Robust spatial extent inference with a semiparametric bootstrap joint inference procedure

Simon N. Vandekar1 | Theodore D. Satterthwaite2 | Cedric H. Xia2 | Azeez Adebimpe2 | Kosha Ruparel2 | Ruben C. Gur2 | Raquel E. Gur2 | Russell T. Shinohara3

1Department of Biostatistics, Vanderbilt University, Nashville, Tennessee
2Department of Psychiatry, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania
3Department of Biostatistics, Epidemiology, and Informatics, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

Correspondence
Simon N. Vandekar, Department of Biostatistics, Vanderbilt University, 2525 West End Ave., #1136, Nashville, TN 37203.
Email: simon.vandekar@vanderbilt.edu

Funding information
National Institute of Neurological Disorders and Stroke, Grant/Award Number: R01NS085211; National Cancer Institute, Grant/Award Number: 5P30CA068485-22; National Institute of Mental Health, Grant/Award Numbers: R01MH107235, R01MH107703, R01MH112847

Abstract
Spatial extent inference (SEI) is widely used across neuroimaging modalities to adjust for multiple comparisons when studying brain-phenotype associations that inform our understanding of disease. Recent studies have shown that Gaussian random field (GRF)-based tools can have inflated family-wise error rates (FWERs). This has led to substantial controversy as to which processing choices are necessary to control the FWER using GRF-based SEI. The failure of GRF-based methods is due to unrealistic assumptions about the spatial covariance function of the imaging data. A permutation procedure is the most robust SEI tool because it estimates the spatial covariance function from the imaging data. However, the permutation procedure can fail because its assumption of exchangeability is violated in many imaging modalities. Here, we propose the (semi-) parametric bootstrap joint (PBJ; sPBJ) testing procedures that are designed for SEI of multilevel imaging data. The sPBJ procedure uses a robust estimate of the spatial covariance function, which yields consistent estimates of standard errors, even if the covariance model is misspecified. We use the methods to study the association between performance and executive functioning in a working memory functional magnetic resonance imaging study. The sPBJ has similar or greater power to the PBJ and permutation procedures while maintaining the nominal type 1 error rate in reasonable sample sizes. We provide an R package to perform inference using the PBJ and sPBJ procedures.

KEYWORDS
bootstrap, FWER, neuroimaging, semiparametric inference, Spatial extent inference

1 | INTRODUCTION

Investigating neuroimage-phenotype associations is important to understand how stimuli, environment, and diseases affect the brain. Various neuroimaging modalities are used to study brain-behavior associations. For example, imaging scientists can study how brain activation measured with functional magnetic resonance imaging (fMRI) is related to performance on an executive functioning task (Satterthwaite et al., 2013). Spatial extent inference (SEI) is a widely used tool across neuroimaging modalities to study image-phenotype associations. SEI views the statistical image as a realization of a stochastic process and thresholds the image at a given value in
order to compute the size of spatially contiguous suprathreshold clusters of activation. A \( p \)-value is determined for each contiguous cluster based on its size by comparing the cluster size (extent) to the distribution of the maximum cluster size under a predetermined null hypothesis (Worsley et al., 1999).

A number of recent studies have shown that the standard tools used for SEI that rely on Gaussian random field (GRF) approximations can have highly inflated false-positive rates (Eklund et al., 2016; Greve and Fischl, 2018). These studies have provoked significant controversy about the conditions necessary for the GRF-based tools to produce nominal family-wise error rates (FWERs; Kessler et al., 2017; Mueller et al., 2017; Slotnick, 2017; Eklund et al., 2019). A good deal of this discussion has led to ad hoc preprocessing and thresholding suggestions for particular data types that are based on simulation results (Eklund et al., 2016; Flandin and Friston, 2016; Mueller et al., 2017; Slotnick, 2017; Greve and Fischl, 2018). For example, simulation results suggest using particular values of preprocessing parameters so that the GRF-based methods yield nominal FWERs (Flandin and Friston, 2016). However, GRF-based methods cannot control the FWER generally because they assume the spatial covariance function only has a local spatial component, when, in fact, there are interhemispheric and intranetwork correlations (see, eg, Satterthwaite et al., 2012; Woodward and Heckers, 2016). These distal correlations may be why Eklund et al. (2016) noted a slower than Gaussian decay in the spatial correlation function in the supplement of their paper.

To date, only multivariate permutation testing procedures have been shown to robustly control the spatial extent false-positive rate in most neuroimaging data sets (Winkler et al., 2014; Eklund et al., 2016; Greve and Fischl, 2018). However, the permutation procedure may not control the false-positive rate at the nominal level in some cases because it relies on the assumption of exchangeability. Exchangeability includes the assumption of homoskedasticity, namely, that the spatial covariance functions of the subjects are unaffected by the covariates. This assumption is often violated in neuroimaging data sets because many types of imaging data sets require multilevel models where parameters are estimated separately for each subject and subsequently analyzed at the group level.

