network

Fig. 5. Inferring the parallel network using SC and SCSG with cross validation from $n = 150$ time samples. Black lines and numbers illustrate true connections and the the percentage of 30 trials in which they are correctly identified. Red dotted lines and text identify the most common false connection and percentage of occurrence over 30 trials.

As predicted by the theoretical arguments of Sec. IV-A, the SG approach does not perform as well on the parallel network (Fig. 5). In particular, the true connections from nodes 3, 4, 5, and 6 to node 1 are never identified, the true connections from node 2 to nodes 3, 4, 5, and 6 are identified about half of the time, and the connection from node 1 to 6 is incorrectly identified in all cases. The next most common false connections (not shown in Fig. 5) are from node 1 to nodes 3–5 with probabilities of 93%, 83%, and 87%, respectively. These four false connections (from node 1 to its parents) have the highest false connection score ($\psi^{FC}_{1\to j} = 1.93$, $j = 3, 4, 5, 6$) for this scenario, according to Table II. The false connection from node 1 to node 2 is the next most common, occurring in 80% of the trials. The false connection score for this link is 1.41. Notice these five most common false connections reverse the true direction of causal influence.

The SCSG approach performs considerably better for the parallel network, consistent with the improvement in the false connection scores given in Table II. The connections from node 2 to nodes 3–6 are almost always discovered, although the true connections from nodes 3–6 to node 1 are missed more frequently. However, SCSG identifies a connection directly from node 2 to node 1 in 70% of the trials. A possible explanation for this error is that a single connection from node 2 to node 1 is a sparser solution than connecting nodes 3–6 to node 1 and accounts for much of the variance at node 1. The connection from node 2 to node 1 has the highest false connection score (see Table II).

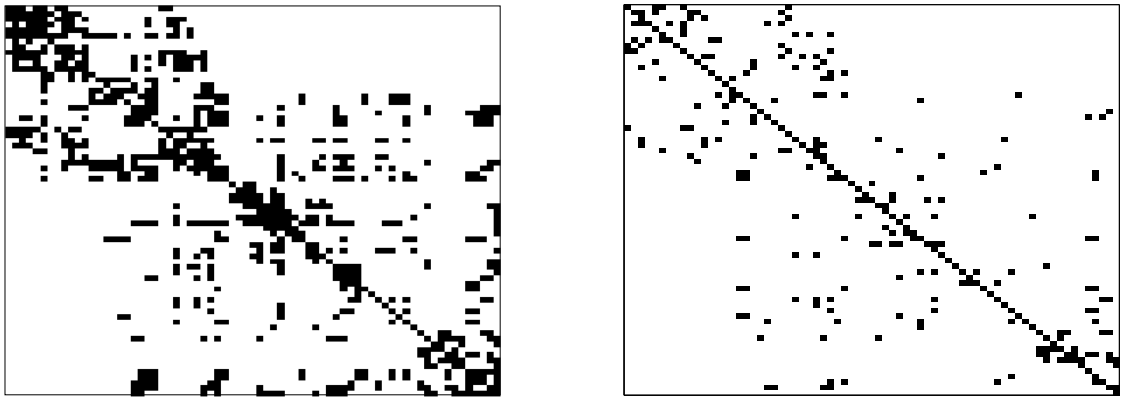
When using SG on the parallel network, none of the true connections to node 1 are identified. While these connections might be recovered by allowing a greater range of λ in the cross validation selection procedure, their absence reveals a downside to penalizing self-connections. As λ is decreased below λ^*, the first connection identified is the self-connection. When SCSG is used, self-connections are always present, so decreasing λ below λ^* activates a connection from a different node. In a sense, the SCSG approach has a "head start" in detecting connections.

Simulations with $n = 50$ and $n = 100$ time samples (not shown) reveal that the ability of SCSG to recover the true connections improves as the number of samples increases. However, the number of trials in which false connections were made between nodes 1 and 2 (both directions) also increases as the number of samples increases. This behavior is consistent with the asymptotic result of Cor. 2 which indicates that the probability of identifying the wrong network goes to one as the number of samples increases.

VII. Macaque Brain Simulation

Lasso-type procedures have recently been applied to MAR model estimation of brain activity [13], [31], [36], [37]. In this section we simulate electrocorticogram (ECoG) recordings with a SMART model using a realistic network topology obtained from tract-tracing studies of a macaque brain [38], [39]. A matrix representing connectivity in the macaque brain – the "macaque71" data set, consisting of 71 nodes and 746 connections – is shown in Fig. 6(a). Each node is an area of the cortex. A connection between areas exists if neuronal axons physically connect respective areas. Figure 6(a) suggests a sparse connectivity structure in the macaque. Including self-connections, there are an average of 11.5 out of 71 possible parents for each node.

We simulate two networks based on this physical connectivity structure. First we assume that every physical connection in the macaque71 data set is actively conveying information. It is unrealistic to model every physical connection as active at a given time, so we also simulate a model in which up to ten randomly selected parents (including the self-connection) are active for each node. For simulation purposes, we choose a model order of six and draw coefficients for nonzero entries of the $\mathbf{A}_i$ matrices independently from a $\mathcal{N}(\mathbf{0}, 0.04\mathbf{I})$ distribution for the full model and a $\mathcal{N}(\mathbf{0}, 0.16\mathbf{I})$ model for the subset model. The first realization for each model that results in a stable network is used.



(a) Full Network (b) Subset Network

Fig. 6. Connectivity matrices of the simulated macaque brain networks: (a) all physical connections are active and (b) up to ten parent nodes are active. A connection from node i to node j exists if the entry in the i^{th} row and j^{th} column is black.

Given these stable SMART models based on physical connections in the macaque brain, we generate time series using Eq. 1 with initial conditions $x_i(0) = 0$ and driving noise $u_i(t)$ distributed i.i.d. $\mathcal{N}(0, 1)$ over all channels and all time samples. The data are normalized, as described in Sec. IV-D using the estimated power at each node.

