Bootstrapped Permutation Test for Multiresponse Inference on Brain Behavior Associations

Bernard Ng, Jean Baptiste Poline, Bertrand Thirion, Michael Greicius

Bootstrapped Permutation Test for Multiresponse Inference on Brain Behavior Associations

Bernard Ng[1,2], Jean Baptiste Poline[2,3], Bertrand Thirion[2], Michael Greicius[1], and IMAGEN Consortium

[1]Functional Imaging in Neuropsychiatric Disorders Lab, Stanford University, United States
[2]Parietal team, Neurospin, INRIA Saclay, France
[3]Hellen Wills Neuroscience Institute, University of California Berkeley, United States
bernardyng@gmail.com

Abstract. Despite that diagnosis of neurological disorders commonly involves a collection of behavioral assessments, most neuroimaging studies investigating the associations between brain and behavior largely analyze each behavioral measure in isolation. To jointly model multiple behavioral scores, sparse multiresponse regression (SMR) is often used. However, directly applying SMR without statistically controlling for false positives could result in many spurious findings. For models, such as SMR, where the distribution of the model parameters is unknown, permutation test and stability selection are typically used to control for false positives. In this paper, we present another technique for inferring statistically significant features from models with unknown parameter distribution. We refer to this technique as bootstrapped permutation test (BPT), which uses Studentized statistics to exploit the intuition that the variability in parameter estimates associated with relevant features would likely be higher with responses permuted. On synthetic data, we show that BPT provides higher sensitivity in identifying relevant features from the SMR model than permutation test and stability selection, while retaining strong control on the false positive rate. We further apply BPT to study the associations between brain connectivity estimated from pseudo-rest fMRI data of 1139 fourteen year olds and behavioral measures related to ADHD. Significant connections are found between brain networks known to be implicated in the behavioral tasks involved. Moreover, we validate the identified connections by fitting a regression model on pseudo-rest data with only those connections and applying this model on resting state fMRI data of 337 left out subjects to predict their behavioral scores. The predicted scores are shown to significantly correlate with the actual scores of the subjects, hence verifying the behavioral relevance of the found connections.

Keywords: Bootstrapping, brain behavior associations, connectivity, fMRI, multiresponse regression, permutation test, statistical inference

1 Introduction

Diagnosis of neurological disorders generally entails assessments of multiple behavioral domains. For instance, Attention Deficit Hyperactivity Disorder (ADHD) is

commonly diagnosed based on a collection of criteria related to inattention, hyperactivity, and impulsivity as specified in the Diagnostic and Statistical Manual of Mental Disorders (DSM). Thus, for most cases, it is the aggregate of multiple criteria that characterizes a neurological disorder. Past neuroimaging studies investigating brain behavior relationships typically analyze each behavioral measure independently [1]. One of the state-of-the-art approaches for jointly modeling multiple response variables is to incorporate a group least absolute shrinkage and selection operator (LASSO) penalty into the regression model [2] to promote selection of features associated with all response variables. However, directly applying this sparse multi-response regression (SMR) technique and assuming that all features corresponding to nonzero regression coefficients are relevant could result in many spurious findings, since SMR alone does not control for false positives [3]. Another approach is to find linear combinations of features and responses that best correlate with each other using partial least square (PLS) or canonical correlation analysis (CCA), which can be cast as a reduced rank regression (RRR) problem [4]. In limited sample settings, especially when the number of features exceeds the number of samples, sparse variants of PLS, CCA, and RRR are often used [5], but these sparse variants in their raw forms suffer the same limitation as SMR in terms of false positives not being controlled.

The growing feature dimensionality of today's problems warrants caution in controlling for false positives [6]. A number of techniques have been put forth for addressing this critical concern in the context of sparse regression [7]. The key idea behind these techniques is to de-bias the sparse regression coefficient estimates, so that parametric inference can be applied to generate approximate p-values. How to de-bias the parameter estimates of SMR as well as sparse variants of PLS, CCA, and RRR is currently unclear. For models with unknown parameter distribution, a widely-used technique is permutation test (PT) [8], which is applicable to any statistics generated from the model parameters since PT requires no assumptions on the underlying parameter distribution. Another flexible technique is stability selection (SS) [9], which operates under the rationale that if we subsample the data many times and perform feature selection on each subsample using e.g. SMR, relevant features will likely be selected over a large proportion of subsamples, whereas irrelevant features will unlikely be repeatedly selected. Importantly, SS has a theoretical threshold that bounds the expected number of false positives. Also, SS eases the problem of regularization level selection in penalized models, such as SMR, in which only a range of regularization levels needs to be specified without having to choose a specific level. However, as we will show in Section 2.2 and 4, the choice of threshold and regularization range has a major impact on the results.

