Spatial-extent inference for testing variance components in reliability and heritability studies

Ruyi Pan[a,b], Erin W. Dickie[b,c], Colin Hawco[b,c], Nancy Reid[a], Aristotle N. Voineskos[b,c], Jun Young Park[a,d,*]

[a] *Department of Statistical Sciences, University of Toronto, Toronto, ON, M5G 1Z5, Canada*
[b] *The Centre for Addiction and Mental Health, Toronto, ON, M5T 1R8, Canada*
[c] *Department of Psychiatry, University of Toronto, Toronto, ON, M5T 1R8, Canada*
[d] *Department of Psychology, University of Toronto, Toronto, ON, M5G 1Z5, Canada*

Abstract

Clusterwise inference is a popular approach in neuroimaging to increase sensitivity, but most existing methods are currently restricted to the General Linear Model (GLM) for testing mean parameters. Statistical methods for testing *variance components*, which are critical in neuroimaging studies that involve estimation of narrow-sense heritability or test-retest reliability, are underdeveloped due to methodological and computational challenges, which would potentially lead to low power. We propose a fast and powerful test for variance components called CLEAN-V (**CLEAN** for testing **V**ariance components). CLEAN-V models the global spatial dependence structure of imaging data and computes a locally powerful variance component test statistic by data-adaptively pooling neighborhood information. Correction for multiple comparisons is achieved by permutations to control family-wise error rate (FWER). Through analysis of task-fMRI data from the Human Connectome Project across five tasks and comprehensive data-driven simulations, we show that CLEAN-V outperforms existing methods in detecting test-retest reliability and narrow-sense heritability with significantly improved power, with the detected areas aligning with activation maps. The computational efficiency of CLEAN-V also speaks of its practical utility, and it is available as an R package.

Keywords: clusterwise inference, heritability, spatial autocorrelation, task-fMRI, test-retest reliability, variance component

*Corresponding author
Email address: junjy.park@utoronto.ca (Jun Young Park)

1. Introduction

Imaging biomarkers obtained by magnetic resonance imaging (MRI) are critical to the understanding of brain functions and structures and their relationship to the diagnosis of brain disorders. Among these, functional MRI (fMRI) has been increasingly used to measure brain activation in response to experimental tasks to understand human cognition and behavior. However, recent studies using large neuroimaging databases (e.g., the Human Connectome Project (HCP) and Adolescent Brain Cognitive Development (ABCD) study) have shown poor test-retest reliability of task-fMRI measures in most brain regions [1, 2]. Since test-retest reliability provides a measure of consistency and stability of data obtained repeatedly over time, the lack of reliability of task-fMRI measures has raised concerns about current efforts using fMRI data to find brain-behavior associations [3]. Blokland et al. [4], with 319 twins in their study, discovered several regions of the brain highlighting heritable fMRI activations, but only with a loose cluster defining threshold (uncorrected voxel-wise $p < 0.05$), which might be prone to inflate false positives. These recent findings in neuroimaging motivate a need for a powerful statistical method for neuroimaging studies that involve test-retest reliability or heritability.

Conducting inference for test-retest reliability and heritability is fundamentally a variance component problem in statistics [5] because both measures can be expressed in terms of random effects for modeling dependencies between observed data (images). However, most existing variance component methods in neuroimaging literature estimate and test each variance component separately (e.g., massive univariate analysis). In contrast to testing for mean parameters, testing for variance component parameters is more challenging due to model specification and intensive computation to address spatial dependence structures. Recently, Risk and Zhu [6] showed that spatial modeling of variance components would lead to improved heritability estimation. However, their method also suffers from computational costs, and conducting statistical inference with their method is not straightforward, as the authors also noted.

We hypothesize that leveraging spatial dependencies in testing for the test-retest reliability or heritability would increase the replicability of identified regions [7]. The spatial-extent inference has shown its utility to increase sensitivity in neuroimaging compared to simple massive univariate analysis. In the context of the General Linear Model (GLM), clusterwise inference improves sensitivity through the convolution of univariate test statistics in the spatial domain, which is exemplified by the threshold-free clusterwise enhancement (TFCE) [8]. Park et al. [9] showed that combining the test statistics across neighboring vertices yielded high statistical power in linear mixed models as well. Also, spatial Bayesian GLMs that parametrically model spatial dependencies of the mean parameters resulted in more accurate and reliable estimates [10, 11]. Bernal-Rusiel et al. [12] and Risk et al. [13] showed that modeling of the local spatial dependence of the residual noise term using a Gaussian process results in increased statistical power. More recently, the CLEAN method proposed by Park and Fiecas [14] showed that additional gain in sensitivity is achieved in GLMs by

cluster-enhanced statistic, which is constructed from multivariate test statistics by applying spatial Gaussian process to the residual noise. This idea has been extended to CLEAN-R [15] in testing and localizing correspondence between two imaging modalities by modeling modality-specific spatial autocorrelation and constructing cluster-enhanced test statistics for correlation parameters. CLEAN and CLEAN-R highlight the importance of using two aspects of spatial dependencies to improve sensitivity.

The aim of this paper is to extend spatial-extent inference to testing variance components, which have been underdeveloped in neuroimaging due to their methodological and computational challenges. To overcome the limitations of massive univariate analysis in variance component testing, we propose a new powerful method called CLEAN-V (**CLEAN** for testing **V**ariance components) that tests and localizes variance components at the vertex level. CLEAN-V shares the same underlying mechanism as CLEAN [14] and CLEAN-R [15] in improving statistical power: (i) an explicit spatial autocorrelation modeling of neuroimaging data, (ii) effective spatial enhancement of test statistics, and (iii) a fast resampling procedure for making statistical inferences. However, CLEAN and CLEAN-R are primarily developed for testing GLM parameters or intermodal correlations, and their direct applications to test variance components in test-retest reliability and heritability studies are limited. Therefore, the proposed method aims to bridge this gap and provide a robust solution for testing and localizing variance components in neuroimaging studies.

The rest of the paper is organized as follows. In Section 2, we describe the details of CLEAN-V and compare it to the existing variance component testing methods. In Section 3, we compare the performance of CLEAN-V to other existing methods by conducting comprehensive data-driven simulation studies. We apply CLEAN-V to multiple tasks from Human Connectome Project (HCP) to compare statistically significant test-retest reliability and heritability regions across different experimental tasks. We conclude with discussion in Section 4.

2. Methods

2.1. Notation and model specifications

We consider a dataset consisting of subject-level continuous neuroimaging data. In this paper, we focus on activation levels in task-fMRI measured on the cortical surface, although other neuroimaging data types such as cortical thickness and cerebral blood flow that are mapped onto the cortex are also applicable. Task activation for each individual at each location is obtained by applying GLM to the observed blood oxygenation level-dependent (BOLD) signal on the convolution of the stimulus function with the hemodynamic response function [16, 17].

Consider N total images where each image has V vertices. Note that N does not necessarily denote subjects in reliability studies, where more than two scans from a subject are obtained. Let $i = 1, \ldots, N$ be the indices for images, and $v = 1, \ldots, V$ be the indices for vertices in a cortical surface. The observed imaging data $\mathbf{y}(v) = (y_1(v), y_2(v), \ldots, y_N(v))'$ are modelled by

$$
\begin{aligned}
&\mathbf{y}(v) = \mathbf{X}\boldsymbol{\beta}(v) + \mathbf{r}(v) + \boldsymbol{\epsilon}(v), \\
&\mathbf{r}(v) \sim \mathcal{MVN}(\mathbf{0}_N, \theta^2(v) \cdot \mathbf{K}), \quad v = 1, \ldots, V, \\
&\boldsymbol{\epsilon}_i = (\epsilon_i(1), \ldots, \epsilon_i(V))' \overset{i.i.d}{\sim} \mathcal{MVN}(\mathbf{0}_V, \boldsymbol{\Sigma}), \quad i = 1, \ldots, N,
\end{aligned} \tag{1}
$$

where $\mathbf{r}(v)$= $(r_1(v), \ldots, r_N(v))'$ and $\boldsymbol{\epsilon}(v) = (\epsilon_1(v), \ldots, \epsilon_N(v))'$ are assumed to be independent of each other. Please refer to Figure 1 for illustrations of the model specification. Here, $\mathbf{X}$ is an $N \times p$ nuisance covariates matrix (e.g., age): each row of the matrix $(\mathbf{x}_i')$ is a p-dimensional nuisance covariates vector for the ith image. Under this model, $\mathbf{r}(v)$ characterizes the dependence among all N images at vertex v (e.g., vertex-level test-retest reliability or heritability) through $\mathbf{K}$. We note that the dependence of images at vertex v is captured through $\theta^2(v)$. Without any true reliable (or heritable) features, we have $\theta^2(v) = 0$ for all v, and images are considered independent.

In this model, $\boldsymbol{\Sigma}$, which is the covariance of the ith residual image $\boldsymbol{\epsilon}_i = (\epsilon_i(1), \ldots, \epsilon_i(V))'$, characterizes both spatial and non-spatial variations within images. We use $\mathbf{b}_i = (b_i(1), \ldots, b_i(V))'$ to characterize the variation due to spatial autocorrelation, whose covariance is modeled by $\sigma^2\boldsymbol{\Phi}(\phi, \mathcal{D})$ with a scale parameter σ^2 and a predefined spatial autocorrelation function (SACF) with a parameter $\phi > 0$ and a $V \times V$ pairwise geodesic distance matrix $\mathcal{D}$. We also use $\boldsymbol{\delta}_i = (z_i(1), \ldots, z_i(V))'$ to characterize the non-spatial variations, or white noise, with its covariance $\tau^2\mathbf{I}_V$. The assumption that $\mathbf{b}_i$ and $\boldsymbol{\delta}_i$ are independent leads to the $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Phi}(\phi, \mathcal{D}) + \tau^2\mathbf{I}_V$, which follows $\boldsymbol{\epsilon}_i = \mathbf{b}_i + \boldsymbol{\delta}_i$ with

$$
\begin{aligned}
&\mathbf{b}_i = (b_i(1), \ldots, b_i(V))' \overset{i.i.d}{\sim} \mathcal{MVN}(\mathbf{0}_V, \sigma^2 \cdot \boldsymbol{\Phi}(\phi, \mathcal{D})), \quad i = 1, \ldots, N, \\
&\boldsymbol{\delta}_i = (z_i(1), \ldots, z_i(V))' \overset{i.i.d}{\sim} \mathcal{MVN}(\mathbf{0}_V, \tau^2 \cdot \mathbf{I}_V), \quad i = 1, \ldots, N.
\end{aligned}
$$

Following exploratory data analyses of Park and Fiecas [14] and Weinstein et al. [15], we assume the exponential spatial autocorrelation function (SACF) for modeling the contrast of parameter estimates in task-fMRI. Thus, the (v, v^*)th element of $\boldsymbol{\Phi}(\phi, \mathcal{D})$ can be expressed by $\exp(-\phi \cdot d_{v,v^*})$ where d_{v,v^*} is the geodesic distance between vertices v and v^*.

