5-13-2009

COVARIATE-ADJUSTED NONPARAMETRIC ANALYSIS OF MAGNETIC RESONANCE IMAGES USING MARKOV CHAIN MONTE CARLO

Haley Hedlin
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, hhedlin@jhsph.edu

Brian S. Caffo
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Ziyad Mahfoud
American University of Beirut, Department of Epidemiology and Population Health

Susan Spear Bassett
Johns Hopkins University School of Medicine, Psychiatric Neuroimaging

Covariate-adjusted Nonparametric Analysis of Magnetic Resonance Images using Markov Chain Monte Carlo

Haley Hedlin, Brian Caffo, Ziyad Mahfoud and Susan Spear Bassett

Abstract

Permutation tests are useful for drawing inferences from imaging data because of their flexibility and ability to capture features of the brain that are difficult to capture parametrically. However, most implementations of permutation tests ignore important confounding covariates. To employ covariate control in a nonparametric setting we have developed a Markov chain Monte Carlo (MCMC) algorithm for conditional permutation testing using propensity scores. We present the first use of this methodology for imaging data. Our MCMC algorithm is an extension of algorithms developed to approximate exact conditional probabilities in contingency tables, logit, and log-linear models. An application of our nonparametric method to remove potential bias due to the observed covariates is presented.

1 Introduction

In this paper we introduce a methodology to identify differences in brain structure or function between two groups. The method we propose uses permutation tests to detect differences in brain structure and function as measured by imaging data. Permutation testing is widely used in the neuroimaging community because of its flexibility and ability to capture features of the observed data that would be difficult to capture parametrically and its ability to account for complex covariance structures (Hayasaka and Nichols, 2003; Nichols and Holmes, 2002; Bullmore et al., 1999; Arndt et al., 1996; Holmes et al., 1996). The methodology we propose extends the permutation tests used in the neuroimaging field by conditioning on the observed covariates using a propensity score model. The theory behind the approach draws upon the conditional permutation test proposed by Rosenbaum (1984), exact conditional tests, and propensity scores as developed by Rosenbaum and Rubin (1983). Our manuscript represents the first application of conditional permutation methods using propensity scores for neuroimaging data. Further, we develop an adaptation of the Diaconis/Sturmfels algorithm for sampling conditional permutations.

Our goal is to compare images across two groups and identify localized regions of group differences that warrant further research, without a priori selection of regions of interest. Our methods apply generally, wherever two-group comparisons are of interest. For example they could apply to contrast maps from a functional magnetic resonance imaging (fMRI) study, volumetric maps from a voxel-based morphometry study, registered diffusion imaging summaries, such as fractional anisotropy, tracer images from a positron emission computed tomography study and so on. In such studies permutation testing is common practice. However, imbalance in treatment assignment for confounding variables is a common problem.

Hence, to remove potential bias due to the observed covariates we devise a method to control for them (Bross, 1964; Gail et al., 1988; Edgington, 1995; Kennedy, 1995; Anderson and Legendre, 1999; Rosenbaum, 2002). The most elementary and common approach to accomplish this goal with permutation tests is through a stratified analysis that permutes treatment labels within strata for each category of the covariates. This approach has the drawback of only being applicable for one or a few categorical covariates. When adjusting for more than one covariate, one must stratify by their crossed levels, reducing counts within bins.

An appealing approach creates subclasses defined by binning estimated propensity scores (Rosenbaum and Rubin, 1984) and then permutes treatment labels within those subclasses to make inferences. Here the propensity score is the probability of treatment assignment given the confounding covariates (Rosenbaum and Rubin, 1983). It has been shown that the propensity score is a balancing mechanism for covariate control (Rosenbaum and Rubin, 1983). If treatment assignment was randomized, the propensity score is known, otherwise it must be estimated. Estimation of the propensity score is often achieved with a logit model on treatment assignment. The estimated scores are then the natural scale mean predictions from the model. Propensity scores are used in a variety of ways, including regression adjustment, weighting, stratification and others (Rosenbaum and Rubin, 1984; D'Agostino, 1998; Rosenbaum, 2002). Most germane to our discussion is stratification, where five (or so) bins of the estimated scores are created and used for covariate control.

A benefit of propensity scores is the reduction of a complex covariate space to the single estimated propensity score. Thus, permuting treatment labels within estimated propensity score strata applies more generally than the covariate stratification discussed above, as it can be used for multiple continuous or categorical covariates. Moreover, analysis of estimated propensity scores forces a discussion on the comparability of the groups. In addition, the technique facilitates a discussion of causal interpretations and assumptions. However, under such an analysis, the uncertainty in estimating the propensity score is not taken into account. Also, the propensity score model itself must be correctly specified and does not control for important omitted or uncollected covariates, assumptions that our proposed methodology shares.

We propose to use conditional permutation testing using a propensity score model. This method conditions on the sufficient statistics for a logit propensity score model and permutes treatment labels under this conditional distribution. Thus one does not need actual estimates of the propensity scores, as the parameters in the model drop out via the conditioning. Also, arbitrary strata bins are not necessary. However, it does require a correctly specified logit model on treatment assignment. The logit link function is specifically necessary, being the canonical link function for binary data and yielding closed form minimal sufficient statistics. Furthermore, conditional permutation testing does not apply universally, as the conditional distribution can be uninformative. For example, when conditioning on several continuous covariates, the observed treatment assignment may be the only permutation that satisfies the sufficient statistics. A final complication is computational. Conditional permutation testing is computationally more intensive and intricate than using propensity score strata. However, when applicable and computationally feasible, conditional permutation testing is preferable to using propensity score bins.

This manuscript represents the first use of conditional permutation testing in neuroimaging and hence we view it as a proof of concept. We connect computational methods for exact testing to generating conditional permutations to perform the relevant computing. We use an example dataset containing MR images of subjects identified as having a high familial risk of Alzheimer's disease and a group of controls to illustrate the methodology.

The paper begins with a discussion of the example dataset (Section 2), permutation tests (Section 3), propensity scores (Section 4), Markov chain Monte Carlo (Section 5.1), and cluster-level tests (Section 5.3). In Section 5.2 we describe our proposed algorithm and we present results of its application in Section 6. Finally, we conclude with a discussion in Section 7.

2 Example dataset

The dataset used as an example consists of contrast maps from a verbal paired associates functional MRI task. The groups in question are either at high familial risk for Alzheimer's disease or control. The at-risk group had at least one autopsy confirmed parent and at least one additional affected first degree relative per probable clinical diagnosis. The control subjects had no diagnosed first degree relatives. At the time of imaging, the control and the at-risk subjects were clinically asymptomatic.

