Sparse System Identification for Discovering Brain Connectivity from fMRI time series

Arnan Pongrattanakul, Puttichai Lertkultanon and Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering, Chulalongkorn University, Thailand
(Tel: +662-2186487; E-mail: arnan_osk127@hotmail.com, L.Puttichai@gmail.com, jitkomut.s@chula.ac.th)

Abstract: This paper presents a convex framework for problems of fitting multivariate autoregressive (AR) models that cooperate Granger causality constraints to fMRI time series which describe the dynamics of the human brain activity level. The Granger causality characterizes a relationship structure of variables in the system and can be explained from a common zero pattern of AR coefficients. We present two important model estimation problems that can be expressed as constrained least-squares and ℓ_1-type regularized least-squares formulations. We show that the first problem has a closed-form solution, while the second one admits a group lasso formulation which can be solved efficiently by a convex optimization technique. In combination with model selection criteria, these two problems produce a sparse AR model whose coefficients's sparsity can reveal a Granger causal inference illustrated as a graphical model for fMRI time series. We verify the proposed method on simulated data sets and find that a desirable performance is obtained if the graph underlying the true model is sparse. The experiment result on an fMRI data set shows that this method leads to a reasonable graphical model of brain dynamics which can be a useful guideline for further studies in neuroscience.

Keywords: fMRI time series, autoregressive models, Granger causality, sparse system identification

1. INTRODUCTION

fMRI (Functional Magnetic Resonance Imaging) is a neuroimaging technique that allows us to measure activity levels of the human brain via blood oxygen-level dependent (known as BOLD signals). One of challenging research topics is the analyzes of brain effective connectivity, *i.e.*, a relationship structure that describes how one neuron affects another. There have been numerous studies that applied the concept in control theory or model estimation in order to find a model for fMRI time series. A model described by nonlinear differential equations (known as dynamic causal modeling or DCM), vector autoregressive models, state-space models (or referred to as multivariate dynamical systems in [1]) are examples of known modeling techniques that have been of interest recently; see a short survey on fMRI in [2], a summary of current techniques of learning brain connectivity in [3] and the original references therein.

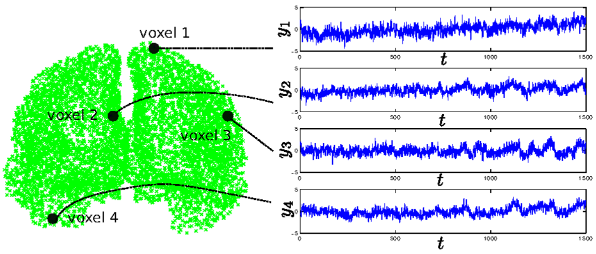

Fig. 1 Example of fMRI time series measured at four voxels of the human brain. Each voxel represents a surrogate for neuronal populations. The acquired data used in this project contain the time records from 6004 voxels, resulting in autoregressive models with dimension $n = 6004$.

Since the human brain is a large complex system, one would like to analyze the brain connectivity and illustrate it as a brain network [3-5]. Nodes in a brain network represent brain regions or neurons and the presence of a link between two nodes describe effective connections. These connectivities can be defined and examined through several statistical concepts such as partial correlation, Granger causal modeling (GCM) and dynamic causal modeling (DCM) [1, 3, 5, 6]. Though it is still a debatable and interesting topic to conclude which approach is the most suitable for fMRI analysis, our focus here is pursue one of the common approaches, *i.e.*, to discover causal relations via the concept of Granger causality because its characterization on a linear model turns out to be computationally simple. The concept of Granger causality states that a time series y_i is Granger-caused by time series y_j if knowing the past values of y_j helps improve the prediction of y_i [7]. There is a nice characterization of Granger causality for autoregressive processes which is widely used to model multivariate time series in many applications; see [8-10]. An n-dimensional autoregressive (AR) process of order p is given by

$$y(t) = \sum_{k=1}^{p} A_k y(t-k) + u(t) \qquad (1)$$

where $y(\cdot) \in \mathbf{R}^n$, $A_k \in \mathbf{R}^{n \times n}$, $k = 1, 2, \ldots, p$ and $u(\cdot)$ is input noise. Let y_i and y_j be the ith and jth voxels, respectively and let n be the number of voxels in the human brain. A brain network has no directed edge from node j to i if y_i is not *Granger-caused* by y_j and this can be characterized in terms of AR coefficients as [7]

$$(A_k)_{ij} = 0, \quad k = 1, 2, \ldots, p \qquad (2)$$

(we denote $(A_k)_{ij}$ the (i, j) entry of the matrix A_k.)