From the model specification, the null hypothesis of interest is

$$
H_0 : \theta^2(v) = 0 \quad \text{for all} \quad v = 1, \ldots, V, \tag{2}
$$

which implies that there is no between-image dependencies. The model shows flexibility in specifying reliability and heritability structures through the specification of the dependence structure $\mathbf{K}$. In testing for test-retest reliability, we specify the (i, i^*)th element of $\mathbf{K}$ with 1 if images i and i^* are from the same subject and 0 otherwise. In testing narrow-sense heritability in twin studies, the (i, i^*)th element of $\mathbf{K}$ with 1 if the subjects i and i^* are monozygotic twins, 0.5 if they are dizygotic twins, 0 otherwise. In larger pedigree studies, $\mathbf{K}$ becomes a genetic relationship matrix.

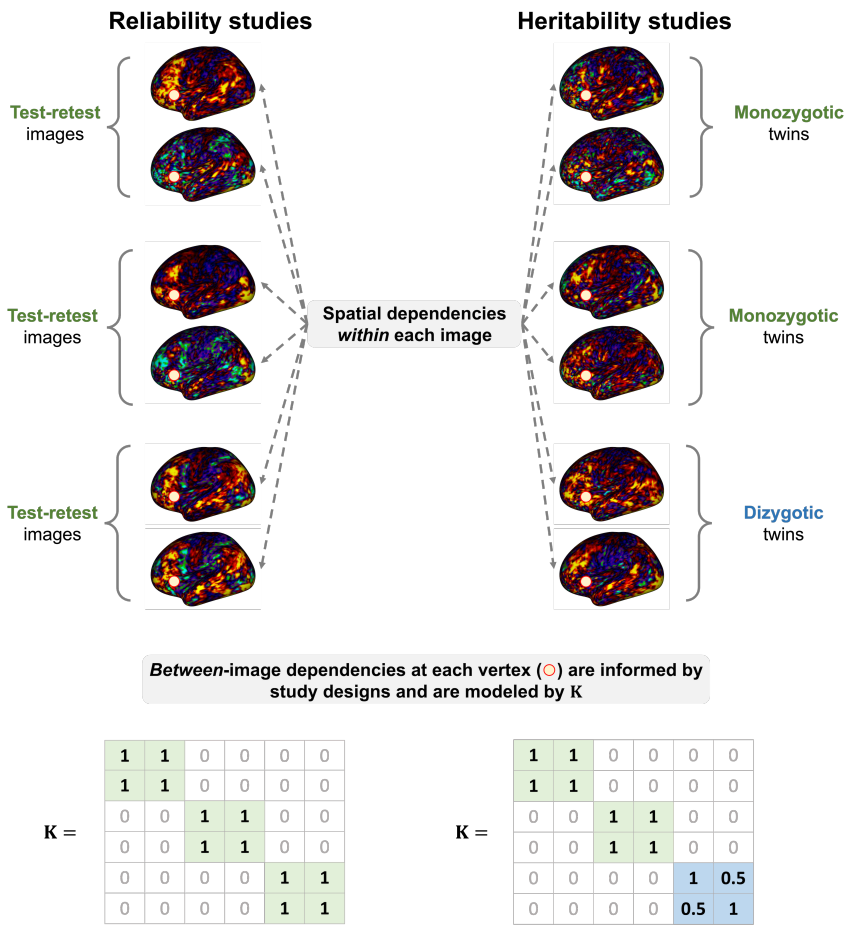

Figure 1: Visualization of two sources of dependencies in reliability and heritability studies. Our model specification captures the spatial dependencies within each image (top) as well as between-image dependencies at each vertex informed by study designs, which are modeled by $\mathbf{K}$.

2.2. Existing approaches

Existing approaches in testing for variance components are mostly massive univariate analyses where spatial dependencies of the imaging features are not considered. Specifically, the $\mathbf{b}(v)$ terms (spatial variations) are dropped, and the model reduces to a massive vertex-wise model as

$$\mathbf{y}(v) = \mathbf{X}\boldsymbol{\beta}(v) + \mathbf{r}(v) + \boldsymbol{\delta}(v), \tag{3}$$

where the null hypothesis remains the same as (2). Note that, from this vertex-level model, narrow-sense heritability is defined by variance due to additive genetic effects divided by the total variance, and test-retest reliability is defined by intraclass correlation coefficient (ICC), both of which giving the same parametrization as $\theta^2(v)/(\theta^2(v) + \tau^2(v))$ based on study designs.

Ganjgahi et al. [18] used the massive univariate method to carry out voxel-wise heritability inference and estimation with the likelihood ratio test (LRT), Wald test, and score test. Their study results showed that the most efficient method, the score test, produced more conservative inferences. They argued that it is likely due to untenable 50:50 χ_1^2 mixture approximation even if all of the tests have a common asymptotic result following a 50:50 mixture of χ_1^2 distribution and χ_0^2 (i.e., point mass at 0).

Ge et al. [19] proposed massively expedited genome-wide heritability analysis (MEGHA) by fitting the model (3) to all vertices with brain structure measures. They used a scaled variance-component score test statistic. The variance-component score test statistic is also known as the sequence kernel association test (SKAT) statistic with a linear kernel in genome-wide association studies (GWASs) [20]. SKAT is a computationally efficient method to test the association between genetic variants since the computation of the test statistics only requires parameter estimates obtained under the null model. The statistic has the form

$$U(v) \propto (\mathbf{y}(v) - \mathbf{X}\hat{\boldsymbol{\beta}}(v))'\mathbf{K}(\mathbf{y}(v) - \mathbf{X}\hat{\boldsymbol{\beta}}(v)) \quad \text{for} \quad v = 1, \ldots, V, \tag{4}$$

where $\hat{\boldsymbol{\beta}}(v)$ is estimated under the null hypothesis (2). Wu et al. [20] showed that the SKAT statistic follows a mixture chi-square distribution under H_0, where the weights of the mixture chi-squares are determined by the eigenvalues of $\mathbf{QKQ}'$ and $\mathbf{Q}$ is a Cholesky decomposition of $\mathbf{P} = \hat{\tau}^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ that satisfies $\mathbf{P} = \mathbf{QQ}'$. Wu et al. [20] also showed that it is the locally most powerful test compared to the log-likelihood ratio test.

2.3. Proposed method: CLEAN-V ('CLEAN' for testing 'V'ariance components)

From the spatial model specified in Equation (1), the score-based variance component test statistic for vertex v is

$$U(v) = (\mathbf{y}(v) - \mathbf{X}\hat{\boldsymbol{\beta}}(v) - \hat{\mathbf{b}}(v))'\mathbf{K}(\mathbf{y}(v) - \mathbf{X}\hat{\boldsymbol{\beta}}(v) - \hat{\mathbf{b}}(v)) \quad \text{for} \quad v = 1, \ldots, V, \tag{5}$$

with $\hat{\boldsymbol{\beta}}(v)$ and $\hat{\mathbf{b}}(v)$ estimated under H_0. Importantly, fitting the null model without any $\theta^2(v)$s is equivalent to fitting CLEAN for GLM [14] because dependencies of images are modeled by $\mathbf{r}(v)$ only. Therefore, we first use ordinary least squares to estimate $\boldsymbol{\beta}(v)$ for each v separately and use residual images $\hat{\boldsymbol{\epsilon}}_i$ to obtain consistent estimates σ^2, ϕ and τ^2 via covariance regression analysis [21, 14] by minimizing

$$\sum_{i=1}^{N} \|\hat{\boldsymbol{\epsilon}}_i\hat{\boldsymbol{\epsilon}}_i' - \sigma^2\boldsymbol{\Phi}(\phi, \mathcal{D}) - \tau^2\mathbf{I}_V\|_F^2, \tag{6}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm of a matrix (Appendix A of Park and Fiecas [14]). Then, we use the parameter estimates to construct $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2\boldsymbol{\Phi}(\hat{\phi}, \mathcal{D}) + \hat{\tau}^2\mathbf{I}_V$ and obtain $\hat{\mathbf{b}}_i =$

$(\hat{b}_i(1), \ldots, \hat{b}_i(V))' = \widehat{E}[\mathbf{b}_i|\mathbf{y}_i]$ under the null model. It can also be shown that $U(v)$ is also a score-based variance component test statistic under the working (spatial) covariance matrix, which implies that $U(v)$ from Equation (5) also follows a mixture chi-square distribution (Supplementary material A).