For example, with task-based fMRI data, subject-level estimates of parameter images are obtained by fitting a linear model designed to capture how the fMRI time series varies in response to experimentally manipulated task demands. Then, a single image of location-specific parameter estimates for each subject is used as the outcome in a group-level analysis, where associations with the diagnostic group or other covariates can be investigated. Because covariates such as in-scanner motion, age, and diagnosis can affect the amount of noise in the parameter estimate images, the subjects may have different variance parameters conditional to the covariates. This difference in subject-level variance implies that the subjects are not exchangeable. Moreover, these covariates may affect the variance differently at separate locations in the image.

Because images are multivariate outcomes, exchangeability can be violated even if the variance at each location is identical across subjects. This is because a covariate can affect the spatial correlation between two locations in a brain image. For example, there is evidence that exchangeability is violated because the spatial covariance structure of imaging data is known to be affected by covariates, such as in-scanner motion, age, and diagnosis (Satterthwaite et al., 2012; Woodward and Heckers, 2016).

In the statistics literature, many procedures have been developed for inference in spatial data sets that could be applied in neuroimaging. Sun et al. (2015) developed false discovery rate (FDR) controlling procedures in a Bayesian context in order to perform inference pointwise or on predefined clusters. Procedures to address large-scale hypothesis testing for correlated tests could also be utilized to perform inference on images (eg, Yekutieli and Benjamini, 1999; Benjamini and Yekutieli, 2001; Romano et al., 2008). These procedures, which perform inference pointwise have the desirable property that they identify locations of the image where the mean is likely to be nonzero. However, due to the high dimensionality, small sample sizes, and complex spatial covariance functions of neuroimaging data, many studies lack the power to perform inference directly on the mean function (Mumford and Nichols, 2008; Durnez et al., 2014). As a solution, other proposed methods for controlling the FDR in spatial data improve power by using predefined clusters and performing inference on the clusters (Pacifico et al., 2004; Benjamini and Heller, 2007). Because the spatial FDR approach requires predefined clusters, the investigator must guess how to define subregions of the image before making an inference. In contrast, SEI does not require that the user defines the clusters a priori, but does require selecting a cluster forming threshold (CFT) for the statistical image to define suprathreshold clusters (see Section 2). SEI is similar to these cluster-based approaches because a significant cluster only guarantees (with a prespecified probability) that an arbitrary nonzero portion of the cluster has a nonzero mean.

Here, we develop two methods to perform SEI that use (semi-) parametric bootstrap joint (PBJ; sPBJ) testing
procedures to compute test statistics and estimate the joint distribution of the statistics. The PBJ procedure accommodates violations of exchangeability by using subject weights to deweight noisy observations at each volumetric pixel (called a “voxel”) in the image. If the selected weights are inversely proportional to the subject-level variances then the covariance function is correctly specified and the parameter estimators have minimum variance. In order to address the case that subject weights are not correctly specified, we introduce the sPBJ procedure that uses a robust “sandwich” estimate of the covariance function, which is a consistent estimator of the covariance function of the stochastic process (estimator HC3 in MacKinnon and White, 1985; Long and Ervin, 2000). Because the covariance function of the stochastic process is complex in imaging data, we leverage a bootstrap procedure to estimate the distribution of the maximum cluster size (Vandekar et al., 2018). These methods are asymptotically robust across any range of preprocessing parameters and CFTs because they estimate the covariance structure from the observed data. The developments presented here are novel in two ways: we extend our previously proposed PBJ procedure to a multilevel and semiparametric setting using an estimating equations approach for stochastic processes. In addition, we utilize this procedure to perform SEI and demonstrate that it controls the FWER in small samples using a realistic simulation analysis.

We utilize a subset of 1000 subjects from the Philadelphia Neurodevelopmental Cohort (PNC; Satterthwaite et al., 2014) and a novel approach to generate realistic simulated data, in which the assumption of exchangeability is violated, in order to show that the PBJ and sPBJ procedures are robust to violations of this assumption (Section 4). We demonstrate the procedures by studying the association between task performance and executive functioning measured with fMRI using the N-back task (Sections S3 and S5; Ragland et al., 2002; Satterthwaite et al., 2013). While the procedure is generally applicable for inference in spatial data sets, we developed an R package to perform the PBJ and sPBJ SEI procedures on NIFTI images that is available for download from Github and Neuroconductor (see Supporting Information; Muschelli et al., 2018).

