Revisiting PLS Resampling: Comparing Significance vs. Reliability Across Range of Simulations

Natasa Kovacevic, Hervé Abdi, Derek Beaton, for the Alzheimer's Disease Neuroimaging Initiative*, and Anthony R. McIntosh

In
Abdi, H., Chin, W., Esposito Vinzi, V., Russolillo, G., & Trinchera, L. (Eds.), 2013, New Perspectives in Partial Least Squares and Related Methods. New York: Springer Verlag

Abstract PLS as a general multivariate method has been applied to many types of data with various covariance structures, signal strengths, numbers of observations and numbers of variables. We present a simulation framework that can cover a wide spectrum of applications by generating realistic data sets with predetermined effect sizes and distributions. In standard implementations of PLS, permutation tests are used to assess effect significance, with or without procrustes rotation for matching effect subspaces. This approach is dependent on signal amplitude (effect size) and, as such, is vulnerable to the presence of outliers with strong amplitudes. Moreover, our simulations show that in cases when the overall effect size is weak, the rate of false positives—and to a lesser extent—false negatives, is quite high. From the applications point of view, such as linking genotypes and phenotypes, it is often more important to detect reliable effects, even when they are very weak. Reliability in such cases is measured by the ability to observe the same effects supported by the same patterns of variables, no matter which sets of observations (subjects) are used. We implemented split-half reliability testing with thresholds based on null distributions and compared the results to the more familiar significance testing.

Key words: PLS, SVD, significance, simulations, reliability

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

1 Introduction

Partial Least Squares (PLS) is a versatile multivariate method that has been applied to many data types in neuroimaging, Psychology, physiology, Genetics, and Chemometrics to name but a few [1,2,14]. In neuroimaging, the standard application of PLS is PLS-correlation (PLSC, see [1]) whose computational core consists in the singular value decomposition of the correlation matrix of the variables from two matrices. In this context, the singular vectors are called *saliences* and the associations between the two data tables are explored with latent variables (LVs) computed as the projection of each table on its corresponding saliences (see, e.g., [1], for details). More recently, researchers have been interested in investigating large scale data sets with weak signals, as found, for example, when relating genotypes and complex phenotypes (e.g., personality traits, body weight). For these problems, PLS methods offer an excellent framework. However, it is difficult to interpret and validate genotype-phenotype associations obtained by PLS because the ground truth is not known. For this reason it is important to generate simulated data and hone the ability of PLS to correctly detect weak but reliable signals.

More generally, PLS validation requires a simulation framework that can cover a wide range of applications that vary in covariance structure, signal strength, number of observations and variables. In this work, we present a set of simulations where significance testing—within the standard PLS implementation—shows a clear propensity to Type I errors, and much more so for certain data types. PLS methods seem to be more prone to this Type I error inflation when the cross-validation approach—used to derive the sampling distribution under the null hypothesis—involves Procrustes rotations to project the LVs from the permuted data onto the original LVs (as done in the current implementation of PLSC, see [2]).

It is well known that significance testing using permutation tests is essentially amplitude driven and vulnerable to the presence of outliers ([4]). We have found, in practice, another weakness of PLS when it is applied to a weak correlation structure between predictor and response variables, where one of these two data sets has a very weak covariance structure while the other set has a very strong covariance structure between variables (as is often the case in "brute force" approaches to genotype-phenotype associations). In such cases, the strength of correlations between response variables (e.g., highly redundant behavioral measures)—even though not necessarily related to the genes—can overpower the permutation tests and falsely identify genotype-phenotype associations. When this is the case, using completely random genetic data will produce similar results to the analysis performed on real genetic data (because the analysis is driven by the other set). Therefore, a more appropriate question then is: Can we detect associations (i.e., LVs) that reliably represent specific genotype-phenotype links, such that any set of subjects would produce similar LVs with simultaneous, better-than-chance similarity for both associated patterns (e.g., genotype and phenotype)?

