

# INFERRING BRAIN DYNAMICS USING GRANGER CAUSALITY ON FMRI DATA

Guillermo A. Cecchi, Rahul Garg and A. Ravishankar Rao

Computational Biology Center, IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

## ABSTRACT

Here we present a scalable method to compute the structure of causal links over large scale dynamical systems that achieves high efficiency in discovering actual functional connections. The method is based on the Granger causality analysis of multivariate linear models, solved by means of a sparse regression approach, and can deal with autoregressive models of more than 10,000 variables.

**Index Terms**— Magnetic Resonance Imaging, Image interpretation, Functional Imaging

## 1. INTRODUCTION

Neuroscience, as most other fields in biology, and experimental science in general, is experiencing a growth bottleneck: inordinate amounts of data (high resolution imaging, multi-electrode recordings, gene expression arrays) meet conceptual frameworks and mathematical models developed for a more scarce context, such as single-unit electrophysiology. Indeed, successful mathematical modeling in neuroscience has been dominated by either low dimensional or linear or quasi-linear models. However, neuroanatomy and neurophysiology indicate that brain function is an emergent property of a highly interconnected and highly non-linear system. In particular, research has shown that the response of early sensory processing cortical units can be non-linearly affected by intra-area connections, as well as by global or non-localized brain states such as perceptual attention, task-oriented behavior, or otherwise non-specific “house-keeping” or “ongoing” activity [1]. Moreover, performance in relatively simple tasks such as face recognition involves the coordination of vast neuronal networks, spanning almost the entire brain [2].

Even more than modeling, analysis of neurophysiological data has been largely constrained to linear, low-dimensional approaches due to a number of reasons: (1) computational intractability of non-linear approaches, such as Independent Component Analysis, for high-dimensional data, (2) experimental infeasibility, in terms of limited statistical sampling for non-linear kernel reconstruction, such as for the Volterra expansion method [6], (3) limited spatial and/or temporal resolution of the data. In the case of functional neuroimaging, the General Linear Model (GLM) method has been applied with success to identify the activation of different areas in response to stimulation, and in correlation with perceptual, motor and cognitive tasks, as well as various other brain conditions such as emotional and attentional states. This model is not only obviously limited to linear mappings between regressor and voxel activation, but it also glosses over potentially discoverable relationships between different voxels or groups of voxels, which may even be independent or not linearly correlated with the condition at hand.

In order to overcome these limitations, several models have been developed in recent years to overcome the co-dependence/non-linearity limitations, including heuristics for binarization of full-voxel network reconstruction [10, 4], and more principled approaches, based on statistical causality measures [7], but only implementable on small number of voxels. Here we present a scalable method to compute the structure of causal links over large scale dynamical systems that achieves high efficiency in discovering actual functional connections. The method is based on the Granger causality analysis of multivariate linear models, solved by means of a sparse regression approach.

## 2. GRANGER CAUSALITY

Granger causality is defined in terms of predictability of stochastic processes. A process  $X_i$  is said to have a causal influence on another process  $X_j$  if *predictability* of  $X_i$  at a given time instant can be improved using the past values of  $X_j$ .

Formally, let  $X$  represent a vector of  $N$  stationary stochastic processes. Let  $X(t)$  represent the  $N$  dimensional vector of random variables at time  $t$  with  $X_i(t)$  as its  $i$ -th component. Let  $\overline{X}(t)$  represent the set of past random variables in  $X$  i.e.,  $\overline{X}(t) = \{X(t-j) : j = 1, 2, \dots, \infty\}$ . Let  $P(A|\overline{B})$  represent the optimal unbiased least-square predictor of the random variable  $A$  using only the random variables in the set  $\overline{B}$ . Thus  $P(X_i(t)|\overline{X}_i(t))$  represent the optimal unbiased least-square predictor of  $X_i(t)$  using only the past values of  $X_i(t)$ . Let  $\sigma^2(A|\overline{B})$  be the variance of  $A - P(A|\overline{B})$ .

