# A Bayesian Morphometry Algorithm

Edward H. Herskovits*, *Member, IEEE*, Hanchuan Peng, *Member, IEEE*, and Christos Davatzikos

***Abstract*—Most methods for structure-function analysis of the brain in medical images are usually based on voxel-wise statistical tests performed on registered magnetic resonance (MR) images across subjects. A major drawback of such methods is the inability to accurately locate regions that manifest nonlinear associations with clinical variables. In this paper, we propose Bayesian morphological analysis methods, based on a Bayesian-network representation, for the analysis of MR brain images. First, we describe how Bayesian networks (BNs) can represent probabilistic associations among voxels and clinical (function) variables. Second, we present a model-selection framework, which generates a BN that captures structure-function relationships from MR brain images and function variables. We demonstrate our methods in the context of determining associations between regional brain atrophy (as demonstrated on MR images of the brain), and functional deficits. We employ two data sets for this evaluation: the first contains MR images of 11 subjects, where associations between regional atrophy and a functional deficit are almost linear; the second data set contains MR images of the ventricles of 84 subjects, where the structure-function association is nonlinear. Our methods successfully identify voxel-wise morphological changes that are associated with functional deficits in both data sets, whereas standard statistical analysis (i.e., t-test and paired t-test) fails in the nonlinear-association case.**

***Index Terms*—Bayesian network, Bayes procedures, computational anatomy, image analysis, image classification, morphology-function analysis, voxel-based morphometry.**

## I. Introduction

VOXEL- and deformation-based morphometry have been increasingly used to identify morphological abnormalities, such as atrophy, without the need to define *a priori* specific regions of interest (ROIs). Many different approaches [1]–[16] have been proposed for generating statistical maps that identify groups of voxels that display differences in morphology, or voxels for which significant correlations exist among morphological and clinical measurements. Morphological measurements can be computed from the deformation field used to spatially normalize subjects into a stereotaxic space [7], [11], [13], [14], [17], from residual variability in the spatial distribution of gray and white matter after spatial normalization [1], [10], or from tissue-density maps obtained after mass-preserving spatial normalization [4], [6].

For the purpose of morphology-function analysis, particularly voxel-wise morphometry, one of the first steps is warping MR images into a normalized space (i.e., registration), to ensure that voxel attributes across subjects can be compared. A widely used brain-image registration technique is the smooth parametric transformation [1], [8], [9], [16], which is provided as part of the SPM99 software package (http://www.fil.ion.ucl.ac.uk/spm/spm99.html). Our group previously developed another method referred to as spatial transformation algorithm for registration (STAR) [4], which utilizes a high-dimensional elastic transformation, coupled with a procedure that preserves information about the volumes of different anatomical structures, by constructing tissue-density maps, in which relatively higher density of a particular structure implies that this structure had a relatively higher volume prior to spatial normalization. This procedure is a key component of our approach, since spatial normalization changes the anatomy of individual subjects, by making each subject's anatomy similar to that of a template. Therefore, applying a registration algorithm that preserves volumetric information during spatial normalization is critical.

Regardless of the type of morphological variables being considered, e.g., voxels or regions, most existing morphology-function analysis methods rely on voxel-wise linear statistics, such as t-tests (TTs), paired t-tests (PTs), or analyses of variance (ANOVAs). Such statistics compare only the means and variances of variables among different groups; therefore, methods based on these statistics may not be able to detect nonlinear morphology-function associations. Second, even for linear associations, these methods usually require a predefined confidence interval, or p-value threshold, to generate ROIs. Third, these statistical tests generally do not directly describe the relationships among the generated ROIs. Other methods, such as principal-component analysis and partial-least-square analysis [18], can be expected to capture more complex morphology-function associations. However, few of these methods can synthesize, without user intervention, complex multivariate morphology-function models. Thus, we distinguish between linear associations among continuous variables, which are readily evaluated using methods based on the general linear model (GLM), such as ANOVA or linear regression, and nonlinear associations among continuous or categorical variables, which may not be captured by GLM-based analysis.

In this paper, we use Bayesian networks (BNs) [19]–[21] to represent probabilistic associations among MR image voxels and clinical variables. A BN is a directed acyclic graph (DAG) describing the probabilistic relationships among variables; each node represents a variable, and directed edges coming into a child node indicate that there are corresponding conditional-probability distributions for the child, given the joint

Manuscript received July 28, 2003; rebised February 20, 2004. This work was supported in part by the Human Brain Project under Grant AG13743, which is funded by the National Institute of Aging, the National Cancer Institute, and the National Institute of Mental Health. The work of E. H. Herskovits was supported in part by a Richard S. Ross Clinician Scientist Award from Johns Hopkins University.

*E. H. Herskovits is with the Department of Radiology, University of Pennsylvania, 3600 Market Street, Suite 370, Room 117, Philadelphia, PA 19104 USA (e-mail: ehh@ieee.org).

H. Peng is with the Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720 USA (e-mail: hpeng@lbl.gov).

C. Davatzikos is with the Department of Radiology, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: christos@rad.upenn.edu).

Digital Object Identifier 10.1109/TMI.2004.826949

0278-0062/04$20.00 © 2004 IEEE

states of its parents. Each node without parents is associated with a prior-probability distribution. In this framework, voxel-morphology variables and clinical variables are nodes, and morphology-function analysis is equivalent to the generation of a BN from MR image data and clinical information for each subject. Because BNs can represent any joint distribution over discrete variables, they provide a powerful foundation for the nonparametric analysis of nonlinear associations among these variables. The BN-based methods that constitute our Bayesian morphological analysis (BMA) algorithm are formalized at the end of this section.

We evaluate our methods by simulating cerebral atrophy in structural MR brain images, in the setting of changes in clinical variables. Cerebral atrophy is a degenerative process that generally occurs after 55 years of age, although it may occur much more rapidly in certain diseases [22]. In this process, the brain loses mass and volume, causing the cerebral sulci and ventricles to dilate. Many cortical, subcortical, and mixed cortical-subcortical encephalopathies, such as Alzheimer's disease and Parkinson's disease, have atrophy as their primary structural manifestation. This application is a typical example of morphology-function analysis, in which we seek to delineate associations among brain morphological changes and clinical variables, such as anomia (inability to name objects) or apraxia (inability to perform tasks).

We have arranged this paper as follows. The remainder of this section briefly introduces BNs. Section II describes an overview of our approach, and presents two methods implemented within the BMA framework for generating sets of equivalent voxels. Section III describes our performance metrics. Sections IV and V illustrate experimental results on a linear-association data set of cerebral MR images, and on a nonlinear-association data set of ventricular MR images, respectively. After discussion in Section VI, we present our conclusions in Section VII.

### A. Bayesian Networks

Suppose we have $n$ random variables $X = \{x_1, \ldots, x_n\}$; a BN for $X$ consists of a DAG structure $\mathbf{S}$, in which nodes correspond to variables in $X$ (we, therefore, use the terms *node* and *variable* interchangeably), and a set of local distribution functions $p(x_i|\pi_i, \boldsymbol{\theta}_\mathbf{S}, \mathbf{S})$, where $\pi_i$ is the set of $x_i$'s parent nodes and $\boldsymbol{\theta}_\mathbf{S}$ is the parameter set of all conditional probabilities.

The structure $\mathbf{S}$ encodes conditional-independence statements, such that $p(X|\boldsymbol{\theta}_\mathbf{S}, \mathbf{S}) = \prod_{i=1}^{n} p(x_i|\pi_i, \boldsymbol{\theta}_\mathbf{S}, \mathbf{S})$ [23]; that is, the structure of a BN defines a decomposition of a joint distribution into the product of conditional-probability distributions, based on the notion of conditional independence, which we elaborate below. Many model-selection algorithms [20], [24]–[27] have been proposed to construct BNs from data. Often these algorithms are based on assumptions similar to the following [20], [24], [27]–[30].

1) Each variable is discrete, having a finite number of states. We use $x_i^k$ and $\pi_i^j$ to denote the $k$th state of $x_i$ and the $j$th joint state of $\pi_i$, respectively. We use $r_i$ to denote the number of states of $x_i$, and $q_i$ to denote the number of joint states of $\pi_i$.
2) Each local distribution function $P(x_i|\pi_i, \boldsymbol{\theta}_\mathbf{S}, \mathbf{S})$ consists of a set of distributions defined as the parameters
$$\boldsymbol{\theta}_{ijk} \equiv P\left(x_i^k|\pi_i^j, \boldsymbol{\theta}_\mathbf{S}, \mathbf{S}\right) \tag{1}$$
where for all $i, j, k, \boldsymbol{\theta}_{ijk} > 0$ and $\sum_{k=1}^{r_i} \boldsymbol{\theta}_{ijk} = 1$. Denote the parameter set $\boldsymbol{\theta}_{ij} = \{\boldsymbol{\theta}_{ij1}, \ldots, \boldsymbol{\theta}_{ijr_i}\}$.
3) The parameter sets $\boldsymbol{\theta}_{ij}$ are mutually independent, so that $P(\boldsymbol{\theta}_\mathbf{S}|\mathbf{S}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} P(\boldsymbol{\theta}_{ij}|\mathbf{S})$.
4) Each parameter set $\boldsymbol{\theta}_{ij}$ assumes a Dirichlet distribution: $P(\boldsymbol{\theta}_{ij}|\mathbf{S}) = \mathrm{Dir}(\boldsymbol{\theta}_{ij}|\alpha_{ij1}, \ldots, \alpha_{ijr_i}) \propto \prod_{k=1}^{r_i} \boldsymbol{\theta}_{ijk}^{\alpha_{ijk}-1}$, where each hyperparameter $\alpha_{ijk} > 0$ for every $i, j, k$.
5) The data set $\mathbf{D}$ is complete, that is, every variable is observed in every case of $\mathbf{D}$.

Under these assumptions, the parameters remain independent given $\mathbf{D}$

$$P(\boldsymbol{\theta}_\mathbf{S}|\mathbf{D}, \mathbf{S}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} P(\boldsymbol{\theta}_{ij}|\mathbf{D}, \mathbf{S}) \tag{2}$$

and the posterior distribution of each $\boldsymbol{\theta}_{ij}$ assumes the Dirichlet distribution

$$\begin{aligned} P(\boldsymbol{\theta}_{ij}|\mathbf{D}, \mathbf{S}) &= \mathrm{Dir}\left(\boldsymbol{\theta}_{ij}|\alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i}\right) \\ &\propto \prod_{k=1}^{r_i} \boldsymbol{\theta}_{ijk}^{\alpha_{ijk}+N_{ijk}-1} \end{aligned} \tag{3}$$

where $N_{ijk}$ is the number of cases in $\mathbf{D}$ in which $x_i = x_i^k$ and $\pi_i = \pi_i^j$. $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ is the total number of cases in which $\pi_i$ assumes the $j$th joint parent configuration.