Further complicating statistical inference on multi-feature models is the problem of multicollinearity [3]. In the face of correlated features, small perturbations to the data can result in erratic changes in the parameter estimates. In particular, sparse models with a LASSO penalty tends to arbitrarily select one feature from each correlated set [3]. One way to deal with this problem is to perturb the data e.g. by subsampling as employed in SS, and examine which features are consistently selected. Complementing this strategy is a technique called Randomized LASSO, which involves deweighting a random subset of features for each subsample. This combined tech-

nique is shown to improve relevant feature identification over pure subsampling [9]. Another strategy is to cluster the features to moderate their correlations, which has the additional advantage of reducing the feature dimensionality [3].

In this paper, we present a technique that combines bootstrapping with permutation test for inferring significant features from models with unknown parameter distribution. We refer to this technique as bootstrapped permutation test (BPT). BPT is originally proposed for inferring significant features from classifier weights [10], but as discussed here and in the next section, BPT is in fact applicable to arbitrary models with a number of properties that makes it advantageous over PT and SS. Bootstrapping is traditionally used for assessing variability in model parameters. In BPT, the variability differences in parameter estimates with and without permutation are exploited. The intuition is that parameter estimates of relevant features are presumably more variable when responses are permuted. Thus, dividing the parameter estimates with and without permutation by their respective standard deviation should magnify their magnitude differences. This intuition is incorporated by using Studentized statistics, as generated by taking the mean of bootstrapped parameter estimates and dividing it by the standard deviation. The Studentized statistics is known to be approximately normally-distributed [11]. Thus, we can generate a null distribution by fitting a normal distribution to Studentized statistics derived from the permuted responses, thereby enabling parametric inference, which is statistically more powerful than pure PT [12]. Also, BPT is more flexible than SS, since it can directly operate on parameter estimates from any models without the need for feature selection, which could be nontrivial for certain non-sparse models that do not possess an inherent feature selection mechanism. In this work, we focus on the SMR model for drawing associations between brain connectivity estimated from functional magnetic resonance imaging (fMRI) data and multiple behavioral measures related to attention deficit hyperactivity disorder (ADHD). Functional connectivity is typically estimated by computing the Pearson's correlation between fMRI time series of brain region pairs, which are highly inter-related. To reduce the correlations between these connectivity features, we cluster them based on the network to which each brain region belongs, thereby accounting for the similarity between time series of brain regions within the same network. To compare BPT against PT and SS, we generate synthetic behavioral scores using network-level connectivity estimates derived from real fMRI data. We also apply these techniques on pseudo-rest fMRI data from 1139 fourteen year olds in identifying significant connections that are relevant to ADHD behavioral measures. The identified connections are validated on resting state fMRI data from 337 left out subjects by comparing their predicted and actual scores.

2 Methods

We first briefly review SMR (Section 2.1) and describe how stability selection (Section 2.2) can be incorporated to control for false positives. We then discuss the properties of PT, and how BPT improves upon PT via the use of Studentized statistics as generated by bootstrapping (Sections 2.3).

2.1 Sparse Multiresponse Model

Let $\mathbf{Y}$ be a $n \times q$ matrix, where n is the number of samples, and q is the number of response variables. Further, let $\mathbf{X}$ be a $n \times d$ matrix, where d is the number of features. The standard way for assessing associations between $\mathbf{Y}$ and $\mathbf{X}$ is via regression:

$$\min_{\boldsymbol{\beta}} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2, \tag{1}$$

where $\boldsymbol{\beta}$ is the $d \times q$ regression coefficient matrix. The optimal $\boldsymbol{\beta}$ obtained by solving (1) is equivalent to regressing each column of $\mathbf{Y}$ on $\mathbf{X}$ independently. Thus, the relations between columns of $\mathbf{Y}$ are ignored. To incorporate this information in estimating $\boldsymbol{\beta}$, one of the state-of-the-art approaches is to employ the SMR model [2]:

$$\min_{\boldsymbol{\beta}} \left\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta}_g \right\|_{2,1}, \text{ s.t. } \left\| \boldsymbol{\beta}_g \right\|_{2,1} = \sum_{i=1}^{d} \left\| \boldsymbol{\beta}_{i,:} \right\|_2 = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{q} \boldsymbol{\beta}_{ij}^2} \tag{2}$$

where $\|\boldsymbol{\beta}_g\|_{2,1}$ is the group LASSO penalty and each row of $\boldsymbol{\beta}$, denoted as $\boldsymbol{\beta}_{i,:}$, corresponds to a feature. With elements of each $\boldsymbol{\beta}_{i,:}$ taken as a group, only features associated with all q response variables would be selected with the corresponding $\boldsymbol{\beta}_{i,j} \neq 0$ for all j. To set λ, we search over 100 λ's in $[\lambda_{\max}, \lambda_{\min}]$, where $\lambda_{\max} = max_j ||\mathbf{X}_{:,j}^{\mathrm{T}} \mathbf{Y}||_2$ and $\lambda_{\min} = c\lambda_{\max}$, $c < 1$. Optimal λ is defined as the one that minimizes the prediction error over 1000 subsamples with the data randomly split into 80% for model training and 20% for error evaluation. A fast solver of (2) is implemented in GLMNET [13].

