Penalized Decomposition Using Residuals (PeDecURe) for Mitigating Nuisance Variables in Multivariate Pattern Analysis

Sarah M. Weinstein [1] Christos Davatzikos [2] Jimit Doshi [2] Kristin A. Linn [1,2,*] Russell T. Shinohara [1,2,*]

[1]Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics [2]Center for Biomedical Image Computing and Analytics, Department of Radiology
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
[*]Contributed equally

Introduction

In neuroimaging studies, multivariate methods provide a framework for studying associations between complex spatial patterns in the brain and neurological, psychiatric, and behavioral phenotypes. However, mitigating the influence of nuisance variables, such as confounders, remains a critical challenge. For example, in studies of Alzheimer's Disease (AD), imbalance in disease rates across age and sex may make it difficult to distinguish between structural patterns in the brain attributable to disease progression and those characteristic of typical human aging or sex differences (Hua et al. 2010).

When not properly accounted for, nuisance variables can obscure interpretations and preclude the generalizability of findings from neuroimaging studies (Linn et al. 2016; Rao et al. 2017). Motivated by this critical issue, in this work we examine the impact of nuisance variables on features extracted from image decomposition methods and propose Penalized Decomposition Using Residuals (PeDecURe), a new method for obtaining nuisance variable-adjusted features in neuroimaging and other complex datasets.

Methods

Let $X_{n\times p}$ be a matrix matrix of p image-derived features, such as volumes of different regions of interest (ROIs) in the brain, for n study subjects. Let $A_{n\times q}$ be a matrix of nuisance variables, which are considered important to account for before we can study the relationship between X and Y (Figure 1(a)), an outcome of interest (say, AD diagnosis). The objective functions of existing decomposition methods (PCA, partial least squares (PLS), and PCA with adjustment for confounders (AC-PCA, (Lin et al. 2016)), are illustrated and provided in Figure 1(b).

Penalized Decomposition Using Residuals (PeDecURe). To implement our proposed method, we first fit the following linear model at each image location ($j = 1, ..., p$):

$$X^{(j)} = \mathbf{1}\beta_0^{(j)} + Y\beta_Y^{(j)} + A\beta_A^{(j)} + \epsilon. \quad (1)$$

Next, we define two sets of residuals:

$$X^{(j)*} = X^{(j)} - A\hat{\beta}_A^{(j)} \quad (2)$$

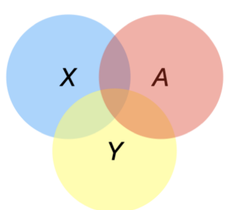
$$\widetilde{X}^{(j)} = X^{(j)} - Y\hat{\beta}_Y^{(j)}. \quad (3)$$

where $\hat{\beta}_A^{(j)}$ and $\hat{\beta}_Y^{(j)}$ are the coefficient estimates for A and Y (conditional on Y and A respectively) in Equation (1) for image locations $v = 1, \ldots, p$. As shown in Figure 1(c), PeDecURe identifies primary components (PCs) which maximize covariance between X^* and Y, while simultaneously penalizing associations between $\widetilde{X}$ and A.

Simulation set-up. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we take 1000 training samples of $n = 200$ using a biased sampling procedure to induce confounding by age (A_1) and sex (A_2) in each training sample. We apply each decomposition method to estimate PCs in each training sample and then compare performance of the methods in both the training samples and in 1000 random testing samples of $n = 80$ (which are balanced with respect to A_1 and A_2) using partial correlation coefficients between PC 1-3 scores and Y, A_1, and A_2.

Figure 1.

(a) Observed data.



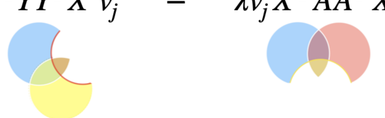
(b) Objective functions for previous decomposition methods.
For j = 1, … , p^* ($p^* \le p$), find v_j that maximizes:

PCA	AC-PCA	PLS
$v_j^T X^T X v_j$	$v_j^T X^T (I - \lambda A A^T) X v_j$	$v_j^T X^T Y Y^T X v_j$

subject to $v_j \perp v_{j-1} \perp \ldots \perp v_1$

(c) Proposed method: Penalized Decomposition Using Residuals (PeDecURe).
For j = 1, … , p^* ($p^* \le p$), find v_j that maximizes:

$$v_j^T X^{*T} Y Y^T X^* v_j \quad - \quad \lambda v_j \widetilde{X}^T A A^T \widetilde{X} v_j$$