The closed-form solution for computing the probability that data D could be generated by BN structure $\mathbf{S}$ was first derived in [28], [29]

$$P(\mathbf{D}|\mathbf{S}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{(\alpha_{ij}-1)!}{(N_{ij}+\alpha_{ij}-1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{4}$$

where $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. As discussed in [27], [31], [32], the choice of $\alpha_{ijk}$ reflects a choice of prior distribution over $\theta$; although most algorithms are based on a noninformative prior distribution, there is no consensus for the correct choice of $\alpha_{ijk}$ to achieve this goal, even for the simplest case, in which $r_i = 2$ (i.e., a binary variable). Two commonly specified values for $\alpha_{ijk}$ are 1 and $1/r_i$; based on our previous experience, we have found that the choice of $\alpha_{ijk}$ affects results only for small (i.e., undersampled) data sets; thus, we have chosen to set $\alpha_{ijk}$ to 1 for BMA and, therefore, $\alpha_{ij} = r_i$

$$P(\mathbf{D}|\mathbf{S}) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}+r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!. \tag{5}$$

There are many heuristic and information-theoretic approaches to measuring how well a particular BN structure represents the joint distribution inherent to the data; we have chosen to base BMA on the metric in (5), since it has polynomial computational complexity in the cardinality of $\mathbf{D}$, and can be combined with prior information, $P(\mathbf{S})$, regarding associations among variables [28], [29].

Conventional statistical approaches to data analysis usually require that the researcher specify a model prior to data collection, or at least prior to data analysis; greedy-search methods are part of stepwise regression, among other conventional statistical methods, but have not been applied to morphology-function analysis. This aspect of the analysis is important because, in the setting of morphology-function analysis, we may have little or no knowledge of which regions undergo morphologic change in the setting of a particular clinical syndrome; thus, we

must specify a space of models to search, and a search algorithm that, in conjunction with the model-scoring metric, will generate the model representing the associations inherent in the data. Because the number of possible BN structures is exponential in the number of variables [33], it is impossible to completely search the space of all possible network structures to find the best one, especially for voxel-wise morphometry, in which there may be hundreds of thousands of voxel variables. As we describe later, we solve this problem by heuristically searching a specific subset of possible BN structures that represent associations among voxels and clinical variables.

As stated above, BNs are based on the concept of conditional independence among variables [21]. Variable $u$ is conditionally independent of $v$ given variable $w$ if $P(u|v,w) = P(u|w)$. In this case, knowledge of $v$ will not alter the probability of $u$, given knowledge of $w$. If $w$ is empty, we say that $u$ and $v$ are marginally independent. In this paper, we utilize the notion of conditional independence to find candidate sets of equivalent variables, in particular, voxels that have similar probabilistic associations with a function variable.

Latent-variable induction in BN models has been presented as a clustering method [34]–[36]. In this paper, we use latent-variable induction to generate sets of equivalent variables from the above-mentioned candidate sets.

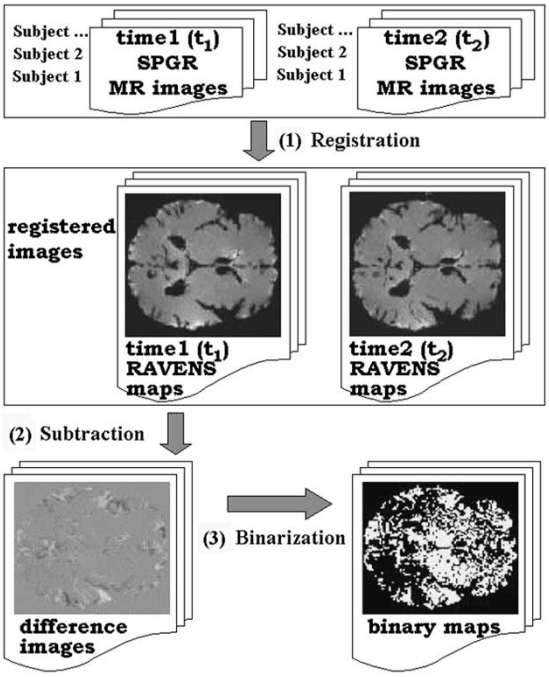


Fig. 1. Three major image preprocessing steps: registration, subtraction and thresholding (binarization).

## II. Bayesian Morphological Analysis

In this section, we propose a BMA algorithm, which detects morphology-function associations between brain morphological measurements and clinical variables. For convenience, we use the terms voxel and voxel variable interchangeably. Like other morphometry methods, BMA requires a preprocessing stage for the images. BMA is based on the Bayesian metric (5) and a heuristic model-selection method to generate a Bayesian-network structure from the data. BMA generates sets of equivalent voxels based on Bayesian thresholding (BT) or Bayesian clustering (BC).

### A. Image Preprocessing

The purpose of data preprocessing is to generate the image portion of data $\mathbf{D}$ for the BMA algorithm. For purposes of illustration, most of our development will be presented in the context of finding associations between longitudinal brain atrophy and a function variable that might be associated with such atrophy. Suppose we have longitudinal MR images (at times $\mathrm{t}_1$ and $\mathrm{t}_2$) of a group of subjects (labeled 1, 2, ...), along with measurements of a categorical clinical variable, $f$, which could reflect performance on a neuropsychiatric battery of tests. The three major image-processing steps in the BMA framework, i.e., registration, subtraction and thresholding, are shown in Fig. 1.

1) *Registration:* In the registration step, brain images of different sizes and shapes are warped to a stereotaxic canonical space. In this paper, we have employed STAR [4], [12], which is a high-dimensional elastic-registration method. Unavoidably, registration introduces a complication, namely that it changes the morphology of an individual's brain. Therefore, it would be pointless to examine the morphology of spatially normalized brain images in a structure-function analysis. In order to overcome this problem, we use an approach referred to as regional analysis of volumes embedded in stereotaxic space (RAVENS), which is described in detail in [4], [6], [12]. In this approach, three-dimensional (3-D) density maps for each tissue class, such as gray matter, white matter, and cerebrospinal fluid, are generated separately. For example, assume that an individual brain has, due to atrophy, larger ventricles than those in the canonical-space template. Then the density of the corresponding CSF map will be high after spatial normalization, reflecting the fact that a relatively larger volume of CSF is forced to fit into a relatively smaller space. More generally, RAVENS maps reflect the regional volumetric structure of the brain of each subject, with the tissue density of any structure being proportional to the actual volume of the structure in an individual brain, prior to spatial normalization. Since these maps are registered and reside in the same canonical space, they can be overlaid and analyzed on a voxel-wise basis.
2) *Subtraction:* We subtract the pair-wise $(\mathrm{t}_1 - \mathrm{t}_2)$ RAVENS maps to generate a difference image for each subject; this image reflects longitudinal voxel-wise morphological changes for each subject. Due to volume contraction, atrophic regions in $\mathrm{t}_2$ images will, on average, have lower intensity than the corresponding regions in $\mathrm{t}_1$ images, since image intensity of the RAVENS maps reflects tissue density. As a result, the difference maps $(\mathrm{t}_1 - \mathrm{t}_2)$ will generally have positive values in these regions, and negative values for regions that dilate over time, such as ventricles in the setting of progressive cerebral atrophy.
3) *Thresholding:* Our current method applies only to categorical variables. Therefore, we binarize the longitudinal difference maps by thresholding them at zero. That is, a voxel with a value larger than 0 is set to state 1 (for "volume contraction") otherwise the voxel is set

to state 0 (for "no volume contraction"). In regions of atrophy, binary-map voxels will in general assume state 1. At all other locations the binary maps would have approximately equal probabilities of assuming state 1 or state 0, because of noise or other factors that may make a voxel appear to have enlarged or shrunk when in fact it did not change. These binary maps are used as the image data provided to BMA.

In BMA, the data satisfy the five assumptions upon which (2) is based. We ensure that the voxel and function data are discrete, to satisfy assumption 1), as described in the previous paragraph. The principal advantage of working with discrete variables is the ability to detect multivariate nonlinear associations among these variables; thus, although there are methods for generating continuous-variable BNs from data, we chose not to implement them because they are restricted to modeling multivariate Gaussian distributions over these variables. We model associations among variables using edges in a BN, which correspond to conditional-probability distributions, as specified in assumption 2). We assume independence of distributions, since we have no domain-specific knowledge that would lead us to assume dependence among these distributions. In fact, this assumption is almost always applied in practice; it is rare to work with data for which knowledge of one variable's conditional-probability distribution will constrain other variables' distributions. There is nothing specific to morphological analysis that makes unreasonable the assumption of a Dirichlet distribution over conditional probabilities; see also Heckerman [20] for a discussion of the application of these prior distributions when generating BNs from data. In addition, it is reasonable to assume that these distributions are stationary across subjects; that is, we assume that the nature of the structure-function relationships that we seek to elucidate do not change over time or across subjects. Finally, we work with complete data in the experiments presented here; in the setting of morphological analysis, there will virtually never be missing values for the voxel or function variables.

### B. Generating a Bayesian-Network Structure From Data

In BMA, a Bayesian-network structure $\mathbf{S}$ is constructed from data $\mathbf{D}$ to achieve two goals: 1) to identify voxels associated with the function variable; 2) to classify these voxels into probabilistically homogeneous subsets or clusters, each of which has a single *representative voxel*, whose association with the function variable is similar to those for the other members of that cluster. The initial network consists of all voxels and the function variable, $f$, with no edges among these variables. The first step results in the addition to this network of a few edges from representative voxels to the function variable; all of the remaining voxels have no edges. In the second step, each representative voxel is associated with a cluster of probabilistically equivalent voxels; this switch from voxels to voxel clusters is a form of data reduction. These two steps are repeated in each iteration of BMA. We next consider three major implementation issues: the subset of possible Bayesian-network structures considered, the metric used to compare network structures, and the search strategy.

We base BMA on a network structure in which we postulate that $f$ possesses associations with representative voxels, whereas each pair $r_i, r_j, i \neq j$, are independent (i.e., there is no edge between $r_i$ and $r_j$). In other words, the final network generated by BMA will contain representative voxels, the function variable, and an edge from each representative voxel to the function variable. This network structure is able to capture complex associations among these variables; that is, the resulting networks can represent multivariate nonlinear associations among representative voxels and the function variable.

We use the Bayesian metric [28], [29] $M(\mathbf{S})$, which is the conditional probability of network structure $\mathbf{S}$ given the data $\mathbf{D}$, to evaluate candidate Bayesian-network structures. $M(\mathbf{S})$ has the following form:

$$M(\mathbf{S}) = P(\mathbf{S}|\mathbf{D}) = \frac{P(\mathbf{D}|\mathbf{S})P(\mathbf{S})}{P(\mathbf{D})}$$

where we use (5) to compute $P(\mathbf{D}|\mathbf{S})$.

A larger Bayesian metric value indicates a higher probability that the corresponding Bayesian-network structure could have generated the observed data. Since we have no *a priori* preference regarding network structures, we assume the prior $P(\mathbf{S})$ is uniform, as in [28], [29]. Furthermore, because the prior probability of observing the data $\mathbf{D}$ is a constant, $M(\mathbf{S})$ is proportional to the likelihood function, i.e., $M(\mathbf{S}) \propto P(\mathbf{D}|\mathbf{S})$, which takes the form of (5) for discrete variables. For computational purposes, we redefine the metric $M$ as the logarithm of (5).