2.2 Stability Selection

A problem with assuming features associated with nonzero $\boldsymbol{\beta}_{i,:}$ in SMR are all relevant is that this property is true only under very restricted conditions, which are largely violated in most real applications [3]. In particular, this guarantee on correct feature selection does not hold when features are highly correlated, which is often the case for real data [3]. With correlated features, perturbations to the data can result in drastic changes in the features that are selected. Based on this observation, an intuitive approach to deal with correlated features is to perturb the data and declare features that are consistently selected over different perturbations as relevant, which is the basis of SS. We describe here SS in the case of SMR, but SS can generally be applied to any models that are equipped with a feature selection mechanism. Given $\mathbf{X}$, $\mathbf{Y}$, and $[\lambda_{\max}, \lambda_{\min}]$, SS combined with Randomized LASSO proceeds as follows [9]:

1. Multiply each column of $\mathbf{X}$ by 0.5 or 1 selected at random.
2. Randomly subsample $\mathbf{X}$ and $\mathbf{Y}$ by half to generate $\mathbf{X}^s$ and $\mathbf{Y}^s$.
3. For each λ in $[\lambda_{\max}, \lambda_{\min}]$, apply SMR to generate $\boldsymbol{\beta}^s(\lambda)$. Let $\mathbf{S}^s(\lambda)$ be a $d \times 1$ vector with elements corresponding to selected features, i.e. nonzero $\boldsymbol{\beta}^s(\lambda)$, set to 1.
4. Repeat steps 2 and 3 S times, e.g. with $S = 1000$.
5. Compute the proportion of subsamples, $\boldsymbol{\pi}_i(\lambda)$, that each feature i is selected for each λ in $[\lambda_{\max}, \lambda_{\min}]$.
6. Declare feature i as significant if $max_\lambda \boldsymbol{\pi}_i(\lambda) \geq \pi_{\text{th}}$.

A π_{th} that controls for the expected number of false positives, $E(V)$, is given by [9]:

$$E(V) \leq \frac{1}{2\pi_{th} - 1} \frac{\gamma^2}{d}, \tag{3}$$

where V is the number of false positives and γ is the expected number of selected features, which can be approximated by: $1/S \cdot \sum_s \sum_i (\mathsf{U}_\lambda \mathbf{S}_i^s(\lambda))$. U_λ denotes the union over λ. We highlight two insights on (3) that have major implications on applying SS in practice. First, (3) is a conservative bound on the family-wise error rate (FWER) = $P(V \geq 1)$, since $E(V) = \sum_{v=1}^{\infty} P(V \geq v) > P(V \geq 1)$. To control FWER at $\alpha = 0.05$ with multiple comparison correction (MCC), i.e. $P(V \geq 1) \leq \alpha/d$, even for $\gamma = 1$, π_{th} based on (3) is >1. Since $\boldsymbol{\pi}_i(\lambda) \in [0,1]$, π_{th} should be clipped at 1, but whether FWER is still controlled is unclear. Second, a key property of SS is that it does not require choosing a specific λ. However, for $n/2 > d$, a "small enough" λ_{min} could lead to all features being selected in all subsamples, resulting in $max_\lambda \boldsymbol{\pi}_i(\lambda) = 1$. Hence, all features would be declared as significant. λ selection is thus translated into λ_{min} selection, which warrants caution. An example from real data (Section 3.2) illustrating the impact of λ_{min} and π_{th} is shown in Fig. 1(a). Even with λ_{min} set to 0.1, i.e. a λ range that would strongly encourage sparsity, a π_{th} of 0.9 (strictest π_{th} in the suggested range of [0.6, 0.9] in [9]) declares >40% of the features as significant, i.e. fails to control for false positives.

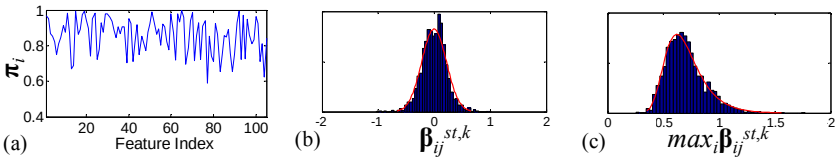

Fig. 1. Behavior of SS and BPT on real data. (a) $\boldsymbol{\pi}_i(\lambda)$ at $\lambda = 0.1$ (for SS). (b) Gaussian fit on Studentized statistics (for BPT). (c) Gumbel fit on maxima of Studentized statistics (for BPT).

2.3 Bootstrapped Permutation Testing

For models with unknown parameter distribution, including those with no intrinsic feature selection mechanisms, PT is often used to perform statistical inference. PT involves permuting responses a large number of times (e.g. 10000 in this work) and relearning the model parameters for each permutation in generating null distributions of the parameters. Features with original parameter values greater than (or less than) a certain percentile of the null, e.g. $>100 \cdot (1 - 0.025/d)^{th}$ percentile (or $<100 \cdot (0.025/d)^{th}$ percentile), are declared as significant. Equivalently, one can count the number of permutations with parameter values exceeding/below the original parameter values to generate approximate p-values. A key attribute of PT is that it does not impose any distributional assumptions on the parameters, but the cost of this flexibility is the need for a large number of permutations to ensure the resolution of the approximate p-values are fine enough for proper statistical testing, i.e. the smallest p-value attainable from N permutation is $1/N$. Also, if the underlying parameter distribution is known, the associated parametric test is statistically more powerful [12].