We construct CLEAN-type cluster-enhanced test statistics in CLEAN-V for its powerfulness and easy implementation in the mesh surface. Compared to other types of cluster enhancement (e.g., TFCE [8]), the CLEAN-type cluster enhancement is constructed from aggregating test statistics from the local neighborhood [14]. To construct cluster-enhanced test statistics using $U(v)$ in Equation (5), we first define $N_r(v)$ as the collection of all vertices within the radius r neighborhood of a central vertex v (i.e., $\{v^* : d_{v,v^*} \leq r\}$). Then, we use all the $U(v^*)$ from vertices v^* within $N_r(v)$ to construct a cluster-enhanced test statistic for the vertex v. For a radius r, we compute the standardized sum of the variance-component score test statistics within $N_r(v)$:

$$T_r(v) = \frac{\sum_{v^* \in N_r(v)} U(v^*) - \widehat{\mathrm{E}}_{H_0}\left(\sum_{v^* \in N_r(v)} U(v^*)\right)}{\sqrt{\widehat{\mathrm{Var}}_{H_0}\left(\sum_{v^* \in N_r(v)} U(v^*)\right)}}. \tag{7}$$

In practice, we obtain the null distribution of $\sum_{v^* \in N_r(v)} U(v^*)$ by permutation and use empirical permuted test statistics to obtain $\widehat{\mathrm{E}}_{H_0}\left(\sum_{v^* \in N_r(v)} U(v^*)\right)$ and $\widehat{\mathrm{Var}}_{H_0}\left(\sum_{v^* \in N_r(v)} U(v^*)\right)$ by the sample mean and sample variance (Section 2.4). Similar to how we used the quadratic form to show that $U(v^*)$ follows a mixture of chi-square distribution, $\sum_{v^* \in N_r(v)} U(v^*)$ also follows a mixture of chi-squares distribution under the null hypothesis (Supplementary material A). A mixture chi-square distribution can be approximated by normal distribution according to the Lyapunov central limit theorem (Supplementary material A). Therefore, after standardization in (7), $T_r(v)$ approximately follows the standard normal distribution for each r and v. This transformation makes sure that each test statistic is mapped onto the same standard normal scale, which allows for data-adaptive testing based on the true areas of signals. Figure 2 is given for illustration.

Statistically, the way CLEAN-V achieves high power aligns with CLEAN [14] developed for GLM. In CLEAN, $T_r(v)$ is defined by assuming that the regression coefficients of interest in $N_r(v)$ are the same [14]. Similarly, it can be shown that $T_r(v)$ in CLEAN-V is constructed by assuming that $\theta^2(v^*)$s in $N_r(v)$ are the same (Supplementary material B). Since the SKAT statistics are known to be locally most powerful, combining test statistics within a neighborhood strengthens the evidence against the null hypothesis when the strength of signals remains consistent within the neighbor. This property, along with Supplementary material B, makes the proposed CLEAN-type enhancement particularly powerful in CLEAN-V. Li et al. [22] used a similar approach, called Quadratic Scan (Q-SCAN) statistic, to identify the sizes and locations of signal regions in genome-wide association studies.

In CLEAN-V, we build a map of test statistics by adopting an adaptive method to benefit from

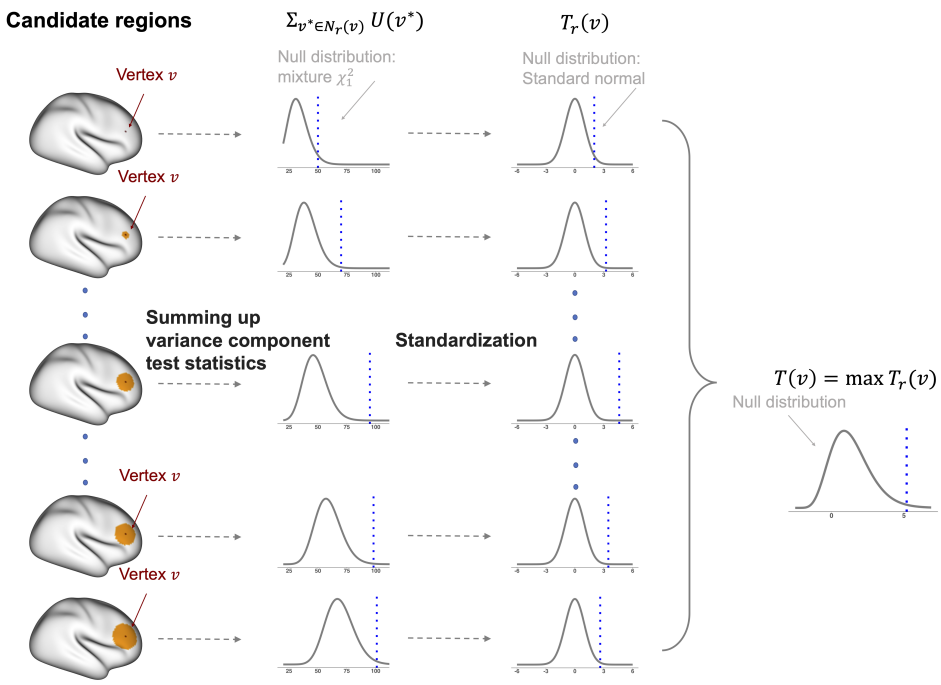

Figure 2: The process to construct clusterwise-enhanced statistics in CLEAN-V.

the 'ideal' local neighbor $N_r(v)$ for each v. Since the best size of the neighborhood area for each vertex might not be the same, we don't fix the radius r. Instead, we consider different clusters by choosing different radii rs from r_1 up to $r_{\max}$ and choose the maximum $T_r(v)$ as the adaptive cluster-enhanced test statistic for vertex v:

$$T(v) = \max \{T_r(v) | r = r_0, r_1, \ldots, r_{\max}\}. \tag{8}$$

In this way, we can find the closely optimal cluster size for each vertex, which is useful as true areas of signals are unknown *a priori*. Following Park and Fiecas [14] and Weinstein et al. [15], we consider $r_1 = 1\text{mm}, r_2 = 2\text{mm}, \ldots, r_{\max} = 20\text{mm}$ as a default in this paper, although $r_{\max}$ can be chosen flexibly based on research questions and contexts. The radius $r_0 = 0\text{mm}$ is included in the radii set since it guarantees that original vertex-level test statistics can also be considered in the maximization process.

2.4. FWER control with permutation

A critical value t_α is needed to control the family-wise error rate (FWER) at level α. We consider the test statistic for H_0 as

$$T = \max\{T(v) : v = 1, \ldots, V\}. \tag{9}$$

Permutation allows us to generate an empirical distribution of test statistics under the null hypothesis. In our setting, it can be achieved by shuffling $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^N$ while keeping the dependence matrix $\mathbf{K}$ unchanged. Note that this is equivalent to shuffling $\hat{\boldsymbol{\delta}}_i$ from the null model across images, which in practice does not require any refitting of the mean model and the covariance model (Equation (6)). For the bth permutation ($b = 1, \ldots, B$), we construct the permuted variance-component test statistics $U^{(b)}(v)$. Then, we use Equation (7) to construct the permuted single cluster-enhanced test statistic $T_r^{(b)}(v)$. $T^{(b)}(v)$ and $T^{(b)}$ follow Equations (8) and (9). Finally, we construct the empirical null distribution by the set $\{T^{(1)}, \ldots, T^{(B)}\}$ and select the $(1-\alpha)$th quantile as t_α. Statistical significance is achieved at $\alpha \times 100\%$ FWER when $T > t_\alpha$.

2.5. Localizing areas of significance

The set of vertices v with statistically significant $\theta^2(v) > 0$ is identified by vertices whose $T(v)$ exceeds t_α.

2.6. Notes on the computational efficiency

Considering the complicated covariance structures implied in Equation (1), CLEAN-V reduces the computational cost dramatically because CLEAN-V uses score-based test statistics that require the estimation σ^2, τ^2, ϕ only. Also, the permutation does not even require estimating σ^2, τ^2, ϕ repeatedly but matrix multiplications from the null model only. We note that other methods, such as likelihood-ratio test or Wald test used by Ganjgahi et al. [18] would require the estimation of $\theta^2(1), \ldots, \theta^2(V), \sigma^2, \tau^2, \phi$ in the original data and each permutation when extended to the spatial autocorrelation modeling, which would be computationally infeasible. In addition, in CLEAN-V, computational costs for obtaining σ^2, τ^2, ϕ under the null model are relaxed by using a method-of-moment approach (Equation (6),[21]), and the calculation of $\hat{\boldsymbol{\Sigma}}^{-1}$ for a large V is relaxed by using the nearest neighbor Gaussian process approximation [23].

3. Data analysis and data-driven simulations

3.1. Data preparation

We collected task-fMRI data from the Human Connectome Project (1200 Subjects Data Release). We focus on 5 experimental tasks: relational processing, emotional processing, social cognition, language processing, and gambling decision-making. Specific contrasts we used for each task are summarized in Table 1. Barch et al. [24] provide a more detailed description of each task. There are more than 1000 subjects in total who completed the tasks and, among these, 44 subjects who completed tasks twice. Additionally, there are 171 twin pairs ($N = 342$ images), of which 68 pairs are dizygotic twins and 103 pairs are monozygotic twins. There are 80 and 116 female twins correspondingly. The average ages for dizygotic and monozygotic twins are 29.16 (SD: 3.42) and 29.29 (SD: 3.37) years, respectively. The task-fMRI data were processed through the HCP's

minimal surface processing pipeline with 2mm surface smoothing [25]. We used the R package ciftiTools [26] to sample 10242 vertices from approximately 32000 vertices resulting in 9394 and 9398 cortex vertices in the left and right hemispheres, respectively, after excluding 848 and 844 vertices in the medial wall in each hemisphere.

Task	**Contrast**
Emotional processing	face vs. shape
Social cognition	social vs. random
Relational processing	relational vs. match
Language	story vs. math
Gambling	reward vs. punishment

Table 1: The list of selected contrasts for each task used in our analysis.

3.2. Competitors

In our data analysis and simulations, we compared CLEAN-V to three competitors, all of which are considered as simplified versions of CLEAN-V. Specifically,

1. **CLEAN-V without spatial correlation**: It constructs cluster-enhanced test statistics $T(v)$ based on the model without considering spatial autocorrelations (Equation (3)). For a fair comparison, we used the same neighbor set as CLEAN-V to construct cluster-enhanced test statistics.
2. **CLEAN-V without cluster enhancement**: It models spatial autocorrelation of the noise following Equation (1), but it does not construct cluster-enhanced statistic. It is equivalent to setting $r = 0$ only for $T_r(v)$ (i.e., the vertex itself).
3. **Massive-univariate analysis**: This method neither models spatial autocorrelation nor constructs cluster-enhanced test statistics, which is equivalent to MEGHA [19].