### C. Model Selection

Here, we wish to obtain a set of clusters, each of which contains voxels that have similar probabilistic associations with the function variable, $f$. In this paper, we employ the concept of *probabilistic equivalence*: two variables $v$ and $u$ are probabilistically equivalent if $v$ and $u$ have the same number of states, and $P(v|u) \approx P(u|v) \approx 1$ for each state. Similarly, we define two voxels $v_1$, $v_2$ to be *probabilistically equivalent* with respect to $f$ and a set of voxels $V$ to mean that the two distributions $P(f|V, v_1)$, $P(f|V, v_2)$ are similar, by some measure. Each cluster is characterized by a representative voxel, to which the remaining cluster members are compared for probabilistic similarity. We propose the method shown in Fig. 2 to generate these clusters of equivalent voxels.

The procedure in Fig. 2 can be understood step by step as follows. We begin with a set, $V$, initially containing all image voxels, and a set of representative voxels, $R$, that is initially empty. During the $n$th iteration of the algorithm we add a representative voxel, $r_n$, to $R$, and determine the set of voxels that are probabilistically similar to $r_n$; this set is called the *equivalence set* for $r_n$. To find this representative voxel, our algorithm first compares pairs of Bayesian-network structures with and without an edge from each voxel $v_i$ to $f$, in the presence of edges from current members of $R$ to $f$ (shown as step 3 in Fig. 2 and structures $\mathbf{S}_a$ and $\mathbf{S}_b$ in Fig. 3). In particular, to find the variable that has the strongest association with $f$, the algorithm compares each pair of BNs shown in Fig. 3(a) and (b) over all $v_i$ by computing the difference metric:

$$dM(\mathbf{S}_a, \mathbf{S}_b) = M(\mathbf{S}_a) - M(\mathbf{S}_b). \tag{6}$$

Since $dM$ is a function of $v_i$, we denote it as $dM_i$. We call the set of all $v_i$ for which an edge from $v_i$ to $f$ is favored (i.e.,

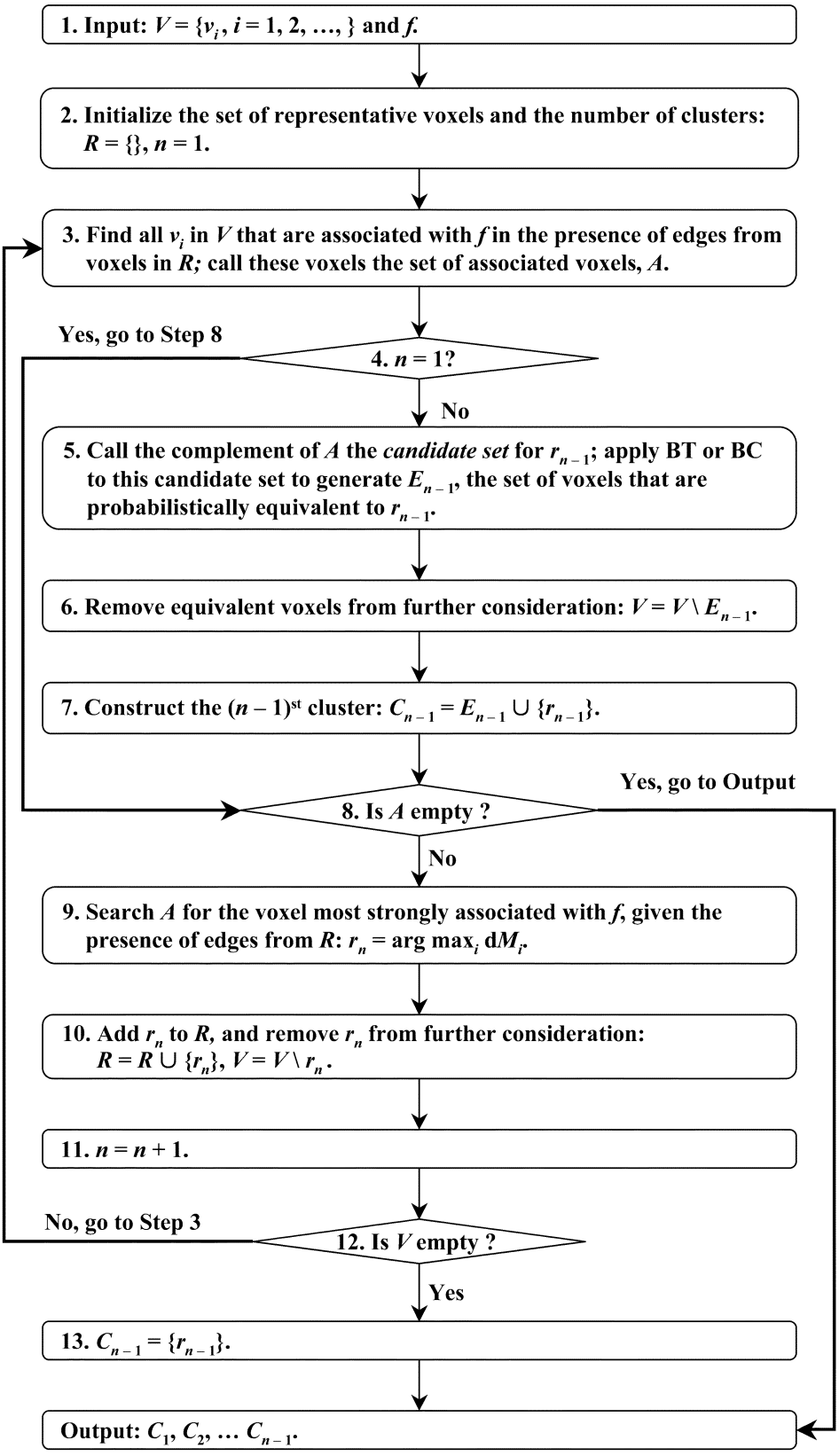


Fig. 2. Flowchart of the algorithm for equivalence-set generation.

$dM_i > 0$) the set of *associated voxels*, A. From this set, we obtain the maximum $dM^*$ and the corresponding voxel $v^*$

$$dM^* = \max dM_i \tag{7}$$

$$v^* = \arg\max_{v_i} dM_i. \tag{8}$$

If there are several voxels whose difference-metric values are equal, we choose the variable with the maximum $M(v_i)$ to be

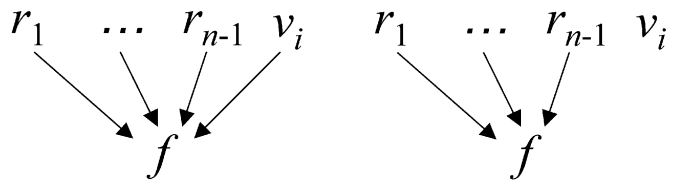


Fig. 3. Alternative structure-function Bayesian-network models $\mathbf{S_a}$ and $\mathbf{S_b}$. The difference between model $\mathbf{S_a}$ and $\mathbf{S_b}$ is the presence of the edge from $v_i$ to $f$.

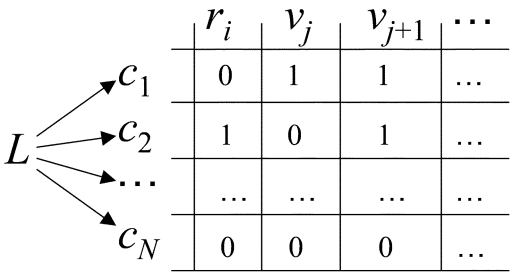


Fig. 4. The latent-variable Bayesian-network structure for the BC method.

$v^*$. If there are still several voxels with the same metric value, we arbitrarily choose one of them to be $v^*$. If $dM^* \leq 0$, the data do not favor an edge from any voxel to $f$ in the presence of $R$, hence our algorithm stops (this condition is equivalent to the judgment in Step 8 of Fig. 2). Otherwise, we set $r_n = v^*$, add $r_n$ to $R$, and remove $r_n$ from $V$.

Subsequent to the first iteration, we must generate the set of voxels that are probabilistically similar to $r_{n-1}$. To do so, we next consider $A'$, the complement of $A$ (with respect to $V$, the set of all voxels); this set contains voxels for which we could not establish an edge to $f$, in the presence of $R$. We call $A'$ the *candidate set* for the representative voxel $r_{n-1}$, because any voxel that is a member of $A$ is still associated with $f$ in the presence of $R$ and is, thus, unlikely to be probabilistically similar to any of the previously added representative voxels, including $r_{n-1}$. We then apply one of the methods described in the next section, BC or BT, to this candidate set $A'$, to generate $E_{n-1}$, the set of voxels equivalent to $r_{n-1}$. We then remove all voxels in $E_{n-1}$ from $V$, and we set cluster $C_{n-1}$ to be the union of $\{r_{n-1}\}$ and $E_{n-1}$.

This process continues until the set of voxels, $V$, is empty (step 12 in Fig. 2). Finally, the algorithm displays all clusters of probabilistically similar voxels, as volumes of interest. In addition, we can examine the conditional-probability distributions $P(f|r_1, \ldots, r_{n-1})$, which are proxies for $P(f|C_1, \ldots, C_{n-1})$.

### D. Bayesian Thresholding

As stated previously, in this paper we employ the concept of *probabilistic equivalence*: two variables $v$ and $u$ are probabilistically equivalent if $v$ and $u$ have the same number of states, and $P(v|u) \approx P(u|v) \approx 1$ for each state. Our goal is to determine the subset of voxels in a representative voxel's candidate set that are equivalent to that representative voxel.

A straightforward method to find the equivalence set of a representative voxel is BT, which determines whether the association between each candidate-set voxel and the representative voxel is strong. In BT, two "equivalent" binary variables $u$ and $v$ are required to satisfy the following conditions:

$$\begin{aligned} &P(v=1|u=1) \approx P(u=1|v=1) \approx 1 \geq P_{\mathrm{BT}} \\ &P(v=0|u=1) \approx P(u=0|v=1) \approx 0 < 1 - P_{\mathrm{BT}} \\ &P(v=0|u=0) \approx P(u=0|v=0) \approx 1 \geq P_{\mathrm{BT}} \\ &P(v=1|u=0) \approx P(u=1|v=0) \approx 0 < 1 - P_{\mathrm{BT}} \end{aligned} \quad (9)$$

where $P_{\mathrm{BT}}$ is a predefined threshold in [0, 1].

### E. Bayesian Clustering

The major drawback of the BT method (as well as other threshold-based methods) is its reliance on a predefined threshold. We, therefore, developed another BMA algorithm, based on latent-variable induction in BNs, to cluster the candidate set and obtain the equivalence set. In this approach, we transpose the data $\mathbf{D}$, i.e., we consider a pseudovariable set $\mathbf{C}$, where each $c_i$ is the variable representing all $i$th cases of every variable in the candidate set (e.g., the $i$th voxel for each subject), and we regard the original variables as pseudocases. The pseudodata are denoted as $\mathbf{D^T}$. Fig. 4 shows this approach, where $L$ is a discrete latent variable with $r_L$ states, and with edges into each of the pseudovariables $c_i$, $i = 1, 2, \ldots, N$. Each state of $L$ corresponds to a set of pseudocases, i.e., all of the pseudocases for which $L$ assumes the $j$th state $L^j$ are clustered as one class. The joint distribution of $c_i$ is given in the following multinomial form:

$$P(C) = \sum_{j=1}^{r_L} P(L = L^j) \prod_{i=1}^{N} P(c_i | L = L^j) P(L = L^j). \quad (10)$$

To perform clustering using this approach, we assume a number of states for the latent variable $L$, and then estimate the unobservable state for $L$ in each case. An approximation method [35], based on the Laplace approximation and either the Bayesian information criterion, a minimum description length metric, or the Cheeseman–Stutz approximation, can be used. An alternative method is based on Monte Carlo approximation, which will produce a more accurate result if given enough time [35]. In this paper, we consider one specific Monte Carlo method, the Gibbs sampler [37], because we want the clustering results to be precise, while reducing computational requirements.

