HHS Public Access

Author manuscript

*Conf Proc IEEE Eng Med Biol Soc*. Author manuscript; available in PMC 2015 March 04.

Published in final edited form as:

*Conf Proc IEEE Eng Med Biol Soc*. 2014 August ; 2014: 1513–1516. doi:10.1109/EMBC.2014.6943889.

# Identification of Patterns of Gray Matter Abnormalities in Schizophrenia using Source-Based Morphometry and Bagging

**Eduardo Castro**[1], **Cota Navin Gupta**[1], **Manel Martínez-Ramón**[2], **Vince D. Calhoun**[1], **Mohammad R. Arbabshirani**[1], and **Jessica Turner**[3]

[1]Mind Research Network, Albuquerque, NM 87131, USA

[2]Electrical and Computer Engineering Department, The University of New Mexico, Albuquerque, NM 87131, USA

[3]Department of Psychology, Georgia State University, Atlanta, GA 30302, USA

## Abstract

Despite its reliable diagnosis, schizophrenia lacks an objective diagnostic test or a validated biomarker, which prevents a better understanding of this disorder. Structural magnetic resonance imaging (sMRI) has been vastly explored to find consistent abnormality patterns of gray matter concentration (GMC) in schizophrenia, yet we are far from having reached conclusive evidence. This paper presents a machine learning approach based on resampling techniques to find brain regions with consistent patterns of GMC differences between healthy controls and schizophrenia patients, these regions being detected by means of source-based morphometry. This work uses multi-site data from the Mind Clinical Imaging Consortium, which is composed of sMRI data from 124 controls and 110 patients. Our method achieves a better classification rate than other algorithms and detects regions with GMC differences between both groups that are consistent with several findings on the literature. In addition, the results obtained on data from multiple sites suggest that it may be possible to replicate these results on other datasets.

## I. INTRODUCTION

Source-based morphometry (SBM) [1] is a technique that uses independent component analysis (ICA) [2] to obtain patterns of common gray matter concentration (GMC) variation among subjects. By applying SBM for schizophrenia studies, GMC deficits clustered into independent spatial regions can be identified. This approach has three main advantages. First, it performs a multivariate analysis of whole-brain data, so it does not restrict the analysis to a single region of interest. Second, it accounts for spatial dependencies between different brain locations, which are not taken into consideration by univariate analyses such as voxel-based morphometry (VBM) [3]. Third, it provides a better interpretation of the location of GMC variations than voxel-based approaches.

Up to date, only univariate tests on the SBM components loading coefficients have been used to identify those that are significantly associated to schizophrenia. This kind of

approaches do not take into account the multivariate contribution of these components to schizophrenia detection, which would better assess SBM's discrimination capacity of healthy controls and schizophrenia patients. Furthermore, such a classification pursuit would benefit from a feature selection strategy, as only some of the SBM components are usually associated to schizophrenia.

While there is evidence that supports the reliability of SBM findings on spatial components that are associated to schizophrenia, doing a sparse selection of informative components is not a trivial task. The reason behind this is that this selection can change if some subjects are set aside of the analysis due to inter-subject variability. If data from multiple sites are incorporated in the analysis, then inter-site variability can further complicate the analysis.

The objective of this work is to find components that have a consistent contribution to the classification model and are less sensitive to inter-subject and inter-site variability. One way of doing so is by using resampling methods such as bagging [4], which trains an ensemble of classifiers with different data samples and gives a prediction based on a combination of them. If the contribution of the components to the classification task is evaluated across these data samples, it is possible to detect the most consistently informative ones.

## II. DATA

### A. Participants

This study uses structural magnetic resonance imaging (sMRI) data from the Mind Clinical Imaging Consortium study of schizophrenia from 234 subjects from 4 participating sites: Massachusetts General Hospital in Boston (MGH), University of Iowa (UI), University of Minnesota (UMN), and the University of New Mexico (UNM).

Schizophrenia was diagnosed according to DSM-IV criteria [5] on the basis of both a structured clinical interview and the review of the medical file. Healthy participants were screened to ensure they were free from DSM-IV Axis I psychiatric diagnosis and were also interviewed to determine that there was no history of psychosis in any first-degree relatives.

The analyzed data sample comprised 124 healthy controls (75 males; mean ± SD age, 32.27 ± 10.89 years) and 110 schizophrenia patients (82 males; mean ± SD age, 34.86 ± 10.99 years).

