Statistica Sinica **18**(2008), 1501-1518

# UNCOVERING SPARSE BRAIN EFFECTIVE CONNECTIVITY: A VOXEL-BASED APPROACH USING PENALIZED REGRESSION

José M. Sánchez-Bornot, Eduardo Martínez-Montes, Agustín Lage-Castellanos, Mayrim Vega-Hernández and Pedro A. Valdés-Sosa

*Cuban Neuroscience Center*

*Abstract:* The processing of massive data generated by bioinformatic and neuroscience studies is a current challenge to statisticians since they require the development of computationally efficient and stable algorithms that can deal with many more variables than observations. In neuroscience, a clear example of this situation is the estimation of brain physiological interactions through the analysis of fMRI time series. The widespread use of the General Linear Model in the resolution of these problems has now been enhanced by the addition of prior assumptions, such as the sparseness and/or the spatiotemporal smoothness of a desirable solution (Valdes-Sosa (2004)). In this context, the use of Local Quadratic Approximation (LQA) (Fan and Li (2001)) and the Minorization-Maximization (MM) Hunter and Li (2005)) algorithms are practical ways for estimating the sparse models. Recently, we have extended these techniques to allow the combination of these attractive properties (Valdes-Sosa et al. (2006)). Here, we further formalize the methods and introduce a feature selection algorithm for feasible implementation. The methodology is then applied to the estimation of voxel-based brain effective connectivity using simulated and neuroimaging data.

*Key words and phrases:* Brain connectivity, FDR, feature selection, fMRI.

## 1. Introduction

There is consensus in the Neuroscience Community that understanding the neural substrates of cognitive processes (language, object recognition, working memory, motor planning, awareness, etc.) will involve much more than just localizing the "functions" subserved by specific brain regions. The objective rather is to identify the widespread and continuously changing networks of neuronal populations that transiently engage with each other to carry out computations. Thus, the problem of estimating brain connectivity has become one of the main current methodological problems in brain research. There are several excellent reviews on this expanding area (Lee, Harrison and Mechelli (2003), Bullmore et al. (2004) and Jirsa and McIntosh (2007)) and, particularly, a project that establishes the study of the "Human Connectome" as a means for uncovering the

structural and functional determinants of brain activity through the combined study of anatomical, functional and effective connectivity (Sporns, Tononi and Kotter (2005)). This fortells an immediate increase in connectivity research as well as community agreement in the development of neuroinformatic databases and statistical tools (Kotter (2001) and Stephan et al. (2001)).

Methods for brain connectivity can be distinguished on the basis of their dealing with the structural (anatomical) or physiological interrelationship between their neuronal populations. Anatomical connectivity studies of the human brain have been guided by investigation in primates, using invasive tracing experiments for accurate fiber tracking inside the brain (Buchel and Friston (1997)). The emergence of non-invasive techniques for studying the brain connectivity in vivo, in particular Diffusion Tensor Imaging (DTI), have allowed the characterization of the major pathways in the human brain (Le Bihan (2003)). However, the insufficient spatial resolution of DTI and some limitations in data acquisition and processing algorithms have hindered the creation of human brain anatomical maps based on DTI data (see however Iturria-Medina et al. (2007)).

Similarly, functional Magnetic Resonance Imaging (fMRI) has been increasingly used to study the principles of functional integration and segregation. Evidence supports that these phenomena have origin in the patchy and selectivity pattern of fibers linking cortico-cortical and thalamo-cortical regions (Friston, Ashburner, Stefan, Thomas and Penny (2006)). In the absence of an accurate structural mapping of these associations, functional and effective connectivity have become the target of novel statistical tools for exploring these connections and their relationships with functional specialization (Table 1).

Table 1. Modes of Brain Connectivity.

| Brain connectivity | | Techniques |
|---|---|---|
| Anatomical | | Lesion studies using manganese (invasive). Diffusion Wighted Imaging (In vivo) |
| Physiological | Functional (Association) Effective (Causality) | EEG, fMRI, Optical Imaging, Transcranial Magnetic Stimulation (TMS) |

A major tool for studying physiological connectivity is fMRI, which is based on the Blood Oxygenation Level Dependent (BOLD) contrast, which in turn is based on the oxygen consumption of the active neuronal population and the magnetic properties of the deoxyhaemoglobin and haemoglobin contained within the red blood cells (Ogawa and Sung (2007)). A temporary increase in neuronal activity produces a smoothed and retarded function known as the haemodynamic response function (HRF) (Heeger and Ress (2002)). This signal is characterized

on some occasions by an initial dip because oxygen is rapidly consumed by active neurons and, in consequence, increases the relative level of deoxyhaemoglobin in the blood. After that, an oversupply of oxygenated blood causes a maximum point in the signal at six seconds after the initial neuronal activity. Finally, in absence of subsequent neuronal activity, the oxygen concentration and the HRF signal return to their baselines (Figure 1).

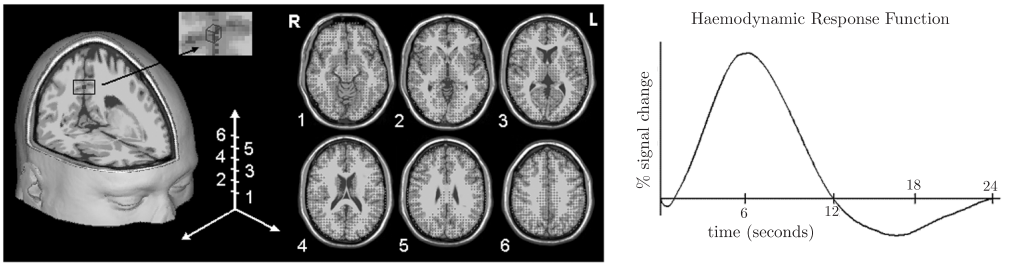


Figure 1. Left: discretization of the brain into voxels as units to reflect the brain activity. Right: the HRF signal produced by transient activity in neuronal populations.

