The Stability of Behavioral PLS Results in Ill-Posed Neuroimaging Problems

Nathan Churchill, Robyn Spring, Hervé Abdi, Natasa Kovacevic, Anthony R. McIntosh, and Stephen Strother

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 Behavioral Partial-Least Squares (PLS) is often used to analyze ill-posed functional Magnetic Resonance Imaging (*f*MRI) datasets, for which the number of variables are far larger than the number of observations. This procedure generates a latent variable (LV) brain map, showing brain regions that are most correlated with behavioral measures. The strength of the behavioral relationship is measured by the correlation between behavior and LV scores in the data. For standard behavioral PLS, bootstrap resampling is used to evaluate the reliability of the the brain LV and its behavioral correlations. However, the bootstrap may provide biased measures of the generalizability of results across independent datasets. We used split-half resampling to obtain unbiased measures of brain-LV reproducibility and behavioral prediction of the PLS model, for independent data. We show that bootstrapped PLS gives biased measures of behavioral correlations, whereas split-half resampling identifies highly stable activation peaks across single resampling splits. The ill-posed PLS solution can also be improved by regularization; we consistently improve the prediction accuracy and spatial reproducibility of behavioral estimates by (1) projecting *f*MRI data onto an optimized PCA basis, and (2) optimizing data preprocessing on an individual subject basis. These results show that significant improvements in generalizability and brain pattern stability are obtained with split-half versus bootstrapped resampling of PLS results, and that model performance can be further improved by regularizing the input data.

Key words: *f*MRI, behavioral PLS, bootstrap, split-half resampling, prediction, reproducibility, PCA

1 Introduction

A central goal of functional magnetic resonance imaging (*f*MRI) studies of the human brain is to identify networks of brain areas that are tightly linked to measures of

behavior [1, 2]. This problem is typically highly ill-posed and ill-conditioned, with the number of variables P being very large (*i.e.,* more than $20,000$ voxels in brain images), and the number of data samples N being quite small (*i.e.,* typically less than one hundred subjects, with behavioral measures and brain images), a configuration of data known as the "$P \gg N$" problem. To address this issue, two general approaches have emerged in the neuroimaging literature to measure behavioral relations with *f*MRI. The first approach defines *a priori* a small number of brain regions expected to relate to the behavior of interest. This provides a much better conditioned problem, because the number of brain regions is now roughly of the same order as the number of observations ($P \approx N$). The second approach uses most of the available voxels, and attempts to find the brain locations that best reflect the behavioral distribution in a data-driven multivariate analysis. This method attempts to control the ill-conditioned nature of the problem, by using resampling and regularization with dimensionality reduction techniques. A leading approach of this second type is behavioral PLS, as provided in the open-source MATLAB™ "PLS package" developed by McIntosh and *et al.* [3].

The closely related problem of building discriminant or so called "mind reading" approaches has also been developed and explored in the neuroimaging community [4–7]. When defined as a data-driven multivariate problem with large P, mind reading is also ill-conditioned. Resampling techniques have been developed to control for instability and optimize the reliability of the voxels most closely associated with the discriminant function [6, 9, 10]. These approaches use cross-validation forms of bootstrap resampling [11] or split-half resampling [6]. Split-half resampling is particularly interesting, because it has been shown theoretically to provide finite sample control of the error rate of false discoveries in general linear regression methods when applied to ill-posed problems, provided certain exchangeability conditions are met [12].

Behavioral PLS and linear discriminant analysis belong to the same linear multivariate class of techniques, as both are special cases of the generalized singular value decomposition or generalized eigen-decomposition problem [20]. Specifically, let $\mathbf{Y}$ be a $N \times K$ matrix of K behavioral measures or categorical class labels for N subjects, and $\mathbf{X}$ be a $N \times P$ matrix of brain images, where $P \gg N$. The eigen-solution of expression:

$$(\mathbf{Y}^{\mathsf{T}}\mathbf{Y})^{-1/2}\mathbf{Y}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1/2} \tag{1}$$

reflects the linear discriminant solution for categorical class labels in $\mathbf{Y}$ [13]. When $P > N$, $(\mathbf{X}^{\mathsf{T}}\mathbf{X})$ will be singular and therefore $(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1/2}$ cannot be computed without some form of regularization. When $\mathbf{X}$ and $\mathbf{Y}$ are centered and normalized (*i.e.,* each column of these matrices has a mean of zero and a norm of 1), and $(\mathbf{X}^{\mathsf{T}}\mathbf{X}) = (\mathbf{Y}^{\mathsf{T}}\mathbf{Y}) = \mathbf{I}$ (*i.e.,* $\mathbf{X}$ and $\mathbf{Y}$ are orthogonal matrices), then Equation 1 corresponds to the general partial least squares correlation approach defined in Krishnan *et al.* [3,14], for which behavioral PLS with $\mathbf{Y}$ containing subject behavioral scores is a special case. Given the similar bivariate form of PLS and linear discriminants, the goal of this study was to use the split-half techniques developed in the discriminant neuroimaging literature to test the stability of solutions from behavioral PLS, which

uses standard bootstrap resampling methods as implemented in the neuroimaging PLS package [3] (code located at: www.rotman-baycrest.on.ca/pls/source/).