The process  $X_j$  is said to have a causal influence on the process  $X_i$  in the context of processes  $X$ , if

$$\sigma^2(X_i(t)|\overline{X}(t)) < \sigma^2(X_i(t)|\overline{X}(t)\setminus\overline{X}_j(t))$$

In general, it is extremely difficult to determine Granger causality (in its purest form) because computing optimal unbiased least-square predictor for arbitrary stochastic processes is a non-trivial task. In practice, Granger causality has been applied in the context of *linear models*.

### 2.1. The Linear Simplification

In the linear simplification, the multivariate stochastic process  $X$  (represented as a  $N$  dimensional row vector) is modeled as a linear combination of its past values and independent, identically distributed (iid) noise. Such representation is also called a multivariate autoregressive model. Formally

$$X(t) = \sum_{\tau=1}^k X(t-\tau)A(\tau) + E(t) \quad (1)$$

where  $k$  is called the *model order*,  $A(\tau)_{\tau=1\dots k}$  are the model parameters in the form of  $k$  matrices of size  $N \times N$  (with coefficients  $a_{ij}(t)$ ),  $E(t)$  is a  $N$ -dimensional row vector of noise with zero mean and a covariance equal to  $R$ . For any  $t_1 \neq t_2$ ,  $E(t_1)$  and  $E(t_2)$  are identically distributed and uncorrelated.

In this model, if  $a_{ij}(t) > 0$  for some  $t$ , then past values of  $X_i$  improve the predictability of  $X_j$  and therefore,  $X_i$  is said to have causal influence on  $X_j$ . The parameter  $t$  is called the causality lag between  $X_i$  and  $X_j$ .

To infer the causal relationships in the linear simplification, we need to know the model parameters  $\{a_{ij}(t)\}$ . These may be estimated from a realization of the process  $X$ . Let  $\{x(t)\}_{t=1\dots T}$  be a realization of the stochastic process  $X$  and  $\{e(t)\}_{t=1\dots T}$  be a realization of the iid noise  $E$ . This realization must satisfy

$$x(t) = [x(t-1) \dots x(t-k)] [A'(1) \dots A'(k)]' + e(t) \quad (2)$$

for all  $t \in [k+1, \dots T]$ . The above set of equations can be written in a compact matrix form as follows. Let  $Y$  be a matrix of size  $(T-k) \times N$ ,  $Z$  be a matrix of size  $(T-k) \times Nk$ ,  $W$  be a matrix of size  $Nk \times N$  and  $\mathcal{N}$  be a matrix of size  $(T-k) \times N$  obtained by stacking the rows in (2) for  $t = T-k+1$  to  $t = T$ . Now, Eq. 2 may equivalently be written as

$$Y = ZW + \mathcal{N} \quad (3)$$

where  $Y$  and  $Z$  are derived from a realization  $x$  of the process  $X$ ,  $W = [A'(1) \dots A'(k)]'$  contains all the model parameters ( $a_{ij}(t)$ ) and  $\mathcal{N}$  is derived from realization  $e$  of the noise  $E$ .

The maximum likelihood estimate ( $W_{MLE}$ ) of model parameters is given by the standard least square solution of Eq. (3) i.e.,

$$W_{MLE} = \arg \min_W \sum_{j=1}^N \|Y_j - ZW_j\|_2^2 = \quad (4)$$

$$\arg \min_{a_{ij}(\tau)} \sum_{j=1, t=k+1}^{N, T} \left[ x_j(t) - \sum_{\tau=1}^k \sum_{l=1}^N x_l(t-\tau) a_{lj}(\tau) \right]^2 \quad (5)$$

where  $Y_j$  represents  $j^{th}$  column of  $Y$  and  $W_j$  represents the  $j^{th}$  column of  $W$ . Eq. (4) has a unique solution only if (3) is not under-determined, i.e.,

$$(T-k)N \geq N^2K \Rightarrow T \geq (N+1)k$$

In general, for reliable estimates of the model parameters, Eq. (3) must be sufficiently overdetermined, i.e., the number of observations of the process  $X$  must be significantly larger than the number of model parameters ( $(T-k)N \gg N^2k$ ).

