Dynamic Causal Modeling Across Network Topologies

Shaza B. Zaghlool

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Engineering

Christopher L. Wyatt, Chair
Aloysius A. Beex
William T. Baumann
Stephen M. LaConte
Paul J. Laurienti

Feb 24, 2014
Blacksburg, Virginia

Keywords: Dynamic Causal Modeling, Missing Data, Graph Topology

Dynamic Causal Modeling Across Network Topologies

Shaza B. Zaghlool

(ABSTRACT)

Dynamic Causal Modeling (DCM) uses dynamical systems to represent the high-level neural processing strategy for a given cognitive task. The logical network topology of the model is specified by a combination of prior knowledge and statistical analysis of the neuro-imaging signals. Parameters of this a-priori model are then estimated and competing models are compared to determine the most likely model given experimental data. Inter-subject analysis using DCM is complicated by differences in model topology, which can vary across subjects due to errors in the first-level statistical analysis of fMRI data or variations in cognitive processing. This requires considerable judgment on the part of the experimenter to decide on the validity of assumptions used in the modeling and statistical analysis; in particular, the dropping of subjects with insufficient activity in a region of the model and ignoring activation not included in the model. This manual data filtering is required so that the fMRI model's network size is consistent across subjects.

This thesis proposes a solution to this problem by treating missing regions in the first-level analysis as missing data, and performing estimation of the time course associated with any missing region using one of four candidate methods: zero-filling, average-filling, noise-filling using a fixed stochastic process, or one estimated using expectation-maximization. The effect of this estimation scheme was analyzed by treating it as a preprocessing step to DCM and observing the resulting effects on model evidence. Simulation studies show that estimation using expectation-maximization yields the highest classification accuracy using a simple loss function and highest model evidence, relative to other methods. This result held for various dataset sizes and varying numbers of model choice. In real data, application to Go/No-Go and Simon tasks allowed computation of signals from the missing nodes and the consequent computation of model evidence in all subjects compared to 62 and 48 percent respectively if no preprocessing was performed. These results demonstrate the face validity of the preprocessing scheme and open the possibility of using single-subject DCM as an individual cognitive phenotyping tool.

Dedication

I dedicate this work to my dear and loving husband Khalid. I love you to eternity.

Acknowledgments

First and foremost, I would like to thank my adviser Dr. Chris Wyatt for supervising me and constantly guiding me in doing my PhD. He was very helpful through out all the stages and gave me great support. Being so knowledgeable and dedicated to his field gave me numerous benefits and I definitely could not have done any of this without him.

I would also like to thank my committee members for providing useful feedback on my work which paved the way to completing my dissertation. Thank you Dr. Beex, Dr. Laurienti, Dr. Baumann, and last but definitely not least, Dr. La Conte.

As for my family, words can not express the endless support I received from everyone. My dearest husband, Khalid, thank you for being so understanding and putting up with all I went through to do a PhD. I will not forget the number of times you would keep encouraging me even though I spent so many nights working late rather than spending time with you and constantly canceling out on our social calendar to study. Although most of the time you were actually playing video games while I suffered, I still could not have done this without your support. My dear parents, you have always been there for me, always believing in me. Thank you Daddy for encouraging me to do this and for being so confident in me. Thank you Mommy for always remembering me in your prayers and believing in me. Thank you Hadeel and Ehab for being the most amazing siblings who also undoubtedly believed in me.

I also want to give a special thanks to Dr. Sedki Riad and Dr. Yasser Hanafy for making this program available to me. Thank you Dr. Riad for always offering to fund me when funds ran low. Thank you for all the encouragement and thank you for making this opportunity happen for me.

I can not forget my dear friends Molly Fadl, Omneya Attallah, Riham Hassan, Nahla Zakzouk, Fayrouz Elsalmy, Nada Abugad, Elnaz Karimian, Anja Halama, Noha Yousri, and Sweetie Matthews. You were all the most supportive friends whether scientifically when I would get stuck on something or even emotionally.

Finally I would like to thank everyone that I may have not mentioned by name for helping me in even the slightest way to move a step forward in my career and life. Every ounce of support counted in some way for me.

Contents

List of Figures

List of Tables

Acronyms

AAL Automated Anatomical Labeling

AIC Akaike's Information Criterion

BIC Bayesian Information Criterion

BMS Bayesian Model Selection

BOLD Blood-Oxygen-Level Dependence

CBF Cerebral Blood Flow

CBV Cerebral Blood Volume

CWD Casewise Deletion

DCM Dynamic Causal Modeling

DTI Diffusion Tensor Imaging

FDR False Discovery Rate

FWE Family-wise Error Rate

EEG Electroencephalography

EM Expectation Maximization

EPI Echo-Planar Imaging

fMRI Functional Magnetic Resonance Imaging

FN False Negative

FOV Field of View

FP False Positive

GLM General Linear Model

HAROLD Hemispheric Asymmetry Reduction in Older Adults

HRF Hemodynamic Response Functions

ICA Independent Component Analysis

MEG Magnetoencephalography

MS Mean Substitution

PCA Principal Components Analysis

PET Positron Emission Tomography

PPI Psycho-Physiological Interactions

ReML Restricted Maximum Likelihood

RF Radio Frequency

RFT Random Field Theory

ROI Region of Interest

SEM Structural Equation Modeling

SNR Signal-to-Noise Ratio

SPM Statistical Parametric Mapping

TE Time Echo

TR Time Repetition

VB Variational Bayes

WGC Wiener-Granger Causality

Chapter 1

Introduction

1.1 Cognitive Phenotyping

The quantitative measurement of cognitive function (cognitive phenotyping) is a central task in computational psychiatry [1] and related proposals [2]. The functional systems perspective [3] is emerging as a promising approach to identifying such phenotypes, leveraging the theory and analysis methods of network science and dynamical systems. A leading method in this dynamical systems approach is Dynamic Causal Modeling (DCM) [4]. DCM models the interaction among brain regions as a dynamical system, giving rise to a generative model of brain activity that is augmented with an observation model that depends on the imaging modality. Inversion of the model provides an estimate of the model parameters and supports the comparison and averaging of models. DCM has had extensive development with several variations, improvements, and applications.

DCM makes a strict separation between where activity occurs and how that activity is coordinated. When fMRI is used as the imaging modality, DCM uses a first-level statistical analysis to locate active regions for the task or contrast under consideration. When performed subject-wise, this commonly reveals variation in the activation among individuals, occurring as missing or extra regions of activation relative to the proposed model topology [6]. The sources of this variation are complex including individual anatomical, functional, and measurement factors that are not related to the cognitive task. However, factors such as cognitive strategy which are important for cognitive phenotyping have also been identified [7].

While the number of nodes resulting from the first-level analysis is itself informative in focal neurological disease (e.g. [8]), other diseases are thought to be related to connectivity or other parameters of the generative DCM model. For instance, in schizophrenia, DCM analysis showed significantly decreased bilateral endogenous connectivity between two regions in comparison to healthy controls [9]. What is unclear in any given disease is whether the number of active nodes itself is a sufficient diagnostic indicator, or whether DCM model parameters add significant information. For example, suppose a DCM for a specific condition has been identified. To apply the model to an individual, their fMRI data is first subjected to a first-level test to identify active regions. Now, suppose that the subject does not have an active region close to that expected. Is this due to a false-negative in the first-level analysis or is this model simply not applicable in any way to this individual?

1.2 Functional connectivity in fMRI

There are two different terms used to describe brain connectivity in neuroimaging [10]. Functional connectivity refers to the descriptive concept of correlation defined as the temporal correlation between remote time-series. Functional connectivity essentially reduces to finding whether activity in two regions shares information. Effective connectivity, by contrast, refers to causation. It is defined as "the influence one neural system exerts over another either directly or indirectly [10]." It does not imply a direct physical connection but rather a causative influence. DCM in fMRI measures effective connectivity by taking regions of interest (ROIs) and using a generative model of measured brain responses that describe their nonlinear and dynamic characteristics. A neuronal model of interacting ROIs is made, then this model is supplemented with a forward model of how neuronal activity is transformed into a measured response enabling effective connectivity (parameters of the model) to be estimated from the observed data.

DCM takes as input a design matrix containing the different stimulus types or "explanatory variables" that represent the different experimental designs and the time-series from regions, and explicitly considers some nonlinear aspects to the experiment: specifically, the connections between ROIs and how they might change with experimental manipulation. A process including Bayesian estimations then produces a set of parameters. That parameter set includes: hemodynamic response functions (HRFs) for each region, resting connection strengths between each region, beta weights describing how the experiment affected each region, and connection beta weights, indicating how the experimental manipulation affects the connection strengths. It also produces an estimate of the statistical significance of each of these.

The general goal of computational neuroscience is to study brain function in terms of its physiology and dynamics by processing information measured from the various brain structures. This can be done by modeling features of the brain's underlying biological system at different scales ranging from neuronal membrane function or protein, to chemical coupling, to brain networks and architecture. Computational models can then be used to frame and test certain hypotheses by estimation and inference. Model comparison is also common in the studying of brain function and information processing of networks.

Numerous modalities are used to study brain function in neuroscience including fMRI (functional magnetic resonance imaging), MEG (Magnetoencephalography), PET (Positron Emission Tomography), and EEG (Electroencephalography). The data used in this thesis includes different fMRI datasets. FMRI measures the hemodynamic response, or change in blood flow, with respect to the neural activity in the brain. The changes in blood flow are correlated with changes in the BOLD (blood-oxygen-level dependence) signal [11]. FMRI has a high spatial resolution ranging from about 1-3 mm. One disadvantage of fMRI is its low temporal resolution compared to MEG or EEG [12]. The BOLD response peaks approximately 5 seconds after neuronal firing in a region making it hard to distinguish between responses to different events which occur within a short time window. This problem can be reduced with careful experimental design, such as ensuring sufficient time gaps between the presentations of different stimuli. Some researchers [13] are even attempting to combine fMRI signals which have relatively high spatial resolution with other techniques such as MEG or EEG which have higher temporal resolution to magnify the benefits of both techniques [14].

1.3 Problem Description and Significance

A first-level single-subject, fMRI activation analysis usually results in variations of activation among individuals. This is discussed in detail in Section 2.1.3. Differences among subjects could appear in the form of missing or extra regions of activation. Existing methodologies have resorted to discarding subjects with missing regions and any extra regions of activation in some subjects, usually in order to proceed with a higher level analysis such as a fixed or random effects analysis.

