Video Article

Basics of Multivariate Analysis in Neuroimaging Data

Christian Georg Habeck
Department of Neurology, Columbia University

Correspondence to: Christian Georg Habeck at ch629@columbia.edu

URL: http://www.jove.com/index/Details.stp?ID=1988
DOI: 10.3791/1988
Citation: Habeck C.G. (2010). Basics of Multivariate Analysis in Neuroimaging Data. JoVE. 41. http://www.jove.com/index/Details.stp?ID=1988, doi: 10.3791/1988

Abstract

Multivariate analysis techniques for neuroimaging data have recently received increasing attention as they have many attractive features that cannot be easily realized by the more commonly used univariate, voxel-wise, techniques[1,4,5,6,7]. Multivariate approaches evaluate correlation/covariance of activation across brain regions, rather than proceeding on a voxel-by-voxel basis. Thus, their results can be more easily interpreted as a signature of neural networks. Univariate approaches, on the other hand, cannot directly address interregional correlation in the brain. Multivariate approaches can also result in greater statistical power when compared with univariate techniques, which are forced to employ very stringent corrections for voxel-wise multiple comparisons. Further, multivariate techniques also lend themselves much better to prospective application of results from the analysis of one dataset to entirely new datasets. Multivariate techniques are thus well placed to provide information about mean differences and correlations with behavior, similarly to univariate approaches, with potentially greater statistical power and better reproducibility checks. In contrast to these advantages is the high barrier of entry to the use of multivariate approaches, preventing more widespread application in the community. To the neuroscientist becoming familiar with multivariate analysis techniques, an initial survey of the field might present a bewildering variety of approaches that, although algorithmically similar, are presented with different emphases, typically by people with mathematics backgrounds. We believe that multivariate analysis techniques have sufficient potential to warrant better dissemination. Researchers should be able to employ them in an informed and accessible manner. The current article is an attempt at a didactic introduction of multivariate techniques for the novice. A conceptual introduction is followed with a very simple application to a diagnostic data set from the Alzheimer s Disease Neuroimaging Initiative (ADNI), clearly demonstrating the superior performance of the multivariate approach.