### B. Image Acquisition

SMRI data were acquired with either a Siemen's 1.5-Tesla (MGH, UI, and UNM) or a Siemen's 3-Tesla (UMN) MR scanners. The $T_1$-weighted structural brain scans at each of the 4 sites were acquired with an in-plane resolution of $0.625 \times 0.625$ mm$^2$, a slice thickness of 1.5 mm, and a flip angle of 7 degrees. MGH and UNM used a Siemen's 1.5-Tesla scanner with repetition time (TR) = 12 ms, echo time (TE) = 4.76 ms, and number of excitations (NEX) = 1. UI used a GE 1.5-Tesla Genesis Signa scanner with TR = 20 ms, TE = 6 ms, and NEX = 3. UMN used a Siemen's 3-Tesla scanner with TR = 2530 ms, inverse time (TI) = 1100, TE = 3.79 ms, and NEX = 1.

### C. Preprocessing

Images were preprocessed using the methods presented in [6]. The SPM5 software package (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/) was used to segment the $T_1$-weighted images into gray matter, white matter and cerebrospinal fluid images, using unified segmentation [3].

Since age and gender affect GMC [1], [6], and SBM components as a consequence, these were regressed out of the images voxel-wise prior to further analysis. Then an isotropic 10 mm full width at half maximum Gaussian filter was used to smooth the images. In order to overcome imaging heterogeneity of this multi-site dataset, we regressed out data collection site ($N$ = 4 sites) from the loading coefficients of the SBM analysis.

## III. METHODS

### A. SBM Analysis

This approach identifies spatial maps with common patterns of GMC by applying ICA to the subjects' GMC. This provides a decomposition of these volumes into a linear combination of spatial maps and loading coefficients. These coefficients, which weight each component in each subject's data as depicted in Fig. 1, are used to examine component differences between controls (Ct) and patients (Sz).

A larger loading coefficient associated to a spatial pattern for an individual or group indicates that this component is more strongly weighted in the data for that individual or group. However, the interpretation of the loading coefficient difference depends on the spatial map of the component. If the spatial component is predominantly positive and if the loading coefficients are greater in Ct than in Sz, we infer that GMC is greater in Ct for the spatial component under consideration [7].

The SBM module of the GIFT Toolbox (http://mialab.mrn.org/software/gift/) was used to perform the ICA decompositions on the aggregated dataset. The number of components was set to 30 as in [1], [6], and we used ICASSO (http://research.ics.aalto.fi/ica/icasso/, 20 runs) to determine the stability of the components [6].

### B. Bagged Support Vector Machine

The machine learning approach we propose to identify components associated to schizophrenia is based on a linear classifier, which is provided with $N$ labeled training pairs of data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. Here, GMC data of subject $i$ is represented by the its $S$ loading coefficients ($S$: number of sources or components) such that $\mathbf{x}_i = [x_{i1}, \ldots, x_{iS}]^\top$, and $y_i$ indicates the subject class (−1 for Ct, +1 for Sz). The estimated class of an unseen example $\mathbf{x}$ by a linear classifier is defined by

$$\hat{y} = f(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x} + b) = \operatorname{sgn}\left(\sum_{d=1}^{S} w_d x_d + b\right), \quad (1)$$

where $\mathbf{w} = [w_1, \ldots, w_S]$ is the weight vector of the classifier, each element of **w** being associated to one component.

Here we provide a brief insight of the interpretation given to the components weights. For the sake of simplicity, let us assume that all of the components spatial maps are predominantly positive. If the loading coefficients of component *d* were greater in Sz than in Ct, then GMC would be greater for Sz for that component. Given the class representation of Ct (−1) and Sz (+1), $w_d$ would take a positive value. On the other hand, if loading coefficients of the same component were lower in Sz, then GMC would be lower for Sz too and $w_d$ would take a negative value. This gives a clear interpretation of the role of each component in the discriminant function defined on (1).

We use a support vector machine (SVM) [8] to train the linear classifier. The SVM, which was coded in MATLAB (http://www.mathworks.com) and solved using the MOSEK optimization toolbox (http://www.mosek.com), finds **w** and the bias term *b* in (1) by solving the following optimization problem:

$$\begin{array}{lll} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} & \|\mathbf{w}\|_p + C\sum\limits_{i=1}^{N}\xi_i & \\ s.t. & y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i & \forall i \in [1, N] \\ & \xi_i \geq 0 & \forall i \in [1, N]. \end{array} \quad (2)$$

Slack variables $\xi_i$ in (2) allow errors in training data, $\|\mathbf{w}\|_p$ is an $l_p$-norm regularization term that controls overfitting and *C* is a parameter that controls the tradeoff between the regularization and the training error terms.

