Computational and Mathematical Methods in Medicine

Effect-Size Estimation Using Semiparametric Hierarchical Mixture Models in Disease-Association Studies with Neuroimaging Data

Ryo Emoto,[1] Atsushi Kawaguchi,[2] Kunihiko Takahashi,[3] and Shigeyuki Matsui[1]

[1] Department of Biostatistics, Nagoya University Graduate School of Medicine, Nagoya 466-0003, Japan
[2] Faculty of Medicine, Saga University, Saga, 849-8501, Japan.
[3] Medical and Dental Data Science Center, Tokyo Medical and Dental University, Tokyo, 101-0062 Japan

Correspondence should be addressed to Ryo Emoto; remoto@med.nagoya-u.ac.jp

Abstract

In disease-association studies using neuroimaging data, evaluating the biological or clinical significance of individual associations requires not only detection of disease-associated areas of the brain, but also estimation of the magnitudes of the associations or effect sizes for individual brain areas. In this paper, we propose a model-based framework for voxel-based inferences under spatial dependency in neuroimaging data. Specifically, we employ hierarchical mixture models with a hidden Markov random field structure to incorporate the spatial dependency between voxels. A non-parametric specification is proposed for the effect size distribution to flexibly estimate the underlying effect size distribution. Simulation experiments demonstrate that compared with a naive estimation method, the proposed methods can substantially reduce the selection bias in the effect size estimates of the selected voxels with the greatest observed associations. An application to neuroimaging data from an Alzheimer's disease study is provided.

1. Introduction

In disease-association studies using neuroimaging data, such as those related to brain magnetic resonance imaging (MRI), screening of disease-associated regions in the brain is a fundamental statistical task to understand the underlying mechanisms of disease and also to develop disease diagnostics. Such screening analysis typically involves detection of disease associations in the framework of hypothesis testing, followed by estimation of the magnitudes of the associations or their effect sizes to determine their biological or clinical significance.

Many statistical methods have been proposed to detect disease associations. In a cluster-level inference, groups of contiguous voxels whose association statistic values are above a certain threshold are defined and then associated with disease status [1, 2]. Another approach is to test every voxel individually, which takes into account the serious multiplicity problem of testing enormous numbers of voxels simultaneously. In this voxel-level inference, several

model-based methods based on random field theory have been proposed. Smith and Fahrmeir proposed to use an Ising prior in a classical Markov random field to model the dependency among contiguous voxels [3]. More recently, Shu et al. (2015) [4] proposed to use hidden Markov random field modelling, and developed a multiple testing procedure based on the local index of significance (LIS) proposed by Sun and Cai (2009) [5] in multiple testing under dependency. Brown et al. (2014) proposed to use a Gaussian random field with conditional autoregressive models [6]. With these voxel-level methods, contiguous voxels may be more prone to rejection than conventional, voxel-level multiple testing procedures. They may also facilitate the interpretation of significant voxels or regions in neuroimaging data, as in cluster-level inference, while circumventing the problems with that approach, including the arbitrariness of the threshold used in initial clustering and the lack of spatial specificity [1].

On the other hand, for the problem of estimating disease associations, traditional neuroimaging studies reported "naive" estimates, such as Cohen's d, for significant voxels. However, several authors have pointed out that such methods may suffer from overestimation, reflecting a selection bias for picking up voxels with the greatest effect sizes, possibly due to random errors [7, 8]. Reddan et al. (2017) recommended several ways to either avoid such bias, for instance by testing predefined regions of interest or integrating effects across multiple voxels into a particular model, or to adjust bias using independent samples [7]. However, in association analysis of neuroimaging data with spatial dependency, the estimation problem has not been well studied compared with the detection problem using multiple testing.

In this paper, we use empirical Bayes estimation and hierarchical modelling of summary statistics from the whole set of features to derive shrinkage estimation for individual features [9, 10], and adapt this method to the analysis of disease-association studies using neuroimaging data with spatial dependence. Specifically, we employ hierarchical mixture models with a hidden Markov random field structure to incorporate the spatial dependency between voxels. We assume a non-parametric distribution for the underlying distribution of voxel-specific effect sizes. With a generalized expectation-maximization (EM) algorithm, we can estimate all the parameters in the model, including the effect size distribution. We then obtain shrinkage estimates for individual voxels and also an estimate of the LIS for control of the false discovery rate (FDR) in the detection problem based on the fitted model.

With an appropriate effect size statistic and its asymptotic sampling distribution, our method is generally applicable to effect size estimations in many neuroimaging association studies where general linear models have been employed, such as those with functional/structural MRI (fMRI/sMRI), diffusion tensor imaging (DTI), and so forth. This paper is organized as follows. We provide the proposed method in Section 2. We describe simulation experiments to evaluate the performance of the proposed methods and an application to neuroimaging data from an Alzheimer's disease study in Section 3. We discuss the details of the methods and results in Section 4. Finally, we conclude this paper in Section 5.

2. Materials and Methods

We propose an estimation method based on a hierarchical mixture model in which the underlying distribution of voxel-specific effect sizes is specified. We suppose a simple

situation where diseased and normal control subjects are compared without any covariates (see Section 2.5 for incorporation of covariates). We introduce a binary disease status variable with a group label of either 1 or 2, for example disease or normal. Let n_1 and n_2 be the numbers of diseased and normal control subjects, respectively, and $n = n_1 + n_2$ be the total number of subjects. We suppose that spatial normalization [1] has been performed for each subject to adjust for differences in the size or shape of the observed image, and the image is divided into voxels by a three-dimensional grid. We also suppose a further normalization to ensure normality of the voxel-level intensity values across subjects within each group. Let S be the set of all voxels in the neuroimaging data and m denote the number of voxels in S. In order to measure the association of the observed intensity values from individual voxels with the disease status variable, we define the standardized mean difference between the two groups. Specifically, for voxel $s \in S, \delta_s = (\mu_{1s} - \mu_{2s})/\sigma_s$, where μ_{1s} and μ_{2s} are the means of voxel s for groups 1 and 2, respectively, and σ_s is the common standard deviation for voxel s across groups. As an estimate of δ_s, we use the following statistic,

$$Y_s = \frac{\bar{\mu}_{1s} - \bar{\mu}_{2s}}{\hat{\sigma}_s} \quad (1)$$

where $\bar{\mu}_{1s}$ and $\bar{\mu}_{2s}$ are sample means of voxel values in the two groups and $\hat{\sigma}_s^2$ is an estimator of the common within-group variance. This statistic is essentially a two-sample *t*-statistic, apart from the sample size term. One may consider a calculation of Y_s from the *t*-value provided by software packages such as Statistical Parametric Mapping (SPM, https://www.fil.ion.ucl.ac.uk/spm/). Let $\boldsymbol{Y} = \{Y_s: s \in S\}$ be the vector of Y_s for all m voxels. Of note, the reason for using the standardized mean difference, rather than test statistics such as *Z*-statistics, is that it is a direct interpretation of the effect size of individual voxels with no dependency on the sample size.