In the Gibbs-sampler approximation, we first randomly initialize the unobservable pseudocases of $L$ (with the assumed number of states $r_L$). Then we sequentially unassign $L$ for each pseudocase, and calculate the probability for each possible state, given the other pseudocases

$$P(L_j = L^i | \mathbf{D^T} \setminus L_j, \mathbf{S}) = \frac{P(L_j = L^i, \mathbf{D^T} \setminus L_j | \mathbf{S})}{\sum_{k=1}^{r_L} P(L_j = L^k, \mathbf{D^T} \setminus L_j | \mathbf{S})} \quad (11)$$

where $\mathbf{D^T} \setminus L_j$ denotes the data set $\mathbf{D^T}$ with $L$ set to "unknown" for $j$th pseudocase, and $\mathbf{S}$ is the model shown in Fig. 5. Since both the numerator and denominator are probabilities that are computed based on an assumption of complete data, they can be computed using (5). Then, the results of (11) are used to sample a new pseudocase. Next, all unobserved data are reassigned to produce the new data $\mathbf{D^T}$. We iterate this procedure until the

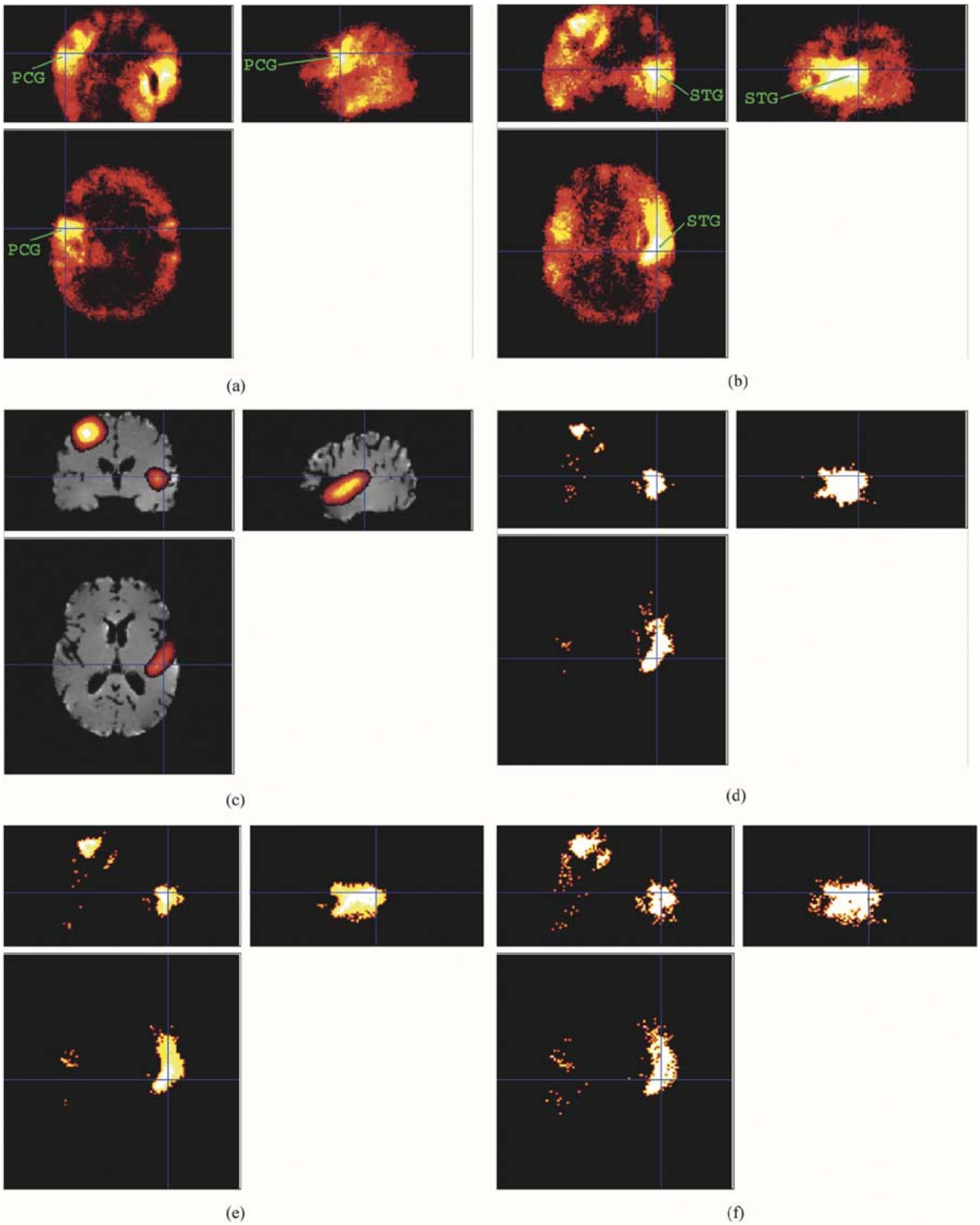


Fig. 5. Detection results for the paired $t$-test (PT), BT, and BC methods for linear morphology-function associations. (a),(b) Intensity plots of the average binary maps for the right PCG and left STG, where atrophy was simulated. Pixel intensity is proportional to the summation of binary maps, where 1 stands for volume loss and 0 for no volume loss. (c) The ground-truth image for the STG. The colored region is the smoothed ground truth, which is overlaid on one subject's image (gray) to aid visualization. (d)–(f) Detection results for the PT, BT, and BC methods.

distribution of model parameters in Fig. 5, which takes the form of (2), converges. We can rewrite such an indexing function as

$$P(\boldsymbol{\theta}_{\mathbf{S}}|\mathbf{D}^{\mathrm{T}},\mathbf{S}) = \prod_{i=1}^{N}\prod_{j=1}^{q_i} P(\boldsymbol{\theta}_{ij}|\mathbf{D}^{\mathbf{T}},\mathbf{S}). \tag{12}$$

To calculate (12), we use (3) and (1). After the indexing function converges, typically most of the unobserved cases of the latent variable $L$ will not change states. Recall that those pseudocases where the latent variable $L$ takes the same state are treated as one cluster. Hence, we can find the pseudocases for which the

corresponding voxels assume the same state as the representative voxel; these voxels constitute the equivalence set for the representative voxel.

We call this clustering method BC; it has also been referred to as the candidate method [35], [38], [39]. In [35], the authors were interested primarily in the indexing function value in (12), as opposed to the clustering results; this value is used to decide when the algorithm has converged. In the setting of morphology-function analysis, we are primarily concerned with whether the clustering results are meaningful. Furthermore, in contrast to [35], we do not use the normalized model parameters $\boldsymbol{\theta}$.

The computational bottleneck of BC is (11). Fortunately, we can simplify this equation as the relative value of (5), which obviates computation of Bayesian metrics over the whole BN in Fig. 5. In our implementation, this optimization greatly accelerates the Monte Carlo method, especially when there are many variables (pseudocases) in a candidate set.

The parameter $r_L$, the number of clusters, must be set; however, it is not difficult to choose this parameter. First, when $r_L$ is larger than a critical value $r_L^*$, which is the real number of clusters, the indexing function, (12), will typically converge to a value that is independent of $r_L$. Second, when $r_L > r_L^*$, the latent variable will usually have $r_L^*$ states after convergence. Therefore, we can simply set $r_L$ to a large value, say 6 or 7, even when we only expect to find 2 or 3 clusters. The only drawback of beginning with a large value for $r_L$ is greater computational burden, which is not critical in practice.

## III. Performance Metrics

To evaluate the performance of BMA, we must choose metrics that reflect the accuracy of detection of morphology-function associations; we do so with the knowledge that, for simulated data, we will have ground truth (or an approximation thereof) available.

Toward this end, we denote the result of BMA as $C^{\cup}$; this set is the union of the clusters, which include representative voxels and their corresponding equivalent voxels. Similarly, let us denote the original image data provided as input to BMA as $C^I$, and the ground truth as $C^*$. A natural way to measure the performance of the algorithm is to define the signal-detection rate (SDR) and the signal-to-noise ratio (SNR) as follows:

$$\mathrm{SDR} = \frac{\Omega(C^{\cup} \cap C^*)}{\Omega(C^*)} \tag{13}$$

$$\mathrm{SNR} = \frac{\Omega(C^{\cup} \cap C^*)}{\Omega\left(C^{\cup} \cap (C^I \setminus C^*)\right)} \tag{14}$$

where $\Omega(.)$ is the operator to calculate a region's volume, i.e., the number of voxels in that region. SDR measures the fraction of ground-truth voxels in $C^*$ that are included in $C^{\cup}$, and SNR indicates the degree of false-positive detection. When the extent of the ground truth $C^*$ is contained within the original image data $C^I$, (14) can also be written as

$$\mathrm{SNR} = \frac{\Omega(C^{\cup} \cap C^*)}{\Omega\left(C^{\cup} \setminus (C^{\cup} \cap C^*)\right)} = \frac{\Omega(C^{\cup} \cap C^*)}{\Omega(C^{\cup}) - \Omega(C^{\cup} \cap C^*)}.$$

We expect that SNR may be larger than 1. Nonetheless, when the ground-truth region $C^*$ is not in accordance with the input data $C^I$, it is possible that the algorithm might detect association regions not belonging to $C^*$. Hence both SDR and SNR can only be used as references, and alternatively a better way to evaluate performance is to compare the input data $(C^I)$ and the detection result $(C^{\cup})$ directly.

Although ideally morphology-function analysis would maximize SDR and SNR, if the data $\mathbf{D}$ contain redundant variables (for example two variables, one in $C^{\cup}$ and the other in $\overline{C^{\cup}}$ (i.e., $C^I \setminus C^{\cup}$), that have the same state for each case), there is no way to distinguish them without spatial information. We can expect $\mathrm{SNR} > 1$ if $C^*$ is in concordance with $C^I$ (i.e., if $C^{\cup}$ is very accurate), in which case it is possible to derive the theoretical SDR and SNR based on Bayes' theorem. Furthermore, we can claim that the morphology-function analysis algorithm performs well if both SDR and SNR are close to their respective theoretically maximal values. However, in this paper we do not have an accurate $C^*$ for our data sets (although it is still very interesting to compare $C^*$ with $C^{\cup}$), hence we omit the derivation of the theoretical SDR and SNR.