In a fixed effects analysis it is assumed that every subject uses the same model, whereas in a random effects analysis different subjects are allowed to use different models [15]. Individual level analyses in DCM are usually followed by a second-level or group-level analysis, as the subtle cognitive effects are often only manifested at the group level. Also, whole subjects and regions can be overlooked when undesired differences appear, posing a limitation to performing single-subject analysis. Therefore, there is a need to focus more on individual level analyses of fMRI data.

Given a group of m DCMs M_1, M_2, ... M_m, and a group of n subjects S_1, S_2, ... S_n, an evidence matrix can be computed where every entry in the matrix represents the evidence that a certain model M_x fits a certain subject S_y. The evidence can be referred to as the

Table 1.1: *Hypothetical Evidence Matrix. This matrix shows the evidence of each subject being fit to each available model given the model priors and loss functions. The highest evidence values in each column indicate the model having the best fit.*

	S_1	S_2	...	S_n
M_1	0.85	0.77	...	0.66
M_2	0.72	0.91	...	0.93
...	...	...	...	...
M_m	0.54	0.63	...	0.89

likelihood of a model and is defined by Bayes theorem [16] as:

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \tag{1.1}$$

where D is the observed data, and M is the hypothesized model.

By combining model priors and a loss function, model selection can be performed by analyzing the computed evidence [17]. Model(s) with the best evidence of fit can be selected for representation of a group or subgroup of subjects. By looking at the hypothetical figures in the evidence matrix (Table 1.1), it can be concluded that Model M_1 best describes subject S_1, and Model M_2 best describes subjects S_2 and so on.

Computing the evidence that a given model fits a given subject for fMRI data is possible as long as the subjects all have the same input data. The network topology of the DCM data includes the number of nodes and their spatial position which is determined according to detected levels of fMRI activation exceeding statistically set thresholds. The number of nodes in a DCM network represents the number of brain regions where voxel-wise statistically significant activation is detected, and the physical location of a brain region determines the spatial parameters of the node in the network. In practice, the signals are acquired from subjects during the presentation of stimuli in response to a specific task. Section 2.1.3 contains an overview of various examples showing evidence of network topology variations.

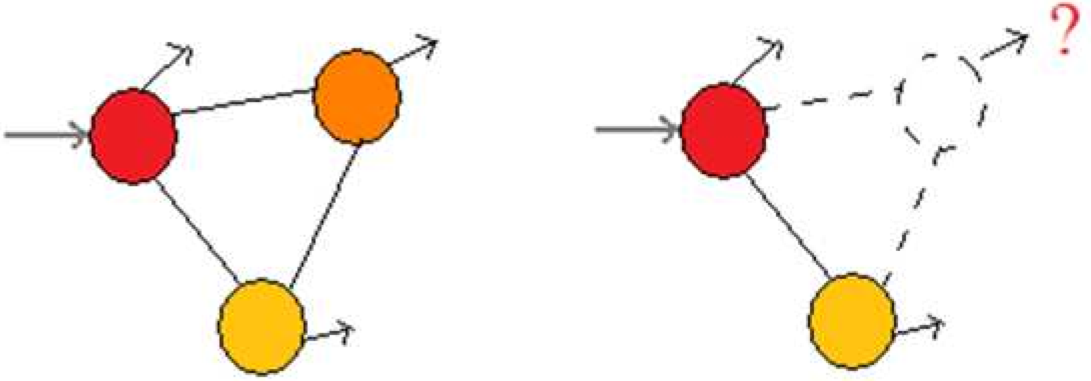

Figure 1.1: *Simplified DCMs for young (left) vs. old (right) person. There is a missing activated node in the old person.*

For example, in one study, given a group of old and young subjects, it was consistently shown that older adults exhibited altered brain activity compared to younger adults to achieve similar levels of cognitive performance [8] [18].

An issue arises, however, when trying to compute the evidence that a certain model fits a certain subject when the DCMs have different topologies. The problem is that when there is a missing node in the fMRI forward model (relatively no activation detected from that anatomical region), one cannot specify the spatial information of that node due to this missing information and the evidence cannot be computed. One question that arises is whether the assumption that the evidence a given model fits a given subject is equivalent to 0 is valid or not, since the model does not appear to fit the subject. If it is not a valid assumption, is there a structured method to compute the evidence in such a case where there is a mismatch in topology? Also, could there be a method to estimate the missing node in order to force the topologies to match and then be able to perform the computation?

The assumption that the evidence a given model fits a given subject is equivalent to 0 can in fact be quite problematic. Having evidence of a certain model being equal to 0 might not seem problematic in a standalone perspective per se. However, after computing the evidence for the different models, model comparison is usually performed between two or more models. Some methods for computing the model evidence are the BIC (Bayesian Information Criterion) and AIC (Akaike's Information Criterion) [19]. Model comparisons based on AIC are asymptotically equivalent to those based on Bayes factors. BIC is observed to be biased towards simple models and AIC to complex models [20].

The Bayes factor K [21] can assess this comparison and is simply a ratio of two likelihoods or evidences. If there are more than two models included in the comparison, then the relative log evidence can be used for various models over subjects.

$$K = \frac{P(D|M_1)}{P(D|M_2)} \tag{1.2}$$

A value of $K > 1$ means that the data indicates that M_1 is more strongly supported than M_2. Computing K with a 0 in the denominator can lead to issues with the misleading information that M_1 is infinitely more strongly supported than M_2. As a result of this comparison, it is questionable whether any model M_2 is considered computationally better than having none. Another possibility to compare models other than using the Bayes factor, is using significance levels or p-values [22]. However, unlike Bayes factors, p-values can only be used to compare nested models and do not allow one to quantify evidence in favor of a null hypothesis [20]. A possible advantage of using p-values is their independence from parameters of the prior distributions in the evidence computation.

The problem of missing features has been addressed in classification and there are methods of dealing with it. Classification involves categorizing a set of instances into different groups by assigning the same label to instances that share the same properties and different labels to instances that do not. The research problem being addressed is a clear example of classification where the different subjects are the instances we are trying to assign labels to, or models of best fit. When attempting to label an unknown pattern, one could face the problem of missing data due to any reason. Simple solutions to such a problem involve substituting a zero for the missing feature or taking the average of the available features. One method [16] involves marginalizing the full joint distribution over the missing features and then using Bayes decision procedure on the resulting distributions. Another method [16] uses the EM algorithm to maximize the log-likelihood of the available data, with the missing data marginalized so that the log-likelihood for the full data (available plus missing) is greater than that for available data alone. The basic idea in the EM algorithm is to iteratively estimate the likelihood given the available data.

Solving this problem is important because using current methods such as voxel-wise significance testing is prone to error due to the multiple comparison problem in considering sets of statistical inferences simultaneously. It is likely that a mistake is made while performing the DCM depending on choices such as noise threshold or p-value. The SPM analysis can thus yield different activation maps based on the threshold chosen. Sometimes, little activation in a given region can not pass the threshold, so the resulting DCM would show total absence of that particular region despite the presence of some activity in the area. For these reasons, a technique that quantifies the values for all the entries in the evidence matrix would be useful, rather than just assuming the values are plain zeros or that there is no evidence of a particular model fitting a particular subject.

DCM can be used to perform group studies between subjects performing the same task. Bayesian model selection can then be performed at the group level allowing inferences about regional effects to be made [23]. The model selection process can be based on the model evidence or the probability of obtaining the observed data y given the model m, $p(y|m)$. Differences in model parameters have been noted to occur among subjects. In a group study, separate evidence for each model and each subject can be computed and given that the data across subjects are independent, evidence values can be multiplied resembling a single evidence for each model. Using computational methods, such as the exceedance probability [15][17] or the group Bayes factor (ratio of evidences) [24] group categorization or analysis can be performed. The group Bayes factor is also known as the fixed effects analysis [25]. Using this approach does not allow formal inferences about the group to be made because it only takes into account the within-subject (between scans) variability. Contrary to that, the random effects analysis [17] takes into account the between-subject variability. Although there is limited information about the recruitment operation, or which brain regions end up being involved in the processing of a given task, human subjects can in fact perform cognitive tasks in many different ways. The computational barrier is in the difficulty in performing group analysis when the subjects do not have the same topology for a given task.

1.4 Thesis Statement

This thesis introduces the application of missing data approaches as a preprocessing step in dynamic causal modeling and shows that the inconsistency in network structure can be dealt with directly. Missing data approaches can compensate for some mismatches in topology specifically when there is a discrepancy in the number of active nodes in networks. This would allow the computation of model evidence on a per subject basis. The usefulness of this approach affects group studies in particular, as there is no longer a need to get rid of some subjects that were considered useless before. The manual filtering of these subjects was mandatory and subjects who did not follow a given DCM model were discarded [15] [23] [17] [24] [25]. A problem would arise when testing a particular DCM model for purposes of classifying the subjects could not be applied to a certain subgroup. Application of the missing data approaches in classification of DCM subjects enables the testing of all subjects rather than not performing the test at all and reducing the statistical power of the analysis. In addition, a solution to this problem also shows that model selection and ranking can now be applied to single-subject analyses as opposed to only filtered group analyses.

1.5 Summary and Organization

The missing data comes from first-level activation studies where some of the subjects might not show all the activated regions in particular networks and some subjects show extra regions of no interest. The traditional approach has been to discard subjects with missing regions and discard extra regions in some of the subjects. Rather than rejecting subjects and regions, this thesis presents a preprocessing approach to allow DCM to be conducted in single subjects by treating missing regions of activation, relative to a candidate model, as missing data.

By considering the application of missing data approaches as a preprocessing step in dynamic causal modeling, the significance of allowing the computation of the evidence that a certain model fits a certain subject when the DCMs have different topologies is first shown. Among its many uses are identifying individual differences/similarities among the group, analysis or prediction on groups to attempt explaining the underlying system, and finally the detection and treatment of disease.

This thesis shows novelty in the application of missing data approaches in the DCM framework. Although the applied missing data techniques are not entirely new, they have never been applied in the context of DCM before. Finally, this thesis shows feasibility and compares the different methods of missing data estimation. An important issue is dealt with when using multiple subjects in the context of single subject network classification. Practical usefulness is demonstrated in the context of classifying individual subjects' DCMs thus improving its use as a cognitive phenotyping tool in terms of classification ability, accuracy, and lastly model ranking.