Although fMRI signals are an indirect result of neuronal activations (Logothetis et al. (2001)) and Heeger and Ress (2002)), they has been established as an indispensable technique for brain imaging (Friston et al. (2006) and Penny and Friston (2007)). The spatial resolution achieved by this technique at 1.5 tesla is 2-5 mm, which is equivalent to having a 3D segmentation of the brain volume with cubes (voxels) of these dimensions. A BOLD measure for a time instant t is recorded at each voxel. This vector is known as a scan or image. In a typical experiment, we have $N$ equal to hundreds of scans, with a time resolution of 1-4 seconds, producing a time series vector for each voxel. The problem is that the number of voxels ($p$) is of the order of hundreds of thousands. Thus, neuroimaging together with bioinformatics are the two main driving forces requiring statistical inferences for $p \gg N$ with $p \to \infty$.

Anatomical, functional and effective connectivity must be gathered from the context of functional imaging as the structural and dynamical relationship between the neuronal populations. In this context brain connectivity analyses can be conducted using a voxel-based strategy, dealing with the voxels' time series, or between segmented Regions Of Interest (ROIs) selected by specialists attending to prior knowledge, where the ROIs' dynamics is estimated by combining the interior voxels' recordings. The latter serves the purpose of dimensionality reduction and focuses the attention on specialized structures that are supposedly involved in some cognitive processes (Figure 2).

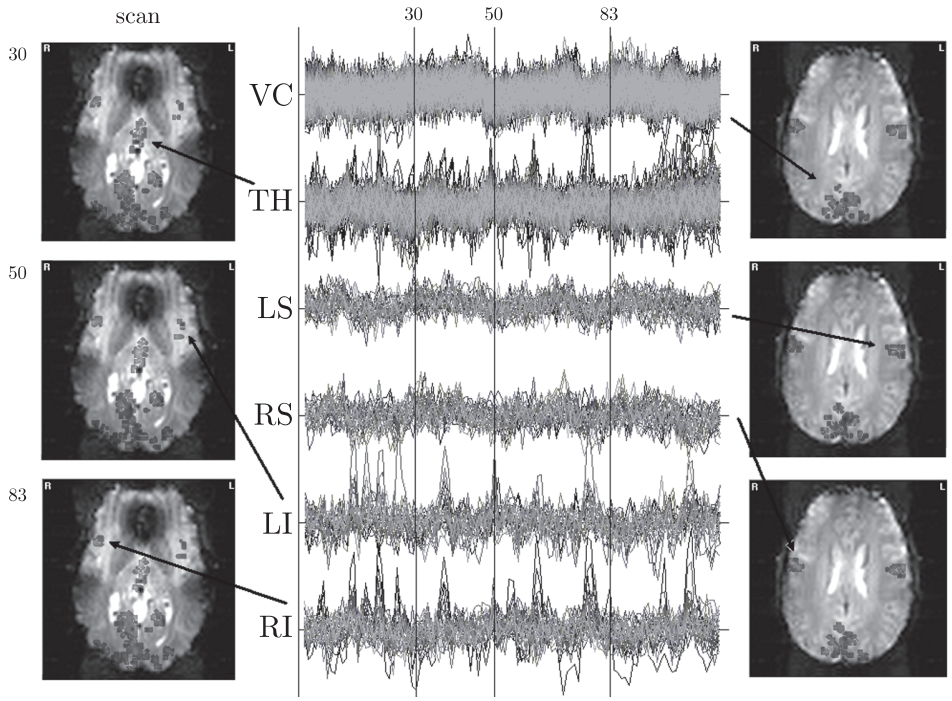


Figure 2. Time series for six preselected ROIs. VC: Visual Cortex, 647 (voxels); TH: Thalamus, 177; LS: Left Somatosensory area, 29; RS: Right Somatosensory area, 31; LI: Left Insula, 28; RI: Right Insula, 33. At both sides are shown ROIs' activity for scans 30, 50 and 83 for a superior (right) and inferior (left) horizontal view.

On the other hand, functional connectivity must be understood as the statistical association between spatially remote neurophysiological events, and can be quantified using covariance or mutual information analyses. This is frequently done in the exploratory mode. Instead, effective connectivity is concerned with the causal influence that neuronal populations exert over others (Friston (1994)). Several algorithms are reported in the literature for the assessment of effective connectivity, in general they estimate the weights of the structural connections assumed in a model, based on a priori selection of ROIs, or guided by anatomical investigation in primates (Buchel and Friston (1997) and Jirsa and McIntosh (2007)). Some explorations of effective connectivity using the fMRI time series deserve special mention: a study in the frequency domain (Salvador et al. (2005)) and the Multivariate Autoregressive (MAR) models (Harrison, Penny and Friston (2003) and Valdés-Sosa et al. (2005)) that directly approach temporal causation in time series, and the Dynamic Causal Modeling (DCM) technique (Friston,

Harrison and Penny (2003)), which models the hidden neural activity that originates the haemodynamic response based on intrinsic and extrinsic connections and neural dynamics. Moreover, MAR models have been used for focal exploration of effective connectivity using voxel-based time series, which constitutes a current challenge for statistical data analysis (Valdés-Sosa et al. (2006)).

The use of Graph Theory is a convenient way to characterize brain connectivity. The anatomical information could be embodied in matrix format with binary entries representing the presence or absence of a structural connection linking two regions. In functional terms, connectivity is represented as a weighted matrix, with entries representing densities or strength of the connections between the regions. Functional connectivity graphs are rendered by symmetric matrices and effective connectivity's are produced by non-symmetric ones. When the subjacent structural model is barely known, effective connectivity could also be used in an exploratory fashion. In this situation, a thresholded version of the latter produces a rough estimation of the structure and directions of the physical connections. The graph elements, known as nodes and edges, represent ROIs (or voxels) and "paths" respectively. Paths are an abstract entity that stand for the anatomical or communication links (synapses or statistical dependencies between the nodes' dynamics) (Figure 3).

