NeuroImage 56 (2011) 1382–1385

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Technical Note

Verification of predicted robustness and accuracy of multivariate analysis ☆

P.J. Markiewicz [a,*], J.C. Matthews [a], J. Declerck [b], K. Herholz [a]
and for the Alzheimer's Disease Neuroimaging Initiative (ADNI)

[a] School of Cancer and Enabling Sciences, University of Manchester/MAHSC, Wolfson Molecular Imaging Centre, Manchester, England, UK
[b] Siemens Molecular Imaging, Oxford, England, UK

ARTICLE INFO

Article history:
Received 15 October 2010
Revised 8 February 2011
Accepted 12 February 2011
Available online 19 February 2011

Keywords:
Resampling
Bootstrap
Robustness
PET
Multivariate image analysis
Alzheimer's disease

ABSTRACT

The assessment of accuracy and robustness of multivariate analysis of FDG-PET brain images as presented in [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.] using a homogeneous sample (from one centre) of small size is here verified using a heterogeneous sample (from multiple centres) of much larger size.

Originally the analysis, which included principal component analysis (PCA) and Fisher discriminant analysis (FDA), was established using a sample of 42 subjects (19 Normal Controls (NCs) and 23 Alzheimer's disease (AD) patients) and here the analysis is verified using an independent sample of 166 subjects (86 NCs and 80 ADs) obtained from the ADNI database.

It is shown that bootstrap resampling combined with the metric of the largest principal angle between PCA subspaces as well as the deliberate clinical misdiagnosis simulation can predict robustness of the multivariate analysis when used with new datasets. Cross-validation (CV) and the .632 bootstrap overestimated the predictive accuracy encouraging less robust solutions.

Also, it is shown that the type of PET scanner and image reconstruction method has an impact on such analysis and affects the accuracy of the verification sample.

Introduction

Small sample sizes in neuroimaging make the extraction of image features meaningful to the population a challenging task when using multivariate analyses. In many cases, due to small sample size and/or other limiting factors (e.g., unavoidable measurement and analysis errors, using different scanners, protocols, methods, recruitment sites, etc.), the samples are unlikely to be fully representative of the populations from which they are taken. Nonetheless, statistical analysis is frequently performed on those samples to extract some limited portion of the robust information representative of the populations.

Estimation of a limited number of robust image features (principal components, PCs) extracted from a sample of 42 subjects (19 NCs and 23 ADs from one European centre) was presented in Markiewicz et al. (2009). Bootstrap resampling (Efron and Tibshirani, 1993) with the metric of the largest principal angle[1] between PCA subspaces was used in the estimation. The angle was measured between a PCA subspace spanned by a given number of PCs obtained from the whole sample (42 subjects with no resampling) and a PCA subspace spanned by the same number of PCs from one of the 1000 bootstrap samples, and thus forming the distribution of the angle. Investigation of the median and dispersion of the distribution (the narrower and closer to zero is the distribution the better) indicated that only the first four PCs can be used in the PCA/FDA discrimination between AD and NC groups. The same procedure has also been used with SPECT data (Merhof et al., 2011).

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://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 writing of this report. A complete listing of ADNI investigators can be found at: http://www.loni.ucla.edu/ADNI/Collaboration/ADNI Authorship list.pdf.

* Corresponding author.
E-mail address: p.markiewicz@manchester.ac.uk (P.J. Markiewicz).

[1] Largest principal angle measures the angle between two multidimensional spaces which is closely related to the distance between the spaces. If the spaces are one-dimensional it is equivalent to the usual angle between two vectors. It is used for measuring the stability/robustness of the resampled PCA subspaces.

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.02.036

However, as shown in Markiewicz et al. (2009), the .632 bootstrap[2] and cross-validation[3] predicted highest accuracy (95% and 97%, respectively) obtained for as many as 10 PCs. However, single deliberate clinical misdiagnosis[4] of each subject of the sample showed decreased robustness (greater sensitivity to a single misdiagnosis) when using more than four PCs, thus confirming that only the first four PCs could be useful. In this brief article, a much larger sample, obtained from many American sites, was used to verify the conclusion that only the first four PCs are reliable.

Methods

An independent verification sample was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). The Principal Investigator is Michael W. Weiner, MD, VA Medical Center and University of California—San Francisco.