PR0001/13/0000-0949 ¥400 © 2013

One can estimate an AR model using a simple method such as a linear regression and develop a statistical test to examine a pairwise Granger causality between any two pair of the variables. There is a freely available MATLAB toolbox developed by [11]. The computation used in this approach may, however, become infeasible when the number of neurons or voxels (n) grows. Moreover, with a limited number of data points, it is known that the problem of estimating a complex system (in high-dimension small-sample setting) will become ill-conditioned. The work in [12] considered the Least Absolute Shrinkage Selection Operator (LASSO) to pre-select voxels before estimating an AR model. A common remedy for such high-dimension small-sample problem is to incorporate a regularization term in the estimation problem [13]. They showed that adding ℓ_1-norm penalty term into linear regression problems induces zeros in the estimated variables. Examples of this approach applied on fMRI data can be found in [8, 14, 15]. Using this approach, the estimated AR coefficients will be sparse matrices, but we do not necessarily obtain a *common* sparsity among A_k's as required in (2). To solve this, we consider a group lasso formulation [16, 17] in the estimation problem. Our problem formulation is also similar to the ones shown in [9, 10]. However, we will show a competitively efficient algorithm for solving the resulting optimization problems.

Section 2 presents two estimation problems categorized by our knowledge of Granger causality patterns. Then we compare the performance of the group lasso approach and a ridge regression technique in section 3. In section 4, we illustrate two model selection methods for choosing a parameter in the estimation problem. These methods will be used in topology selection of a graphical model for time series. Section 5 presents experimental results on a real fMRI data set.

2. ESTIMATION PROBLEMS

The least-squares (LS) method is a common approach used for fitting an AR model (1) to the measurements $y(1), y(2), \ldots, y(N)$. The model parameters A_k's are chosen such that the quadratic loss $\sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|_2^2$ is minimized. If we define $A = \begin{bmatrix} A_1 & \cdots & A_p \end{bmatrix} \in \mathbf{R}^{n \times np}$ then the quadratic loss can be rewritten more compactly as $\|Y - AH\|_2^2$ where

$$Y = \begin{bmatrix} y(p+1) & y(p+2) & \cdots & y(N) \end{bmatrix}, \qquad (3)$$

$$H = \begin{bmatrix} y(p) & y(p+1) & \cdots & y(N-1) \\ y(p-1) & y(p) & \cdots & y(N-2) \\ \vdots & \vdots & \vdots & \vdots \\ y(1) & y(2) & \cdots & y(N-p) \end{bmatrix}. \qquad (4)$$

AR estimation problems that take the Granger causality (2) into account can be divided into two categories depending on whether a causal inference is given or not. We will show that both of the two problems fall into a convex optimization framework which can be solved by efficient algorithms presented in the appendix.

2.1 Known Granger Causality

If a Granger causality structure is given (for example, a brain network topology is known), formulating the problem of estimating AR model subject to the zero pattern of A_k's as in (2) is straightforward and given by

$$\begin{array}{ll} \text{minimize} & (1/2)\|Y - AH\|_2^2 \\ \text{subject to} & (A_1)_{ij} = (A_2)_{ij} = \cdots = (A_p)_{ij} = 0 \end{array} \qquad (5)$$

for $(i, j) \notin \mathcal{V}$, where $\mathcal{V}$ is the index set of common nonzero entries in A_k's. In other words, if we present the given Granger causal inference as a graph, then $\mathcal{V}$ is the set of edges in such graph. Though it appears unlikely how one could know a priori about the causal inference, the problem (5) becomes more important if one wishes to find an AR model whose estimated parameters have less variance since some of them are shrunk to zero [13].