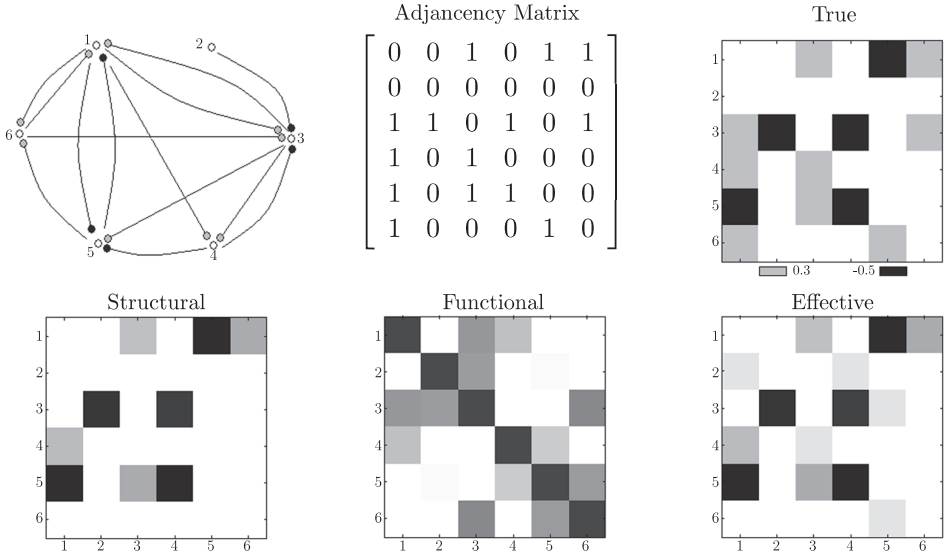


Figure 3. Graphical and matrix representation of a connected network (top row). Nodes represent voxels or ROIs and edges stand for bundles of fibers (anatomical) or statistical dependencies between the constituent elements. An edge linking $i$ and $j$ with a filled circle at the tip of $i$, indicates influence of $j$ onto $i$. In the bottom row, matrices of structural, functional and effective connectivity are shown.

The multivariate autoregressive (MAR) modeling of the nodes' time series allows the computation of the network weights and must be understood as the estimation of effective connectivity. Autoregressive models take into account the causation of past activations on current dynamics, addressing the temporal aspects of causality in time series. Thus, the activation at time instant $t$, derived from the BOLD signal in fMRI, is modeled as the linear combination of the $N_k$ previous activations vector:

$$z_t = \sum_{k=1}^{N_k} A_k z_{t-k} + e_t,\ t = N_k + 1, \ldots, N_t.$$

The fMRI activations for the $p$ nodes at time instant $t$ are collected in $z_{t\ (p\times 1)}$ and the white noise input $e_t \sim N(0, \mathbf{\Sigma})$ is introduced for modeling spatial correlation between nodes. The matrices of autoregressive coefficients $\mathbf{A}_{k\ (p\times p)} = \{a_{ij}^k\}_{1\le i,j\le p}$ for the different time lags, $k = 1, \ldots, N_k$, represent the influence that node $j$ exerts on node $i$ after $k$ time instants. In general, a row (column) of $\mathbf{A}_k$ corresponds to the influence field of ingoing (outgoing) connections from (to) the other nodes in the graph, and this can be represented as the corresponding rows (columns) of the topologies shown as matrices in Figure 3. Setting

$$\mathbf{Y} = \begin{bmatrix} z_{N_k+1}^T \\ \vdots \\ z_{N_t}^T \end{bmatrix}, \mathbf{X} = \begin{bmatrix} z_{N_k}^T & \cdots & z_1^T \\ \vdots & \ddots & \vdots \\ z_{N_t-1}^T & \cdots & z_{N_t-N_k}^T \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \mathbf{A}_1^T \\ \vdots \\ \mathbf{A}_{N_k}^T \end{bmatrix} \text{ and } \mathbf{E} = \begin{bmatrix} e_{N_k+1}^T \\ \vdots \\ e_{N_t}^T \end{bmatrix},$$

we can recast the dynamics of the network as a multivariate regression model

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

where $\mathbf{Y}$ is $(N_t - N_k)\times p$, $\mathbf{X}$ is $(N_t - N_k)\times pN_k$, and $\mathbf{B}$ is $pN_k \times p$. In this formula, each row of $\mathbf{Y}$ corresponds to a typical scan of the fMRI and the columns contain the time series for each node.

Solving MAR models for a large $p$ is a current challenge in statistics. In fMRI this corresponds to the number of voxels, which can be as many as 50,000. Currently it is common to estimate effective connectivity based on ROIs. However, omission or cancellation of activity due to this strategy can lead to miss-estimation or bias. In this paper, we argue that the use of methods that deal with huge dimensionalities through exploiting appropriate priors is a valid and useful procedure to surmount these difficulties. In this case, realistic priors are those reflecting the patchy pattern of inter-area connectivity (smoothness) and the selectivity of the connections between areas (sparseness).

The Bayesian framework and penalized least squares regression are two of the most appealing approaches. In the latter, Fan and Li (2001) proposed the use of non-convex functions (continuous and with a singularity at the origin) as penalties, in order to guarantee the estimation of sparse and stable solutions. For estimating models involving non-convex penalties, the Local Quadratic Approximation (LQA) (Fan and Li (2001) and Fan and Peng (2004)) and the Minorization-Maximization (MM) (Hunter and Li (2005)) algorithms have been introduced. They use quadratic approximation in order to apply a modified Newton-Raphson (MNR) technique and approximate the solution in successive steps. Therefore, LQA and MM share the convergence properties of the MNR algorithm, using a robust local quadratic approximation. In both cases, the Hessian matrix is guaranteed to be positive definite, driving convergence at least to a local minimum.

Recently, our group has used the penalized least squares approach in the analysis of brain functional connectivity (Valdés-Sosa et al. (2005)). This was assessed using a two-step process involving (i) penalized regression using the MM algorithm, then (ii) edge removal using the false discovery rate (FDR) to detect spurious connections. The FDR (Benjamini and Hochberg (1995)) allows one to remove statistically non-significant connections, highlighting the results and the interpretability.