2 SPATIAL EXTENT INFERENCE

Our approach is designed for group-level analysis and assumes a single image for each subject: let \( Y_i(v) \) denote the image for subject \( i = 1, ..., n \), which can be described as stochastic processes on \( \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}) \), where \( \mathbb{B} \subset \mathbb{R}^3 \) denotes the bounded space of interest (in neuroimaging, the brain), and \( \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}) \) denotes the space of functions from \( \mathbb{B} \) to \( \mathbb{R} \) with the finite infinity norm. With reference to the \( N \)-back analysis in Section S5, \( Y_i(v) \) represents the image that is the difference between the parameter estimates for the 2 and 0 back conditions for subject \( i \). We assume that

\[
Y_i(v) = X_{i0}\alpha(v) + X_{i1}\beta(v) + E_i(v) = X_i\gamma(v) + E_i(v),
\]

where \( X_{i0} \in \mathbb{R}^{m_0} \) is a row vector of nuisance covariates including the intercept, \( X_{i1} \in \mathbb{R}^{m_1} \) is a row vector of the variables of interest, \( m = m_0 + m_1 \), \( X_i = [X_{i0}, X_{i1}] \), \( \alpha(v) \in \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}^{m_0}) \), and \( \beta(v) \in \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}^{m_1}) \) are parameter image vectors, \( \gamma(v) = [\alpha(v)^T, \beta(v)^T]^T \), and \( E_i(v) \) is an error term with \( \mathbb{E}[E_i(v)] = 0 \) and spatial covariance function \( \Sigma_i(v, w) = \text{Cov} \{E_i(v), E_i(w)\} < \infty \) for all \( v, w \in \mathbb{B} \). Here, all capital letters denote random variables or stochastic processes. Finally, let \( X_0 = (X_{00}^T, ..., X_{0n}^T) \in \mathbb{R}^{m_0 \times n} \), \( X_i = (X_i^T, ..., X_{i1}^T) \in \mathbb{R}^{m_1 \times n} \), \( X = [X_0, X_i] \), \( Y(v) = (Y_i(v), ..., Y_n(v)) \).

Let \( Z_0(v) \) denote the test statistic image (eg, a normal or chi-square) under the null hypothesis

\[
H_0(v) : \beta(v) = 0 \quad \text{for all } v.
\]

\( Z_0(v) \) is a stochastic process on \( \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}) \). In words, the hypothesis (2) states that the image is unassociated with the covariates \( X_{i1} \). Let \( z(v) \in \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}) \) be the observed statistical image that is sampled from a stochastic process \( Z(v) \) on \( \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}) \) where the assumption (2) does not hold over a measurable subset of \( \mathbb{B} \). We assume that the sample paths of \( Z_0(v) \) and \( Z(v) \) are continuous almost everywhere. Here, \( Z(v) \) denotes the distribution of the statistical image under the true distribution and \( Z_0(v) \) denotes the distribution of the statistical image under the null (2).

Let \( C(z(v), z_0) : \mathcal{L}^\infty(\mathbb{B}, \mathbb{R}) \times \mathbb{R} \to \mathbb{R}^\infty \) be a function that thresholds the image \( [z(v)] \) with a given CFT, \( z_0 \in \mathbb{R} \), and returns the volumes of all suprathreshold contiguous clusters in the image. \( C(z(v), z_0) \) takes values in \( \mathbb{R}^\infty \) because there is an arbitrary number of contiguous clusters possible. To give the clusters a natural ordering, we assume that the clusters are indexed in decreasing order and there are a finite number of nonzero clusters.

**Procedure 1 SEI.** For a fixed \( z_0 \in \mathbb{R} \) we set

\[
p_j = \mathbb{P}\{C(Z_0(v), z_0) \geq C(Z(v), z_0)\}
\]

and call cluster \( j \) “significant” if \( p_j < \alpha \) for some predetermined rejection threshold \( \alpha \in [0, 1] \).
3 | STATISTICAL THEORY FOR THE BOOTSTRAP PROCEDURES

3.1 | Overview

We propose two approaches to model the spatial covariance function of the statistical image. The PBJ procedure is essentially a multivariate weighted least squares approach, which assumes the within subject variance estimates are correct. The sPBJ procedure utilizes a working spatial covariance function. The term joint in the name of the procedures refers to the fact that the PBJ and sPBJ procedures estimate the spatial covariance function of the statistical image, instead of assuming a restrictive spatial structure as in the classical GRF-base approaches. An advantage of the sPBJ is that the procedure yields consistent standard errors even if the working spatial covariance function is incorrect. We use the simulations in Section 4 to assess finite sample performance.

3.2 | PBJ: Parametric covariance functions

The PBJ model assumes that the spatial correlation function is the same for all subjects, but the covariance function may differ:

\[ \Sigma_i(\mathbf{v}, \mathbf{w}) = W_i(\mathbf{v})^{-1/2}W_i(\mathbf{w})^{-1/2}\Delta(\mathbf{v}, \mathbf{w}) \]  

for \( i = 1, \ldots, n \), where \( \Delta(\mathbf{v}, \mathbf{w}) = \text{Cor}\{E_i(\mathbf{v}), E_i(\mathbf{w})\} \) denotes a common spatial correlation function for all subjects. The weight functions \( W_i(\mathbf{v}) \) are assumed to be proportional to the inverse of the variance, \( \Sigma_i(\mathbf{v}, \mathbf{v}) \). Inference using the weighted procedure is conditional to the weights, which are usually estimates (or approximations) of the standard deviation of \( E_i(\mathbf{v}) \).