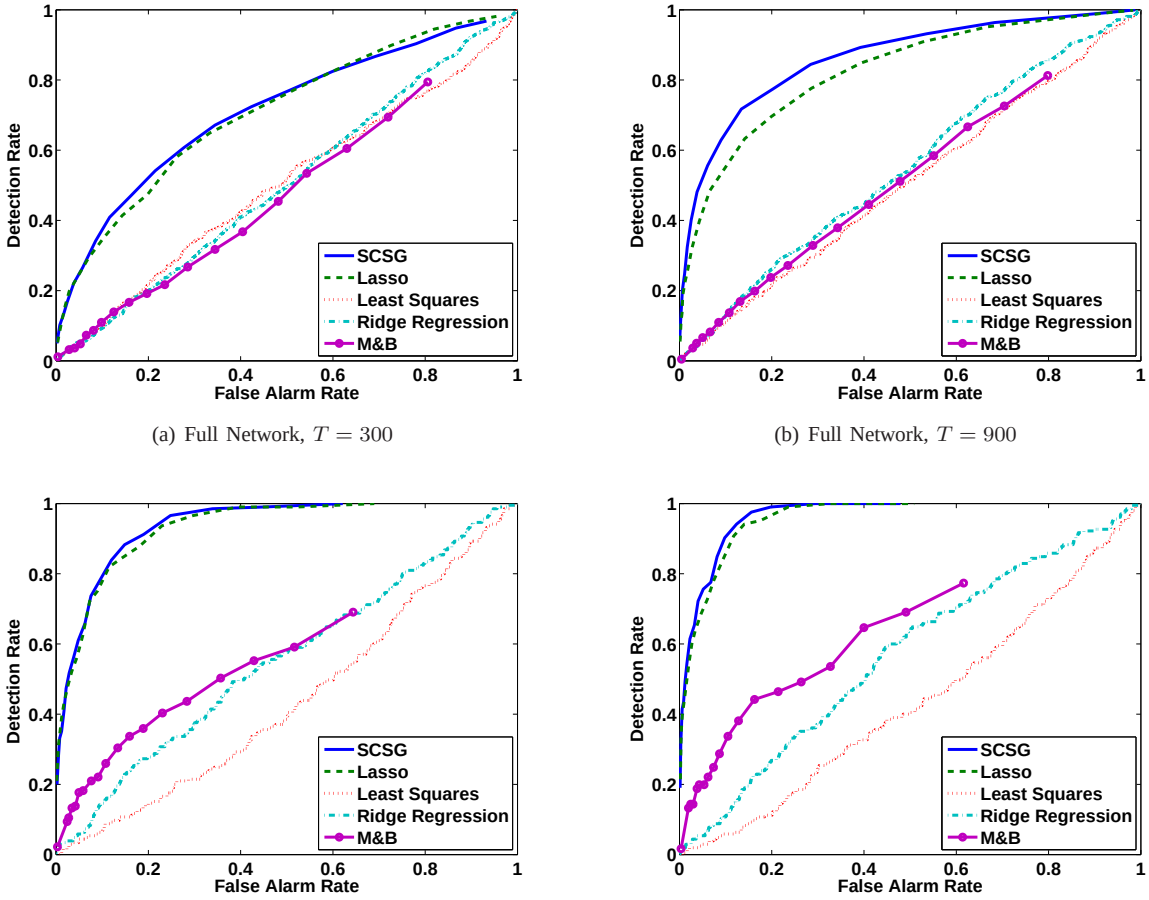
Normalization reduces the worst case SCSG false connection score of the full network from 1.73 to 1.25. Hence the SCSG estimate will be inconsistent as the number of samples increases. Note however, that SCSG can still consistently recover the parents of nodes i for which $\psi^{FC}_{j\to i} < 1$ for all $j \in \mathcal{S}^C_i$. In this example, only four nodes i have $\psi^{FC}_{j\to i} > 1$, meaning that the parents of 67 of the nodes can be recovered accurately. Interestingly the neighborhoods of the four nodes which violate the false connection score condition exhibit a topology very similar to the parallel network described in Sec. V. Each of these four nodes has many parent nodes which provide an indirect link to the same "grandparent" node. If only some of these paths are active at a given time, the network may be recoverable. This is indeed the case in the subset model where the false connection score is reduced from 4.07 to 0.54 by normalization.

We illustrate the performance of several network estimation techniques in Fig. 7 using receiver operating characteristic (ROC) curves. We simulate the SCSG approach, the standard Lasso which promotes sparse coefficients as opposed to sparse connections (see Sec. V), least squares estimation (Yule-Walker equations for $n > pN$), ridge regression, and an approach for estimating sparse non-causal networks described in [18] which we call the Meinshausen and Bühlmann (M&B) approach. The poor performance of the M&B approach illustrates that it is not appropriate for causal network inference[2]. The performance of the SG technique is similar to that of the SCSG for these networks, so we do not include it here.

Using ROC curves to evaluate performance removes the difficult task of selecting regularization parameters (which relate, sometimes directly, to significance level) for different techniques. The ROC curve is obtained for the SCSG, Lasso, and M&B approaches by varying the penalty weight λ (using the same solver with group size of one when necessary). A detection occurs when a nonzero estimate $\hat{\mathbf{a}}_{i,j}$ coincides with a true connection from node j to node i, while a miss occurs when $\hat{\mathbf{a}}_{i,j} = \mathbf{0}$ despite a true connection from j to i. False postives and true negatives are similarly defined. For least squares and ridge regression approaches, we use the simultaneous inference method proposed in [13] which makes use of adjusted p-values [40]; however, we threshold the normalized test statistics directly (rather than the p-values) to produce ROC curves in order to avoid compuationally intensive Monte Carlo sampling of multivariate integrals. This yields the same curve due to the monotonic ralationship between test statistic and associated p-value. Since SCSG has additional knowledge that all self-connections are non-zero, we do not include self-connections when calculating ROC curves for any method. The ROC is defined as the percentage of true connections detected versus the percentage of false positive connections.

We simulate both $n = 300$ and $n = 900$ time samples from all 71 nodes. In the first case we have fewer samples ($300 \times 71 = 21300$) than coefficients ($6 \times 71^2 = 30246$), so enforcing a sparse solution is essential. This is clearly seen in Figs. 7(a) and 7(c) where SCSG and Lasso clearly outperform the other methods. In fact, least squares, ridge regression, and the Meinshausen and Bühlmann approach perform similarly to coin flipping. The SCSG performs better than the Lasso because the group assumption of the gLasso better matches the true model. In the second case with $n = 900$ time samples for each node, we have a few more than two samples for every coefficient. The results are shown in Figs. 7(b) and 7(d). Both SCSG

[2]Readers familiar with [18] will observe that the M&B technique is not meant to recover nonzero connections as defined here, but rather nonzero entries in the inverse covariance matrix. We point out that although the MAR networks presented here are sparse in the number of parent nodes, the inverse covariance matrices are not sparse.



(a) Full Network, $T = 300$

(b) Full Network, $T = 900$

(c) Subset Network, $T = 300$

(d) Subset Network, $T = 900$

Fig. 7. Fraction of connections identified vs. fraction of zero valued $\mathbf{a}_{i,j}$ misidentified as nonzero (ROC curve) in simulated macaque brain networks. Top row: all physical connections active. Bottom row: subset of connections active.

and Lasso perform better with more samples, as expected. The other methods still perform similarly to coin flipping. In the case of least squares and ridge regression, there are still too few samples to reliably estimate the covariance matrices.