Protocol

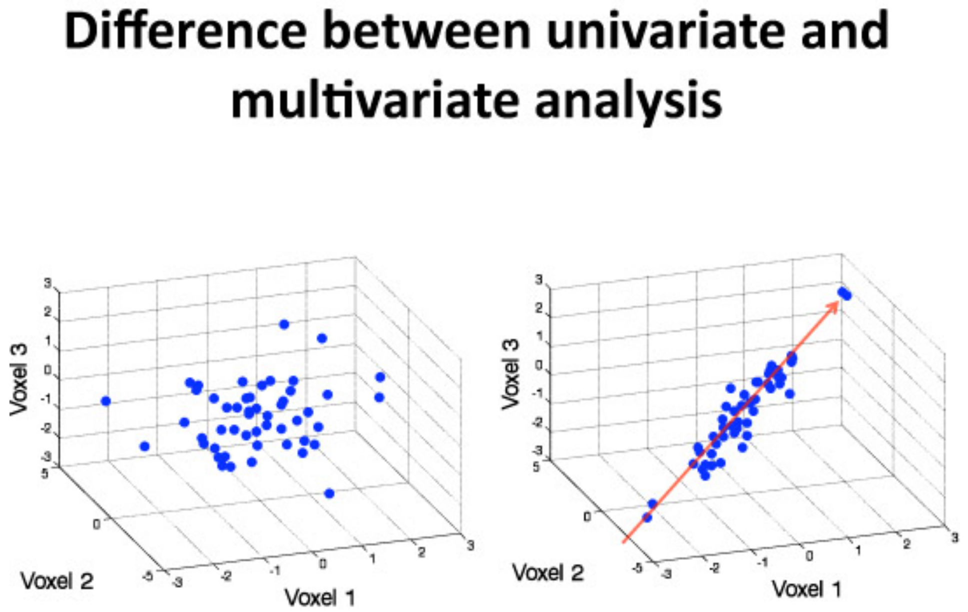
1. To give a conceptual overview of multivariate analysis we can picture a very simple situation: a hypothetical data set for 50 human participants, where only three regions, denoted as voxels (=3-dimensional pixels in Figure 1) in the brain were measured. (Insert Figure 1 here, read caption as voice over.)
2. The general aim of multivariate analysis is to identify the major sources of variance in the data, and then describing the major effects of interest in the data in terms of these sources of variance. Figure 2 shows a simplistic example. (Insert Figure 2 here, read caption as voice over.)
3. We now apply both univariate and multivariate analysis to a clinical data set. We downloaded FDG-PET resting scans for 95 early Alzheimer's patients and 102 age-matched controls from the website of the Alzheimer's Disease Neuroimaging Initiative (http://www.loni.ucla.edu/ADNI/). We randomly picked 20 scans of both patients and controls and designated them as our derivation sample. The remaining 75 and 82 scans, respectively, constitute our replication sample. Univariate and multivariate Alzheimer's disease (AD) markers will now be derived in the derivation sample, and their diagnostic efficacy tested in the replication sample.
4. For the univariate marker, we contrast the 20 AD scans with the 20 controls scans in the derivation sample and pick the brain location that shows the largest decrease in PET signal in the AD patients as shown by a T-test. To test the diagnostic efficacy of this region, we check the data in the replication sample at this location and plot its PET signal as a function of disease status.
5. For the multivariate marker, we first perform a PCA on the combined 40 scans in the derivation sample, and then construct a covariance pattern from the first 5 Principal Components whose subject scaling factor shows a maximal mean difference between AD patients and healthy controls. (Details can be found in these representative papers [2].) The diagnostic covariance pattern obtained form the derivation sample is then prospectively applied to the replication sample. The resulting subject scaling factors are plotted as a function of disease status.
6. To provide a more general comparison of both univariate and multivariate approaches from step 4 and 5, we perform a "split sample" simulation and repeat both steps 1,000 times on resampled data, each time forming a 20/20 derivation sample and a 75/82 replication of AD patients and healthy controls afresh. Univariate and multivariate disease markers are computed from the derivation sample and the decision threshold is set such that at most 1 healthy control is misclassified as AD (= specificity 95%). The disease markers with their specific decision thresholds are then prospectively applied to the replication samples. The classification error rates in the replication sample are recorded for all resampling iterations.

Representative Results

Univariate performance The results can be seen in detail in Figure 3. The area of the largest AD-related FDG deficit was found in the super temporal gyrus, Brodmann area 38. The area under the ROC-curve achieved was AUC=0.90. The generalization of this contrast to the replication sample was quite good with an area under the ROC curve of AUC=0.84.