All these competitors used the same permutation step outlined in Section 2.4. We included *CLEAN-V without spatial correation* and *CLEAN-V without cluster enhancement* to examine the effect of modeling the spatial structure of features and the impact of the novel clusterwise enhancement method, separately.

3.3. Data analysis

3.3.1. Test-retest reliability and heritability for five different cognition tasks

We first fitted CLEAN-V to (i) 44 subjects with test-retest images to obtain the test-retest reliability map and (ii) 171 twin pairs to obtain the heritability map, with FWER controlled at 5%. These maps are shown in Figure 3(a). CLEAN-V detected wide areas of statistically significant test-retest reliability on the brain surface for most of the task-fMRI. However, we obtained different areas of significance for different tasks. The language task showed strong evidence of test-retest

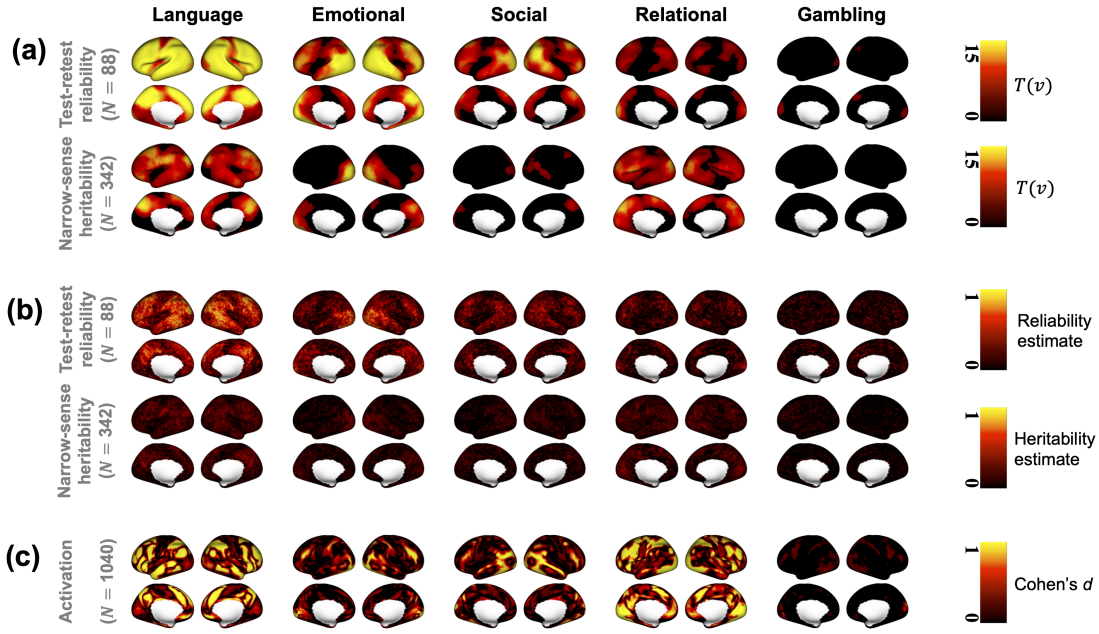

Figure 3: (a) CLEAN-V's localization results ($T(v)$s) for the five cognition tasks. For visualizations, $T(v)$s were shrunk to zero if they were less than the threshold $t_{0.05}$. (b) The effect size estimates of the test-retest reliability and narrow-sense heritability $[\hat{\theta}^2(v)/(\hat{\theta}^2(v)+\hat{\tau}^2(v))]$, estimated based on the vertex-level (non-spatial) model in Equation (3). (c) The effect size (Cohen's d) maps for the group-level activation, which is thresholded at 0.2 for visualizations.

reliability, but the gambling task showed tiny and sparse test-retest reliability. However, the five tasks displayed a common result: anterior central gyrus and postcentral gyrus showed low or no test-retest reliability.

We also provided estimates of test-retest reliability and narrow-sense heritability (Figure 3(b)) for comparison. Their effect size computed by the massive univariate method was small and specific to tasks, as expected. However, CLEAN-V detected test-retest reliability and narrow-sense heritability with high significance even if the estimates were small, and the detected regions corresponded well to the effect size estimates. Also, By qualitatively comparing the test-retest reliability maps to the activation maps (Cohen's d; Figure 3(c)), we found that there is a strong correspondence between the strength of reliability and activation [27]. Specifically, the language tasks showed a strong positive and negative activation map which agrees with our results: the strongest evidence of test-retest reliability came from the language task. The weak evidence of activations in gambling tasks also agrees with the test-retest reliability map of CLEAN-V.

3.3.2. Localization results from CLEAN-V and its competitors

Figure 4 shows the CLEAN-V and the competitors' binarized localization results of both test-retest reliability and narrow-sense heritability for the emotional processing task. CLEAN-V without cluster enhancement produced almost the same result as the massive univariate analysis, where

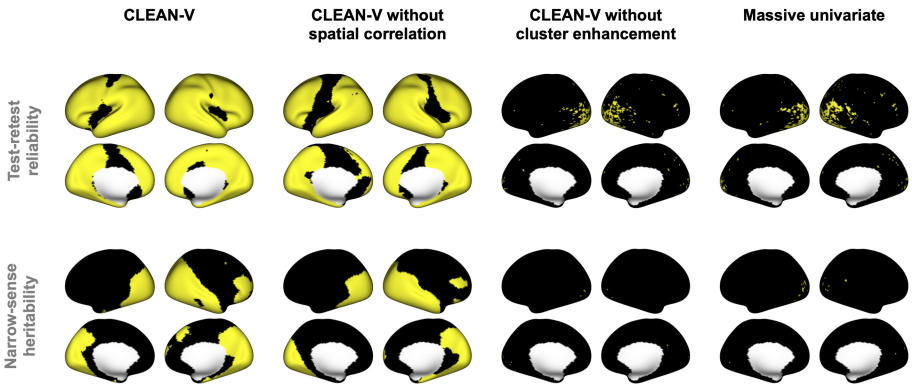

Figure 4: Localization results from CLEAN-V and its competitors for emotional task are displayed on the inflated surface. The yellow vertices denote for statistically significant vertices detected by each method.

very limited number of vertices were detected. Compared to these two methods, CLEAN-V without spatial correlation detected obviously more regions, which shows using cluster-enhanced test statistics improved signal detection. CLEAN-V detected most of the areas for the reliability and heritability, which validates the benefits of cluster-enhanced tests statistics and spatial autocorrelation modeling. The analysis results for other tasks are provided in Figures S1 and S2 in the supplementary material C, and the performances were analogous to the emotional task which supported the most sensitive performance of CLEAN-V and the least sensitive performance of massive univariate analysis.

3.3.3. Comparisons to massive univariate analysis with smoothed images

In this section, we also conducted massive univariate analyses (MEGHA) with different levels of surface smoothing (e.g., 5mm, 8mm, 10mm). We compared them to CLEAN-V applied to the original data but with $r_{\max}$ = 0mm, 5mm, 8mm, 10mm, although we acknowledge that a direct comparison between 'smoothing radius' in massive univariate analysis and $r_{\max}$ in CLEAN-V is not feasible. From Figure 5, massive univariate analysis applied to presmoothed data improved the number of detected vertices. For CLEAN-V, it also consistently increased the detected regions as $r_{\max}$ increased. At the same time, it suggests that even a small $r_{\max}$ (e.g., 5mm) seems to outperform massive univariate analyses with different smoothing levels. These results demonstrate the relative efficiency of CLEAN-V.

A cautionary note is warranted. Both imaging smoothing and cluster enhancement involve a tradeoff between sensitivity and specificity. Therefore, just as choosing the optimal smoothing level is an open question, choosing an optimal $r_{\max}$ in CLEAN-V, similar to CLEAN and CLEAN-R, is an open question which should be determined carefully based on expected signal-to-noise ratios, sample sizes, and research purposes.

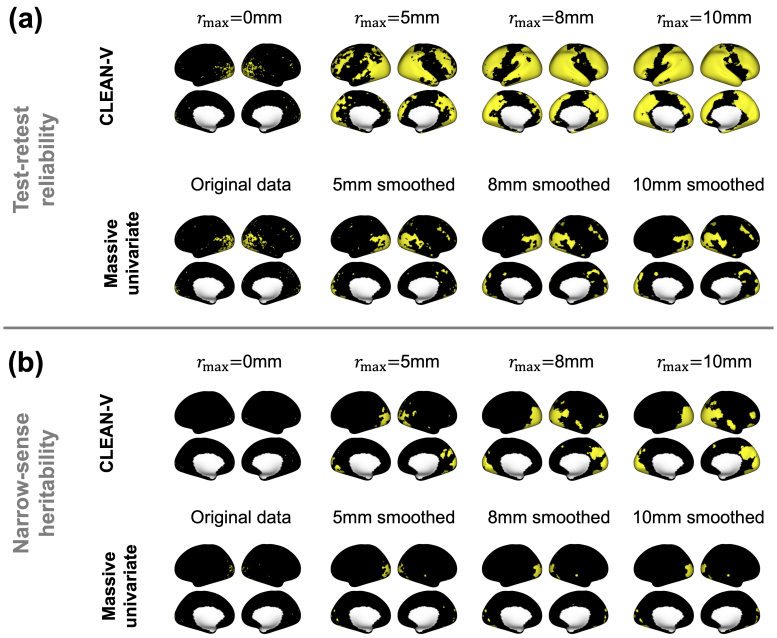

Figure 5: Binarized localization results from CLEAN-V and massive univariate analysis for emotional task. CLEAN-V models the original data but with different maximum radius $r_{\max}$ for cluster-wise enhancement process. MEGHA models smoothed data under different levels of surface smoothing.

.