2 Methods and Results

2.1 Functional magnetic resonance imaging (fMRI) data set

Twenty young normal subjects (20–33 years, 9 male) were scanned with *f*MRI while performing a forced-choice, memory recognition task of previously encoded line drawings [15], in an experiment similar to that of Grady *et al.* [16]. We used a 3 Tesla *f*MRI scanner to acquire axial, interleaved, multi-slice echo planar images of the whole brain ($3.1 \times 3.1 \times 5$ mm voxels, TE/TR = 30/2000 ms). Alternating scanning task and control blocks of 24 s were presented 4 times, for a total task scanning time per subject of 192 s. During the 24 s task blocks, every 3 s subjects saw a previously encoded figure side-by-side with two other figures (semantic and perceptual foils) on a projection screen, and were asked to touch the location of the original figure on an *f*MRI-compatible response tablet [17]. Control blocks involved touching a fixation cross presented at random intervals of 1–3 s.

The resulting 4D *f*MRI time series were preprocessed using standard tools from the AFNI package, including rigid-body correction of head motion (`3dvolreg`), physiological noise correction with RETROICOR (`3dretroicor`), temporal detrending using Legendre polynomials and regressing out estimated rigid-body motion parameters (`3dDetrend`, see [8] for an overview of preprocessing choices in *f*MRI). For the majority of results (see Sections 2.2 and 2.3), we preprocessed the data using a framework that optimizes the specific processing steps independently for each subject, as described in [18, 19], within the split-half NPAIRS resampling framework [6]. In Section 2.4, we provide more details of pipeline optimization, and demonstrate the importance of optimizing preprocessing steps on an individual subject basis in the PLS framework.

We performed a two-class linear discriminant analysis separately for each dataset (Class 1: Recognition scans; Class 2: Control scans), which produced an optimal Z-scored statistical parametric map [$SPM(Z)$] per subject. For each subject, the Z-score value of each voxel reflects the extent to which this voxel's brain location contributes to the discrimination of recognition vs. control scans, for that subject.

2.2 Split-half behavioral PLS

The 20 subjects' $SPM(Z)$s were stacked to form a $20 \times 37,284$ matrix $\mathbf{X}$ as described in Equation 1, and a 20×1 $\mathbf{y}$ vector was formed from the differences of the mean (Recognition – Control) block reaction times per subject (in milli-seconds).

After centering and normalizing $\mathbf{X}$ and $\mathbf{y}$, a standard behavioral PLS was run, as outlined in [3], with $1,000$ bootstrap replications. The resulting distribution is reported in Figure 1 (left) under "Bootstrapped PLS." For each bootstrap sample, a latent variable (LV) brain map was also calculated. At each voxel, the mean was divided by the standard error on the mean (SE), computed over all bootstrap measures; this is reported as a bootstrap ratio brain map SPM_{boot} (horizontal axes of Figure 3).

The behavioral PLS procedure was modified to include split-half resampling as follows. After centering and normalizing $\mathbf{X}$ and $\mathbf{y}$, subjects were randomly assigned $1,000$ times to split-half matrices $\mathbf{X}_1$ and $\mathbf{X}_2$, and behavioral vectors $\mathbf{y}_1$ and $\mathbf{y}_2$. For each split-half matrix/vector pair, we obtained the projected brain pattern LV defined by $\mathbf{e}_i = \mathbf{y}_i^{\mathsf{T}}\mathbf{X}_i$ that explained the most behavioral image variance for $i = 1,2$. The correlation $r_{(i,\text{train})} = \rho(\mathbf{y}_i, \mathbf{X}_i\mathbf{e}_i^{\mathsf{T}})$ reflects the correlation between behavior and expression of the latent brain pattern $\mathbf{e}_i$, for each split-half training set. The distribution of the $2,000$ split-half $r_{(i,\text{train})}$ values is plotted in Figure 1 (middle). We also obtained an independent test measure of the behavioral prediction power of each $\mathbf{e}_i$ by calculating $r_{(i,\text{test})} = \rho(\mathbf{y}_{j\neq i}, \mathbf{X}_{j\neq i}\mathbf{e}_i^{\mathsf{T}})$ for i and $j = 1,2$. The distribution of these $2,000$ $r_{(i,\text{test})}$ values is plotted in Figure 1(right). The test $r_{(i,\text{test})}$ behavioral correlations are consistently lower than both training and bootstrap estimates. The reproducibility of the two split-half brain patterns may also be measured as the correlation of all paired voxel values $r_{\text{spatial}} = \rho(\mathbf{e}_1, \mathbf{e}_2)$; this measures the stability of the latent brain pattern across independent datasets. The overall reproducibility of this pattern is also relatively low but consistently greater than zero, with median r_{spatial} of .025 (ranging from .014 to .043; plotted in Figure 4).