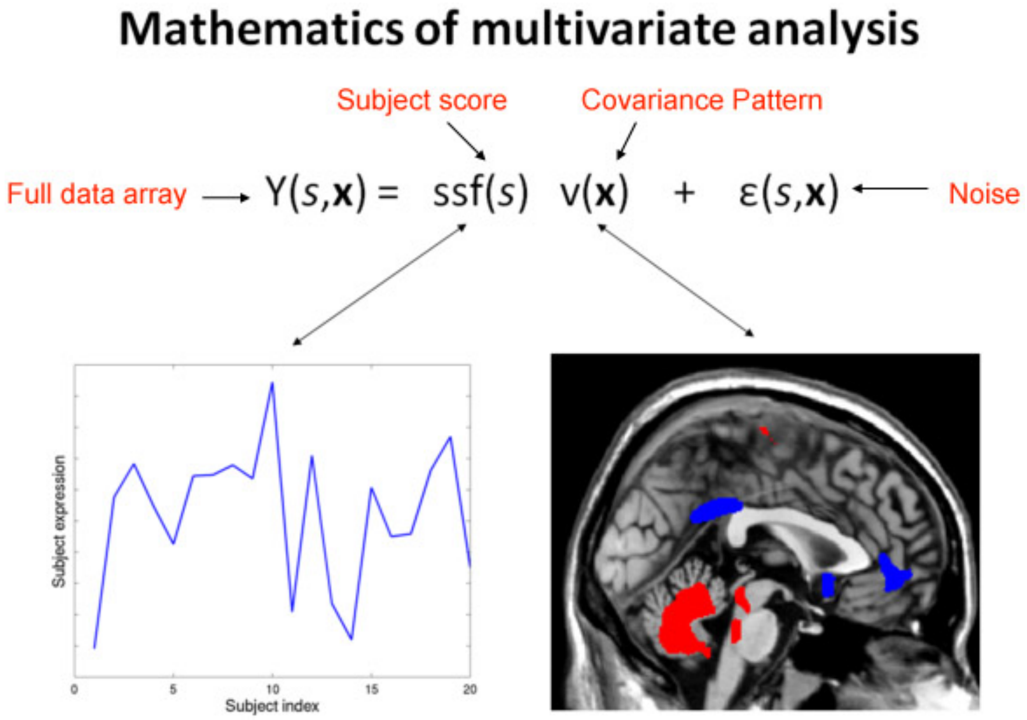
Multivariate performance The results can be seen in detail in Figure 4. Areas with positive loadings, hinting at a relative preservation of signal in the face of disease were found in the cerebellum, while associated signal loss was found the parietotemporal and frontal areas, and the posterior cingulate gyrus. The areas under ROC-curves in both derivation and replication samples were slightly better than the univariate marker at 0.96 and 0.88, respectively.

Split-sample simulations The results can be seen in detail in Figure 5. The figure shows that the multivariate marker gives better replication of diagnostic performance than the univariate marker. The mean total error rate for the multivariate marker is 0.203, whereas for the univariate marker it is 0.307.



Difference between univariate and multivariate analysis

Figure 1. This simple figure describes the difference between univariate and multivariate analytic strategies: a hypothetical 3-dimensional data set is displayed in this illustration. On the left side, there is no correlation between the 3 variables plotted. On the right side in contrast, one can see a major source of variance indicating a positive correlation between all three voxels. A univariate analysis that just considered mean values on a voxel-by-voxel basis could not tell any difference between these two scenarios. Multivariate analysis, in contrast, identifies the major sources of variance in the data (red arrow) before proceeding to construct neural activation patterns form these sources.



Mathematics of multivariate analysis

Subject score

Covariance Pattern

Full data array

$$Y(s,\mathbf{x}) = ssf(s)\ v(\mathbf{x}) + \varepsilon(s,\mathbf{x})$$

Noise

Figure 2. This slide shows in a simplified form the basic accomplishment of any multivariate analysis in neuroimaging data. The data array Y(s,x), which depends on a subject index s, and a voxel index x, indicating the location of the voxel in the brain, is decomposed into a sum of several terms. First, a product of a purely subject-dependent factor score, ssf(s), and a purely voxel-dependent covariance pattern, v(x). Second, activation that cannot be accounted for by the covariance pattern is captured in a subject- and voxel-dependent noise term, e(s,x). The two graphics below the equation give an example of the subject scaling factor and the covariance pattern. Every participant manifests the covariance pattern, just to a different degree as shown by the subject factor score. Rather than having to keep track of every voxel's behavior separately, the covariance pattern and its subject expression provide a parsimonious summary of the major source of variance. As the subject scaling factor increases in magnitude, the areas denoted in blue in the covariance pattern decrease their associated activation, while the areas indicated in red simultaneously increase their associated activation. The subject factor score can be correlated with external variables of interest like subject age or behavioral performance in a cognitive task, and no correction for multiple comparisons has to be applied to this correlation.

Several techniques for such decomposition exist, but the most common one is Principal Components Analysis (PCA). This is the technique of choice for us. Note that subject scaling factors can be obtained by projecting the covariance pattern into any data set of equal dimensionality, not just the data set that produced the covariance pattern in the first place. This makes covariance patterns suitable for testing whether brain-behavioral relationships that were observed in one data set can be replicated in a different data set.