The remainder of this thesis is organized as follows: the first part of Chapter 2 reviews the uses of functional connectivity for cognitive phenotyping and provides some background information on the methods of processing fMRI and DCM. Then a literature review of the DCM work related to this thesis is presented along with a review of the existing missing data approaches. Chapter 2 also presents the various methods for all the experiments in detail. Four simple methods for estimating the missing data are presented and compared using simulated networks in Section 2.5. In Chapter 3, experiments are used to compare the missing data approaches. The results are interpreted in light of other published results and related to the thesis. Application to representative phenotyping tasks is described and analyzed, showing the missing data approaches enable classification of all subjects using DCM, letting the model evidence rather than the number of first-level nodes alone be the determining factor for phenotyping. In Chapter 4, the methodologies used to estimate missing nodes are discussed with respect to the simulations and real experiments. Finally, the thesis is concluded with a comparison to using a relaxed statistical threshold in the first-level analysis, and future work is proposed.

Chapter 2

Background and Methods

2.1 Background

2.1.1 Functional Magnetic Resonance Imaging

MRI is used for non-invasively building three-dimensional images based on differences between nuclear spin relaxation times in different molecules [26]. The subject is first placed into a large magnetic field which causes nuclear spins to align. Radio Frequency (RF) signals are then used to excite nuclear spins away from the base alignment. In the process of nuclei precessing back to the alignment of the magnetic field, they emit detectable RF signals. The nuclear spins actually return to their original states at different rates. These rates are referred to as the longitudinal relaxation time or T1. In addition, the coherence of spins also decays differently based on the properties of the region. This process yields two main methods of contrast and forms the basis of T1 and T2 weighted images.

Moreover, dephasing occurs at two different rates, namely the T2 relaxation time and the T2* relaxation time. T2 is unrecoverable, but T2* is much faster and possible to recover from special RF signals. T1 relaxation times are usually on the order of seconds while T2* relaxation times are usually less than 100 ms.

Functional MRI (fMRI) can be used to acquire an entire brain image. For rapid acquisition time, a single large excitation pulse is applied to the whole brain and the entire volume can be acquired in a single T1 relaxation period [27]. Acquiring the entire k-space (spatial frequency) volume in a single excitation causes the Signal-to-Noise Ratio (SNR) to be very low in Echo-planar imaging (EPI). To increase the spatial resolution of EPI, more time or faster magnetic field switching is needed. However, increasing the switching rates can also yield more artifacts and lower SNR. Hence, the highest temporal resolution fMRI can yield is approximately one second.

2.1.2 Methods of processing fMRI

The images produced from fMRI can be interpreted in different ways [28] and there are numerous ways to model connectivity in fMRI. As mentioned earlier in Section 1.2, effective connectivity is defined as the influence one neural system exerts over another either at a synaptic or cortical level. The data input is essentially the same - consisting of time-series data, and the underlying computations look for patterns in the data that are similar between regions. Some methods attempt to do the same thing, but the different methods examine slightly different aspects of connectivity. Connectivity analyses can be categorized into model-driven and non-model-driven analyses. Non-model-driven analyses lack information about the details of the experiment and use blind algorithms because they search for patterns in the data without knowing the structure of the experiment. Principal components analysis (PCA) [29] and independent component analysis (ICA) [30] are examples of non-model-driven types of analyses.

On the other hand, examples of model-driven analysis include DCM and Psycho-physiological interactions (PPIs). DCM also includes SPM or Statistical Parametric Mapping [31] in which the brain is examined for differences in activity. Parametric statistical models are assumed at each voxel using the general linear model to describe the variability in data, and univariate statistics are used to assess the model parameters at each voxel. The goal of SPM fMRI data analysis is to determine how the brain processes information, which is done by detecting correlations between brain activation and the task the subject performs during the scan. Although correlations between all brain voxels can be computed to determine which brain regions are related, correlation does not imply causality since brain processes are very complex and often non-localized. Statistical methods like SPM thus can be used to assume relationships between brain regions.

There are also tools that measure connectivity based on autoregressive modeling [32] and Granger causality [33] - a way of ruling out some directions of causality. Structural equation modeling (SEM) [34] [35] is used when there is a set of ROIs under investigation, but little information is available concerning the links between them. It searches among the possible graphs that connect the ROIs and rules out some connections but includes others. SEM analyses start with a set of ROIs and estimate connection strengths between those ROIs that make up the best possible model of connections between them. The connection strengths are correlational (not directional), but represent the straightforward degree of correlation between the time-series of those regions. This method can be considered as a path analysis of time-series data. The logic behind the use of SEM originates from the suggestion that brain function is the result of changes in the covariances of activity among neural elements.

Standard fMRI experiments investigate the functional organization of the brain by contrasting the response to two or more sets of stimuli that are hypothesized to be treated differently by the brain. An activation map is generated by statistically comparing the response of each voxel to one stimulus versus another. The consistency of these activation maps across subjects is commonly evaluated by aligning the brain data across multiple subjects in a common anatomical space using spatial normalization. Using voxel-wise correspondence across subjects, statistical analyses can test whether each voxel consistently produces a higher hemodynamic response in one condition than another. The limitations of these standard methods are that they can only test hypotheses generated by the experimenter and that they assume consistency across the subjects in the spatial pattern of activation.

Moreover, the concept of a selectivity profile [36] is used to represent data as profiles of selectivity using linear regression estimates, and employ mixture model density estimation to identify functional systems with distinct types of selectivity. This method allows the discovery of patterns of functional response found across subjects without an a priori hypothesis or spatial consistency across subjects. Systems are characterized by their selectivity patterns and spatial maps, which are both estimated simultaneously using the Expectation Maximization (EM) algorithm.

Other methods [37] estimate brain networks from fMRI data by identifying a set of functional nodes or ROIs and performing a connectivity analysis between the nodes based on the fMRI time-series. Analysis can be as simple as correlation between two nodes' time series, or as complex as considering all the nodes simultaneously and estimating one global network model. Resting-state functional connectivity [38], which evaluates regional interactions that occur when a subject is not performing an explicit task, has also been quite prevalent in the analysis of fMRI data. This approach detects temporal correlations in the BOLD signal oscillations while the subject is at rest in the scanner. DTI (Diffusion Tensor Imaging) [39] can also be combined to test whether the resting-state functional connectivity hypotheses reflect structural connectivity as well. DTI extracts a measure of directionality of the white-matter tracts in a given voxel, indicating direction information of white matter in the brain. This can be used to infer which areas are strongly connected to each other and which are less so.

2.1.3 Evidence of Occurrence of Topological Differences

Apart from which method is used to process fMRI data, differences in topology are bound to occur. These differences could be due to age-related differences, gender-related differences, or abnormalities resulting from certain diseases. Even subject-specific processing can lead to such differences. This section provides a review of the literature presenting various sources of topological difference.

Differences in topology have consistently been shown to achieve similar levels of cognitive performance [8] [18], in addition to differences in dynamics when comparing younger to older adults. Having similar levels of cognitive performance means being able to successfully complete the same task with the same requirements. There is evidence that older adults recruit bilateral frontal brain areas when performing difficult cognitive tasks while younger subjects tend toward unilateral processing. To investigate this issue [8], fMRI scans were collected while performing three different tasks: working memory, visual attention, and episodic retrieval. The findings were consistent with the HAROLD (Hemispheric Asymmetry Reduction in Older Adults) model.

Age-related changes in anatomy and physiology are thought to be causes for brain reorganization of functions [40][41]. In [40], PET was used to measure activity in certain regions during four episodic retrieval tasks. The effects of the task conditions at each voxel were estimated using a general linear model and statistical contrasts were used to determine regions of activation by yielding t statistics (expressed as a Z score) for a given comparison at each voxel. [41] used PET scans to compare regional cerebral blood flow in young and old subjects while they were encoding, recognizing, and recalling word pairs. A multi-variate partial-least-squares analysis of the data was used to identify age-related neural changes.

Several others [42] [43] [44] showed that other tasks with different paradigms also yielded age-related topological changes, as opposed to a single task being responsible for this occurrence. [42] collected PET images to investigate verbal and spatial short-term storage in older and younger adults and used volume of interest analyses to specifically compare activation at sites identified with working memory to their homologous twin in the opposite hemisphere. A series of correlation analyses were conducted and different scores were derived. [43] and [44] also used the same methodology. This ruled out the doubt that age-related topological change was a specific phenomenon that would not reoccur for different kinds of tests (recall and recognition) and different kinds of stimuli (verbal and pictorial).

It was also shown that between a group of adolescents and adults performing a visual Go/No-

Go task, differing brain network architectures resulted [45][46]. They examined whole brain functional connectivity using independent component analysis. Older adolescents were able to respond faster making fewer mistakes when compared to younger ones. There are obvious maturity influences on cognitive control abilities that seem to increase with age up to a certain limit before they start declining again at the senior level. Different brain networks were found to be engaged with correctly withholding the planned response as opposed to committing an error. Networks engaged by errors usually involve some form of cognitive mechanism for error detection or error-based behavior correction.

There are other scenarios involving disease yielding variation in topology among subjects. Evidence shows that after unilateral brain damage, recovery of function can be facilitated by recruiting homologous regions in the unaffected hemisphere [8]. Moreover, according to [47], subjects that had undergone a stroke showed significant changes in activity in the unaffected hemisphere. Other cases of healthy hemisphere involvement were also found in the recovery of language abilities. [48] and [49] both showed healthy hemisphere involvement in aphasic patients in the process of language recovery. They performed statistical analysis on fMRI images by a combination of complex temporal cross-correlation and cluster-size thresholding which was based on true neural activity tending to stimulate signal changes over contiguous pixels. Levitan et al. [50] also showed network architecture changes under certain conditions for patients with brain lesions. Neural network models were used to simulate the changes in topographic maps that existed in the sensory and motor cortex due to brain damage. The computational models of map reorganization following cortical damage were used to study how a lesion in one cerebral hemisphere affects the organization of the corresponding maps in the opposite, intact hemisphere.

Finally, there is evidence that gender also influences the lateralization of brain activity during cognitive performance [51], shown during phonological processing between males and females. Using fMRI, large differences between women and men were reported in the lateralization of activation during a task in which subjects determined whether two printed words rhyme. When this task was compared with non-linguistic visual control tasks involving consonant letter string matching and line orientation matching, women showed symmetric activation of the frontal lobes while men showed left-lateralized activation. Voxel-wise tests were performed to show overall activation patterns, and regions of interest and hemispheres were examined to study the group differences.

2.1.4 Dynamic Causal Modeling

There are many tools used for the analysis of neuro-imaging data, such as FSL [52], AFNI [53], and BrainVoyager [54]. In this thesis the SPM tool [55] is used for neuro-imaging data analysis with a specific focus on fMRI data.