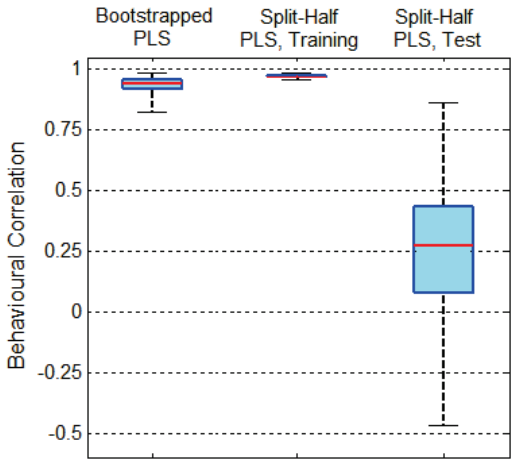

Fig. 1 Behavioral correlation distributions for standard bootstrapped behavioral PLS **(left)**, and split-half training **(middle)** and test **(right)** distributions. Distributions are plotted as box-whisker plots with min.– max. whisker values, a 25th-75th percentile box and the median (red bar); results shown for $1,000$ bootstrap or split-half resampling iterations.

Figure 2 plots the median latent variable (LV) score of each subject, as training-data ($\mathbf{X}_i\mathbf{e}_i^{\mathsf{T}}$ scores, Figure 2a) and as test-data ($\mathbf{X}_{j\neq i}\mathbf{e}_i^{\mathsf{T}}$ scores, Figure 2b); we plotted the median LV scores vs. behavior over the 1000 resamples. The median training

scores show a consistently stronger linear trend than for test. In addition, there is a subject (red circle) whose brain-behavior relation cannot be predicted by the other subjects' data in the test space (it is a significant outlier by Cooks D test, with statistic $d = .90$ exceeding the outlier threshold $4/N$ [21]). By comparison, in the training space, this subject is not a significant outlier.

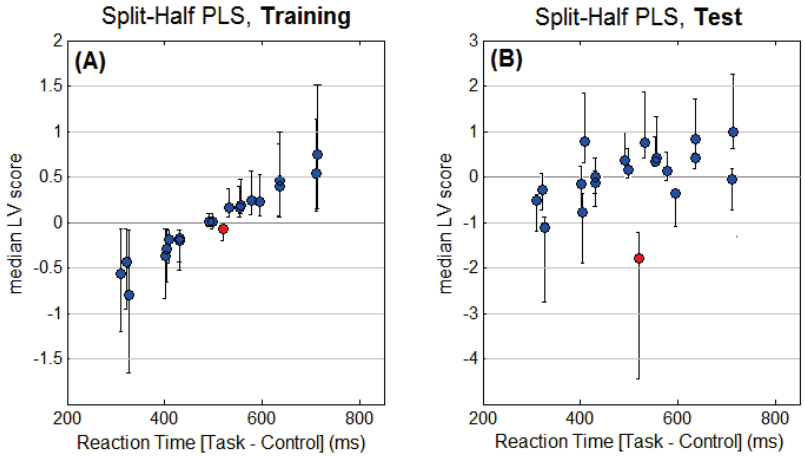

Fig. 2: Median subject behavioral LV scores are plotted against difference in reaction time between (Task-Control) experimental conditions. The error bars give upper and lower quartile ranges on LV scores, for each subject (for $1,000$ split-half resamples). Results are shown for **a.** the training-space LV scores, and **b.** the test-space LV scores. The subject represented by a *red* dot is a significant outlier in test-space, based on Cook's D statistic (see text for details).

The split-half brain patterns $\mathbf{e}_1$ and $\mathbf{e}_2$ can also be used to estimate a behavioral *SPM* that is robust to subject heterogeneity. As described in [6], this is done by normalizing each $\mathbf{e}_i$ to mean zero and variance one, and then projecting the pairwise voxel values onto the line of identity (the first component of a principal component analysis (PCA) on the scatterplot of $\mathbf{e}_1$ vs. $\mathbf{e}_2$ voxel values), which gives a signal-axis estimate: $\mathbf{e}_{\text{signal}} = (\mathbf{e}_1 + \mathbf{e}_2)/\sqrt{2}$. The orthogonal, minor-axis projection (second component of a PCA on the scatter-plot), forms the noise axis. This measures uncorrelated, non-reproducible signal at each voxel, giving noise vector: $\mathbf{e}_{\text{noise}} = (\mathbf{e}_1 - \mathbf{e}_2)/\sqrt{2}$. This is used to estimate a reproducible Z-scored map $rSPM(Z)_{\text{split}} = \mathbf{e}_{\text{signal}}/SD(\mathbf{e}_{\text{noise}})$, where $SD(\mathbf{e}_{\text{noise}})$ provides a single spatially global noise estimator. The average of the $1,000$ $rSPM(Z)_{\text{split}}$ voxel values are plotted on the vertical axis in Figure 3a, against SPM_{boot} values. The $rSPM(Z)_{\text{split}}$ shows generally lower signal values than SPM_{boot}, with a nonlinear relationship. However,