Under assumption (4) we can compute the \( \chi^2_{m_i} \)-statistic image:

\[ Z(\mathbf{v}) = \Phi^{-1}_{m_i}[\Phi_F(T(\mathbf{v}))], \]

where \( T(\mathbf{v}) \) is the F-statistic image for the test of the hypothesis (2), \( \Phi_F \) denotes the cumulative distribution function (CDF) of an F-distribution on \( m_1 \) and \( n - m \) degrees of freedom, and \( \Phi^{-1}_{m_i} \) is the inverse CDF of a \( \chi^2_{m_i} \)-statistic (Vandekar et al., 2018). The transformation in (5) makes \( Z_0(\mathbf{v}) \) asymptotically a diagonal Wishart process (Vandekar et al., 2018):

\[ Z_0 \sim \text{diag}\{W(m_1, \Sigma_Z)\} \]

under the null (2), where \( \Sigma_Z(\mathbf{v}, \mathbf{w}) = \Delta(\mathbf{v}, \mathbf{w}) \). The transformation on the right hand side of (5) improves the finite sample performance of the PBJ procedure (Vandekar et al., 2018). The PBJ procedure, given in Procedure S2.1, samples observations from a diagonal Wishart process, conditional to an estimate of \( \Sigma_Z \). Procedure S2.1 can be used to estimate the distribution of the maximum cluster size to compute the cluster \( p \) values (3).

3.3 | sPBJ: Sandwich covariance functions

The PBJ procedure is not robust to model misspecification as it assumes that the correlation structure (4) is the same for all subjects. Moreover, if the weights used in the PBJ procedure are incorrectly specified, then the standard errors for the parameter estimate image \( \hat{\beta}(\mathbf{v}) \) in model (1) are asymptotically biased and the FWER may not be controlled at the nominal level. This follows directly from the simpler case where the weight function and outcome are not images (see, eg, chapter 7 of Boos and Stefanski, 2013). The sPBJ procedure utilizes an estimating equations approach to obtain parameter estimates and standard errors that are robust to misspecification of the covariance function, which can occur with fMRI data because the spatial covariance function \( \Sigma_i(\mathbf{v}, \mathbf{w}) \) can differ across subjects. We use the following estimating equation:

\[ \Psi(\xi, \mathbf{v}, \mathbf{Y}) = \sum_{i=1}^{n} W_i(\mathbf{v})(Y_i(\mathbf{v}) - X_i(\mathbf{v})\mathbf{X}_i^T) = 0, \]

where \( W_i(\mathbf{v}) \) is a weight for subject \( i \) that is an estimator for \( \mathbb{E}[\Sigma_i(\mathbf{v}, \mathbf{v})|X_i]^{-1} \) for \( \mathbf{v} \in \mathbb{R} \). \( W_i(\mathbf{v}) \) does not need to be
an estimator of $\mathbb{E}\{\Sigma_i(v, v)|X_i\}^{-1}$ for the sPBJ procedure to be asymptotically valid; however, the best choices are estimators that are nearly unbiased for $\Sigma$ and have small variance. Because our parameters are spatial functions indexed by $v \in \mathbb{B}$, (7) is a stochastic process on $\mathbb{B}$.

We define $\hat{\zeta}(v)$ as the solution to (7), which is equivalent to the least squares estimate for $\zeta(v)$ at each location $v$ in the image using the weight function $W(v)$. $\hat{\zeta}(v)$ is unbiased in finite samples and has covariance function equal to

$$\Sigma_{\hat{\zeta}}(v, w) = \text{Cov}[n^{1/2}\{\hat{\zeta}(v) - \zeta(v)\}, n^{1/2}\{\hat{\zeta}(w) - \zeta(w)\}] = A(v)^{-1}\Omega_{\hat{\zeta}}(v, w)A(w)^{-1},$$

where

$$A(v) = \lim_{n \to \infty} n^{-1}\sum_{i=1}^{n} \mathbb{E}[W_i(v)X_i^2 X_i], \quad \Omega_{\hat{\zeta}}(v, w)$$

and we assume both limits (9) exist. We define (9) as limits because the underlying observations are not identically distributed due to having different subject-level variances. The covariance (8) is derived by performing a Taylor expansion of the estimating equation (7) centered at the true value of the parameter $\zeta(v)$ and then computing their asymptotic covariances (see, eg, Boos and Stefanski, 2013, p. 300). Theorem S2.1 gives a useful means to compute an estimate of the sandwich covariance function for the parameter of interest $\hat{\Sigma}_{\hat{\zeta}}(v, w)$. We use a modification (by Long and Ervin, 2000) of the covariance estimate of MacKinnon and White (1985), which was shown in simulations to be best at achieving nominal type 1 error rates in small samples over several heteroskedasticity-consistent estimators (Long and Ervin, 2000).