VIII. Conclusion

We have analyzed application of the Group Lasso to the SMART model and proposed a modified gLasso for SMART model estimation. The gLasso groups together all p coefficients which comprise a connection from one node to another and penalizes the sum of the ℓ_2 norm of these coefficient groups. Such an approach tends to yield estimated networks with only a few nonzero connections. Our proposed SCSG removes the penalty for self-connections so that a node's own past is always used to predict its next state. We have shown that both the SG and SCSG approaches are capable of recovering the true network structure under certain conditions, the most crucial of which we term the false connection score, ψ_{max}. MAR networks are identifiable when $\psi_{max} < 1$, but not when $\psi_{max} > 1$. To our knowledge, this is the first attempt to quantify the characteristics of MAR networks that result in gLasso based recovery.

The false connection score condition (and to some degree Assumption 4) implies that the network under study must be not only sparse, but also have the property that each node in the network is independent enough from other nodes (then $\mathbf{\Psi}_{i,j}$ will be small). Clearly, a network with only self-connections satisfies this condition, but these are not very interesting or realistic. On the other hand, small world networks [41] have the type of structure that seems likely to meet the false connection condition (again depending on the connection coefficients). In small world networks, each node is connected to most of its nearest neighbors, but also has a few long range connections (short path lengths). It has been shown that such networks efficiently transmit information to all nodes [41], [42] and suggested that the brain may have a small-world network structure. In fact, the structural connectivity pattern of the macaque brain used for simulations in Sec. VI represents a small-world network [39], [43]. Small-world networks have sparse structure, though each node may have a somewhat large number of local connections.

The false connection score indicates whether a false positive connection is likely to occur. False negatives or missed connections are also of concern. Our analysis shows that, for fixed parameter networks (m, p, and N constant), the penalty

weight λ can be set small enough that false negatives are improbable. The false connection score determines whether this small λ will avoid false positives. Our experience suggests that misses are more likely to occur for weak connections. Our examples indicate that the SCSG approach is effective at recovering network structure and that the false connection score is a an informative indicator of recovery performance for even relatively small sample sizes n. Finally, note that the result of Theorems 1 and 2 apply to any gLasso application which satisfy the assumptions. In a generic application the false connection score may be interpreted as a statistical property of the $\mathbf{X}$ matrix.

APPENDIX A
PROOF OF ASYMPTOTIC CONSISTENCY

To prove Theorem 1, we consider applying gLasso (5) to a single node (without loss of generality, node 1), and use the union bound to achieve the desired result. We restate Assumption 1 in terms of positive constants $c_1 - c_4$ to facilitate the proof: number of nodes $N = \mathcal{O}(n^{c_1})$, maximum number of parent nodes $m = \mathcal{O}(n^{c_2})$, model order $p = \mathcal{O}(n^{c_3})$, and regularization parameter $\lambda = \Theta(n^{-c_4/2})$ with $c_2 < c_4$ and $c_3 + c_4 < 1$.

KKT Conditions

The Karush-Kuhn-Tucker (KKT) conditions for a solution to (5) follow from the theory of subgradients. The subgradient of $\|\mathbf{v}\|_2$ is any vector whose ℓ_2-norm is less than one for $\mathbf{v} = \mathbf{0}$, while it is simply the gradient $\frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ when $\mathbf{v} \neq \mathbf{0}$. Thus the KKT conditions are given by:

$$\mathbf{X}_i^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1) = \frac{\lambda n \hat{\mathbf{a}}_{1,i}}{2\|\hat{\mathbf{a}}_{1,i}\|_2} \qquad \forall\ i \text{ s.t. } \hat{\mathbf{a}}_{1,i} \neq \mathbf{0} \tag{11}$$

$$\|\mathbf{X}_i^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1)\|_2 \leq \frac{\lambda n}{2} \qquad \forall\ i \text{ s.t. } \hat{\mathbf{a}}_{1,i} = \mathbf{0}. \tag{12}$$

For convenience, we define $\hat{\mathbf{z}}_1 = [\,\hat{\mathbf{z}}_{1,1}^T \cdots \hat{\mathbf{z}}_{1,N}^T\,]^T$ with $\hat{\mathbf{z}}_{1,i} = \frac{2}{\lambda n}\mathbf{X}_i^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1)$. The vector $\hat{\mathbf{z}}_1$ restricted to the active set is denoted $\hat{\mathbf{z}}_{\mathcal{S}_1}$. We assume without loss of generality that $\mathcal{S}_1 = \{1, 2, \ldots, m\}$.

Limiting False Negatives

We start with conditions assuring that all nonzero coefficients are estimated as nonzero. To do so, we follow the arguments used by [28]. We consider the "oracle" solution; e.g., we consider the solution to the group sparse penalized estimator if the active set were known:

$$\begin{aligned}
\hat{\mathbf{a}}_1^*(\lambda) &= \arg \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}_{\mathcal{S}_1^C}=0} \frac{1}{n}\|\mathbf{y}_1 - \mathbf{X}\boldsymbol{\alpha}_1\|^2 + \frac{\lambda}{2}\sum_{i=1}^{N}\|\boldsymbol{\alpha}_{1,i}\|_2 \\
&= \arg\min_{\boldsymbol{\alpha}_{\mathcal{S}_1}} \frac{1}{n}\left\|\mathbf{y}_1 - \begin{bmatrix}\mathbf{X}_{\mathcal{S}_1} & \mathbf{X}_{\mathcal{S}_1^C}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}_{\mathcal{S}_1} \\ \mathbf{0}\end{bmatrix}\right\|^2 \qquad (13) \\
&+ \frac{\lambda}{2}\sum_{i\in\mathcal{S}_1}\|\boldsymbol{\alpha}_{1,i}\|_2. \qquad (14)
\end{aligned}$$

We must ensure that all coefficient subvectors in $\mathcal{S}_1$ are nonzero in the oracle estimate $\hat{\mathbf{a}}_1^*$. Since all subvectors $\hat{\mathbf{a}}_{1,i}^*$ of $\hat{\mathbf{a}}_1^*$ will be zero for large enough λ, this means we must make sure that λ is not too big.

All nonzero blocks must satisfy (11), so we consider:

$$\begin{aligned}
\frac{\lambda n}{2}\hat{\mathbf{z}}_{\mathcal{S}_1} &= \mathbf{X}_{\mathcal{S}_1}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}^*) \\
&= \mathbf{X}_{\mathcal{S}_1}^T(\mathbf{X}_{\mathcal{S}_1}\mathbf{a}_{\mathcal{S}_1} + \mathbf{u}_1 - \mathbf{X}_{\mathcal{S}_1}\hat{\mathbf{a}}_{\mathcal{S}_1}^*) \\
&= (\mathbf{X}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1}(\mathbf{a}_{\mathcal{S}_1} - \hat{\mathbf{a}}_{\mathcal{S}_1}^*) + \mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1 \qquad (15)
\end{aligned}$$

from which we obtain:

$$\hat{\mathbf{a}}_{\mathcal{S}_1}^* = \mathbf{a}_{\mathcal{S}_1} - \frac{\lambda n}{2}(\mathbf{X}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1})^{-1}\hat{\mathbf{z}}_{\mathcal{S}_1} + (\mathbf{X}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1})^{-1}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1 \tag{16}$$