Motivated by such examples, we designed a Monte-Carlo simulation framework, flexible enough to mimic many realistic scenarios, with the advantage that we could manipulate the ground truth. We also introduced a new split-half resampling frame-

work for reliability testing, similar to [5], as an alternative to significance (null hypothesis) testing within the PLS approach. We then compared results obtained with classical PLSC analysis with and without Procrustes rotations to those obtained using split-half reliability testing.

Although the work presented here is general in nature and applies to different data types, our starting point was the PLSC methodology as implemented in the PLSC software package ([2]), which focuses on neuroimaging applications. We will therefore use terminology appropriate for neuroimaging applications, where predictor variables are typically some sort of brain imaging data, as in [2]. Typically, subjects are split across several groups and their data are collected under different experimental conditions. In this spirit, observations are condition specific subject data and predictor variables are voxels. Task-PLS refers to the data driven approach where stacked subjects voxel data are tested for group/condition membership patterns, called task effects. Seed-PLS refers to data driven analysis of the correlation matrix between entire brain data and a (typically small) subset of voxels, called seeds, where correlations are calculated across group and condition specific subject data. In this case, PLS also extracts group/condition patterns in correlations (see [1] for details).

2 Simulations

We used real data as a starting point for our simulations in order to create realistic scenarii while manipulating effect sizes and noise sizes and distributions. These data sets were chosen from brain imaging, behavior, and genetics to represent a wide range of data dimensions, specifically number of observations and number of predictor variables. These synthetic data sets were then tested with two main flavors of PLS, data driven task-PLS and seed-PLS.

2.1 Real data sources

We used three different types of real data: electro-encephalogram, behavior, and genetics.

2.1.1 Event related potentials (ERP) data

The first set consists of electroencephalogram (EEG) data from a total of 48 subjects whose data were collected across 2 experimental conditions. In addition, subjects were divided into 3 age groups, with 16 subjects in each group: Young (mean age 22 ± 3 years), Middle (mean age 45 ± 6 years) and Older (mean age 66 ± 6 years). For the purposes of the present work, we used 2 visual perceptual matching tasks

from the larger study that involved 6 conditions. Visual stimuli were presented simultaneously in a triangular array. In the perceptual matching task (PM), subjects indicated which of the three bottom stimuli matched the one on the top by pressing one of three buttons. In the delayed match to sample task (DMS), the instructions were the same as in the PM, except that the three bottom row stimuli were presented after a 2.5s delay following the presentation of the top row stimulus.

EEG recordings from 76 electrodes were collected using BioSemi Active Two system with a bandwidth of 99.84 (0.16 100) Hz and sampling rate of 512 Hz. Data were recorded reference-free, but were converted to an average reference at Cz during the pre-processing. We utilized standard preprocessing steps for ERP data analysis. Continuous EEG recordings were bandpass filtered from 0.5 to 55 Hz. Data from trials with correct responses were "epoched" and base-lined into $[-500\ 2000]$ ms epochs with a $[-500\ 0]$ ms pre-stimulus baseline. Artifact removal was performed using Independent Component Analysis (ICA). The data were averaged across trials for each condition separately. For our simulations we considered only $[0\ 500]$ms time window (257 time points) of the averaged data.

This represents a scenario with a small number of subjects (48), a large number of predictors (EEG channels $\times$ time points = $76 \times 257 = 19,532$), that are somewhat strongly correlated (see Figure 1A) and a small number of group/condition dimensions (3 age groups $\times$ 4 conditions = 12).