In addition, we can perform receiver operating characteristic (ROC) curve analysis [40], [41], which involves computing the true-positive rate (TPR) and the false-positive rate (FPR) while varying algorithm parameters, such as the BT threshold. TPR indicates the sensitivity of the method, and the false-negative rate $(\mathrm{FNR} = 1 - \mathrm{FPR})$ indicates the specificity of the method. We can write TPR and FPR in terms of SDR and SNR

$$\mathrm{TPR} = \frac{\Omega(C^{\cup} \cap C^*)}{\Omega(C^*)} = \mathrm{SDR} \tag{15}$$

$$\mathrm{FPR} = \frac{\Omega(C^{\cup} \cap \overline{C^*})}{\Omega(\overline{C^*})} = \frac{\Omega(C^*)}{\Omega(\overline{C^*})} \cdot \frac{\mathrm{SDR}}{\mathrm{SNR}}. \tag{16}$$

These equations demonstrate that ROC-curve analysis is equivalent to SDR/SNR analysis. Therefore, we present only ROC curves in the following experimental sections.

## IV. Experiments for Linear Detection

This section addresses the problem of detection of a linear morphology-function association: a subject has a functional deficit whenever there is significant atrophy in a certain region. The goal of this experiment is to establish that, despite discretization, BMA has high sensitivity for the detection of morphology-function associations, approximating that of a standard statistical method, such as the TT.

### A. Data

For this experiment we used a set of simulated cerebral-atrophy MR images, which were based on T1-weighted gradient-echo SPGR images of 11 normal elderly subjects (average age is 70.1 years, standard deviation 5.9). We selected two gyri, the right precentral gyrus (PCG) and the left superior temporal gyrus (STG), in all subjects; both gyri were manually defined using the DISPLAY software package distributed by the Brain Imaging Center, Montreal Neurological Institute. We then introduced a 30% uniform contraction of the labeled gyri, and created 11 additional images with localized atrophy in these gyri. We called the labeled region of each subject the *atrophy mask*. For each subject, we called the image without atrophy the $\mathrm{t}_1$ image, and we called the image with simulated atrophy the $\mathrm{t}_2$ image. These simulated data are similar to those expected in a longitudinal study, because each ($\mathrm{t}_1$ and $\mathrm{t}_2$) pair of images belongs to the same subject, the only difference

between the two being localized atrophy. These 22 images, as well as the corresponding atrophy masks, have also been used in [6].

We registered these 22 3-D images with the STAR algorithm, generating RAVENS maps. The size of each RAVENS map was $256 \times 256 \times 129$. The voxel resolution for each spatial dimension was 0.9375 mm. As explained in Section II-A, the 3-D elastic warping transform in the STAR algorithm preserves the brain mass of each image volume; therefore, atrophic regions in a subject's $\mathrm{t}_2$ image had lower mean intensity than did the corresponding regions in that subject's $\mathrm{t}_1$ image. Due to memory limitations of our workstation, we down-sampled each image by a factor of 2, and cropped all images to the largest brain-region bounding box across all images. Each of the smaller images ($74 \times 91 \times 65$ voxels) contained $\Omega(C^I) = 231\,091$ voxels. For each subject, we repeated the warping, down-sampling and cropping procedure (with the same parameters) on the corresponding atrophy mask and obtained 11 RAVENS maps of atrophy masks (we still call them atrophy masks for convenience). These atrophy masks are slightly different from each other; therefore, we superimposed these atrophy masks and binarized them with the threshold $5.5\ (= 11/2)$ to generate a binary "ground truth" atrophy volume, $C^*$.

To correct registration errors, we applied an isotropic Gaussian-smoothing kernel to these images, as is customary in voxel-based morphometry [1], [8], [9], [16]. We applied the same smoothing kernel to $C^*$, in order to generate a ground-truth mask for the results obtained from a statistical analysis applied to smoothed images. We previously found the optimal diameter, in the sense of atrophy detection, for these data set to be 9 mm [6]; therefore, we show experimental results using a 9-mm smoothing kernel only, although we have tested other smoothing kernels, as reported below. In this context, $\Omega(\text{nonsmoothed } C^*) = 3888 < \Omega(\text{smoothed } C^*) = 10\,670$. Of note, $C^*$ is the thresholded mean value of the original atrophy masks, however some brain regions other than $C^*$ will also have significant intensity differences. Hence, for this data set, neither the nonsmoothed $C^*$ nor the smoothed $C^*$ is completely accurate, although they do indicate the locations of greatest morphological change.

We then applied the subtraction and binarization preprocessing steps, as shown in Fig. 1, to the smoothed images, to generate 11 binary maps, and included a value for $f$ for each subject, to indicate whether that subject had a functional deficit. In addition to the binary maps corresponding to abnormal function, we required binary maps for subjects with normal function. Hence, we created an additional 11 binary maps in which all voxels, and $f$, are normal (i.e., no atrophy in $\mathrm{t}_2$ images), assuming the value zero. Another reason for these zero maps is due to Gaussian smoothing. Because the Gaussian-smoothing operation includes a large region around each voxel (i.e., 9-mm kernel diameter relative to 0.9375-mm voxel size), the value of each smoothed voxel is calculated based on the values of hundreds of neighboring voxels. If there were no atrophy in $\mathrm{t}_2$ images, we would expect that after smoothing, a given voxel's intensity in the $\mathrm{t}_1$ and $\mathrm{t}_2$ images would be approximately equivalent. Therefore, after subtraction and binarization, most voxel values would be 0.

The entire data set, with 22 simulated binary maps, consisting of 11 cases for which $f = \mathrm{Abnormal}$ and 11 cases for which $f = \mathrm{Normal}$, were used as input to the BMA algorithm.

### *B. Results*

We compared our Bayesian methods with the standard paired $t$-test (PT) method (provided in SPM99). For simplicity, we call our methods BT and BC, although they differ only in Step 5 in the algorithm listed in Fig. 2.

Fig. 5(a) and (b) shows intensity plots of the average binary maps at the locations of simulated atrophy. Comparing the smoothed ground truth for the STG in Fig. 5(c) with results in Fig. 5(b), we see that there are false-positive regions outside $C^*$ (the ground-truth region) and false-negative regions inside $C^*$.

In Fig. 5(c)–(f), we show atrophy-detection results versus ground truth for the association between the status of $f$ and the status of the STG. The parameters are $P_{\mathrm{BT}} = 0.8$ for BT; initial number of clusters $r_L = 7$, and maximum number of $\mathrm{iterations} = 100$ for BC; and significance threshold $t_{\mathrm{PT}} = 5$ for the PT $t$-statistic map. Note that in each step of BC, although $r_L$ was initialized to 7, the BC algorithm returned only those voxels with the same state as the current representative voxel. The colored region in Fig. 5(c) is $C^*$ for STG atrophy, behind which the grayscale image corresponds to a randomly chosen subject's brain, which serves as an anatomical reference. Comparing Fig. 5(d)–(f) with Fig. 5(b), we see that both BT and BC correctly identify most ground-truth voxels, whereas PT finds only a subset of $C^*$, even with a relatively low threshold $t_{\mathrm{PT}} = 5$.

As expected, the results of BT and BC overlap, except for a few noisy regions in BC. This noise exists partly because BC is based on Monte Carlo iterations, so a few pseudocases of the latent variable may change states even when the maximum number of iterations has been reached.

In Fig. 6(a) and (c), we show ROC curves for both nonsmoothed and smoothed gold-standard regions, respectively. Fig. 6(b) and (d) show the portions of the corresponding ROC curves in Fig. 6(a) and (c) for which we have restricted the analyses to parameter values for each method that would be used in practice. As seen in Fig. 6(b) and (d), for comparable FPRs, BT is more sensitive than PT, although the difference is small. Furthermore, although BC always has a higher FPR than BT and PT, BC also has a higher TPR than BT and PT. When $P_{\mathrm{BT}}$ is lowered to 0.6, BT produces results very similar to those of BC. However, for BT we do not know the optimal threshold beforehand, whereas BC does not depend on a user-defined threshold. Our current implementation of BC requires a value for $r_L$; however, we could automate this choice, because when $r_L$ is large enough (for example, $>4$ for these experiments), the resulting clusters are similar in each trial. The difference in the FPRs for BC for different values of $r_L$ is primarily due to a value for $r_L$ that is too small, which forces incorrect clustering, whereas a large $r_L$ value distributes errors across clusters. As for PT, changing the threshold $t_{\mathrm{PT}}$ to yield higher TPRs in Fig. 6(b) and (d) results in an unacceptable value for $t_{\mathrm{PT}}$ ($t_{\mathrm{PT}} = 3$ is too small to reflect significant variations in the $t$-map) for realistic applications. However, when $t_{\mathrm{PT}}$ is set to values typically used in practice (e.g., 10), only a small fraction