where the invertibility of $\mathbf{X}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1}$ is assured for large n since n grows faster than mp by Assumption 1. At this point the following notation is convenient. Let $\widehat{\mathbf{G}}_{\mathcal{S}_1} = n(\mathbf{X}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1})^{-1}$, with columns partitioned as $\widehat{\mathbf{G}}_{\mathcal{S}_1} = [\,\widehat{\mathbf{G}}_{\mathcal{S}_{1,1}} \cdots \widehat{\mathbf{G}}_{\mathcal{S}_{1,m}}\,]$, where each

sub-matrix is mp x p. Since $n^{-1}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1}$ is an empirical covariance matrix (maximum likelihood estimate of $\mathbf{R}_{\mathcal{S}_1,\mathcal{S}_1}$), we denote the true inverse covariance matrix of signals from the active set by $\mathbf{G}_{\mathcal{S}_1} = \mathbf{R}_{\mathcal{S}_1,\mathcal{S}_1}^{-1} = [\,\mathbf{G}_{\mathcal{S}_{1,1}} \cdots \mathbf{G}_{\mathcal{S}_{1,m}}\,]$.

To show that each subvector $\hat{\mathbf{a}}_{1,i}^* \neq 0$ for $i \in \mathcal{S}_1$ in the limit, it suffices to show that $\|\widehat{\mathbf{G}}_{\mathcal{S}_{1,i}}^T(\frac{\lambda}{2}\hat{\mathbf{z}}_{\mathcal{S}_1} - \frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1)\|_2 < C_{con} \leq \|\mathbf{a}_{\mathcal{S}_{1,i}}\|_2$. Applying the triangle inequality, we instead show that $\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T(\frac{\lambda}{2}\hat{\mathbf{z}}_{\mathcal{S}_1} - \frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1)\|_2 < C_{con}$ with the following lemma.

Lemma 1: Given Assumptions 1–5, $\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T(\frac{\lambda}{2}\hat{\mathbf{z}}_{\mathcal{S}_1} - \frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1)\|_2 = \mathcal{O}(\max(n^{\frac{c_2-c_4}{2}}, n^{c_2+\frac{c_3-c_4-1}{2}}, n^{-\frac{1}{2}}))$ with probability exceeding $1-\exp(-\Theta(n))$.

Proof: Using $\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T(\frac{\lambda}{2}\hat{\mathbf{z}}_{\mathcal{S}_1} - \frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1)\|_2 \leq \frac{\lambda}{2}\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T\hat{\mathbf{z}}_{\mathcal{S}_1}\|_2 + \frac{1}{n}\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1\|_2$, we bound the two terms separately. First:

$$\begin{aligned}
\frac{\lambda}{2}\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T\hat{\mathbf{z}}_{\mathcal{S}_1}\|_2 &\leq \frac{\lambda}{2}\|\widehat{\mathbf{G}}_{\mathcal{S}_1}\|_2\|\hat{\mathbf{z}}_{\mathcal{S}_1}\|_2 \\
&\leq \frac{\lambda\sqrt{m}}{2}\|\widehat{\mathbf{G}}_{\mathcal{S}_1}\|_2 \\
&\leq \frac{\lambda\sqrt{m}}{2}\left(\|\mathbf{G}_{\mathcal{S}_1}\|_2 + \|\widehat{\mathbf{G}}_{\mathcal{S}_1} - \mathbf{G}_{\mathcal{S}_1}\|_2\right) \\
&\leq \frac{\lambda\sqrt{m}}{2}\left(C_{min}^{-1} + \|\mathbf{G}_{\mathcal{S}_1}\|_2\left\|\left(\left(\frac{\mathbf{W}^T\mathbf{W}}{n}\right)^{-1} - \mathbf{I}\right)\right\|_2\right) \\
&< \frac{\lambda\sqrt{m}}{2}\left(C_{min}^{-1} + \mathcal{O}\left(\sqrt{\frac{mp}{n}}\right)\right) \\
&< \mathcal{O}(n^{(c_2-c_4)/2}) + \mathcal{O}(n^{c_2-c4/2+c_3/2-1/2})
\end{aligned}$$

where $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The second inequality is simply the triangle inequality applied to $\hat{\mathbf{z}}_{\mathcal{S}_1}$ since $\|\hat{\mathbf{z}}_{1,i}\|_2 = 1$ for all $i \in \mathcal{S}_1$ and each node has no more than m parents. The second to last inequality holds with probability greater than $1-\exp(-\Theta(n))$ [28]. Given the conditions on constants c_2–c_4, the last line goes to zero.

Next, consider:

$$\begin{aligned}
\frac{1}{n}\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1\|_2 &= \|(\mathbf{X}_{\mathcal{S}_1}^T\mathbf{X}_{\mathcal{S}_1})^{-1}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1\|_2 \\
&= \|(\mathbf{X}_{\mathcal{S}_1}^+)^T\mathbf{u}_1\|_2 \\
&\leq \frac{\sigma_1^2}{n}\left\|\widehat{\mathbf{G}}_{\mathcal{S}_1}\right\|_2^{1/2}\|\mathbf{u}_1\|_2 \qquad (17)
\end{aligned}$$

where $\mathbf{X}_{\mathcal{S}_1}^+$ denotes the pseudoinverse and $\mathbf{u}_1 \sim \mathcal{N}(\mathbf{0}, \sigma_1^2\mathbf{I})$ since we have assumed independent time samples. Inequality (17) can be easily seen by considering the singular value decomposition of $\mathbf{X}_{\mathcal{S}_1}$. Obozinski et al. [28] provide the following bound for the inverse sample covariance matrix:

$$\mathbb{P}\left(\|\widehat{\mathbf{G}}_{\mathcal{S}_1}\|_2 \leq 2C_{min}^{-1}\right) \geq 1 - 2\exp(-\Theta(n))$$

and [44] provide a bound for the chi-square variate:

$$\mathbb{P}\left(\|\tilde{\mathbf{u}}_1\|_2^2 - n \geq 2\sqrt{nt} + 2t\right) \leq \exp(-t)$$

which holds for any $t > 0$. In particular, $\|\tilde{\mathbf{u}}_1\|_2^2 < 5n$ for $t = n$ with probability exceeding $1-\exp(-n)$. Combining these bounds with (17) and Assumption 2 gives us:

$$\frac{1}{n}\|\widehat{\mathbf{G}}_{\mathcal{S}_{1,i}}^T\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1\|_2 < \frac{C_{power}}{\sqrt{n}}\sqrt{10C_{min}^{-1}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

with probability greater than $1-2\exp(-\Theta(n))$.