Dynamic Causal Modeling (DCM) [4] is used for the interpretation of functional neuro-imaging data. The aim of this modeling is to estimate the coupling among brain regions and how experimental changes affect that coupling. DCM involves constructing models of cognitive processes, representing the interaction between cortical regions or nodes, and modeling how neuronal activity is transformed into a measured response. DCM is based on an approach called TSI (Time Series Inference) but involves additional more complicated assumptions. TSI is based on temporal relations existing between time series of neural activity and depends on the statistical predictability of one time series by another. If the two time series represent neural activity from different neurons, then it is possible to infer about causal relations between them. WGC (Wiener-Granger Causality) [33] is the most common type of TSI used in neuroscience. WGC relies on the estimation of causal statistical influences between simultaneously recorded time series. Causality is based on the statistical predictability of one series that derives knowledge of one or more others.

DCM is a framework for modeling neuronal activity of brain imaging data using Bayesian inference. It uses Bayesian methods to fit dynamic models (represented as a system of differential equations) to functional imaging data, making inferences about model parameters and performing model comparison. Bayesian model comparison can use an approximation to the model evidence, which quantifies the properties of a good model.

DCM analyses in particular are considered highly model-driven. Given a set of ROIs and some hypothesis regarding connectivity such as “fully connected” (i.e. every ROI is connected to every other ROI), some connections can be eliminated immediately. DCM explicitly considers some nonlinear aspects of the experiment: specifically, the connections between the ROIs and how they might change with experimental manipulation. Friston et al. [56] consider all the other analyses special cases of DCM. Even the standard GLM analysis of activation is a special case where it is assumed that there are no connections between ROIs, and the ROIs are the voxels. DCM is relevant given a set of ROIs, a hypothesis about how they might work, and interest in how some areas or conditions might influence the connections between others. Mechelli et al. [57] looked at whether differences in visual activation due to categories of stimuli were mediated from the bottom up or top down. Although it is complicated and its results may be difficult to interpret, DCM is currently considered the state of the art in fMRI connectivity research for model-driven analyses.

The DCM approach can be applied to various modalities such as fMRI, EEG, and MEG. A DCM is a dynamic multiple-input multiple-output system with one output per anatomical brain region. DCMs rely on two classes of states, namely “neuronal” and “hemodynamic” states which encode the neurovascular coupling required to model variations in fMRI signals generated by neural activity. In this work, the focus is on DCM for fMRI, which involves a bilinear model for neurodynamics (defined in Equation 2.2) and an extended Balloon model [58] for hemodynamics. The parameters are bilinear in the sense that an input-dependent change in connectivity can be constructed as a second-order interaction between the input and activity in a source region, when causing a response in a target region. The bilinear interaction is between states and inputs, allowing quantitative inference on input-state and state-state (within region) coupling parameters. Inputs are used to induce responses that represent changes in stimulation. The inputs correspond to the stimulus functions that represent experimental manipulation. The state variables include the neuronal activities and the physiological variables required to produce outputs. The outputs are the measured hemodynamic responses from the brain regions being studied.

DCM uses experimental design principles to obtain region specific interactions that are used in experiments to extract region-specific activations. The causal variables that include the conventional design matrix become the inputs and the parameters become measures of effective connectivity [10], which is the influence one neural system exerts over another either at a synaptic or cortical level. There are two other qualitatively different types of connectivity, namely structural and functional connectivity. Structural connectivity [59] determines which neural units interact directly with each other through the anatomical layout of axons and synaptic connections while functional connectivity [60] regards non-mechanistic whole brain descriptions of statistical dependencies between measured time series.

The extrinsic effects of inputs are usually restricted to a single input region. Each of the regions produces a measured output that corresponds to the observed BOLD signal. A diagram of the BOLD signal is shown in Figure 2.1. These time-series from the regions would normally be taken as the average or first eigenvariate of key regions, selected on the basis of a conventional analysis. Each region has five state variables. Four of these are of secondary importance and correspond to the state variables of the hemodynamic model first presented in [4]. These hemodynamic states comprise a vasodilatory signal, normalized flow, normalized venous volume, and normalized deoxyhemoglobin content. These variables are required to compute the observed BOLD response and are not influenced by the states of other regions. Critical to the estimation of effective connectivity or coupling parameters are the state variables of each region. These correspond to average neuronal or synaptic activity and are a function of the neuronal states of other brain regions.

The results of DCM are specific to the tasks performed and stimuli presented during an experiment. Designed inputs can produce responses in one of two ways. Inputs can bring forth changes in the state variables (neuronal activity) directly. For example, sensory input could be modeled as causing direct responses in primary visual or auditory areas. The second way in which inputs affect the system is through changing the effective connectivity or interactions. An example of this second sort of contextual input would be time. Time-dependent changes in connectivity correspond to plasticity.

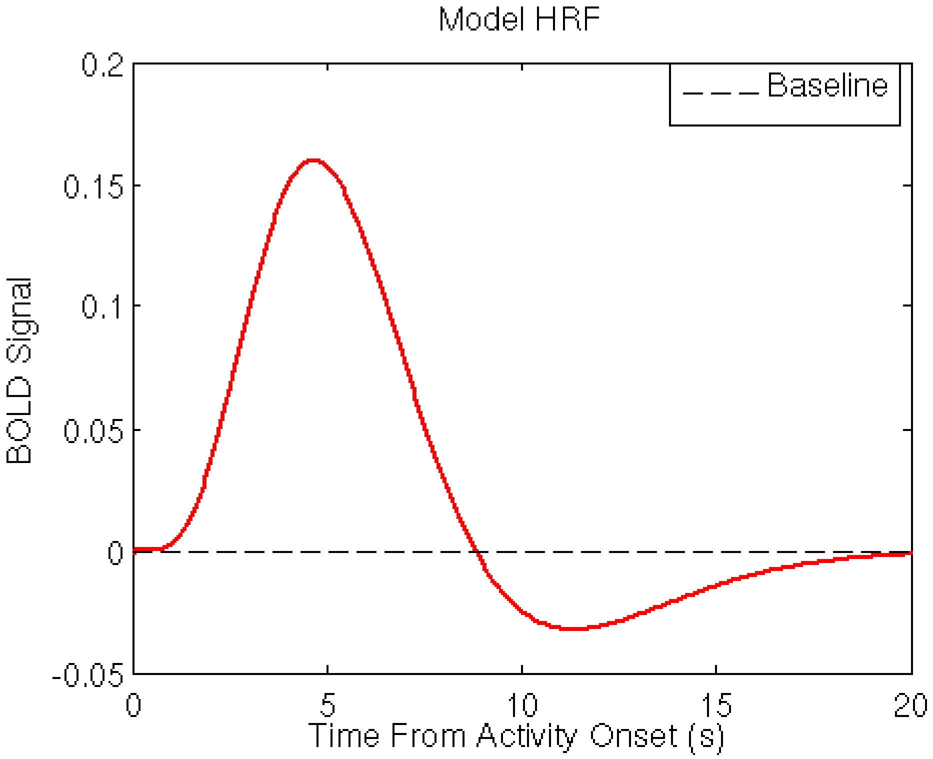

Figure 2.1: *The BOLD Signal. After a stimulus is applied, the BOLD signal rises till it reaches a peak. The input blood flow then starts to drop below normal for a typical post-stimulus undershoot. Following that is a full return to normal. The whole process takes approximately 20-30 seconds.*

The bilinear approximation of DCM reduces the parameters to three sets controlling three different aspects. The first set is the direct or extrinsic influence of inputs on brain states in any particular area; the second is the intrinsic or latent connections that couple responses in one area to the states of others; and the third is the change in this intrinsic coupling induced by inputs. Usually, DCM analysis focuses on the changes in connectivity embedded in bilinear parameters.

As mentioned earlier, DCMs are estimated using Bayesian estimators and inferences about connections are made using the posterior or conditional density. Having established the posterior density, the probability that the connection exceeds some specified threshold is computed. The posterior density is computed using the likelihood and prior densities. The DCM specifies the likelihood of the data, given some parameters. The prior densities of the connectivity parameters offer suitable constraints to ensure robust and efficient estimation reflecting some information about their likely ranges. These priors harness natural constraints about the dynamics of coupled systems but also allow the user to specify which connections are likely to be present and which are not. An example of prior constraints is restricting where inputs can elicit extrinsic responses. Not all inputs have unconstrained access to all brain regions and activations are not necessarily caused directly by experimental factors, but can be mediated by afferents from other brain areas. Some priors are principled priors that ensure that certain parameters can not have negative values, and these are considered conservative priors. On the other hand, some can be shrinkage priors, which reflect the fact that coupling parameters are zero, while some others can be empirical or based on previous independent measures.

2.1.5 The State Equation of DCM for fMRI

Given neuronal states $z = [z_1, ..., z_l]^T$ where l is the number of regions, the nonlinear model for effective connectivity is:

$$\dot{z} = F(z, u, \theta) \tag{2.1}$$

F is a nonlinear function describing the neurophysiological influences that activity in all brain regions z and inputs u, exert upon changes in the others. θ are the parameters of the model whose posterior density is required for inference.

The bilinear form of Equation 2.1 that describes the neurodynamics by the following multivariate differential equation is given below where u_j is the j^{th} experimental input and the dot notation denotes a time derivative.

$$\dot{z} = (A + \sum u_j B^j)z + Cu \tag{2.2}$$

$$A = \frac{\partial F}{\partial z} = \frac{\partial \dot{z}}{\partial z} \tag{2.3}$$

$$B^j = \frac{\partial^2 F}{\partial z \partial u_j} = \frac{\partial}{\partial u_j}\frac{\partial \dot{z}}{\partial z} \tag{2.4}$$

$$C = \frac{\partial F}{\partial u} = \frac{\partial \dot{z}}{\partial u} \tag{2.5}$$

A is the set of connections that characterize the DCM and specify which regions are connected and whether these connections are uni-directional or bi-directional. B is the set of modulatory connections that specify which intrinsic connections can be changed by which inputs. C is the set of input connections that specify which inputs are connected to which regions.