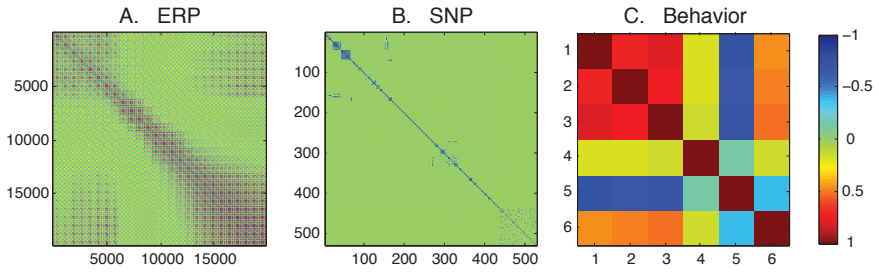

Fig. 1: Correlation matrices for 3 real data sets. Each matrix was derived from all available observations. Notice the wide variety in voxel space dimensionality and correlation strengths.

2.1.2 Genes and behavior: genetic data

Genetic and associated behavioral data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering

(NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California at San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S.A. and Canada. The initial goal of ADNI was to recruit 800 adults, aged from 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see `www.adni-info.org`.

In the ADNI database, the genetic information of each participant is a long list of pairs (one per chromosome) of DNA nucleotides (A, T, C, and G)—which could occur in $2^4 = 16$ different configurations—grouped in 23 chromosomes, amounting to roughly $600,000$ genetic markers. However, after standard preprocessing with PLINK (`pngu.mgh.harvard.edu/purcell/plink/`, e.g., with a call rate of 90% and minor allele frequency of 5%, [6]) we were left with approximately $500,000$ genomic locations that show enough variability in a population. These locations of variability are called single nucleotide polymorphisms (SNPs).

Because our goal was to understand the effects of resampling-based inference tests for PLS, we selected only some of the top reported, clinically relevant genetic markers, consisting of 178 SNP's. Besause the work presented here is not concerned with data inperpretation, we skip the details of clinical relevance. Each SNP has a major allele (e.g., A) which is the most frequent nucleotide (in a population) and a minor allele (e.g., T; rare in the population but required to be found in at least 5% of the population to be considered worth exploring). Thus, in practice only 3 variants for each location are used: the major homozygote (e.g., AA), the minor homozygote (e.g., TT), and the heterozygote (e.g., AT). Multivariate data sets of SNPs are most often re-coded through a process of counting the number of minor alleles. So, in our data: 0 represents the major allele homozygote (e.g., AA), 1 codes for the heterozygote (e.g., AT), and 2 represents the minor allele homozygote (e.g., TT). In most analyses, the SNPs are treated as quantitative data since most statistical methods used rely upon quantitative measures. Because the assumptions of a quantitative coding scheme seem unrealistic, we have decided to use a qualitative coding scheme and to consider that the values 0, 1, and 2 represent three different levels of a nominal variable and to code each possible variants with a 3 by 1 vector of binary variables (i.e., $AA = [100]$, $AT = [010]$, and $TT = [001]$).

The data were extracted from 756 subjects comprising three clinical groups and each clinical group was further split by sex. This produced a total of six, approximately equally populated, groups of subjects. This pattern of data represents a data analytic scenario with comparable numbers of observations (756) and predictors (SNPs $\times$ variants = $178 \times 3 = 528$). Here, the predictor data are binary and weakly correlated (Figure 1B). The number of group/condition dimensions is also small (6 groups based on clinical diagnosis and sex).

2.1.3 Genes and behavior: behavioral data

We extracted 6 behavioral measures from the same subjects as for the genetic data. Once again, as the intperpretation of the behavioral data is not important for our simulations, we skip the details pertaning to the choice of behavioral measures. This represents a scenario with a large number of observations (756), small number of highly correlated predictors (6 behavioral measures, see Figure 1C), and a small number of group/condition dimensions.