The choice of norm *p* depends on the nature of the classification task. An $l_1$-norm regularization enforces sparsity in **w**, making the weights associated to non-informative features equal to zero. Since the data dimensionality is significantly reduced from hundreds of thousands voxels to *S* components through SBM, an $l_1$-norm regularization term is well-suited to find informative components. Nonetheless, a selection of informative components cannot rely on an $l_1$-norm SVM approach only, as its sparse selection can vary significantly if some subjects are set aside of the analysis due to the inherent inter-subject variability of the data. This problem can be solved to a certain extent by using a pool of diverse linear classifiers, as suggested by the starplots method proposed by [9]. However, our approach follows the rationale of the bagged-selection presented in [10], which looks for features with consistent weight patterns.

The way the ensemble of classifiers is constructed closely resembles that of bagging classifiers [4]. Each classifier is trained using a subset of $M < N$ examples randomly sampled from the training data. However, unlike [4], the data is sampled with no replacement. The idea is to select those components that exhibit a high degree of sign consistency across the ensemble, disregarding those ones with unstable sign fluctuations or that are rarely selected by the pool of classifiers. The degree of sign consistency is defined by parameter *r*, which establishes a lower bound of the fraction of classifiers in which informative components weights should be consistently positive or negative.

### C. Experimental Setup

The classification accuracy rate obtained by bagged SVM was estimated using a leave-one-out cross-validation (loo-CV) procedure. More specifically, 234 classifiers were trained by using the data from all subjects except one, each of them being tested on that left-out example. The fraction of correctly classified examples using this experimental setting defined the method's accuracy rate.

There are also several parameters that need to be either fixed to a given value or selected using some validation procedure. These parameters are the number of classifiers of the ensemble ($P$), the number of examples drawn to train each of these classifiers ($M$), the sign consistency threshold of the components ($r$) and the SVM parameter $C$. The proposed approach did not seem to be sensitive to parameters $P$ and $M$, which were fixed to 250 and 80% of the available data, respectively. However, the sign consistency threshold defines the number of selected components, so it needs to be validated. Furthermore, $C$ also plays an important role in determining the number of components to be selected, so it needs to be properly estimated. To do so, an inner loo-CV was ran inside each round of the loo-CV used to estimate the classification accuracy. In other words, at each round of the main loo-CV, a nested loo-CV selected the appropriate values of $C$ and $r$ for that specific round. $C$ was picked from 3 logarithmically spaced points in the range [0.1, 1], while $r$ was selected from the range [50%, 100%] in steps of 1%.

## IV. RESULTS

### A. Classification Accuracy

Table I shows the classification accuracy, sensitivity and specificity rates achieved by Bagged SVM. This table also shows the results obtained by a linear SVM with $l_1$- and $l_2$-norm regularization terms, as well as for partial least squares (PLS[1]) [11] and random forests (RF[2]) [12]. PLS and RF were also implemented using MATLAB.

### B. Relevant Components Statistics

The multivariate association of the components to schizophrenia was evaluated in terms of their weights mean and selection frequencies across CV rounds. Table II shows the components that achieved a selection frequency greater than or equal to 0.5 and the $p$-values associated to a two-sample $t$-test across Sz and Ct on their loading coefficients.

### C. Components Loadings Directionality and Spatial Extent

Three out of the four listed components contain areas were GMC is greater on Ct than Sz, while the other one shows a reversed pattern. Table III also shows the set of brain regions spanned by these components, which are graphically depicted in Fig. 2.

[1] the parameter #factors $\in [1, S/2]$ was selected on the inner CV

[2] trained with fixed parameters: #trees=250, $\mathrm{mtry} = \sqrt{S}$

## V. DISCUSSION AND CONCLUSIONS

To the best of our knowledge, this is the first work that performs classification of schizophrenia using GMC patterns obtained by SBM. In addition, the results obtained on data from multiple sites indicate that it may be possible to replicate these results on other datasets.

We present evidence that the effects of imaging heterogeneity due to multi-site acquisition are successfully dealt with our data analysis pipeline based on the accuracy rates obtained by different classifiers (Table I). Site differences can significantly deteriorate prediction capacity when working directly with GMC maps, potentially giving poor accuracy rates, as we discovered on preliminary analyses.