The fMRI paradigm included encoding (learning) and recall phases in a blocked paradigm. In the encoding phase, subjects were presented with unrelated word pairs. In the recall phase, subjects were presented the first word and asked to recall the second. This paradigm was chosen as loss of verbal memory is one of the early symptoms of Alzheimer's disease (Bookheimer et al., 2000).

We analyze the contrast map comparing recall blocks to rest. The fMRI time series was smoothed using a Gaussian filter with a 5 mm full width at half maximum, coregistered within a subject and normalized to the Montreal Neurological Institute template. The normalization allows for the comparison of images across subjects in a standard space. The preprocessing methods are described further in Bassett et al. (2006). The design matrix was convolved with the default haemodynamic response function in SPM (Friston et al., 2007). The fMRI time series was regressed on the design matrix voxel-by-voxel to obtain contrast maps which were retained for inter-subject group-level analysis. In this manuscript we only consider the recall versus rest contrast.

The images focus on a coronal band encompassing the medial temporal lobe and surrounding structures, where group differences were hypothesized to exist. Each contrast map is a 79x95x68 array of 2mm^3 voxels in template space conceptually representing the change in regional cerebral blood flow between the recall and rest conditions.

The example dataset consists of 161 right-handed subjects between the ages of 48 and 83. We consider two important potential confounders, the apolipoprotein E (APOE) gene and gender. Presence of $\epsilon 4$ alleles of the APOE gene has been linked with the risk of Alzheimer's disease (Corder et al., 1993; Saunders et al., 1993; Strittmatter et al., 1993). Gender is a potential confounder, as females have been shown to have a higher incidence of Alzheimer's disease (see Gao et al., 1998, for example). However, gender is fairly balanced between the groups, so its inclusion is primarily to illustrate the algorithm. Age, another important confounder for risk status, was not necessary to add as a confounder, as the two groups had very similar age distributions. Figure 1 shows the cross-classification of at-risk status with gender and $\epsilon 4$ status. Eighty five are at-risk for late-onset Alzheimer's disease, 75 of the subjects are male and 46 have at least one $\epsilon 4$ allele of the APOE gene.

	Control		At-risk	
	No $\epsilon 4$	At least 1 $\epsilon 4$	No $\epsilon 4$	At least 1 $\epsilon 4$
Female	31	7	32	16
Male	31	7	21	16

Table 1: Alzheimer's disease data set, cross-classification of at-risk status with gender and $\epsilon 4$ status.

3 Permutation tests

Below we motivate permutation testing using counterfactual notation as part of the motivation. We note that our example dataset is not ideal for this discussion, as it is difficult to conceptualize the meaning of a subject having a different (counterfactual) group status, a problem that does not exist for assignable treatments. Therefore, we use the generic term "treatment" as the labeling being permuted. However, we propose the use of these methods for covariate control in our non-assignable setting in the same way that propensity scores are frequently used as a balancing mechanism unrelated to any causal discussion.

Consider the possibility that each subject i has an voxel-specific outcome that would be observed if they received a treatment of interest and an outcome that would be observed if they were in a control group, $r_{1i}(v)$ and $r_{0i}(v)$ for voxel v, respectively (see Rosenbaum, 1984; Rubin, 1974, 1977). Only $r_{1i}(v)$ or $r_{0i}(v)$ can be observed, as each subject either received the treatment or control; $r_{1i}(v) - r_{0i}(v)$ cannot be measured directly.

A strong null hypothesis specifies that the observed outcome for a subject does not differ depending on which treatment he or she received, i.e $H_0 : r_{1i}(v) = r_{0i}(v)$ for all voxels v. If this were the case, treatment labels among "similar" subjects are arbitrary. Hence, under the null hypothesis we assume that the labels are exchangeable among subjects with similar covariates (Good, 2006). This assumption justifies permuting the labels while keeping the covariates fixed to create the null distribution from which inferences can be drawn about the mean population difference (Nichols and Holmes, 2002; Pitman, 1937).

We emphasize that, in imaging applications, separate permutation tests are not performed at each voxel. Instead, treatment labels are permuted to images and a map of statistics is created. Performed in this manner, an appealing feature of permutation tests is their ability to capture features and account for correlation without making stringent assumptions (Rabe-Hesketh et al., 1997; Holmes et al., 1996). In addition, permutation tests allow the researcher to pick an image-wide test statistic (Nichols and Holmes, 2002; Bullmore et al., 1999). Notably, in imaging applications this allows researchers to choose statistics operating on the image obtained after calculating voxel-wise statistics, such as supra-threshold cluster sizes (see section 5.3).

The null distribution used in permutation testing is ideally formed by enumerating each of the possible permutations of the data. However, for a large number of observations it is not feasible to enumerate all of the possibilities, even when permuting within crossed levels of covariates. For example, the data presented in Table 1 has 161 subjects and four levels of two crossed dichotomous covariates yielding

$$\binom{63}{31} * \binom{23}{7} * \binom{23}{7} * \binom{52}{31} \approx 10^{43}$$

permutations possible, which is clearly too large to be enumerated. Thus permutation testing is usually performed via Monte Carlo. For ordinary permutation testing, or permuting within levels of strata, this process is trivial. However, using Monte Carlo to generate permutations for conditional permutation testing is more difficult. We propose the use of Markov chain Monte Carlo (MCMC) to generate conditional permutations.

4 Conditional permutation and propensity scores

Consider a comparison of two groups with membership denoted by y_i where $y_i = 1$ if subject i belongs to the group of interest and $y_i = 0$ if subject i belongs to the control group for $i = 1, \ldots, n$. For example, we may be interested in a group receiving a specific treatment, an exposed subset of the population, or a group that has or is at-risk for a certain disease. Suppose we also have a set of observed covariates associated with group membership for which we would like to control. Denote the d-vector of covariates for subject i by $\mathbf{x}_i$. Our aim is to identify significant differences between the group of interest and the control group while removing bias due to the observed covariates.

Data gathered from randomized studies are assumed to have treatment labels that are exchangeable due to the random assignments to treatment groups (Good, 2006). Observational studies lack a randomization mechanism and, as a result, the distribution of covariates may differ between the group of interest and the control group. Rosenbaum and Rubin introduced the propensity score in 1983 to account for such bias in observational studies. The propensity score for subject i (p_i) is defined to be the conditional probability of being a member in the group of interest given the observed covariates, i.e. $p_i = P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i)$. Given p_i, Y_i and $\mathbf{X}_i$ are conditionally independent, i.e. $Y_i \perp \mathbf{X}_i \mid p_i$ (Rosenbaum and Rubin, 1983). Hence, conditioning on propensity scores allows us to control for any underlying bias that may be present in the two groups due to the observed covariates. Unlike randomization, propensity scores do not balance the unobserved covariates, unless the unobserved covariates are strongly correlated with the observed covariates (Rosenbaum and Rubin, 1984).