If the model is sparse, i.e., the number of non-zero coefficients in  $\{a_{ij}(\tau)\}$  is significantly smaller than the total number of coefficients ( $Nk$ ), then it might be possible to find a reliable solution to (3) using techniques of sparse regression.

## 2.2. Sparse Regression

Consider a multivariate linear regression model of the form  $Y = ZW$  where  $Y$  is a known  $n_1 \times 1$  response vector,  $Z$  is a known  $n_1 \times n_2$  regressor matrix and  $W$  is the unknown model vector of size  $n_2 \times 1$  to be determined using the response  $Y$  and regressor  $Z$ . The usual techniques to solve this are ordinary least square regression [11], ridge regression and subset selection. For these techniques, it

is usually required to have  $n_1 \gg n_2$ . However, there is a growing body of work indicating that if  $W$  is sparse then it may be recovered even if  $n_2 > n_1$  using the lasso regression [9].

The lasso regression [9] solves the problem

$$\min_W \|Y - ZW\|_2^2 \quad (6)$$

$$\text{s.t. } \|W\|_1 \leq t \quad (7)$$

where  $\|\cdot\|_2^2$  represents the square of L2 norm and  $\|\cdot\|_1$  represents the L1 norm of the respective vectors. The parameter  $t$  is the regression parameter usually chosen after cross-validation.

It can be verified that for any  $t$ , there exist a  $\lambda$  such that the program (6, 7) is equivalent to the following optimization problem:

$$\min_W \|Y - ZW\|_2^2 + \lambda \|W\|_1 \quad (8)$$

The programs (6, 7) and (8) can be solved efficiently using a technique called least angle regression [3] in time no longer than time required to carry out the ordinary least square computation.

The estimation of multivariate autoregressive coefficients in (3) may be viewed as a regression problem where  $Y$  is the response variable,  $Z$  is the matrix containing the regressors and  $W$  is the model to be determined. In this case, the maximum likelihood estimate of (4) becomes the least square solution to the regression problem. The lasso formulation thus becomes

$$W^{sparse} = \arg \min_W \sum_{j=1}^N [\|Y_j - ZW_j\|_2^2 + \lambda \|W_j\|_1] \quad (9)$$

Note that the coefficients of  $W_j$  only appear in the  $j^{th}$  term of the above sum. So, this problem may be decomposed into  $N$  independent lasso regression problems of size  $(T-k) \times Nk$  as follows

$$W_j^{sparse} = \arg \min_{W_j} \|Y_j - ZW_j\|_2^2 + \lambda \|W_j\|_1$$

The goodness of fit of the regression is captured using the notion of *predictability* ( $p_j$ ), which is defined as  $p_j = 1 - Q \sum_{t=k+1}^T [x_j(t) - \sum_{\tau=1}^k \sum_{l=1}^N x_l(t-\tau) a_{lj}(\tau)]^2$  where  $Q = [\sum_{t=k+1}^T (x_j(t))^2]^{-1}$ . It may be verified using the properties of the lasso regression that the predictability varies from 0 to 1. If the predictability of a voxel is 1 then its time course can be predicted exactly using the past  $k$  values of other voxels. On the other hand, if a voxel has zero predictability then its time course is orthogonal to (independent of) the shifted time course of all the other voxels.