In DCM, neuronal activity leads to fMRI activity by a dynamic process in each region (extended Balloon model). Changes in neuronal activity cause changes in the blood oxygenation which is measured with fMRI. Neuronal activity in a certain region increases the vasodilatory signal. Inflow responds in proportion to this signal causing changes in the blood volume and deoxyhemoglobin content and outflow is related to volume [61]. The Blood Oxygenation Level Dependent (BOLD) signal is then taken to be a static nonlinear function of volume and deoxyhemoglobin that includes a volume-weighted sum of extra- and intra-vascular signals. The set of hemodynamic state variables, state equations, and hemodynamic parameters are referred to as h and are specific to each brain region.

$$h = \{\kappa, \gamma, \tau, \alpha, \rho\} \tag{2.6}$$

It is possible to predict the BOLD signal directly from a stimulus time course

$$\dot{s} = \epsilon u(t) - \frac{s}{\tau_s} - \frac{f-1}{\tau_f} \tag{2.7}$$

$$\dot{f} = s \tag{2.8}$$

$$\dot{v} = \frac{1}{\tau_0}(f - v^{\alpha}) \tag{2.9}$$

$$\dot{q} = \frac{1}{\tau_0}(\frac{f(1-(1-E_0)^f)}{E_0} - \frac{q}{v^{1-\frac{1}{\alpha}}}) \tag{2.10}$$

where s is a flow inducing signal, f is the input Cerebral Blood Flow (CBF), v is the normalized Cerebral Blood Volume (CBV), and q is the normalized local deoxygenated hemoglobin content. The parameters controlling blood flow are ϵ - a neuronal efficiency term, $u(t)$ - the stimulus, τ_f, and τ_s which are time constants. The parameters for the evolution of blood volume are E_0 which is the resting metabolic rate and α which is Grubb's parameter controlling the balloon model. τ_0 is a single time constant controlling the speed of v and q.

The most important property of the BOLD model is that the system is dissipative and eventually converges to a constant value. This is shown by analyzing the eigenvalues of the state equation Jacobian [62]. The steady state of the Balloon model equations gives the following where the parameters are all the same as in Equation 2.10:

$$s_{ss} = 0 \tag{2.11}$$

$$f_{ss} = \tau_f \epsilon u + 1 \tag{2.12}$$

$$v_{ss} = (\tau_f \epsilon u + 1)^{\alpha} \tag{2.13}$$

$$q_{ss} = \frac{(\tau_f \epsilon u + 1)^{\alpha}}{E_0}(1 - (1 - E_0)^{\frac{1}{\tau_f \epsilon u + 1}}) \tag{2.14}$$

$$y_{ss} = V_0((k_1 + k_2)(1 - q_{ss}) - (k_2 + k_3)(1 - v_{ss})) \tag{2.15}$$

The model parameters $\theta = \{A, B, C, h\}$ in DCM, are estimated using Bayesian methods. The B parameters are considered of most interest since they describe how connections between brain regions depend on alterations in the experiment. For a given DCM indexed by m, a prior distribution, $p(\theta|m)$ is specified using biophysical and dynamic constraints. The likelihood, $p(y|\theta, m)$ can be calculated by integrating the neurodynamic and hemodynamic processes. The posterior density $p(\theta|m, Y)$ is then estimated using a nonlinear variational method described in [4]. The Appendix contains a more detailed description of the above process.

2.1.6 Model Evidence

A method for approximating the evidence for a model m, fitted to a dataset y, is the Variational Bayes (VB) approach. Bayesian estimation provides estimates of two quantities

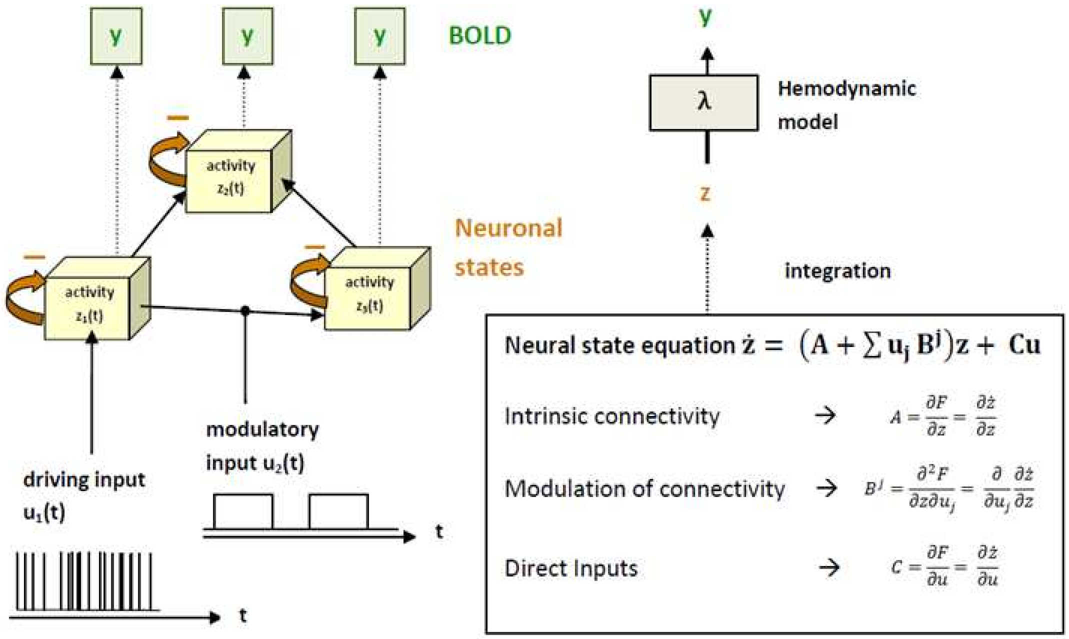
y
y
y
BOLD
activity $z_2(t)$
activity $z_1(t)$
activity $z_3(t)$
–
–
–
Neuronal states
modulatory input $u_2(t)$
driving input $u_1(t)$
t
t
y
λ
Hemodynamic model
z
integration
Neural state equation $\dot{z} = (A + \sum u_j B^j)z + Cu$
Intrinsic connectivity → $A = \frac{\partial F}{\partial z} = \frac{\partial \dot{z}}{\partial z}$
Modulation of connectivity → $B^j = \frac{\partial^2 F}{\partial z \partial u_j} = \frac{\partial}{\partial u_j}\frac{\partial \dot{z}}{\partial z}$
Direct Inputs → $C = \frac{\partial F}{\partial u} = \frac{\partial \dot{z}}{\partial u}$

Figure 2.2: *DCM diagram for fMRI. There are three interacting regions. The driving inputs are connected to certain regions, and each region has a state variable z that describes the neurodynamics of the system. A is the set of "intrinsic connections", B is the set of "modulatory connections", and C specifies which inputs are connected to which regions. Integrating the state equation produces the predicted BOLD signal y.*

the posterior distribution over model parameters $p(\theta|m, y)$ and the probability of the data given the model, or the evidence. The former can be used to make inferences about the model parameters θ. Computation of the model evidence is complex and involves integrating out the dependence on the model parameters.

$$p(y|m) = \int p(y|\theta, m)p(\theta|m)d\theta \tag{2.16}$$

[4] considers the VB technique ideal because the high dimensional integrals required for Bayesian parameter estimation can not be evaluated analytically, and it is computationally costly to evaluate them using numerical brute force or Monte-Carlo sampling schemes. VB is an iterative algorithm that optimizes an approximation to both the model evidence $p(y|m)$ and the posterior density $p(\theta|y, m)$. The key step involves decomposition of the log model evidence into

$$\ln p(y|m) = F(q) + D_K L(q(\theta); p(\theta|y, m)) \tag{2.17}$$

where $q(\theta)$ is any density over the model parameters, $D_K L$ is the Kullback-Leibler divergence and the free energy $F(q)$ is defined as:

$$F(q) = \langle \ln p(\theta|m) + \ln p(y|\theta, m) \rangle_q + S(q) \tag{2.18}$$

where $S(q)$ is the Shannon entropy of q.

Maximizing the functional $F(q)$ indirectly minimizes the Kullback-Leibler divergence between an approximate posterior density $q(\theta)$ and the true posterior $p(\theta|y, m)$. This quantity is always positive or zero when the densities are identical. Maximizing $F(m)$ minimizes the KL divergence making the free energy a lower bound on the log-model evidence. The VB approach is considered an extension of the Expectation-Maximization (EM) algorithm from maximum a posteriori estimation of the single most probable value of each parameter to fully Bayesian estimation which analytically approximates the posterior distribution of the parameters and latent variables. A more detailed derivation of the Variational Bayes (VB) approach is listed in the Appendix.

2.2 Review of Related Work

2.2.1 DCM Related Work

The first-level analysis is susceptible to false negatives (the underlying problem we are addressing here) due to multiple comparisons and the associated corrections. Control of false positives and negatives is a critical step in the interpretation of statistical parameter maps of activation. Typically, to reduce the chance of false positives, a conservative (corrected) p-value can be used. Alternatively, to reduce false negatives a loose or relaxed p-value can be used. Although not explicitly stated, this technique appears to be common practice when applying DCM when nodes are missing and dropping subjects is undesirable, in particular for exploratory analysis. When an expected node does not appear, the corrected p-value can be relaxed in stages until activation is produced at or close by the expected location. Although ad-hoc in nature, we can place this idea on a more sound theoretical base by treating the missing node as missing data, a common problem in statistics and machine learning, giving rise to numerous approaches [63].

According to the literature on DCM, the problem of differing DCM topologies has not been addressed directly. In a study by [64], models were compared at the group level and not the individual level. Group studies are preferred over individual studies because it is possible to identify individual differences or similarities among the group. Potentially, classification of subjects can be based on their model evidence. VB was used to infer the posterior density of the models and the exceedance probability (the probability that a particular model was more likely than any other model) was derived given the group data. To accommodate for inter-subject variability in the exact location of activation maxima in brain regions under study, subject-specific anatomical and functional constraints were set. A regional time series was extracted if it passed a prior set threshold and if it was located within a certain distance from the group maximum. Thus, the individual maxima were required to be within a fixed distance of the group maximum or they would be excluded from the analysis. They were able to extract time series in the same regions for only three-quarters of the subjects for a set threshold in the respective contrast of the GLM analysis. Because they constrained the extraction of time series to be located within a certain distance from the group maximum, they could not obtain time series from several subjects in certain regions. Constraining the distance to be bounded within a certain distance from the group maximum does not take into account the greater scale inter-subject variability. It only accommodates for a relatively small amount of inter-subject variability, which as shown in the experiments in Section 2.4, is not always sufficient.

The previous work was extended to compare families of DCMs [15]. Models were compared across a group of subjects using the Bayes factor under the implicit assumption that every subject used the same model. A number of DCMs were compared after being partitioned into families. This work focused on how to deal with families that did not contain the same number of models as opposed to different model topologies. Inferences about parameters, independent from assumptions about model topology, were done using Bayesian model averaging within families[65]. The families were composed of different groups of DCMs. All DCMs had three fully connected regions (same number of nodes). However, differences within the groups included which regions received direct input (thus affecting entries in the C matrix) and which connections were modulated by factors (affecting entries in the B matrix). The groups were arranged into families according to different patterns of input connectivity. The focus in [15] was on overcoming the numerical issues faced when there is a large number of models in the comparison. They resorted to using the Gibbs sampling method as opposed to Variational Bayes to overcome that problem.