The central idea behind BPT is to generate Studentized statistics via bootstrapping to exploit how the variability of the parameter estimates associated with relevant features are likely higher with responses permuted. Similar to PT, BPT can be applied to any models. We describe here BPT in the context of SMR, which proceeds as follows.

Estimation of Studentized statistics, $\boldsymbol{\beta}_{ij}^{st}$:

1. Bootstrap $\mathbf{X}$ and $\mathbf{Y}$ with replacement for $B = 1000$ times, and denote the bootstrap samples as $\mathbf{X}^b$ and $\mathbf{Y}^b$.
2. Multiply each column of $\mathbf{X}^b$ by 0.5 or 1 selected at random.
3. Select the optimal λ for SMR by repeated random subsampling on $\mathbf{X}$ and $\mathbf{Y}$, and apply SMR on $\mathbf{X}^b$ and $\mathbf{Y}^b$ for each bootstrap b with this λ to estimate $\boldsymbol{\beta}^b$.
4. Compute Studentized statistics, $\boldsymbol{\beta}_{ij}^{st} = 1/B \cdot \sum_b \boldsymbol{\beta}_{ij}^b \,/\, std(\boldsymbol{\beta}_{ij}^b)$, where $std(\boldsymbol{\beta}_{ij}^b)$ is the standard deviation over bootstrap samples.

Estimation of the null distribution of $\boldsymbol{\beta}_{ij}^{st}$:

5. Permute $\mathbf{Y}$ for $N = 500$ times.
6. For each permutation k, compute $\boldsymbol{\beta}_{ij}^{st,k}$ with the same λ, samples, and feature weighting used in each bootstrap b as in the non-permuted case.
7. p-value $= 2 \cdot min(1 - \Phi(\boldsymbol{\beta}_{ij}^{st} | 1/N \cdot \sum_k \boldsymbol{\beta}_{ij}^{st,k}, std(\boldsymbol{\beta}_{ij}^{st,k})), \Phi(\boldsymbol{\beta}_{ij}^{st} | 1/N \cdot \sum_k \boldsymbol{\beta}_{ij}^{st,k}, std(\boldsymbol{\beta}_{ij}^{st,k})))$, where $\Phi(\cdot)$ = the cumulative distribution function of the normal distribution.

To account for multiple comparisons, one can apply Bonferroni correction, but this technique tends to be too conservative [8]. A more sensitive technique is to use maximum statistics [8], which entails finding the maximum $\boldsymbol{\beta}_{ij}^{st,k}$ over i for each response variable j and permutation k. Since $\boldsymbol{\beta}_{ij}^{st,k}$ is approximately normally-distributed [11], its maximum statistics should follow a Gumbel distribution. We can thus fit a Gumbel distribution to the maxima of $\boldsymbol{\beta}_{ij}^{st,k}$ for each j and declare features associated with $\boldsymbol{\beta}_{ij}^{st}$ exceeding a certain percentile of the fitted Gumbel distribution as significant. Negative $\boldsymbol{\beta}_{ij}^{st}$ can be similarly tested with maximum replaced by minimum. An example from real data (Section 3.2) illustrating a Gaussian and a Gumbel distribution fit to the empirical distribution of $\boldsymbol{\beta}_{ij}^{st,k}$ and its maxima, respectively, for an exemplar feature and response is shown in Fig. 1(b) and (c). Another sensitive MCC technique is false discovery rate (FDR) correction [8], which involves sorting the p-values in ascending order, and testing the l^{th} p-value against $l \cdot \alpha / d$, e.g. $\alpha = 0.05$.

We highlight here properties of BPT that make it advantageous over PT and SS. First, $std(\boldsymbol{\beta}_{ij}^b)$ of relevant features are likely larger with responses permuted. By using Studentized statistics, i.e. dividing the bootstrapped mean of $\boldsymbol{\beta}_{ij}$ by $std(\boldsymbol{\beta}_{ij}^b)$, the magnitude differences in $\boldsymbol{\beta}_{ij}$ between the permuted and non-permuted cases would be magnified. Second, Studentized statistics approximately follows a normal distribution [11], hence justifying the use of parametric inference, which is more powerful than PT. It is worth noting that normal approximation on Studentized statistics is more accurate than on conventional mean [11], which provides another reason for dividing by the standard deviation. Lastly, in contrast to SS, BPT does not require a feature selection mechanism. Instead, BPT can directly operate on the parameter estimates, which additionally accounts for the magnitude of $\boldsymbol{\beta}_{ij}$. Also, BPT facilitates greater flexibility in the choice of statistical inference procedures, e.g. MCC with maximum statistics cannot be easily incorporated into SS.