For fixed $v$, the robust Wald statistical image, $Z_0(v)$, for the test of the hypothesis image (2) is asymptotically chi-squared (under the null):

$$Z_0(v) = n\hat{\beta}(v)^T\hat{\Sigma}_{\hat{\zeta}}(v)^{-1}\hat{\beta}(v) \sim \chi^2_{m},$$

by Theorem 8.3 of Boos and Stefanski (2013, p. 345), where $\hat{\Sigma}_{\hat{\zeta}}(v) = \hat{\Sigma}_{\hat{\zeta}}(v, v)$ is defined in (S1). In practice, we make use of the transformation (5) using $m_1$ and $n - m$ degrees of freedom to compute (10) because we observed that this transformation produces type 1 error rates that are closer to the nominal level in small sample simulations.

Theorem S2.2 describes the joint distribution of the statistical image $Z_0(v)$ from (10) when $m_1 = 1$ and Procedure S2.2 describes how to sample from the joint distribution. When $m_1 > 1$, $Z_0(v)$ is a multivariate stochastic process that is challenging to rapidly sample from in high dimensions. Future work will investigate fast sampling for multivariate parameters.

### 3.4 Selection of weights for neuroimaging studies

For the PBJ and sPBJ procedures, the optimal weights $W_i(v)$ are proportional to the inverse of the variance at location $v$. The sPBJ procedure still yields consistent standard errors if the weights are misspecified (White, 1980). In neuroimaging, several weighting schemes are reasonable:

1. An estimate of the inverse of the subject’s variance image $\Sigma_i(v, v)$, which can be obtained for fMRI analyses from the standard output of some imaging software, such as FSL’s FEAT program (Jenkinson et al., 2012). This is the approach used in Section S5.

2. A proxy for the subject’s variance. Because the sPBJ procedure is consistent regardless of the chosen weights, an approximation to the subject’s variance image can be used in place of an estimate of the variance image. One example is $W_i(v) = \text{mot}_i$, where $\text{mot}_i$ is the mean in-scanner motion measured from the fMRI time series. This approach has the advantage that the weighting does not depend on the image location, so reduces computation time. This procedure is implemented in the simulation analysis in Section 4 and denoted PBJ(mot) and sPBJ(mot).

3. Uniform/equal weights. For the PBJ procedure, this implies the assumption of homoskedasticity, which is likely violated in many imaging data sets. For the sPBJ procedure, while not optimal, uniform weights still yield consistent standard errors. This procedure is implemented in the simulation analysis in Section 4 and denoted PBJ(1) and sPBJ(1).

### 4 RESAMPLING-BASED SIMULATION ANALYSIS

In this section, we perform realistic simulation analyses to assess the spatial extent FWER by resampling imaging data from the PNC. Synthetic simulations where the spatial covariance function is explicitly specified are given in Section S4.
4.1 Methods

Detailed data processing steps are given in Section S3 of the online supporting information. Our analyses utilize a subset of 1000 subjects, who satisfied image quality and clinical exclusionary criteria from the total of 1601 scanned as part of the PNC. In order to generate realistic nonexchangeable data with heteroskedastic covariance functions, we first residualize the entire N-back neuromaging data set to a set of independent variables that have known associations with the imaging variables, including age, sex, in-scanner motion (mot), and task performance (d’; Macmilling, 2002), using the unpenalized spline model:

\[
Y_i(v) = \alpha_0(v) + \alpha_i(v) \times \text{sex}_i + g(v, \text{age}_i) \\
+ h(v, \text{mot}_i) + \epsilon_i(v),
\]

(11)

where sex\(_i\) is an indicator for males, and \(g, h\) and \(\text{sex}\) are fit using thin plate splines with 10 knots (Wood and Augustin, 2002). Fixed degree spline bases were used to model continuous variables to ensure that the mean of the residuals was completely unassociated with the covariates.

As a result of fitting this model, the residualized images have little to no mean association with the independent variables. However, the residualized images do not remove the effect the independent variables can have on the covariance structure of the imaging data. By drawing subsamples of subjects from the residualized images together with the independent variables, we generate data that constitutes the global null hypothesis because the covariates are unassociated with the mean; however, these covariates can still have an effect on the spatial covariance function of the data. To demonstrate how the covariates affect the spatial covariance function of the residualized data, we create covariance matrices using nodes defined in the Power atlas for the 100 lowest and 100 highest valued subjects for motion and \(d’\) (Figure 1; Power et al., 2011). For example, high motion is associated with stronger interhemispheric and anterior-posterior correlations than low motion (Figure 1).