On the other hand, combinations of penalty functions have also been explored in the literature. The Elastic Net method (Zou and Hastie (2005)) proposes a penalty that is a combination of $l_1$ an $l_2$ norms, rendering solutions where covariates having high correlation are selected from background. The Fused Lasso (Tibshirani et al. (2005)) uses a Lasso modification that penalizes differences between coefficients to produces sparseness patches as solutions. In general, this family of regression methods can be joined in a common framework, and a generalization of LQA and MM algorithms can be used for parameter estimation (Valdés-Sosa et al. (2006)).

In this paper, we put forward the combined use of several penalties in an approach called the Multiple Penalized Least Squares (MPLS) method. This is presented in Section 2. Moreover, we introduce an algorithm based on the mixture of a feature selection strategy with the generalized LQA and FDR techniques. We argue that this methodology allows us to deal with the high-dimensional fMRI data, as in the estimation of voxel-wise connectivity as an alternative to the use of ad-hoc ROI analysis. In Section 3 this algorithm, named LQA-Fext, is explained and applications to simulated and actual neuroimaging connectivity data are presented in Section 4. Finally, some considerations and recommendations are posed in the conclusion.

## 2. The MPLS Model

Consider the multivariate linear regression model

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E},$$

where $\mathbf{Y}$ is an $N \times q$ matrix of responses or observations, $\mathbf{X}$ is the $N \times p$ design matrix, and $\mathbf{B}$ is the $p \times q$ matrix of coefficients to be estimated. As usual, rows of the residual matrix $\mathbf{E}$ are settled in the model as independent samples from a $q$-variate normal distribution $N(0, \mathbf{\Sigma})$, with finite covariance matrix $\mathbf{\Sigma}$. Assume that response columns and covariates (columns of $\mathbf{X}$) have been centered, so the intercepts are excluded from the regression model.

The least squares solution to this model is obtained by the minimization of the functional $f(\mathbf{B}) = trace\{(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})\mathbf{\Sigma}^{-1}\}$, and can be analytically expressed as $\mathbf{B} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$.

The solution assumes that $\mathbf{X}$ is of full column rank, so $\mathbf{X}^T\mathbf{X}$ is invertible. However, this is not a feasible solution if we are dealing with an ill-conditioned problem where $\mathbf{X}^T\mathbf{X}$ is almost singular or, in the case of $p \gg N$, when this is a singular matrix. To avoid this situation, it is convenient to make assumptions on the shape of the coefficients. These can be imposed applying hard-edge or soft-edge constraints on the matrix $\mathbf{B}$ (Ramsay and Silverman (1997)). The use of soft edge constraints is equivalent to adding a penalty term to the least squares fitting criterion. This penalization produces the estimator

$$\widehat{\mathbf{B}} = \underset{\mathbf{B}}{\operatorname{argmin}} \left\{ trace\left( (\mathbf{Y}-\mathbf{XB})^T (\mathbf{Y}-\mathbf{XB}) \right) + \lambda \Psi(\mathbf{B}) \right\},$$

where $\Psi$ is a scalar penalization function that imposes some constraints on the space of coefficients, while the positive scalar $\lambda$ quantifies the relative importance between the error fitting term and the term of constraints. Note that $\mathbf{\Sigma}$ is omitted since its influence can be modeled by the penalty term. The simplest case is Ridge regression (Hoerl and Kennard (2000)), which corresponds to the penalty $\Psi(\mathbf{B}) = trace\left(\mathbf{B}^T\mathbf{B}\right)$, producing the solution

$$\mathbf{B} = \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

The ridge regression technique is frequently used to obtain stable estimators. The penalty term can also be enriched with smoothness constraints on $\mathbf{B}$, for example to produce smooth variations of its rows or columns. Smoothness can be required on rows of $\mathbf{B}$ by using $\Psi(\mathbf{B}) = trace\left(\mathbf{B}^T\mathbf{L}^T\mathbf{LB}\right)$, where $\mathbf{L}$ is the prior information matrix, usually taken as the first or second order difference operator (Valdés-Sosa et al. (2005)). In both cases, estimators can also be obtained by

solving the equation separately for each column of $\mathbf{Y}$. However, this cannot be done if constraints are applied on columns of $\mathbf{B}$, since the problem becomes non-separable, leading to a more computationally difficult case that we do not detail here. The MPLS model for a single column of $\mathbf{B}$ can be stated as

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \sum_{k=1}^{N_k} \lambda_k \Psi_k (\beta) \right\},$$

with a mixture of $N_k$ penalty terms that can impose different kind of constraints through $\Psi_k(\beta)$. Particular cases of this model involve the LASSO, SCAD, Elastic Net, Fused Lasso and other regularization techniques. For example, the LASSO (Tibshirani (1996)) is represented with the penalty $\Psi(\beta) = \sum |\beta_i| ; i = 1, \ldots, p$, where $\beta_i$ represent the elements of $\beta$; the Elastic Net can be represented in this general model using $N_k = 2, \Psi_1(\beta) = \sum |\beta_i|, \Psi_2(\beta) = \sum \beta_i^2$. Similarly, the Fusion and Fused LASSO can be modeled using MPLS as

**Fusion Lasso:** $N_k = 1, \Psi_1(\beta) = \sum_i |\mathbf{L}(i,:)\beta|$,
**Fused Lasso:** $N_k = 2, \Psi_1(\beta) = \sum_i |\beta_i|, \Psi_2(\beta) = \sum_i |\mathbf{L}(i,:)\beta|$,

where $\mathbf{L}(i,:)$ represents the $i$-th row of prior information matrix $\mathbf{L}$, here taken as the first order difference operator.