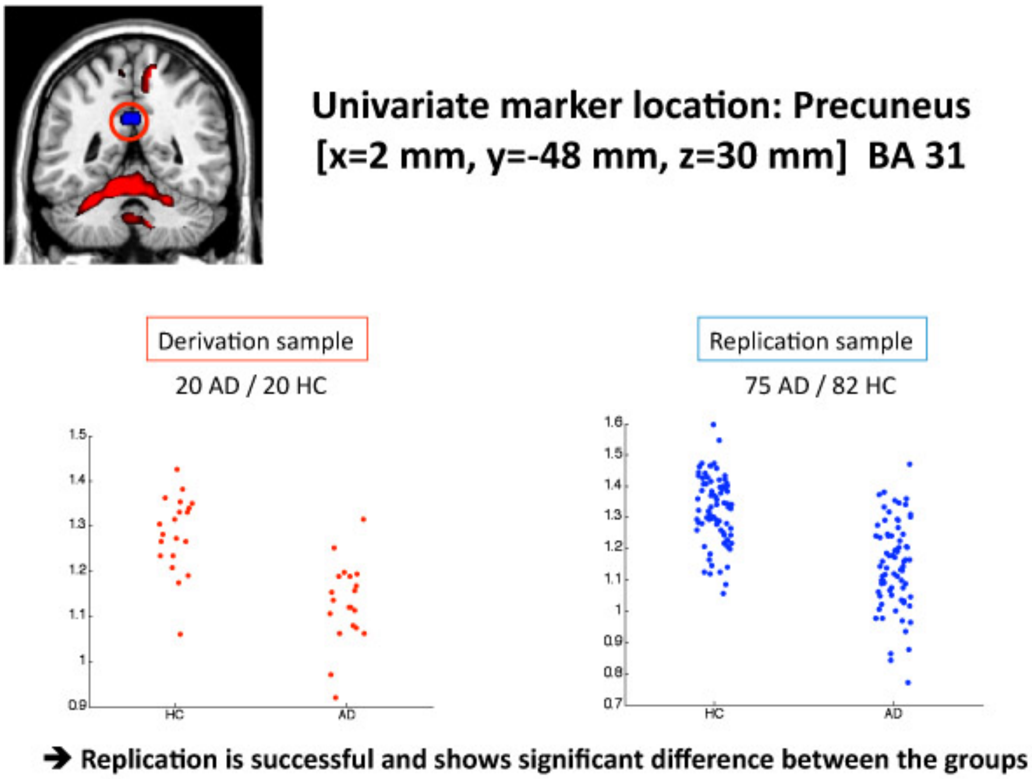


Figure 3. This figure shows the result of the univariate analysis. In the lower left panel, the FDG signal values are plotted for the area that shows the largest AD-related deficit in the derivation sample. Its MNI coordinates are X=2 mm ,Y= -48 mm , Z= 30mm (Precuneus/PCG, Brodmann Area 31). The lower right panel shows the FDG signal at this very location in the replication sample. One can appreciate that the FDG differences between AD patients and controls in the replication sample, while still significant overall, are reduced with more overlap between the groups.