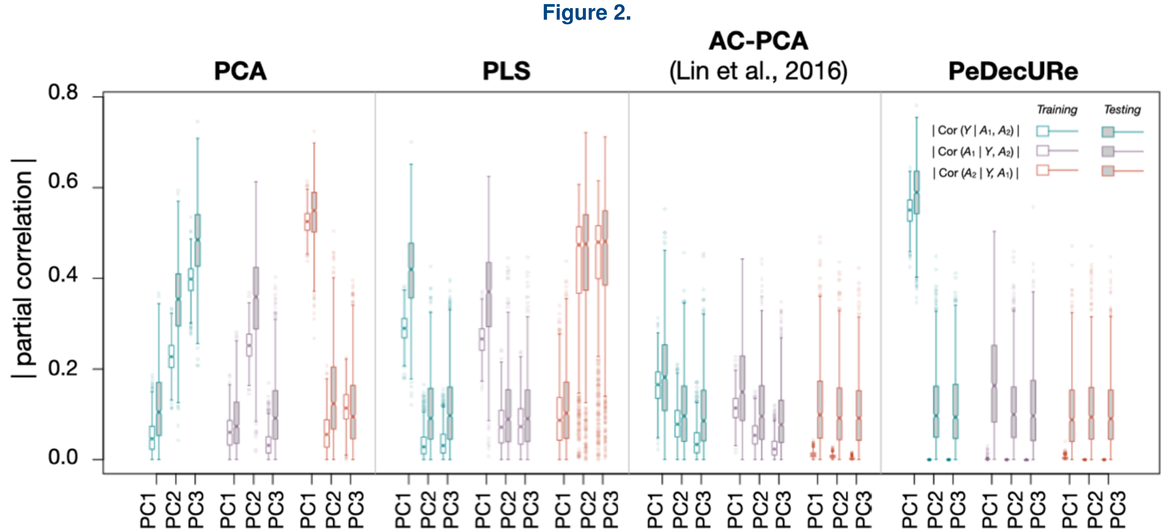


subject to $v_j \perp v_{j-1} \perp \ldots \perp v_1$

Results

Scores on PeDecURe's PC1 had high partial correlation with Y, and scores on the top 3 PCs were not correlated with A_1 or A_2, conditional on Y (Figure 2). PCA, PLS, and AC-PCA all had higher distributions of partial correlations with A_1 or A_2. PeDecURe's performance was similar (although slightly more variable) in held-out testing data that came from a different sampling distribution than the training samples.

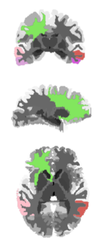
Figure 2.



Conclusions

PeDecURe can be used to reduce the influence of nuisance variables, including confounders, in neuroimaging data, without losing valuable information about an outcome of interest. PeDecURe's PC1 is also highly predictive of the outcome of interest (see additional results in our preprint, which is linked below). In addition, PeDecURe's robustness to the distribution of confounders supports the method's generalizability. While PeDecURe is primarily motivated by MVPA in the context of neuroimaging, it is broadly applicable to datasets where the dimensionality or complexity of the covariance structure calls for novel methods to handle sources of nuisance variation. In future work, we will further investigate the interpretability of features derived using PeDecURe (Figure 3).

Figure 3.



Acknowledgements

This work is supported by the following grants: R01MH112847, R01MH123550, R01NS112274, R01NS060910, U01AG068057, R01MH112070, and RF1AG054409. This work is also supported by the National Science Foundation Graduate Research Fellowship Program.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

References

Hua, Xue et al. (2010). "Sex and age differences in atrophic rates: an ADNI study with n= 1368 MRI scans". In: *Neurobiology of aging* 31.8, pp. 1463–1480.

Lin, Zhixiang et al. (2016). "Simultaneous dimension reduction and adjustment for confounding variation". In: *Proceedings of the National Academy of Sciences* 113.51, pp. 14662–14667.

Linn, Kristin A et al. (2016). "Addressing confounding in predictive models with an application to neuroimaging". In: *The international journal of biostatistics* 12.1, pp. 31–44.

Rao, Anil et al. (2017). "Predictive modelling using neuroimaging data in the presence of confounds". In: *NeuroImage* 150, pp. 23–49.

10.1101/2022.01.27.4
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

CONTACT

Sarah Weinstein, Biostatistics PhD Candidate
sarah.weinstein@pennmedicine.upenn.edu
http://smweinst.github.io