The strongly ignorable treatment assumption must be satisfied to use propensity scores to make causal inference in observational studies (Rosenbaum, 1984; Rosenbaum and Rubin, 1983). This assumption requires that $(r_{1i}, r_{0i}) \perp Y_i \mid \mathbf{X}_i = \mathbf{x}_i$ and $0 < P(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) < 1$ for all subjects i. This is not the case, however, when the group of interest consists of diseased individuals and the control group is healthy, as is often true in neurological studies. In this situation the propensity scores are used only to balance the covariates and the results no longer have a causal interpretation (Joffe and Rosenbaum, 1999; Rosenbaum and Rubin, 1983).

Propensity scores are used in statistical analyses in various ways. It is common to create subclasses of propensity scores by binning similar p_i's. Subclasses can be used to define the subpopulations in stratified analyses or can be used in conditional permutation tests to define groups of similar covariates among which the group labels can be permuted. For a small number of covariates with only a few levels it is straightforward to bin the propensity scores. For example, if we wish to control for two dichotomous covariates there are four possible propensity scores that can arise, i.e. $p_{00} = P(Y = 1|X_1 = 0, X_2 = 0)$, $p_{01} = P(Y = 1|X_1 = 0, X_2 = 1)$, $p_{10} = P(Y = 1|X_1 = 1, X_2 = 0)$, and $p_{11} = P(Y = 1|X_1 = 1, X_2 = 1)$. If we were to define four subclasses, each corresponding to a propensity score, permuting labels within subclass would be equivalent to permuting labels among individuals with the same covariates. In the example dataset from Section 2 these four subclasses would be females with no $\epsilon4$ alleles, males with no $\epsilon4$ alleles, females with at least one $\epsilon4$ allele, and males with at least one $\epsilon4$ allele.

As the complexity and/or number of covariates increases, the number of propensity scores grows. When the number of propensity scores is large there is no longer a clear choice of cutoff to define the subclasses, resulting in a subjective process that varies between researchers. If the bins are too wide, error will be introduced because individuals in the same subclass may no longer have similar probabilities of $Y_i = 1$. On the other hand, too many bins will reduce the number of observations per bin, which hinders conditional permutation tests or, in the case of stratified analyses, greatly increases the number of subpopulation analyses (D'Agostino, 1998).

Under the assumption that a logit model describes the relationship between $\mathbf{y}$ and $\mathbf{x}$, then the propensity score satisfies:

$$\text{logit}\{p_i\} = \mathbf{x}_i\boldsymbol{\beta}. \tag{1}$$

Furthermore, we assume that the observations are independent, yielding the likelihood

$$P(\mathbf{Y} = \mathbf{y}) = \prod_i P(Y_i = y_i) = \frac{\exp(\mathbf{y}^T\mathbf{x}\boldsymbol{\beta})}{\prod_i \{1 + \exp(\mathbf{x}_i\boldsymbol{\beta})\}}, \tag{2}$$

where $\mathbf{y}$ is the vector of treatment assignments and $\mathbf{x}$ is the matrix of covariates with $\mathbf{x}_i$ for each row. From Equation (2) $\mathbf{x}^T\mathbf{y}$ is a sufficient statistic for β (that is also minimal, see Cox and Snell, 1989). That is, by assuming the logit model we arrive at a relatively simple, closed form minimal sufficient statistic that can be used to derive the conditional distribution (Rosenbaum, 1984; Rosenbaum and Rubin, 1983).

The existence of these minimal sufficient statistics implies that $\mathbf{Y} \perp \mathbf{X} \mid \mathbf{X}^T\mathbf{Y} = \mathbf{s}$. Furthermore,

$$P(\mathbf{Y} = \mathbf{y} \mid \mathbf{X}^T\mathbf{Y} = \mathbf{s}) = 1/|\Gamma| \text{ for } \mathbf{y} \in \Gamma$$

where $\Gamma = \{\mathbf{y} \in \{0,1\}^n | \mathbf{X}^T\mathbf{Y} = \mathbf{s}\}$ is the space of treatment assignments satisfying the sufficient statistics. This is the distribution that conditional permutation testing uses for inference. Notice that if $\mathbf{x}$ contains only an intercept, then the sufficient statistic is equivalent to the total number of treated and controls and the conditional distribution is uniform on this space. Hence in this case, conditional permutation testing reduces to standard permutation testing.

In Section 5, we introduce an algorithm that implicitly creates subclasses by conditioning on the covariates, entirely bypassing the calculation and binning of propensity scores. For algorithmic reasons, described subsequently, we also utilize the information in $\mathbf{x}$ to run the chain. We describe the impact of this choice below.

5 Methods

5.1 Markov chain Monte Carlo

Markov chain Monte Carlo is a method of sampling from a density using Markovian samples. A valid MCMC algorithm starts at an initial point and proceeds to traverse the sample space through a Markov chain where asymptotically the chain reaches its stationary distribution (Robert and Casella, 2004; Gilks et al., 1996; Chib and Greenberg, 1995). Thus, realizations of the chain can be used to estimate features of the stationary distribution (Gilks et al., 1996).

Several properties are necessary to ensure that the Markov chain will appropriately explore its stationary distribution. First, the chain must be aperiodic and irreducible. Aperiodicity is satisfied if there are no deterministic visits to subsets of the chain. Irreducibility is the condition that any state a must be reachable from state b in a finite number of transitions for all states b. Finally, we require that the target distribution is the stationary distribution of the chain. That is, if the chain is started via a simulation from the target distribution, then the marginal distribution of every iterate is also the target distribution.

For simple moment estimators, stationarity of the chain is not required, as consistency and Markov chain central limit theorems can produce valid interval estimates for moments of the stationary density provided standard error estimates (Jones et al., 2006; Jones and Hobert, 2001; Jones, 2004; Hobert et al., 2002). Moreover, standard MCMC practice suggests that subsampling the chain, i.e. only retaining every m^{th} iteration, is wasteful and unnecessary (MacEachern and

Berliner, 1994). However, our problem is unique by MCMC standards. Running of the chain is trivial and billions of samples (treatment assignments) are easy to produce. In contrast, creating the statistical map for each treatment assignment on the collection of images is computationally burdensome. Therefore, a high quality sample of nearly independent permutations is desired. Hence, contrary to standard MCMC practice, we subsample the chain quite heavily. To monitor the chain, we examine trace plots and estimated autocorrelation functions of the statistic of interest evaluated at the subsampled chain.