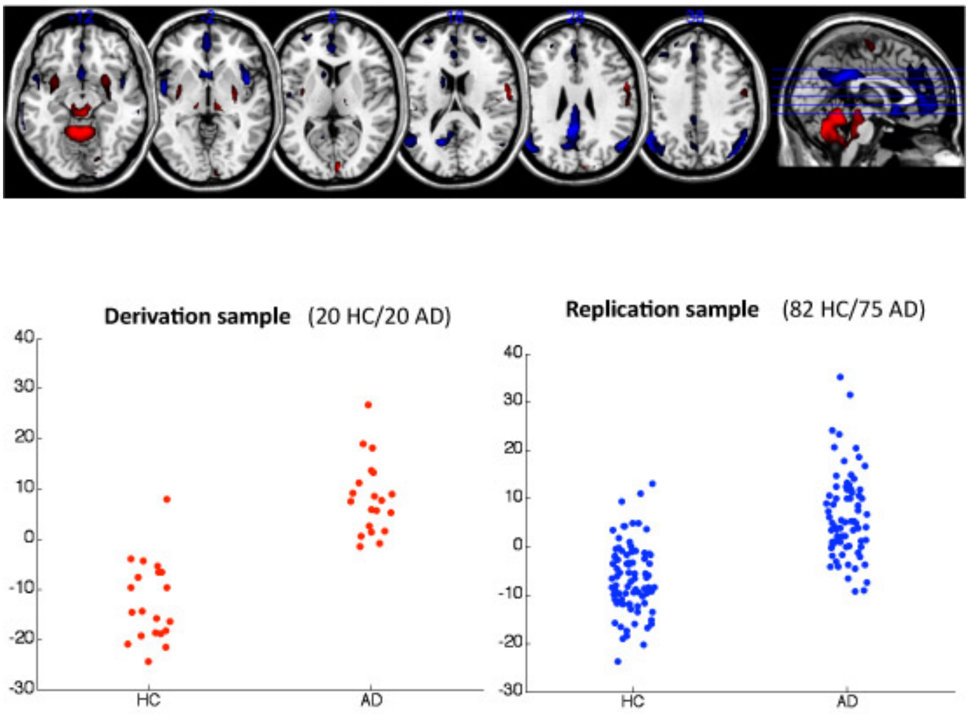
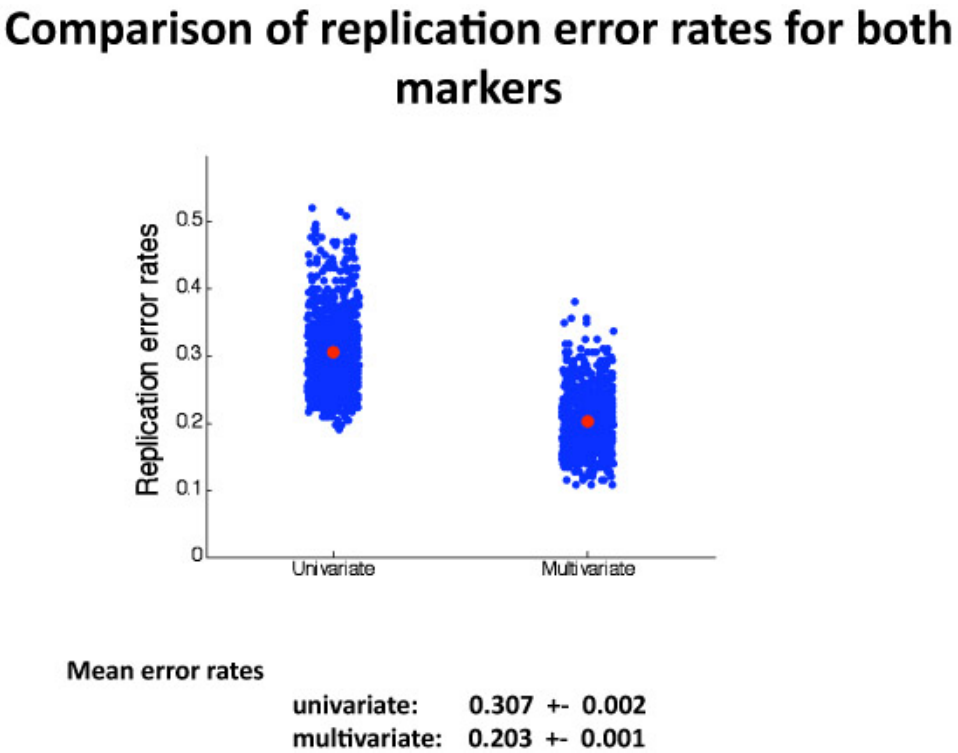