The model also captures Ridge regression or more generally, any quadratic function for $\Psi_k$. We analogously define two versions of Ridge that are called the Fusion and Fused Ridge as

**Fusion Ridge:** $N_k = 1, \Psi_1(\beta) = \sum_i (\mathbf{L}(i,:)\beta)^2 = \beta^T \mathbf{L}^T \mathbf{L} \beta$,
**Fused Ridge:** $N_k = 2, \Psi_1(\beta) = \sum_i \beta_i^2 = \beta^T \beta, \Psi_2(\beta) = \beta^T \mathbf{L}^T \mathbf{L} \beta$.

Fusion Ridge and Fusion Lasso will be used as examples of MPLS models to study the physiological interactions between the brain regions' voxels in the application section.

## 3. The LQA-Fext Algorithm

Dealing directly with MPLS regression is computationally infeasible in the presence of a large amount of parameters. Gradient descent algorithms are commonly used in these circumstances in order to avoid the matrix inverse operation, and for exploiting the problem properties. Here, under the assumption that the number of covariates that truly explain the data are much less than the number of parameters, we follow a procedure that basically consists of identifying the relevant features (covariates) and estimating the corresponding coefficients using

direct methods. Specifically, we implement this process with the successive application of the following stages (i) Feature insertion: inclusion as relevant of the "most promising" covariates not included in the subset of salient features; (ii) MPLS parameter estimation using a MNR algorithm on this subset; (iii) Feature extraction: statistical determination of significant features using local FDR, non-significant covariates being removed from the salient features subset.

The objective function to minimize is

$$\begin{aligned} f(\beta) &= RSS(\beta) + \Psi(\beta) \\ &= (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) + \textstyle\sum_{k=1}^{N_k} \lambda_k \Psi_k(\beta). \end{aligned}$$

The purpose is to look for a sub-optimal but computationally feasible solution of this quadratic problem, considering the procedure explained above.

Perkins, Lacker and Theiler (2003) introduced the Grafting technique to automate feature selection in the context of mixing Ridge, Lasso, and subset selection. In this procedure, predictors were divided into two subsets, one containing the active covariates (those corresponding to non-null coefficients), and the Z subset containing the covariates whose coefficients are still zero in the current iteration. At each step of the algorithm, a gradient-based heuristic is applied to select a Z covariate, whose coefficient could be adjusted away from zero in order to render the greatest descent of the objective function. This covariate is then considered active, and a minimization MNR algorithm is applied on the active covariates subset (salient features) until a local minimum is reached. Successive iterations of this procedure are carried out while there are covariates in Z that could be activated.

The feature insertion stage of our procedure uses a modification of the Grafting heuristic in order to select a group of Z covariates at each step. Here, the Grafting technique is reproduced to improve understanding, and a modification to include group of covariates is explained. We start the procedure with the null vector as the initial solution and consider that, after the $k$-th step, the procedure is positioned in a local minimum with respect to the active covariates. Therefore

$$\frac{\partial f}{\partial \beta_j} = 0; \; x_j \text{ is active.}$$

In the feature insertion stage, we heuristically choose a subset of the covariates with corresponding coefficient equal to zero, whose inclusion in the active set could further improve the function minimization. If $x_j$ is a candidate covariate and $\partial f/\partial \beta_j > 0$, then $\beta_j$ is decreased rendering it negative, the contrary if $\partial f/\partial \beta_j < 0$. For the non-negative, continuous, monotone increasing in $(0, +\infty)$,

and symmetric functions considered here, this derivative is (for the case of one penalty function):

$$\frac{\partial f}{\partial \beta_j} = \frac{\partial \mathrm{RSS}}{\partial \beta_j} + \frac{\partial \Psi\left(|\beta_j|\right)}{\partial \beta_j} \mathrm{sgn}\left(\beta_j\right).$$

Thus if $|\partial RSS/\partial \beta_j| > |\partial \Psi/\partial \beta_j|$, $\beta_j$ can be considered as candidate. Instead, if $|\partial RSS/\partial \beta_j| < |\partial \Psi/\partial \beta_j|$, moving $\beta_j$ away from zero leads to a contradiction: setting $\beta_j$ slightly negative produces $\partial f/\partial \beta_j < 0$ and $\beta_j$ must be increased; setting $\beta_j$ slightly positive has $\partial f/\partial \beta_j > 0$ and $\beta_j$ must be decreased.

Non-convex functions of the type considered here are singular at the origin, which means that $\lim_{\beta \to 0}(\partial \Psi\left(|\beta|\right)/\partial \beta) > 0$. By the explanation above, it is clear why these penalty functions produce sparse solutions. On the contrary, the derivative of convex penalty functions is well defined at the origin, with value 0. Thus, when $|\partial RSS/\partial \beta_j| > 0$, covariates can be selected from Z to explain the residual data. Also, it is easy to see why some penalty function combinations, like the Elastic Net, are variable selection algorithms.

Finally, the candidate covariates are defined by $\{x_j : C_j = |\partial RSS/\partial \beta_j| - |\partial \Psi/\partial \beta_j| > 0\}$. Grafting includes in the active set only the covariate corresponding to $\widehat{j} = \mathrm{argmax}_j\left(C_j\right)$. However, in our case, this would result in a slow procedure because the computational advantages of introducing covariates step by step cannot be exploited. Therefore, we select $\{x_j : C_j > 0.5 \max\left(C_j\right)\}$ to be included in the active set.

After the selection of active covariates, parameter estimation is carried out with a variant of LQA, using a MNR algorithm with a robust approximation of the Hessian rendering it positive definite (Li, Dziak, Ma (2006)). In addition, an estimate of the degree of freedom (df), residual sum of squares (RSS), and the coefficients' standard deviation are used in order to compute the generalized cross-validation (GCV) criterion, and Student test, to identify the relevant features. In the simplest case, GCV is computed using a set or range of regularization parameters ($\lambda$):

$$\mathrm{GCV} = \frac{\mathrm{RSS}\left(\mathbf{X}^*, \lambda\right)}{N\left(1 - \frac{df(\mathbf{X}^*, \lambda)}{N}\right)^2},$$

where $\mathbf{X}^*$ means that GCV is estimated on the active set, so the degree of freedom of the estimated model will be lower than the number of active covariates. The regularization parameters that minimize the GCV function are those used for the estimation of the iterated solution.