■

Since both terms of $\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T(\frac{\lambda}{2}\hat{\mathbf{z}}_1 - \frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1)\|_2$ go to zero as n grows, their sum will be less than C_{con} with high probability for large n. This implies that each $\|\widehat{\mathbf{G}}_{\mathcal{S}_{1,i}}^T(\frac{\lambda}{2}\hat{\mathbf{z}}_1 - \frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1)\|_2$, $i \in \mathcal{S}_1$ will also be less than C_{con}, so for all $i \in \mathcal{S}_1$, $\|\hat{\mathbf{a}}_{1,i}^*\|_2 > \|\mathbf{a}_{1,i}\|_2 - C_{con} \geq 0$.

We have shown that $\hat{\mathbf{a}}_{1,i}^* \neq 0$ for each $i \in \mathcal{S}_1$ with probability greater than $1-\exp(-\Theta(n))$. We next show that the oracle solution is in fact the overall solution with high probability.

Limiting False Positives

Assuming that the oracle solution from (15) has all nonzero subvectors $\hat{\mathbf{a}}^*_{1,i}$, we must ensure that $\hat{\mathbf{a}}^* = \left[\, (\hat{\mathbf{a}}^*_{\mathcal{S}_1})^T \; \mathbf{0}^T \,\right]^T$ is a solution to the full problem with high probability. In other words, we must show that $\frac{2}{\lambda n}\|\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}^*)\|_2 \leq 1$ for all $j \in \mathcal{S}_1^C$. To do so, we adopt a technique used in [18]. Write $\mathbf{X}_j = \sum_{i\in\mathcal{S}_1} \mathbf{X}_i \mathbf{\Psi}_{j,i} + \mathbf{V}_j$, where

$$\mathbf{\Psi}_{j,\mathcal{S}_1} = \begin{bmatrix} \mathbf{\Psi}_{j,1} \\ \vdots \\ \mathbf{\Psi}_{j,m} \end{bmatrix} = \arg\min \mathbb{E}\left[\left\| \mathbf{X}_j - \sum_{i\in\mathcal{S}_1} \mathbf{X}_i \mathbf{\Psi}_{j,i} \right\|_F^2 \right], \tag{18}$$

and $\mathbf{V}_j$ is a random variable representing the portion of $\mathbf{X}_j$ that can't be predicted by $\mathbf{X}_i$, $i \in \mathcal{S}_1$. Now we have:

$$\begin{aligned}
&\frac{2}{\lambda n}\left\|\mathbf{X}_j^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1^*)\right\|_2 && (19)\\
&= \frac{2}{\lambda n}\left\| \left(\sum_{i\in\mathcal{S}_1} \mathbf{X}_i\mathbf{\Psi}_{j,i} + \mathbf{V}_j \right)^T (\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1^*) \right\|_2 \\
&= \frac{2}{\lambda n}\left\| \sum_{i\in\mathcal{S}_1} \mathbf{\Psi}_{j,i}^T\mathbf{X}_i^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1^*) + \mathbf{V}_j^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1^*) \right\|_2 \\
&= \left\| \sum_{i\in\mathcal{S}_i} \mathbf{\Psi}_{j,i}^T \frac{\hat{\mathbf{a}}^*_{1,i}}{\|\hat{\mathbf{a}}^*_{1,i}\|_2} + \frac{2}{\lambda n}\mathbf{V}_j^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1^*) \right\|_2 && (20)\\
&\leq \left\| \sum_{i\in\mathcal{S}_i} \mathbf{\Psi}_{j,i}^T \left(\frac{\hat{\mathbf{a}}^*_{1,i}}{\|\hat{\mathbf{a}}^*_{1,i}\|_2} - \frac{\mathbf{a}^*_{1,i}}{\|\mathbf{a}^*_{1,i}\|_2} \right) \right\|_2 \\
&+ \left\| \sum_{i\in\mathcal{S}_1} \mathbf{\Psi}_{j,i}^T \frac{\mathbf{a}^*_{1,i}}{\|\mathbf{a}^*_{1,i}\|_2} \right\|_2 + \frac{2}{\lambda n}\left\|\mathbf{V}_j^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1^*)\right\|_2 && (21)
\end{aligned}$$

where (20) follows from the KKT condition (11). The second term of (21) is less than one by Assumption 6. We bound the remaining terms separately. In order to bound the first term, we use the following lemma:

Lemma 2: $\left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} - \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\| < \frac{2\|\mathbf{v}-\mathbf{w}\|}{\|\mathbf{w}\|}$

Proof:

$$\begin{aligned}
\left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} - \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\| &\leq \left\| \frac{\mathbf{v}}{\|\mathbf{v}\|} - \frac{\mathbf{v}}{\|\mathbf{w}\|} \right\| + \left\| \frac{\mathbf{v}}{\|\mathbf{w}\|} - \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\| \\
&= \|\mathbf{v}\| \left| \frac{1}{\|\mathbf{v}\|} - \frac{1}{\|\mathbf{w}\|} \right| + \frac{\|\mathbf{v}-\mathbf{w}\|}{\|\mathbf{w}\|} \\
&\leq \frac{|(\|\mathbf{w}\| - \|\mathbf{v}\|)|}{\|\mathbf{w}\|} + \frac{\|\mathbf{v}-\mathbf{w}\|}{\|\mathbf{w}\|} \\
&\leq \frac{2\|\mathbf{v}-\mathbf{w}\|}{\|\mathbf{w}\|}
\end{aligned}$$

■

We now bound the first term of (21):

$$\begin{aligned}
&\left\| \sum_{i\in\mathcal{S}_1} \mathbf{\Psi}_{j,i}^T \left(\frac{\hat{\mathbf{a}}^*_{1,i}}{\|\hat{\mathbf{a}}^*_{1,i}\|_2} - \frac{\mathbf{a}_{1,i}}{\|\mathbf{a}_{1,i}\|_2} \right) \right\|_2 \\
&\leq \|\mathbf{\Psi}_{j,\mathcal{S}_1}\|_2 \left(\sum_{i\in\mathcal{S}_1} \left\| \frac{\hat{\mathbf{a}}^*_{1,i}}{\|\hat{\mathbf{a}}^*_{1,i}\|_2} - \frac{\mathbf{a}_{1,i}}{\|\mathbf{a}_{1,i}\|_2} \right\|_2^2 \right)^{1/2} \\
&\leq \|\mathbf{\Psi}_{j,\mathcal{S}_1}\|_2 \left(\sum_{i\in\mathcal{S}_1} \frac{2\left\|\hat{\mathbf{a}}^*_{1,i} - \mathbf{a}_{1,i}\right\|_2^2}{\|\mathbf{a}_{1,i}\|_2^2} \right)^{1/2}
\end{aligned}$$