2.1. Hierarchical Mixture Models in a Hidden Markov Random Field

We assume a hidden Markov random field model [4] for $\boldsymbol{Y}$. Let $\boldsymbol{\Theta} = \{\Theta_s: s \in S\} \in \{0,1\}^m$ be a set of latent variables, where $\Theta_s = 0$ if the voxel s is null (i.e., no association with disease) and $\Theta_s = 1$ otherwise (i.e., association with disease). The dependence structure across contiguous voxels is modeled assuming that this latent variable $\boldsymbol{\Theta}$ is generated from the following Ising model with two parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T$,

$$\Pr(\boldsymbol{\Theta} = \boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{\gamma}^T \boldsymbol{H}(\boldsymbol{\theta})\}}{C(\boldsymbol{\gamma})},$$

where $\boldsymbol{H}(\boldsymbol{\theta}) = \left(\sum_{(s,t)\in S_1} \theta_s \theta_t, \sum_{s\in S} \theta_s\right)^{\mathrm{T}}$ and $C(\boldsymbol{\gamma})$ is the normalizing constant. In the vector $\boldsymbol{H}(\boldsymbol{\theta})$, the first component pertains to a summation over all pairs of contiguous voxels, S_1, and the second component to a summation over all voxels, S.

Given the latent status $\boldsymbol{\Theta} = \boldsymbol{\theta}$, we assume that the statistics Y_ss are mutually independent, such that

$$\Pr(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{\Theta} = \boldsymbol{\theta}) = \prod_{s\in S} \Pr(Y_s = y_s | \Theta_s = \theta_s). \quad (2)$$

For the component $\Pr(Y_s = y_s|\Theta_s = \theta_s)$, we define f_0 as the null density function, $f_0(y_s) = \Pr(Y_s = y_s|\Theta_s = 0)$, and f_1 as the non-null density function, $f_1(y_s) = \Pr(Y_s = y_s|\Theta_s = 1)$. We assume the distribution of Y_s as the mixture of null and non-null distributions,

$$\Pr(Y_s = y_s) = Pr(\Theta_s = 0)f_0(y_s) + Pr(\Theta_s = 1)f_1(y_s), \quad (3)$$

Of note, this is an instance of the so-called "two-groups model" [11] when the hidden Markov random field model is introduced. When the sample size n is sufficiently large, it is reasonable to employ asymptotic normality for Y_s. For the null voxels, we assume f_0 to be a normal distribution, $N(0, c_n^2)$, where $c_n = \sqrt{n/n_1 n_2}$. For the non-null voxels, we assume the hierarchical structure with two levels:

$$\begin{aligned} Y_s|\delta_s, \Theta_s = 1 &\sim N(\delta_s, c_n^2), \\ \delta_s &\sim g(\cdot). \end{aligned} \quad (4)$$

At the first level, the conditional distribution of Y_s for effect size δ_s is normal with mean δ_s and variance c_n^2, again based on asymptotic normality for Y_s. At the second level, the voxel-specific effect size δ_s has an effect size distribution g. From this hierarchical structure, we can express the non-null density function as the marginal density function, $f_1(y_s) = \int f(y_s|\delta, \theta_s = 1)g(\delta)d\delta$, where $f(y_s|\delta, \theta_s = 1)$ is a conditional density function in the first level of Equation (4). Note that Equation (4) is the Brown-Stein model for estimating effect sizes [9, 12, 13].

If the sample size is not large enough, as occurs in many exploratory neuroimaging studies, it is reasonable to use the t-distribution rather than the normal distribution. In this case, the statistic Y_s/c_n follows a t-distribution with $n-2$ degrees of freedom for the null voxels, and we consider the following hierarchical model for the non-null voxels:

$$\begin{aligned} \frac{Y_s}{c_n}|\delta_s, \Theta_s = 1 &\sim t_{n-2,\delta_s/c_n}, \\ \delta_s &\sim g(\cdot). \end{aligned} \quad (5)$$

where $t_{n-2,\delta_s/c_n}$ represents a non-central t-distribution with $n-2$ degrees of freedom and noncentrality parameter δ_s/c_n.

2.2. Non-Parametric Effect Size Distribution

We can consider both parametric and non-parametric specifications for the effect size distribution g. However, the information regarding the parametric form of g is generally limited because of the exploratory nature of disease-association studies that observe neuroimaging data with a large number of voxels (see Section 4 for discussion of the technical difficulty of specifying parametric mixture models for the effect size distribution). We therefore consider a non-parametric specification and estimate it based on presumed parallel association structures across a large number of voxels. For this estimation, we propose to perform the smoothing-by-roughening method [14], in the same way this method has been used for analyzing genomic data [15]. We approximate that g has discrete probabilities $\boldsymbol{p} = (p_1, \ldots, p_B)$ at each mass point $t = (t_1, \ldots, t_B)$,

$$g(t_b; \boldsymbol{p}) = p_b, \quad b = 1, \ldots, B, \quad (6)$$

where B is a sufficiently large number of mass points and discrete probability p_b satisfies $p_1 + \cdots + p_B = 1$. In practice, we set $B = 200$, following the guideline by Shen and Louis (1999) [14]. The mass point t may be specified to cover a possible range of Y and $t_b \neq 0$ for any b.

When asymptotic normality is assumed, then based on Equations (4) and (6), the marginal non-null distribution of Y_s, f_1, can be expressed as a mixture of normal distributions,

$$f_1(y; \boldsymbol{p}) = \sum_{b=1}^{B} p_b \, \phi(y; t_b, c_n^2), \tag{7}$$

where $\phi(\cdot; \mu, \sigma^2)$ represents the density function of normal distribution, $N(\mu, \sigma^2)$. If the sample size is not large enough, the non-central *t*-distribution, $\phi_t(y/c_n; n-2, t_b/c_n)$, is substituted for the normal distribution, $\phi(y; t_b, c_n^2)$, in Equation (7), where $\phi_t(\cdot; \nu, \delta)$ represents the density function of the non-central *t*-distribution $t_{\nu,\delta}$. In this case, the marginal non-null distribution of Y_s, f_1, is a mixture of non-central *t*-distributions.

The parameter set specifying the above hierarchical model is $\boldsymbol{p}$. We use the vector $\boldsymbol{\varphi} = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{p}^{\mathrm{T}})^{\mathrm{T}}$ to represent the set of all parameters, including those in the Ising model. The parameter set $\boldsymbol{\varphi}$ is estimated by a generalized EM algorithm. Details of the algorithm are provided in Appendix A. Another approach to estimating the effect size distribution g is a non-parametric Bayes estimation with a Dirichlet process (DP) prior [16]. Assuming a DP prior for the discretized version of g, Equation (4) forms a DP mixture model that is equivalent to an infinite mixture model. It is pointed out that the estimated non-parametric distribution based on the smoothing-by-roughening algorithm with initial distribution $G^{(0)}$ behaves similarly to the one based on DP hyper-prior with mean $G^{(0)}$, where the number of repetitions in the smoothing-by-roughening algorithm is related to prior precision of the DP [15].