In each of 1000 simulations we draw a bootstrap sample of size \(n \in \{25, 50, 200\}\) and fit the model:

\[
\epsilon_i(v) = \beta_0(v) + \beta_1(v) \times \text{sex}_i + \beta_2(v) \times \text{age}_i \\
+ \beta_3(v) \times \text{mot}_i + \beta_4(v) \times d’_i + \epsilon_\text{d}(v),
\]

(12)

where \(\epsilon_i(v)\) are the residuals from model (11). We then create a \(\chi^2\) statistical image for the test of \(\beta_j(v) = 0\) for \(j \in \{3, 4\}\). We perform inferences for each parameter image using SEI with two CFTs \((p < .01, .001; \chi^2 > 6.63, 10.83)\). The less stringent threshold was used to show which procedures control the SEI FWER under a CFT where the classical GRF-based procedure fails.

We compare the PBJ and sPBJ (using 500 bootstraps) procedures to a standard GRF-based method (easythresh; Jenkinson et al., 2012) and to the permutation procedure (using 500 permutations; Winkler et al., 2014). For the simulation study, we use the PBJ and sPBJ with uniform weights (PBJ(1); sPBJ(1)) and with the inverse of motion as the weight (PBJ(mot); sPBJ(mot)), which downweights subjects with high motion. These weighting schemes are described in detail in Section 3.4. The permutation procedure is fit using a single-group variance estimate (Perm) as well as a two-group variance (Perm Grp), which attempts to weaken the assumption of exchangeability by allowing for group differences in variance at each location (Winkler et al., 2014). The groups were determined by the upper and lower quantiles of motion. The group-based procedure estimates a separate variance for each group and normalizes out the variances when computing the test statistic in each permutation (Winkler et al., 2014). The SEI FWER is estimated by computing the proportion of simulations where a cluster is determined significant at two rejection thresholds \(\alpha = .01\) and \(.05\).

We use a novel procedure to assess power: for each of 1000 simulations, we randomly select four voxels within the brain and create an image of spheres, \(s(v)\), with radii of 8, 10, 11, and 12 voxels centered at each of the four voxels. We set \(s(v) = 1\) when \(v\) is inside a sphere (and inside the brain) and 0 otherwise. \(s(v)\) represents the locations in the image where there is a true effect. Because the four voxels are randomly selected, the spheres could overlap or be cropped depending on the proximity to the edge of the brain, thus the number and size of clusters is random across simulations. For locations where \(s(v) = 1\), the \(d’\) variable is first orthogonalized to the other covariates and then we set \(\beta_d(v) = \sigma_d(v)/\sigma_d \times 0.4\), where \(\beta_d\) as defined in (12) and \(\sigma_d\) is the variance of the orthogonalized \(d’\) variable. This approach ensures that the voxel level effect size of \(d’\) is approximately equal to 0.4. Power curves as a function of cluster size are estimated for each sample size using shape constrained additive models (SCAMs) using the \texttt{scam} package in R (Pya and Wood, 2015). SCAMs are used to ensure the estimated power curves are monotonically increasing. A true positive is considered when the suprathreshold statistical map intersects with \(\{v: s(v) = 1\}\) by at least one voxel and has a spatial extent \(p\)-value less than .01. Power was only assessed for procedures that had nearly nominal type 1 error rate.
Results

Simulations under the global null (2) were used to assess FWER for each of the SEI procedures. For the test of the motion variable, all methods have inflated FWER for smaller sample sizes \((n = 25, 50)\), except the PBJ and sPBJ procedures with motion deweighting (Figure 2). Because the sPBJ procedure without weights is asymptotically valid, the FWER falls to the nominal level with increasing sample size. The permutation procedures and the PBJ procedure with uniform weights actually increase FWER with increasing sample size. The GRF-based method always has a FWER above 0.7. This pattern holds across both CFTs \((0.01, 0.001)\) and both nominal FWER \((0.01, 0.05)\).

Our novel null simulation approach allows us to assess how the heteroskedasticity of real variables of interest affect the FWER for each procedure. For the test of \(d'\), most procedures control the FWER at the nominal level (Figure 2). The FWER for the test of \(d'\) using the PBJ procedure with motion deweighting increases above 0.1 with both CFTs when the desired level is 0.05. The GRF-based procedure maintains a FWER above 0.2.

Simulations where the known signal is added to the image are used to assess the power of each method. Because the parameter image was randomly generated in each simulation, cluster sizes varied across simulations, so power can be plotted as a function of the true parameter image’s cluster sizes (Figure 3). Across sample sizes and CFTs, the PBJ and sPBJ with motion deweighting tend to have a greater power at a nominal FWER of \(\alpha = 0.01\), with the exception of the CFT = 0.01 when \(n = 25, 50\) (Figure 3). This FWER was chosen to assess power because most methods were approximately maintaining the nominal FWER. The use of the motion weighted sPBJ procedure may decrease power slightly relative to the motion weighted PBJ procedure. However, the PBJ(mot) method had a slightly increased error rate (Figure 2; \(d'\), Nom FWER = 0.01), so the increased power of that method may be due in part to the slightly higher FWER. The permutation procedure with group variances...
has a detrimental effect on the power of the procedure. The more conservative CFT has a greater power to detect small clusters because small clusters with a strong signal are less likely under the null hypothesis (2).