2.2 Simulation of group/condition effects

We start with a real data set stacked in a standard manner as a two-dimensional matrix $\mathbf{X}$ whose every row contains data for one subject (observation) in one condition ([2]). The rows are arranged such that observations are nested within condition blocks, which are in turn nested within group membership. From $\mathbf{X}$ we extracted two parameters: the covariance matrix $\mathbf{C}$ of the voxel space (covariance calculated across observations) and the group/condition specific mean signal $\mathbf{m}$ of the predictor variables across the real data observations. The mean signal $\mathbf{m}$ is further centered by subtracting the grand mean of all groups and conditions. To generate comparable simulated data with a controlled number of group/condition effects, we first decomposed $\mathbf{m}$ using a principal component analysis and then rebuilt the modified signal (denoted $\mathbf{m}_1$) using only the first K principal components. This allowed us to control the number of expected group/condition effects. In the simulations presented here, we chose $K = 3$ as the reasonable number of effects that can be expected with this type of data. To create a simulated voxel data set similar to the real data, we drew observations from a multivariate normal distribution with covariance $\mathbf{C}$ and mean $\mathbf{m}_1$. However, we wanted to test how well we can detect reliable task effects depending on the signal strength ($\mathbf{m}_1$ amplitude across voxels) and the noise distribution. For this reason, we used the signal amplitude as a scalar parameter denoted γ that was manipulated in order to vary the intensity of the signal as $\gamma\,\mathbf{m}_1$. In order to explore the effects of noise we removed the signal from a proportion (denoted *np*) of randomly selected voxels. To summarize, we designed a simulations scheme where we controlled:

1. the number of expected group/condition effects (set to 3 for all simulations)

2. the signal strength measured by γ, where $\gamma \in \{0, 0.5, 1, 3\}$
3. percentage of noise-voxels (i.e., voxels for which $\mathbf{m}_1 = 0$) measured by *np*, where $np \in \{80\%, 40\%, 10\%, 0\%\}$.

2.3 Simulation of correlation effects

Once again we start with a real voxel data set $\mathbf{X}$, with voxel covariance matrix $\mathbf{C}$, as above. We create simulated versions of the data by drawing observations from a multivariate normal distribution with covariance $\mathbf{C}$ and zero mean vector. This produces a matrix $\mathbf{Y}$ with same dimensions as the real data $\mathbf{X}$. We then selected a small set of voxels as seeds (i.e., we extracted columns of $\mathbf{Y}$ corresponding to the selected voxels). Seed-PLS analyzes the correlation between $\mathbf{Y}$ and the seeds and searches for the group/condition effects within the correlation structure which is stacked by group and condition specificity in the same way as for task-PLS. In this case, the strength of the signal reflects the strength of the correlations between the columns of $\mathbf{Y}$ and the seeds. Note that the correlations are exactly 1 for the columns corresponding to the seeds across all groups and conditions. In this scenario, we manipulated the strength of the signal by permuting a random subset of rows of the seed matrix, while keeping the voxel data matrix $\mathbf{Y}$ unperturbed. The percentage of rows that were permuted, denoted *pp*, is inversely related to the strength of the correlations: if only few rows are permuted (e.g., $pp < 5\%$), the correlations change only slightly; if all rows are randomly permuted ($pp = 100\%$), all the correlations are destroyed. In the results presented here, we tested a range of *pp* values with $pp \in \{0\%, 30\%, 60\%, 100\%\}$.

3 Split-Half Reliability

The reliability of the latent variables is implemented in a split-half resampling framework similar to [5]. Here we give a brief description for the data driven PLS methods. The overview of the algorithm is shown in Figure 2. We start by first decomposing the signal $\mathbf{D}$ (whether mean-centered group/condition mean signal in task-PLS or correlation signal between predictors and responses in seed-PLS) using the singular value decomposition (SVD). Specifically, assuming that $\mathbf{D}$ is in a group/condition by voxel format, then the SVD of $\mathbf{D}^T$ is obtained as:

$$\mathbf{D}^T = \mathbf{USV}^T.$$

The columns of $\mathbf{U}$ store the left singular vectors (voxel patterns), the columns of $\mathbf{V}$ store the right singular vectors (group/condition effects) and $\mathbf{S}$ is the diagonal matrix of the singular values. In our framework, we will consider that a latent variable $\boldsymbol{\ell}_i$ comprises a matching set of right and left singular vectors (*i*-th column of $\mathbf{U}$, and $\mathbf{V}$