The problem (5) has a closed-form solution due to its simple linear constraints. The details of calculating a closed-form solution will be shown in the appendix.

2.2 Unknown Granger Causality

In most applications including the fMRI study, the goal is to learn a Granger causal inference from the data, so the graph topology is commonly *unknown*. The topology can be induced from a *common* zero pattern of matrices A_k's. Therefore, we consider a formulation that favors a *group* sparsity in A_k's. This can be done by introducing a sum of ℓ_2-norm in the cost objective as

$$\text{minimize} \quad (1/2)\|Y - AH\|_2^2 + \lambda g(A) \qquad (6)$$

where $g(A) = \sum_{i \neq j} \|[(A_1)_{ij} \; (A_2)_{ij} \; \cdots \; (A_p)_{ij}]\|_2$.

The optimization variable is $A = \begin{bmatrix} A_1 & \cdots & A_p \end{bmatrix}$ where $A_k \in \mathbf{R}^{n \times n}$ for $k = 1, \ldots, p$. The scalar $\lambda > 0$ is called the regularization parameter which controls a trade-off between the quadratic loss and the penalty term. If we define $\mathbf{a}_{ij} = \|[(A_1)_{ij} \; (A_2)_{ij} \; \cdots \; (A_p)_{ij}]\|_2$, we can write $g(A) = \sum_{i \neq j} \mathbf{a}_{ij}$. This notation suggests that $g(A)$ plays a role of ℓ_1-norm of the matrix $[\mathbf{a}_{ij}]$, so for a sufficiently large λ, $g(A)$ will be small and this will cause *some* (i, j) entries $\mathbf{a}_{ij}$ to zero. Furthermore, using the ℓ_2 norm of p-tuple of $(A_k)_{ij}$ will force all p matrices A_k's to have the same sparsity pattern, *i.e.*, $\mathbf{a}_{ij} = 0 \Leftrightarrow (A_k)_{ij} = 0$ for all k. This is a common technique to force a group sparsity pattern and is known as a *Group Lasso* problem introduced in [17]. The formulation (6) is also independently considered in [9, 10, 18]. In these studies, they have shown an advantage of using group lasso formulation over the standard lasso where the estimated A_k's may have different zero patterns.

While the problem (6) is an unconstrained convex program, it is quite challenging to solve it in a large-scale setting due to the nondifferentiability of $g(A)$. We will briefly describe in the appendix a widely-used algorithm for large-scale convex problems called the alternating direction method of multipliers for solving (6). The implementation details and its numerical performance were shown in our related paper [19].

3. CLASSIFICATION PERFORMANCE

Using several values of λ in (6) results in several estimates of A_k's, where each of them corresponds to a sparsity pattern, ranging from dense to sparse. The effectiveness of the formulation (6) for learning sparse models can be explained from receiver operating characteristic (ROC) curve [20]. If we make a comparison of the true and estimated sparsity patterns, classifying the *common* nonzero entries in A_k's has two types of errors: 1) the misclassified entries as nonzero (False Positive) and 2) the misclassified entries as zero (False Negative). We can compute the total error by

$$\text{error} = \frac{\text{False Positive} + \text{False Negative}}{n^2 - n} \tag{7}$$

(note that the total number of entries in the off-diagonal of matrices A_k is $n^2 - n$.) Evaluating the performance of a binary classifier (detecting whether an entry is zero or nonzero) is commonly done via a receiver operating characteristic curve [20], which is a plot between the true positive rate (number of correctly identified nonzeros) versus the false positive rate. Each point along an ROC curve corresponds to a value of the classifier parameter. Using this evaluation technique, we can view (6) as a classifier with parameter λ, and the ROC curve is obtained by varying λ from zero to a large value. When $\lambda = 0$, we obtain the least-squares solution of A_k's and they are typically dense matrices. Hence, we expect a high true positive rate and also a high false positive rate. On the other hand, if λ is large, the formulation (6) returns a sparse solution, so we expect a decrease in the false positive rate. A good classifier should yield an ROC curve that is above the diagonal line (a random guess classifier) and tends towards the top-left corner.