The reliability of Bagged SVM is supported by the results presented in Table I. It can be seen that applying a linear $l_1$-norm SVM to the data is not beneficial, as its prediction accuracy is lower than that obtained by an $l_2$-norm SVM. However, bagging significantly improves the performance of $l_1$-norm SVM, making it surpass the one achieved by $l_2$-norm SVM and PLS, which show equivalent performance. This reinforces the notion that resampling can reduce the effect of inter-subject variability of the data for feature selection if it is used appropriately on the analyzed domain. For instance, RF, which also apply resampling and perform feature selection, show a suboptimal performance.

Table II shows the components that were more informative for schizophrenia classification, along with the multivariate statistics retrieved from Bagging SVM and the results of two-sample *t*-tests across Sz and Ct on their loading coefficients. Components 1, 2 and 4 are significantly associated to schizophrenia according to the univariate tests, which is not the case of component 3. This result suggests that the multivariate analysis performed by Bagging SVM is capable of detecting relevant components that would be rendered irrelevant to schizophrenia based on univariate tests. In fact, a similar component that covers the middle occipital gyrus is reported to be significant for schizophrenia on a very large data study [7].

Fig. 2 shows the spatial extent of the components, the labels of these regions being listed in Table III. Component 1 presents the largest GMC differences between Ct and Sz according to Bagging SVM and includes, among other regions, the superior temporal gyrus, which represents one of the most consistent structural abnormalities found on schizophrenia [13]. In addition, several of the regions identified by the VBM meta-analysis in [14] as having GM deficits in schizophrenia are detected in this analysis across components 1 and 2.

On the other hand, component 4 shows increased GM in Sz. This component covers the ventral tegmentum, which is responsible for dopamine production. Since most antipsychotic treatments tend to block receptors in the brain's dopamine pathways, this may result in compensatory increases in GM in some brainstem regions [7], [15].

## Acknowledgments

This work was mainly funded by NIMH 1R01MH094524-01A1 and DOE DE-FG02-99ER62764 grants.

## References


1. Xu L, Groth K, Pearlson G, Schretlen D, Calhoun V. Source-based morphometry: The use of independent component analysis to identify gray matter differences with application to schizophrenia. Hum Brain Mapp. 2009; 30(3):711–724. [PubMed: 18266214]
2. Bell A, Sejnowski T. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 1995; 7:1129–1159. [PubMed: 7584893]
3. Ashburner J, Friston K. Voxel-based morphometry—the methods. NeuroImage. 2000; 11(6):805–821. [PubMed: 10860804]
4. Breiman L. Bagging predictors. Mach Learn. 1996; 24(2):123–140.
5. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 4. APPI; 1994.
6. Turner J, Calhoun V, Michael A, van Erp T, Ehrlich S, et al. Heritability of multivariate gray matter measures in schizophrenia. Twin Res Hum Genet. 2012; 15(3):324–335. [PubMed: 22856368]
7. Gupta C, Calhoun V, Rachakonda S, Chen J, Liu J, Segall J, et al. Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis. Schizophr Bull. Under Review.
8. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995; 20(3):273–297.
9. Bi J, Bennett K, Embrechts M, Breneman C, Song M. Dimensionality reduction via sparse support vector machines. J Mach Learn Res. 2003; 3:1229–1243.
10. Parrado-Hernández E, Gómez-Verdejo V, Martínez-Ramón M, Shawe-Taylor J, et al. Discovering brain regions relevant to obsessive–compulsive disorder identification through bagging and transduction. Medical Image Analysis. 2014; 18(3):435–448. [PubMed: 24556078]
11. de Jong S. Simpls: An alternative approach to partial least squares regression. Chemometr Intell. 1993; 18(3):251–263.
12. Breiman L. Random forests. Mach Learn. 2001; 45(1):5–32.
13. Ross C, Margolis R, Reading S, Pletnikov M, Coyle J. Neurobiology of schizophrenia. Neuron. 2006; 52(1):139–153. [PubMed: 17015232]
14. Glahn D, Laird A, Ellison-Wright I, Thelen S, Robinson J, et al. Meta-analysis of gray matter anomalies in schizophrenia: Application of anatomic likelihood estimation and network analysis. Biol Psychiatry. 2008; 64(9):774–781. [PubMed: 18486104]
15. Laruelle M, Abi-Dargham A, Gil R, Kegeles L, Innis R. Increased dopamine transmission in schizophrenia: relationship to illness phases. Biol Psychiatry. 1999; 46(1):56–72. [PubMed: 10394474]