3.4. Data-driven simulations

3.4.1. Setup

Motivated by existing efforts in neuroimaging literature [14, 15, 28, 29], we used task-fMRI data from HCP to design data-driven simulations to evaluate the performance of CLEAN-V compared to the competitors. It aligns with the approach of Eklund et al. [30], which used real data (resting-state fMRI) to evaluate false positive rates of different statistical methods for task-fMRI. Also, this approach is particularly useful in evaluating the robustness of CLEAN-V under potential model misspecifications (in covariance structures) because it assumes stationarity and isotropy of underlying spatial autocorrelation with the exponential SACF, which may not hold. For our data-driven simulations, we used images from the emotional task to construct 'null' data and 'signal' data with different signal-to-noise ratios (SNRs), which are described below.

3.4.2. Simulating null data

The null data consists of independent images only. Among all subjects in HCP, we only considered test images from non-twin subjects to construct 'pseudo' test-retest pairs or twin pairs. For example, a set of test-retest images for a subject was generated by randomly sampling two images from test images from different independent (non-twin) subjects. Since there are no true test-retest pairs or twins in the null data, we expect no test-retest reliability or narrow-sense heritability.

3.4.3. Simulating test-retest data with different signal-to-noise ratios

We fixed the number of subjects to be 20 ($N = 40$ images) throughout the simulation studies for test-retest reliability, and we changed the proportion of subjects whose test and retest images are sampled (without replacement) from the 44 subjects available from HCP. The remaining subjects' images were sampled from test images from different subjects from HCP. We expect stronger SNR when more test-retest pairs are included. These evaluations are shown in Section 3.5.1.

3.4.4. Simulating twins data with different signal-to-noise ratios

The simulation process for twins data is the same as the procedure for test-retest data except for the larger sample size. We fixed the number of twin pairs to be 60 ($N = 120$ images) as the overall effect size of narrow-sense heritability is smaller than test-retest reliability. We changed the proportion of true twins whose images were sampled from the 171 twins. The rest of the image pairs were sampled from different subjects. These evaluations are also shown in Section 3.5.1.

3.5. Data-driven simulation results

In each scenario, we generated 1000 sets of simulated data to evaluate performances.

3.5.1. Power analysis and family-wise error rate

Figure 6(a) shows the empirical FWER of the CLEAN-V, which demonstrates that CLEAN-V and its competitors controlled FWER accurately at 0.05 under different maximum radii $r_{\max}$ of cluster enhancement for testing both test-retest reliability and narrow-sense heritability.

Figure 6(b) shows the power of the four competitors. The empirical power is calculated by the number of times rejecting the null hypothesis divided by the number of simulations (1000). We see that CLEAN-V's power was uniformly higher than the competitors, and the power of CLEAN-V without spatial correlation was substantially higher than the massive univariate analysis. For test-retest reliability, however, the power of CLEAN-V without cluster enhancement showed a similar power as massive univariate analysis and even lower when the signal-to-noise ratio was high. Notably, CLEAN-V achieved strong power (80%) when the number of test-retest pairs was only 3. As the number of test-retest pairs increased to 5, almost 100% power was obtained by CLEAN-V. For narrow-sense reliability, CLEAN-V also achieved the highest power. Specifically, it reached 80% power when the number of twin pairs was 35. And it attained 100% power as the number of twin pairs increased to 40.

3.5.2. Reproducibility

We computed vertex-level reproducibility from the simulation results. Specifically, for each vertex, we calculated the proportion of the vertex identified out of 1000 simulations. Figure 7(a) shows the reproducibility maps of three settings for both test-retest reliability and narrow-sense heritability. Clearly, CLEAN-V outperformed its competitors under all settings. CLEAN-V and

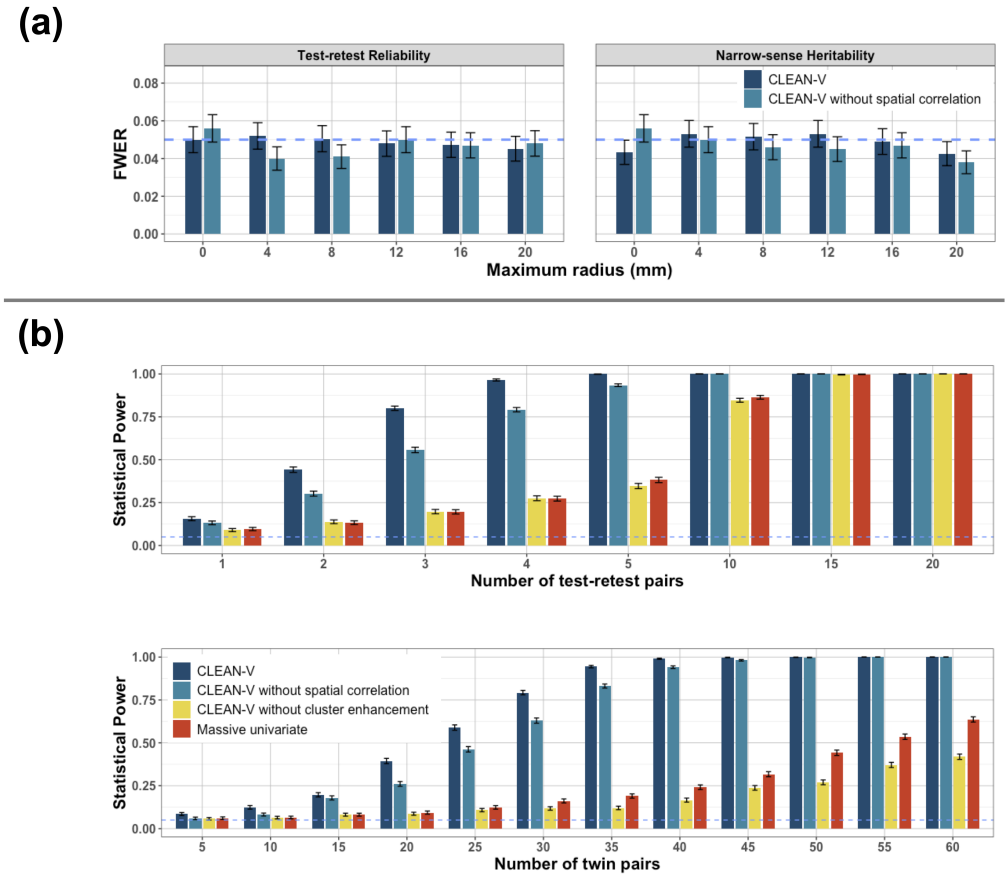

Figure 6: (a) Empirical FWER of CLEAN-V and its competitors under different maximum radii $r_{\max}$s for clusterwise enhancement from the data-driven simulation studies under the null hypothesis. Note that since massive univariate analysis does not involve the clusterwise procedure, massive univariate analysis is the same as CLEAN-V without spatial correlation with 0mm maximum radius. Similarly, CLEAN-V without cluster enhancement is equivalent to CLEAN-V with 0mm maximum radius. (b) Empirical power of CLEAN-V and its competitors under different signal-to-noise ratios. The error bars denote standard errors.

CLEAN-V without spatial correlation identified substantial signal regions for test-retest reliability when sufficient test-retest pairs and twins were included. CLEAN-V was better than CLEAN-V without spatial correlation in all cases. Regardless of the number of test-retest pairs and twin pairs included in the samples of simulations, Massive univariate and CLEAN-V without cluster enhancement nearly cannot detect the signals of test-retest reliability and heritability. Because the signal region of heritability for the emotional task is relatively small, both CLEAN-V and CLEAN-V without spatial correlation identified small signal regions. However, CLEAN-V still identified more signal regions with higher reproducibility.

Figure 7(b) included the proportion of vertices whose reproducibility is over 80% or 50% for each approach. CLEAN-V outperformed in all scenarios. When there were 20 test-retest pairs in the sample for test-retest reliability analysis (all test-retest data), CLEAN-V remarkably produced

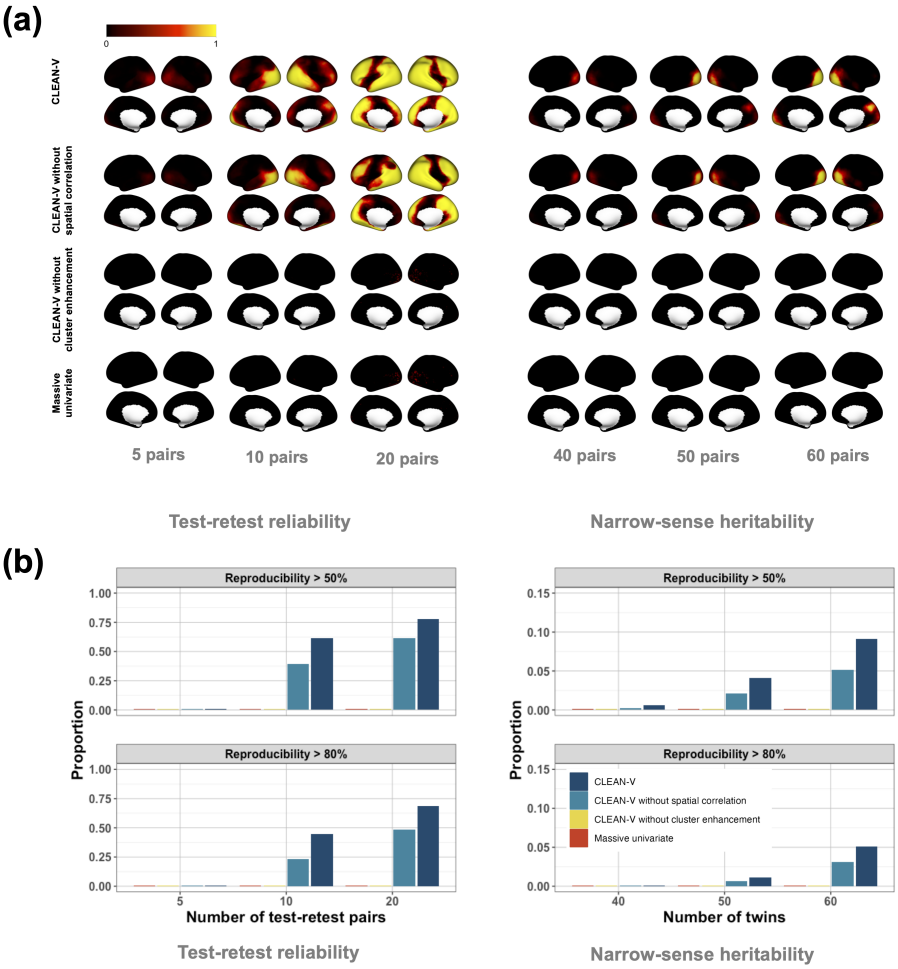

Figure 7: Reproducibility analysis from the data-driven simulation studies. The left and right figures describe the results for test-retest reliability and narrow-sense heritability. (a) The value of each vertex is the proportion of rejecting the null hypothesis among the 1000 simulated datasets. (b) Proportions of vertices with over 80% and 50% reproducibility for each method.