this difference is partly a function of the global versus local noise estimators. We can instead estimate the mean $\mathbf{e}_{\text{signal}}$ value at each voxel, and normalize by the SD on $\mathbf{e}_{\text{noise}}$ for each voxel (each computed across $1,000$ resamples), generating voxel-wise estimates of noise in the same manner as SPM_{boot}. This $rSPM(Z)$ is plotted against SPM_{boot} in Figure 3b, demonstrating a strong linear trend, albeit with increased scatter for high-signal voxels. This scatter is primarily due to differences in the local noise estimates: the mean bootstrap LV and $\mathbf{e}_{signal}$ patterns are highly consistent (correlation equal to .99), whereas the local noise estimates are more variable between the two methods (plotted in Figure 3c; correlation equal to .86).

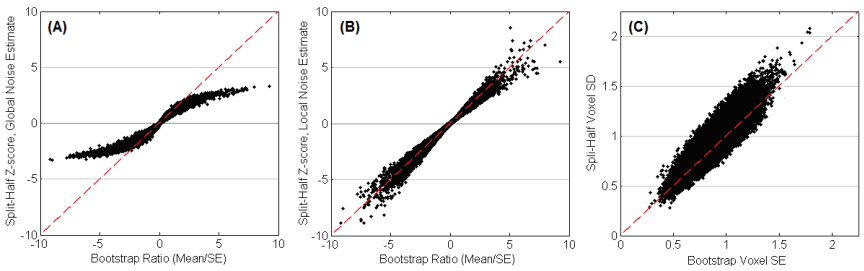

Fig. 3: Scatter plot of pairs of voxel *SPM* values: we compare standard bootstrapped behavioral PLS analysis producing (mean voxel salience)/(standard error), to split-half signal/noise estimates. This includes **a.** standard NPAIRS estimation of voxel signal, normalized by global noise standard deviation (Z-scored) for each resample, and **b.** voxel signal, normalized by standard error (bootstrap ratios) or standard deviation (split-half Z-scores) estimated at each voxel. **c.** plot of voxels' standard error (bootstrap) against standard deviation (split-half). Results are computed over $1,000$ split-half/bootstrap resamples.

2.3 Behavioral PLS *on a principal component subspace*

For standard behavioral PLS, we project the behavioral vector $\mathbf{y}$ directly onto $\mathbf{X}$ (the subject *SPM*s) to identify the latent basis vector $\mathbf{e} = \mathbf{y}^{\mathsf{T}}\mathbf{X}$. However, taking our cue from the literature on split-half discriminant analysis in *f*MRI (see, *e.g.,* [7, 10, 18, 19, 22]), we can regularize and de-noise the data space in which the analysis is performed, by first applying PCA to $\mathbf{X}$, and then running a PLS analysis on a reduced PCA subspace.

The singular value decomposition [20] produces $\mathbf{X} = \mathbf{USV}^{\mathsf{T}}$, where $\mathbf{U}$ is a set of orthonormal subject-weight vectors, $\mathbf{S}$ is a diagonal matrix of singular values, and $\mathbf{V}$ is a set of orthonormal image basis vectors. We represent $\mathbf{X}$ in a reduced k-

dimensional PCA space ($k \leq N$), by projecting onto the subset of 1 to k image bases, $\mathbf{V}^{(k)} = [\mathbf{v}_1 \mathbf{v}_2 \ldots \mathbf{v}_k]$, giving $\mathbf{Q}^{(k)} = \mathbf{X}\mathbf{V}^{(k)}$. We performed PLS analysis on $\mathbf{Q}^{(k)}$, by normalizing and centering subject scores of each PC-basis, and then obtaining the projection $\mathbf{w}_i = \mathbf{y}_i^{\mathsf{T}} \mathbf{Q}_i$ that explained the most behavior variance in the new PC basis.