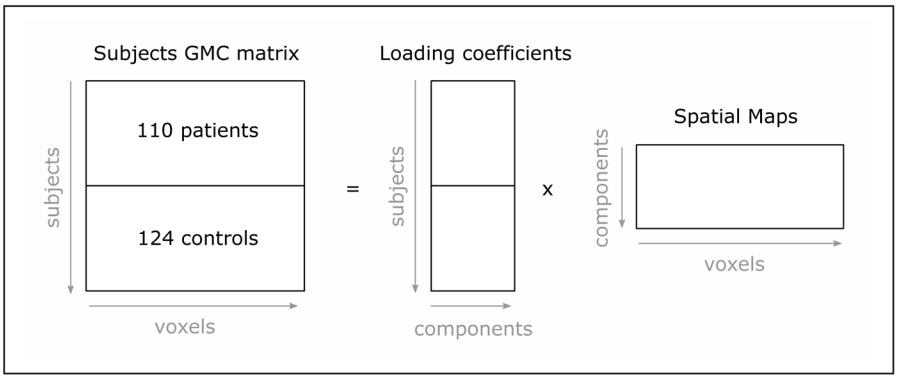


**Fig. 1.**

GMC images are concatenated for each subject. SBM decomposes the subject/voxel matrix into loading coefficients (mixing matrix) and a source matrix (spatial maps). The loading coefficients are unique for each subject and permit group comparisons.

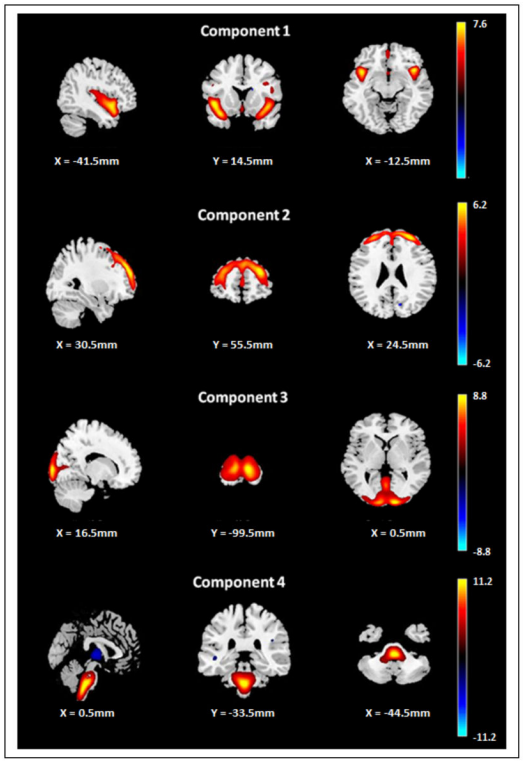


**Fig. 2.**

Spatial maps of the top 4 components showing (Ct/Sz) group difference ranked by product of mean weight and selection frequency. All maps are thresholded at |*z*-score| > 3.5. The colorbars indicate the color mapping of the component weights.

Author Manuscript

Author Manuscript

Author Manuscript

**TABLE I**

Classification Performance of Bagged SVM

| Approach | Class. Acc. | Sens/Spec |
|---|---|---|
| RF | 0.67 | 0.59/0.74 |
| $l_1$-norm SVM | 0.68 | 0.61/0.73 |
| $l_2$-norm SVM | 0.70 | 0.64/0.75 |
| PLS | 0.70 | 0.66/0.73 |
| Bagged SVM | 0.73 | 0.65/0.80 |

**TABLE II**

Relevant Components Statistics

| Component Number | Mean Weight | Selection Frequency | *t*-test (*p*-value) |
|---|---|---|---|
| 1 | −0.64 | 0.98 | 1.50E-8 |
| 2 | −0.44 | 0.79 | 4.42E-8 |
| 3 | −0.23 | 0.97 | 0.01 |
| 4 | 0.26 | 0.50 | 9.09E-6 |

**TABLE III**

Loadings Directionality and Components Regions

| Component Number | Loadings Directionality | Brain Regions |
|---|---|---|
| 1 | Ct>Sz | Superior Temporal Gyrus<br>Inferior Frontal Gyrus<br>Insula |
| 2 | Ct>Sz | Superior Frontal Gyrus<br>Middle Frontal Gyrus<br>Medial Frontal Gyrus |
| 3 | Ct>Sz | Middle Occipital Gyrus<br>Calcarine Cortex<br>Lingual Gyrus |
| 4 | Sz>Ct | Brainstem |