All subjects of the sample (166 in total, 86 NCs and 80 ADs) underwent FDG-PET scans in 36 different American centres. Within these centres there were different types of PET scanners (i.e., General Electric Medical Systems: (1) Advance, (2) Discovery HR, (3) Discovery LS, (4) Discovery RX, (5) Discovery ST; Philips Medical Systems: (1) Allegro Body, (2) G-PET Brain, (3) Guardian Body, (4) Gemini TF; Siemens/CTI: (1) ACCEL, (2) ECAT EXACT, (3) HRRT, (4) ECAT EXACT HR+.) with associated different image reconstruction methods (filtered back projection (FBP) or iterative methods). Mean age at the time of PET scan was 75.94 ± 4.61 for NCs and 75.40 ± 6.96 for ADs. Mean MMSE score for AD patients was 23.56 ± 2.36.

The preprocessing of this sample, including spatial and intensity normalisation as well as smoothing, were exactly the same as for the original sample of 42 subjects. Briefly, this consisted of 12-parameter affine normalisation followed by nonlinear iterative spatial transformation in SPM5 (statistical parametric mapping, Ashburner and Friston (1999)) resulting in images with voxel size of 2 mm. The images were smoothed with a Gaussian with a FWHM of 8 mm. All images were normalized to the global mean of brain intensities. (Markiewicz et al., 2009).

The impact of different scanners and reconstruction methods was assessed by limiting the ADNI test sample to (i) only those scanners which were very similar or the same to the scanners used in the original sample (Siemens ECAT EXACT and ECAT EXACT HR) and (ii) the reconstruction method used in the original sample (FBP). The number of cases in ADNI sample were: (i) 65 (36 NCs and 29 ADs) for the matching of scanners and (ii) 50 (22 NCs and 28 ADs) for the matching of reconstruction method, out of a total of 166 subjects from the available ADNI data.

Results and discussion

The PCA/FDA analysis with all its parameters established on the small sample was applied to the larger ADNI data. The number of PCs included in the analysis was varied from 1 to 15 to find those PCs which give best results. As shown in Fig. 1, the maximum obtained accuracy for the ADNI test sample (shown with thick solid black curve in all three plots) is between the first two and four PCs. Note that the maximum obtained accuracy (82%) is achieved for a smaller number of PCs (2 PCs) compared to up to 4 PCs predicted using angles between PCA subspaces[5] (Markiewicz et al., 2009) and more than 10 PCs as predicted with the .632 bootstrap and CV on the original 42-subject sample. The predicted .632 bootstrap accuracy is given in the top and bottom plots and the CV predicted accuracy is shown only in the bottom plot in Fig. 1 (Markiewicz et al., 2009). Also, it can be noticed that the maximum accuracy for the verification sample is significantly lower than the accuracy predicted by CV and the .632 bootstrap. This may, at least partially, be due to the different age distribution between the two samples (the average ADNI age is significantly higher, Haense et al. (2009)).

Further, the greater heterogeneity of the ADNI sample and different methodological factors of the two samples will also have an effect on the accuracy. The middle plot of Fig. 1 shows the impact of scanner type and reconstruction method on the accuracy of the verification sample (curves with the triangular markers). If the scanner types and reconstruction methods are matched in the training and verification samples, the obtained verification accuracy is higher (by 4% for matched scanners and 6% for matched reconstruction), which means that methods/protocols can have an impact on such analysis and should be accounted or corrected for when possible. Notice that the maximum accuracy for the same scanner type is achieved for the first two PCs whereas for the same reconstruction method the maximum accuracy is achieved for the first four PCs. Although, the ADNI sample does not have cases with matching both the scanner type and reconstruction method (for matching scanners all the ADNI data was reconstructed using iterative methods whereas the original sample was reconstructed using FBP only), it is anticipated that the obtained accuracy would be even higher.