2.3. FDR Estimation

In our framework, multiple testing methods can be derived based on the estimated model. We employ the LIS [5] to estimate the FDR to incorporate the spatial dependency between voxels. As a function of the parameter $\boldsymbol{\varphi}$, the LIS is defined as the posterior probability that the voxel is null given all Y_ss,

$$\mathrm{LIS}_s(\boldsymbol{y}) = \Pr(\Theta_s = 0 | \boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{\varphi}).$$

Note that the LIS corresponds to the local FDR [17] when independence across voxels is assumed. Multiple testing is based on the LIS. Let $\mathrm{LIS}_{(1)}(\boldsymbol{y}) \leq \cdots \leq \mathrm{LIS}_{(m)}(\boldsymbol{y})$ represent a series of ordered LISs across voxels and let $H_{(i)}$ be the null hypothesis (representing no association with disease) on the voxel corresponding to $\mathrm{LIS}_{(i)}(\boldsymbol{y})$. A LIS-based, oracle LIS procedure was proposed for minimizing the false negative rate subject to a constraint on FDR under hidden Markov chain dependence [5]; this procedure was then extended under a hidden Markov random field for analyzing neuroimaging data [4]. The oracle LIS procedure determines rejected voxels using the following rule:

$$\text{let } k = \max\left\{ i: \frac{1}{i}\sum_{j=1}^{i} \text{LIS}_{(j)}(\boldsymbol{y}) \le \alpha \right\}, \quad \text{then reject all } H_{(i)}, i = 1, \dots, k. \tag{8}$$

This procedure controls the FDR level at α. Since the parameter $\boldsymbol{\varphi}$ is unknown, a plug-in estimator, $\widehat{\text{LIS}}_s(\boldsymbol{y}) = \Pr(\Theta_s = 0|\boldsymbol{y}; \widehat{\boldsymbol{\varphi}})$, is used. This probability, $\widehat{\text{LIS}}_s(\boldsymbol{y}) = \Pr(\Theta_s = 0|\boldsymbol{y}; \widehat{\boldsymbol{\varphi}})$, can be calculated using the Gibbs sampler from the distribution of $\boldsymbol{\Theta}|\boldsymbol{Y}$ [4],

$$\Pr(\boldsymbol{\Theta} = \boldsymbol{\theta}|\boldsymbol{Y} = \boldsymbol{y}; \widehat{\boldsymbol{\varphi}}) \propto \exp\left[\widehat{\gamma_1} \sum_{(s,t)\in S_1} \theta_s \theta_t + \sum_{s\in S} \{\hat{\gamma}_2 - \log f_0(y_s) + \log f_1(y_s; \widehat{\boldsymbol{p}})\} \theta_s \right]. \tag{9}$$

In applying the aforementioned FDR estimation procedure to neuroimaging data, it is generally reasonable to divide all voxels into neurologically defined sub-regions with distinct functional or structural features, such as Automated Anatomical Labeling (AAL, [18]) (thus resulting in plausible heterogeneity in effect size and dependence structure across sub-regions). We apply the pooled LIS [19] and fit the model separately for each sub-region, thereby obtaining LIS values within sub-region. We then determine rejected voxels by equation (8), where $\text{LIS}_{(1)}(\boldsymbol{y}) \le \cdots \le \text{LIS}_{(m)}(\boldsymbol{y})$ is the ordered LIS for a pool of all sub-regions.

2.4. Effect Size Estimation

As mentioned in Section 1, estimation of effect sizes for selected voxels is important for evaluating their biological or clinical significance. Of note, the naive estimator given by $\tilde{\delta}_s = Y_s$ generally overestimates the true effect size (absolute δ_s) for the selected "top" voxels with the highest statistical significance. This estimation bias reflects the selection bias, caused by random variation, that is inherent in selecting voxels with the largest absolute Y_s. We consider shrinkage estimation for selected voxels. Specifically, we extend posterior indices originally developed in the case of independent Y_ss [10] to the case of dependent Y_ss.

The posterior mean of δ_s for a non-null voxel s is given by

$$E[\delta_s|y_s, \Theta_s = 1; \boldsymbol{\varphi}] = \int_{-\infty}^{\infty} \delta\, f(\delta|y_s, \theta_s = 1; \boldsymbol{p}) d\delta, \tag{10}$$

where $f(\delta|y_s, \theta_s = 1; \boldsymbol{p})$ is the posterior probability,

$$f(\delta|y_s, \theta_s = 1; \boldsymbol{p}) = \frac{\phi(y_s; \delta, c_n^2) g(\delta; \boldsymbol{p})}{f_1(y_s; \boldsymbol{p})}, \tag{11}$$

when the normal approximation is employed for the sampling distribution of Y_s. Since the effect size under the null hypothesis is zero, the posterior mean of the effect size of the voxel s is given by

$$E[\delta_s|y; \boldsymbol{\varphi}] = E[\delta_s|y_s, \Theta_s = 1; \boldsymbol{\varphi}] \Pr(\Theta_s = 1|y; \boldsymbol{\varphi}).$$

Based on these formulas, we can then estimate the effect size using the following posterior indices,

$$\hat{\delta}_s = d_s \ell_s, \tag{12}$$

where d_s and ℓ_s are plug-in estimators, $d_s = E[\delta_s | y_s, \Theta_s = 1; \widehat{\boldsymbol{\varphi}}]$ and $\ell_s = \Pr(\Theta_s = 1 | \boldsymbol{y}; \widehat{\boldsymbol{\varphi}}) = 1 - \widehat{\mathrm{LIS}}_s(\boldsymbol{y})$. Based on equations (10) and (11), we have the following form for the estimator d_s,

$$d_s = \sum_{b=1}^{B} t_b \, \hat{p}_b \phi(y_s; t_b, c_n^2) / \sum_{b=1}^{B} \hat{p}_b \, \phi(y_s; t_b, c_n^2).$$

This posterior mean d_s is the shrinkage estimate of effect sizes, given that the voxel is non-null. The probability ℓ_s depends on the multiple testing index $\widehat{\mathrm{LIS}}_s(\boldsymbol{y})$, which incorporates spatial dependency and is calculated using the Gibbs sampler from the distribution of $\boldsymbol{\Theta}|\boldsymbol{Y}$, presented in equation (9). These two posterior indices adjust for two different errors. The first is overestimation of effect sizes, and the shrinkage estimate d_s is used to adjust this bias. The second is incorrect selection of the null voxel, and ℓ_s is used to correct for this error. Again, if the sample size is not large enough, the *t*-distribution $\phi_t(y/c_n; n-2, t_b/c_n)$ is substituted for the normal distribution $\phi(y; t_b, c_n^2)$.

2.5. Incorporating Additional Covariates

We shall now address adjustment for additional subject-level covariates (other than the disease status) by employing general linear models. For each voxel, we first standardize all the intensity values across subjects based on the common within-group variance ($\hat{\sigma}_s^2$) such that the within-group variance equal to 1. Let $x_{i,s}$ be the standardized intensity value of voxel s on subject i ($s \in S, i = 1, \ldots, n$). We then assume a general linear model for $x_{i,s}$ as the observed intensity values for voxel s,

$$x_{i,s} = \beta_{0,s} + \beta_{1,s} w_{i,1} + \cdots + \beta_{p,s} w_{i,p} + \varepsilon_{i,s}, \quad i = 1, \ldots, n, \tag{13}$$

where $w_{i,1}$ is the binary variable on disease status, $w_{i,2}, \ldots, w_{i,p}$ represents the additional covariates on subject i, and $\varepsilon_{i,s}$ is an error term. As an estimate of the effect size for voxel s (with adjustment for the additional covariates), we use $Y_{s,\mathrm{adj}} = \hat{\beta}_{1,s}$ with the variance $\widehat{\mathrm{Var}}(\hat{\beta}_{1,s})$ (in the first level of the hierarchical model). When $p = 1$ (no additional covariates), $Y_{s,\mathrm{adj}}$ may reduce to Y_s in equation (1). We approximate that the distribution of $Y_{s,\mathrm{adj}}$ is normal, $N(\beta_{1,s}, (W^{\mathrm{T}}W)^{-1}_{\{22\}})$, where $W = (w_1^{\mathrm{T}}, \ldots, w_n^{\mathrm{T}})^{\mathrm{T}}$ and $w_i = (1, w_{i,1}, \ldots, w_{i,p})$, and $(W^{\mathrm{T}}W)^{-1}_{\{22\}}$ represents the (2, 2) entry of the inverse matrix $W^{\mathrm{T}}W$. If the sample size is not large enough, we assume $Y_{s,\mathrm{adj}}/(W^{\mathrm{T}}W)^{-1}_{\{22\}} \sim t_{\nu,\delta}$ where $\nu = n - p - 1$ and $\delta = \beta_{1,s}/(W^{\mathrm{T}}W)^{-1}_{\{22\}}$.