3 Materials

3.1 Synthetic Data

To generate synthetic data, we used the network connectivity matrix, $\mathbf{X}_{real}$, estimated from pseudo-rest fMRI data (Section 3.2), which comprised 1139 subjects and 105 features. This way, the feature correlations present in the real data would be retained to enable method evaluation under more realistic settings. Two scenarios were considered: $n = 50 < d = 105$ and $n = 200 > d = 105$. For each scenario, we generated 50 $n \times 3$ response matrices, $\mathbf{Y}_{syn}$. Each $\mathbf{Y}_{syn}$ was created by randomly selecting n out of 1139 subjects and 10 out of 105 features taken as ground truth. Denoting the resulting $n \times 10$ feature matrix as $\mathbf{X}_{syn}$, we generated $\mathbf{Y}_{syn}$ as $\mathbf{X}_{syn}\boldsymbol{\beta}_{syn} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}_{syn}$ is a $d \times 3$ matrix with each element randomly drawn from a uniform distribution, $U(0.01, 0.1)$, and $\boldsymbol{\varepsilon}$ corresponds to Gaussian noise. Each $\mathbf{Y}_{syn}$ and the corresponding n rows of $\mathbf{X}_{real}$ (i.e. with all features kept, not $\mathbf{X}_{syn}$) constituted as one synthetic dataset.

3.2 Real Data

Neuroimaging and behavioral data from ~1,000 fourteen year olds were obtained from the IMAGEN database [14]. Details on data acquisition can be found in [14]. In this work, we focused on 3 behavioral measures of the Cambridge Neuropsychological Test Automated Battery (CANTAB) associated with ADHD [15]. Specifically, we examined the between error and strategy scores of the spatial working memory (SWM) task as well as the response accuracy score of the rapid visual information processing (RVP) task. For estimating connectivity, we used fMRI data acquired during a localizer task (140 volumes) as well as at rest (187 volumes) for model training and validation, respectively. For the task fMRI data, slice timing correction, motion correction, and spatial normalization to MNI space were performed using SPM8. Motion artifacts, white matter and cerebrospinal fluid confounds, principal components of high variance voxels found using CompCor [16], and their shifted variants as well as task paradigm convolved with the canonical hemodynamic response and discrete cosine functions (for highpass filtering at 0.008 Hz) were regressed out from the task fMRI time series. The task regressors were included to decouple co-activation from connectivity in generating pseudo-rest data. The resting state fMRI data were similarly preprocessed except a bandpass filter with cutoff frequencies of 0.01Hz and 0.1Hz was used. Taking the intersection of subjects with all 3 behavioral scores and task fMRI data, while excluding those who also have resting state fMRI data to ensure that subjects for model training and validation are independent, 1139 subjects were available for model training. For validation, 337 subjects with all 3 behavioral scores and rest data were available. To estimate connectivity, we generated brain region time series by averaging the voxel time series within the 90 regions of interest (ROIs) of a publicly available functional atlas that span 14 large-scale networks [17]. The Pearson's correlation between ROI time series was taken as estimates of connectivity. Since time series of ROIs within the same network would be similar, the magnitude of their correlation with other ROIs would also be similar. To reduce the correlations

between these connectivity features, we computed the within and between network connectivity from the 90×90 Pearson's correlation matrix. For estimating within network connectivity, we averaged the Pearson's correlation values between all ROI pairs within each network, which resulted in 14 features. For estimating between network connectivity, we averaged the Pearson's correlation values of all between network ROI connections for each network pair, which resulted in 91 features. Age, sex, scan site, and puberty development scores were regressed out from both the behavioral scores and the network connectivity features (separately for the training and validation subjects), which were further demeaned and scaled by the standard deviation.

4 Results and Discussion

On synthetic (Section 4.1) and real data (Section 4.2), we compared BPT at $p < 0.05$ with maximum statistics-based MCC and FDR correction against SMR with features associated with non-zero regression coefficients assumed significant, SS at $p < 0.05/d$ with π_{th} set based on (3) and $\pi_{th} = 0.75$ (midpoint of suggested range of [0.6, 0.9] in [9], and PT at $p < 0.05$ with maximum statistics-based MCC and FDR correction. Central to applying SMR is the choice of λ. We thus examined results for $\lambda_{min} = 0.001$, 0.01 and 0.1. Setting λ_{min} to 0.1 produces a very narrow λ range that tends to generate overly-sparse $\boldsymbol{\beta}$. This value of λ_{min} was considered due to SS's failure to control for false positives for smaller λ_{min} in our real data experiments.