68% and 77% vertices over 80% and 50% reproducibility correspondingly, which were better than CLEAN-V without spatial correlation (48% and 61%) and considerably higher than CLEAN-V without cluster enhancement and Massive univariate (0% and 0.1% for both). For narrow-sense heritability analysis with 60 pairs of twins, 9% and 5% vertices had over 50% and 80% reproducibility from CLEAN-V, which is also notably higher than the proportion of vertices with CLEAN-V without spatial correlation (5% and 3%) and the other two methods (0% for both). For other settings, CLEAN-V also significantly outperformed the others.

4. Discussion

In this paper, we proposed a new method, CLEAN-V, for statistical inference and localization of test-retest reliability and narrow-sense heritability for task-based fMRI. Our method directly adjusts the global spatial autocorrelation, constructs spatially-enhanced test statistics in a locally powerful way, and controls FWER efficiently by permutations. We showed that CLEAN-V controlled family-wise error rate and achieved high power simultaneously.

Our results showed that it is necessary to use both cluster-wise enhancement and spatial autocorrelation modeling to increase sensitivity in testing variance components, which aligns with previous findings [14, 15]. It was supported by comparing CLEAN-V to simpler methods: massive univariate analysis, CLEAN-V without spatial correlation, and CLEAN-V without cluster enhancement. The comparison results showed that there is no power improvement if only spatial dependence is considered (without cluster enhancement). Our cluster-wise enhancement method with univariate statistics led to increased statistical power by flexibly collecting neighbors' information and adaptation. With a combination of the two modeling strategies (clusterwise enhancement and spatial autocorrelation modeling), CLEAN-V sufficiently detected test-retest reliability even if the proportion of test-retest data in samples was very small (15%) and achieved almost perfect power when the proportion increased to 25%. Considering that the number of images used in studies for test-retest reliability or heritability is mostly small ($N < 100$), CLEAN-V provides a powerful and sensitive approach that partially relaxes such concerns.

In the analysis of 5 experimental tasks of the HCP task-fMRI data, CLEAN-V, in general, distinguished the strength of test-retest reliability and narrow-sense heritability on the brain surface and among tasks. We also found a strong relationship between test-retest reliability and activation by comparing the CLEAN-V's test-retest maps to activation maps. Thus, CLEAN-V offers possibilities to discover the correspondence between reliability regions and activation regions.

Combining the score-based variance component test and permutations, CLEAN-V is much more computationally efficient than the classical likelihood-based tests since constructing the test statistics and finding FWER-controlled threshold only need fitting the null model once instead of refitting numerous models. The computational efficiency of CLEAN-V speaks of its practical utility.

This paper focused primarily on analyzing task-fMRI data. Still, CLEAN-V is widely applicable to many neuroimaging data types (e.g. cortical thickness) where 'borrowing' information across spatial domains is expected to improve power. Also, a comprehensive exploratory analysis by Weinstein et al. [15] reveals that applying exponential spatial autocorrelation function (SACF) is reasonable in many imaging modalities mapped onto the cortical surface. While applying the same assumption to volumetric (3D) imaging data could be questioned, we showed that a simplified version of CLEAN ('CLEAN-V without spatial correlation') also results in a dramatic increase in power and sensitivity compared to massive univariate analysis, which can be adopted easily.

Lastly, CLEAN-V provides useful insight into recent concerns on low test-retest reliability or heritability in neuroimaging studies. Although CLEAN-V does not provide effect size estimates, the fact that the adoption of 'spatial variations' resulted in increased power suggests that the current definition of test-retest reliability or heritability in neuroimaging studies might be too simple to characterize all sources of variations, such as spatial variations, which would provide limited understanding of variance components and lead to suboptimal estimations. For example, Risk and Zhu [6] showed how spatial modeling of heritability provides better estimates than mass univariate analysis. Extending the 'inference' problem of CLEAN-V to the 'estimation' problem is worth further investigations, which we leave as future work.

5. Software

Without parallel computing, implementing CLEAN-V for 342 images (each image containing 10,000 vertices) with 5,000 permutations took less than 10 minutes on a laptop, which supports the practical utility of the method. We are working on additional optimization of the software with parallel computing, which is publicly available at `https://github.com/junjypark/CLEAN`.

Declaration of Competing Interest

None.

Acknowledgement

We would like to thank reviewers for their helpful comments and suggestions. EWD receives funding from the Brain and Behavior Research Foundation Young Investigator Award (NARSAD), the Canadian Institutes of Health Research. CH receives research funding from the National Institute of Mental Health (NIMH), the Brain and Behavior Research Foundation, the University of Toronto, and the Centre for Addiction and Mental Health (CAMH) Foundation. NR receives funding from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-05897). ANV receives funding from the National Institute of Mental Health (1/3R01MH102324 & 1/5R01MH114970), Canadian Institutes of Health Research, Canada Foundation for Innovation, CAMH Foundation, and University of Toronto. JYP receives funding from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2022-04831), and the University of Toronto's Data Science Institute, McLaughlin Centre, and Connaught Fund.

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. This research was supported in part by

NIH grants R01EB016061 and R01EB026549 from the National Institute of Biomedical Imaging and Bioengineering and R01 MH116026 from the National Institute of Mental Health.

Supplementary materials

A. Normal approximation of mixture chi-square distribution

Combing A.1 and A.2, the null distribution of $\sum_{v^* \in N_r(v)} U(v^*)$ is approximated by the normal distribution when N is large. This implies that $T_r(v)$ (the standardized $\sum_{v^* \in N_r(v)} U(v^*)$) follows the standard normal distribution approximately under H_0.

A.1 Proof that $\sum_{v^* \in N_r(v)} U(v^*)$ follows mixture chi-square under H_0

Based on our model specification, we define $\mathbf{y} = (y_1(1), y_2(1), \ldots, y_N(1), \ldots, y_1(V), y_2(V), \ldots, y_N(V))'$, which leads to

$$\mathbf{y} \sim \mathcal{MVN}((\mathbf{I}_V \otimes \mathbf{X})\boldsymbol{\beta}, \boldsymbol{\Sigma} \otimes \mathbf{I}_N)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}(1)', \ldots, \boldsymbol{\beta}(V)')'$ and $\otimes$ denotes for the Kronecker product. Because OLS is used to estimate $\boldsymbol{\beta}(v)$, we have

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - (\mathbf{I}_V \otimes \mathbf{X})\hat{\boldsymbol{\beta}} = \underbrace{[\mathbf{I}_{NV} - (\mathbf{I}_V \otimes \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')]}_{\mathbf{P}}\mathbf{y}.$$

After estimating $\hat{\sigma}^2, \hat{\tau}^2$ and $\hat{\phi}$ by covariance regression analysis, we get $\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N$ where $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2\boldsymbol{\Phi}(\hat{\phi}, \mathcal{D}) + \hat{\tau}^2\mathbf{I}_V$. Then, we can get the $\hat{\mathbf{b}}$ by estimated conditional expectation $\hat{E}(\mathbf{b}|\mathbf{y})$ from the joint distribution of $(\mathbf{b}, \mathbf{y})$:

$$\begin{pmatrix} \mathbf{b} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} \mathbf{0}_{NV} \\ (\mathbf{I}_V \otimes \mathbf{X})\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{NV} & \sigma^2\boldsymbol{\Phi}(\phi, \mathcal{D}) \otimes \mathbf{I}_N \\ \sigma^2\boldsymbol{\Phi}(\phi, \mathcal{D}) \otimes \mathbf{I}_N & \boldsymbol{\Sigma} \otimes \mathbf{I}_N \end{pmatrix}\right),$$

which provides

$$E(\mathbf{b}|\mathbf{y}) = \left(\sigma^2\boldsymbol{\Phi}(\phi, \mathcal{D}) \otimes \mathbf{I}_N\right)(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_N)(\mathbf{y} - (\mathbf{I}_V \otimes \mathbf{X})\boldsymbol{\beta}).$$

Plugging in the estimated spatial parameters, we obtain

$$
\begin{aligned}
&\mathbf{y} - (\mathbf{I}_V \otimes \mathbf{X})\hat{\boldsymbol{\beta}} - \hat{\mathbf{b}} \\
&= \mathbf{y} - (\mathbf{I}_V \otimes \mathbf{X})\hat{\boldsymbol{\beta}} - \left(\hat{\sigma}^2 \boldsymbol{\Phi}(\hat{\phi}, \mathcal{D}) \otimes \mathbf{I}_N\right) (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N) \left(\mathbf{y} - (\mathbf{I}_V \otimes \mathbf{X})\hat{\boldsymbol{\beta}}\right) \\
&= \left[\mathbf{I}_{NV} - \left(\hat{\sigma}^2 \boldsymbol{\Phi}(\hat{\phi}, \mathcal{D}) \otimes \mathbf{I}_N\right) (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\right] \hat{\boldsymbol{\epsilon}} \\
&= \hat{\tau}^2 (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\hat{\boldsymbol{\epsilon}}.
\end{aligned}
$$