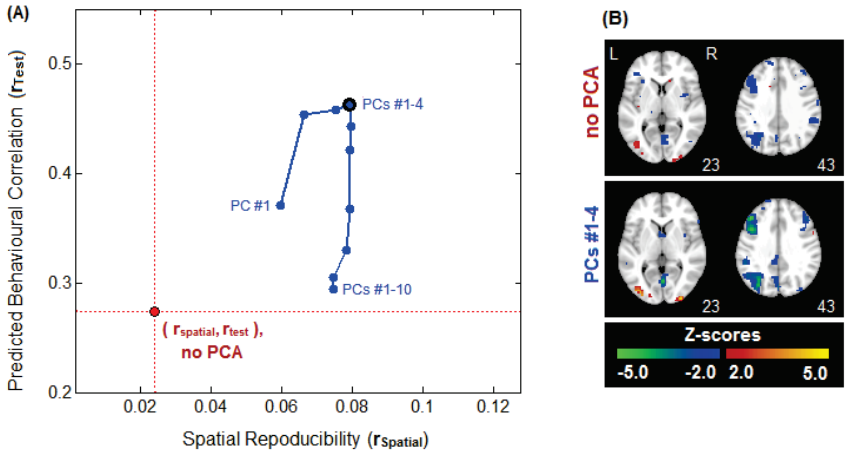

Fig. 4: **a.** Plot of median predicted behavioral correlation r_{test} and spatial reproducibility r_{spatial} of the LV brain map, for PLS performed on a PCA subspace of the subject data (*blue*). These subspaces include the 1 to k Principal Components (PCs), where we vary ($1 \leq k \leq 10$). The (r_{spatial}, r_{test}) values are plotted for each k (subspace size) as points on the curve; a subspace of PCs 1–4 simultaneously optimized (r_{spatial}, r_{test}), circled in *black*. We also plot the median (r_{test},r_{spatial}) point, estimated directly from matrix $\mathbf{X}$ for reference (*red circle*). **b.** Plots of split-half Z-scored *SPM*s with global noise estimation, for no PCA estimation (*red*), and an optimized PCA dimensionality $k = 4$ (*blue*). Positive Z-scores indicate positive correlation with the behavioral measure of reaction time, and negative Z-scores indicate negative correlation. Voxel values are computed as the mean over $1,000$ split-half resamples, with spatially global noise estimation from each split-half pair.

The predicted behavioral correlation is measured by projecting the test data onto the training PC-space, and then onto $\mathbf{w}_i$, giving behavioral correlations $r_{(i,\text{test})} = \rho(\mathbf{y}_{j\neq i}, \mathbf{w}_i(\mathbf{X}_{j\neq i}\mathbf{V}_i)$. We also obtained eigen-images by projecting back onto the voxel space (*i.e.,* $\mathbf{e}_i = \mathbf{w}_i \mathbf{V}_i^{(k)}$), to compute the $rSPM(Z)_{\text{split}}$ and reproducibility, r_{spatial}. The resulting median behavioral prediction $r_{(\text{test})}$ and reproducibility $r_{(\text{spatial})}$ are plotted in Figure 4a, as a function of the number of PC bases k. From this curve, we identify the PC subspace $k = 4$, that maximizes both r_{test} and r_{spatial}. Note that the median r_{test} and r_{spatial} are consistently higher when performed on a PCA basis

than PLS performed directly on $\mathbf{X}$, for all subspace sizes $k = 1 \ldots 10$. The predicted behavioral correlation is generally higher for the $k = 4$ PC subspace than PLS performed directly on $\mathbf{X}$ (median $\Delta r_{\text{test}} = .17$; increased for 891 of the 1,000 resamples), as is spatial reproducibility ($\Delta r_{\text{spatial}} = .05$; increased for all 1,000 resamples). Figure 4b depicts slices from the mean $rSPM(Z)s$ of PLS performed directly on $\mathbf{X}$ (top) and in an optimized PC subspace (bottom). The PCA optimization tends to increase mean Z-scores in the same areas of activation previously identified by voxel-space results, indicating that the optimized PC basis increases sensitivity of the PLS model to the same underlying set of brain regions.

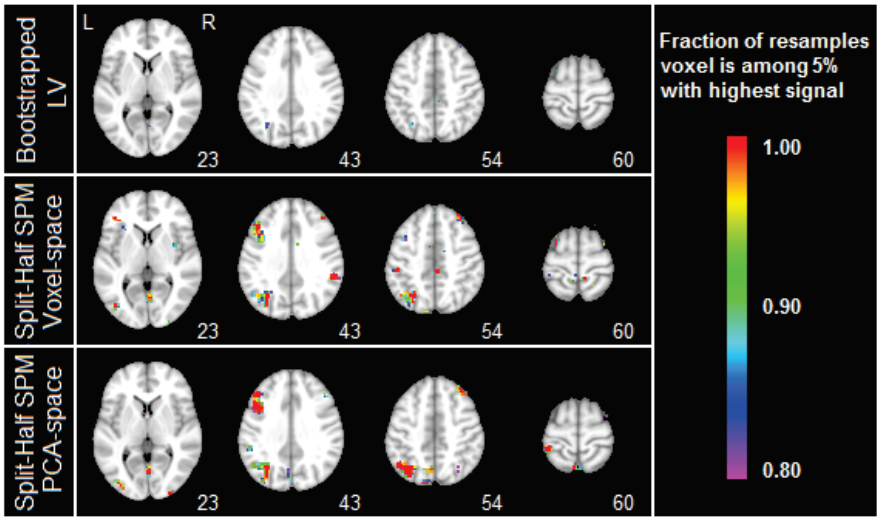

Fig. 5: Plot showing the reliability of peak voxel values. **(top)** peak LV values are shown across standard PLS bootstrap replications. Peak voxels of the split-half reproducible $rSPM(Z)s$, with global noise estimated at each split, are shown across resamples for **(middle)** voxel-space estimation, and **(bottom)** estimation on an optimized PCA subspace. For each of the 1,000 bootstrap/split-half resamples, we identified the top 5% highest-signal voxels (LV values for bootstrap estimation; Z-scores for split-half estimation). This plot measures the fraction of resamples where each voxel is part of the top 5%.