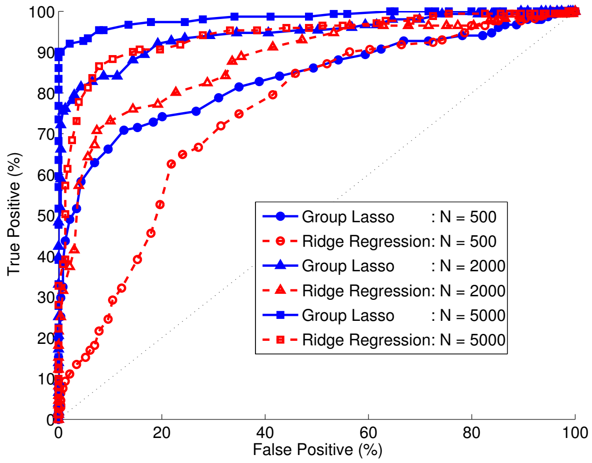

Fig. 2 Receiver operating curves (ROC) of Group lasso (**blue solid line**) and Ridge regression (**red dashed line**). We vary the number of time samples used in each method; $N = 500$ (circles), $N = 2000$ (triangles), and $N = 5000$ (square).

We randomly generate sparse AR models with $n = 20, p = 3$ and compare ROC curves between group lasso (6) and ridge regression:

$$\text{minimize} \quad (1/2)\|Y - AH\|_2^2 + \lambda\|A\|_2^2. \tag{8}$$

The latter problem is also known as the ℓ_2-regularized least-squares [13] which is a common and better based-line approach than the traditional least-squares. The regularization λ in the ridge regression is chosen via 10-fold cross validation (with respect to prediction accuracy in a 2-norm sense). To determine the estimated sparsity pattern (or Granger causality) from the solution to (8), we construct

$$B = [\mathbf{a}_{ij}], \quad \mathbf{a}_{ij} = \left\| \begin{bmatrix} (A_1)_{ij} & \cdots & (A_p)_{ij} \end{bmatrix} \right\|_2. \tag{9}$$

A Granger causal inference can then be read from the zero pattern in B by comparing its off-diagonals with a threshold value ϵ. We normalize B so that it has unit diagonals and the ROC curve is constructed by varying ϵ.

Figure 2 illustrates that the ROC curves of group lasso formulation (6) lie above the ridge regression curve, meaning our approach yields a higher true positive rate, and lower false positive rate. By varying the number of time samples (N), an improvement of classification performance is also expected, but our approach performs better than the ridge regression even when N is small.

4. MODEL SELECTION & VALIDATION

From Figure 2, our goal is to achieve an operating point near the top-left corner. This is equivalent to finding λ corresponding to that point. In practice, it is not possible to calculate the true and false positive rates in advance since the true sparsity pattern is unknown. In this study, we consider a model selection problem by incorporating a Bayes information criterion (BIC) score [13] for ranking a subset of candidate topologies (group sparsity patterns of A_k's) obtained by solving (6) for M values of λ.

We generate three AR models with dimension $n = 20, p = 3$ and $N = 1500$ and the matrices A_k's in each of the models have the density of nonzero entries as $5\%, 35\%$ and 70% respectively (ranging from sparse to dense models.) Solving (6) by using M values of λ results in M estimated group sparsity patterns of A_k's, ranging from densest to sparsest topologies. Then we use each of these topologies as a Granger causality constraint in (5) and solve it to obtain M model candidates whose complexity is varied from low to high. The chosen λ corresponds to the Granger causality constrained AR model that minimizes the BIC score. We show in the appendix that the problem (5) has a closed-form solution, so it can be computed very efficiently. For this reason, it allows us to consider a large set of model candidates, says we choose $M = 400$. In this experiment, the top row in Figure 3 shows that the best model according to BIC yields the error of 1.58% in the estimated topology when the true AR model has the density of nonzero entries of 5%, and the errors increase if the true AR models tend to be denser. The bottom row in Figure 3 shows the result when λ is chosen by a cross validation technique.