To get corresponding test statistics for each vertice v, we define $\mathbf{S}_v : \mathbb{R}^{NV} \to \mathbb{R}^N$ to identify the subset of all images' element for vertice v (e.g., $\mathbf{S}_{\{1\}}\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}(1)$). Then, the test statistic becomes

$$
\begin{aligned}
U(v) &= \hat{\boldsymbol{\epsilon}}^{'} (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\hat{\tau}^2 \mathbf{S}_v^{'} \mathbf{K} \mathbf{S}_v \hat{\tau}^2 (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\hat{\boldsymbol{\epsilon}} \\
&= \mathbf{y}'\mathbf{P} \underbrace{\left[(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)' \hat{\tau}^2 \mathbf{S}_{\{v\}}' \mathbf{K} \mathbf{S}_{\{v\}} \hat{\tau}^2 (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\right]}_{\mathbf{A}} \mathbf{P}\mathbf{y}.
\end{aligned}
$$

Similarly, define $\mathbf{S}_{N_r(v)} : \mathbb{R}^{NV} \to \mathbb{R}^{N \cdot |N_r(v)|}$ to identify the subset of all subjects' elements for vertices in $N_r(v)$ (e.g., $\mathbf{S}_{\{1,2\}}\hat{\boldsymbol{\epsilon}} = (\hat{\boldsymbol{\epsilon}}(1)^{'}, \hat{\boldsymbol{\epsilon}}(2)^{'})^{'}$. Then, for cluster-enhanced test statistic, we have

$$
\sum_{v^* \in N_r(v)} U(v^*) = \mathbf{y}'\mathbf{P} \underbrace{\left[(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)' \hat{\tau}^2 \mathbf{S}_{N_r(v)}' (\mathbf{I}_{|N_r(v)|} \otimes \mathbf{K}) \mathbf{S}_{N_r(v)} \hat{\tau}^2 (\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_N)\right]}_{\mathbf{A}} \mathbf{P}\mathbf{y}. \tag{10}
$$

Since $\mathbf{P}\mathbf{y}$ follows multivariate normal with zero mean vector and the covariance matrix $\mathbf{P}(\boldsymbol{\Sigma} \otimes \mathbf{I}_N)\mathbf{P}$, $\mathbf{y}^{'}\mathbf{P}\mathbf{A}\mathbf{P}\mathbf{y}$ follows a mixture chi-square distribution with weights being the eigenvalues of $\mathbf{Q}\mathbf{A}\mathbf{Q}^{'}$ where $\mathbf{Q}$ corresponds to the Cholesky decomposition of $\mathbf{Q}\mathbf{Q}^{'} = \mathbf{P}(\hat{\boldsymbol{\Sigma}} \otimes \mathbf{I}_N)\mathbf{P}$ [31].

A.2 Proof that the mixture chi-square distribution converges in distribution to a normal distribution

We use the Lyapunov Central Limit Theorem to prove the mixture chi-square distribution converges in distribution to a normal distribution as the number of components in the mixture chi-square goes to infinity. Assume $\{Y_1, \ldots, Y_k\}$ is a sequence of independent random variables, each with finite expected value μ_i and variance σ_i^2, and $s_k^2 = \sum_{i=1}^k \sigma_i^2$. The Lyapunov Central Limit Theorem states that if the following condition is satisfied for the sequence:

$$
\lim_{k \to \infty} \frac{1}{s_k^{2+\delta}} \sum_{i=1}^{k} \mathrm{E}[\mid Y_i - \mu_i \mid^{2+\delta}] = 0 \quad \text{for some } \delta > 0,
$$

then $\sum_{i=1}^k (Y_i - \mu_i)/s_k$ will converge weakly to the standard normal distribution.

Suppose $X_1, \ldots, X_k \overset{i.i.d}{\sim} \chi_1^2$ are the components of the mixture chi-square for our test statistic (10) with their corresponding weights $\lambda_1, \ldots, \lambda_k$ which are the eigenvalues of $\mathbf{Q}\mathbf{A}\mathbf{Q}^{'}$. Then it is sufficient to prove that the sequence of $\lambda_i X_i$ satisfies the Lyapunov's condition. Because the fourth

central moment of χ_1^2 is 48, we can get the fourth central moment of $\lambda_i X_i$: $\mathrm{E}[(\lambda_i X_i - \lambda_i \mu_i)^4] = \lambda_i^4 \mathrm{E}[(X_i - \mu_i)^4] = 48\lambda_i^4$. Then, choosing $\delta = 2$, we get

$$\sum_{i=1}^{k} \mathrm{E}[(\lambda_i X_i - \lambda_i \mu_i)^4] = 48 \sum_{i=1}^{k} \lambda_i^4.$$

Because $\mathrm{Var}(\lambda_i X_i) = \lambda_i^2 \mathrm{Var}(X_i) = 2\lambda_i^2$, we have

$$s_k^4 = \left(\sum_{i=1}^{k} \mathrm{Var}(\lambda_i X_i) \right)^2 = 4 \left(\sum_{i=1}^{k} \lambda_i^2 \right)^2.$$

The result above leads to

$$\frac{1}{s_k^4} \sum_{i=1}^{n} \mathrm{E}[(\lambda_i X_i - \lambda_i \mu_i)^4] = \frac{48 \sum_{i=1}^{k} \lambda_i^4}{4 \left(\sum_{i=1}^{k} \lambda_i^2 \right)^2} \leq \frac{12}{k} \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^4,$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues, respectively. As k goes to infinity,

$$\lim_{k \to \infty} \frac{1}{s_k^4} \sum_{i=1}^{n} \mathrm{E}[(\lambda_i X_i - \lambda_i \mu_i)^4] \leq \lim_{k \to \infty} \frac{12}{k} \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right)^4 = 0,$$

provided that $\lambda_{\max}/\lambda_{\min}$ is bounded. Therefore, the Lyapunov's condition is satisfied and $\sum_{i=1}^{k} (\lambda_i X_i - \lambda_i \mu_i)/s_k$ converges in distribution to a standard normal random variable which is equivalent to $\sum_{i=1}^{k} \lambda_i X_i$ converges weakly to a normal distribution.

B. Proof that $T_r(v)$ is constructed by assuming $\theta^2(v^*)$s in the neighborhood $N_r(v)$ are the same

Let $v_1, \ldots, v_{|N_r(v)|}$ be indices of vertices in $N_r(v)$. From our model assumptions, the model within $N_r(v)$ is

$$\begin{pmatrix} \mathbf{y}(v_1) \\ \mathbf{y}(v_2) \\ \vdots \\ \mathbf{y}(v_{|N_r(v)|}) \end{pmatrix} = \begin{pmatrix} \mathbf{X}\boldsymbol{\beta}(v_1) \\ \mathbf{X}\boldsymbol{\beta}(v_2) \\ \vdots \\ \mathbf{X}\boldsymbol{\beta}(v_{|N_r(v)|}) \end{pmatrix} + \begin{pmatrix} \mathbf{r}(v_1) \\ \mathbf{r}(v_2) \\ \vdots \\ \mathbf{r}(v_{|N_r(v)|}) \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}(v_1) \\ \boldsymbol{\epsilon}(v_2) \\ \vdots \\ \boldsymbol{\epsilon}(v_{|N_r(v)|}). \end{pmatrix}$$

Assuming that $\theta^2(v^*)$s in $N_r(v)$ have the same value θ^2, then

$$\begin{pmatrix} \mathbf{r}(v_1) \\ \mathbf{r}(v_2) \\ \vdots \\ \mathbf{r}(v_{|N_r(v)|}) \end{pmatrix} \sim \mathcal{MVN}\left(\mathbf{0}_{N\cdot|N_r(v)|}, \theta^2 \times \mathbf{K} \otimes \mathbf{I}_{|N_r(v)|}\right).$$

Since the $\hat{\boldsymbol{\beta}}(v)$s and $\hat{\mathbf{b}}(v)$ are obtained from the null model, they keep the same estimated value as we described in section 2.3. Within the $N_r(v)$, the score-based variance component test statistic for θ^2 becomes

$$\begin{pmatrix} \mathbf{y}(v_1) - \mathbf{X}\hat{\boldsymbol{\beta}}(v_1) - \hat{\mathbf{b}}(v_1) \\ \mathbf{y}(v_2) - \mathbf{X}\hat{\boldsymbol{\beta}}(v_2) - \hat{\mathbf{b}}(v_2) \\ \vdots \\ \mathbf{y}(v_{|N_r(v)|}) - \mathbf{X}\hat{\boldsymbol{\beta}}(v_{|N_r(v)|}) - \hat{\mathbf{b}}(v_{|N_r(v)|}) \end{pmatrix}' \mathbf{K} \otimes \mathbf{I}_{|N_r(v)|} \begin{pmatrix} \mathbf{y}(v_1) - \mathbf{X}\hat{\boldsymbol{\beta}}(v_1) - \hat{\mathbf{b}}(v_1) \\ \mathbf{y}(v_2) - \mathbf{X}\hat{\boldsymbol{\beta}}(v_2) - \hat{\mathbf{b}}(v_2) \\ \vdots \\ \mathbf{y}(v_{|N_r(v)|}) - \mathbf{X}\hat{\boldsymbol{\beta}}(v_{|N_r(v)|}) - \hat{\mathbf{b}}(v_{|N_r(v)|}) \end{pmatrix}.$$

Note that, this matrix multiplication is summarized by

$$\sum_{v^* \in N_r(v)} (\mathbf{y}(v^*) - \mathbf{X}\hat{\boldsymbol{\beta}}(v^*) - \hat{\mathbf{b}}(v_i))' \mathbf{K} (\mathbf{y}(v^*) - \mathbf{X}\hat{\boldsymbol{\beta}}(v^*) - \hat{\mathbf{b}}(v^*))$$

which is equal to $\sum_{v \in N_r(v)} U(v^*)$. This result suggests that $T_r(v)$ is constructed by assuming the $\theta^2(v)s$ are same in the neighborhood $N_r(v)$ as we stated in Section 2.3.

C. Supplementary figures

Figures S1 and S2 show binarized maps of statistical significance when CLEAN-V, CLEAN-V without spatial correlation, CLEAN-V without cluster enhancement, and massive univariate analysis are applied to language, social cognition, relational processing, and gambling tasks.