4.1 Synthetic Data

We evaluated the contrasted techniques by computing the average true positive rate (TPR) and average false positive rate (FPR) over the 50 synthetic datasets for each n/d scenario. TPR was defined as the proportion of ground truth significant features that were correctly identified, and FPR was defined as the proportion of ground truth non-significant features that were incorrectly found. Results for $n/d = 50/105$ and $n/d = 200/105$ are shown in Figs. 2 and 3. PT declared all features as non-significant, hence its results were not displayed. Also, $\lambda_{min} = 0.1$ led to degraded performance for all techniques, which was likely due to the resulting λ range enforcing overly-sparse solutions. We thus focused our discussion on results for $\lambda_{min} = 0.001$ and 0.01.

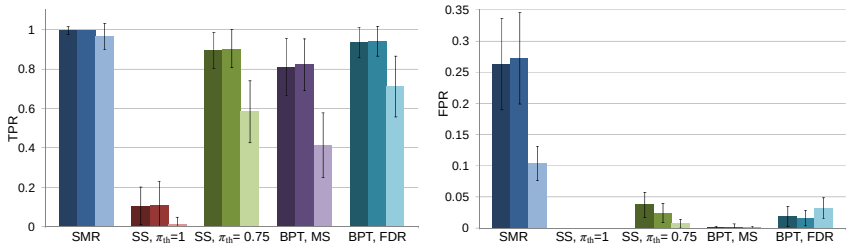

Fig. 2. Synthetic data results for $n/d = 50/105 < 1$. Each set of three bars (left to right) correspond to $\lambda_{min} = 0.001$, 0.01, and 0.1, respectively. MS = maximum statistics.

For n/d = 50/105, SMR achieved a TPR close to 1, but also included many false positives. Using SS with π_{th} set based on (3) had FPR well controlled, but TPR was merely 0.1. Note that π_{th} was >1 in all cases even without MCC, and was thus clipped at 1. By relaxing π_{th} to 0.75, SS's TPR increased to ~0.9, and FPR was <0.04, despite FPR control was not guaranteed. Using BPT with maximum statistics-based MCC, which exerts strong control over FPR, attained a TPR of ~0.8 with FPR being close to 0. Relaxing the control on FPR using FDR correction resulted in TPR of ~0.94 and FPR < 0.02, which is half the FPR of SS with π_{th} = 0.75. Thus, for similar control on FPR, BPT provides higher sensitivity than SS. For n/d = 200/105, all contrasted techniques (except PT) achieved a TPR of ~1, but FPR was not well controlled with SMR and SS. In particular, SS with π_{th} = 0.75 resulted in a FPR of 1 for λ_{min} = 0.001due to all features being selected for small λ's near λ_{min} across all subsamples. Thus, declaring features as significant if $\boldsymbol{\pi}_i(\lambda) \geq \pi_{th}$ for any λ could be erroneous for small λ_{min} in $n > d$ settings. Also worth noting was the lack of sensitivity with PT, which clearly demonstrates the enhanced sensitivity gained by using Studentized statistics in BPT. Nevertheless, PT with n = 1139 attained a TPR of ~1 and FPR close to 0.

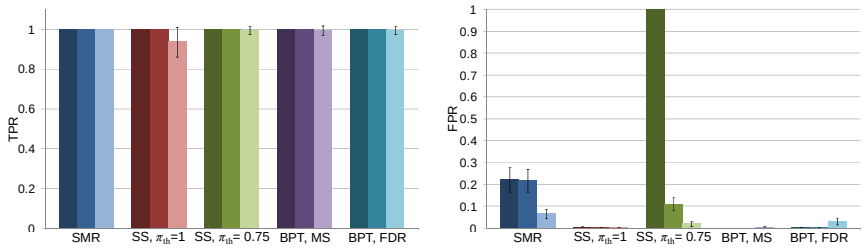

Fig. 3. Synthetic data results for n/d = 200/105 > 1. Each set of three bars (left to right) correspond to λ_{min} = 0.001, 0.01, and 0.1, respectively. MS = maximum statistics.

4.2 Real Data

We applied the contrasted techniques to the pseudo-rest fMRI data of 1139 subjects to identify the significant network connections associated with ADHD behavioral measures. Since the ground truth is unknown, we validated the identified connections by first fitting a regression model (1) to the pseudo-rest data but with only the identified connections retained. We then applied these models to the resting state fMRI data of 337 left out subjects to predict their behavioral scores. The Pearson's correlation between the predicted and actual scores of these subjects was computed with significance declared at the nominal α level of 0.05.

Using BPT with FDR correction, connections between the right executive control network (ECN) and the anterior salience network (ASN), the language network and the auditory network, the visuospatial network and the dorsal default mode network (DMN), as well as the precuneus and the language network were found to be significant (Fig. 3). These findings were consistent for λ_{min} = 0.001 and 0.01. The Pearson's correlation between the predicted and actual scores was significant for all three be-

havioral measures. The ECN comprises the dorsolateral prefrontal cortex (DLPFC) and the parietal lobe, which are involved with vigilance, selective and divided attention, working memory, executive functions, and target selection [18]. Strongly connected to the DLPFC is the dorsal anterior cingulate cortex (part of the ASN), which plays a major role in target and error detection [18]. Hence, the finding of the connection between the ECN and the ASN to be significant matches well with the cognitive processes required for the SWM and RVP tasks. Also, the visuospatial manipulation and memory demands involved with these tasks would explain the detection of the connection between the visuospatial network and the dorsal DMN. The detection of the connection between the precuneus and the language network might relate to the variability in the level of linguistic strategy employed by the subjects to process spatial relations [19]. Interestingly, although all subjects have subclinical ADHD scores, there is moderate variability in their values [20], which might explain the resemblance between the found networks and those affected by ADHD [18, 21, 22]. We note that BPT with the stricter maximum statistics-based MCC detected only the connection between the Precuneus and the language network.