We use the Metropolis/Hastings algorithm to guarantee the appropriate invariant density for the chain (Chib and Greenberg, 1995; Hastings, 1970). As the desired stationary density is uniform, our Metropolis coin flip accepts the proposed state with probability $\min\left\{1, \frac{P(\mathbf{Y}_c \rightarrow \mathbf{Y}_p)}{P(\mathbf{Y}_p \rightarrow \mathbf{Y}_c)}\right\}$ where $\mathbf{Y}_c$ is the current state of the chain and $\mathbf{Y}_p$ is the proposal.

5.2 MCMC algorithm

Our proposed algorithm applies to any linear predictor with polytomous confounding variables and no interactions in the linear predictor of the logit model on the propensity score. Below, we describe the algorithm in generality then describe it via a specific example with two binary covariates. We first cover existing methods for generating from Γ.

The algorithm we present below grew from theory developed to approximate conditional probabilities in contingency tables, logit and log-linear models. Agresti (1992) surveys these methods and Mehta and Patel (1998) review exact procedures for contingency tables. There are currently several approaches for simulating from conditional distributions for logit and loglinear models (Chen et al., 2005; Caffo and Booth, 2001; Booth and Butler, 1999; McDonald et al., 1999; Smith et al., 1996; Forster et al., 1996). Agresti (1992) gives a historical review of methods for exact inference for situations that do not require Markov chain algorithms. More recently, Caffo and Booth (2003) surveyed algorithms for Monte Carlo conditional inference for logit and log-linear models. In particular, the survey includes Diaconis and Sturmfels (1998) who developed a theory for generating conditional distributions for contingency tables and logit models given the sufficient statistic. Their theory generalizes the random walk algorithm consisting of a series of $+/-$ steps, which is the basis of the algorithm we propose.

Our basic strategy for generating conditional permutations is as follows. First, conditioning on $\mathbf{x}^T\mathbf{y}$ and knowledge of $\mathbf{x}$ fixes the margins of the contingency table obtained by cross-classifying treatment assignment and the confounding variables. Our algorithm first operates on this contingency table. Given a current permutation, we calculate its associated contingency table. Next, we find a new contingency table satisfying the observed margins using the Diaconis/Sturmfels algorithm. We then randomly draw the new permutation from all of the permutations that are consistent with the new contingency table.

To elaborate, let $\mathbf{y}_c$ be the current state (permutation) with associated contingency table $\mathbf{c}_c$. Let $\mathbf{e}_1, \ldots, \mathbf{e}_k$ be the Markov basis from the Diaconis/Sturmfels algorithm. We randomly select an element from the basis to create a new table, say $\mathbf{c}_p = \mathbf{c}_c + \mathbf{e}_j$. If $\mathbf{c}_p$ contains negative entries, the current state is retained and the algorithm moves on to the next iteration. If not, then a proposal permutation is generated from all permutations whose associated contingency table is $\mathbf{c}_p$. Generating such a permutation is exactly generating a permutation from the crossed levels of the covariates with counts given by $\mathbf{c}_p$. The forward probability for the proposal, $\mathbf{y}_p$, satisfies

$$P(\mathbf{y}_c \to \mathbf{y}_p) = P(\mathbf{c}_c \to \mathbf{c}_p)P(\mathbf{y}_p \mid \mathbf{c}_p)$$

and the backward probability is then

$$P(\mathbf{c}_p \to \mathbf{c}_c)P(\mathbf{y}_c \mid \mathbf{c}_c).$$

As $P(\mathbf{c}_p \to \mathbf{c}_c) = P(\mathbf{c}_c \to \mathbf{c}_p)$ the Metropolis coin flip probability is

$$\min\left\{1, \frac{P(\mathbf{y}_c \mid \mathbf{c}_c)}{P(\mathbf{y}_p \mid \mathbf{c}_p)}\right\} = \min\left\{1, \frac{|\mathbf{c}_p|}{|\mathbf{c}_c|}\right\},$$

where $|\mathbf{c}_c|$ and $|\mathbf{c}_p|$ are the number of permutations satisfying that contingency table, respectively.

An example below shows that this is a trivial quantity to calculate.

The algorithm described above creates an aperiodic, irreducible, reversible chain that will converge to the uniform stationary distribution on Γ. Aperiodicity of the chain is clear, we argue for irreducibility below. Pick any two treatment assignments from Γ, say $\mathbf{y}_1$ and $\mathbf{y}_2$. Suppose that these elements have different associated contingency tables. As guaranteed by the Diaconis/Sturmfels algorithm, the chain can move between the two contingency tables. Then, as we randomly generate permutations given the new contingency table, generating $\mathbf{y}_2$ is possible. In the event that $\mathbf{y}_1$ and $\mathbf{y}_2$ have the same associated contingency table, the same argument applies with the caveat that the algorithm must move away and return to the common table in order to generate $\mathbf{y}_2$, an event that has non-zero probability.

This hybrid strategy offers many benefits over simply applying the Diaconis/Sturmfels algorithm directly to the logit propensity score model. First, deriving the Diaconis/Sturmfels Markov chain for moving between contingency tables is much easier to derive than the chain for the underlying logit model. Establishing the set of basic moves (Markov bases) for a given problem has been solved generally by Diaconis and Sturmfels (1998). Unfortunately, their method requires knowledge of computational algebra and often the computational complexity of the algebraic problem rivals that of avoiding MCMC and simply enumerating the space of permutations. However, Markov bases for large classes of log-linear models have been created (Dobra, 2003). These models include all of the contingency tables with fixed margins considered in this manuscript. Moreover, computationally calculating the available permutations satisfying the new contingency table is a trivial problem compared to deriving the chain for the logit model.

The use of $\mathbf{x}$ in the running of the chain does not impact the permutations. That is, the result of the chain are permutations uniformly distributed on Γ regardless of the use of $\mathbf{x}$ in the algorithm. Hence the propensity score interpretation of the permutations remains appropriate.