In this case, the data are split into estimation and validation sets. We vary λ for M values, and pick the one that yields the highest prediction accuracy which is evaluated on the validation data set [13]. The plots illustrate that the cross validation technique tends to perform better than BIC when the true models are dense. This agrees with a known result that BIC prefers to favor the model with low complexity.

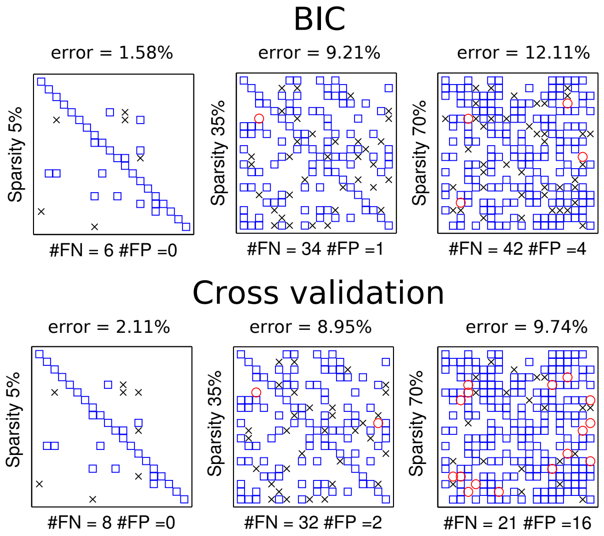

Fig. 3 Binary matrix of the *common* zero patterns in the estimated A_k's. The **blue** squares are the correctly estimated nonzero entries (TP). The **red** circles are misclassified entries as nonzero (FP). The **black** cross signs are misclassified entries as zero (FN).

5. EXPERIMENT ON FMRI DATA

The fMRI time series considered in this project are the commonly-used BOLD (blood oxygen-level dependent) signals for analyzing brain effectivity, which were recorded while a subject was being in a resting a state. The details of data acquisition through an fMRI machine can be found in [21]. The data contain the 1500 time samples recorded from 6004 voxels of interest. Note that for an n-dimensional AR model of order p, the parameters are $A_1, A_2, \ldots, A_p$, so the total number of free parameters is n^2p. This means if we were to fit the time series from all voxels with a full AR model, we would require a memory space for storing more than 4×10^7 parameters (or at least 300 MB just for a single matrix A_k). Thus, we decide to reduce the number of voxels of interest to $n = 201$ by sampling 201 voxels that spatially cover almost all areas of the brain. We applied the model selection method using the BIC score explained in section 2. There are 200 candidate models corresponding to 50 different graph topologies and model order of $p = 1, 2, 3, 4$. The selected AR model has order 1 and the density of nonzero entries in the AR coefficients is 7.04%. The corresponding Granger graphical model that best explains the fMRI time series is shown in Figure 4. We compare this result with some studies in neuroscience, where the network of brain activity during a resting state is often called *the default mode network (DMN)* [22-25]. In these papers, it has revealed that the main components of the DMN are the precuneus/posterior cingulate cortex/retrosplenial cortex (pC/PCC/RSC), the ventral anterior cingulate cortex (vACC), the medial prefrontal cortex (MPFC) and the medial temporal lobes (MTLs). By considering the sagittal view in Figure 4, we believe we found many active nodes in vACC, MTLs, and a few dominant nodes in MPFC and PCC/RSC. Moreover, it seems there are strong connections between MTLs and PCC, and a connection between MPFC and PCC, which agree with the findings in [22]. In [24], it was shown that vACC has a significant connectivity with PCC, which is also found in our result. The coronal view gives the expression of strong connections between left and right medial temporal lobes, which are also found in [25]. There are other connection pairs that are not discussed in the previous work. For example, the sagittal view shows connections between the prefrontal cortex and the temporal lobes and connections between parietal lobes and temporal lobes. It will be our best interest to interpret these findings in the future work.