In Figure 5, we depict the stability of bootstrap and split-half resampling estimates. We compared the reliability of peak voxels across bootstrap LVs (top), relative to split-half $rSPM(Z)_{\text{split}}$ estimates with global noise estimation; the split-half model estimates a Z-scored SPM from each resampling split. Results are shown for $rSPM(Z)_{\text{split}}$ estimated directly from data matrix $\mathbf{X}$ (middle), and $rSPM(Z)_{\text{split}}$ estimated from the optimized PCA subspace of $\mathbf{X}$, $k = 4$ PCs (bottom). We measured peak signal as the top 5% of voxel signal values, for each resample (bootstrap-

estimated LV scores or split-half-estimated Z-scores). At each voxel, we measured the fraction of resamples where it was a peak voxel (*i.e.,* among the top 5%). For bootstrap LVs, only 2 of $37,284$ voxels (less than .001%) were active in more than 95% of resamples, compared to split-half Z-scored estimates of 324 voxels (0.87%; PLS computed on $\mathbf{X}$) and 343 voxels (0.92%; PLS on an optimized PCA basis). This demonstrates that although $rSPM(Z)_{\text{split}}$ with global noise estimation produces lower mean signal values than SPM_{boot} (Figure 3a), the location of peak $rSPM(Z)_{\text{split}}$ values are highly stable across resampling splits. We can therefore identify reliable *SPM* peaks with relatively few resampling iterations.

2.4 Behavioral PLS *and optimized preprocessing*

For results in Sections 2.2 and 2.3, we preprocessed the *f*MRI data to correct for noise and artifact, as outlined in [18, 19]. For this procedure, we included/excluded every combination of the preprocessing steps: (1) motion correction, (2) physiological correction, (3) regressing head-motion covariates and (4) temporal detrending with Legendre polynomial of orders 0 to 5, evaluating $2^3 \times 6 = 48$ different combinations of preprocessing steps ("pipelines").

For each pipeline, we performed an analysis in the NPAIRS split-half framework [6], and measured spatial reproducibility and prediction accuracy (posterior probability of correctly classifying independent scan volumes). We selected the pipeline that minimized the Euclidean distance from perfect prediction and reproducibility:

$$D = \sqrt{(1 - \text{reproducibility})^2 + (1 - \text{prediction})^2}, \tag{2}$$

independently for each subject. This may be compared to the standard approach in *f*MRI literature, which is to apply a single fixed pipeline to all subjects. We compared the current "individually optimized" results with the optimal "fixed pipeline," of motion correction and 3^{rd}-order detrending; this was the set of steps that, applied to all subjects, minimized the D metric across subjects (details in [18]).

Figure 4 compares fixed pipeline results (*red*) to individually optimized data (*blue*), for PLS on a PCA subspace. Figure 4a show, for both pipelines, median behavioral prediction $r_{(\text{test})}$ and reproducibility $r_{(\text{spatial})}$ plotted as a function of PCA dimensionality. Data with fixed preprocessing (*red*) optimized $r_{(\text{test})}$ and $r_{(\text{spatial})}$ at PC #1, a lower dimensionality than individually optimized preprocessing (*blue*), at PCs #1–4. For the optimized PC bases (circled in *black*), individual pipeline optimization improves over fixed pipelines with median $\Delta r_{(\text{test})} = .11$ (increased for 898 out of the $1,000$ resamples), and $\Delta r_{(\text{spatial})} = .06$ (increased for 810 out of the $1,000$ resamples). Figure 4b shows sample slices from the mean Z-scored *SPM*s, in the optimized PC subspaces. Individual subject pipeline optimization generally produces higher peak Z-scores, and sparser, less noisy SPMs, than fixed preprocessing.

