Cell Biochem Biophys (2010) 58:53–67
DOI 10.1007/s12013-010-9093-0

REVIEW

Multivariate Data Analysis for Neuroimaging Data: Overview and Application to Alzheimer's Disease

Christian Habeck · Yaakov Stern · the Alzheimer's Disease Neuroimaging Initiative

Published online: 24 July 2010

Abstract As clinical and cognitive neuroscience mature, the need for sophisticated neuroimaging analysis becomes more apparent. Multivariate analysis techniques 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. 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 functional connectivity in the brain. The covariance approach can also result in greater statistical power when compared with univariate techniques, which are forced to employ very stringent, and often overly conservative, corrections for voxel-wise multiple comparisons. 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 following article attempts to provide a basic introduction with sample applications to simulated and real-world data sets.

Keywords Alzheimer's disease · Multivariate analysis · Principal components analysis · Brain reading · Classification · Cross validation · Nonparametric inference · Split-sample simulations

Introduction

Multivariate techniques have made substantial inroads into cognitive and clinical neuroimaging and are bound to become the accepted *modus operandi* as people have realized the limiting factors of the more commonly used mass-univariate analysis (for a recent review see [1]). The topographic interpretation of multivariate analysis is less clear than of univariate activation maps, which complicates judging the relative merits of both approaches for research questions of cognitive neuroscience aimed at the discovery of neural substrates of brain processes.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or production of this report. A listing of ADNI authors is available at http://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Manuscript_Citations.pdf.

Matlab code for spatial covariance analysis is downloadable at http://groups.google.com/group/gcva.

C. Habeck (✉) · Y. Stern
Cognitive Neuroscience Division, The Taub Institute for Research on Aging and Alzheimer's Disease, Columbia University, New York, NY 10032, USA
e-mail: ch629@columbia.edu

However, multivariate techniques have an empirically verifiable advantage over univariate approaches when it comes to predicting outcome measures from independent data on the basis of previously identified brain–behavioral relationships. This slightly different focus from the traditional goal of cognitive neuroscience has become more prominent in recent years under the rubric of "brain reading." In the context of brain reading, multivariate approaches have been shown to be both more sensitive and more specific than univariate approaches. This is not surprising since multivariate techniques achieve sparse representations of complex data and can identify the robust features that are most important for classification and prediction problems. Non-parametric techniques [2, 3] or standard machine learning techniques like *k*-fold cross validation [4] can aid in this endeavor and are easily performed on modern computers, obviating the pragmatic and historical advantage of easily available parametric statistical inference and model selection that univariate techniques have enjoyed.

However, some disadvantages of the multivariate approach remain, mainly pertaining to higher demands of computational and mathematical literacy on the data analyst, which presents an effective barrier to the more widespread use. Further, after finding the resolve for serious engagement with multivariate techniques, the neuroscience researcher might find herself lost in a large variety of approaches and software packages (as well as acronyms). While the advantages of multivariate over univariate analysis are relatively easy to formulate and demonstrate, the same cannot be said for the large number of approaches within the field of multivariate analysis. In our experience of applying multivariate techniques to a variety of data sets from clinical and cognitive neuroscience, it appears to us that favoring one particular approach and software package across the board, while seemingly comforting and understandable, will often result in a less than optimal way of analyzing the data, i.e., sensitivity and specificity might be less than what they could have been. Frustratingly, the relative merits of different multivariate techniques always depend on the variance structure of the particular data set under consideration, meaning that absolute statements about merits and drawbacks of different multivariate techniques are impossible. The most promising strategy, in our view, is to equip the clinical and cognitive neuroscientist with the tools to arrive at an optimal selection of multivariate approaches herself for the particular data set under consideration. Conceptually, this is not difficult to do: it involves (1) a choice of a meaningful performance metric for methodological comparisons, (2) a variety of pre-determined multivariate prediction or classification tools, and (3) two data sets, one for the derivation of the optimal predictor/classifier, and one for the testing of the prediction/classification (=brain reading) performance of the previously derived classifiers. The choice of the best classifier (or an ensemble of different classifiers) is then readily made on the basis of the best performance in the test data set, as judged by the adopted performance metric. Usually, there is not sufficient data to provide both a derivation and a test data set, but the use of *k*-fold cross validation [4] enables the estimation of generalization performance in *one* data set.

The current article strives to provide a simple introduction to multivariate approaches based on Principal Components Analysis (PCA).[1] Further, it adds to the large body of evidence of the superiority of multivariate techniques by showing a simple application to a clinical data set from the Alzheimer's Disease Neuroimaging Initiative (ADNI). A few disclaimers are probably in order: the review aspects of our article are selective not only in content, but also when it comes to referencing other authors' contributions. Even for all PCA-based approaches, we cannot possibly do justice to all major contributions of recent years; this acknowledgment itself shows the promising and quickly expanding scope of multivariate approaches. We strove to give a basic introduction to PCA-based approaches with some citations of landmark articles and encourage the reader to follow up on these citations independently. Further, since the current article is methodological in nature, the traditional distinction between the Methods and Results is somewhat blurred. We hope that reader whether novice or seasoned practitioner of multivariate techniques will profit from our article.

Basics of PCA and Notational Conventions

First, we give a simple explanation of PCA, the most basic and well-understood form of multivariate decomposition. As we mentioned before, there are many types of multivariate decompositions—we picked PCA since in our opinion it is the best understood of all multivariate decompositions and computationally fast to run, with a clear ordering of the results in terms of variance accounted for. This simplicity is seen by some authors as a vice rather than a virtue, with the justification that neuroimaging data are of such complexity that a simple algorithm like PCA cannot be adequate for illuminating analysis. While this statement is superficially appealing, it neglects to mention that complex tools and algorithms can have a "life of their

[1] The literature on PCA is vast. A good didactic exposition with a historical overview and references can be found at http://en.wikipedia.org/wiki/Principal_component_analysis.

own" and introduce artifacts whose proper assessment demands rigorous pre-testing with Monte-Carlo simulations and test runs on simple real-world data sets that are understood perfectly in terms of their variance structures. Otherwise, the analyst runs the risk of unleashing poorly understood, but complex, tools on even more complex brain data with an insufficient understanding of the ensuing results. For this article, such techniques are thus beyond the scope of our investigation.

Some notational conventions first: matrices are given in capital bold-face, while column vectors are given in lower-case bold-face. Row vectors are just transposed versions of column vectors and no separate notation will be introduced for them. Scalar variables are given in italics. Furthermore, we follow the conventions of the software package Matlab for concatenation of vectors. [**x y**] denotes the assembly of the two column vectors **x** and **y** into a matrix that has 2 columns and as many rows as **x** and **y. [x; y]**, on the other hand, is column vector that has twice as many rows as **x** and **y.** Dimensions of matrix are denoted with a curly bracket, for instance for a 40-by-2 matrix **X,** we can write

$$\{\mathbf{X}\} = 40 \times 2.$$

Transposition is expressed as $\mathbf{X}^{\mathrm{T}}$, so

$$\left\{\mathbf{X}^{\mathrm{T}}\right\} = 2 \times 40.$$

Any data array analyzed in this article assumes a data matrix **Y** with *R* rows, i.e., one row per image voxel (=3-dimensional pixel), and *N* columns, i.e., one column per brain image included in the data set. Usually *N* is several orders of magnitude smaller than *R*. A typical neuroimaging experiment might comprise 40 human participants who are scanned in a functional MRI experiment in 2 experimental conditions. In this case, $N = 2 * 40 = 80$, and the number of voxels *R* usually is on the order of several hundreds of thousands. Thus, the rank of the data matrix **Y** is *N*, and this determines the number of Principal Component (PCs) that follow from a PCA. It is customary, although not necessary, to remove the grand mean image from the data array, reducing the rank to $N - 1$. Further, we assume that all columns of **Y** have been mean-centered. These normalizations assure that voxel-by-voxel and subject-by-subject covariance matrices are just scaled versions of $\mathbf{YY}^{\mathrm{T}}$ and $\mathbf{Y}^{\mathbf{T}}\mathbf{Y}$, respectively.