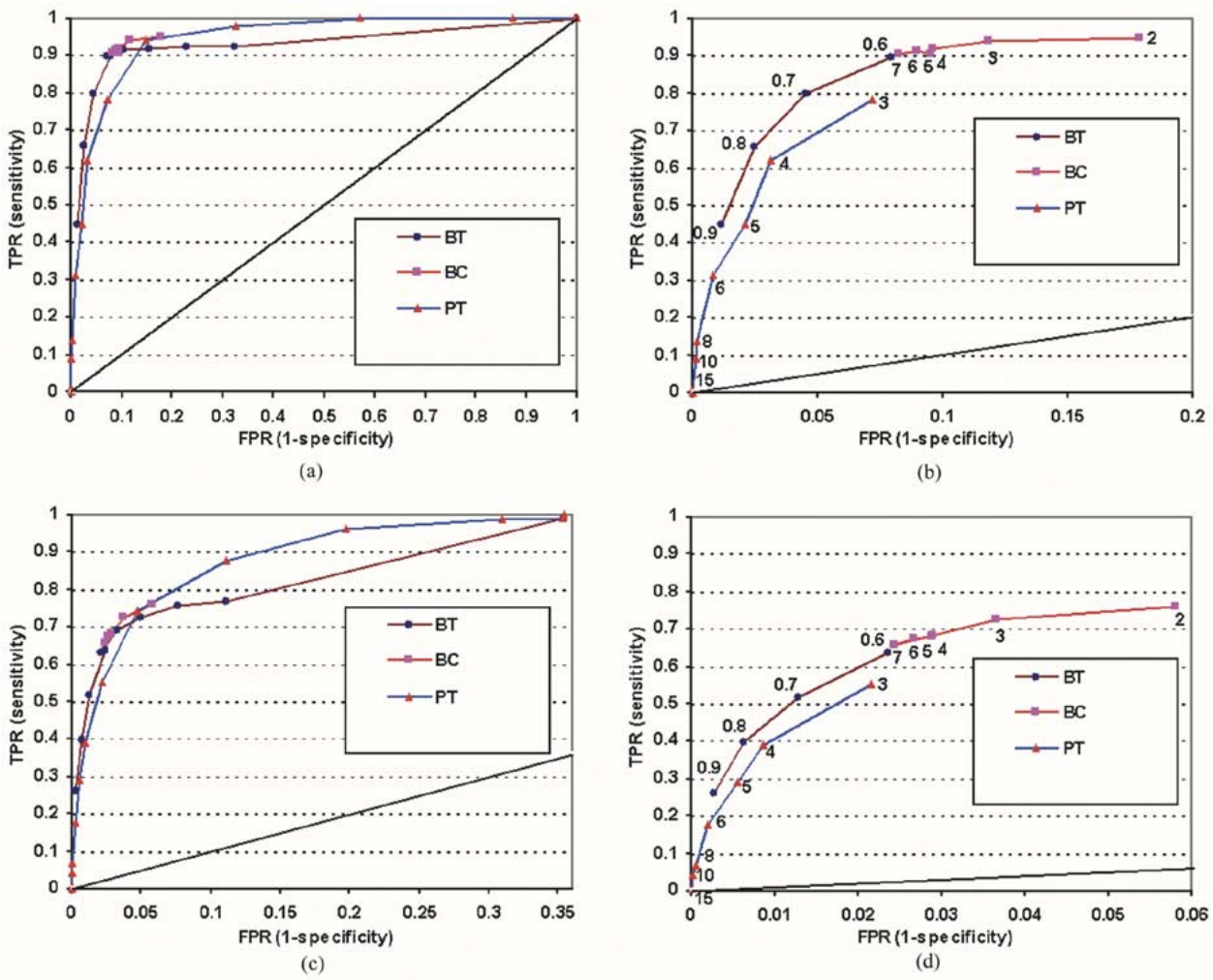


Fig. 6. ROC curves for the PT, BT, and BC methods. Labels for the PT method are t-score thresholds ($t_{\mathrm{PT}}$); labels for the BT method are conditional-probability thresholds ($P_{\mathrm{BT}}$), as defined in (9); labels for the BC method are initial numbers of clusters ($r_L$), as defined in (11). (a) ROC curve (nonsmoothed $C^*$) for all parameters; (b) ROC curve (nonsmoothed $C^*$) for meaningful parameters; (c) ROC curve (smoothed $C^*$) for all parameters; (d) ROC curve (smoothed $C^*$) for meaningful parameters.

of ground-truth voxels (<10%) are detected. Thus, settings for which PT has low FPR also result in low TPR. For example, when $t_{\mathrm{PT}} = 15$, the corresponding TPR is 0.0034, which means that only about $0.0034 \times \Omega(\mathrm{smoothed}\, C^*) \approx 36$ voxels are detected. In addition, PT shares BT's drawback of requiring a user-set threshold.

One problem that all three methods share is a relatively high FPR. Recall that $\Omega(\mathrm{smoothed}\ C^*) = 10\,670$, and $\Omega(C^I) = 231\,091$; thus, if we wanted no more than 10% of detected voxels to represent false-positive detections, this would require an FPR of $(1067/231\,091 - 10\,670) \approx 0.004$, which is not achieved by any of the three methods at realistic parameter settings. However, Fig. 5 demonstrates that the false-positive voxels do not impede visual interpretation of detection results; it is clear that all three methods detect both the right precentral gyrus and the left STG.

## V. Experiments for Nonlinear Detection

In this section, we further evaluate the performances of these methods on a more difficult problem, in which ventricular enlargement, a proxy for cerebral atrophy, in only a specific nonlinear combination of locations, is associated with a functional deficit. This structure-function relationship is one that cannot be captured by standard linear statistical tests, and analysis of these data, therefore, demonstrates the principal strength of our approach.

### A. Data

For this experiment, we used 168 T1-weighted SPGR images obtained from 84 normal elderly subjects. These subjects have different degrees of cerebral atrophy. For each subject, there are two images that were scanned with a 5-year interval between times $\mathrm{t}_1$ and $\mathrm{t}_2$. We manually segmented these images and obtained a lateral-ventricle (LV) mask for each image. For the purposes of this experiment, we called the smaller LV image of a subject the $\mathrm{t}_1$ LV, and the larger LV image the $\mathrm{t}_2$ LV; that is, we arranged the data to make all $\mathrm{t}_2$ LVs larger than the corresponding $\mathrm{t}_1$ LVs. On average, $\mathrm{t}_2$ LVs were approximately 20% larger than $\mathrm{t}_1$ LVs. We normalized these LV images using the STAR algorithm, and obtained two RAVENS maps for each subject. Because cerebral atrophy is indirectly manifested as LV enlargement, the $\mathrm{t}_2$ RAVENS maps have higher values than the

TABLE I
DESIGN OF A NONLINEAR-ASSOCIATION DATA SET, IN WHICH ENLARGEMENT OF BOTH LATERAL VENTRICLES IS ASSOCIATED WITH A FUNCTIONAL DEFICIT, HOWEVER, ONLY LEFT LATERAL-VENTRICULAR ENLARGEMENT EXHIBITS A LINEAR ASSOCIATION WITH THE FUNCTION VARIABLE. L = LEFT LATERAL VENTRICLE, R = RIGHT LATERAL VENTRICLE, $f$ = FUNCTION VARIABLE, A = ABNORMAL (SIZE OR FUNCTION), AND N = NORMAL (SIZE OR FUNCTION)

| Group | Number of Subjects | Pattern Name | Lateral-Ventricular Status | | Function Variable Status | Fraction of Subjects with this Pattern that are Abnormal | Other Statistics |
|---|---|---|---|---|---|---|---|
| | | | Left (L) | Right (R) | | | |
| 1 | 13 | $P_{NN}$ | N | N | N | $P(f = \text{A} \mid P_{NN}) =$ | $P(f = \text{A}) = 0.50$ |
| 2 | 16 | | | | A | $16 / (13 + 16) = 0.55$ | $P(\text{L=A} \mid f = \text{A}) = 0.50$ |
| 3 | 8 | $P_{AN}$ | A | N | N | $P(f = \text{A} \mid P_{AN}) =$ | $P(\text{L=A} \mid f = \text{N}) = 0.26$ |
| 4 | 5 | | | | A | $5 / (8 + 5) = 0.38$ | $P(\text{R=A} \mid f = \text{A}) = 0.50$ |
| 5 | 19 | $P_{NA}$ | N | A | N | $P(f = \text{A} \mid P_{NA}) =$ | $P(\text{R=A} \mid f = \text{N}) = 0.50$ |
| 6 | 5 | | | | A | $5 / (19 + 5) = 0.21$ | $P(f = \text{A} \mid \text{L} = \text{A}) = 0.68$ |
| 7 | 2 | $P_{AA}$ | A | A | N | $P(f = \text{A} \mid P_{AA}) =$ | $P(f = \text{A} \mid \text{L} = \text{N}) = 0.40$ |
| 8 | 16 | | | | A | $16 / (2 + 16) = 0.89$ | $P(f = \text{A} \mid \text{R} = \text{A}) = 0.50$ |
| | | | | | | | $P(f = \text{A} \mid \text{R} = \text{N}) = 0.50$ |

respective $t_1$ maps. Then we defined the left and right LVs in the spatially normalized RAVENS maps.

Our aim in this experiment was to use synthetic data to evaluate the performance of the BC, BT, and TT (i.e., standard $t$-test) approaches on the following nonlinear detection problem: the function variable $f$ has associations with the states of both LVs, however only left lateral-ventricular enlargement is linearly associated with $f$. We designed the data set, swapping $t_1$ and $t_2$ maps when necessary, and selecting values for $f$, such that right lateral-ventricular enlargement has no univariate association with the function variable; that is, $P(\text{Enlarged right LV}|f = \text{Abnormal}) = P(\text{Normal right LV}|f = \text{Normal}) = 0.5$. Since we had ensured that all $t_2$ LV RAVENS maps have higher values than the corresponding $t_1$ maps, we next constructed 8 groups of images that displayed different patterns of ventricular enlargement.

To construct these 8 groups, we performed the following four steps:

1) We generated four ventricular-enlargement patterns, $P_{NN}$, $P_{AN}$, $P_{NA}$, and $P_{AA}$, which stand for normal lateral ventricles, left lateral-ventricular enlargement only, right lateral-ventricular enlargement only, and bilateral lateral-ventricular enlargement, respectively. These patterns are shown in the third, fourth, and fifth columns of Table I. Since we refer to ventricular enlargement as the situation in which the $t_2$ ventricular RAVENS map has a larger value than the corresponding $t_1$ ventricular RAVENS map, to create the normal lateral ventricle, we swapped the $t_1$ and $t_2$ lateral-ventricular RAVENS maps to guarantee that the $t_1$ map has the larger value.
2) We randomly assigned the 84 subjects among 8 groups, as described in Table I. For example, group 3 has 8 subjects, each of which has pattern $P_{AN}$, i.e., abnormal (enlarged) left lateral ventricle $(\text{L} = \text{A})$ and normal right lateral ventricle $(\text{R} = \text{N})$.
3) For each group of subjects, we swapped the $t_1$ and $t_2$ ventricular RAVENS maps according to the patterns of ventricular enlargement listed in the fourth and fifth columns of Table I, and set the respective function-variable state according to the sixth column of Table I. For each case in which lateral-ventricular enlargement was not present (i.e., normal), we swapped maps, to ensure that the $t_2$ map had a smaller ventricle than the $t_1$ map. For each case in which there was lateral-ventricular enlargement, we performed no swap, since we started with cases in which the $t_2$ ventricles were larger than the $t_1$ ventricles. For example, for both lateral ventricles in the first group (with 13 subjects), we swapped the $t_1$ and $t_2$ ventricular RAVENS maps so that $t_2$ maps ended up having smaller ventricles than those in the corresponding $t_1$ maps, thus ensuring that this group displayed no ventricular enlargement. We set the function variable $f$ for this group to N (normal). Another example is the 6th group (with 5 subjects), for which we swapped only the left lateral ventricular $t_1$ and $t_2$ RAVENS maps, but did not swap the right lateral-ventricular maps, thus ensuring that the data for this group displayed only right-sided ventricular enlargement. We set the function variable $f$ for this group to A (abnormal).
4) From the statistics listed in the seventh and eighth columns in Table I, we designed a simulated data set in which morphological changes in both lateral ventricles have strong associations with the presence of a functional deficit. When both lateral ventricles are enlarged, i.e., pattern $P_{AA}$, a functional deficit is highly likely $(P(f = \text{A}|P_{AA}) = 0.89)$, whereas unilateral ventricular enlargement is associated with a much lower risk of functional deficit $(P(f = \text{A}|P_{AN}) = 0.38$ and $P(f = \text{A}|P_{NA}) = 0.21)$. Of note, there is no univariate association between right lateral-ventricular enlargement and the presence of a functional deficit $(P(f = \text{A}|\text{R} = \text{A}) = 0.5$ and $P(f = \text{A}|\text{R} = \text{N}) = 0.5)$, whereas left lateral-ventricular enlargement is associated with the presence of a functional deficit $(P(f = \text{A}|\text{L} = \text{A}) = 0.68$ and $P(f = \text{A}|\text{L} = \text{N}) = 0.40)$.

Subsequently, we obtained 84 binary maps after subtraction (we used $t_2 - t_1$ here, in contrast to the experiment in the last section) and binarization steps. As indicated by the state of $f$ for each simulated subject in Table I, we have 42 normal subjects and 42 abnormal subjects, with functional deficits arising under specific spatial patterns of cerebral atrophy (and, therefore, ventricular enlargement). The average binary map for these 84 subjects is shown in Fig. 7(a).