where we have applied Lemma 2. From Assumption 3 we have $\|\mathbf{a}_{1,i}\|_2 \geq C_{con}$. Using this and $\|\mathbf{\Psi}_{j,\mathcal{S}_1}\|_2 = \|\mathbf{R}^{-1}_{\mathcal{S}_1,\mathcal{S}_1}\mathbb{E}[\mathbf{X}^T_{\mathcal{S}_1}\mathbf{X}_j]\|_2 \leq \|\mathbf{R}^{-1}_{\mathcal{S}_1,\mathcal{S}_1}\mathbf{R}_{\mathcal{S}_1,\mathcal{S}_1^C}\|_2 \leq C_{max}C^{-1}_{min}$, we have:

$$\begin{aligned}
&\left\|\sum_{i\in\mathcal{S}_1}\mathbf{\Psi}^T_{j,i}\left(\frac{\hat{\mathbf{a}}^*_{1,i}}{\|\hat{\mathbf{a}}^*_{1,i}\|_2}-\frac{\mathbf{a}_{1,i}}{\|\mathbf{a}_{1,i}\|_2}\right)\right\|_2 \\
&\leq \sqrt{2}\|\mathbf{\Psi}_j\|_2 C^{-1}_{con}\left(\sum_{i\in\mathcal{S}_1}\left\|\hat{\mathbf{a}}^*_{1,i}-\mathbf{a}_{1,i}\right\|_2^2\right)^{1/2} \\
&\leq \sqrt{2}\frac{C_{max}}{C_{min}C_{con}}\left\|\hat{\mathbf{a}}^*_{\mathcal{S}_1}-\mathbf{a}_{\mathcal{S}_1}\right\|_2 \\
&= \mathcal{O}(\max(n^{\frac{c_2-c_4}{2}}, n^{c_2+\frac{c_3-c_4-1}{2}}, n^{-\frac{1}{2}}))
\end{aligned}$$

where the last inequality follows from (16) and Lemma 1.

Finally, we show that the last term of (21) goes to zero faster than $\mathcal{O}(n^{(c_3+c_4-1)/2})$. Since they are linear combinations of zero mean Gaussian random vectors, the p columns of $\mathbf{V}_j$ as well as vector $\mathbf{y}-\mathbf{X}\hat{\mathbf{a}}^*$ are Gaussian. Though these $p+1$ vectors will be correlated for most interesting networks, the entries in any one of these vectors are i.i.d. Gaussian with variance less than C_{power}. We establish the following lemma.

Lemma 3: Let $\mathbf{V}$ be an n by p random matrix and $\mathbf{w}$ an n dimensional random vector. For each $i = 1, 2, \ldots, n$, let the i^{th} row of $\mathbf{V}$ concatenated with the i^{th} entry of $\mathbf{w}$ be i.i.d. Gaussian vectors with distribution $\mathcal{N}(\mathbf{0}, \mathbf{C})$, for some covariance matrix $\mathbf{C}$ whose maximum (diagonal) entry is C_m. Then with probability exceeding $1 - p\exp(-n)$, $\|\mathbf{V}^T\mathbf{w}\|_2 < C_m\sqrt{5np}$.

Proof: The entries in any column of $\mathbf{V}$ are i.i.d. Gaussian with variance less than C_m, as are the entries of $\mathbf{w}$. With this in mind, we bound each entry of $\mathbf{z} \equiv \mathbf{V}^T\mathbf{w}$ by C_m times a chi-squared random variable with n degrees of freedom (denoted $\tilde{z}_i \sim \chi^2_n$ for $i = 1, 2, \ldots, p$) and use the union bound:

$$\begin{aligned}
\mathbb{P}(\|\mathbf{z}\|_2^2 \geq C_m^2 5np) &\leq \mathbb{P}\left(\|\tilde{\mathbf{z}}\|_2^2 \geq 5np\right) \\
&\leq p\mathbb{P}\left(\tilde{z}_1^2 \geq 5n\right) \\
&\leq p\exp(-n)
\end{aligned}$$

where we have used the same chi-squared bound as in Lemma 1. Thus with probability exceeding $1 - p\exp(-n)$, $\|\mathbf{V}^T\mathbf{w}\|_2 < C_m\sqrt{5np}$.

■

Using Lemma 3, we have with probability exceeding $1 - p\exp(-n)$, $\|\mathbf{V}_j^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}^*)\|_2 < C_{power}\sqrt{5np}$. Dividing by λn and using Assumption 1, we have:

$$\frac{2}{\lambda n}\|\mathbf{V}_j^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{a}}^*)\|_2 < \frac{C_{power}\sqrt{5p}}{\lambda\sqrt{n}} = \mathcal{O}(n^{(c_3+c_4-1)/2}). \tag{22}$$

By (12), there will be no false positives if (21) is less than one. With high probability, the second term is less than $C_{fc} < 1$ by Assumption 6. The first and third terms go to zero with large n with high probability.

Union Bound

We have shown that (5) recovers the correct parents of node 1 (set $\mathcal{S}_1$) with probability exceeding $1 - \exp(-\Theta(n))$. To obtain the result for the whole network, we apply the union bound:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^N \hat{\mathcal{S}}_i \neq \mathcal{S}_i\right) &\leq \sum_{i=1}^N \mathbb{P}\left(\hat{\mathcal{S}}_i \neq \mathcal{S}_i\right) \\
&\leq N\exp(-\Theta(n)) \\
&\leq n^{c_1}\exp(-\Theta(n)) \\
&\leq \exp(c_1\ln n - \Theta(n)) \\
&\leq \exp(-\Theta(n))
\end{aligned}$$

APPENDIX B
PROOF OF NECESSARY CONDITION

We must show that (5) will not recover the correct set of nonzero $\mathbf{a}_{1,i}$ when Assumptions 2–5 hold but $\left\|\sum_{i\in\mathcal{S}_1}\mathbf{\Psi}_{j,i}^T\frac{\mathbf{a}_{1,i}}{\|\mathbf{a}_{1,i}\|_2}\right\|_2 > 1+c$. We do so by contradiction.

Suppose λ scales with n such that all the coefficient blocks in $\mathcal{S}_1$ of the oracle solution are nonzero and the probability of false positives goes to zero as n grows. Then KKT condition (12) must hold with high probability for large n. This implies the following bound must hold with high probability for all $j\in\mathcal{S}_1^C$:

$$\begin{aligned}
\frac{\lambda}{2} &\geq n^{-1}\|\mathbf{X}_j^T(\mathbf{y}_1-\mathbf{X}\hat{\mathbf{a}}_1^*)\|_2 \\
&= \frac{\lambda}{2}\left\|\sum_{i\in\mathcal{S}_1}\mathbf{\Psi}_{j,i}^T\frac{\hat{\mathbf{a}}_{1,i}^*}{\|\hat{\mathbf{a}}_{1,i}^*\|_2}+\frac{2}{\lambda n}\mathbf{V}_j^T(\mathbf{y}_1-\mathbf{X}\mathbf{a}_1^*)\right\|_2 \\
&\geq \frac{\lambda}{2}\left\|\sum_{i\in\mathcal{S}_1}\mathbf{\Psi}_{j,i}^T\frac{\mathbf{a}_{1,i}}{\|\mathbf{a}_{1,i}\|_2}\right\|_2-\frac{\lambda}{2}\|\mathbf{\Psi}_j^T\mathbf{w}\|_2 \\
&\quad -n^{-1}\left\|\mathbf{V}_j^T(\mathbf{y}_1-\mathbf{X}\mathbf{a}_1^*)\right\|_2 \\
&> \frac{\lambda}{2}(1+c)-\frac{\lambda}{2}\left\|\mathbf{\Psi}_j^T\mathbf{w}\right\|_2-n^{-1}\left\|\mathbf{V}_j^T(\mathbf{y}_1-\mathbf{X}\mathbf{a}_1^*)\right\|_2
\end{aligned} \tag{23}$$

where $\mathbf{w}=[\,\mathbf{w}_1\cdots\mathbf{w}_m\,]^T$ and $\mathbf{w}_{1,i}=\left(\frac{\hat{\mathbf{a}}_{1,i}^*}{\|\hat{\mathbf{a}}_{1,i}^*\|}-\frac{\mathbf{a}_{1,i}}{\|\mathbf{a}_{1,i}\|}\right)$. From Eq. (22) we have $n^{-1}\left\|\mathbf{V}_j^T(\mathbf{y}_1-\mathbf{X}\mathbf{a}_1^*)\right\|_2=\mathcal{O}(\sqrt{p/n})$, which goes to zero since $p/n\leq mp/n$, which goes to zero as n goes to infinity by assumption. We have also shown that $\left\|\mathbf{\Psi}_j^T\mathbf{w}\right\|_2$ goes to zero; however, this term is now multiplied by $\frac{\lambda}{2}$ for some unknown λ scaling. To proceed, Eq. (23) implies:

$$\frac{c\lambda}{2}<\frac{\lambda}{2}\left\|\mathbf{\Psi}_j^T\mathbf{w}\right\|_2+\mathcal{O}(\sqrt{p/n})$$

Since the second term goes to zero, this implies:

$$c<\|\mathbf{\Psi}_j^T\mathbf{w}\|_2\leq\|\mathbf{\Psi}_j\|_2\|\mathbf{w}\|_2\leq\frac{C_{max}}{C_{min}}\sqrt{m}\max_i\|\mathbf{w}_i\|_2$$

where the last inequality follows from the definition of $\mathbf{\Psi}_j$ and the triangle inequality. This means there is at least one $i\in\mathcal{S}_1$ for which $\left\|\frac{\hat{\mathbf{a}}_{1,i}^*}{\|\hat{\mathbf{a}}_{1,i}^*\|_2}-\frac{\mathbf{a}_{1,i}}{\|\mathbf{a}_{1,i}\|_2}\right\|_2\geq\frac{cC_{min}}{\sqrt{m}C_{max}}$. Combining this with Lemma (2) implies that $\|\hat{\mathbf{a}}_{1,i}-\mathbf{a}_{1,i}\|_2\geq\frac{cC_{min}\|\mathbf{a}_{1,i}\|_2}{2\sqrt{m}C_{max}}$.

Now we use Assumption 3 and (16):

$$\begin{aligned}
\frac{cC_{min}C_{con}}{2\sqrt{m}C_{max}} &\leq \frac{cC_{min}\|\mathbf{a}_{1,i}\|_2}{2\sqrt{m}C_{max}} \\
&\leq \|\mathbf{a}_{1,i}-\hat{\mathbf{a}}_{1,i}\|_2 \\
&= \left\|\widehat{\mathbf{G}}_{\mathcal{S}_{1,i}}^T\left(\frac{\lambda}{2}\hat{\mathbf{z}}_{\mathcal{S}_1}-\frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1\right)\right\|_2 \\
&= \left\|\begin{bmatrix}\mathbf{I}_p & \mathbf{0}\end{bmatrix}\widehat{\mathbf{G}}_{\mathcal{S}_1}^T\left(\frac{\lambda}{2}\hat{\mathbf{z}}_{\mathcal{S}_1}-\frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1\right)\right\|_2 \\
&\leq \left\|\widehat{\mathbf{G}}_{\mathcal{S}_1}^T\left(\frac{\lambda}{2}\hat{\mathbf{z}}_{\mathcal{S}_1}-\frac{1}{n}\mathbf{X}_{\mathcal{S}_1}^T\mathbf{u}_1\right)\right\|_2 \\
&< \frac{\lambda\sqrt{m}}{2}\left(C_{min}^{-1}+\mathcal{O}\left(\sqrt{\frac{mp}{n}}\right)\right)
\end{aligned}$$

where the last inequality follows from the proof of Lemma 1. Since mp/n goes to zero, we have the following lower bound on λ:

$$\lambda>\frac{cC_{min}^2C_{con}}{mC_{max}} \tag{24}$$

Since $\hat{\mathbf{a}}_{1,i}\neq 0$ for at least one i by assumption, KKT condition (11), repeated here for readability, must hold for at least one i:

$$\mathbf{X}_i^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1) = \frac{\lambda n \hat{\mathbf{a}}_{1,i}}{2\|\hat{\mathbf{a}}_{1,i}\|_2} \qquad \forall\ i \text{ s.t. } \hat{\mathbf{a}}_{1,i} \neq \mathbf{0} \tag{25}$$

Using Lemma 3 (with $\mathbf{V} = \mathbf{X}_i$ and $\mathbf{w} = \mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1$), the norm of the left hand side of (25) is less than $C_{power}\sqrt{5np}$ with high probability for large n. On the other hand, (24) implies that the norm of the right hand side of (25) is $\Omega(n/m)$. Given that n grows faster than m^2p, this is a contradiction.

The scaling law $n > m^2p$ for large n (equivalently $2c_2 + c_3 < 1$) was not required to prove asymptotic consistency. Other proof techniques may result in matching scaling laws.

Appendix C
Proof of Corollary 1

The proof is the same as that of Theorem 1 with a few minor changes. The KKT condition (11) for $l = 1$ becomes $\mathbf{X}_1^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1) = \mathbf{0}$, which implies $\hat{\mathbf{z}}_{1,1} = \mathbf{0}$. The results of Lemma 1 still apply with m replaced by $m-1$ in the proof. In App. A, $\psi_{j\to 1}^{FC} = \left\|\sum_{i\in\mathcal{S}_1} \mathbf{\Psi}_{j,i}^T \frac{\hat{\mathbf{a}}_{1,i}}{\|\hat{\mathbf{a}}_{1,i}\|_2}\right\|_2$ is simply replaced with $\tilde{\psi}_{j\to 1}^{FC} = \left\|\sum_{i\in\mathcal{S}_1, i\neq 1} \mathbf{\Psi}_{j,i}^T \frac{\hat{\mathbf{a}}_{1,i}}{\|\hat{\mathbf{a}}_{1,i}\|_2}\right\|_2$ since $\mathbf{X}_1^T(\mathbf{y}_1 - \mathbf{X}\hat{\mathbf{a}}_1) = \mathbf{0}$ instead of $\frac{\lambda n \hat{\mathbf{a}}_{1,1}}{2\|\hat{\mathbf{a}}_{1,1}\|_2}$.