Feature extraction is carried out after the parameter estimation stage. With a partial estimation of the coefficients and their standard deviation, using the

sandwich formula (Li et al. (2006)), a Student test is computed. Then we can estimate the p-values corresponding to each coefficient using the normal standard approximation of the t-distribution and, from this, a subset of the significant coefficients, taking into account the problem of multiple comparisons. This is done using the FDR procedure (Benjamini and Hochberg (1995)), in which the rate of false discoveries ($q$) is kept under control. An implementation of this procedure is shown below. We use a generalization of the LQA technique in the estimation step, so we called it the LQA-Fext (LQA and feature extraction) algorithm.

**LQA-Fext**

1. Start with $\beta = 0$ as initial solution.
2. Feature insertion: select a group of covariates whose inclusion in the active subset could mean a further decrease of the objective function using the modified Grafting heuristic.
3. Estimate the coefficients for the active covariates for a range of regularization parameters using the generalized LQA algorithm. Evaluate the GCV function and select the solution corresponding to its minimum value.
4. Feature extraction: Find the significant coefficients using the FDR technique, and remove covariates corresponding to non-significant ones from the active subset.
5. With the solution obtained at step $k$, $\beta_k$, repeat (2)−(4) as the next step. If there is not any feature in the group of inactive covariates or all inactive covariates were discarded in a previous step using the FDR, stop.

If at Step 5 the algorithm stopped because all possible candidates covariates were discarded by the FDR, then we can use a restart strategy where covariates in Z are tagged as not rejected by the FDR, and so would be considered for reintroduction. This variant allows one to continue the insertion-extraction strategy until all candidates covariates in Z are marked as rejected again, or introduced as active. We refer to this as stabilizing. In the application section, at most five repetitions were allowed for obtaining stabilized versions. A current implementation of this methodology is available in Matlab®, and can be obtained upon request to the authors.

## 4. Application

### 4.1. Simulation connectivity data

We construct a synthetic network consisting of 200 voxels grouped into 5 regions. Connectivity matrices were simulated using a voxel-based approach, where

regions were connected indirectly through the establishment of connections between their voxels. The established connections followed a nearest neighborhood pattern, in that a node of a target region received connections from neighbors' nodes of the outgoing region; also an auto-connection for each node was allowed. The synthetic model corresponds to an autoregressive model of order one. Time series were generated for ten thousand time instants in order to render stabilized network dynamics, and the last 101 time instants were used to build the MAR model. Stationarity of the time series was guaranteed by ensuring that absolute singular values of the connectivity matrices were lower than one. Figure 4 shows the true connectivity matrix and those estimated using the LQA-Fext algorithm for the Fusion Ridge and Fusion Lasso models. The algorithm converged after 7 iterations for both methods (17 and 88 seconds, respectively).

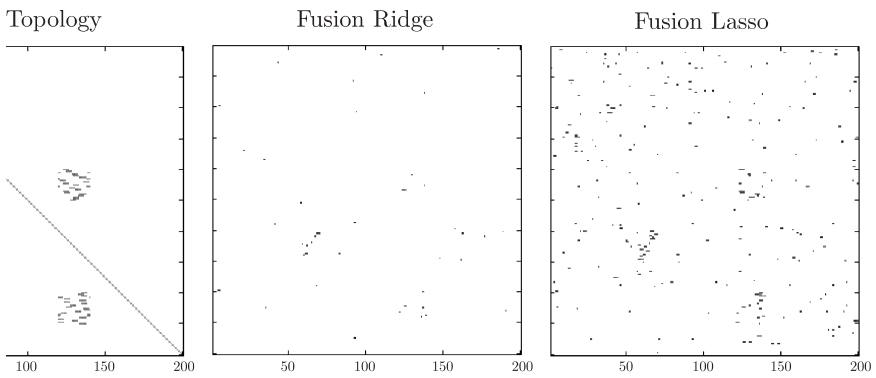


Figure 4. Connectivity matrices representing the true and estimated topologies using the Fusion Ridge and Fusion Lasso algorithms. Five regions were connected through their voxels using a neighborhood pattern. The plots show that the recovered solutions are very sparse. In the case of the Fusion Lasso estimator, the recovering of true connections with a low level of false discoveries is noticeable.

A statistical validation of the LQA-Fext algorithm for these models was carried out using 20 simulations from this synthetic network and a classic Receiver Operator Curve (ROC) analysis. For each simulation, arbitrary thresholds were applied to standard estimators of Fusion Ridge and Fusion Lasso solutions guided from percentiles 0-100% in order to produce different specificity-sensibility values. This allowed the plotting of ROC curves corresponding to these methods. The mean and 95% confident intervals were estimated and plotted, and solutions obtained using the LQA-Fext algorithm for each simulation were summarized using the boxplot technique (Figure 5).

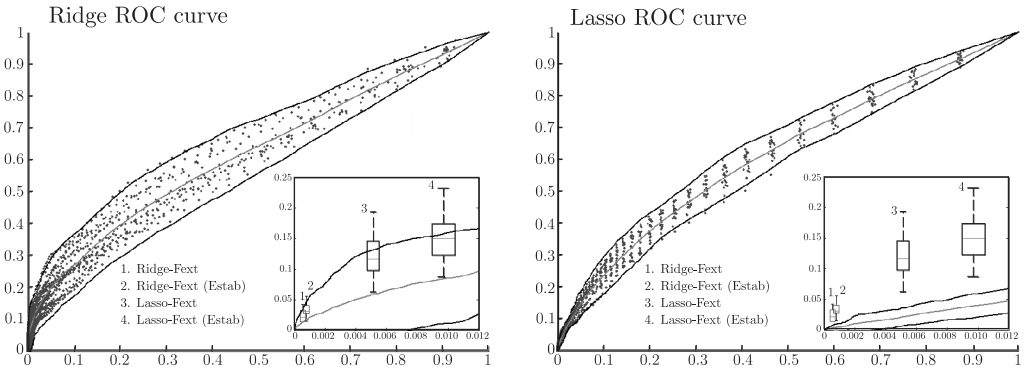


Figure 5. ROC curves for the Fusion Ridge and the Fusion Lasso using standard algorithms. The insets show a selected region of the ROC curves with boxplots of the 20 solutions from the LQA-Fext versions. Stabilized versions are also shown, are not clearly superior to their original estimators. The boxplot width is proportional to the dispersion in specificity values for each method.