### *B. Results*

We applied BT, BC, and TT to these data. From the design of this experiment, we knew that the states of both lateral ventri-

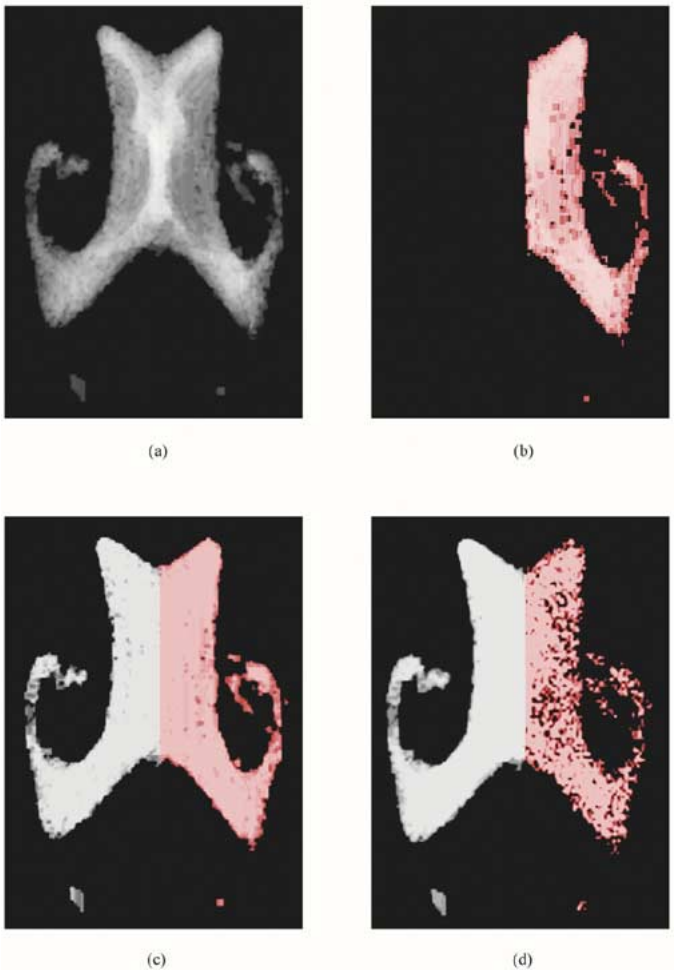


Fig. 7. Average binary map of the lateral ventricles over all subjects, and detection results of TT, BT, and BC for nonlinear morphology-function associations. The red and white regions correspond to two clusters found by BT and BC. The result for TT is painted red because it corresponds to the red regions in BT and BC results. (a) Average binary lateral ventricles over all 84 subjects (axial view from below); (b) TT ($p_{\mathrm{TT}} = 0.05$) (axial view from below); (c) BT ($p_{\mathrm{BT}} = 0.8$) (axial view from below); (d) BC ($r_L = 3$, $\mathrm{iteration} = 50$) (axial view from below).

cles are jointly associated with the functional deficit, but that the state of the right lateral ventricle is not linearly associated with $f$. Hence, a standard linear statistical test, such as the $t$-test, barely detects right lateral-ventricular enlargement, although as expected it detects left lateral-ventricular enlargement. Fig. 7(b) shows the results for the TT method, where only the left lateral ventricle is detected on the thresholded $p$-map (with threshold $p$-value $= 0.05$). In contrast, results for BMA, shown in Fig. 7(c) and (d) (with parameters $p_{\mathrm{BT}} = 0.8$ and $r_L = 3$), both BT and BC detect morphology-function associations for both lateral ventricles. These results confirm our expectation that BT and BC can detect nonlinear morphology-function associations.

In the BC results shown in Fig. 7(d), left lateral-ventricular voxels do not appear as bright as those for the BT results in Fig. 7(c), due to the imperfect convergence of the Monte Carlo iteration in BC. Despite this, the results of BT and BC agree well with each other.

## VI. Discussion

We have demonstrated that Bayesian methods for morphology-function analysis can detect linear and nonlinear associations among voxels and function variables. Although standard methods of analysis, such as the production of $t$-maps, may detect lower order associations as well as BMA, they may not detect nonlinear multivariate associations as well as BMA. In addition, BMA is generally applicable to similar problems, such as detection of morphology-demographic (e.g., age, socioeconomic status) associations. Although it might be reasonable to expect PT to outperform BMA for linear-association detection, it may be the case that discretization of the voxel data increases statistical power by eliminating Gaussian noise.

For the data in Section IV-A, in which we used a 9-mm-diameter Gaussian smoothing kernel, the intensity difference between smoothed $\mathrm{t}_1$ and $\mathrm{t}_2$ voxels was very small, typically within the range $[-10, 10]$. Hence, the naïve binarization threshold 0 appeared to be necessary, because a threshold larger than 0 would have resulted in too much signal loss and, thus, very low TPR. One solution would be to increase the contrast of the comparison maps. For example, we could voxel-wise divide the $\mathrm{t}_2$ image by the $\mathrm{t}_1$ image; this operation would produce ratio maps with much higher contrast than difference maps. This higher contrast would allow us to manually or automatically choose a better binarization threshold to maximize theoretical SNR and/or SDR. Furthermore, in those cases for which we have better ground-truth information, we could calculate the best SNR and/or SDR for many different Gaussian-kernel diameters to find the optimal kernel.

We had previously evaluated the image data presented here, with regard to the effects of smoothing on subsequent morphometric analysis, and found that a 9-mm Gaussian kernel provides the best results [6]. However, we expect that data acquired from different scanners, using different sequences, will require different smoothing kernels, or perhaps no smoothing at all. In addition, there is clearly an interaction between the characteristics of the registration algorithm applied to the image data, and the optimal smoothing kernel. The principal limitation of the application of larger smoothing kernels is loss of sensitivity for morphological changes in small structures. Thus, if the goal of morphometry is to analyze changes in a small structure, such as the hippocampus, the optimal smoothing kernel will almost certainly be much smaller than 9 mm; in fact, it may be optimal to forego the increased SNR that smoothing provides, in order to prevent loss of sensitivity due to averaging with large number of neighboring (i.e., nonhippocampal) voxels. It is important to note that these smoothing effects are not unique to BMA; they would obtain regardless of which morphometry-function analysis paradigm is subsequently applied.

In addition to smoothing, registration of image data to a standard is a crucial preprocessing step that can affect the accuracy (and statistical power) of subsequent morphometry; in this sense, our results are conditioned on the accuracy of our registration algorithm. In addition, we have emphasized the importance of mass-preserving transformation, which allows us to perform RAVENS analysis; a registration method that does not preserve mass would confound subsequent morphometry. Although we have found, in previous validation experiments, that our registration methods have accuracies comparable to or better

than those generally reported, we expect that for morphometry of very small structures, such as the hippocampus or hypothalamus, additional constraints on registration (perhaps provided by identification of fiducial points) will be required to provide sufficient registration accuracy. Furthermore, we note that we expect to be cautious when interpreting the results of clinical studies in which we have registered images, because regions in which no morphological effects are detected might be involved to an extent not observable given registration accuracy.

Although we present experiments for binary morphometric variables only, BMA can also analyze multi-state discrete variables. This choice of binary voxels variables confers the advantage of voxel states having clear meanings, (i.e., atrophy/normal), simplifies thresholding in image preprocessing, and simplifies implementation of the BT algorithm. However, there are no constraints on the number of states for any variable in (5), or for the BC method, so there is no inherent reason that we could not increase the number of states considered for each voxel variable. For example, we could threshold 3 states for each voxel variable, representing contraction, no change, and expansion. This or similar modifications might be useful to detect more general morphology-function changes. In conjunction with this extension, we would threshold RAVENS maps using an unsupervised clustering method, thus providing a sound basis for thresholding of voxel-wise volume changes. Even for binary variables, one could argue that thresholding at 0 (i.e., enlargement/no enlargement) might not be optimal; for example, if there were a structure that did not cause functional impairment until it lost 25% of its neurons, thresholding to detect any volume loss would reduce sensitivity to functional impairment.

Modifying BMA to support the analysis of two or more function variables is straightforward; in particular, BMA would return a BN in which a subnetwork is generated for each function variable. As in the previous paragraph, there is no inherent limitation in (5), or in the BT or BC methods, aside from computational burden, on the number of function (or voxel) variables. In addition, adding a final search for additional or redundant edges among cluster and function variables would not significantly increase the computational requirements of BMA, since the number of variables (clusters + function variables) is greatly reduced relative to the number of voxels provided as input.

We have assumed that we need not compare all possible BNs to generate an adequate network structure. Regardless of whether this assumption is correct, the model-selection algorithm will produce the correct first representative voxel. However, this assumption is probably incorrect for networks with more than one representative voxel, due to the limitations of our heuristic-search algorithm. The greatest limitation of our search algorithm would be in detection of the probabilistic equivalent of an AND gate; the existence of even a slight lower order association between representative voxel(s) and a function variable would enable BMA to detect the next higher order associations, if they exist. Extending the depth of search to $n$th higher order associations would increase the computational complexity of search polynomially (degree $n$), and exhaustive search of the space of models, as described previously, is exponential in the number of variables. Empirically, we have observed that a reasonable upper bound on the order of multivariate associations is 4 (i.e., 4 parent nodes for a child node); therefore, we could expect to capture almost all clinically important multivariate models if we extended BMA's search depth to 4, which would result in a quartic increase in search time; although this increase would render the algorithm intractable for even $10^4$ voxels, we could greatly reduce computational requirements by clustering voxels into groups based on their states in each case, selecting a representative voxel for each cluster, and performing higher order search on representative voxels and function variables. In addition, we have found that most of these higher order associations also manifest lower order associations and are thus amenable to greedy search, which would obviate exhaustive fourth-order search.

Another related challenge is that the BN structures we can generate might not be sufficiently complex to represent associations among representative voxels and a function variable. For example, it is possible that one or more edges *among* representative voxels could yield better models, as judged by (5); our current search algorithm does not examine such models, although such an extension would be straightforward, at the cost of increasing computational burden. Because our primary goal is the identification of associations between voxel clusters and the function variable, our current implementation can achieve this goal even if it does not compare associations among representative voxels.

An additional potential limitation of the BMA approach is its requirement that variables be discrete. There is always the risk of loss of information during discretization, and we expect that when this is the case, statistics based on continuous variables will perform better than BMA. Fortunately, in many applications, the data can be discretized without much loss of information; for example, in the setting of brain morphometry, loss, gain, or no change in volume are important biological categories, which can guide discretization. Furthermore, proper discretization may yield more robust analysis. For example, although density estimation methods such as Parzen windows [42] might be used to approximate the dependency between voxel variables and the function variable, these methods require a large number of samples (subjects) and suffer from high computational complexity. For our linear-detection experiment, in which there are only 22 images, it seems that discretization, rather than density estimation, is a more reasonable choice from the perspective of robust analysis. For our nonlinear-detection experiment, where there are 84 images, the enormous computational burden of density estimation makes it impractical. If the data were drawn from a multivariate Gaussian distribution, we could implement continuous-variable BNs [43] under the BMA framework to analyze these data; however, it is our experience that these image and function data rarely assume a multivariate Gaussian distribution; we, therefore, must transform the offending variable(s), or apply nonparametric analytic methods, as alternatives to discretization and application of BMA.