References

[1] L. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biological Cybernetics*, vol. 84, pp. 463–474, 2001.

[2] M. Kamiński, "Determination of transmission patterns in multichannel data," *Philosophical Transactions of the Royal Society B*, vol. 360, pp. 947–952, 2005.

[3] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistritz, D. Klan, R. Bauer, J. Timmer, and H. Witte, "Comparisson of linear signal processing techniques to infer directed interactions in multivariate neural systems," *Signal Processing*, vol. 85, no. 11, pp. 2137–2160, 2005.

[4] H. Lütkepohl, *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag, 1991.

[5] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, pp. 267–288, 1996.

[6] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B*, vol. 68, no. 1, pp. 49–67, 2006.

[7] A. Bolstad, B. Van Veen, and R. Nowak, "Space-time sparsity regularization for the magnetoencephalography inverse problem," in *4th IEEE International Symposium on Biomedical Imaging*, Arlington, VA, April 2007, pp. 984–987.

[8] ——, "Magneto-/electroencephalography with space-time sparse priors," in *IEEE Statistical Signal Processing Workshop*, Madison, WI, August 2007, pp. 190–194.

[9] L. Ding and B. He, "Sparse source imaging in EEG," in *Proceedings of NFSI & ICFBI*, Hangzhou, China, October 2007.

[10] W. Ou, M. Hämäläinen, and P. Golland, "A distributed spatio-temporal eeg/meg inverse solver," *NeuroImage*, vol. 44, no. 3, pp. 932–946, 2009.

[11] S. Haufe, V. Nikulin, A. Ziehe, K.-R. Müller, and G. Nolte, "Combining sparsity and rotational invariance in eeg/meg source reconstruction," *NeuroImage*, vol. 42, no. 2, pp. 726–738, 2008.

[12] A. Bolstad, B. van Veen, and R. Nowak, "Space-time event sparse penalization for mangeto-/electroencephalography," *NeuroImage*, vol. 46, no. 4, pp. 1066–1081, July 2009.

[13] S. Haufe, K. Müller, G. Nolte, and N. Krämer, "Sparse causal discovery in multivariate time series," *Journal of Machine Learning Research: Workshops & Conference Proceedings (JMLR W&CP)*, vol. 6 (NIPS 2008), pp. 97–106, 2010.

[14] A. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical granger modeling for gene expression regulatory networks discovery," *Bioinformatics*, vol. 25, pp. i110 – i118, 2009.

[15] S. van de Geer and P. Bühlmann, "On the conditions used to prove oracle results for the lasso," *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[16] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, 2008.

[17] A. Lozano, G. Swirszcz, and N. Abe, "Grouped orthogonal matching pursuit for variable selection and prediction," in *Advances in Neural Information Processing Systems (NIPS)*, no. 22, 2009.

[18] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[19] P. Ravikumar, G. Raskutti, M. Wainwright, and B. Yu, "Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularizedmle," in *Advances in Neural Information Processing Systems (NIPS)*, no. 21, 2008.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[21] A. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical granger modeling methods for temporal causal modeling," 2009, pp. 577 – 585.

[22] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, pp. 1418–1429, 2006.

[23] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part ii: Convex relaxation," *Signal Processing, special issue on Sparse approximations in signal and image processing*, vol. 86, pp. 589–602, April 2006.

[24] L. Meier, S. van de Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society Series B*, vol. 70, no. 1, pp. 53–71, 2008.

[25] S. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2477–2488, July 2005.

[26] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, December 2006.

[27] J. Tropp, A. Gilbert, and M. Strauss, "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing, special issue on Sparse approximations in signal and image processing*, vol. 86, pp. 572–588, April 2006.

[28] G. Obozinski, M. Wainwright, and M. Jordan, "Union support recovery in high-dimensional multivariate regression," UC Berkeley, Tech. Rep. 761, August 2008.

[29] H. Wang and C. Leng, "A note on adaptive group lasso," *Computational Statistics and Data Analysis*, vol. 52, pp. 5277–5286, 2008.

[30] H. Liu and J. Zhang, "Estimation consistency of the group lasso and its applications," *Journal of Machine Learning Research: Workshops & Conference Proceedings (JMLR W&CP)*, vol. 5, pp. 376–383, 2009.

[31] P. Valdes-Sosa, J. Sanchez-Bornot, J. Lage-Castellanos, M. Vega-Hernandez, J. Bosch-Bayard, L. Melie-Garcia, and E. Canales-Rodriguez, "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B*, vol. 360, pp. 969–981, 2005.
[32] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *J. Machine Learning Research*, no. 11, pp. 2671–2705, 2010.
[33] J. Bento, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," `arXiv:1011.0415v1 [math.ST]`, 2010.
[34] N. Meinshausen and P. Bühlmann, "Stability selection," *J. Royal Statistical Society B*, vol. 72, no. 4, pp. 417–473, 2010.
[35] E. Pereda, R. Q. Quiroga, and J. Bhattacharya, "Nonlinear multivariate analysis of neurophysiological signals," *Progress in Neurobiology*, vol. 77, no. 1, pp. 1–37, 2005.
[36] M. Carroll, G. Cecchi, R. Rish, R. Garg, and A. Rao, "Prediction and interpretation of distributed neural activity with sparse models," *NeuroImage*, vol. 44, no. 1, pp. 112–122, January 2009.
[37] S. Haufe, R. Tomioka, G. Nolte, K. Müller, and M. Kawanabe, "Modeling sparse connectivity between underlying brain sources for eeg/meg," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 8, pp. 1954–1963, 2010.
[38] M. Young, "The organization of neural systems in the primate cerebral cortex," *Proc. Biol. Sci.*, no. 252, pp. 13–18, 1993.
[39] O. Sporns, *Graph theory methods for the analysis of neural connectivity patterns*, R. Kütter, Ed. Boston: Klüwer, 2002.
[40] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical Journal*, vol. 50, pp. 346 – 363, 2008.
[41] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Letters to Nature*, vol. 393, pp. 440–442, June 1998.
[42] O. Sporns, C. Honey, and R. Kötter, "Identification and classification of hubs in brain networks," *PLoS ONE*, vol. 2, p. e1049, October 2007.
[43] O. Sporns, G. Tononi, and G. Edelman, "Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices," *Cerebral Cortex*, vol. 10, pp. 127 – 141, 2000.
[44] B. Laurent and P. Massart, "Adaptive estimation of a quadratic function by model selection," *Annals of Statistics*, vol. 28, no. 5, pp. 1302–1338, 2000.