3. Results

3.1. Simulation Experiments

We conducted simulation experiments to evaluate the performance of effect size estimation in the proposed method. We simulated the values of the summary statistic Y_s according to the hierarchical mixture model in a hidden Markov random field, as given in Section 2.1. With this simulation, we supposed implementation of appropriate preprocessing normalization procedures for various neuroimaging analysis platforms and devices to obtain normally distributed intensity data across subjects for individual voxels. We considered a simple situation where disease and normal control subjects were compared with no additional covariates. The numbers of disease and normal control subjects, n_1 and n_2, were set as $n_1 = n_2 = n/2$. We specified the total number of subjects n as $50,100$, or 200. Further, we specified the number of voxels m as 3375 $(= 15 \times 15 \times 15)$, which was the number of voxels per sub-region defined based on brain parcellation in effect size estimation within sub-region (see the application in Section 3.2). We generated the true latent variables θ from an Ising model with parameter values $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^{\mathrm{T}}$. We considered that the parameter values $\gamma_1 = 0.05, 0.15$, and 0.25 represented weak, intermediate, and strong degrees of dependency across voxels, respectively. Another parameter, γ_2, was determined such that the proportion of disease-associated voxels accounted for $10\%, 20\%$, and 50% of all the voxels. When $\theta_s = 0$ (i.e., the voxel s was not associated with the disease status), the true effect size was set as $\delta_s = 0$; otherwise, the true effect size was set as $\delta_s \neq 0$ and generated from $N(0.3, 0.1^2)$. Here it is reasonable to assume positive effects only (i.e., one-sided detection) when studying the loss of neurological function after disease onset. The statistics Y_s were generated from a *t*-distribution, $Y_s/c_n|\delta_s \sim t_{n-2,\delta_s/c_n}$.

For simulated data for m voxels, we applied a counterpart of the proposed estimation method with normal approximation for the sampling distribution of Y_s (given the true effect size for voxel s), and also a method assuming a *t*-distribution without normal approximation for the sampling distribution of Y_s (see Section 2). To reduce the computational burden when performing the proposed methods, we assumed that the parameters in the Ising model were constant. We ascertained similar simulation results for a small number of simulation repetitions when the parameters in the Ising model were estimated (results not shown). Following the guideline on the smoothing-by-roughening method [14], we used $B = 200$ in these simulation experiments. We also ascertained similar results in estimating effect sizes of individual voxels when we used a smaller number $B = 20$ (results not shown), indicating that the estimation is relatively insensitive to the selection of B.

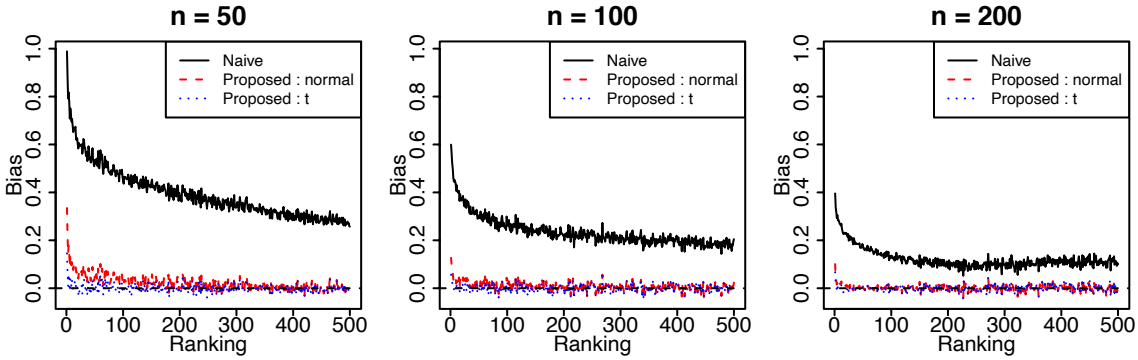
In evaluating the proposed method's performance regarding effect size estimation, estimation biases for voxels with the greatest statistical significance (i.e., greatest values of Y_s) were compared between the naive estimator $\tilde{\delta}_s = Y_s$ and the proposed estimators. We conducted 100 simulations for each configuration of the parameter values in the Ising model and the total sample size. Figure 1 plots average bias values, each defined as the estimate minus the true value of effect size, over 100 simulations at each voxel ranking for the naive estimator and the two counterparts of the proposed posterior mean in Equation (12), for the case in which the proportion of disease-associated voxels was 20% of all the voxels. Note that the top-ranked voxels differed across the 100 simulated datasets, but the three estimates pertained to the same voxels (based on the ranking based on Ys) for each simulated dataset. We also note that we had similar results for the other proportions of disease-associated voxels, i.e., 10% and 50% (see Appendix B).

From Figure 1, we can see that naive estimators suffered from serious overestimation. The proposed estimators were generally less biased. Moreover, we can see that the counterpart of

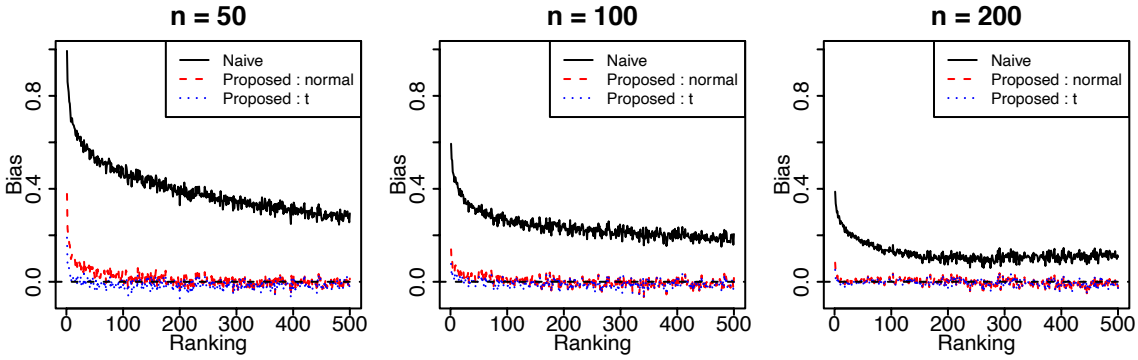
the proposed method, based on a t-distribution, generally gave less biased estimates for $n = 50$ and 100 compared with the method based on normal distribution.