To facilitate a demonstration of the algorithm we consider our example data with the two dichotomous covariates, $\mathbf{x}_1$ and $\mathbf{x}_2$ and no interaction. Along with our dichotomous group variable $\mathbf{y}$, $\mathbf{x}_1$ and $\mathbf{x}_2$ give rise to a 2x2x2 contingency table. Conditioning on the sufficient statistic from

(a) $\mathbf{e}_1$, the first element of the Markov basis

	$y=0$		$y=1$	
	$x_2=0$	$x_2=1$	$x_2=0$	$x_2=1$
$x_1=0$	$+$	$-$	$-$	$+$
$x_1=1$	$-$	$+$	$+$	$-$

(b) $\mathbf{e}_2$, the second element of the Markov basis

	$y=0$		$y=1$	
	$x_2=0$	$x_2=1$	$x_2=0$	$x_2=1$
$x_1=0$	$-$	$+$	$+$	$-$
$x_1=1$	$+$	$-$	$-$	$+$

Table 2: The two elements in the Markov basis from the Diaconis/Sturmfels algorithm for a 2x2x2 table

Section 4 and $\mathbf{x}^T\mathbf{1}$ in the 2x2x2 contingency table implies fixing the margins (see the Appendix).

Table 2 contains the two elements in the Markov basis for a 2x2x2 table from the Diaconis/Sturmfels algorithm. A coin flip at each iteration determines which of the elements will be added to the current table, $\mathbf{c}_c$, to generate the proposed table, $\mathbf{c}_p$. As an example, consider the initial iteration of the chain. The contingency table associated with the observed treatment labels is given in Table 1, $\mathbf{c}_c$ in this initial iteration. Suppose that $\mathbf{e}_1$ was the element randomly chosen from the Markov basis. The proposed contingency table, $\mathbf{c}_p = \mathbf{c}_c + \mathbf{e}_1$, is given in Table 3. $\mathbf{c}_p$ contains no negative entries, so a Metropolis coin flip sets $\mathbf{c}_p = \mathbf{c}_c$ with probability α. In general,

$$|\mathbf{c}_q| = \begin{pmatrix} n_q^{+00} \\ n_q^{100} \end{pmatrix} * \begin{pmatrix} n_q^{+01} \\ n_q^{101} \end{pmatrix} * \begin{pmatrix} n_q^{+10} \\ n_q^{110} \end{pmatrix} * \begin{pmatrix} n_q^{+11} \\ n_q^{111} \end{pmatrix}$$

for $q = c, p$. It follows that $\alpha = \frac{u}{w}$ for

$$u = \frac{1}{|\mathbf{c}_c|} = \left\{ \begin{pmatrix} 63 \\ 31 \end{pmatrix} * \begin{pmatrix} 23 \\ 7 \end{pmatrix} * \begin{pmatrix} 52 \\ 31 \end{pmatrix} * \begin{pmatrix} 23 \\ 7 \end{pmatrix} \right\}^{-1}$$

and

$$w = \frac{1}{|\mathbf{c}_p|} = \left\{ \binom{63}{32} * \binom{23}{6} * \binom{52}{30} * \binom{23}{8} \right\}^{-1}$$

where the superscripts refer to the cells in the 2x2x2 tables by their (y, x_1, x_2) indices and $+$ denotes sum over a dimension. Suppose that according to the flip, $\mathbf{c}_p$ is accepted. For iterations that will be subsampled, $\mathbf{y}_p$ is randomly chosen from among the $|\mathbf{c}_p|$ permutations associated with $\mathbf{c}_p$. If the iteration is not being subsampled, there is no need to draw $\mathbf{y}_p$ because it would immediately be converted back to its associated contingency table to begin the following iteration.

	Control		At-risk	
	No $\epsilon4$	At least 1 $\epsilon4$	No $\epsilon4$	At least 1 $\epsilon4$
Female	32	6	31	17
Male	30	8	22	15

Table 3: Proposed contingency table in initial iteration.

5.3 Cluster-level tests

We wish to identify differences at the level of several voxels for various reasons. First and foremost, the scientific interest is often of differences at the cluster level as opposed to individual voxel-level or regional differences. Cluster-level inferences are often pursued because the spatial correlation between voxels inherent in MR images can mask any voxel-level differences (Wager et al., 2007). Clusters of voxels, on the other hand, tend to be independent under the null hypothesis (Bullmore et al., 1999). Finally, analyses at the cluster level are more sensitive than regional or global tests (Bullmore et al., 1999; Poline and Mazoyer, 1992).

The choice of which test statistic to use depends on the hypothesis being considered. In the following application, we use the maximum cluster size test statistic to test the null hypothesis of no difference between the at-risk group and the control group. The cluster size has the drawback that strong but small very localized differences are penalized. Another suprathreshold cluster test that combats this problem is the exceedance mass which uses the integral of the cluster above the threshold as its test statistic (Bullmore et al., 1999). Yet, another approach considers the maximum statistic within clusters, i.e. considering cluster height instead of extent. Single threshold tests, which reject the null hypothesis of no difference when any voxel exceeds a threshold (Nichols and Holmes, 2002), also avoid this issue, yet have the drawback of not considering spatial contiguity of significant results.

With hundreds of thousands of voxels in each image, multiplicity presents a problem, regardless of the chosen statistic. Bonferroni corrections are a potential solution, but, as a consequence of ignoring the spatial correlation, are too conservative (Brett et al., 2007). We use the distribution of the maximum suprathreshold cluster size to combat multiplicity. That is, at each permutation, we calculate all contiguous clusters and take the largest. Each individual cluster from the observed treatment assignment is then compared to this distribution. This offers control for the familywise error rate (see Nichols and Holmes, 2002) and can be applied generally, for example to the exceedance mass or peak value testing.

5.4 Application to MRI dataset

We apply our proposed method to the example dataset introduced in Section 2 to test the null hypothesis that there are no differences between the at-risk (AR) group and the control group (CTL) while controling for the gender and APOE $\epsilon 4$ status of the subjects. The goal of the analysis is to locate clusters of voxels where there is evidence of a difference across the two groups within the portion of the brain that was imaged.

Throughout we assume a logit model to characterize the relationship between group membership and the two covariates, gender and APOE $\epsilon 4$ status

$$logitP[AR_i] = \beta_0 + \beta_1 Gender_i + \beta_2 APOE4_i \quad i = 1, 2, \ldots, 161$$

where $Gender_i = 1$ if subject i is male and $APOE4_i = 1$ if subject i has at least one $\epsilon 4$ allele of the APOE gene. By conditioning on the sufficient statistic for β, we control for the effects gender and APOE $\epsilon 4$ status (Mehta and Patel, 1998; Rosenbaum, 1984). Next, we conditionally permute the group labels on the images using the algorithm outlined in Section 5.