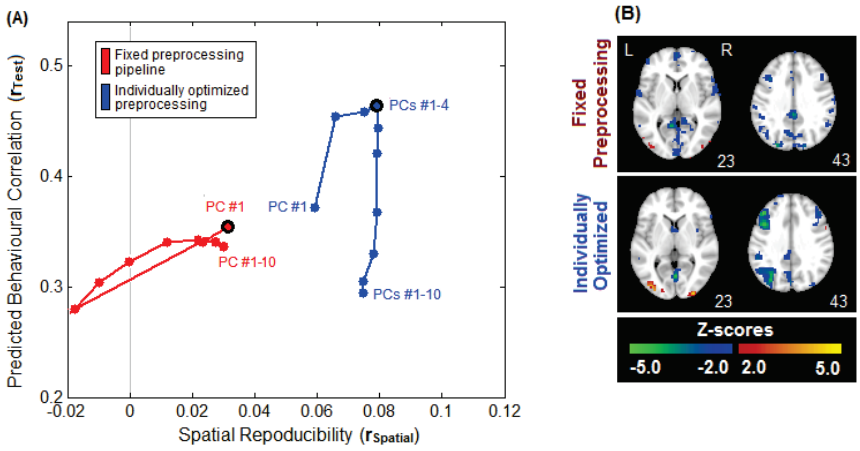

Fig. 6: **a.** Plot of median predicted behavioral correlation r_{test} and spatial reproducibility r_{spatial} of the LV brain map for PLS, when performed on a PCA subspace of the subject data. Results are plotted for data preprocessed with a fixed set of steps (*red*; all subjects have the same preprocessing applied), and with preprocessing individually optimized for each subject (*b*lue; this is the pipeline used for all previous results). For both datasets, subspaces include the 1 to k principal components (PCs), where we vary ($1 < k < 10$). The (r_{spatial}, r_{test}) values are plotted for each k (subspace size) as points on the curve; we circled in *black* the PC-space that optimized (r_{spatial}, r_{test}) for each pipeline set. **b.** Plots of split-half Z-scored *SPM*s with global noise estimation under the optimal PC subspace, for the optimal fixed pipeline (*red*; PC #1), and individually optimized pipelines (*blue*; PCs #1-4). Positive Z-scores indicate areas of positive correlation with the behavioural measure (reaction time), and negative Z-scores indicate negative correlation. Voxel values are computed as the mean over $1,000$ split-half resamples.

3 Discussion and Conclusions

The results presented in Figure 1 indicate that bootstrapping behavioral PLS values may result in a large upward bias in estimated behavioral correlation values (Figure 1, left) that is similar to the prediction biases encountered from training sets (Figure 1, middle) in training-test frameworks such as split-half resampling. Based on our prior experience with such prediction models, this upward bias is caused by over-fitting a low-dimensional categorical or behavioral vector in the high dimensional space spanned by the brain images, without appropriate model regularization. Therefore, the measured correlations from bootstrapped behavioral PLS apply only to the data set used for their estimation and cannot be generalized. In contrast, the

much lower split-half test estimates of behavioral correlation in Figure 1 (right) are generalizable but are potentially biased downwards, being based on relatively small training/test groups of only 10 subjects.

Non-generalizable training bias is also reflected in the plots of median LV scores vs. behavioral measures, in Figure 2. If the scores are computed from the training-space estimates, we obtain a stronger linear trend and less variability across splits, compared to independent test data projected onto the training basis. As shown in Figure 2, plotting the test-space scores may also reveal potential prediction outliers that are not evident in the training plots.

The Figure 3a plot also shows that bootstrapped peak *SPM* signals are consistently higher than standard split-half global *Z*-scoring. However, Figure 3b shows that on this is primarily a function of the different noise estimators, as the voxel-wise, split-half noise estimation *SPM* is highly correlated with the bootstrap estimated *SPM*. Both of the scatter-plots show a strong monotonic relation between SPM_{boot} and the $rSPM(Z)s$, indicating that regardless of the estimation procedure, approximately the same spatial locations drive both bootstrap and split-half analyses. Even for voxel-wise noise estimation, the difference between split-half and bootstrap *SPMs* is primarily driven by the local noise estimates (plotted in Figure 3c), whereas mean signal values are highly similar.

Figure 4 shows that the original $\mathbf{X}$ data space can be better regularized and stabilized, by projecting data onto a PC subspace prior to analysis. By adapting the number of PC dimensions, we trace out a behavioral correlation vs. reproducibility curve as a function of the number of PCs, similar to the prediction vs. reproducibility curves observed in discriminant models [10, 22]. These results highlight, again, the ill-posed nature of the PLS data-analysis problem, and the importance of regularizing *f*MRI data. We also note that even a full-dimensionality PC-space model (*e.g.,* PCs 1-10 included in each split-half) outperforms estimation directly on the matrix $\mathbf{X}$. The PCA projects data onto the bases of maximum variance, prior to standard PLS normalization (giving zero mean and unit variance to scores of each PC basis). The superior performance of PCs 1-10 over no PC basis (Figure 4) indicates that the variance normalization in voxel space may significantly limit the predictive generalizability of behavioral PLS results for some analyses.

Figure 5 demonstrates the advantages of split-half resampling with global noise estimation. For each split, we generate a single Z-scored $rSPM(Z)$, for which peak voxels tend to be highly consistent across $rSPM(Z)s$ of individual resampling splits. This allows us to measure voxel *Z*-scores on a little as one resampling split. The stability of the peak activations also allows us to identify reliable brain regions from a single split, which is not available to voxel-wise bootstrap estimation. The cross-validation framework is therefore particularly useful when only limited *f*MRI data is available, and has been previously used to optimize preprocessing in brief task runs of less than 3 minutes in length (*e.g.,* [18, 19]).