We also evaluated the performance in effect size estimation for two scenarios where the model was misspecified. Specifically, for Scenario 1, the true latent variables θ were generated independently across voxels as in Brown et al (2014) [6], but the true effect sizes were smoothed with a Gaussian kernel after initial effect sizes were independently generated from $N(0.3, 0.1^2)$ across voxels. For Scenario 2, the true effect sizes were smoothed with a Gaussian kernel as in Scenario 1, but the true latent variables θ were generated from an Ising model to reflect special dependency. We ascertained similar performance in effect size estimation for these two scenarios where the model was misspecified.

(a) Strong dependency: $\boldsymbol{\gamma} = (0.25, -0.01)$



(b) Intermediate dependenc $\boldsymbol{\gamma} = (0.15, -0.01)$



(c) Week dependency: $\boldsymbol{\gamma} = (0.05, -0.01)$

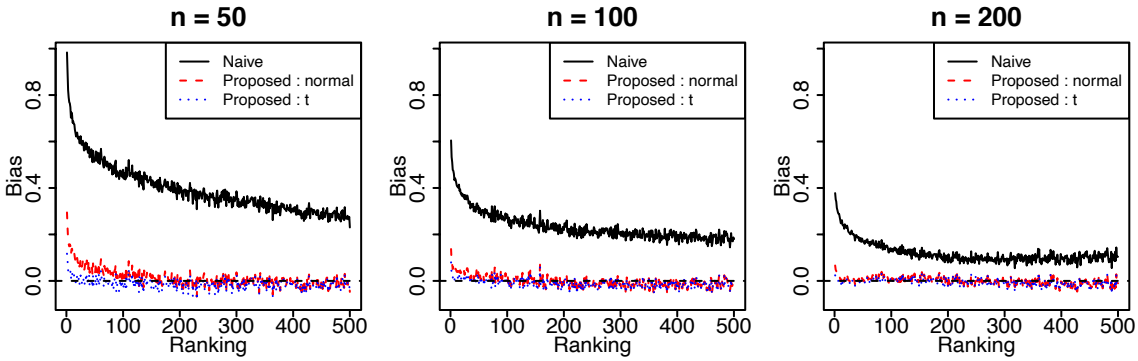


Figure 1: Average bias in estimating effect sizes for each of the top 500 voxels across 100 simulations when the sample size n is 50 (left), 100 (center), and 200 (right). Panels (a), (b), and (c) represent scenarios with

various degrees of dependency among contiguous voxels specified by the parameter γ of the Ising model when the proportion of disease-associated voxels is 20%.

3.2. Application

Alzheimer's disease (AD) is one of the most common neurodegenerative disorders responsible for dementia with brain atrophy. We illustrated our method using a dataset on T1-weighted MRI images from the Open Access Series of Imaging Studies (OASIS), including longitudinal MRI measurements from 150 subjects aged 60 to 96 years (website: https://www.oasis-brains.org/; dataset: "OASIS-2") [20]. Each subject underwent MRI scans using the same scanner with identical sequences at two or more visits with intervals of at least one year. At each subject visit, three or four individual T1-weighted MRI images were obtained during a single imaging session, and the Clinical Dementia Rating (CDR) scale was administered. Here, we evaluated whether assessment of brain sub-regions at the first visit (baseline) could be used for early diagnosis of AD, by associating the baseline MRI measurements with the conversion from mild cognitive impairment (MCI) at baseline to AD at the second visit, where MCI was defined as CDR $= 0.5$ and AD was defined as CDR ≥ 1. Specifically, in the original dataset we identified $n = 51$ MCI subjects (with CDR $= 0.5$) at baseline; of those 51, at the second visit there were $n_1 = 38$ non-converters with CDR $= 0.5$ and $n_2 = 13$ converters with CDR ≥ 1. Of note, $n_2 = 13$ converters were diagnosed as CDR $= 1$ at the second visit within 2 years after the baseline visit. We thus compared baseline MRI data between the non-converter and converter groups.

The baseline MRI data were obtained as follows. In order to make the subject-specific MRI data comparable in assessing brain atrophy at each coordinate across subjects, we utilized the SPM software (https://www.fil.ion.ucl.ac.uk/spm/) to obtain a $91 \times 109 \times 91$ voxel image grid with 2-mm cubic voxels for each subject. Specifically, three or four individual scan images were obtained during single imaging sessions at baseline for each subject and were then co-registered (to make them comparable across each subject's scan images), and image intensity values at respective coordinates were averaged across scan images. The software was then used to achieve the following: segmenting the images into different tissue classes; co-registration of segmented gray and white matter (to make the averaged images comparable among subjects) using the algorithm Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL, [21]); normalization to a standard brain space (MNI-space, developed by Montreal Neurological Institute); modulation of the transformation of intensity values of gray and white matter images into the tissue volume for each coordinate; and smoothing across contiguous voxels based on an 8-mm cube of full-width at half maximum of the Gaussian blurring kernel. After the processing by SPM, gray matter intensity normalization was performed based on white matter intensity using R package WhiteStripe [22] to obtain comparable images across subjects. See Appendix F for more details of the aforementioned processes used to transform the original raw data to normalized data eligible for association analysis using the proposed method.

In the association analysis after the preprocessing of MRI data, the summary statistic $Y_{s,\mathrm{adj}}$ in Section 2.5 was calculated from a *t*-statistic for testing $\beta_{1,s} = 0$ in the general linear model in Equation (13) with the gray matter intensity as the dependent variable and sex, age, and total intracranial volume as covariates. Owing to plausible heterogeneity in voxel intensity across brain regions, we divided the whole brain image into 116 sub-regions based on the AAL, and fit the model for each sub-region separately. Of note, we can consider brain sub-regions other than those based on AAL. We then obtained the effect size estimate $\hat{\delta}_s$ in Equation (12) and

the LIS statistic $\widehat{\mathrm{LIS}}_s(\boldsymbol{y})$ in Section 2.4 for individual voxels based on the estimated model within each sub-region. We used $B = 200$ as the number of mass points used to estimate the effect size distribution g. We also used a smaller number, $B = 20$, for some sub-regions with small sizes, but obtained similar results for $\hat{\delta}_s$ and $\widehat{\mathrm{LIS}}_s(\boldsymbol{y})$. We detected disease-associated voxels at FDR $= 5\%$ by applying the pooled LIS procedure [19], where all the LIS values were pooled across sub-regions and ordered to determine rejection of voxels based on the criterion in Equation (8).