Next, we will perform the PCA on the data array. Since the rank of the data matrix is *N*, an Eigen decomposition of the voxel-by-voxel covariance matrix $\mathbf{YY}^{\mathrm{T}}$ is impossible since this matrix is rank-deficient. Instead, we perform the Eigen decomposition on the scan-by-scan covariance matrix $\mathbf{Y}^{\mathrm{T}}\mathbf{Y}$, and then obtain the Eigen images by projection. The Eigen equation reads:

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} \quad \mathbf{w}_i = \lambda_i \mathbf{w}_i \quad i = 1, \ldots, N-1,$$

where $\mathbf{w}_i$ are the Eigen vectors in subject space, i.e., the dual of the voxel space, and the associated Eigen values are λ_i. The Eigen vectors have $N - 1$ rows each and can be assembled in a matrix **W**

$$\mathbf{W} = [\mathbf{w}_1 \mathbf{w}_2 \mathbf{w}_3, \ldots,].$$

One can see easily how the projection into voxel space works by multiplying with **Y** from the left to obtain:

$$\mathbf{YY}^{\mathrm{T}} \quad \mathbf{Y}\mathbf{w}_i = \lambda_i \quad \mathbf{Y}\mathbf{w}_i = \lambda_i \mathbf{v}_i.$$

This is the Eigen equation for the voxel-by-voxel covariance matrix $\mathbf{YY}^{\mathrm{T}}$ and the Eigen vectors in voxel space (=brain images or PCs) are conveniently obtained by a simple multiplication of $\mathbf{w}_i$ with **Y** from the left. Again, the PCs can be assembled in a matrix according to

$$\mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \mathbf{v}_3, \ldots,].$$

This matrix has *R* rows and $N - 1$ columns.

With the final assembly of all Eigen values into a matrix according to

$$\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots,)$$

we can express the full data array as

$$\mathbf{Y} = \mathbf{V}\,\mathrm{sqrt}(\Lambda)\mathbf{W}^{\mathrm{T}}.$$

A few noteworthy observations can be made: the PCA achieves a decomposition of the data into one factor (**V**) that is only dependent on the voxel locations in the brain and one factor (**W**) that is only dependent on the subject index. The PCs assembled in **V** are invariant across the group and can serve as basis vectors for a coordinate system in terms of which the data **Y** can be conveniently summarized. They can be visualized as brain images and assign loadings to every voxel location in the brain. For our purposes we will combine the square root of the Eigen value matrix and **W** into on matrix **Z** and rewrite the previous equation as

$$\mathbf{Y} = \mathbf{V}\,\mathbf{Z}^{\mathrm{T}} = \mathbf{v}_1 \mathbf{z}_1^{\mathrm{T}} + \mathbf{v}_2 \mathbf{z}_2^{\mathrm{T}} + \mathbf{v}_3 \mathbf{z}_3^{\mathrm{T}} + \cdots$$

We term the column vectors in **Z,** subject score vectors. The normalization for both PCs and subject score vectors are

$$\mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j = \delta_{ij}, \quad \mathbf{z}_i^{\mathrm{T}} \mathbf{z}_j = \lambda_i \delta_{ij}.$$

Let us summarize what we accomplished:

- the data matrix **Y** was expressed as a product of subject-invariant PCs in **V**, and voxel-invariant subject scores in **Z**;
- PCs are mutually orthogonal, subject score vectors are mutually orthogonal;

- the Eigen value λ_i indicates how much variance the associated *i*th PC accounts for in the data array **Y**; the fraction of the variance accounted for by this PC is computed through division of λ_i by the sum of all Eigen values.

We stress again that PCA is just *one* way of achieving a multivariate decomposition, and we chose it for its relative simplicity and transparent nature. Obviously, for the expression $\mathbf{Y} = \mathbf{V}\ \mathbf{Z}^{\mathrm{T}}$ there is an infinity of choices for **V** and **Z**. PCA imposes orthogonality on the columns of both **V** and **Z**. Other choices like independent component analysis impose statistical independence beyond just second-order moments on either **V** or **Z** or both. Other decompositions might be reasonable and conceivable too, particularly if furnished with clear algorithmic formulations that can be executed on null-data to empirically generate the null-distribution for any test statistic of choice.

One last thing to notice is the following: we explained how PCA achieves the decomposition $\mathbf{Y} = \mathbf{V}\mathbf{Z}^{\mathrm{T}}$, a representation of the data matrix in terms of PCs and their subject scores. The PCs in **V** form an orthonormal basis set; this means that *any* data set **Y*** can be expressed in terms of these components with modified subject scores **Z*** plus a residual term of unaccounted variance, regardless whether **Y*** is the "derivation data set," i.e., original data set from which **V** was derived:

$$\mathbf{Y}^{*} = \mathbf{V}\mathbf{Z}^{*\mathrm{T}} + \mathbf{E}.$$

In an independent "replication data set," subject scores of the PCs assembled in **V** are easily computed according to:

$$\mathbf{Z}^{*} = \mathbf{Y}^{*\mathrm{T}}\mathbf{V}.$$

Any brain–behavioral relationship that was discovered in the derivation data set **Y** and involves the subject scores in **Z**, can now be tested in the replication data set, using the subject scores in **Z*** and the subject variable of interest. This means that rather than relying on statistical inference in the derivation data set, one can check empirically whether the findings hold up in a replication data set—a very powerful additional validity test. We will make use of this feature of prospective application extensively in this article.

One important caveat about PCA that needs to be brought to the practitioner's attention is its susceptibility to outliers. Since PCA operates on the parametrically computed variance–covariance matrix, this susceptibility is not surprising and we have observed it in PET and fMRI data numerous times in practice. Single brain images might contribute an overwhelming portion of the variance, resulting in an abnormally large variance concentration in the first PC (>90%). Essentially, one participant's brain image contributes an overwhelming amount of variance to the data and captures the first PC all by itself. The remaining PCs can account for all remaining brain images, but not in an optimal way since everything is predicated on being orthogonal to the unrepresentative and pathological first PC. Clearly, a better strategy would be to down-weight the contribution of the problematic brain image such that a better representation of *all* images in the sample is achieved in the first few PCs of an optimized PCA. In the field of computer vision, a large variety of just such approaches has been proposed using iterative algorithms or exact closed-form solutions (for instance [5–8]). We will not try to pursue these approaches any further for the sake of brevity. However, when reviewing PCA results the analyst should look for signs of trouble and abnormally large variance contributions. If a brain image produces its' own first PC, the common-sense first line of attack would be to just re-run the analysis without the problematic data point.

A Toy Example

First, we would like to demonstrate a scenario for which multivariate analysis performs better than univariate analysis with a Monte-Carlo simulation. The toy problem defined for this simulation is a simple classification between two age groups. The data set is constructed as follows:

$$\mathbf{Y} = \mathbf{v}\left[\mathbf{z}_{\mathrm{young}}; \mathbf{z}_{\mathrm{old}}\right]^{\mathrm{T}} + \mathrm{E}.$$

The pattern **v** is a two-dimensional binary 0–1 pattern of adjoining squares, comprising 100 * 100 = 10,000 pixels overall. The subject scores for the two age groups are sampled from two normal distributions with equal variance but different means, every subject *i* in the age group is treated identically (Fig. 1):

$$z^{(i)}_{\mathrm{young}} \sim N(-1, 1), \quad z^{(i)}_{\mathrm{old}} \sim N(1, 1) \quad i = 1, \ldots, 50.$$

Further we add Gaussian identically and independently distributed voxel and subject noise, i.e., for any voxel *k* and any subject *i*, we have:

$$\mathrm{E}_{ik} \sim N(0, \sigma).$$

For our simulation, we will now vary the noise amplitude σ and observe the performance of a standard mass-univariate group comparison using a T-test and multivariate PCA.

In Fig. 2, we display the results for the univariate T-test for three noise levels $\sigma = 2, 5, 10$ and also perform a PCA for noise level $\sigma = 10$. The T-test has been corrected for 10,000 comparisons with a Bonferroni correction, setting a threshold of $T = 4.66$.