6. CONCLUSIONS

We have presented a useful application of system identification on a human brain study. Exploring relationship structures in fMRI time series can be casted as a model estimation problem with Granger causality constraints. We have considered a problem of fitting autoregressive models that favors sparse AR coefficient matrices. It was shown that the formulation is in the form of a least-squares problem with a sum of 2-norm regularization term, to which we refer as a group lasso formulation. An advantage of this approach is that we are able to obtain solutions $A_1, A_2, \ldots, A_p$ that have a common sparsity pattern, revealing a Granger causal inference. Numerical results on simulated data show that the group lasso formulation yields a better performance than the conventional ridge regression in classifying whether a pair of two variables are Granger-caused to each other or not (learning whether A_k's have a common zero). Moreover, we have described a model selection method for learning the most suitable sparsity pattern (or graph topology) for the given data. Using BIC score tends to pick a sparse model, which result in a low estimation error if the true model is also sparse, while the cross validation technique favorably selects a denser model. Finally, the result on the fMRI data set suggested that the posterior cingulate cortex, ventral anterior cingulate cortex, temporal lobes, and the prefrontal cortex are the main elements of brain functional in the resting state. The graphical model also showed connections between some regions that are also discovered in some previous brain connectivity studies.

7. ACKNOWLEDGEMENT

Thanks to Prof. Dr. Pedro Antonio Valdés Sosa and Dr. Jose Miguel Sánchez Bornot from Cuban Neuroscience Center (CNEURO) for an fMRI data set and useful comments on the fMRI experiment. Thanks to Prof.

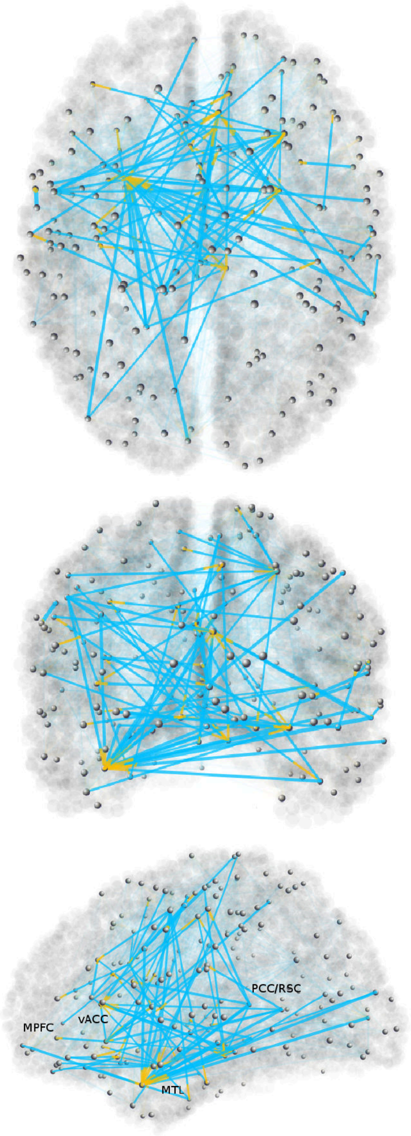

Fig. 4 The estimated Granger graphical model of fMRI time series shown from axial view (top), coronal view (middle) and sagittal view (bottom). The gray cloud covers the whole brain volume and the dark grey dots (graph nodes) locate where the selected 201 voxels are. The width of a link between a node pair (i, j) is proportional to $\mathbf{a}_{ij}$. The orange color painted at the link end towards node j represents that the node j is Granger-caused by node i.

Jafri Malin Abdullah, Universiti Sains Malaysia for explanations on brain regions. We also would like to thank Tanachai Bunlutangtum, Chulalongkorn University for a 3D graphical model plotting program.