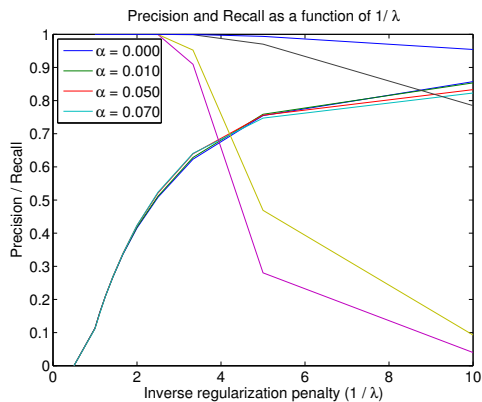
## 3. RESULTS AND DISCUSSION

### 3.1. Simulations

We carry out simulations of sparse multivariate autoregressive models (MAR) to generate its realizations. The parameters of the MAR model and the size of its realization were chosen to reflect the sizes of typical fMRI data sets.

We then use the lasso regression to estimate model parameters (as discussed in Section 2.2). We compare the estimated parameters with the actual parameters of the MAR model.

Let  $\{a_{ij}(t)\}$  be the parameters of the MAR model and  $\{\hat{a}_{ij}(t)\}$  be the model parameters estimated by the regression. Let  $S(\alpha) = \{(i, j, t) : |a_{ij}(t)| > \alpha\}$  and  $\hat{S}(\alpha) = \{(i, j, t) : |\hat{a}_{ij}(t)| > \alpha\}$ . We use the following metric for our evaluation:



**Fig. 1.** Precision and recall as a function of  $\lambda$

**Precision.** It is defined as the ratio of numbers of true non-zero coefficients estimated to the total number of non-zero coefficients estimated. Formally precision  $p = |S(0) \cap \hat{S}(0)| / |\hat{S}(0)|$ .

**Recall.** It is defined as the ratio of the number of true non-zero coefficients estimated to the total number of non-zero coefficients present in the model. Formally recall  $r = |S(0) \cap \hat{S}(0)| / |S(0)|$ .

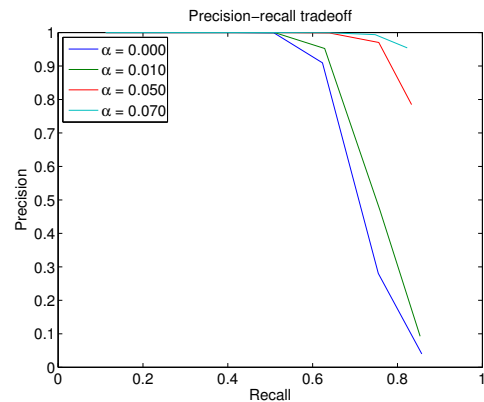
**Thresholding.** Generally, it is more important to discover the causal relationships that are strong. This can be done by considering only the coefficients that are above a given threshold. The precision and recall may respectively be defined with respect to a threshold  $\alpha$  as  $p(\alpha) = |S(\alpha) \cap \hat{S}(\alpha)| / |\hat{S}(\alpha)|$ ,  $r(\alpha) = |S(\alpha) \cap \hat{S}(\alpha)| / |S(\alpha)|$ .

**Correlations.** We use two measures of Pearson correlations between estimated and actual model parameters. The first measure  $c_{true}$  measures the correlations between  $a_{ij}(t)$  and  $\hat{a}_{ij}(t)$  over the true non-zero coefficient estimates (i.e. over *true positives* only). The second measure  $c_{non-zero}$  corresponds to the correlations between actual and estimated parameters over *true positives*, *false positives* and *false negatives*. Formally,  $c_{true} = \langle a_{ij}(t), \hat{a}_{ij}(t) \rangle_{(i,j,t) \in S(0) \cap \hat{S}(0)}$  and  $c_{non-zero} = \langle a_{ij}(t), \hat{a}_{ij}(t) \rangle_{(i,j,t) \in S(0) \cup \hat{S}(0)}$ .