The bottom plot of Fig. 1 relates the results of model (PCA subspace) selection with the performance of the model in the ADNI verification sample. Note that both the .632 bootstrap and CV overestimated the predicted accuracy encouraging higher number of PCs to be included in the model. However, the results of the simulation of clinical misdiagnosis shown in the bottom plot (shown are the median of the distribution of the predicted accuracy with its range defined as $1.5 \times$IQR of the lower and upper quartiles, where IQR is the interquartile range) indicate that up to four PCs can be considered robust (the predicted .632 bootstrap accuracy is out of the range of the accuracy distribution found through the misdiagnosis simulation[6]). It is worth noticing that after the first four PCs the median of the distribution levels out as opposed to the .632 bootstrap and CV estimators. Also, the median largest principal angle between PCA subspaces is plotted in gray in the bottom plot with different y-axis on the right. The metric of the largest principal angle shows the rapid loss of robustness of the PCA subspace with more PCs being included in the model (for the first four PCs the angle already exceeds 50°). Comparing the results of the .632 bootstrap and CV accuracy estimators with the ADNI accuracy it appears that in this case the number of PCs could also be chosen based on significant improvement in the predicted accuracy when using the .632 bootstrap or CV.

[2] In bootstrap resampling the original sample is resampled with replacement resulting in training samples of the same size as the original sample. In any bootstrap sample the probability of a given instance being included is 0.632 which means that on average 37% of the original sample is not included in the bootstrap sample. The remaining instances are used for estimating the predictive accuracy over many bootstrap replications.

[3] In CV resampling the original sample is multiply resampled without replacement by dividing the original sample into a training and validation set. The validation set is a combination of instances (here one AD and one NC with all possible combinations being considered) taken out for estimating the average predictive accuracy. Note that the size of training sample is always smaller than the size of the original sample.

[4] In the deliberate clinical misdiagnosis simulation the diagnosis of one subject at a time is deliberately changed to simulate the limited accuracy of clinical diagnosis and its impact on the .632 bootstrap predicted accuracy. Since in the original sample there are 42 subjects the simulation will result in 42 predicted accuracies whose distribution is informative about the impact of such clinical misdiagnosis. Note, that clinical diagnosis has limited accuracy and its possible errors should be accounted for in image analysis.

[5] Although it was found using the metric of angle between PCA subspaces that up to 4 PCs can be regarded as robust, the metric however, does not indicate how useful PCA subspaces are for a given task of classification. For instance, third and fourth PCs in this case may be robust but of no or little use for discrimination between NCs and ADs.

[6] Note that the clinical misdiagnosis simulation provides complimentary information to that of CV and .632 bootstrap estimators. The simulation is not used for accuracy estimation but rather for assessing the robustness of the predicted accuracy.