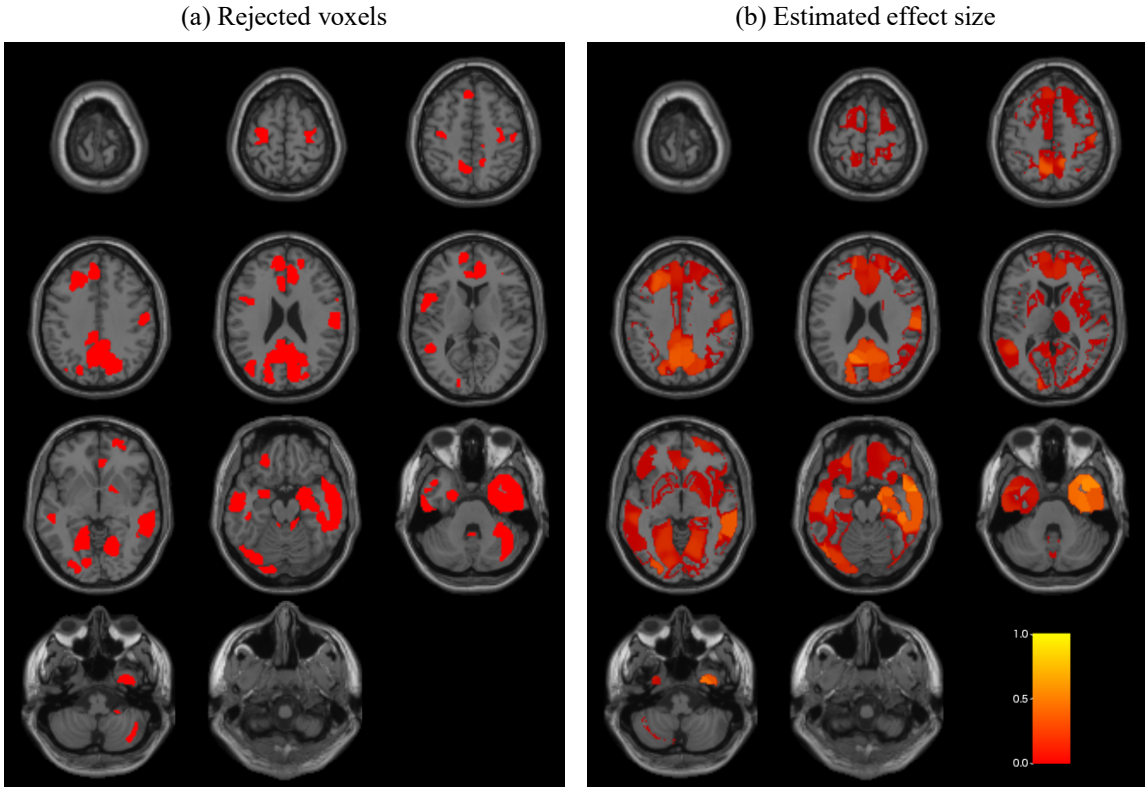


Figure 2: Application to Alzheimer's disease. Panel (a) displays rejected voxels for the nominal FDR level of 0.05. Panel (b) displays positive effect size estimates.

Since the total sample size $n = 51$ was relatively small, we provide the estimation results based on the proposed method with t-distribution for the sampling distribution of Y_s (see Appendix D for results based on the proposed method with normal sampling distribution). Figures 2(a) and (b) display significant voxels at FDR $= 5\%$ by the pooled LIS procedure and all positive effect size estimates $\hat{\delta}_s$ in Equation (12), based on the region-specific estimated models. We note that there were few voxels with negative effects; this is reasonable because brain atrophy should be linked to positive effects. In comparison with Figure 2(a), Figure 2(b) on effect size estimation apparently provides more information about the variation in the strength in disease association. As a reference, we also fit the counterpart of the proposed method based on normal distribution, but similar results were obtained (Appendix D).

For each sub-region, we then calculated average effect sizes for significant voxels based on the proposed method with t-distribution. Table 1 shows 10 sub-regions with the greatest average effect sizes. As expected, the effect size estimates based on proposed method were generally smaller than those based on the naive estimation method for top voxels. See

Appendix E for the differences in effect size estimates for top voxels within sub-region between the proposed and naive methods. The top sub-region, corresponding to the right middle temporal pole (TPOmid.R), has been reported by a connectivity analysis to be a region in which converters exhibited a decreased short-range degree of functional connectivity [23]. The other regions have already been associated with conversion to Alzheimer's disease. For example, the left medial occipital lobe including the left cuneus (CUN.L) has been reported to be associated with MCI conversion [24], and the fusiform gyrus (including FFG.R) and parahippocampal gyrus (including PHG.R) have been reported as the regions with reduced volume in converters [25]. The right anterior portion of the parahippocampal gyrus (part of PHG.R) and left precuneus (PCUN.L) have been used to predict conversion [26]. The amygdala (including AMYG.R) has been used as a predictor of conversion from MCI to AD in many studies [27, 28, 29]. The middle and inferior temporal gyri (including MTG.R and ITG.R) have been reported as the regions with reduced volume in converters [30]. Hypometabolism in the inferior parietal lobe (including SMG.R) has been used as a predictor of cognitive decline from MCI to AD dementia [31]. Although the right superior temporal pole (TPOsup.R) has not been examined in association studies based on the AAL, the temporal pole has been reported to be associated with disease conversion [32].

Table 1: List of the top 10 atlases with the greatest effect size estimates.

Index	Name	Number of voxels	Number of rejected voxels	Proportion rejected	Average of proposed effect size estimate for rejected voxels
88	TPOmid.R	581	577	99.3%	0.540
84	TPOsup.R	743	502	67.6%	0.464
45	CUN.L	939	158	16.8%	0.450
56	FFG.R	2327	708	30.4%	0.443
40	PHG.R	1097	719	65.5%	0.415
42	AMYG.R	248	242	97.6%	0.371
86	MTG.R	2964	1723	58.1%	0.340
67	PCUN.L	2380	1217	51.1%	0.340
90	ITG.R	2368	1597	67.4%	0.339
64	SMG.R	1326	201	15.2%	0.335

4. Discussion

This research was motivated by the growing recognition of the importance of effect size estimation for detected brain areas in disease-association studies using neuroimaging data [7, 8]. In order to permit flexible modelling of effect size distribution across a large number of voxels, while also incorporating the inherent spatial structure among voxels in neuroimaging data, we have integrated the frameworks of semi-parametric hierarchical mixture modelling and hidden Markov random field modelling. The integrated framework allows for more accurate effect size estimation for individual voxels, and also facilitates the accurate estimation of false discovery rates when detecting disease-associated voxels through multiple testing. With this framework, we could assess both voxel-level effect sizes and false discovery rates based on the integrated model without needing additional independent datasets. As shown in Figure 2(b), voxel-level effect size estimates can provide detailed and

unbiased information about the association between detected brain areas and the disease, which may be helpful for biological or clinical analysis of the identified areas. We stress that the effect size index in Equation (1) allows for evaluation without dependency on sample size. This feature may be particularly useful for comparing effect size estimates across different studies with distinct sample sizes. Note that our proposed framework is generally applicable to many neuroimaging analyses where general linear models have been employed.

Although we have supposed a particular effect size statistic, i.e., the standardized mean difference between two groups as in Equation (1), and its sampling distributions, i.e., the normal or *t*-distributions as in Equations (4) and (5), we can consider another effect size statistic and its sampling distribution. With specification of the appropriate effect size statistic and its sampling distribution, our method is widely applicable to many neuroimaging association studies where general linear models have been employed, such as those with fMRI/sMRI, DTI, and so forth. Related to this point, we can accommodate unequal variances between diseased and healthy brain images, rather than equal variance represented in Equation (1). Specifically, we may define the fold change, $\bar{\mu}_{1s} - \bar{\mu}_{2s}$, as the effect size estimate, and assume asymptotic normality with fixed variances specified using reasonable estimators of the group-specific variances, although in our original formulation equal variance could be achieved by an adjustment for appropriate covariates in the framework of general linear models (see Section 2.5). Similarly, in fMRI analyses an absolute effect size such as percent signal change can be evaluated, and asymptotic normality is assumed for the sampling distribution (see Desmond and Glover (2002) [33] for the specification of the asymptotic variance).