Combining BMS results from several subjects relies on simple (fixed effects) metrics or the group Bayes factor, both of which do not account for group heterogeneity or outliers [66]. Stephan et al. [17] used the random effects method for BMS between subjects and at the group level. Their methods provided inference on model-space using the log model evidence as a subject-specific summary statistic. This enabled the use of analysis of variance to test for differences in log-evidences over models relative to the subject differences. The differences in model parameters were emphasized, as well as how to formally compare these models using the model evidence.

The review of DCM by Stephan et al. [17] briefly mentions the missing region problem and questions whether the unmodeled activity in the brain is actually sufficiently weak or unspecific to invalidate the DCM analysis. They suggest that solutions to this problem may encompass diagnostic procedures examining the model residuals for structure indicative of the existence of unmodeled processes. Whether this unmodeled activity is too weak or unspecific brings up other issues regarding the validation of the DCM analysis even before the handling of topological differences. But should there be a solution indicating explanations as to why certain processes could be unmodeled, this could be a step in the process of dealing with the problem of having a missing/extra region across subjects in the DCM analysis. Understanding the reason behind topological differences in DCM could be used towards solving the problem of missing nodes.

A reasonable method to deal with the missing node problem would be to use a higher p-value when performing the statistical tests. This method appears to be standard practice when it comes to fine-tuning the optimum p-value to use. The selection of an appropriate p-value can either be based on the standards used in literature or by experimental trial and error. Increasing the p-value results in more voxels passing the test and appearing as active when in fact they could be false positives. This can be shown in Figure 2.3 where the p-value is varied from 0.001 to 0.01 to 0.1. Using a higher p-value yields more widespread activations allowing users to locate peak centers of activation for every node they would desire. This can greatly reduce the number of omitted subjects, but it is questionable whether increasing the error to achieve this goal is acceptable or not.

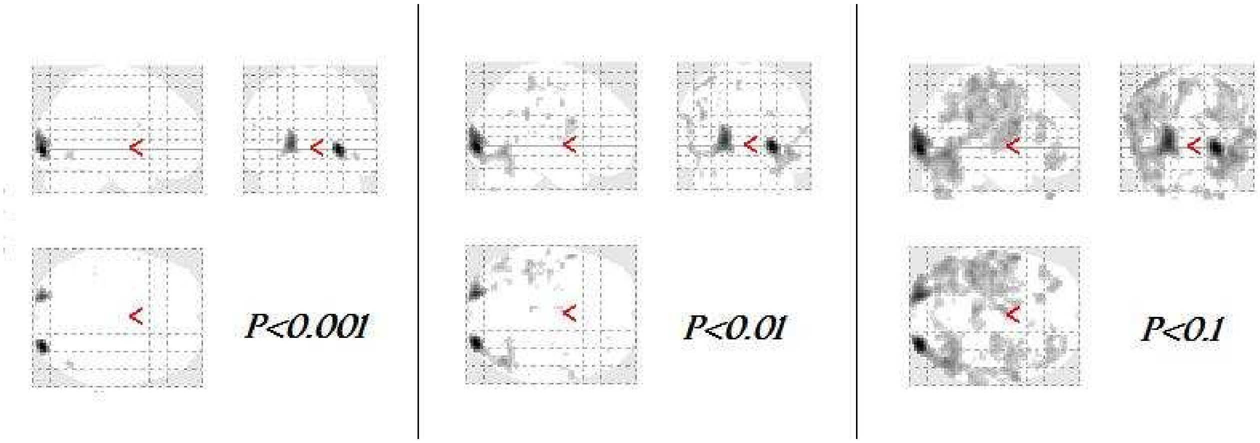

Figure 2.3: *For a particular subject, using the Go vs. Baseline contrast, it is shown that varying the p-value can result in significantly varying levels of activation. Certain regions of activation can disappear all together when reducing the p-value drastically and vice versa.*

Another possibility for dealing with the missing node problem is to perform uncorrected tests, as opposed to doing correction, to locate activation for every node. Again this could solve the missing node problem and greatly reduce the number of subjects that would have to be omitted due to missing nodes. The effect of correction on the size of activation volume and error rates is discussed in detail in Sections 2.4 and 3.1. Table 3.1 particularly shows how correction results in a drastic reduction of the percentage of overall volume of activation. This would still require a significant amount of error to be incorporated in order to avoid missing nodes and does not seem like the best approach to take.

2.2.2 Review of Missing Data Approaches

In classification, there are numerous methods for dealing with the problem of missing features [63]. Simple methods involve substituting a zero for the missing feature or taking the average of the available features. The most trivial method is the casewise deletion method (CWD) [63]. It is based on simply ignoring the features that have some missing values. Problems with this method, particularly for the case of linear regression problems, include having an observed bias in the parameters and large standard errors. A better approach is the so-called mean substitution method (MS) [67]. In this approach the missing values are filled by calculating the mean value of the given input or feature component, and this mean is used to replace all missing values for that particular input. This method is simple, efficient and gives generally good results.

An alternative to the mean substitution method is using linear regression to predict what the missing data should be on the basis of other variables that are present [67]. One advantage of this approach over mean substitution is that it is somehow conditional on other present information, which would likely improve results. The problem of error variance still remains, however, since by substituting a value that is perfectly predictable from other variables, no new information was necessarily added. Instead, the sample size was increased and the standard error was reduced.

Another method [16] uses the EM algorithm to maximize the log-likelihood of the available data, with the missing data marginalized so that the log-likelihood for the full data (available plus missing) is greater than that for available data alone. Here, the available data is the principal eigenvector of the nodes which match the model and survived the first-level activation test, while the missing data is that from nodes in the model with no corresponding activation. The basic idea in the EM algorithm is to iteratively estimate the likelihood given the available data. Iteratively, the features with the current best estimate of the full distribution are updated by candidate features that improve the estimate. Given such a candidate, the likelihood of the data including the missing feature marginalized with respect to the current best distribution is computed. However, there may be particular values of the missing data that give a different solution and an even greater log-likelihood.

The EM algorithm is described as follows where the parameter vector θ^i is the current best estimate for the full distribution and θ is the candidate for an improved estimate. The variable i is the iteration counter and T is the convergence criterion. Given the available data X, the missing data Z, and the unknown parameters θ, along with the likelihood function

$L(\theta; X, Z) = p(X, Z|\theta)$, the maximum likelihood estimate of the unknown parameters is determined by the marginal likelihood of the available data $L(\theta; X) = p(X|\theta) = \sum_Z p(X, Z|\theta)$. The EM algorithm finds the maximum likelihood estimate of the marginal likelihood by iteratively performing the following steps:

```
Begin initialize $\theta^0$, $T$, $i = 0$
Do: $i \leftarrow i + 1$
E Step: compute $Q(\theta; \theta^i)$
M Step: $\theta^{i+1} \leftarrow \arg\max Q(\theta; \theta^i)$
Until $Q(\theta^{i+1}; \theta^i) - Q(\theta^i; \theta^{i-1}) <= T$
Return $\hat{\theta} \leftarrow \theta^{i+1}$
End
```

$Q(\theta; \theta^i)$ is the likelihood of the data including the missing data Z marginalized with respect to the current best distribution described by θ^i. $Q(\theta; \theta^i)$ is the expected value of the log likelihood function with respect to the conditional distribution of the missing data given the available data under the current estimate of the parameters.

$$Q(\theta; \theta^i) = E_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)] \tag{2.19}$$

Given our noise-filling model, the likelihood function $Q(\theta; \theta^i) = p(X, Z|\theta)$ and Q has a Gaussian form. The error being minimized is the sum of squares between θ^i and θ. Once the algorithm converges, θ^i (which contains the distribution properties i.e. mean and covariance) is used to generate the missing data Z so it can be used as input into the DCM.

One alternative method to maximum likelihood is Multiple Imputation [68]. In the EM algorithm, a value of the missing data is estimated and substituted based on the available variables. In multiple imputation, however, a random value is substituted instead. Imputed values are generated on the basis of the existing data, just as in the EM algorithm. Suppose we are estimating Y on the basis of X, for every situation that X has the same value, the same value of Y will be imputed. This leads to an underestimate of the standard error of regression coefficients, due to less variability in the imputed data than if those values had not been missing. One solution, used in the EM algorithm, is to alter the calculation by adding error. With multiple imputation, we take our predicted values of Y and then add an error component drawn randomly from the residual distribution of $Y - \hat{Y}$. This is known as random imputation. However, this still underestimates the standard errors, so the imputation problem is repeated several times to generate multiple sets of new data whose coefficients vary from set to set, capturing this variability and revising the estimate.

2.3 Motivation

This thesis solves the problem of missing data (in the form of missing nodes) in fMRI DCM group studies. It is first necessary to show that this is a genuine problem and that the occurrence of missing nodes is indeed unavoidable. The first step was to design experiments to demonstrate that variability in real subject fMRI data is inescapable. This is shown in Section 2.4.

In order to show practical feasibility and because of the ease of implementation, subjects are simulated to model the variability in available nodes in a network. Then various missing data approaches are investigated as a preprocessing step. To assess the efficiency of the selected algorithms, classification tests are designed and carried out. These classification tests support individual level analysis ability, which is a gain in that previously, greater focus was on group analyses [15]. This is shown in Section 2.5.

To support the practicality of the proposed methods for data estimation in the context of DCM, real datasets are used to test the solution. The rationale behind using two real datasets is to show that simulation results can be supported by real evidence. If analogous results are achieved, this could be considered additional support for the validity of the solution to the problem. Sections 2.6 and 2.7 cover the datasets used and the methods applied to these datasets.

To further back up the solution's usefulness, the effect of the approach on the model ranking is also studied in Section 2.8. Model comparison is an important benefit in both individual and group analyses. Model comparison is not even possible when the model and subject have different topologies, particularly with regards to the number of active nodes in the two networks. A study is designed and performed to show the effect of using missing data approaches on the model selection process and the eventual ranking of models to subjects. This would be a very important contribution to individual analyses by allowing the computation of the entire evidence matrix rather than having to assume no evidence or hindering the computation of evidence altogether.