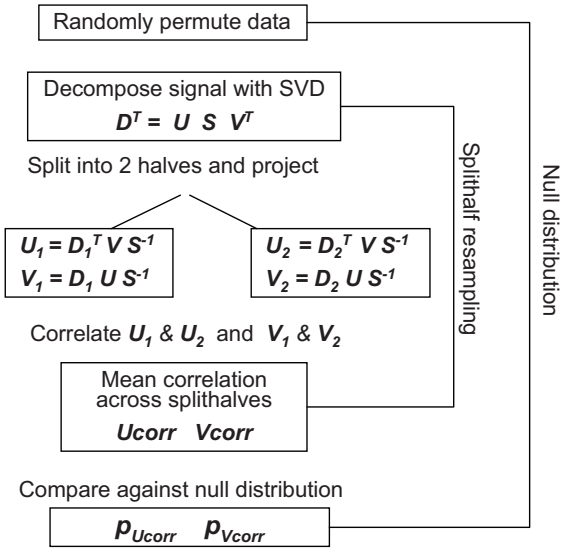

Fig. 2: Diagram of the algorithm for split-half reliability testing.

respectively) and singular value (i-th diagonal element of $\mathbf{S}$). In standard permutation tests, the significance of a given LV is focused on the amplitude of the singular value. However, in split-half reliability testing we are interested in the stability of the pairings between left and right singular vectors. Therefore, we randomly split every group of subjects and calculated the signals $\mathbf{D}_1$ and $\mathbf{D}_2$ by applying the same group and condition specific averaging/correlation procedure as originally performed on $\mathbf{D}$, working with only half of the subjects. We projected the original matrices $\mathbf{U}$ and $\mathbf{V}$ onto each half of $\mathbf{D}$ to obtain the corresponding half-sample matching pairings. Specifically, we computed:

$$\mathbf{U}_1 = \mathbf{D}_1^T\mathbf{V}\mathbf{S}^{-1} \quad \text{and} \quad \mathbf{U}_2 = \mathbf{D}_2^T\mathbf{V}\mathbf{S}^{-1}$$
$$\mathbf{V}_1 = \mathbf{D}_1\mathbf{U}\mathbf{S}^{-1} \quad \text{and} \quad \mathbf{V}_2 = \mathbf{D}_2\mathbf{U}\mathbf{S}^{-1}$$

The correlations between projected left and right split-half patterns (i.e., correlation between the matrices $\mathbf{U}_1$ and $\mathbf{U}_2$, and the matrices $\mathbf{V}_1$ and $\mathbf{V}_2$) are taken as measures of the correspondence between the voxel space and the $\mathbf{V}$ patterns, on one hand, and group/condition membership and the $\mathbf{U}$ patterns, on the other hand. By repeating this procedure many times, we obtain a robust estimate of split-half correlations for both left and right singular vectors. Note that this procedure uses the full sample to decompose the data structure into latent variables. This is particularly important for weak signals, where a half-sample may not reveal the signal. The purpose of the procedure is different from a standard split-half cross-validation, where each half-sample is independently analyzed. Instead, our focus is to evalu-

ate the reliability of the associations—captured by the LV's—between voxel patterns and group/condition effects. In other words, our main question is: Given a group/condition effect, how reliable is the corresponding voxel pattern? Would the same group/condition effect links with a similar voxel pattern if we were to chose a different set of subjects? In an analogous way, given a voxel pattern (left singular vector), we want to estimate the reliability of the associated condition/group effect. For example, in the analysis of genotype/phenotype associations, the SVD decomposes the correlation matrix into latent variables, where each latent variable links a particular weight from the SNPs with a particular weight of from the phenotype measures. In this case, our split-half procedure tests the reliability of this link.