We ran the MCMC algorithm for 1,100,000,000 iterations, discarded the first 100,000,000, and kept every 1,000,000$^{\text{th}}$ iteration after the burn-in. At each iteration that we kept, the images were labeled AR or CTL, according to the current permutation. Then the z-statistic of the mean difference between the AR and CTL groups was calculated at every voxel. This process is repeated for each of the 1,000 permutations remaining after the burn-in and subsampling. We visually examined the chain of z-statistics for a few randomly chosen voxels and the autocorrelation at lags up to 500 to check that convergence had been achieved.

The maximum cluster is calculated from each of the 1,000 z-maps to create a null distribution of the test statistic. To determine the suprathreshold cluster test significance, the maximum cluster size of the original data is compared to the null distribution. Clusters were found using hierarchical clustering of voxels in the z-map beyond a threshold of ± 3.10. This threshold was chosen a priori because it corresponds to a probability of 0.001 in each of the tails of a Normal density. Finally, the p-value for a cluster is simply the proportion of permutations with maximum cluster sizes as or more extreme than that obtained from the original observation. If a statistically significant difference is found between the two groups, the voxels in the original image that fall beyond the threshold indicate the areas that would warrant closer inspection in future research.

6 Results

The null distribution of the maximum cluster size calculated from the conditional permutations is displayed as a histogram in Figure 1. In the example dataset there were two significant clusters of voxels, one containing 4 voxels and the other containing 5 voxels. Hence, the maximum cluster size in the z-map is 5 voxels, indicated in Figure 1 by the vertical line. From the null distribution generated by the MCMC algorithm we calculate that the probability of observing data as or more extreme than the example data to be 0.664 under the hypothesis of no difference between the two groups. Traces such as those shown in Figure 2 were visually examined to verify that convergence had been achieved.

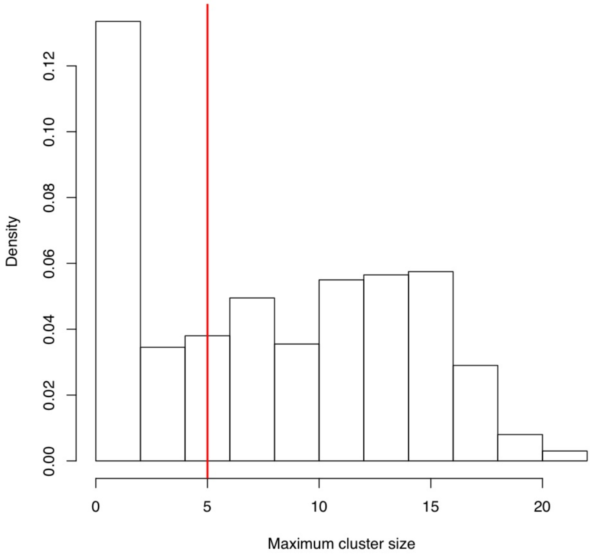

Figure 1: Null distribution created from conditional permutations with the observed maximum cluster size indicated with the vertical red line.

These results differ from an earlier wave of the same data (Bassett et al., 2006), where significant differences between the groups were seen. However, the results for this second wave are confirmed by independent analyses using different methodology. Potential explanations for the discrepant results in the second wave include differential attention to the task, informative dropout, an actual decline in group differences and so on. However, we relegate a full scientific explanation of the longitudinal differences in this study set to other work.