5 | DISCUSSION

The sPBJ procedure is a robust inference procedure that uses spatial covariance function estimators that are consistent when the covariance model is misspecified. The permutation, GRF, and PBJ procedures that do not have consistent covariance estimates can have FWERs that grow as the sample size increases. The sPBJ procedure with noise deweighting has equal or greater power and produces nominal error rates at CFTs where classical GRF-based methods have been shown to fail, and so, is an appealing option for SEI of multilevel imaging data.

Our findings support those presented in other papers (Silver et al., 2011; Eklund et al., 2016; Greve and Fischl, 2018): GRF-based methods do not have robust FWER control across a range of parameter settings, regardless of whether heteroskedasticity is present. The only variable that produces enough heteroskedasticity to affect the permutation procedure is the motion covariate; the permutation procedure performs well with \(d'\), which is a typical covariate of interest. Thus, a model that assumes subject exchangeability is appropriate for the covariate we consider here, albeit with a slight loss in power. It is possible that covariates that are highly correlated with motion (eg ADHD diagnosis) may have FWERs that are far from the nominal level using the permutation procedure.

Here, we perform analyses and simulations of task fMRI data, which is one of the most common multilevel neuroimaging models, where subject-level parameter estimate images are obtained from subjects’ fMRI time...
series and then the parameter estimate images are analyzed across subjects at the group level. The (s)PBJ procedures are appropriate for other types of imaging modalities that are not typically considered multilevel models. For example, estimates of voxel level variance can be used to deweight noisy cerebral blood flow images computed from arterial spin labeled MRI.

There are a few apparent limitations of the current work: First, the sPBJ procedure relies on consistency, which only provides guarantees for FWER control as the sample size gets large. Our simulations demonstrate that the procedure has nearly nominal FWER control in relatively small samples (eg, n = 50), however this convergence is likely slower in very heavily skewed data sets. Second, the distribution of the statistical image (S5) is challenging to sample from when \( m_1 > 1 \), so the procedure is not able to perform \( F \) tests of multiple parameters at each location. Future work is necessary to derive the distribution of the statistical image in the case when \( m_1 > 1 \) and develop efficient methods to sample from the joint distribution. Finally, we only assessed the effect of heteroskedasticity in a group-level model on error rates. Other factors such as spatial smoothing and

**FIGURE 3** Power results for the SEI procedures that had nearly FWER control in the null simulations at a nominal type 1 error rate of \( \alpha = .01 \). Each plot shows the power vs cluster size for CFTs of 0.01 and 0.001. CFT, cluster forming threshold; FWER, family-wise error rate; SEI, spatial extent inference [Color figure can be viewed at wileyonlinelibrary.com]
subject-level processing are known to affect FWER of GRF-based SEI procedures. Future work will study the effect of these other features on the error rates of the (s) PBJ methods. Despite these limitations, the sPBJ procedure is a robust multilevel method for SEI that can be used to study image-covariate associations.

ACKNOWLEDGMENTS

Thanks to Shawn Garbett for helpful thoughts on some theoretical aspects of this paper. We used a BrainNet Viewer to produce the figures (Xia et al., 2013). RC2 grants from the National Institute of Mental Health MH089983 and MH089924. The National Institutes of Health grants: 5P30CA068485-22 (to SNV), R01MH107235 (to RCG, REG, RTS), R01MH107703 (to TDS, RTS), R01MH112847 (to TDS, KR, and RTS) and R01NS085211 (to RTS).

ORCID

Simon N. Vandekar https://orcid.org/0000-0002-7457-9073

REFERENCES

Benjamini, Y. and Heller, R. (2007) False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102, 1272–1281.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4), 1165–1188.

Boos, D.D. and Stefanski, L.A. (2013) *Essential Statistical Inference: Theory and Methods*. New York: Springer-Verlag.

Durnez, J., Moerkerke, B. and Nichols, T.E. (2014) Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, 84, 45–64.

Eklund, A., Knutsson, H. and Nichols, T.E. (2019) Cluster failure revisited: Impact of first level design and physiological noise on cluster false positive rates, 40(7), 2017–2032. https://doi.org/10.1002/hbm.24350

Eklund, A., Nichols, T.E. and Knutsson, H. (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28), 7900–7905.

Flandin, G. and Friston, K.J. (2016) Analysis of family-wise error rates in statistical parametric mapping using random field theory. *Human Brain Mapping*, 40(7), 2052–2054.