Since we have $N = 100$ and $p = 200$, which render a problem with high correlations between the covariates, it is not surprising that the Fusion Ridge's mean ROC has greater area below the curve than the Fusion Lasso's. Also, in the insets it is noticeable that LQA-Fext versions are better than plain estimators for this kind of problem, with the LQA-Fext versions for Fusion Lasso (3 and 4) the most promising. Estimators 2 and 4 correspond to stabilized version for both models (five stabilized iterations). These do not show a clear superiority over their counterparts (Estimators 1 and 3) but require an incremental computational time. This statistical method helps to assess the superiority of the estimators proposed, in particular, for the Fusion Lasso.

### 4.2. Neuroimaging data

Based on the results for synthetic networks, we decided to test the Fusion Lasso method with the LQA-Fext algorithm in actual neuroimaging data. For this purpose we used a concurrent recording of EEG and fMRI time series corresponding to a resting state, which was previously used for the analysis of the origin of resting brain rhythms (Goldman, Stern, Engel and Cohen (2002)). Structured patterns of correlations have been found between time-varying spectral components in different EEG bands and the BOLD signal at different voxels. These revealed widely distributed functional systems apparently involved in the generation of these oscillations (Martínez-Montes et al. (2004)). The fMRI data has a total of 12,642 voxels and 108 time points, and was already used for obtaining the influence field for a representative voxel of the visual cortex (Valdés-Sosa et al. (2006), Figure 1) using a generalization of the MM algorithm.

The fMRI data was segmented into six ROIs according to previous analyses of these data (Goldman et al. (2002) and Martínez-Montes et al. (2004)). However, after segmentation we do not restrict our analysis to ROI time series but use the voxel time series instead. Thus, we estimate a voxel-based effective connectivity between the voxels corresponding to Visual Cortex (647 voxels), Thalamus (177), Left Somatosensory Cortex (29), Right Somatosensory Cortex (31), Left Insula (28) and Right Insula (33) (Figure 6, left panel). This kind of map might allow one to infer the voxels actually connected in the brain without assuming particular ad-hoc segmentations, and can be used in subsequent clustering analysis for revealing highly localized regions involved in cognitive processes. For illustrative purposes, we summarize the voxel-based connectivity information into a ROI-based connectivity map, assuming that a connection from ROI 2 to 1 is well represented as an average of the connections from voxels in ROI 2 to voxels in 1 (Figure 6, middle panel). Auto-connections (connection between voxels of the same ROI) are explicitly shown. Figure 6, right panel, shows a graph of the network connectivity in which auto-connections were omitted to improve visualization. This can be compared with the connectivity network found by Eichler (Eichler (2005)) using the same data.

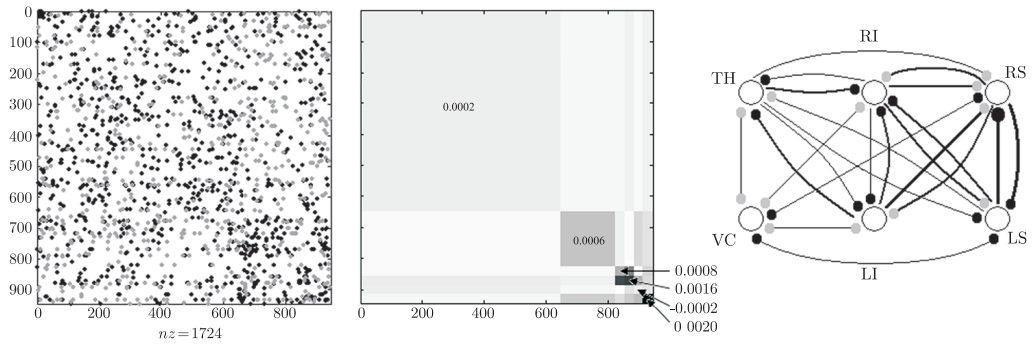


Figure 6. Effective connectivity matrices among six segmented ROIs: Visual Cortex (VC), Thalamo (TH), Left Somatosensory Cortex (LS), Right Somatosensory Cortex (RS), Left Insula (LI) and Right Insula (RI), in this order. Left panel: voxel-wise connectivity matrix for the 945 voxels selected, found after 20 iterations (710 seconds). Middle panel: ROI-wise connectivity matrix found as the average of voxel connectivity values by regions. Grey levels that codify increases in connectivity and auto-connection values are explicitly shown. Right panel: graph of the estimated ROI network; filled circles represent incoming connections and the width of edges their strength. In left and right panels, black and grey dots represent positive and negative significant connections, respectively.

## 5. Conclusions

In this work we have formalized the Multiple Penalized Least Squares approach, with origin in LQA and MM algorithms, as a common framework to carry out several regression methods. This allows the combination of different regression techniques to take advantage of their particular properties (variable selection, good behavior for correlated data). MPLS focuses directly on the use of appropriate priors assumed by users (smoothness, sparseness) in order to render solutions that widely enhance interpretability, and make easier the use of statistical analysis tools. We also proposed an algorithm for dealing with statistical inferences in the case $p \gg N$. This assumes that only a few features are generating the data, and that the way to disclose covariates influences is to detect those features, and then to estimate their corresponding contributions. Thus, the proposed algorithm is a combination of a feature selection heuristic (Grafting), the generalized LQA for parameter estimation, and the removal of non-significant covariates using FDR.