The results of Figure 4 compared data with preprocessing choices optimized on an individual subject basis, relative to the standard *f*MRI approach of using a single fixed pipeline. Results indicate that optimizing preprocessing choices on an individual subject basis can significantly improve predicted test correlation and the spatial

reproducibility of LV maps in behavioral PLS. Note that pipeline optimization was performed independently of any behavioral measures, as we chose preprocessing steps to optimize *SPM* reproducibility and prediction accuracy of the linear discriminant analysis model. These results demonstrate that improved preprocessing may help to better detect brain-behavior relationships in *f*MRI data.

References

1. D. Wilkinson, and P. Halligan, "The relevance of behavioral measures for functional-imaging studies of cognition," *Nature Review Neuroscience* **5**, pp. 67–73, 2004.
2. A. R. McIntosh, "Mapping cognition to the brain through neural interactions," *Memory* **7**, pp. 523–548, 1999.
3. 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.
4. N. Morch, L. K. Hansen, S. C. Strother, C. Svarer, D. .A. Rottenberg, B. Lautrup, R. Savoy, and O. B. Paulson, "Nonlinear versus linear models in functional neuroimaging: Learning curves and generalization crossover," *Information Processing in Medical Imaging*, J. Duncan and G. Gindi, eds.; Springer-Verlag, New York, pp. 259–270, 1997.
5. A. J. O'Toole, F. Jiang, H. Abdi, N. Pénard, J. P. Dunlop, and M. A. Parent, "Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data," *Journal of Cognitive Neuroscience* **19**, pp. 1735–1752, 2007.
6. 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.
7. S. C. Strother, S. LaConte, L. K. Hansen, J. Anderson, J. Zhang, S. Pulapura, and D. Rottenberg, "Optimizing the *f*MRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis," *Neuroimage* **23 Suppl 1**, pp. S196–S207, 2004.
8. S. C. Strother, "Evaluating *f*MRI preprocessing pipelines," *IEEE Engineering in Medicine and Biology Magazine* **25**, pp. 27–41, 2006
9. H. Abdi, J. P. Dunlop, and L. J. Williams, "How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS)," *Neuroimage* **45**, pp. 89–95, 2009.
10. S. Strother, A. Oder, R. Spring, and C. Grady, "The NPAIRS Computational statistics framework for data analysis in neuroimaging," *presented at the 19th International Conference on Computational Statistics, Paris, France*, 2010.
11. R. Kustra, and S. C. Strother, "Penalized discriminant analysis of [15O]-water PET brain images with prediction error selection of smoothness and regularization hyperparameters," *IEEE Transactions in Medical Imaging* **20**, pp. 376–387, 2001.
12. N. Meinshausen, and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, pp. 417–473, 2010.
13. K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
14. H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS Regression)," *WIREs Computational Statistics* **2**, pp. 97–106, 2010.
15. J. G. Snodgrass, and M. Vanderwart, "A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity," *Journal of Experimental Psychology: Human Learning* **6**, pp. 174–215, 1980.
16. C. Grady, M. Springer, D. Hongwanishkul, A. R. McIntosh, and G. Winocur, "Age-related changes in brain activity across the adult fifespan: A failure of inhibition?," *Journal of Cognitive Neuroscience* **18**, pp. 227–241, 2006.

17. F. Tam, N. W. Churchill, S. C. Strother, and S. J. Graham, "A new tablet for writing and drawing during functional MRI," *Human Brain Mapping* **32**, pp. 240–248, 2011.
18. N. W. Churchill, A. Oder, H. Abdi, F. Tam, W. Lee, C. Thomas, J. E. Ween, S. J. Graham, and S. C. Strother, "Optimizing preprocessing and analysis pipelines for single-subject *f*MRI: I. Standard temporal motion and physiological noise correction methods," *Human Brain Mapping* **33**, pp. 609–627, 2012.
19. N. W. Churchill, G. Yourganov, A. Oder, F. Tam, S. J. Graham, and S. C. Strother, "Optimizing preprocessing and analysis pipelines for single-subject *f*MRI: 2. Interactions with ICA, PCA, task contrast and inter-subject heterogeneity," *PLoS One* **7**, (e31147), 2012.
20. H. Abdi, "Singular value decomposition (SVD) and generalized singular value decomposition (GSVD)," in *Encyclopedia of Measurement and Statistics*, N. Salkind, ed., pp. 907–912, Sage, Thousand Oaks, 2007.
21. K. A. Bollen, and R. W. Jackman, "Regression diagnostics: An expository treatment of outliers and influential cases," in J. Fox, and J.S. Long, (eds.), *Modern Methods of Data Analysis*, pp. 257–291. Sage, Newbury Park, 2012.
22. P. M. Rasmussen, L. K. Hansen, K. H. Madsen, N. W. Churchill, and S. C. Strother, "Pattern reproducibility, interpretability, and sparsity in classification models in neuroimaging," *Pattern Recognition* **45**, pp. 2085–2100, 2012.