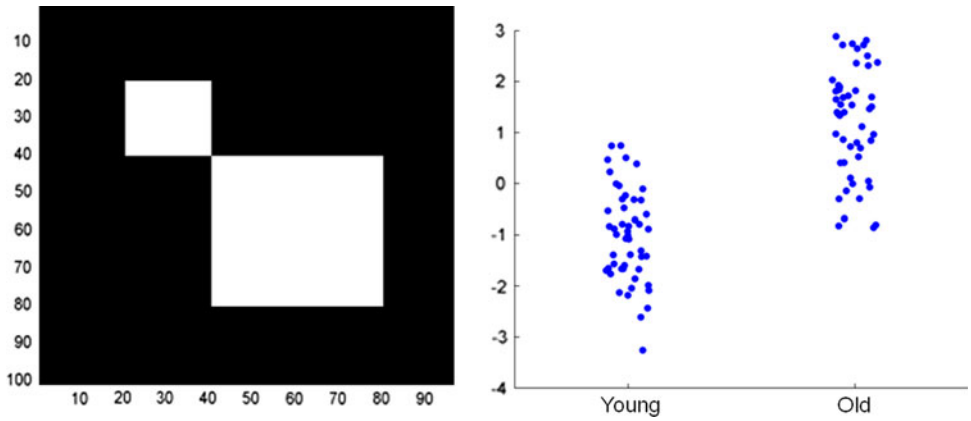

Fig. 1 Visual illustrations of the binary pattern **v** (*left panel*) and the subject scores $\mathbf{z}_{young}$ and $\mathbf{z}_{old}$

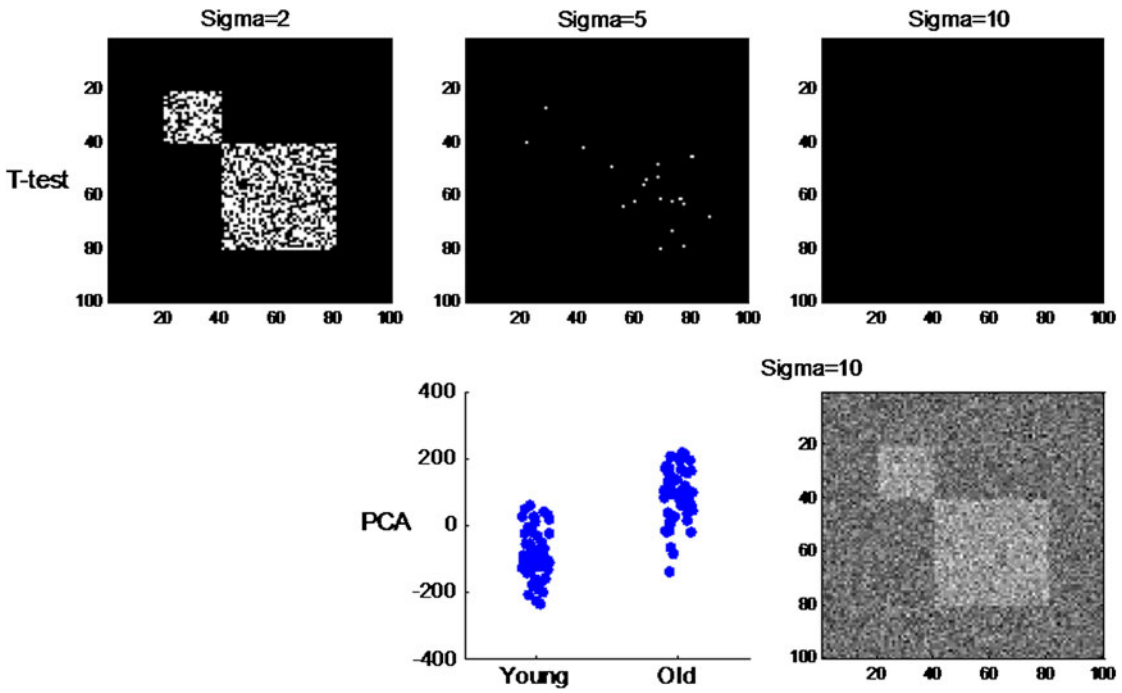

Fig. 2 Simulation results: *upper row*, the thresholded univariate T-fields are shown for noise levels $\sigma = 2, 5, 10$. One can appreciate the decreasing true-positive rate. At $\sigma = 10$, no signal is recovered, while the stringent Bonferroni correction for 10,000 comparisons makes sure there are no false positives. *Lower row*: the results of the PCA are shown for $\sigma = 10$: the subject scores of the first PC show a significant group difference between old and young. Further, the topographic composition of the first PC is visually similar to the binary target pattern

From the figure, one can appreciate how the increasing noise levels gradually cause the univariate T-test to suffer from increasing number of false positives. For $\sigma = 10$, no voxel are caught by the T-test any longer, and the true-positive rate drops to zero. The PCA, on the other hand, as can be seen in the lower panel in Fig. 2, still established a clear group-difference between old and young in the subject scores of PC 1. Also, one can clearly visually recognize a noisy version of the binary pattern in PC 1 itself.

In Fig. 3, the noise level is varied more comprehensively across the range $\sigma = [0, 10]$ in increments of 0.01. The plot shows the true-positive rate for the univariate T-test, as well as the R^2 of the correlations between PC 1 and the binary pattern, and their subject scores. One can appreciate that by $\sigma = 6$, the univariate analysis fails to identify any of the signal voxels. The multivariate analysis, on the other hand, even at $\sigma = 10$ retrieves a first PC that looks topographically similar to the binary pattern, and whose subject scores are very highly correlated with the scores of the binary pattern ($R^2 = 0.91$).

Some observers might be skeptical whether the increased sensitivity of the multivariate analysis looks

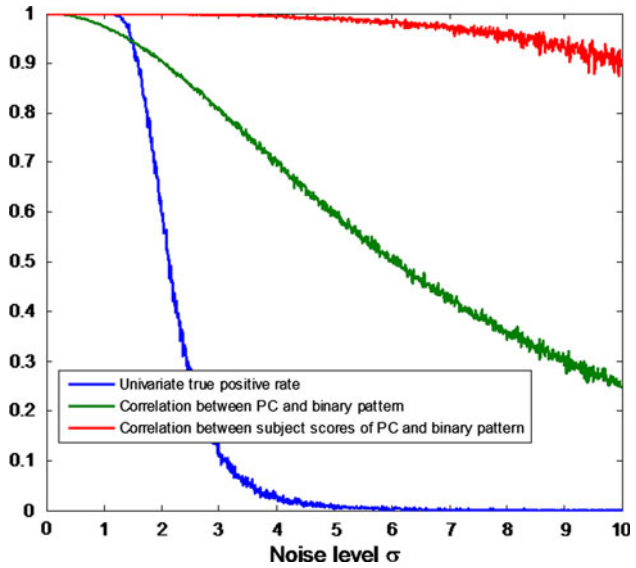

Fig. 3 Comprehensive display of univariate true-positive rate (*blue*), topographic correlation between first PC and binary pattern (*green*), and correlation between subject scores of first PC and binary pattern (*red*)

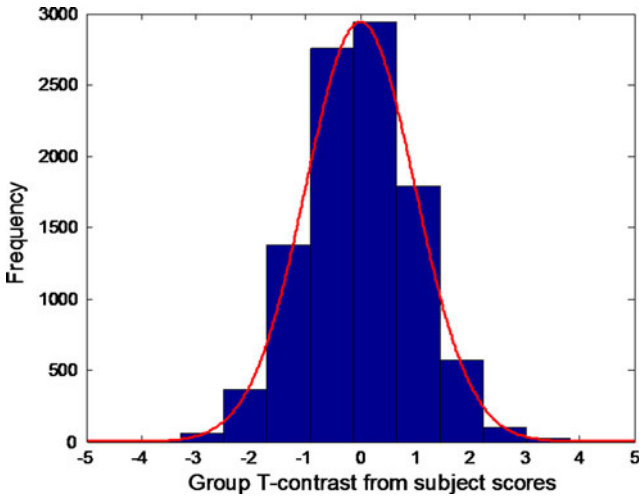

Fig. 4 Empirical histogram generated for the subject score of the first PC obtained in 10,000 Monte-Carlo simulations of Gaussian IID noise and theoretical curve for a T-distribution with 99 degrees of freedom. Increased false positives for the multivariate technique would imply "fat" tails, i.e., a histogram that was much wider than the theoretical T-distribution; fortunately, this is not the case

good only by virtue of increasing the false positives. This can be easily checked by generating pure Gaussian noise, i.e., only retaining the error terms in the Monte-Carlo construction of the data array, and re-applying the PCA.