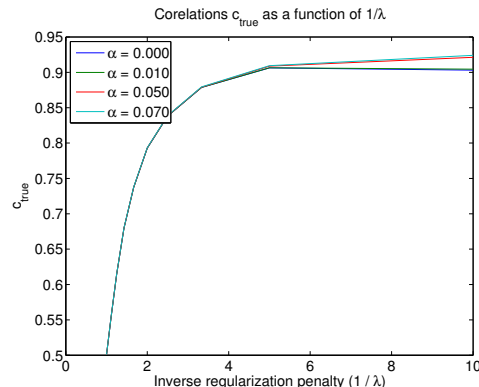
We generated  $k$  random sparse graphs with 10,000 vertices and 50,000 edges. Edges in these graphs were assigned random weights normally distributed with zero mean and unit variance. A MAR(k) model was constructed using these graph. The edge weights were used to define the coefficients  $\{a_{ij}(t)\}$  of the MAR(k) model. A realization of the MAR process was obtained using iid noise with zero mean and identity covariance matrix. If the MAR process was not convergent, weights were scaled down uniformly until the process became convergent. We obtained 500 time points of the realization. This realization was used to estimate the model parameters using the lasso regression. We report results only for  $k = 1$  in this paper.

### 3.2. Results

Figure 1 shows the precision and recall curves as a function of the lasso regularization penalty  $\lambda$  (see (8)). As the penalty is increased, the lasso regression estimates fewer non-zero coefficients. This improves the precision (i.e., the coefficients estimated to be non-zero are more likely to be non-zero) at the cost of recall (i.e., many non-zero coefficients are estimated to be zero). Figure 2 shows the precision-recall trade-off for different thresholds. It is evident



**Fig. 2.** Precision-recall tradeoff



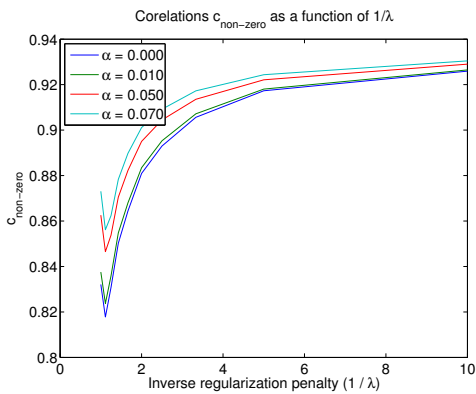
**Fig. 3.** Correlations ( $c_{true}$ ) as a function of  $\lambda$

from these figures that as the threshold increases the precision and recall improve. Thus, the regression consistently makes less error in estimating larger MAR coefficients.

The precision and recall measures only indicate the accuracy in estimating whether a coefficient is non-zero or not. These measures do not convey any information about the accuracy in the values of the MAR coefficients. The correlation measures indicate the accuracy in estimating MAR coefficient values. Figure 3 shows the  $c_{true}$  as a function of regularization penalty  $\lambda$  for different thresholds. For all thresholds,  $c_{true}$  increases as  $\lambda$  is increased. This indicates that higher regularization penalty ( $\lambda$ ) results in fewer non-zero estimates, but with better accuracy. This measure might be a bit misleading because it is only based on true positives and does not consider false positives and false negatives. The measure  $c_{non-zero}$  considers all of these and is plotted in Figure 4 as a function of  $\lambda$ . Note that with a suitable choice of  $\lambda$ , the correlations can be made as high as 0.9. This demonstrates that MAR coefficients (and hence the causality relationships) may be inferred with reasonable degree of accuracy using the lasso regression.

### 3.3. Analysis of fMRI Data

The Granger causality analysis gives  $k$  connectivity matrices  $A(\tau)$  and a map  $[1 \dots N] \rightarrow [0, 1]$  indicating the predictability of voxels in brain using the past values of other voxels. Initial analysis of fMRI



**Fig. 4.** Correlations ( $c_{non-zero}$ ) as a function of  $\lambda$

data points to the hypothesis that predictability is a good indicator of brain activations.

It is very difficult to ascertain the “ground truth” relating to brain activations by using non-invasive measures such as fMRI. The most common approach to analyze fMRI data uses the general linear model (GLM) [5] which identifies activations in brain regions as a response to external stimuli represented in the form of a design matrix. By definition, GLM leaves out activations that are not linearly correlated with the design matrix. If the predictability is an indicator of brain activity, the regions found active by GLM analysis must have high predictability. Moreover, one may also find regions of high predictability that are not found to be active by GLM analysis. Our initial analysis suggests that this is indeed the case.