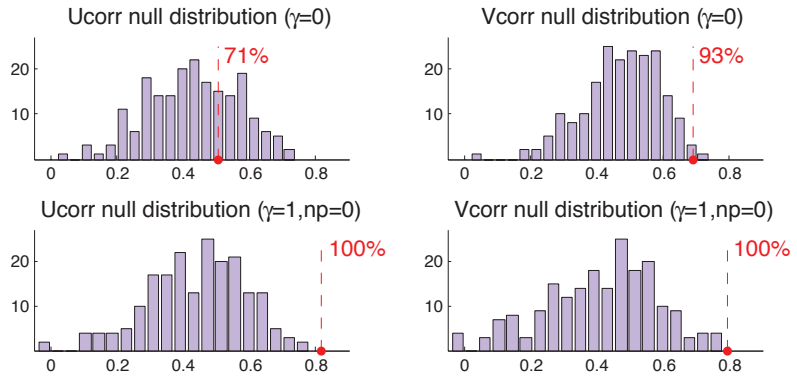

Fig. 3: Null-distribution for p_{Ucorr} and p_{Vcorr}. For illustration we chose the first LV from two ERP-type simulations of task effects. The top row corresponds to the "no signal" scenario with $\gamma = 0$. The bottom row corresponds to the simulation with the most realistic signal strength and distribution, with $\gamma = 1$ and $np = 0$. The red dot marks the split-half correlation of the original un-permuted data. The red dotted line and red percent value indicate the corresponding percentile of the null distribution. In both scenarios, the distributions are strongly skewed towards positive values, however the p_{Ucorr} and p_{Vcorr} percentile values suggest rejection of the null hypothesis for the realistic signal only.

It is important to notice that the distribution of the correlations between projected split-half patterns will be skewed even in a completely random data set. After all, the original SVD decomposition reflects the full sample, so it is not surprising that, on the average, the distribution of the values of the correlation between split halves is biased towards positive values (see Figure 3). To deal with this systematic bias, we create a null distribution for the split-half correlations. This is done by randomly permuting observations (i.e., the rows of $\mathbf{X}$) and repeating the split-half correlation estimation for each permuted data set. This allows us to estimate the probability of surpassing the correlations from the original un-permuted data set. We denote these

probabilities by p_{Ucorr} and p_{Vcorr} and treat them as p-values that describe the stability of voxel patterns associated with **U** and group/condition patterns associated with **V**, respectively. In the present simulations, we performed 200 half-splits and 200 permutations to create the null distributions, and considered that a latent variable was reliable when both probabilities were smaller than .05 (i.e., $p_{\text{Ucorr}} < .05$ and $p_{\text{Vcorr}} < .05$).

4 Results and Discussion

Each of the three real data sets were used to generate simulations for the two flavors of PLS. In the case of task-PLS, simulations were designed to have exactly 3 significant LVs, however the strength of the signal captured by these LVs was varied from no signal ($\gamma = 0$) to weak signal (e.g, $\gamma = 0.5, np = 40\%$) and strong signal ($\gamma = 3, np = 0\%$). Simulations for seedPLS were simpler, where partial permutations of the seed data resulted in a reduction of the initial correlations, going from no reduction ($pp = 0\%$) to more reduction ($pp = 30\%, 60\%$) and full reduction ($pp = 100\%$). For each simulation, we calculated two standard p-value estimates of the LV significance, p_{rot} and p_{nonrot} depending on weather Procrustes rotation was used or not. In addition, we calculated p-values of LV reliability estimates based on split-half resampling, p_{Ucorr} and p_{Vcorr}. The results are presented in Tables 1 and 2.

Acknowledgements 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; Alzheimers Association; Alzheimers Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (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 and K01 AG030514.