Figure 4. This figure shows the results of the multivariate analysis. In the top panel, we display several axial slices that show significantly positively and negatively weighted areas (p<0.001) in the covariance pattern in red and blue, respectively. Note that we scaled every scan by its global mean value, so red and blue color indicate relative rather and absolute increases and decreases of PET signal with disease severity. Red areas thus hint at relative preservation in the face of the disease, while blue indicates a loss of signal as a consequence of the disease. Red areas are mainly found in the cerebellum, while blue areas appear in the posterior cingulate gyrus, parietotemporal and frontal regions. Lower left panel: the subject factor scores of the AD-related covariance pattern are displayed in the derivation sample. Higher subject scores are found for the AD patients. Lower right panel: the subject factor scores resulting from the prospective application of the AD-related covariance pattern to the replication sample are plotted here. One can appreciate a slight worsening of the diagnostic contrast with increased overlap in the replication sample, but the generalization of the diagnostic efficacy is noticeably better than in the univariate case.



Comparison of replication error rates for both markers

Mean error rates

univariate: 0.307 +- 0.002

multivariate: 0.203 +- 0.001

Figure 5. This figure shows the results of the 1,000 split-sample simulations. Listed are means and standard deviations of the univariate and multivariate diagnostic error rates in the replication samples. One can appreciate that the multivariate marker's generalization of performance is considerably better, although somewhat more variable than the univariate marker's.

Disclosures

No conflicts of interest declared.

Discussion

We hope to have given the viewer a flavor of the basics of multivariate analysis; interested viewers are encouraged to check out our website. A few choices for parameters in the multivariate analysis were made that can be subject debate to considerable debate. We spared the discussion of these issues in this article to avoid distraction from the major issues. First, we chose the first 6 Principal Components to construct our AD-related covariance pattern. There are theoretical reasons for this choice that we did not discuss [3]. The particular choice of 6 Principal Components though is not critical for our argument: one can chose in the range from 2 to 20 PCs and still obtain superior generalization performance of the multivariate marker in the split-sample simulations. The results are similarly robust with respect to the choice of numbers of subjects in derivation and replication samples. We chose 20 subjects for both groups in the replication sample, but this was purely for mathematical convenience to speed up the computations. Our results about the relative merits of both techniques would hold similarly if the numbers of subjects in the derivation samples were increased.

Second, we only presented the most basic kind of multivariate analysis. Considerable complication with techniques borrowed from the Machine-Learning literature, linear and non-linear transformations prior to the PCA, and various other wrinkles are feasible that could boost the generalization performance even more. For simplicity we have not touched on these possibilities in this article.

Acknowledgements

The author is grateful for NIH grant support:

- NIH/NIBIB 5R01EB006204-03 Multivariate approaches to neuroimaging analysis
- NIH/NIA 5R01AG026114-02 Early AD Detection with ASL MRI & Covariance Analysis

ADNI: Imaging data was provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (NIH U01AG024904). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics ,Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., and Wyeth, as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health(http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

1. Moeller, J. R. and Strother, S. C., J Cereb Blood Flow Metab 11 (2), A121 (1991)
2. Scarmeas, N. *et al*., Neuroimage 23 (1), 35 (2004); Siedlecki, K. L. *et al*., J Int Neuropsychol Soc 15 (6), 973 (2009).
3. Burnham, K. P., Anderson, D. R., and ebrary Inc., (Springer, New York, 2002), pp. xxvi.
4. Moeller, J. R., Strother, S. C., Sidtis, J. J., and Rottenberg, D. A., J Cereb Blood Flow Metab 7 (5), 649 (1987).
5. Habeck, C. *et al*., Neuroimage 40 (4), 1503 (2008); Habeck, C. *et al*., Neural Comput 17 (7), 1602 (2005).
6. McIntosh, A. R., Bookstein, F. L., Haxby, J. V., and Grady, C. L., Neuroimage 3 (3 Pt 1), 143 (1996)
7. McIntosh, A. R. and Lobaugh, N. J., Neuroimage 23 Suppl 1, S250 (2004).