In Fig. 4, we display the results of 10,000 such noise simulations, and plot the resulting group T-contrast values computed from the subject scores of the first PC. Superimposed on the empirical histogram is the theoretical curve for a T-distribution with 99 degrees of freedom. Increased false positives would imply "fat" tails, i.e., a histogram that was much wider than the theoretical T-distribution; fortunately, this is not the case.

The little simulation just presented in favor of multivariate analysis of course contains one crucial assumption, namely that the pattern of activated voxels is widely distributed. If a substantial part of all image voxels are participating in the pattern, their mutual correlation can be used to separate signal from noise. This does not work for focal activation. For instance, when re-running the simulation with a much reduced signal area in the binary pattern, a 3×3 voxel patch in the center, the univariate analysis—as expected—is unaffected, while the results are changed radically for the multivariate analysis. Once the noise level reaches $\sigma = 1$, both topographic correlation as well as correlation of subject scores between the first PC and the binary pattern have fallen to zero. For such focal activation, i.e., 9 activated voxels out of 10,000, the first PC is dominated by the noise of the remaining 9,991 voxel that carry no signal. The lesson for brain imaging is that if truly focal activation is expected, univariate analysis is more effective than multivariate analysis.

Multivariate Extensions

We outline a variety of sophistications of the simple framework presented above. The first extension will concern the data array used in the multivariate decompositions. The second extension concerns possible transformations of the data array *prior* to any multivariate decomposition.

Data Formats

In all applications shown in this article, we will keep to the simple data structure explained above, i.e., for R voxels and N subjects the format of the data matrix is

$$\{\mathbf{Y}\} = R \times N,$$

implying that there is only one brain image per subject in this data array. This restriction can be relaxed: in most experiments of cognitive neuroimaging a whole time series of T brain images is acquired for each participant. In the standard approach of hierarchical linear modeling, each participant's time-series is reduced two several contrast maps through the use of linear time series analysis, before moving onto the group-level analysis. For our analysis purposes, nothing much changes. The only complication might be that there is now more than one experimental condition. If the experimental design offers C conditions, the group-level data array, after estimation of within-subject contrast maps, would have the format:

$$\{\mathbf{Y}\} = R \times (N \cdot C),$$

that is, the data array features one brain image per participant per condition, and the subject and condition indices have been nested in the column index of $\mathbf{Y}$.

If, on the other hand, one refrains from reducing the full data array by time-series analysis or any other means, the full data array is a third-order tensor,

$$\{\mathbf{Y}\} = R \times T \times N$$

and any condition information is implicitly contained in the time dimension. Such a data array is more complicated than a simple matrix and a multivariate decomposition less straightforward. Possible approaches to deal with these three-way data are the parallel factor analysis (PARAFAC) framework [9] or tensorial ICA [10].

More easily and more commonly, the time and voxel dimensions are collapsed into the row index, leading to a reduced data matrix,

$$\{\mathbf{Y}\} = (R \cdot T) \times N.$$

Mathematically, the treatment of such a data matrix is identical to our example above. The resulting PCs though are no more merely brain images, but rather "brain movies," i.e., $N - 1$ time series of brain images. The important feature of group-invariance though is preserved: the series of $N - 1$ Eigen movies do not contain any subject information. This information is still contained in the subject scores. Each score now quantifies to what extent a subject's time series expresses the associated Eigen movie.

Pre-Transformation with Design Matrix

Apart from considerations about the format of the data array, multivariate analysis often uses a linear pre-transformation in form of a design matrix $\mathbf{X}$. $\mathbf{X}$ has the same number of rows as $\mathbf{Y}$, but already achieves a dimensionality reduction by having fewer columns than $\mathbf{Y}$. We denote the number of columns in $\mathbf{X}$ as P, and call the columns "predictors."

The data matrix $\mathbf{Y}$ is multiplied with $\mathbf{X}$ from the right, and we find

$$\{\mathbf{YX}\} = R \times P.$$

For our PCA discussion above, this does not really introduce many complications. In the above formalism, matrix $\mathbf{Y}$ can just be substituted by $\mathbf{YX}$. Instead of submitting N brain images to a PCA, we are now only analyzing P images, i.e., the number of predictors in the design matrix determines how many PCs can be recovered. The representation of the transformed data matrix $\mathbf{YX}$ can then be written as before as

$$\mathbf{YX} = \mathbf{VZ}^{\mathrm{T}}$$

but the score matrix $\mathbf{Z}$ now consequently only has P rows and columns.

This approach is widespread in the literature can be found with the labels multivariate linear modeling (MLM) [11, 12] or partial least squares (PLS) [13, 14]. PLS also has a well-formulated spatiotemporal version (stPLS) that uses the full voxel- and time-information to produce Eigen movies [15].

We have our own approach, ordinal trend canonical variates analysis (OrT/CVA) [16] that seeks to derive monotonically changing activation patterns on a subject-by-subject basis in repeated-measured design with a specially formulated design matrix.

The purpose of the pre-transformation with $\mathbf{X}$ is a simplification and prior dimensionality reduction *before* a PCA is even applied. The implicit assumption is that the multiplication with $\mathbf{X}$ removes data variance that is uninformative and only contributes noise that might otherwise hamper the detection of interesting effects if the full data array $\mathbf{Y}$ was submitted to a PCA. The dimensionality reduction achieved by some common design matrices can be quite substantial. A hypothetical, but nevertheless representative, example might be the following: assume 40 human participants, scanned in 2 experimental conditions. Based on 3 locations of interest, "voxel seeds" are used to compute across-subjects correlational images of the activation in these 3 voxel locations with the rest of the brain in both experimental conditions. This means that the effective data matrix $\mathbf{YX}$ has the format

$$\{\mathbf{XY}\} = R \times 6.$$

The rank of the data matrix is thus reduced from $40*2 = 80$ to 6, and the PCA is now executed on six brain images, rather than 80. The score matrix $\mathbf{Z}$ now contains 6 Eigen vectors that quantify to what extent the 6 PCs of $\mathbf{YX}$ load onto the predictors of $\mathbf{X}$, rather than participants in $\mathbf{Y}$.

The judgment whether a particular design matrix is appropriate or not cannot be answered by mathematical criteria. Using a low-dimensional design matrix is never "wrong" on mathematical grounds, but it might unhelpful in recovering interesting activation, particularly if it rids the data of precisely such information in the first place. Projecting the data into a low-dimensional sub space ensures that the results are easily interpretable for the analyst, but it might not reveal the most informative aspects of the data. For exploratory analyses where a priori guidance of prior literature or well-established models is lacking, overly restrictive design matrices, therefore, have less of a role to play.

However, we do not want to be too critical of the design-matrix approach. Strong a priori insight might provide

enough guidance in selecting low-dimensional design matrices. For our hypothetical example, the analyst might have a well-founded interest in a multivariate description of the similarities and differences of the six correlational seed images, rather than getting a full description of all subject and task effects in the data, some of which are bound to represent task-unrelated variance. In the absence of such strong guiding information though, we would advise care in selecting design matrices. Further, it does not hurt to do the PCA on the full data array and check whether the subject scores of the PCs with the largest variance contribution show any correlation with either task or nuisance variables. In our opinion, more exhaustive knowledge and understanding of the data is a good thing, whether the data suffer from artifacts or not.

Statistical Inference

For completeness, we quickly sketch current practices of performing statistical inference for multivariate PCA-based techniques, with the caveat that, again, we cannot speak for all possible PCA-based approaches in existence.

For multivariate analysis, statistical inference can concern (1) the topographic composition of the covariance patterns, and (2) the subject scores of the covariance patterns. The first item mainly targets the question "Which voxels are reliably activated/contained in my covariance pattern?" The second item has more facets, and one could ask a variety of questions that each entail a test of a different null-hypothesis and involve the pattern scores, like "Is the relationship between subject scores of my pattern and a particular subject variable statistically significant?," "Are subject scores of my pattern significantly related to the experimental design?," "Does my pattern account for a statistically significant portion of the variance in the data," or "Are subject scores of my pattern significantly different from zero in a particular experimental condition?"