Table 1: Results for task-PLS simulations. For each of the 3 data types, the signal was constructed to have exactly 3 LVs, and its strength was manipulated with the values of the parameters γ and np. For each simulated data set, we computed two standard p-value estimates of LV significance, p_{rot} and p_{nonrot} depending on weather Procrustes rotation was used or not. In addition, we calculated p-values of LV reliability estimates based on split-half resampling, p_{Ucorr} and p_{Vcorr}.

				$\gamma = 0$				$\gamma = 0.5$				$\gamma = 1$				$\gamma = 3$			
		***np*(%)**		**80**	**40**	**10**	**0**	**80**	**40**	**10**	**0**	**80**	**40**	**10**	**0**	**80**	**40**	**10**	**0**
ERP	LV1	p_{rot}	.01	.03	.01	.01	.01	.01	.00	.01	.00	.00	.01	.01	.01	.00	.00	.00	.00
		p_{nonrot}	.46	.35	.27	.20	.17	.20	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.34	.35	.23	.17	.12	.20	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.10	.04	.04	.01	.04	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	LV2	p_{rot}	.41	.17	.03	.01	.01	.04	.00	.01	.00	.01	.00	.00	.01	.00	.00	.00	.00
		p_{nonrot}	.78	.29	.03	.01	.01	.03	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.78	.41	.08	.02	.01	.05	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.18	.04	.02	.03	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	LV3	p_{rot}	.72	.62	.41	.34	.31	.35	.11	.06	.04	.01	.01	.00	.01	.00	.00	.00	.00
		p_{nonrot}	.50	.34	.10	.02	.04	.03	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.58	.29	.07	.01	.04	.04	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.97	.90	.78	.10	.07	.15	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
SNP	LV1	p_{rot}	.21	.14	.00	.01	.00	.02	.01	.01	.01	.01	.00	.01	.00	.00	.00	.00	.00
		p_{nonrot}	.48	.40	.13	.04	.02	.14	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.65	.56	.12	.04	.02	.24	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.97	.86	.97	.93	.95	.83	.11	.04	.04	.00	.00	.00	.00	.00	.00	.00	.00
	LV2	p_{rot}	.24	.29	.10	.07	.04	.14	.01	.01	.01	.01	.00	.01	.00	.00	.00	.00	.00
		p_{nonrot}	.63	.53	.16	.04	.04	.17	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.34	.33	.23	.15	.09	.16	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.82	.83	.26	.32	.12	.12	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	LV3	p_{rot}	.55	.48	.54	.47	.50	.14	.07	.01	.01	.00	.01	.01	.01	.00	.00	.00	.00
		p_{nonrot}	.15	.20	.22	.21	.23	.07	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.47	.40	.54	.41	.51	.08	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.63	.53	.46	.42	.42	.25	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
Behavior	LV1	p_{rot}	.01	.00	.01	.01	.01	.01	.00	.00	.01	.02	.00	.01	.00	.00	.00	.00	.00
		p_{nonrot}	.18	.17	.00	.00	.00	.20	.00	.00	.00	.18	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.28	.28	.00	.00	.00	.26	.00	.00	.00	.19	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.17	.13	.00	.00	.00	.20	.00	.00	.00	.17	.00	.00	.00	.00	.00	.00	.00
	LV2	p_{rot}	.56	.51	.18	.04	.26	.43	.12	.05	.17	.07	.09	.21	.10	.00	.00	.00	.00
		p_{nonrot}	.26	.21	.00	.00	.00	.15	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00
		p_{Ucorr}	.23	.17	.00	.00	.00	.12	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.20	.16	.00	.14	.01	.07	.00	.07	.00	.00	.00	.00	.00	.00	.00	.00	.00
	LV3	p_{rot}	.91	.92	.71	.87	.74	.74	.78	.87	.72	.63	.72	.42	.51	.00	.00	.00	.00
		p_{nonrot}	.51	.51	.07	.34	.10	.27	.17	.28	.07	.01	.00	.00	.00	.00	.00	.00	.00
		p_{Ucorr}	.31	.21	.02	.21	.08	.12	.02	.04	.23	.01	.00	.00	.00	.00	.00	.00	.00
		p_{Vcorr}	.47	.42	.07	.17	.01	.27	.04	.08	.01	.13	.00	.00	.00	.00	.00	.00	.00