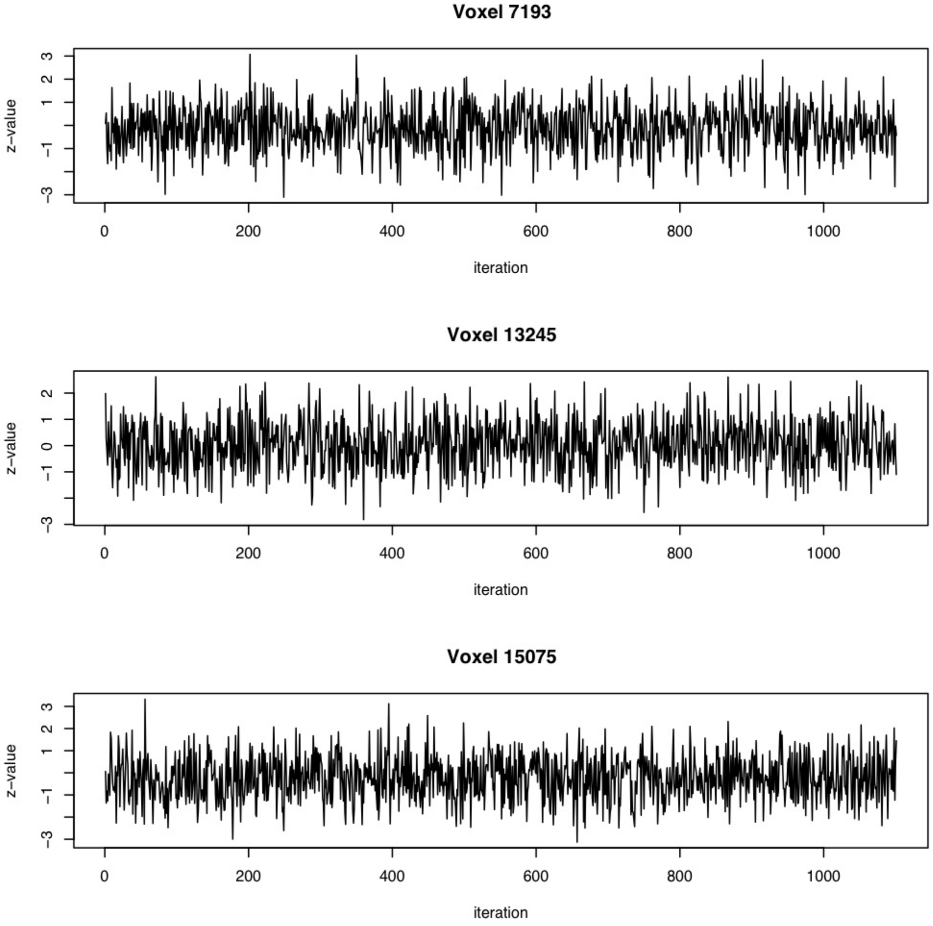

Figure 2: Traces of the z-statistic from three randomly chosen voxels

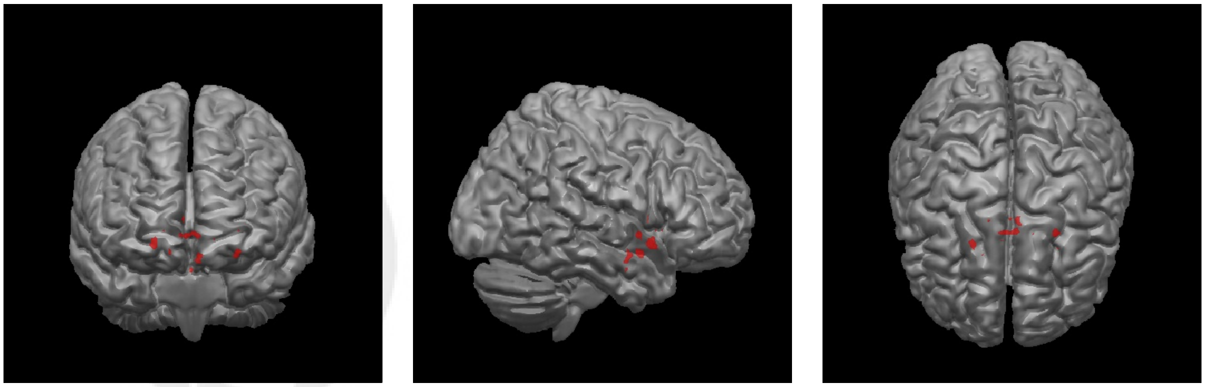

Figure 3: The clusters beyond the $z = \pm 2.33$ threshold are projected to the surface and shown in red. Note that this threshold is lower than that used in the analysis. The entire brain is pictured, but only a coronal band encompassing the medial temporal lobe and surrounding structures was imaged.

7 Discussion

We have introduced a flexible method for covariate control in neuroimaging studies. The method immediately applies to any image-based study with two groups and multiple categorical confounders in a logit propensity score model. The method uses an algorithm to permute the group labels conditionally on the covariates. We introduced a novel MCMC implementation that is applicable to a large class of models and settings. We applied the algorithm to an example fMRI dataset comparing at-risk and control subjects. The application was not ideal for a causal discussion, hence application to a study with an assignable treatment with a full causal discussion is a next step.

As is, the algorithm swaps the group labels on at most four individuals in each permutation. To reduce the resulting high correlation between each iteration, we plan to implement a more general version of the algorithm that alters the cells in the 2x2x2 by $\pm\epsilon$. At each iteration of the algorithm, ϵ is chosen randomly from the set of non-negative integers such that the resulting table has no negative cells. Increasing the number of subjects who are relabeled in each iteration decreases the numbers of iterations required to sufficiently cover the support.

The algorithm currently applies to any setting with two treatments and multiple categorical predictors with sufficient permutations after conditioning. However, we require knowledge of the Markov basis for the associated contingency table/log-linear model. Hence, further characterization of the Markov bases in these settings is of interest. In addition, other Markov chain algorithms, not based on the Diaconis/Sturmfels algorithm may produce more desirable chains. Also, for large numbers of covariates and small, yet complex, space of permutations, network algorithms may provide a fast method for enumeration (Hirji et al., 1987; Mehta et al., 2000).

Another extension is to consider more than two treatment levels. Potentially, baseline category logit models could be used similarly to the logit models here. Further extensions would include methods for continuous covariates. It is possible that methods of approximate conditioning (Pierce and Peters, 1999) could be used.

Finally, extensions to longitudinal data, matched data and other settings with multiple images per subject are of interest. Parallels with traditional rank-based permutation methods (Mahfoud and Randles, 2005a,b) provide an important foundation for future work along these lines.

References

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

Anderson, M. and Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of statistical computation and simulation*, 62(3):271–303.

Arndt, S., Cizadlo, T., Andreasen, N., Heckel, D., Gold, S., and O'Leary, D. (1996). Tests for comparing images based on randomization and permutation methods. *Journal of Cerebral Blood Flow & Metabolism*, 16(6):1271–1279.

Bassett, S., Yousem, D., Cristinzio, C., Kusevic, I., Yassa, M., Caffo, B., and Zeger, S. (2006). Familial risk for Alzheimer's disease alters fMRI activation patterns. *Brain*, 129(5):1229.

Bookheimer, S., Strojwas, M., Cohen, M., Saunders, A., Pericak-Vance, M., Mazziotta, J., and Small, G. (2000). Patterns of brain activation in people at risk for Alzheimer's disease. *New England Journal of Medicine*, 343(7):450–456.

Booth, J. and Butler, R. (1999). An importance sampling algorithm for exact conditional tests in log-linear models. *Biometrika*, 86(2):321–332.

Brett, M., Penny, W., and Kiebel, S. (2007). Parametric procedures. In Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W., editors, *Statistical Parametric Mapping: The Analysis of Functional Brain Imaging*, chapter 17. Elsevier Ltd., 1 edition.

Bross, I. D. J. (1964). Taking a covariable into account. *Journal of the American Statistical Association*, 59(307):725–736.

Bullmore, E., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E., and Brammer, M. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *Medical Imaging, IEEE Transactions on*, 18(1):32–42.

Caffo, B. and Booth, J. (2003). Monte Carlo conditional inference for log-linear and logistic models: a survey of current methodology. *Statistical Methods in Medical Research*, 12(2):109.

Caffo, B. S. and Booth, J. G. (2001). A Markov chain Monte Carlo algorithm for approximating exact conditional probabilities. *Journal of Computational and Graphical Statistics*, 10(4):730–745.

Chen, Y., Diaconis, P., Holmes, S., and Liu, J. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–121.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49:327–335.

Corder, E., Saunders, A., Strittmatter, W., Schmechel, D., Gaskell, P., Small, G., Roses, A., Haines, J., and Pericak-Vance, M. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261(5123):921–923.

Cox, D. and Snell, E. (1989). *Analysis of Binary Data*. Chapman & Hall/CRC.

D'Agostino, R. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19):2265–2281.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26(1):363–397.

Dobra, A. (2003). Markov bases for decomposable graphical models. *Bernoulli*, 9(6):1093–1108.

Edgington, E. (1995). *Randomization tests*. CRC Press.

Forster, J., McDonald, J., and Smith, P. (1996). Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(2):445–453.

Friston, K., Ashburner, J., Stefan, K., Nichols, T., and Penny, W., editors (2007). *Statistical Parametric Mapping The Analysis of Functional Brain Images*. Academic Press.

Gail, M. H., Tan, W. Y., and Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika*, 75(1):57–64.