Greve, D.N. and Fischl, B. (2018) False positive rates in surface-based anatomical analysis. *NeuroImage*, 171, 6–14.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W. and Smith, S.M. (2012) Fsl. *NeuroImage*, 62, 782–790.

Kessler, D., Angstadt, M. and Sripada, C.S. (2017) Reevaluating “cluster failure” in fMRI using nonparametric control of the false discovery rate. *Proceedings of the National Academy of Sciences*, 114, E3372–E3373.

Long, J.S. and Ervin, L.H. (2000) Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54, 217–224.

MacKinnon, J.G. and White, H. (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 305–325.

Macmillan, N.A. (2002) Signal detection theory. *Stevens’ Handbook of Experimental Psychology*, 4, 43–90.

Mueller, K., Lepsien, J., Möller, H.E. and Lohmann, G. (2017) Commentary: cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Frontiers in Human Neuroscience*, 11, 345.

Mumford, J.A. and Nichols, T.E. (2008) Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *NeuroImage*, 39, 261–268.

Muschelli, J., Gherman, A., Fortin, J.-P., Avants, B., Whitzer, B. and Clayden, J.D. et al. (2018) Neuroconductor: an R platform for medical imaging analysis. *Biostatistics*, 20(2), 218–239.

Pacifico, M.P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004) False discovery control for random fields. *Journal of the American Statistical Association*, 99, 1002–1014.

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A. et al. (2011) Functional network organization of the human brain. *Neuron*, 72, 665–678.

Pya, N. and Wood, S.N. (2015) Shape constrained additive models. *Statistics and Computing*, 25, 543–559.

Ragland, J.D., Turetsky, B.I., Gur, R.C., Gunning-Dixon, F., Turner, T., Schroeder, L. et al. (2002) Working memory for complex figures: an fMRI comparison of letter and fractal n-back tasks. *Neuropsychology*, 16, 370–379.

Romano, J.P., Shaikh, A.M. and Wolf, M. (2008) Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17, 417.

Satterthwaite, T.D., Elliott, M.A., Ruparel, K., Loughead, J., Prabhakaran, K., Calkins, M.E. et al. (2014) Neuroimaging of the Philadelphia neurodevelopmental cohort. *NeuroImage*, 86, 544–555.

Satterthwaite, T.D., Wolf, D.H., Erus, G., Ruparel, K., Elliott, M.A., Gennatas, E.D. et al. (2013) Functional maturation of the executive system during adolescence. *The Journal of Neuroscience*, 33, 16249–16261.

Satterthwaite, T.D., Wolf, D.H., Loughead, J., Ruparel, K., Elliott, M.A., Hakonarson, H. et al. (2012) Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *NeuroImage*, 60, 623–632.

Silver, M., Montana, G., Nichols, T.E. and Initiative, A.D.N. (2011) False positives in neuroimaging genetics using voxel-based morphometry data. *NeuroImage*, 54, 992–1000.

Slotnick, S.D. (2017) Cluster success: fMRI inferences for spatial extent have acceptable false-positive rates. *Cognitive Neuroscience*, 8, 150–155.

Sun, W., Reich, B.J., Tony Cai, T., Guindani, M. and Schwartzman, A. (2015) False discovery control in large-scale spatial multiple testing. *Journal of the Royal Statistical Society, Series B*, 77, 59–83.

Van der Vaart, A.W. (2000) *Asymptotic Statistics*. New York, NY: Cambridge University Press.

Vandekar, S.N., Satterthwaite, T.D., Rosen, A., Ciric, R., Roalf, D.R., Ruparel, K. et al. (2018) Faster family-wise error control for neuroimaging with a parametric bootstrap. *Biostatistics*, 19, 497–513.
White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4), 817–838.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M. and Nichols, T.E. (2014) Permutation inference for the general linear model. *Neuroimage*, 92, 381–397.

Wood, S.N. and Augustin, N.H. (2002) GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157, 157–177.

Woodward, N.D. and Heckers, S. (2016) Mapping thalamocortical functional connectivity in chronic and early stages of psychotic disorders. *Biological Psychiatry*, 79, 1016–1025.

Worsley, K.J., Andermann, M., Kouls, T., MacDonald, D. and Evans, A.C. (1999) Detecting changes in nonisotropic images. *Human Brain Mapping*, 8, 98–101.

Xia, M., Wang, J. and He, Y. (2013) BrainNet viewer: a network visualization tool for human brain connectomics. *PLOS One*, 8(7), e68910.

Yekutieli, D. and Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82, 171–196.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Vandekar SN, Satterthwaite TD, Xia CH, et al. Robust spatial extent inference with a semiparametric bootstrap joint inference procedure. *Biometrics*. 2019;75: 1145–1155. [https://doi.org/10.1111/biom.13114](https://doi.org/10.1111/biom.13114)