The sensitivity of the ranking of models as the population size increases is first studied. This is followed by an analysis showing the changes that occurred as a result of having to drop certain subjects not conforming to the correct network topology. It is also determined whether the usage of EM contributes to the restoration of original model ranking or has any sort of positive effect on the model selection and ranking processes. Again, this is all shown using simulated subjects and models.

To determine the practicality of using real datasets, different analyses (both fixed effects and random effects) can be done. This would show feasibility in using the missing data approaches as a preprocessing method before allowing the computation of evidence and the model posterior probabilities for all subjects and all models. If model selection and ranking can both be performed for a complete group of subjects and models without having to discard any subjects not matching the set criteria beforehand, that would be considered a fundamental advantage to DCM studies. In addition, Bayesian model averaging would also be possible since inclusion of subjects with missing nodes would be allowable.

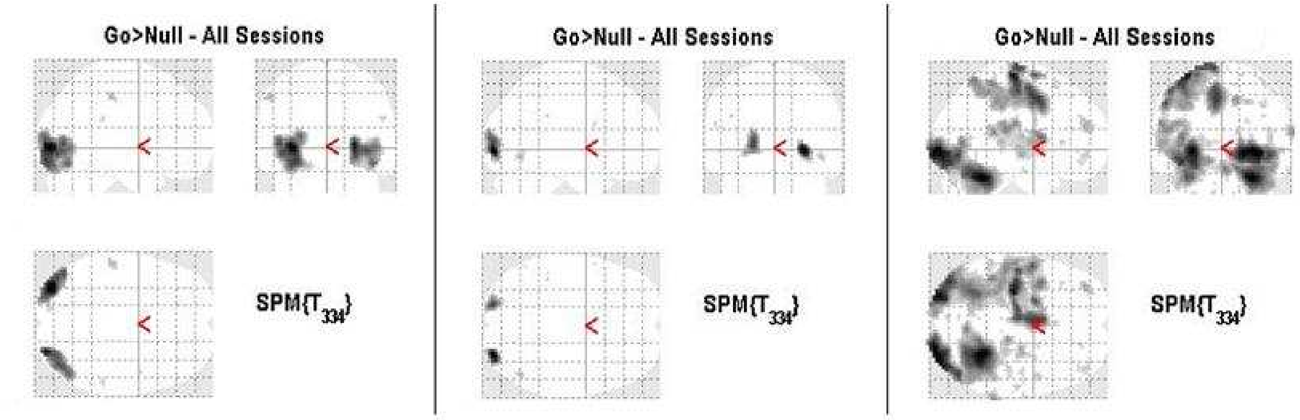

Figure 2.4: *Subject Variability in Activation Maps. These are the SPMs from three different subjects using the same contrast (GO-NULL) The first two subjects show some similarities in their SPMs while the third subject appears to have a whole lot more activation. The p-value that was used for these individual analyses was 0.001.*

2.4 Probability of False Positives and False Negatives

After producing the SPM's for the different subjects, different levels of activation were indicated in the various brain regions. The tables of statistics that were generated showed differences across the subjects in both the number and size of clusters where activation was indicated. The Automated Anatomical Labeling (AAL) [69] for SPM8 was used for the anatomical labeling of the activated areas during the functional brain mapping experiments. A p-value of 0.001 was chosen for the individual SPM analyses. Although some subjects showed similar activation maps, some still showed considerable differences as shown in the following figure which compares the glass brains (3D activation brain maps) of three different subjects. As shown in Figure 2.4, the first two subjects appear to have peak activations in roughly the same locations although some differences in the range/size of the activation can be seen. As for the third subject, there appears to be activation peaks in brain regions that did not occur in the first two subjects.

Figure 2.5 shows the combined variance in the SPM levels of activation at each voxel

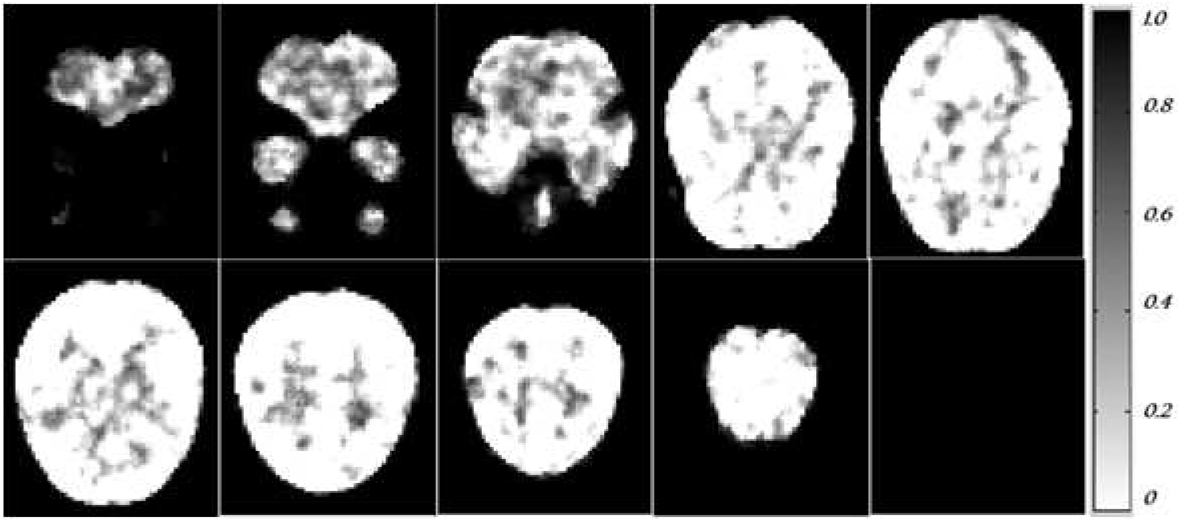

Figure 2.5: *Voxel-wise variance in SPM's across subjects. This illustrates the voxel-wise group variance of the z-scores across all the subjects. The darker levels of gray show the areas with highest group variances in comparison with the lighter shades of gray according to the gray scale-bar.*

in several different slices across all twenty-one subjects. As shown in the figure, there is considerable variability among subjects both in terms of magnitude of activation and the spatial parameters of those regions of activations. Some voxels show no variation among the subjects in terms of their z-scores (shown in white). Other voxels show different levels of variance in the z-scores across subjects that is proportional to the gray-scale level (i.e. darker voxels indicating greater variance).

Furthermore, a group SPM analysis was performed on the 21 subjects contrasting the Go versus No-Go conditions. A family-wise p-value of 0.05 was chosen for this analysis. Figure 2.6 shows both a glass view and a slice view of voxels that were significantly active in the population. Regions of activations from the group analysis were clustered according to connectivity forming certain clusters.

When doing the SPM there is a likelihood that nodes are missing (false negatives) or present when they should not be (false positives). This could be due to errors in the first level

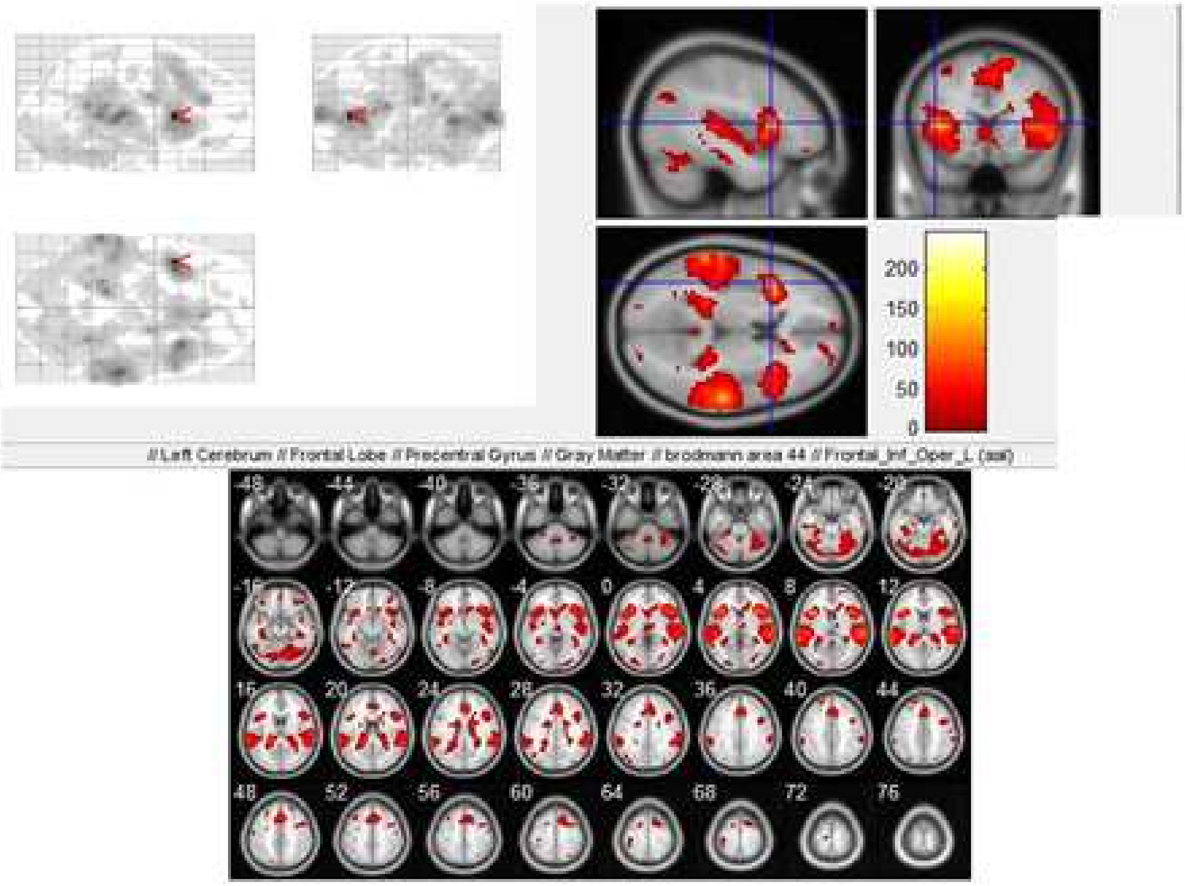
200
150
100
50
0
// Left Cerebrum // Frontal Lobe // Precentral Gyrus // Gray Matter // brodmann area 44 // Frontal_Inf_Oper_L (aal)
-48
-44
-40
-36
-32
-28
-24
-20
-16
-12
-8
-4
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
64
68
72
76

Figure 2.6: *Group SPM Analysis. This involved treating all the data from all subjects at once, rather than individually. A p-value of 0.05 was used and the voxels that survived the group-wise analysis are shown. The axial, sagital, and coronal views of a combined group brain are shown. Also various individual axial slices are taken from the overall brain volume and displayed.*

statistical analysis. The multiple comparisons problem occurs when one considers a set of statistical inferences simultaneously. This occurs due to testing all the voxels simultaneously. If only one individual pixel is being looked at, there can be confidence to a certain value of p that the time course is correlated to the stimulus. But when a whole image is formed from a statistical test, the confidence that all of the active pixels are not active due to random fluctuations is reduced. For example, if an image has 100 pixels and a p-value greater than 0.1 is used, then it is probable that 10 pixels will, by chance, be false positively labeled as active when they are not.