First, we consider statistical inference concerning the pattern's topographic composition. As we mentioned before, for most neuroimaging experiments there are far fewer observations (=images) than voxels. This means that there is no parametric formula that can be applied to decide whether a voxel is significantly activated or not. Several approaches use a semi-parametric bootstrap estimation procedure [2] to assess the reliability of individual voxels' contribution in the covariance pattern. The advantage is that the bootstrap is conceptually easy to understand. It consists of the repeated execution of the derivation of the point-estimate covariance patterns, but each time the data is re-sampled with replacement from the original pool of subjects. This means that some subjects are represented more than once in the bootstrap sample, while others are totally dropped. On this re-sampled data, all steps that were employed for the derivation of the point-estimate pattern are executed again.

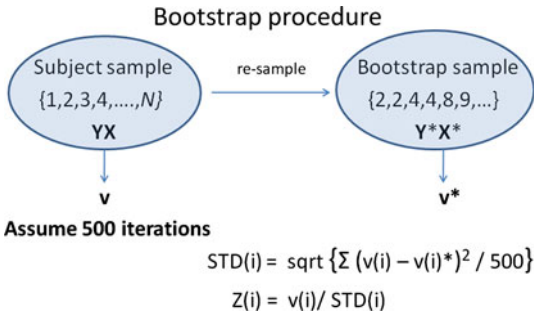

Fig. 5 Schematic figure for illustration of the bootstrap procedure for assessing the robust of individual voxel weights in the covariance pattern. Sampling from the pool of subjects with replacement results in some subjects being dropped, while others are represented more than once in the associated data and design matrix **Y*** and **X***, respectively. The algorithm that was applied to **XY** to derive a covariance pattern **v** is performed on **Y*X*** to obtain **v***. Resampling and subsequent pattern derivations are repeated $\sim$ 500 times. From all 500 bootstrap patterns, a *Z*-map can finally be computed

Figure 5 shows the bootstrap procedure schematically. We assume a covariance pattern **v** was derived from the unperturbed data matrix **Y.** The data sample is re-sampled with replacement to produce the bootstrap data matrix **Y*** and design matrix **X***, and the pattern derivation algorithm is applied to derive a new bootstrap pattern **v*.** The re-derivation step is repeated many times ($\sim$500 times). Finally a Z-score can be computed for each voxel location *i* as the ratio of the voxel weight divided by the bootstrap-incurred standard deviation around this point estimate

$$Z(i) = v(i)/\mathrm{STD}(v(i)),\ i = 1, \ldots, R.$$

This computed *Z*-statistic roughly follows a standard-normal distribution.[2] A one-tailed p-level of 0.001 implies a threshold of $|Z| > 3.09$.

The second type of statistical inference involves pattern scores. As our sample questions demonstrated many more tests and associated null-hypotheses can be investigated. Most of these will demand non-parametric permutation tests, i.e., the null-hypothesis distribution for any statistic of choice is generated from the data itself by destroying the subject-group or –condition assignment, particularly when a design matrix **X** is used that encodes subject information subsequently used in the particular test of interest. The following exceptions to the non-parametric testing requirement are easy and clear to formulate and follow

[2] The larger the number of voxels in the data array, the more the empirical bootstrap distribution of individual voxel weights looks standard-normal. When the number of brain regions in the array is small, i.e., similar to, or a low-integer multiple of, the number of observations, the bootstrap distribution can deviate substantially from a standard-normal distribution.—Repeated personal observation by the authors.

directly from the avoidance of any independence violation ("double dipping" [17]):

1. When no design matrix is applied prior to the PCA, the statistical significance of a brain-behavioral relationship between pattern scores and subject variables can be assessed using standard parametric statistics;
2. When a covariance pattern is applied prospectively to a replication data set, i.e., a data set that it was *not* derived from in the first place, any brain-behavioral relationship between the resulting pattern scores and subject variables can also be assessed using standard parametric statistics. This is even possible when the pattern derivation utilized a design matrix that incorporated the particular subject variables in question for the derivation data set.

A Caveat About the Interpretation of Multivariate Patterns

Before we go on and apply PCA and a variety of classifiers on some real-world data, we close with a remark about the interpretation of covariance patterns. These considerations are equally valid for Independent Component Analysis [18] or any other multivariate decomposition routine that uses a feature like orthogonality or statistical independence to derive components in terms of which the data can be described. Although it is tempting, one should be careful in assigning biological meaning to these components, particular in absence of any observed brain-behavioral relationships. After all, the feature of orthogonality/statistical independence follows necessarily from the PCA/ICA step itself; even when applied to meaningless statistical noise, the resulting PCs or ICs will display mutual independence, but in this case it is obvious that they cannot serve as the neural substrates of any meaningful cognitive or biological processes. For real-world data the problem is only slightly better: now we have meaningful signal mixed in with statistical noise, but it is unlikely that the particular decomposition adopted achieves a neat break-down into separate components that exclusively capture either signal or noise. The components are most likely made up of varying mixtures of both. Further, whether neural substrates of different cognitive or biological processes in the brain display statistical independence in the way that PCA/ICA demand is a research question with an empirical answer, and cannot be taken as a given.

In summary, it is worth keeping this in mind: PCA and ICA are useful tools for dimensionality reduction and achieve sparse representations of complex data. ICA in particular has been used successfully for artifact detection and source dimensionality estimation [19]. However, it is less clear whether the feature of uncorrelated or statistically independent sources (and the metaphor of the "cocktail party problem" [18]) is appropriate for brain function, and consequently whether the components resulting from an application of PCA/ICA to brain data can themselves be interpreted as neural substrates of brain processes. In other words: although PCA/ICA will always come up with separate components, these might not represent separate networks. Additional converging evidence for heaving identified a network is the successful prediction of subject variables associated with the brain processes in question. Such prediction has fortunately become more of a focus in neuroimaging in recent years, and, in our opinion, is more valuable than the fitting of sophisticated data models without any subsequent prediction in independent data [20].

A Real-World Example from the Alzheimer's Disease Neuroimaging Initiative

We now put univariate and multivariate analysis to the test on real-world data. The analysis that follows is similar to already published results on a different data set [21], but in the following we investigate replication more extensively in split-sample simulations and also include additional model selection tools.

We downloaded 40 FDG-PET scans of early Alzheimer's disease (AD) patients and 40 FDG-PET scans from healthy control (HC) participants from the website of the ADNI.[3] The mean age of Alzheimer patients at the time of the scan was 75.2 ± 1.1 years, and for the HCs it was 75.5 ± 0.7 years. All participants had a comprehensive clinical and neuropsychological evaluation performed on them, but for the simple application discussed here only their overall Clinical Dementia Rating (CDR) scale measurement was important ($CDR = 1$ for Alzheimer's, $CDR = 0$ for HCs).

Comparing Classifiers Through Split-Sample Simulations

Our data set of 40 AD and 40 HC scans is an ideal playground to test out a variety of univariate and multivariate classifiers in split-sample simulations. For these simulations, we divide the data randomly into a derivation sample of 30 AD and 30 HC scans, while the remaining 10 AD and 10 HC scans serve as the replication sample. Any diagnostic classifier can be derived in the derivation sample, and subsequently tested in the replication sample. The whole procedure, i.e., random partitioning of the data, derivation of classifiers in one part of the data with subsequent test of the classifier in the remaining part of the

[3] The website is: http://www.loni.ucla.edu/ADNI/.