Gao, S., Hendrie, H., Hall, K., and Hui, S. (1998). The relationships between age, sex, and the incidence of dementia and Alzheimer disease a meta-analysis. *Archives of General Psychiatry*, 55(9):809–815.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice*, chapter 1, pages 1–19. Chapman and Hall.

Good, P. (2006). *Resampling methods: a practical guide to data analysis*. Birkhäuser, 3 edition.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hayasaka, S. and Nichols, T. E. (2003). Validating cluster size inference: random field and permutation methods. *NeuroImage*, 20(4):2343 – 2356.

Hirji, K., Mehta, C., and Patel, N. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, pages 1110–1117.

Hobert, J., Jones, G., Presnell, B., and Rosenthal, J. (2002). On the applicability of regenerative simulation in markov chain monte carlo. *Biometrika*, 89(4):731–743.

Holmes, A., Blair, R., Watson, G., and Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, 16(1):7–22.

Joffe, M. M. and Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, 150(4):327–333.

Jones, G. (2004). On the Markov chain central limit theorem. *Probability surveys*, 1:299–320.

Jones, G., Haran, M., Caffo, B., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101(476):1537–1547.

Jones, G. and Hobert, J. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, 16(4):312–334.

Kennedy, P. E. (1995). Randomization tests in econometrics. *Journal of Business & Economic Statistics*, 13(1):85–94.

MacEachern, S. and Berliner, L. (1994). Subsampling the Gibbs sampler. *American Statistician*, 48(3):188–190.

Mahfoud, Z. and Randles, R. (2005a). On multivariate signed rank tests. *Journal of Nonparametric Statistics*, 17(2):201–216.

Mahfoud, Z. and Randles, R. (2005b). Practical tests for randomized complete block designs. *Journal of Multivariate Analysis*, 96(1):73–92.

McDonald, J., Smith, P., and Forster, J. (1999). Exact tests of goodness of fit of log-linear models for rates. *Biometrics*, 55(2):620–624.

Mehta, C. and Patel, N. (1998). Exact inference for categorical data. *Encyclopedia of Biostatistics*, 2:1411–1422.

Mehta, C., Patel, N., and Senchaudhuri, P. (2000). Efficient Monte Carlo methods for conditional logistic regression. *Journal of the American Statistical Association*, pages 99–108.

Nichols, T. and Holmes, A. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, 15(1):1–25.

Pierce, D. and Peters, D. (1999). Improving on exact tests by approximate conditioning. *Biometrika*, 86(2):265–277.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any population. *Journal of the Royal Statistical Society*, 4(2):119–130.

Poline, J. and Mazoyer, B. (1992). Analysis of individual positron emission tomography activation maps by high signal to noise ratio pixel clusters (HSC) detection. In *Nuclear Science Symposium and Medical Imaging Conference, 1992., Conference Record of the 1992 IEEE*, volume 2, pages 1259–1261.

Rabe-Hesketh, S., Bullmore, E. T., and Brammer, M. J. (1997). The analysis of functional magnetic resonance images. *Statistical Methods in Medical Research*, 6(3):215–237.

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer, 2 edition.

Rosenbaum, P. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574.

Rosenbaum, P. (2002). *Observational studies*. Springer.

Rosenbaum, P. and Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of The American Statistical Association*, 79(387):516–524.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26.

Saunders, A. M., Strittmatter, W. J., Schmechel, D., St. George-Hyslop, P. H., Pericak-Vance, M. A., Joo, S. H., Rosi, B. L., Gusella, J. F., Crapper-MacLachlan, D. R., Alberts, M. J., Hulette, C., Crain, B., Goldgaber, D., and Roses, A. D. (1993). Association of apolipoprotein E allele epsilon4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*, 43(8):1467–1472.

Smith, P., Forster, J., and McDonald, J. (1996). Monte Carlo exact tests for square contingency tables. *Journal of the Royal Statistical Society: Series A (Statistics in society)*, 159(2):309–321.

Strittmatter, W. J., Saunders, A. M., Schmechel, D., Pericak-Vance, M., Enghild, J., Salvesen, G. S., and Roses, A. D. (1993). Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences of the United States of America*, 90(5):1977–1981.

Wager, T., Hernandez, L., Jonides, J., and Lindquist, M. (2007). Elements of functional neuroimaging. In Cacioppo, J., Tassinary, L., and Berntson, G., editors, *Handbook of Psychophysiology*, pages 19–55. Cambridge University Press, 4 edition.

Appendix

Explanation of fixed margins

The margins in a 2x2x2 table are n_{0+0}, n_{0+1}, n_{1+0}, n_{1+1}, n_{00+}, n_{01+}, n_{10+}, n_{11+}, n_{0++}, and n_{1++} where the first subscript denotes the value of y, the second subscript denotes the value of x_1, the third subscript denotes the value of x_2, and a $+$ subscript indicates the sum over the corresponding variable (see Table 4). We assume the logit model in equation (1) and condition on the sufficient statistic S to eliminate the nuisance parameter β from the null distribution. Specifically, when $\mathbf{x} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x_2})$, $S = (\sum_i y_i, \sum_i x_{1i} y_i, \sum_i x_{2i} y_i)$. Note that $\sum_i x_{1i} y_i = n_{11+}$ and $\sum_i x_{2i} y_i = n_{1+1}$. Hence n_{11+} and n_{1+1} are fixed and, because $\sum_i y_i$ is fixed, so are n_{10+} and n_{1+0}. We assume that $\mathbf{x}$ and the total number of observations, $n = n_{1++} + n_{0++}$, are fixed. Therefore fixing $\sum_i y_i$ implies that $n - \sum_i y_i$ must also be fixed. Fixing $\sum_i x_{1i}$ implies that n_{01+} is fixed because n_{11+} is fixed and fixing $\sum_i x_{2i}$ implies that n_{0+1} is fixed because n_{1+1} is fixed. Finally, because $n - \sum_i y_i$ is fixed, we have that n_{00+} and n_{0+0} are fixed.

	$y = 0$		
	$x_2 = 0$	$x_2 = 1$	
$x_1 = 0$	n_{000}	n_{001}	n_{00+}
$x_1 = 1$	n_{010}	n_{011}	n_{01+}
	n_{0+0}	n_{0+1}	n_{0++}

	$y = 1$		
	$x_2 = 0$	$x_2 = 1$	
$x_1 = 0$	n_{100}	n_{101}	n_{10+}
$x_1 = 1$	n_{110}	n_{111}	n_{11+}
	n_{1+0}	n_{1+1}	n_{1++}

Table 4: Cell counts and margins in a 2x2x2 table