SMR found connections between the visuospatial network and the language network as well as connections within the left ECN to be behaviorally relevant, in addition to those found by BPT. These extra connections, although seem relevant, resulted in the Pearson's correlation for the RVP task to fall below significance. As for SS, $\pi_{th} > 1$ based on (3) for all λ_{min} tested, even for $E(V) < 0.05$ instead of $0.05/d$, i.e. no MCC. Setting λ_{min} to 0.001 and 0.01 resulted in almost all network connections declared as significant for $\pi_{th} = 1$. For $\lambda_{min} = 0.1$, only one connection survived with $\pi_{th} = 1$, and >1/3 of the connections declared significant with $\pi_{th} = 0.9$ (Fig. 1(a)). To obtain sensible results, we used $\pi_{th} = 1$ with $\lambda_{min} = 0.05$, which declared the connections between the language network and the auditory network, the primary visual network and the precuneus, as well as connections within the left ECN as significant. With these connections, only the Pearson's correlation for the RVP task was significant. Using PT detected connections between the Precuneus and the language network as well as the language network and the auditory network, which constitutes a subset of the connections found by BPT, and the Pearson's correlations obtained were similar.

We highlight here several notable observations. First, our results show that models built from pseudo-rest data can generalize to true rest data. This finding provides further support for the hypothesis that intrinsic brain activity is sustained during task performance [23]. The correlations between the predicted and actual scores, however, are rather small in absolute terms, which might limit practical applications. Second, SS is gaining popularity due to its generality and claimed robustness to regularization settings. Our results show that SS is actually sensitive to the choice of threshold and regularization range, especially for $n > d$. Thus, SS should be applied with caution. Third, applying BPT with $1/B \cdot \sum_b \boldsymbol{\beta}_{ij}^{\,b}$ without diving by $std(\boldsymbol{\beta}_{ij}^{\,b})$ resulted in degraded performance (similar to PT's), which indicates that modeling differences in variability between the permuted and non-permuted cases is key to BPT's superior sensitivity. Lastly, Studentized bootstrap conference intervals are known to have lower coverage errors than its empirically derived counterpart [24]. The observed improvements with BPT over PT could partly be attributed to this property of Studentized statistics.

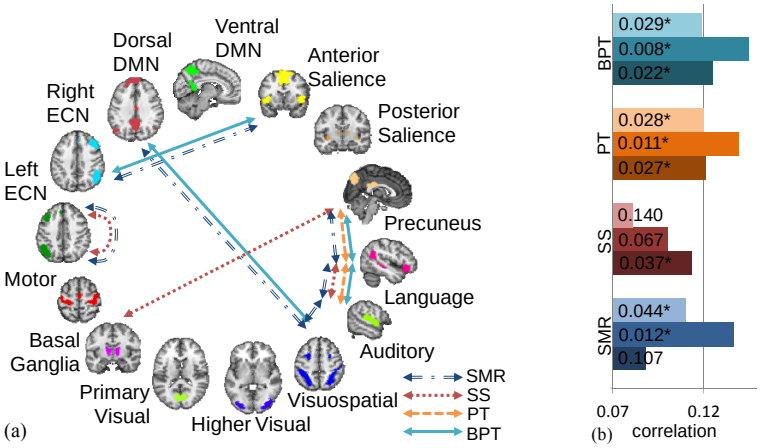

Fig. 4. Real data results. $\lambda_{min} = 0.001$, except SS where $\lambda_{min} = 0.05$ used. (a) Significant network connections found on pseudo-rest fMRI data. (b) Pearson's correlation between predicted and actual scores with p-values noted. Each set of three bars (top to bottom) correspond to SWM strategy, SWM between errors, and RVP accuracy scores. *Significance declared at $p < 0.05$.

5 Conclusions

We presented BPT for statistical inference on models with unknown parameter distributions. Superior performance over PT and SS was shown on both synthetic and real data. The resemblance of the found networks with those implicated in ADHD suggests the associated network connections might be promising for ADHD classification, which currently has accuracy <70% with most neuroimaging-based classifiers.

Acknowledgements. Bernard Ng is supported by the Lucile Packard Foundation for Children's Health, Stanford NIH-NCATS-CTSA UL1 TR001085 and Child Health Research Institute of Stanford University. Jean Baptiste Poline is partly funded by the IMAGEN project (E.U. Community's FP6, LSHM-CT-2007-037286).