We have proposed two counterparts of the proposed method; one uses normal approximation and the other is based on t-distribution for the sampling distribution of the voxel-level summary statistic Y_s (or $Y_{s,\mathrm{adj}}$), for both null and non-null voxels (see Section 2.2). Our simulation experiments demonstrated that the proposed method with normal approximation could substantially overestimate voxel-level effect sizes when the sample size was small ($n = 50$), due to the erroneous assumption of a smaller dispersion of the sampling distribution of the statistic Y_s (or $Y_{s,\mathrm{adj}}$) for both null and non-null voxels, such that greater mass probabilities would be assigned for large effect sizes in estimating the effect size distribution g. However, this problem disappears as the sample size becomes large, as demonstrated in our simulations. One advantage of the proposed method with normal approximation is shorter computational time for model estimation, compared with the counterpart with *t*-distribution and heavier tails. We recommend using the proposed method with normal approximation if the sample size is sufficiently large (say, $n > 100$); otherwise, use the its counterpart with *t*-distribution.

As for the specification of the null distribution $f_0(y_s)$ in Equation (3), we have specified the theoretical null, represented by $N(0, c_n^2)$ or central *t*-distribution, with the Ising model to incorporate spatial dependency in the association status across voxels. To accommodate residual dependency, we could assume the empirical null, say $N(\mu, \tau^2)$, and estimate the null parameters using the central matching method that fits an estimated curve $h(y_s)$ for the frequency distribution of y_s, such that we obtain an estimate $\hat{\mu} = \mathrm{argmax}\{h(y_s)\}$ [34]. However, for many neuroimaging data, the central peak may not pertain to a "null" distribution, rather a "non-null" distribution, because moderate to large non-null effects can dominate over small null effects, especially when the estimation is performed within subregion, as seen in our application example in Section 3.2.

With respect to specification of the effect size distribution g, we have employed a flexible, non-parametric specification because the information about the distributional form of g is generally limited in exploratory disease-association studies. Other flexible specifications may include the use of a parametric effect size distribution with several components, such as finite normal mixture models. When this type of model is assumed, the marginal distribution of Y_s may also have a finite normal mixture form when the sampling distribution of Y_s is normal, as in Equation (4). In this case, the model parameters can be estimated using the method described by Shu et al. (2015) [4], where a penalized likelihood is used to avoid an unbounded likelihood function (or non-identifiability of the variances of the individual normal components) and Bayesian information criteria are used for selecting the number of components. However, a fundamental problem with this approach is that it lacks a natural constraint preventing the variance of the particular normal component in the marginal distribution of Y_s from becoming no smaller than the variance of the sampling distribution of Y_s (i.e., c_n^2 in Equation (4)). By contrast, the non-parametric specification incorporates this constraint in principle; each of a large number of mass points corresponds to a "component", as seen in Equation (7), and the variance of the marginal distribution corresponding to each component is specified as the variance of the sampling distribution (c_n^2). In addition, the non-parametric specification does not need a penalized likelihood maximization or repeated model fitting to select the number of components based on a model selection criterion, and thus the computational burden is much lower.

Our method with a non-parametric effect size distribution, in principle, can capture any forms of the effect size distribution, and voxel-level effect sizes will be estimated based on the fitted effect size distribution. In practice, however, it is reasonable to consider estimation within sub-regions (e.g., those based on the AAL in Section 3.2) to take account of a large heterogeneity in the effect size distribution across sub-regions or to avoid influence of the heterogeneity on the estimation of voxel-level effect sizes in a particular sub-region. Although our model could be extended to incorporate the heterogeneity, e.g., by introducing a hidden structure on the effect size distribution across sub-regions, estimation results may become difficult to interpret. We therefore simply recommend sub-region analysis based on biologically relevant and interpretable brain parcellations in which effect sizes within sub-region are deemed relatively homogeneous.

One inherent feature of the Ising model is that there is a critical value for the spatial interaction term γ_1, beyond which the model has a so-called phase transition, in which almost all binary (null or non-null) indicators will have the same value. Thus the algorithm for estimating $\boldsymbol{\gamma}$ does not converge, while the parameters $\boldsymbol{p}$ in the hierarchical mixture model converge since the plug-in estimate $\widehat{\mathrm{LIS}}_s(\boldsymbol{\gamma})$ assumes values close to 0 or 1 in such a situation. In implementing our algorithm, for the samples of $\boldsymbol{\Theta}$ under candidate new values of $\boldsymbol{\gamma}$, we reject the values of $\boldsymbol{\gamma}$ if all the samples of $\boldsymbol{\Theta}$ are equal. Details of the algorithm and its implementation, including specification of the number of iterations, are provided in Appendix A.

It is interesting to discuss different approaches to modelling the association status (null/non-null) and effect size distribution. Brown et al. (2014) [6] considered a parametric model where the association status and effect size follow a Bernoulli distribution and a conditional normal distribution, respectively, independently across voxels, but the mean of the conditional distribution is a weighted mean or smoothed across adjacent voxels, like the misspecified model investigated in our simulation (see Appendix C). On the other hand, our proposed model incorporates spatial dependency in the association status, but not the effect

size, using the Ising model. Further, for effect sizes, a non-parametric marginal distribution is specified as in Equations (3) or (4). Even under the absence of the specification of dependence in effect sizes across voxels, our method worked well under various simulation models in Section 3.1. This could be explained by the feature of our method that it can yield similar effect size estimates for similar values of the observed association statistic Y from relatively adjacent voxels. However, integration of different modelling approaches for more efficient estimation is an interesting area for future study.

Lastly, another important aspect of the proposed framework for disease-association studies with neuroimaging data is that it can provide a flexible statistical model for the distribution of all neuroimaging data with a large number of voxels. Based on such a whole-brain, voxel-based model, it is appropriate to make a formal inference for a particular group of brain areas or contiguous voxels. In addition, power and sample size calculations of disease-association studies involving neuroimaging are another important direction based on whole-brain modelling.

5. Conclusions

The proposed method allows for accurate estimation of voxel-level effect sizes, as well as detection of significant voxels with disease association, based on the flexible, hierarchical semi-parametric model incorporating spatial dependency across voxels. Our method can be generally applicable for many neuroimaging disease-association studies where general linear models can be assumed for voxel-level intensity values.

Data Availability

This research uses a publicly available dataset “OASIS-2: Longitudinal MRI Data in Nondemented and Demented Older Adults” available at: https://www.oasis-brains.org/ [20]. The codes used in this research are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by a Grant-in-Aid for Scientific Research (16H06299) and JST-CREST (JPMJCR1412) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. For the Alzheimer’s disease data in Section 3.2, we appreciate the Open Access Series of Imaging Studies (OASIS) supported under grants P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, and R01 MH56584.

Supplementary Materials

Appendix A (referenced in Section 2.2) shows details of generalized EM algorithm for parameter estimation. Appendices B and C (referenced in Section 3.1) show the simulation results for other simulation settings. Appendices D and E (referenced in Section 3.2) show the results of the proposed method with normal approximation. Appendix F (referenced in Section 3.2) show the details of the preprocess conducted in application of proposed method.