To formalize the above observation, we define a measure called *weighted coverage* of a set  $S_1$  of voxels in space by another set  $S_2$ . First, the voxels in the two sets are clustered into groups of spatially connected 3-d components (two voxels are said to be connected if they share a common face, i.e. we are using 6-connectedness in three dimensions). The weighted coverage of  $S_1$  using  $S_2$  is defined as ratio of the weight of clusters of  $S_1$  that intersect with clusters of  $S_2$  to the total weight of clusters of  $S_1$ . Clearly, the weighted coverage ranges from 0 to 1. A weighted coverage of zero implies that  $S_1$  and  $S_2$  have no voxels in common. Large coverage indicates that many connected components of  $S_1$  intersect with connected components in  $S_2$ . A coverage of one implies that every connected component in  $S_1$  intersects with a component in  $S_2$ .

In order to select the voxels that form set  $S_1$  or  $S_2$ , we select the top  $p$ -percent of active voxels given by the predictability maps on the finger-tapping data-set considered in [4]. We also use the voxels given by the GLM maps using the same top  $p$ -percent of active voxels. We used  $p = 1.25\%$  to compute  $S_1$  and  $S_2$ , and performed 3d connected component analysis on these voxels. The coverage of GLM maps by the predictability maps has a mean of 82% and standard deviation (across subjects and conditions) of 16%. On the other hand the coverage of predictability maps by the GLM maps has a mean of 41% and standard deviation of 30%. This confirms our hypothesis that the predictability maps discover most of the active regions found by the GLM analysis along with many other regions that are not found by the GLM analysis.

### 3.4. Discussion

Our work represents a significant improvement over previous approaches based on causality and sparse regression; in particular, our method can be scaled up by a factor of 100 over recent work by Valdes-Sosa et al. [8], albeit with the use of massively parallel environments (analysis was implemented in one rack of IBM’s BlueGene supercomputer). Our results imply that it is possible to study the full extent of the internal structure of the brain’s dynamics - as measured by fMRI - without analytic compromises. Applications of the method to task and dysfunction related experimental conditions are currently underway, and will be reported in coming publications.

## 4. REFERENCES

- [1] A. Arieli et al. Dynamics of Ongoing Activity: Explanation of the Large Variability in Evoked Cortical Responses. *Science*, 273(5283):1868–1871, 1996.
- [2] E. Rodriguez et al. Perception’s shadow: long-distance synchronization of human brain activity. *Nature*, 397(6718):430–3, 1999.
- [3] B. Efron et al. Least angle regression. *Ann. Statist.*, 32(1):407–499, 2004.
- [4] G.A. Cecchi et al. Identifying directed links in large scale functional networks: application to brain fMRI. *BMC Cell Biology*, 8(Suppl 1:S5), 2007.
- [5] K. J. Friston et al. Statistical parametric maps in functional imaging - a general linear approach. *Human Brain Mapping*, 2:189–210, 1995.
- [6] K.J. Friston and C. Büchel. Attentional modulation of effective connectivity from V2 to V5 in humans. *Proc. Natl. Acad. Sci. USA*, 97:7591–7596, 2000.
- [7] L.A. Baccal, K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.*, 84(6):463–74, 2001.
- [8] P.A. Valdes-Sosa et al. Estimating brain functional connectivity with sparse multivariate autoregression. *Philos Trans R Soc Lond B Biol Sci.* 2005 May 29;360(1457):969–81, 360(1457):969–81, 2005.
- [9] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [10] V.M. Eguiluz et al. Scale-free functional brain networks. *Phys. Rev. Lett.*, 94(018102), 2005.
- [11] S. Weisberg. *Applied Linear Regression*. wiley, New York, 1980.