The new methodology allows the analysis of the voxel-based effective connectivity that challenges statistical tools with the estimation of a huge number of parameters. Although here we presented only an exploratory analysis on effective connectivity of resting state networks, results are in agreement with those obtained in previous studies (Eichler (2005)). Some of these networks are widely explored in the literature, but analyses are restricted to ROIs. The problem discussed here must be interpreted as an assessment of the basal effective connectivity, i.e., active communication links in a functional brain in resting state.

A thorough validation of the methodology requires a more extensive study using synthetic networks with more or less realism, but with different topologies and organizational structures. This work can also be extended to deal with more realistic models that involve external input to the systems in the way of DCM (Friston et al. (2006)), so we can estimate effective connectivity in association with a realization of a cognitive task. It must be pointed out that the proposed estimation procedure deals with an autoregressive model using traditional linear regression methods. The simulations and practical applications seem to indicate that is a valid first approximation. Further theoretical work is in order to fully justify the methods presented.

## Acknowledgement

Authors gratefully acknowledge Lester Melie-García for useful comments and detailed revision of the manuscript.

## References


Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Buchel, C. and Friston, K. J. (1997). Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex* **7**, 768-778.

Bullmore, E., Harrison, L., Lee, L., Mechelli, A., and Friston, K. (2004). Brain Connectivity Workshop, Cambridge UK, May 2003. *Neuroinformatics* **2**, 123-125.

Eichler, M. (2005). A graphical approach for evaluating effective connectivity in neural systems. *Phil. Trans. Roy. Soc. B* **360**, 953-967.

Fan, J. Q. and Li, R. Z. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statis. Assoc.* **96**, 1348-1360.

Fan, J. Q. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Friston, K. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping* **2**, 56-78.

Friston, K., Ashburner, J., Stefan, K., Thomas, E. and Penny, W. (2006). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier, London.

Friston, K. J., Harrison, L. and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* **19**, 1273-1302.

Goldman, R. I., Stern, M., Engel, J., and Cohen, M. S. (2002). Simultaneous EEG and fMRI of the alpha rhythm. *Neuroreport* **13**, 2487-2492.

Harrison, L., Penny, W. D., and Friston, K. (2003). Multivariate autoregressive modeling of fmri time series. *Neuroimage* **19**, 1477-1491.

Heeger, D. J. and Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience* **3**, 142-151.

Hoerl, A. E. and Kennard, R. W. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **42**, 80-86.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Iturria-Medina, Y., Canales-Rodriguez, E. J., Melie-Garcia, L., Valdes-Hernandez, P. A., Martinez-Montes, E., Aleman-Gomez, Y., and Sanchez-Bornot, J. M. (2007). Characterizing brain anatomical connections using diffusion weighted MRI and graph theory. *Neuroimage* **36**, 645-660.

Jirsa, V. K. and McIntosh, A. R. (2007). *Handbook of Brain Connectivity*. Springer, Berlin.

Kotter, R. (2001). Neuroscience databases: tools for exploring brain structure-function relationships. *Phil. Trans. Roy. Soc. B* **356**, 1111-1120.

Le Bihan, D. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nature Reviews Neuroscience* **4**, 469-480.

Lee, L., Harrison, L. M. and Mechelli, A. (2003). A report of the functional connectivity workshop, Dusseldorf 2002. *Neuroimage* **19**, 457-465.

Li, R. Z., Dziak, J., Ma, H. Y. (2006). Nonconvex penalized least squares: characterizations, algorithm and application. Manuscript.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, 150-157.

Martínez-Montes, E., Valdés-Sosa, P. A., Miwakeichi, F., Goldman, R. I., and Cohen, M. S. (2004). Concurrent EEG/fMRI Analysis by Multi-way Partial Least Squares. *Neuroimage* **22**, 1023-1034.

Ogawa, S. and Sung, Y. W. (2007). Functional Magnetic Resonance Imaging. Scholarpedia.

Penny, W. and Friston, K. (2007). Functional Imaging. Scholarpedia.

Perkins, S., Lacker, K., and Theiler, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *J. Mach. Learn Res.* **3**, 1333-1356.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer, New York.

Salvador, R., Suckling, J., Schwarzbauer, C., and Bullmore, E. (2005). Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Phil. Trans. Roy. Soc. B* **360**, 937-946.

Sporns, O., Tononi, G. and Kotter, R. (2005). The human connectome: A structural description of the human brain. *Plos Computational Biology* **1**, 245-251.

Stephan, K. E., Kamper, L., Bozkurt, A., Burns, G., Young, M. P., and Kotter, R. (2001). Advanced database methodology for the collation of connectivity data on the Macaque brain (CoCoMac). *Phil. Trans. Roy. Soc. B* **356**, 1159-1186.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. Roy. Statist. Soc. B* **67**, 91-108.

Valdés-Sosa, P. A. (2004). Spatio-temporal autoregressive models defined over brain manifolds. *Neuroinformatics* **2**, 239-250.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Lage-Castellanos, A., Vega-Hernández, M., Bosch-Bayard, J., Melie-García, L. and Canales-Rodríguez, E. (2005). Estimating brain functional connectivity with sparse multivariate autoregression. *Phil. Trans. Roy. Soc. B* **360**, 969-981.

Valdés-Sosa, P. A., Sánchez-Bornot, J. M., Vega-Hernández, M., Melie-García, L., Lage-Castellanos, A., and Canales-Rodríguez, E. (2006). *Granger Causality on Spatial Manifolds: Applications to Neuroimaging*. Handbook of Time Series Analysis: Recent Theoretical Developments and Applications, Chapter 18. ISBN: 3-527-40623-9.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301-320.

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: bornot@cneuro.edu.cu

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: eduardo@cneuro.edu.cu

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: agustin@cneuro.edu.cu

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: mayrim@cneuro.edu.cu

Neurostatistics Department, Cuban Neuroscience Center, Havana, Cuba.

E-mail: peter@cneuro.edu.cu

(Received April 2007; accepted February 2008)