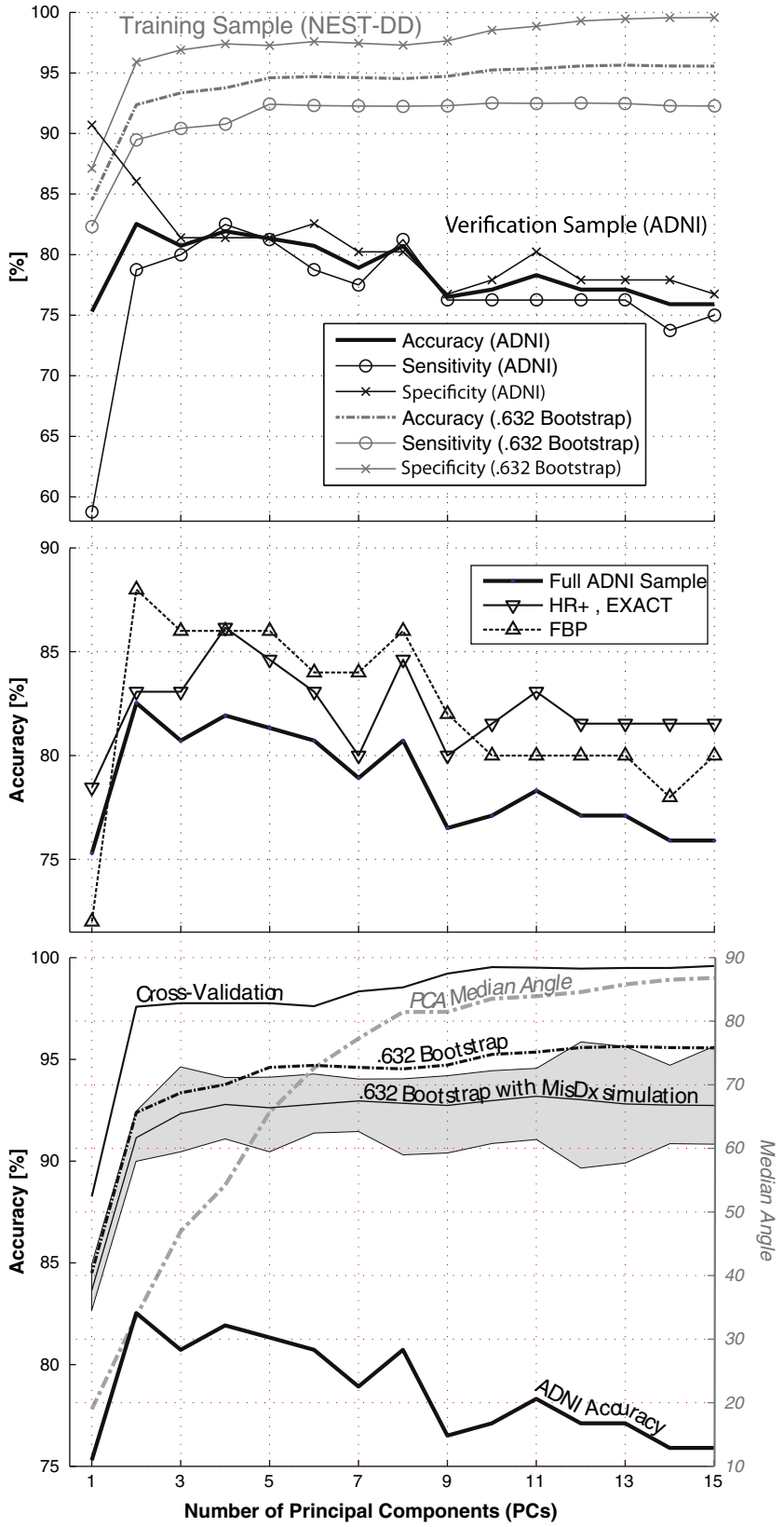

Fig. 1. Top: Accuracy, sensitivity and specificity of the PCA/FDA discrimination analysis trained on the original sample of 42 subjects (grey) and verified on the ADNI sample (black). Middle: Accuracy of the ADNI sample with matched scanner type and reconstruction method to that of the original sample. Bottom: Model selection using CV, the .632 bootstrap, the largest principal angle between PCA subspaces (median angle shown) and deliberate clinical misdiagnosis simulation (shown are the median with the range of the dispersion of the accuracy distributions). The performance of the different metrics derived from the original 42-subject sample are compared with the accuracy of the model in the ADNI sample for each choice of the number of PCs.

Conclusion

The verification with the larger and heterogeneous ADNI dataset supports the findings obtained with bootstrap resampling and the metric of the largest principal angle applied to the small sample of 42 subjects that only the first four PCs can be regarded as robust and useful for future statistical analysis (Markiewicz et al., 2009). Although, the maximum accuracy for the whole verification sample is achieved for the first two PCs the first four PCs may be still considered for image analysis since for the same scanner type the maximum accuracy is achieved for the first four PCs. It has been shown that the scanner type and reconstruction method can affect the analysis resulting in higher accuracies if such methods are matched. Additionally, standard cross-validation or the .632 bootstrap estimators can fail in cases of small sample size as such samples may not well represent the population suggesting that resampling with more refined metrics would have to be used.

Acknowledgments

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., 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

Ashburner, J., Friston, K.J., 1999. Nonlinear spatial normalization using basis functions. Hum. Brain Mapp. 7, 254–266.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall/CRC.

Haense, C., Herholz, K., Jagust, W.J., Heiss, W.D., 2009. Performance of FDG PET for detection of Alzheimer's Disease in Two Independent Multicentre Samples (NEST-DD and ADNI). Dement. Geriatr. Cogn. Disord. 28, 259–266.

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.

Merhof, D., Markiewicz P. J., Platsch, G., Declerck, J., Weih, M., Kornhuber, J., Kuwert, T., Matthews, J. C., Herholz, K., 2011. Optimized data preprocessing for multivariate analysis applied to 99mTc-ECD SPECT datasets of Alzheimer's patients and asymptomatic controls. J. Cerebral Blood Flow Metab 31, 371–383.