APPENDIX

Problem (5): Known Granger causality

We will show that this problem can be, in fact, divided into n independent least-squares problems. Hence, the solution to this problem has a closed-form expression and can be solved efficiently. To this end, we rearrange the optimization variables, $A_1, \dots, A_p$ as

$$\mathbf{a}_{ij} = \begin{bmatrix}(A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij}\end{bmatrix}^T \in \mathbf{R}^p \qquad (10)$$

for $i, j = 1, 2, \dots, n$. The goal is to write the cost objective (5) in terms of $\mathbf{a}_{ij}$ and we will make some $\mathbf{a}_{ij}$ zero for those (i, j) that are *not* in $\mathcal{V}$. This also requires rearranging the entries of Y and H in (3)-(4) as follows. Let $\mathbf{y}_k$ be the k^{th} column of Y^T in (3), *i.e.*, $Y^T \triangleq \begin{bmatrix}\mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n\end{bmatrix}$ and let

$$X_j = \begin{bmatrix} H_{j,1} & H_{j+n,1} & \cdots & H_{j+(p-1)n,1} \\ H_{j,2} & H_{j+n,2} & \cdots & H_{j+(p-1)n,2} \\ \vdots & \vdots & \ddots & \vdots \\ H_{j,N-p} & H_{j+n,N-p} & \cdots & H_{j+(p-1)n,N-p} \end{bmatrix}.$$

By using the above notations, the problem (5) can be equivalently solved through the following n independent subproblems:

$$\begin{array}{ll} \text{minimize} & (1/2) \left\| \mathbf{y}_i - \sum_{j=1}^{n} X_j \mathbf{a}_{ij} \right\|_2^2 \\ \text{subject to} & \mathbf{a}_{ij} = 0, \quad (i, j) \notin \mathcal{V}, \end{array} \qquad (11)$$

for $i = 1, 2, \dots, n$. For those index pairs $(i, j) = (i_0, j_0)$ that $(i_0, j_0) \notin \mathcal{V}$, we have $\mathbf{a}_{ij} = 0$ which makes the matrix X_{j_0} irrelavant to the cost objective in (11). When $i = i_0$, the i^{th} subproblem can therefore be reduced to

$$\text{minimize} \quad (1/2) \left\| \mathbf{y}_{i_0} - \mathbf{X}\mathbf{a} \right\|_2^2 \qquad (12)$$

where $\mathbf{X}$ is formed by horizontally concatenating matrices X_j and $\mathbf{a}$ is formed by vertically concatenating vectors $\mathbf{a}_{i_0 j}$ for $j = 1, 2, \dots, n$ except for $j = j_0$. Thus, the closed-form solution of (12) is given by $\mathbf{a} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}_{i_0}$ provided that $\mathbf{X}$ is full rank.

Problem (6): Unknown Granger Causality

This problem falls in to a group lasso formulation where many algorithms have been proposed [16, 17, 26, 27]. We selectively describe a competitive algorithm, namely, the alternating direction method of multiplier (ADMM) [27] for solving (6). The algorithm repeats the following steps for $k = 0, 1, 2, \dots$

$$\begin{aligned} A^{(k+1)} &= \underset{A}{\operatorname{argmin}} \, \|Y - AH\|_F^2 + \rho \|A - Z^{(k)} + U^{(k)}\|_F^2, \\ Z^{(k+1)} &= \underset{Z}{\operatorname{argmin}} \, \frac{\rho}{2} \left\| Z - \left(A^{(k+1)} + U^{(k)}\right) \right\|_F^2 + \\ &\qquad \lambda \sum_{i \neq j} \left\| \left[(Z_1)_{ij} (Z_2)_{ij} \cdots (Z_p)_{ij} \right] \right\|_2, \\ U^{(k+1)} &= U^{(k)} + A^{(k+1)} - Z^{(k+1)}, \end{aligned}$$

until the residual error $A - Z$ is small. Here, U and Z are the auxiliary variables and $\rho > 0$ is an ADMM parameter. The algorithm format and the details of efficient numerical implementation can be found in [19].

REFERENCES

[1] S. Ryali, K. Supekar, T. Chen, and V. Menon, “Multivariate dynamical systems models for estimating causal interactions in fMRI,” *Neuroimage*, vol. 54, no. 2, pp. 807–823, 2011.

[2] K. E. Stephan and A. Roebroeck, “A short history of causal modeling of fMRI data,” *NeuroImage*, vol. 62, no. 2, pp. 856–863, 2012.