Table 2: Results for seedPLS simulations. For each of the 3 datatypes correlation strengths were manipulated with parameter *pp*. For each simulated data set, we computed two standard p-value estimates of the LV significance, p_{rot} and p_{nonrot} depending on wether Procrustes rotation was used or not. In addition we calculated p-values of LV reliability estimates based on split-half resampling, p_{Ucorr} and p_{Vcorr}.

		***pp*(%)**	**100**	**60**	**30**	**0**
ERP	LV1	p_{rot}	.00	.01	.01	.00
		p_{nonrot}	.15	.00	.00	.00
		p_{Ucorr}	.27	.00	.00	.00
		p_{Vcorr}	.34	.02	.01	.00
	LV2	p_{rot}	.03	.04	.12	.04
		p_{nonrot}	.17	.24	.86	.14
		p_{Ucorr}	.32	.06	.94	.04
		p_{Vcorr}	.52	.30	.55	.06
	LV3	p_{rot}	.15	.23	.45	.14
		p_{nonrot}	.58	.60	.88	.14
		p_{Ucorr}	.12	.20	.81	.01
		p_{Vcorr}	.85	.16	.20	.21
SNP	LV1	p_{rot}	.01	.01	.00	.01
		p_{nonrot}	.32	.01	.00	.00
		p_{Ucorr}	.07	.00	.00	.00
		p_{Vcorr}	.96	.79	.00	.00
	LV2	p_{rot}	.00	.01	.00	.01
		p_{nonrot}	.81	.01	.00	.00
		p_{Ucorr}	.12	.01	.00	.00
		p_{Vcorr}	.82	.07	.00	.00
	LV3	p_{rot}	.12	.01	.00	.00
		p_{nonrot}	.84	.05	.00	.00
		p_{Ucorr}	.99	.32	.00	.00
		p_{Vcorr}	.01	.70	.00	.00
Behavior	LV1	p_{rot}	.07	.00	.01	.01
		p_{nonrot}	.85	.00	.00	.00
		p_{Ucorr}	.91	.00	.00	.00
		p_{Vcorr}	.62	.22	.00	.00
	LV2	p_{rot}	.28	.00	.00	.01
		p_{nonrot}	.20	.00	.00	.00
		p_{Ucorr}	.32	.00	.00	.00
		p_{Vcorr}	.10	.00	.00	.00
	LV3	p_{rot}	.74	.36	.78	.90
		p_{nonrot}	.69	.02	.78	.86
		p_{Ucorr}	.41	.05	.66	.41
		p_{Vcorr}	.82	.02	.46	.26

References

1. A. Krishnan, L. J. Williams, A. R. McIntosh, and H. Abdi, "Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review,"*NeuroImage* **56**, pp. 455–475, 2011.
2. A. R. McIntosh and N. Lobaugh, "Partial least squares analysis of neuroimaging data: Applications and advances,"*NeuroImage* **23**, pp. 250–263, 2004.
3. H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS Regression)," *WIREs Computational Statistics* **2**, pp. 97–106, 2010.
4. R. Wilcox, "Introduction to Robust Estimation and Hypothesis Testing," *Academic Press*, New York, 2012.
5. S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg, "The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework,"*NeuroImage* **15**, pp. 747–771, 2002.
6. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas,MAR Ferreira, D. Bender, J. Maller, P. Sklar, PIW de Bakker, M. J. Daly, and P. C. Sham, ' 'PLINK: a toolset for whole-genome association and population-based linkage analysis," *American Journal of Human Genetics* **81**, 559–575.