References

1. Seeley, W.W., Menon, V., Schatzberg, A.F., Keller, J., Glover, G.H., Kenna, H., Reiss, A.L., Greicius, M.D.: Dissociable Intrinsic Connectivity Networks for Salience Processing and Executive Control. J Neurosci. 27, 2349–2356 (2007)
2. Simon, N., Friedman, J., Hastie, T.: A Blockwise Descent Algorithm for Group-penalized Multiresponse and Multinomial Regression. arXiv:1311.6529 (2013)
3. Varoquaux, G., Gramfort, A., Thirion, B.: Small-sample Brain Mapping: Sparse Recovery on Spatially Correlated Designs with Randomization and Clustering. Int. Conf. Machine Learning (2012)
4. De la Torre, F.: A Least-Squares Framework for Component Analysis. IEEE Trans. Patt. Ana. Mach. Intell. 34, 1041–1055 (2012)

5. Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.B., Duchesnay, E.: Significant Correlation between a Set of Genetic Polymorphisms and a Functional Brain Network Revealed by Feature Selection and Sparse Partial Least Squares. NeuroImage 63, 11–24. (2012)
6. MacArthur, D.: Methods: Face up to False Positives. Nature 487, 427–428 (2012)
7. Javanmard, A., Montanari, A.: Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. arXiv:1306.3171 (2013)
8. Nichols, T., Hayasaka, S.: Controlling the Familywise Error Rate in Functional Neuroimaging: a Comparative Review. Stat. Methods Med. Research 12, 419–446 (2003)
9. Meinshausen, N., Bühlmann, P.: Stability Selection. J. Roy. Statist. Soc. Ser. B 72, 417–473 (2010)
10. Ng, B., Dresler, M., Varoquaux, G., Poline, J.B., Greicius, M.D., Thirion, B.: Transport on Riemannian Manifold for Functional Connectivity-based Classification. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8674, pp. 405–413. Springer, Heidelberg (2014)
11. Delaigle, A., Hall, P., Jin, J: Robustness and Accuracy of Methods for High Dimensional Data Analysis based on Student's t-statistic. arXiv:1001.3886 (2010)
12. Tanizaki, H.: Power Comparison of Non-parametric Tests: Small-sample Properties from Monte Carlo Experiments. J. Applied Stat. 24, 603–632 (1997)
13. http://www.stanford.edu/~hastie/glmnet_matlab/
14. Schumann, G., et al.: The IMAGEN Study: Reinforcement-related Behaviour in Normal Brain Function and Psychopathology. Mol. Psychiatr. 15, 1128–1139 (2010)
15. Chamberlain, S.R., Robbins, T.W., Winder-Rhodes, S., Müller, U., Sahakian, B.J, Blackwell, A.D., Barnett, J.H.: Translational Approaches to Frontostriatal Dysfunction in Attention-Deficit/Hyperactivity Disorder Using a Computerized Neuropsychological Battery. Biol Psychiatry 69, 1192–1203 (2011)
16. Behzadi, Y., Restom, K., Liau, J., Liu, T.T.: A Component based Noise Correction Method (CompCor) for BOLD and Perfusion based fMRI. NeuroImage 37:90–101(2007)
17. Shirer, W.R., Ryali, S., Rykhlevskaia, E., Menon, V., Greicius, M.D.: Decoding Subject-driven Cognitive States with Whole-brain Connectivity Patterns. Cereb. Cortex 22, 158−165 (2012)
18. Bush, G., Valera, E.M., Seidman, L.J.: Functional Neuroimaging of Attention-Deficit/Hyperactivity Disorder: A Review and Suggested Future Directions. Biol Psychiatry 57, 1273–1284 (2005)
19. Wallentin, M., Weed, E., Østergaard, L., Mouridsen, K., Roepstorff, A.: Accessing the Mental Space-Spatial Working Memory Processes for Language and Vision Overlap in Precuneus. Hum. Brain Mapp. 29, 524–532 (2008)
20. Whelan, R., et. al.: Adolescent Impulsivity Phenotypes Characterized by Distinct Brain Networks. Nat Neurosci. 15, 920–925 (2012)
21. Westerberg, H., Hirvikoski, T., Forssberg, H., Klingberg, T.: Visuo-spatial Working Memory Span: A Sensitive Measure of Cognitive Deficits in Children with ADHD. Child Neuropsychol. 10, 155–161 (2004)
22. Ghanizadeh, A.: Sensory Processing Problems in Children with ADHD, A Systematic Review. Psychiatry Investig. 8, 89–94 (2011)
23. Fox, M.D., Raichle, M.E.: Spontaneous Fluctuations in Brain Activity Observed with Functional Magnetic Resonance Imaging. Nat. Rev. Neurosci. 8, 700–711 (2007)
24. Kuonen, D.: Studentized Bootstrap Confidence Intervals based on M-estimates. J. Applied Stats. 32, 443–460 (2005)