[3] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston, “Effective connectivity: Influence, causality and biophysical modeling,” *Neuroimage*, vol. 58, no. 2, pp. 339–361, 2011.

[4] M. Rubinov and O. Sporns, “Complex network measures of brain connectivity: uses and interpretations,” *Neuroimage*, vol. 52, no. 3, pp. 1059–1069, 2010.

[5] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, “Network modelling methods for FMRI,” *Neuroimage*, vol. 54, no. 2, pp. 875–891, 2011.

[6] S. Ryali, K. Supekar, T. Chen, and V. Menon, “Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty,” *NeuroImage*, vol. 59, no. 4, pp. 3852–3861, 2011.

[7] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*, Springer, 2005.

[8] P. A. Valdés-Sosa, J.M. Bornot-Sánchez, M. Vega-Hernández, L. Melie-García, A. Lage-Castellanos, and E. Canales-Rodríguez, “Granger causality on spatial manifolds: Applications to neuroimaging,” in *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, B. Schelter, M. Winterhalder, and J. Timmer, Eds. Wiley, 2006.

[9] S. Haufe, G. Nolte, and N. Kräemer, “Sparse causal discovery in multivariate time series,” *Proceedings of JMLR Workshop and Conference*, vol. 6, pp. 97–106, 2008.

[10] A.C. Lozano, N. Abe, Y. Liu, and S. Rosset, “Grouped graphical granger modeling for gene expression regulatory networks discovery,” *Bioinformatics*, vol. 25, pp. 110–118, 2009.

[11] A. Seth, “A MATLAB toolbox for Granger causal connectivity analysis,” *Journal of Neuroscience Methods*, vol. 186, no. 2, pp. 262–273, 2010.

[12] W. Tang, S. Bressler, C. M. Sylvester, S. L. Gordon, and M. Corbetta, “Measuring Granger causality between cortical regions from voxelwise fMRI bold signals with lasso,” *PLoS Computational Biology*, vol. 8, no. 5, pp. e1002513, 2012.

[13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2nd edition, 2009.

[14] R. Garg, G. A. Cecchi, and A. R. Rao, “Full-brain auto-regressive modeling (FARM) using fMRI,” *Neuroimage*, vol. 58, no. 2, pp. 416–441, 2011.

[15] B. Ng and R. Abugharbieh, “Generalized sparse regularization with application to fMRI brain decoding,” in *Information Processing in Medical Imaging*. Springer, 2011, pp. 612–623.

[16] J. Friedman, T. Hastie, and R. Tibshirani, “A note on the group lasso and a sparse group lasso,” 2010, Preprint available at `arXiv.org` (1001.0736).

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

[18] A. A. Bolstad, B. Van Veen, and R. Nowak, “Causal network inference via group sparse regularization,” *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[19] J. Songsiri, “Sparse autoregressive model estimation for learning Granger causality in time series,” in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2013.

[20] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, 2004.

[21] D. A. Feinberg, S. Moeller, S. M. Smith, E. Auerbach, S. Ramanna, M. F. Glasser, K. L. Miller, K. Ugurbil, and E. Yacoub, “Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging,” *PLoS One*, vol. 5, no. 12, pp. e15710, 2010.

[22] M. D. Greicius, K. Supekar, V. Menon, and R. F. Dougherty, “Resting-state functional connectivity reflects structural connectivity in the default mode network,” *Cerebral Cortex*, vol. 19, no. 1, pp. 72–78, 2009.

[23] P. Fransson, “Spontaneous low-frequency bold signal fluctuations: An fMRI investigation of the resting-state default mode of brain function hypothesis,” *Human Brain Mapping*, vol. 26, no. 1, pp. 15–29, 2005.

[24] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon, “Functional connectivity in the resting brain: a network analysis of the default mode hypothesis,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 1, pp. 253–258, 2003.

[25] P. Fransson and G. Marrelec, “The precuneus/posterior cingulate cortex plays a pivotal role in the default mode network: Evidence from a partial correlation network analysis.,” *Neuroimage*, vol. 42, no. 3, pp. 1178–1184, 2008.

[26] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009.

[27] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.