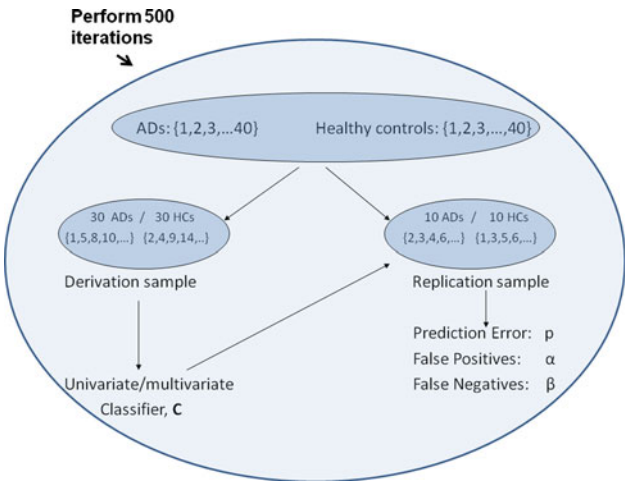

Fig. 6 Schematic figure to illustrate our split sample simulations for the empirical comparison of different classifier's prediction performance. The data sample of 40 ADs and 40 HCs is split into a 30/30 derivation, and a 10/10 replication sample. A classifier **C** is derived in the derivation sample and then prospectively applied to the replication sample with predictions of the class labels $\{\pm 1\}$, corresponding to the diagnostic status "AD" (label = 1), or "HC" (label = −1). Total prediction error, false-positive rate and false-negative rate are recorded each time and enable an empirical comparison of different classifiers' performances

data, can be repeated many times to get a better idea about the generalization of performance of the classifiers. The success of the diagnostic prediction in the replication samples can be recorded and enables an empirical comparison of all classifiers included in the simulations. The advantage of this approach is the total absence of any reliance on data models and the corresponding statistical inference.

Figure 6 shows the split-sample procedure. The full data set of 40 AD and 40 HC images is divided randomly into a 30/30 derivation sample and a 10/10 replication sample. Any classifier of interest can be derived in the derivation sample and then tested in the 10/10 replication sample that was left out of the derivation. Total error rate p, false-positive rate α, and false-negative rate β are then recorded for the prediction made in the replication sample and can be compared for all classifiers that were included in the split-sample simulation. 500 iterations are run of this procedure to avoid any particular sampling biases.

We give a minimum of symbolic notation for our classifiers used in this article. In general a classifier is a mapping from the voxel space of neural images to a binary label with a particular algorithm C applied to a neural image $\mathbf{y}$, such that a prediction of a label $\{\pm 1\}$ results,

$$C(\mathbf{y}) = \{\pm 1\}.$$

In the definition, we can include a vector of parameters $\boldsymbol{\theta}$, and modify this expression to

$$C(\mathbf{y};\theta) = \{\pm 1\}.$$

Different classifiers have different sets of parameters and we will give an exhaustive listing of all parameters for each classifier below.

Univariate Classifier

Definition of the univariate classifier is the simplest. We give the steps of the algorithmic recipe for the derivation below:

1. perform a T-test between the 30 AD and 30 HC images in the derivation sample;
2. pick the voxel j that shows the largest relative deficit in the AD patients;
3. choose a decision threshold T with maximum sensitivity in the derivation sample such that at most one HC subject is misclassified as AD.

This means there are two parameters in our univariate classifier: a voxel location j and a decision threshold T. The classifier and its application a brain image $\mathbf{y}$ in the replication sample can then be denoted as

$$C(\mathbf{y};j,T) = \text{sign}(T - \mathbf{y}(j)) = \{\pm 1\}.$$

Basically, voxel j is checked in the replication brain image and if its signal level falls below threshold T, the image is classified as AD, i.e., it is assigned a label of +1. This is done for every image in the replication sample.

Multivariate Linear-Discriminant Classifier

Derivation of the multivariate linear-discriminant classifier is slightly more involved.

1. Perform PCA on combined 30 AD and 30 HC images in the derivation sample and obtain matrices **V** of PCs and **Z** of subject scores.
2. Pick a set of PCs **SET**.
3. Perform linear discriminant regression and use the labels $\{\pm 1\}$ as the dependent variable and the subject scores **Z**(:,**SET**) as the independent variables; obtain regression weights β_i, $i = 1,\ldots, s$.
4. Construct corresponding linear discriminant pattern **v** as a linear combination of the PCs indicated in **SET**.

$$\mathbf{v} = \sum \beta_i \mathbf{v}_i$$

5. For the expression of the discriminant pattern in the derivation sample, $\mathbf{Y}^{\mathrm{T}}\mathbf{v}$, choose a decision threshold with maximum sensitivity such that at most one HC is misdiagnosed as AD.

One can appreciate that now we have two parameters: a covariance pattern **v**, and a decision threshold T. The linear-discriminant classifier can now be applied to an image **y** from the replication sample as

$$C(\mathbf{y};\mathbf{v},T) = \mathrm{sign}\left(\mathbf{y}^{\mathrm{T}}\mathbf{v} - T\right) = \{\pm 1\}.$$

Thus, prospective application of this classifier to a brain image in the replication sample entails computing the level of expression of the pattern in the brain image and comparing it to the threshold T.

We have conveniently neglected one more implicit parameter: the set of PCs **SET** used in the pattern construction. Once a choice for **SET** has been made, a covariance pattern can be constructed and **SET** itself does not appear as a parameter in the final form of the classifier. However, the optimal choice for SET is far from trivial since there are so many possibilities that an exhaustive search quickly becomes impossible. For instance, if there are 20 PCs, there are

$$2^{20} - 1 = 1048575$$

possible choices to select a subset among these 20 PCs. This combinatorial explosion necessitates some strategies to limit the search space. We confine our search to contiguous sets of PCs like {1,2,3,4,…,s}. This drastically reduced our number of possible choices from $2^N - 1$ to N. As expected, there are many approaches to settle on an optimal choice for **SET** [22]. One could choose an information-theoretic criterion like AIC, BIC, Minimum Description Length or C–p Mallow's criterion [23, 24]. Further, one could chose an empirical approach of k-fold cross validation [4] to derive the optimal PC-set that yields the best replication in the left out data folds, already in the derivation sample. The advantage of the latter is that it is completely model free and does not rely on any assumptions; the drawback, however, is that it is computationally more expensive. For a k-fold cross-validation procedure, any PCA and classifier derivation has to be run k times, and the prediction error is computed on the average across the k left out data folds. K-fold cross validation thus roughly represents a k-fold increase in computational expense compared to any information-theory based approaches.

The problem of the optimal subspace selection is very important, but not the main focus of this paper. To give a quick flavor of the possibilities and enable a simple comparison of model-free and information-theory based approaches, we constructed the linear-discriminant classifier both using fivefold cross validation and minimization of the AIC criterion. For the fivefold cross-validation procedure, we picked the total prediction-error rate as the loss function to be minimized, i.e., we chose the set of PCs that gives the lowest prediction error in the left out data fold, regardless of whether the error is a false positive or a false negative. For the AIC computation, we employed a small sample correction computed for the linear discriminant regression as explained in [23], and picked the PC-set with the lowest AIC value. For both cross validation and AIC optimization, we only considered the PCs that have Eigen values bigger than unity, further restricting the number of PC-sets that need to be tested. We tested the effect of admitting all PCs, but found that, in addition to prolonging the computation time, it always resulted in over-fitting for both approaches (results not shown).

For notational clarity, we dropped argument and parameters in the notation, and settled on the following classifiers,

1. **C-UNI**: univariate classifier;
2. **C-LD/AIC**: multivariate linear-discriminant classifier using AIC-based subspace selection;
3. **C-LD/5CV**: multivariate linear-discriminant classifier using fivefold cross validation for subspace selection.

Results of Split-Sample Simulations

We performed 200 iterations of the split-sample simulations and recorded the total-error rate p, the false-positive rate α, and the false-negative rate β. We display the mean results in Table 1.

The table shows that both linear-discriminant classifiers do better in all aspects of performance than the univariate method. Further, for both multivariate techniques it appears that the added effort of fivefold cross validation pays off and gives lower total error rates and false negatives than for the classifier constructed using the AIC criterion, while the false-positive rate is virtually the same.