References

[1] R. A. Poldrack, J. A. Mumford, and T. E. Nichols, *Handbook of functional MRI data analysis.* Cambridge University Press, 2011.

[2] S. M. Smith and T. E. Nichols, "Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference," *NeuroImage*, vol. 44, no. 1, pp. 83 – 98, 2009.

[3] M. Smith and L. Fahrmeir, "Spatial Bayesian variable selection with application to functional magnetic resonance imaging," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 417–431, 2007.

[4] H. Shu, B. Nan, and R. Koeppe, "Multiple testing for neuroimaging via hidden Markov random field," *Biometrics*, vol. 71, no. 3, pp. 741–750, 2015.

[5] W. Sun and T. T. Cai, "Large-scale multiple testing under dependence," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 71, no. 2, pp. 393–424, 2009.

[6] D. A. Brown, N. A. Lazar, G. S. Datta, W. Jang, and J. E. McDowell, "Incorporating spatial dependence into Bayesian multiple testing of statistical parametric maps in functional neuroimaging," *NeuroImage*, vol. 84, pp. 97–112, 2014.

[7] M. C. Reddan, M. A. Lindquist, and T. D. Wager, "Effect size estimation in neuroimaging," *JAMA Psychiatry*, vol. 74, no. 3, pp. 207–208, 2017.

[8] M. A. Lindquist and A. Mejia, "Zen and the art of multiple comparisons," *Psychosomatic Medicine*, vol. 77, no. 2, pp. 114–125, 2015.

[9] B. Efron, "Empirical Bayes estimates for large-scale prediction problems," *Journal of the American Statistical Association*, vol. 104, no. 487, pp. 1015–1028, 2009.

[10] S. Matsui and H. Noma, "Estimating effect sizes of differentially expressed genes for power and sample-size assessments in microarray experiments," *Biometrics*, vol. 67, no. 4, pp. 1225–1235, 2011.

[11] B. Efron, "Microarrays, empirical bayes and the two-groups model," *Statistical Science*, vol. 23, no. 1, pp. 1–22, 2008.

[12] L. D. Brown, "Admissible estimators, recurrent diffusions, and insoluble boundary value problems," *The Annals of Mathematical Statistics,* vol. 42, no. 3, pp. 855–903, 1971.

[13] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[14] W. Shen and T. A. Louis, "Empirical Bayes estimation via the smoothing by roughening approach," *Journal of Computational and Graphical Statistics*, vol. 8, no. 4, pp. 800–823, 1999.

[15] S. Matsui and H. Noma, "Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics," *Biostatistics*, vol. 12, no. 2, pp. 223– 233, 2011.

[16] P. Mu¨ller and R. Mitra, "Bayesian nonparametric inference–why and how," *Bayesian analysis*, vol. 8, no. 2, 2013.

[17] B. Efron, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Institute of Mathematical Statistics Monographs, Leiden: Cambridge University Press, 2010.

[18] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated Anatomical Labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273 – 289, 2002.

[19] Z. Wei, W. Sun, K. Wang, and H. Hakonarson, "Multiple testing in genome-wide association studies via hidden Markov models," *Bioinformatics*, vol. 25, no. 21, pp. 2802–2808, 2009.

[20] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies: Longitudinal MRI data in nondemented and demented older adults," *Journal of Cognitive Neuroscience*, vol. 22, no. 12, pp. 2677–2684, 2010.

[21] J. Ashburner, "A fast diffeomorphic image registration algorithm," *NeuroImage*, vol. 38, no. 1, pp. 95 – 113, 2007.

[22] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu, et al., "Statistical normalization techniques for magnetic resonance imaging," *NeuroImage: Clinical*, vol. 6, pp. 9–19, 2014.

[23] Y. Deng, K. Liu, L. Shi, Y. Lei, P. Liang, K. Li, W. C. W. Chu, and D. Wang for the Alzheimer's Disease Neuroimaging Initiative, "Identifying the alteration patterns of brain functional connectivity in progressive mild cognitive impairment patients: A longitudinal whole-brain voxel-wise degree analysis," *Frontiers in Aging Neuroscience*, vol. 8, p. 195, 2016.

[24] P. J. Modrego, N. Fayed, and M. Sarasa, "Magnetic resonance spectroscopy in the prediction of early conversion from amnestic mild cognitive impairment to dementia: a prospective cohort study," *BMJ open*, vol. 1, no. 1, 2011.

[25] E. Pravatà, J. Tavernier, R. Parker, H. Vavro, J. E. Mintzer, and M. V. Spampinato, "The neural correlates of anomia in the conversion from mild cognitive impairment to Alzheimer's disease," *Neuroradiology*, vol. 58, no. 1, pp. 59–67, 2016.

[26] S. F. Eskildsen, P. Coupé, V. S. Fonov, J. C. Pruessner, and D. L. Collins, "Structural imaging biomarkers of Alzheimer's disease: predicting disease progression," *Neurobiology of Aging*, vol. 36, pp. S23–S31, 2015.

[27] Y. Liu, T. Paajanen, Y. Zhang, E. Westman, L.-O. Wahlund, A. Simmons, C. Tunnard, T. Sobow, P. Mecocci, M. Tsolaki, B. Vellas, S. Muehlboeck, A. Evans, C. Spenger, S. Lovestone, and H. Soininen, "Analysis of regional MRI volumes and thicknesses as predictors of conversion from mild cognitive impairment to Alzheimer's disease," *Neurobiology of Aging*, vol. 31, no. 8, pp. 1375–1385, 2010.

[28] X. Tang, D. Holland, A. M. Dale, L. Younes, M. I. Miller, and A. D. N. Initiative, "Shape abnormalities of subcortical and ventricular structures in mild cognitive impairment and Alzheimer's disease: detecting, quantifying, and predicting," *Human Brain Mapping*, vol. 35, no. 8, pp. 3701–3725, 2014.

[29] H.-A. Yi, C. Möller, N. Dieleman, F. H. Bouwman, F. Barkhof, P. Scheltens, W. M. van der Flier, and H. Vrenken, "Relation between subcortical grey matter atrophy and conversion from mild cognitive impairment to Alzheimer's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 87, no. 4, pp. 425–432, 2016.

[30] G. Karas, J. Sluimer, R. Goekoop, W. van der Flier, S. Rombouts, H. Vrenken, P. Scheltens, N. Fox, and F. Barkhof, "Amnestic mild cognitive impairment: Structural MR imaging findings predictive of conversion to Alzheimer disease," *American Journal of Neuroradiology*, vol. 29, no. 5, pp. 944–949, 2008.

[31] T. Kato, Y. Inui, A. Nakamura, and K. Ito, "Brain fluorodeoxyglucose (FDG) PET in dementia," *Ageing research reviews*, vol. 30, pp. 73–84, 2016.

[32] C. Salvatore, A. Cerasa, and I. Castiglioni, "MRI characterizes the progressive course of AD and predicts conversion to Alzheimer's dementia 24 months before probable diagnosis," *Frontiers in Aging Neuroscience*, vol. 10, p. 135, 2018.

[33] J. E. Desmond and G. H. Glover, "Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses," *Journal of Neuroscience Methods*, vol. 118, no. 2, pp. 115 – 128, 2002.

[34] B. Efron, "Large-scale simultaneous hypothesis testing: the choice of a null hypoth- esis," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 96–104, 2004.