Our experiments demonstrate that, although the FPRs of BC, BT, and PT may be greater than we would wish, these methods, and BMA in particular, yield clusters that clearly correspond to the original structures (e.g., PCG) in which cerebral atrophy (or ventricular enlargement) was induced, when visualized, as in Fig. 5. The principal reason for this discrepancy lies in the spatial distributions for false-positive and true-positive voxels; the former will be scattered, and the latter will be spatially clus-

tered in one or more structures. Clearly, the addition of spatial information into BMA would decrease the FPR, and would thus increase the utility of BMA.

Accordingly, we plan the following major areas of further development:

1) *Spatial information*: incorporating spatial information into the algorithm should increase SNR. For example, an isolated bright point on the average binary map is usually noise. Therefore, it may be reasonable to eliminate such voxels, using erosion, neighborhood filtering, or probability distributions over the spatial distribution of regions associated with function variables. More sophisticated methods for incorporating spatial information, such as Gaussian fields [44], or Markov random fields [45], could be used to ensure spatial contiguity of clusters or components of clusters, at the cost of increased computational burden.
2) *Effects of smoothing kernel*: Given the effects of smoothing on subsequent morphometry [6], quantification of the interactions among registration methods, data-acquisition parameters (e.g., modality, sequence), sizes of structures being analyzed, and smoothing-kernel sizes is a prerequisite for widespread application of morphometry.
3) *Number of cases*: Although we omit the derivation here, analytically the number of cases will play an important role in improving both SDR and SNR, as one would expect. Because we expect to work with images from hundreds, or perhaps thousands, of subjects, we have emphasized scalability in the design of BMA.
4) *Different metric and model-search heuristics*: Although we employ the Bayesian metric in (5) and propose the heuristic method in Fig. 2, the proposed BMA paradigm itself is independent of particular BN metrics and heuristic model-search methods. We plan to investigate several other approaches described in the BN learning literature, including alternative choices for the hyperparameters $\alpha_{\mathrm{ijk}}$, nonuniform prior distributions over BN models, and information-theoretic metrics [20], [24], [26], [27], [29].

To extend our validation of these methods, we are currently evaluating data from the Baltimore Longitudinal Study of Aging, to determine whether there are morphological indicators of cognitive impairment.

## VII. Conclusion

We have described a framework for morphology-function analysis, based on a Bayesian-network model of associations among image and function variables. The algorithms based on this framework generate sets of voxels whose members have similar probabilistic associations with the function variable(s). Two methods implemented within this framework, BT and BC, are examples of how this framework can be used to generate equivalent-voxel sets. The BT method is simpler and faster, however it requires a predefined threshold for determining whether two voxels have similar conditional probability distributions given the function variable. The BC method utilizes a latent-variable Bayesian-network model, and the Monte Carlo algorithm, to generate equivalent-voxel sets. BC takes more time than BT, however the former does not require a user-defined threshold, which is the principal limitation of the latter.

We compared these Bayesian methods to the $t$-test and paired $t$-test for both linear and nonlinear morphology-function detection problems. Our methods succeeded in both cases, whereas the $t$-test failed to detect nonlinear associations.

## Acknowledgment

The authors would like to thank D. Chickering for assistance with latent-variable BC methods. They would also like to thank the reviewers for their insightful comments.

## References

[1] J. Ashburner and K. J. Friston, "Voxel-based morphometry-the methods," *Neuroimage*, vol. 11, pp. 805–821, 2000.
[2] F. L. Bookstein, "Principal warps—thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 567–585, June 1989.
[3] C. Davatzikos, "Spatial transformation and registration of brain images using elastically deformable models," *Comput.Vis. Image Understanding*, vol. 66, pp. 207–222, 1997.
[4] ——, "Mapping image data to stereotaxic spaces: applications to brain mapping," *Human Brain Mapping*, vol. 6, pp. 334–338, 1998.
[5] ——, "Spatial normalization of 3D brain images using deformable models," *J. Comput. Assist. Tomogr.*, vol. 20, pp. 656–665, 1996.
[6] C. Davatzikos, A. Genc, D. Xu, and S. M. Resnick, "Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy," *Neuroimage*, vol. 14, pp. 1361–1369, 2001.
[7] C. Davatzikos, M. Vaillant, S. M. Resnick, J. L. Prince, S. Letovsky, and R. N. Bryan, "A computerized approach for morphological analysis of the corpus callosum," *J. Comput. Assist. Tomogr.*, vol. 20, pp. 88–97, 1996.
[8] K. J. Friston, J. Ashburner, C. D. Frith, J. B. Poline, J. D. Heather, and R. S. J. Frackowiak, "Spatial registration and normalization of images," *Human Brain Mapping*, vol. 3, pp. 165–189, 1995.
[9] K. J. Friston, A. P. Holmes, K. J. P. Worsley, J. B. C. D. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging—a general linear approach," *Human Brain Mapping*, vol. 2, pp. 189–210, 1995.
[10] C. Gaser, H. P. Volz, S. Kiebel, S. Riehemann, and H. Sauer, "Detecting structural changes in whole brain based on nonlinear deformations—application to schizophrenia research," *Neuroimage*, vol. 10, pp. 107–113, 1999.
[11] J. C. Gee, M. Reivich, and R. Bajcsy, "Elastically deforming 3D atlas to match anatomical brain images," *J. Comput. Assist. Tomogr.*, vol. 17, pp. 225–236, 1993.
[12] A. F. Goldszal, C. Davatzikos, D. L. Pham, M. X. Yan, R. N. Bryan, and S. M. Resnick, "An image-processing system for qualitative and quantitative volumetric analysis of brain images," *J. Comput. Assist. Tomogr.*, vol. 22, pp. 827–837, 1998.
[13] P. Thompson and A. W. Toga, "A surface-based technique for warping 3-dimensional images of the brain," *IEEE Trans. Med. Imag.*, vol. 15, pp. 402–417, Aug. 1996.
[14] P. M. Thompson and A. W. Toga, "Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations," *Med. Image Anal.*, vol. 1, pp. 271–294, 1997.
[15] P. M. Thompson, D. MacDonald, M. S. Mega, C. J. Holmes, A. C. Evans, and A. W. Toga, "Detection and mapping of abnormal brain structure with a probabilistic atlas of cortical surfaces," *J. Comput. Assist. Tomogr.*, vol. 21, pp. 567–581, 1997.
[16] F. G. Woermann, S. L. Free, M. J. Koepp, J. Ashburner, and J. S. Duncan, "Voxel-by-voxel comparison of automatically segmented cerebral gray matter: A rater-independent comparison of structural MRI in patients with epilepsy," *Neuroimage*, vol. 10, pp. 373–384, 1999.
[17] D. Shen, E. H. Herskovits, and C. Davatzikos, "An adaptive-focus statistical shape model for segmentation and shape modeling of 3-D brain structures," *IEEE Trans. Med. Imag.*, vol. 20, pp. 257–270, Apr. 2001.

[18] A. R. McIntosh, F. L. Bookstein, J. V. Haxby, and C. L. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *Neuroimage*, vol. 3, pp. 143–157, 1996.
[19] C. Glymour and G. F. Cooper, *Computation, Causation, and Discovery*. Menlo Park, CA: AAAI Press, 1999.
[20] D. Heckerman, "Bayesian networks for data mining," *Data Mining and Knowledge Discovery*, vol. 1, pp. 79–119, 1997.
[21] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.
[22] C. G. Goetz and E. J. Pappert, *Textbook of Clinical Neurology*. Philadelphia, PA: Saunders, 1999.
[23] H. Kiiveri, T. Speed, and J. B. Carlin, "Recursive causal models," *J. Austr. Math. Society—A*, vol. 36, pp. 30–52, 1984.
[24] R. Bouckaert, "Properties of measures for Bayesian belief network learning," presented at the *10th Conf. Uncertainty in Artificial Intelligence*, Seattle, WA, 1994.
[25] J. Cheng, D. A. Bell, and W. Liu, "Learning belief networks from data: An information theory based approach," presented at the *Proc. 6th ACM Int. Conf. Information and Knowledge Management*, Las Vegas, Nevada, 1997.
[26] W. Buntine, "A guide to the literature on learning probabilistic networks from data," *IEEE Trans. Knowledge Data Eng.*, vol. 8, pp. 195–210, 1996.
[27] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: the combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.
[28] G. F. Cooper and E. H. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–347, 1992.
[29] E. H. Herskovits, "Computer-based probabilistic-network construction," Ph.D. dissertation, Medical Informatics, Stanford Univ., Stanford, CA, 1991.
[30] D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell, "Bayesian analysis in expert systems," *Statist. Sci.*, vol. 8, pp. 219–247, 1993.
[31] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *J. Roy. Statist. Soc. B*, vol. 41, pp. 113–147, 1979.
[32] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1985.
[33] R. W. Robinson, "Counting unlabeled acyclic digraphs," in *Combinatorial Mathematics*, C. H. C. Little, Ed. Berlin, Germany: Springer-Verlag, 1977, vol. 622, Lecture Notes in Mathematics, pp. 28–43.
[34] P. Cheeseman and J. Stutz, "Bayesian classification (autoclass): theory and results," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. Cambridge, MA: AAAI/MIT Press, 1996, pp. 153–180.
[35] D. M. Chickering and D. Heckerman, "Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables," *Machine Learning*, vol. 29, pp. 181–212, 1997.
[36] C. C. Clogg, "Latent class models," in *Handbook of Statistical Modeling for Social and Behavioral Sciences*, G. Arminger, C. C. Clogg, and M. E. Sobel, Eds. New York: Plenum, 1995, pp. 311–359.
[37] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, 1984.
[38] S. Chib, "Marginal likelihood from the Gibbs output," *J. Amer. Statist. Assoc.*, vol. 90, pp. 1313–1321, 1995.
[39] A. E. Raftery, "Hypothesis testing and model selection," in *Markov Chain Monte Carlo in Practice*, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds. London, U.K.: Chapman and Hall/CRC, 1995, pp. 163–188.
[40] C. E. Metz, "Basic principles of ROC analysis," *Sem. Nucl. Med.*, vol. 8, pp. 283–298, 1978.
[41] ——, "ROC methodology in radiologic imaging," *Investigat. Radiol.*, vol. 21, pp. 720–733, 1986.
[42] E. Parzen, "Estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.
[43] D. Geiger and D. Heckerman, "Learning Gaussian networks," in *10th Conf. Uncertainty in Artificial Intelligence*, Seattle, WA, 1994.
[44] K. J. Friston, K. J. Worsley, R. S. J. Frackowiak, J. C. Mazziotta, and A. C. Evans, "Assessing the significance of focal activations using their spatial extent," *Human Brain Mapping*, vol. 1, pp. 214–220, 1994.
[45] N. A. C. Cressie, *Statistics for Spatial Data*. New York: Wiley, 1993.