Because it is easy to compute, we also took a majority vote of all 3 classifiers as

Table 1 Replication performance of 3 classifiers and a majority vote of all 3 classifiers as recorded in a split-sample simulation with 500 iterations

	C-UNI	C-LD/AIC	C-LD/5CV	Vote
Total error p	0.301 ± 0.005	0.220 ± 0.004	0.209 ± 0.004	0.213 ± 0.005
False positives α	0.122 ± 0.006	0.097 ± 0.005	0.103 ± 0.005	0.090 ± 0.005
False negatives β	0.480 ± 0.008	0.340 ± 0.009	0.315 ± 0.008	0.336 ± 0.008

The univariate classifier performs noticeably worse than the alternatives

Table 2 Prediction error of 3 classifiers and a majority vote of all 3 classifiers as recorded in a split-sample simulation with 500 iterations, with deliberate labeling errors that were introduced in the derivation sample

	C-UNI	C-LD/AIC	C-LD/5CV	Vote
SWAP = 1	0.317 ± 0.004	0.251 ± 0.004	0.230 ± 0.004	0.238 ± 0.004
SWAP = 2	0.331 ± 0.004	0.277 ± 0.004	0.245 ± 0.004	0.260 ± 0.004
SWAP = 3	0.345 ± 0.004	0.309 ± 0.005	0.277 ± 0.005	0.295 ± 0.005

"SWAP = *N*" implies that *N* patients were labeled as HCs, and *N* HCs were labeled as AD patients, resulting in 2 *N* labeling errors in total

$$\text{Vote} = \text{sign}(\{\text{C}-\text{UNI}+\text{C}-\text{LD}/\text{AIC}+\text{C}-\text{LD}/5\text{CV}\}/3).$$

One can see from the table that the this majority vote comes close to the performance of C-LD/5CV. Combining classifiers in this manner might be a good strategy to hedge against any particular deficiencies in any one of them and make the results more robust. We can test this further by introducing deliberate labeling errors into the derivation sample in our simulations. We conducted the split-sample simulations again with the same parameters as before, but one crucial difference: varying numbers of AD patients and HC subjects were swapped between the groups, keeping the overall number of nominal AD patients and controls constant. We swapped 1, 2, or 3 persons between the groups, meaning we introduced a total of 2, 4, or 6 labeling errors.

Table 2 shows the results for these simulations for the total prediction error. One can appreciate that the linear-discriminant classifier with fivefold cross validation is still the best performer, but the majority vote presents a good strategy to limit the impact of the labeling errors on the prediction success in independent data. While this example might appear somewhat contrived since subjects of uncertain diagnostic status should normally always be left out of any classification, there are other factors that might hamper the classification and are hard for the analyst to correct for. Taking the majority vote of several classifiers is an easy first defense against such difficulties.

Conclusion

This article gave a basic overview of PCA-based multivariate approaches with demonstrations on simulated and real-world data that demonstrated better sensitivity and replicability of multivariate over mass-univariate approaches. This fact by now has become well established in the neuroimaging community, and our demonstrations partly underline what has been shown in numerous other publications as well (e.g., [13, 14, 25–30]). Comparative surveys of univariate and multivariate approaches to map neural substrates of cognitive processes have often suffered from clear and understandable performance metrics, and instead had to invoke appeals to functional connectivity arguments and Occam's razor to make the case for multivariate approaches. Applications of brain reading, which has the goal of predicting subject information, like diagnostic AD status, *from* brain data, rather than mapping it *in* brain data, offers predictive success in independent data as a metric for comparative apples-to-apples methods' evaluation, and thus does not need appeals to functional connectivity etc. Our split-sample analysis of FDG-PET data obtained from the ADNI study demonstrated that diagnosis of AD in independent data clearly necessitates a multivariate approach.

We can speculate why spatially correlated brain signals carry more information about the diagnostic status than any voxel-wise signal, and the reasons are three-fold.

1. *Low-rank data*: this first reason applies to any neuroimaging study that acquires brain-wide data and has nothing to do with functional connectivity or co-activation of areas; usually in neuroimaging experiments the number of observations (=number of brain images) is small compared to the number of variables (=number of voxels). This means that conducting analyses on a voxel-by-voxel basis has to involve redundancy, by definition. *There has to be correlation between the voxels*, purely on account of the grave imbalance between the number of variables and observations. Picking key voxels to predict outcome measures, like diagnostic status, is thus bound to be inefficient compared to an approach that uses a dimensionality reduction like PCA first.
2. *Spatially spreading pathology*: the second reason involves the assumption of spreading disease pathology, for any neurodegenerative disease; even if no brain areas share any mutual "communication" of any sort, focal deficits induced by the disease process that gradually spread to neighboring areas will be more easily detectable with multivariate analysis. Particularly in the face of noisy signals, the whole pattern of the spreading signal deficit can probably be picked up earlier and more reliably before a single regions' deficit has reached a detectable threshold.
3. *Functional connectivity*: the last reason involves genuine communication of brain areas beyond just spatially spreading activation or de-activation. The

scenario of brain areas that actively interact with one another lends an even stronger rationale for multivariate analysis and additional verification of network activity that can account for behavioral performance in cognitive tasks on a subject-by-subject basis. Admittedly, this is an ambitious research program, and the examples presented in this study cannot speak to functional connectivity in this strict sense, since they only involved group-level derived covariance patterns. Nevertheless, as we have seen, even without postulating functional connectivity, multivariate analysis has a major role to play in neuroimaging.

We close our report with several comments about multivariate analysis in the context of the rapidly expanding field of brain reading: numerous studies and comparative surveys of different classifier and predictors have been presented in recent years that exceed our simple linear-discriminant classifiers in complexity and power, and have been applied both at the group level and within subjects (e.g., [31–37]). It is impossible to do justice to all contributions here, but it might helpful to point out some caveats and challenges for further methodological research in this exciting field.

As we mentioned, with the premise of brain reading the focus of neuroimaging analysis is no longer to find a neural correlate $\boldsymbol{y}$ of a cognitively or clinically relevant subject variable x as

$$\mathbf{y}(x)$$

but instead to find a reasonable classifier or predictor function such that

$$C(\mathbf{y}) = x$$

is true. The reader can appreciate that, mathematically, C looks like an inverse function to the neural correlate $\boldsymbol{y}$, but in practice there might be many instantiations of C, many of which might involve formulation of complicated algorithms that ensure very good predictive success and many of which deliberately ignore features in the data. A strict mathematical inversion of C might therefore be practically impossible or only constitute part of the data, i.e., the part that is most relevant to the diagnostic classification. The formulation of a reasonable classifier C is easier, and its success easier to evaluate, than the formulation of an appropriate data model $\boldsymbol{y}(x)$, particularly in exploratory data analysis. For $\boldsymbol{y}(x)$ there is usually no gold standard which enables empirical evaluation with techniques like cross validation or replication in independent data. This means that exploratory neuroimaging studies that cannot build on a well established literature and employ estimation techniques with a single a priori defined data model are in danger of informing the analyst about the data model at the expense of neuroscientific phenomenon to be studied [20]. However, concerning the selection of an appropriate parameterization and technique for classification (which is usually called "model selection" and has nothing to do with the term "data model" used above), this is not a problem as we have seen in this paper. The internal structure of the classifier is not of great importance in any case: this might make the classifier un-interpretable or weaken the stability of any associated neural substrates [38], but simultaneously frees the analyst of unrealistic or unfounded assumptions. Accurate prediction of the outcome measure of interest takes priority, allowing fusion of data from different modalities as well as meta-algorithms like bootstrap aggregating [39] or boosting [40].

There are some caveats and challenges: the first practical point concerns the appropriate loss function when trying to perform cross validation for model and technique selection when deriving the best classifier. In our example, we had equal number of AD patients and HCs. Our loss function that was used in the subspace selection with cross validation was the total prediction error. This error was just the average of both type-I and type-II error

$$p = (\alpha + \beta)/2,$$

which is appropriate when selecting the best model. As has been pointed out though [41], total prediction error is not a good performance metric in the presence of large imbalances in the class strengths of the training labels. For instance, if the derivation sample consists of 90 AD patients and 10 HCs, the prediction error changes to

$$p = \alpha * (10/100) + \beta * (90/100).$$

A total prediction error of 10% could be the consequence of a very liberal threshold that tolerates 100% false positives (all 10 HCs diagnosed as AD) in order to ensure that the false-negative rate is zero. The total prediction error for this scenario is obviously quite misleading and would produce a biomarker with terrible characteristics—anybody would be diagnosed with AD. Equally weighted averages of α and β are would be preferable and more meaningful.