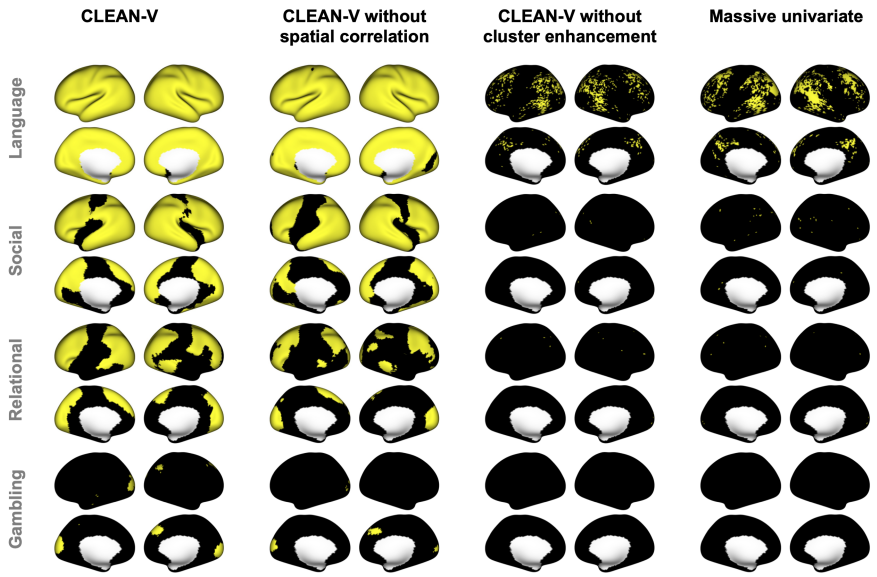

Figure S1: Test-retest reliability localization results from the four methods for the other four tasks.

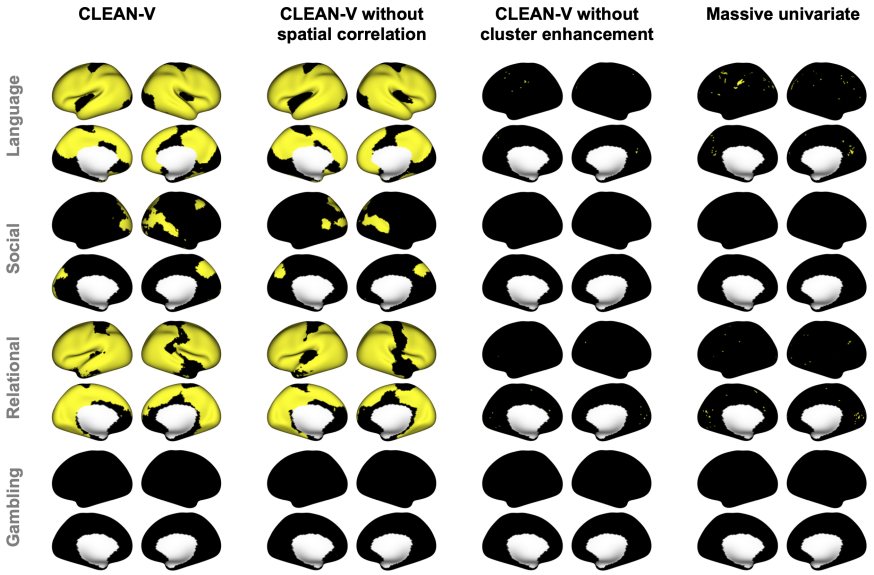

Figure S2: Narrow-sense heritability localization results from the four methods for the four other tasks.

References

[1] M. L. Elliott, A. R. Knodt, D. Ireland, M. L. Morris, R. Poulton, S. Ramrakha, M. L. Sison, T. E. Moffitt, A. Caspi, A. R. Hariri, What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis, Psychological Science 31 (2020) 792–806.

[2] J. T. Kennedy, M. P. Harms, O. Korucuoglu, S. V. Astafiev, D. M. Barch, W. K. Thompson, J. M. Bjork, A. P. Anokhin, Reliability and stability challenges in ABCD task fMRI data, NeuroImage 252 (2022) 119046.

[3] S. Marek, B. Tervo-Clemmens, F. J. Calabro, D. F. Montez, B. P. Kay, A. S. Hatoum, M. R. Donohue, W. Foran, R. L. Miller, T. J. Hendrickson, et al., Reproducible brain-wide association studies require thousands of individuals, Nature 603 (2022) 654–660.

[4] G. A. Blokland, K. L. McMahon, P. M. Thompson, N. G. Martin, G. I. de Zubicaray, M. J. Wright, Heritability of working memory brain activation, Journal of Neuroscience 31 (2011) 10882–10890.

[5] M. A. Lindquist, J. Spicer, I. Asllani, T. D. Wager, Estimating and testing variance components in a multi-level GLM, NeuroImage 59 (2012) 490–501.

[6] B. B. Risk, H. Zhu, ACE of space: estimating genetic components of high-dimensional imaging data, Biostatistics 22 (2021) 131–147.

[7] S. Noble, D. Scheinost, R. T. Constable, A guide to the measurement and interpretation of fMRI test-retest reliability, Current Opinion in Behavioral Sciences 40 (2021) 27–32.

[8] S. M. Smith, T. E. Nichols, Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference, NeuroImage 44 (2009) 83–98.

[9] J. Y. Park, M. Fiecas, A. D. N. Initiative, et al., Permutation-based inference for spatially localized signals in longitudinal MRI data, NeuroImage 239 (2021) 118312.

[10] A. F. Mejia, Y. Yue, D. Bolin, F. Lindgren, M. A. Lindquist, A Bayesian general linear modeling approach to cortical surface fMRI data analysis, Journal of the American Statistical Association 115 (2020) 501–520.

[11] D. Spencer, Y. R. Yue, D. Bolin, S. Ryan, A. F. Mejia, Spatial Bayesian GLM on the cortical surface produces reliable task activations in individuals and groups, NeuroImage 249 (2022) 118908.

[12] J. L. Bernal-Rusiel, M. Reuter, D. N. Greve, B. Fischl, M. R. Sabuncu, A. D. N. Initiative, et al., Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data, NeuroImage 81 (2013) 358–370.

[13] B. B. Risk, D. S. Matteson, R. N. Spreng, D. Ruppert, Spatiotemporal mixed modeling of multi-subject task fMRI via method of moments, NeuroImage 142 (2016) 280–292.

[14] J. Y. Park, M. Fiecas, CLEAN: Leveraging spatial autocorrelation in neuroimaging data in clusterwise inference, NeuroImage 255 (2022) 119192.

[15] S. M. Weinstein, S. N. Vandekar, E. B. Baller, D. Tu, A. Adebimpe, T. M. Tapera, R. C. Gur, R. E. Gur, J. A. Detre, A. Raznahan, et al., Spatially-enhanced clusterwise inference for testing and localizing intermodal correspondence, NeuroImage (2022) 119712.

[16] M. A. Lindquist, The statistical analysis of fMRI data, Statistical Science 23 (2008) 439 – 464.

[17] M. M. Monti, Statistical analysis of fMRI time-series: a critical review of the GLM approach, Frontiers in human neuroscience 5 (2011) 28.

[18] H. Ganjgahi, A. M. Winkler, D. C. Glahn, J. Blangero, P. Kochunov, T. E. Nichols, Fast and powerful heritability inference for family-based neuroimaging studies, NeuroImage 115 (2015) 256–268.

[19] T. Ge, T. E. Nichols, P. H. Lee, A. J. Holmes, J. L. Roffman, R. L. Buckner, M. R. Sabuncu, J. W. Smoller, Massively expedited genome-wide heritability analysis (MEGHA), Proceedings of the National Academy of Sciences 112 (2015) 2479–2484.

[20] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, X. Lin, Rare-variant association testing for sequencing data with the sequence kernel association test, The American Journal of Human Genetics 89 (2011) 82–93.

[21] T. Zou, W. Lan, H. Wang, C.-L. Tsai, Covariance regression analysis, Journal of the American Statistical Association 112 (2017) 266–281.

[22] Z. Li, Y. Liu, X. Lin, Simultaneous detection of signal regions using quadratic scan statistics with applications to whole genome association studies, Journal of the American Statistical Association 117 (2022) 823–834.

[23] A. Datta, S. Banerjee, A. O. Finley, A. E. Gelfand, Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets, Journal of the American Statistical Association 111 (2016) 800–812.

[24] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, et al., Function in the human connectome: task-fMRI and individual differences in behavior, NeuroImage 80 (2013) 169–189.

[25] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, M. Jenkinson, The minimal preprocessing pipelines for the Human Connectome Project, NeuroImage 80 (2013) 105–124.

[26] D. D. Pham, J. Muschelli, A. F. Mejia, ciftitools: A package for reading, writing, visualizing, and manipulating cifti files in R, NeuroImage 250 (2022) 118877.

[27] C. Hawco, E. W. Dickie, G. Jacobs, Z. J. Daskalakis, A. N. Voineskos, Moving beyond the mean: Subgroups and dimensions of brain activity and cognitive performance across domains, NeuroImage 231 (2021) 117823.

[28] S. M. Weinstein, S. N. Vandekar, A. Adebimpe, T. M. Tapera, T. Robert-Fitzgerald, R. C. Gur, R. E. Gur, A. Raznahan, T. D. Satterthwaite, A. F. Alexander-Bloch, et al., A simple permutation-based test of intermodal correspondence, Human brain mapping 42 (2021) 5175–5187.

[29] S. Geuter, G. Qi, R. C. Welsh, T. D. Wager, M. A. Lindquist, Effect size and power in fMRI group analysis, Biorxiv (2018) 295048.

[30] A. Eklund, T. E. Nichols, H. Knutsson, Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates, Proceedings of the national academy of sciences 113 (2016) 7900–7905.

[31] P. Duchesne, P. L. De Micheaux, Computing the distribution of quadratic forms: Further comparisons between the liu–tang–zhang approximation and exact methods, Computational Statistics & Data Analysis 54 (2010) 858–862.