In general, the choice of the loss function in the cross-validation procedure has an impact on what model and classifier is selected. For symmetry, we advocate that loss functions always encode the predictive success that matters most to the analyst. In our case, this was the successful diagnosis of AD. A different loss function, for instance the residual unaccounted variance in the data, in our opinion does not make sense to use here, even if it is easier to compute. After all, we are interested in the correct diagnosis of Alzheimer's, rather than giving a complete account of the neuroimaging data.

Secondly, with the easy empirical evaluation of different classifiers' prediction performance, one quickly realizes that the ranking of classifiers' performance can differ

substantially across different data since it critically depends on the variance–covariance structure of the data under consideration. This means that across-the-board statements about the relative merits of different classifiers are suspect. In our own anecdotal experience with Support Vector Machines, Linear Indicator regression, Linear and Quadratic Discriminants, Decision Trees, Naïve Bayes Classifiers and Nearest-Neighbor techniques [4] applied to different clinical and cognitive data sets in both fMRI and PET, we have seen that some crude and very general rules emerge (for instance, "Nearest Neighbor and Naive Bayes are often worse than everything else"), but that otherwise no fixed conclusions can be drawn. Further, and more disconcertingly, relative performance differences as estimated by cross validation within a derivation sample are often not vindicated when testing the classifiers in independent replication data. This means that classifier *A* might give better cross-validation performance than classifier *B* in the derivation sample, but *B* might perform better than *A* when applying both to independent data. In absence of a better theoretical understanding how the comparative performance of different classifiers depends on the variance–covariance structure in the data, it behooves the analyst to be careful and always look at a variety of classifiers, and possibly take an ensemble vote of all of them. For methodological papers, the challenge is similar: unveiling a new multivariate technique is most informative when compared, at least in discussion, if not in actual performance, with other *multivariate* techniques on a variety of representative data sets. The superiority of multivariate over univariate analysis for the majority of neuroimaging applications has by now been unequivocally established. The next step for the community is a better understanding of relative merits within the large class of multivariate techniques.

Acknowledgments 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(www.fnih.org <http://www.fnih.org/> <http://www.fnih.org <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. C. Habeck acknowledges grant support from NIH/NIBIB 5R01EB006 204-03 (Multivariate approaches to neuroimaging analysis) and NIH/NIA 5R01AG026114-02 (Early AD Detection with ASL MRI & Covariance Analysis).

References

1. O'Toole, A. J., Jiang, F., Abdi, H., Penard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience, 19*, 1735–1752.
2. Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York: CRC Press LLC.
3. Good, P. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York: Springer.
4. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
5. Heo, G., Gader, P., & Frigui, H. (2009). RKF-PCA: Robust kernel fuzzy PCA. *Neural Networks, 22*, 642–650.
6. Hubert, M., & Engelen, S. (2004). Robust PCA and classification in biosciences. *Bioinformatics, 20*, 1728–1736.
7. Rajagopalan, A. N., Chellappa, R., & Koterba, N. T. (2005). Background learning for robust face recognition with PCA in the presence of clutter. *IEEE Transactions on Image Processing, 14*, 832–843.
8. Serneels, S., & Verdonck, T. (2008). Principal component analysis for data containing outliers and missing elements. *Computational Statistics & Data Analysis, 52*, 1712–1727.
9. Harshman, R. A., & Lundy, M. E. (1994). PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis, 18*, 39–72.
10. Beckmann, C. F., & Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *Neuroimage, 25*, 294–311.
11. Worsley, K. J., Poline, J. B., Friston, K. J., & Evans, A. C. (1997). Characterizing the response of PET and fMRI data using multivariate linear models. *Neuroimage, 6*, 305–319.
12. Zarahn, E., Rakitin, B., Abela, D., Flynn, J., & Stern, Y. (2007). Age-related changes in brain activation during a delayed item recognition task. *Neurobiology of Aging, 28*, 784–798.
13. McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. (1996). Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage, 3*, 143–157.
14. McIntosh, A. R., Chau, W. K., & Protzner, A. B. (2004). Spatiotemporal analysis of event-related fMRI data using partial least squares. *Neuroimage, 23*, 764–775.
15. McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *Neuroimage, 23 Suppl 1*, S250–S263.
16. Habeck, C., Krakauer, J. W., Ghez, C., Sackeim, H. A., Eidelberg, D., Stern, Y., et al. (2005). A new approach to spatial covariance modeling of functional brain imaging data: ordinal trend analysis. *Neural Computation, 17*, 1602–1645.
17. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience, 12*, 535–540.
18. Stone, J. V. (2002). Independent component analysis: An introduction. *Trends in Cognitive Sciences, 6*, 59–64.

19. Beckmann, C. F., & Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging, 23*, 137–152.
20. Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*, 199–231.
21. Habeck, C., Foster, N. L., Perneczky, R., Kurz, A., Alexopoulos, P., Koeppe, R. A., et al. (2008). Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease. *Neuroimage, 40*, 1503–1515.
22. Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis, 49*, 974–997.
23. Burnham, K.P., Anderson, D.R., & ebrary Inc. (2002). *Model selection and multimodel inference a practical information-theoretic approach (Vol. xxvi)*. New York: Springer, 488 pp.
24. Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.
25. Moeller, J. R., & Habeck, C. (2006). Reciprocal Benefits of Mass-Univariate and Multivariate Modeling in Brain Mapping: Applications to Event-Related Functional MRI, H215O-, and FDG-PET. *International Journal of Biomedical Imaging*, *2006*, 13, Article ID 79862.
26. Moeller, J. R., & Strother, S. C. (1991). A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *Journal of Cerebral Blood Flow and Metabolism, 11*, A121–A135.
27. Moeller, J. R., Strother, S. C., Sidtis, J. J., & Rottenberg, D. A. (1987). Scaled subprofile model: a statistical approach to the analysis of functional patterns in positron emission tomographic data. *Journal of Cerebral Blood Flow and Metabolism, 7*, 649–658.
28. Frutiger, S. A., Strother, S. C., Anderson, J. R., Sidtis, J. J., Arnold, J. B., & Rottenberg, D. A. (2000). Multivariate predictive relationship between kinematic and functional activation patterns in a PET study of visuomotor learning. *Neuroimage, 12*, 515–527.
29. Grady, C. L., Protzner, A. B., Kovacevic, N., Strother, S. C., Afshin-Pour, B., Wojtowicz, M., et al. (2010). A multivariate analysis of age-related differences in default mode and task-positive networks across multiple cognitive domains. *Cerebral Cortex, 20*, 1432–1447.
30. Bergfield, K. L., Hanson, K. D., Chen, K., Teipel, S. J., Hampel, H., Rapoport, S. I., et al. (2009). Age-related networks of regional covariance in MRI gray matter: Reproducible multivariate patterns in healthy aging. *Neuroimage, 49*, 1750–1759.
31. LaConte, S., Strother, S., Cherkassky, V., Anderson, J., & Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage, 26*, 317–329.
32. Martinez-Ramon, M., Koltchinskii, V., Heileman, G. L., & Posse, S. (2006). fMRI pattern classification using neuroanatomically constrained boosting. *Neuroimage, 31*, 1129–1141.
33. Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *Neuroimage, 28*, 980–995.
34. Chen, X., Pereira, F., Lee, W., Strother, S., & Mitchell, T. (2006). Exploring predictive and reproducible modeling with the single-subject FIAC dataset. *Human Brain Mapping, 27*, 452–461.
35. Tripoliti, E. E., Fotiadis, D. I., & Argyropoulou, M. (2008). A supervised method to assist the diagnosis and classification of the status of Alzheimer's disease using data from an fMRI experiment. *Conference Proceedings of the IEEE Engineering in Medicine and Biology Society, 2008*, 4419–4422.
36. De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage, 43*, 44–58.
37. Etzel, J. A., Gazzola, V., & Keysers, C. (2009). An introduction to anatomical ROI-based fMRI classification analysis. *Brain Research, 1282*, 114–125.
38. Markiewicz, P. J., Matthews, J. C., Declerck, J., & Herholz, K. (2009). Robustness of multivariate image analysis assessed by resampling techniques and applied to FDG-PET scans of patients with Alzheimer's disease. *Neuroimage, 46*, 472–485.
39. Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 123–140.
40. Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation, 121*, 256–285.
41. Wood, I. A., Visscher, P. M., & Mengersen, K. L. (2007). Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics, 23*, 1363–1370.