The probability of identifying at least one significant result due to chance increases as more hypotheses are tested. For example, if we are testing 20 hypotheses simultaneously using a p-value of 0.05 then the probability of one significant result occurring is 1 minus the probability of no significant results occurring. The probability of no significant result occurring is 1 minus the p-value which is 0.05. This figure is multiplied 20 times since there are 20 tests. This yields a 64% chance of false positives occurring. However if the p-value is reduced to 0.001, then the chance of false positive occurrence drops to about 2%.

A false positive is also called a type I error and is defined as the error of rejecting a true null hypothesis such as telling a patient he is sick when he is not. A false negative is also called a type II error and is defined as the error of failing to reject a false null hypothesis such as telling a patient he is not sick when he actually is, or in the SPM case, concluding that there is no activation when there actually is. Activation could be present but not detectable due to poor signal to noise ratio and could be interpreted as no activation at all.

Random Field Theory (RTF) is used to resolve the multiple-comparisons problem when making inferences over volumes. It provides a method for adjusting p-values for the volume and plays the same role for SPMs as the Bonferroni correction for discrete statistical tests. If all the pixels in an image can be considered independent, then a Bonferroni correction can be applied. However, with a large number of pixels in an image, the risk of false rejection of true activations is very high. There is also doubt whether each pixel in the image can truly be regarded as independent because of factors such as spatial correlation of neighboring pixels due to smoothing.

Two approaches were used for the correction of multiple comparisons control of family-wise error rate (FWE) and control of false discovery rate (FDR). FWE controls the number of false positive regions rather than voxels. The disadvantage is that many true positives could be missed. FDR has more false positives than FWE but fewer false negatives (more true positives). The results of effects of correction are shown in Section 3.1 and particularly in Table 3.1 which shows the percentage of voxels that survived the threshold averaged over the population for the uncorrected versus the corrected images.

2.5 Simulated Dataset Performing a Standard Go/No-Go Task

A network was designed to replicate a standard Go/No-Go task based on extensive literature (e.g. [46]). Networks were used to simulate realistic BOLD time series using the methods and code described in [37] [70]. The simulations were based on the DCM fMRI hemodynamic forward model [4], which uses the nonlinear balloon model for the vascular dynamics, on top of a causal model.

A network simulating a Go/No-Go task was generated. The task consisted of a visual and auditory stimulus. Absence of the auditory stimulus is accompanied with a motor function while the presence of the auditory stimulus, inhibits the motor function. The visual input u_1 and the auditory input u_2 together form u. Input u_1 consisted of a box function of zeros and ones resembling 2 different visual inputs. Input u_2 also consisted of a box function representing the absence/presence of an auditory stimulus. Two plausible (i.e. acceptably valid) networks were designed and used (Figures 2.7 and 2.8).

The simulated Go/No-Go task involved four brain nodes: a visual node (V), an auditory node (A), a motor node (M), and a prefrontal cortex node (P). Input u_1 was connected to V, u_2 was connected to A, and output was measured at M. The topology of the network was designed to give output similar to the expected output of the standard Go/No-Go task. The Go/No-Go task is a task where the presence of a visual stimulus and the absence of an auditory stimulus are accompanied with a motor function; however, when the auditory stimulus is present, inhibition of the motor function takes place. The output of logical 0 represents "No Go" response and logical 1 represents "Go" response. The Go/No-Go task truth-table is listed as in Table 2.1.

The inputs (visual and auditory) were Poisson processes with the rate controlled by the presence or absence of the stimulus. The auditory node had an external binary input and was generated based on a Poisson process that controls the likelihood of switching state. Neural noise/variability of standard deviation 1/20 of the difference in height between the two states is added to the signal. The mean durations of the states were 2.5 s (up) and 10 s (down), with the asymmetry representing longer average "no beep" than "beep" durations. This external input into the auditory node is viewed as a signal feeding directly into the auditory node.

The visual node also had an external binary input and that was generated based on a Poisson process that controls the likelihood of switching state. Neural noise/variability of standard deviation 1/20 of the difference in height between the two states is added to the signal. The mean durations of the states were 5 s (up) and 5 s (down), representing the input durations. This external input into the visual node is viewed as a signal feeding directly into the visual node.

In the bilinear DCM state equation $\dot{z} = (\alpha A + \sum u_j B_j)z + Cu$, $\dot{z}$ is the rate of change in z with respect to time (the rate of change of each brain region's output). A is the connectivity matrix describing which regions are connected. Coming into the network, there are j modulatory inputs which modify the connections between regions. For each modulatory input there is a B matrix which is similar to the A connectivity matrix but denotes where

Table 2.1: *Typical Go/No-Go Task Truth Table. The presence/absence of a visual stimulus are indicated by 1 and 0 respectively, and the presence/absence of an auditory stimulus are indicated by 1 and 0 respectively. An output of 0 stands for a No Go response, while a 1 stands for a Go response.*

Visual Input u_1	Auditory Input u_2	Output
0	0	0
1	0	1
0	1	do not care
1	1	0

the modulatory input connects to the network. Each of the B-matrices is multiplied by a vector of the modulatory inputs strengths, u. The vector of input strengths u, is a vector of the inputs. The C matrix determines which external input connects to which region. It has one row per region, and one column per input contained in the input u. The Go/No-Go task that was simulated involved four brain nodes. Four discrete parameter sets drove the time series. There was a visual node (V), an auditory node (A), a motor node (M), and a prefrontal cortex node (P). The visual input was connected to the visual node and the auditory input was connected to the auditory node. The motor node was where the output was measured.

The non-diagonal terms in A determine the network connections between nodes. To model the within-node temporal decay, the diagonal terms in A are all set to -1. Consequently, the term α is the rate of change of activity of mass-neuronal processing within a region (which inhibits itself) as described in [37]. The effect of the within-node dynamics (exponential temporal decay) is to create a lag between the input and output of every node. Although the original DCM forward model [61] includes a prior on α that results in a mean one second lag between neural time series from directly connected nodes, this unrealistically long lag was originally coded into DCM for practical algorithmic purposes in the Bayesian modeling.

Even though this is not a problem when DCM is applied to real data, it produces unrealistic lags in a simulation based on this model. Therefore it is changed to a more realistic time constant of a mean neural lag of approximately 50 ms [37].

Following that, every node's neural time series was fed through the nonlinear balloon model for vascular dynamics and this allows it to respond to changing neural demand. The amplitude of the neural time series was set so that the amount of nonlinearity (with respect to both changing neural amplitude and duration) matched what is seen in typical 3T fMRI data, and BOLD % signal change amplitudes of approximately 4% resulted (relative to mean intensity of simulated time courses). The balloon model parameters were set according to the prior means in DCM. There are differences in hemodynamic processes across brain areas and subjects which lead to different lags between the neural processes and the BOLD signal. Variations that occur could be up to one second or more ([71] [72]). This was taken into account by adding randomness to the balloon model parameters at each node, producing variations in the HRF (hemodynamic response function) delay of standard deviation 0.5 second. Lastly, thermal white noise with a standard deviation 0.1-1% (of mean signal level) was added.

The BOLD data was sampled with a TR of 3 seconds (simulation output sampling rate), and the simulations included 100 realizations/subjects all using the same simulation parameters, having randomized external input time series, randomized HRF parameters at each node and (slightly) randomized connection strengths. Each subject's data was a 10-min fMRI session (200 time points) and the time step of the integrator was 5 ms. Example simulated neural and fMRI BOLD time series can be seen in Figure 2.9 for a simple 4-node network with 2 nodes having external inputs.

The networks consisted of 4 nodes with 2 external inputs, and connection strengths were set randomly to have mean 0.4 and standard deviation 0.1 (with maximum range limited to 0.2:0.6). Each 4-node network can also be represented as a 4x4 connection matrix where each element (i,j) determines the presence of a connection from node i to node j.

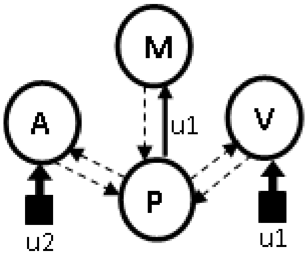

Figure 2.7: *Network 1 simulating Go/No-Go Response. Inputs are connected to nodes V and A and output is measure at M. The connection between P and M is modulated by u1.*

The matrices in 2.20, 2.21, and 2.22 were the chosen DCM matrices used for the design. From the top-left corner of the matrices, the rows and columns are ordered A, V, P, and M. The connection directions are from the row node to the column node. The C matrix first column is u_1 and second column is u_2.

$$A = \begin{bmatrix} -1 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0.5 & -0.5 & -1 & 1 \\ 0 & 0 & -0.5 & -1 \end{bmatrix} \tag{2.20}$$

$$B^1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.21}$$

$$C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{2.22}$$

Another set of DCM matrices that was used for another design is shown in (2.23), (2.24),

(2.25), and (2.26). This also produced the sought after Go/No-Go response.

$$A = \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ -0.5 & -0.5 & -1 & 1 \\ 0 & 0 & -0.5 & -1 \end{bmatrix} \tag{2.23}$$

$$B^1 = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.24}$$

$$B^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.25}$$

$$C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{2.26}$$

In the simulation, iterative methods were used for the approximation of solutions of the differential equations. The fourth order Runge-Kutta method was used to integrate the differential equations. After simulating DCMs all through to the hemodynamic model to get ROI time series for the subjects, the next step was feeding the simulated time series into SPM for estimation. The main aim in doing the simulations was to use the simulated time series to create the disconnected time series and then using EM-estimation to fill in the missing nodes. At that point, different models can be tested to see the evidence relative to the true simulated model.

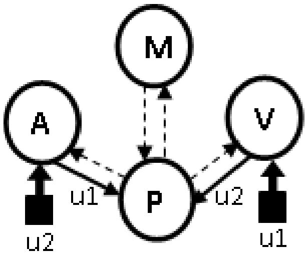

Figure 2.8: *Network 2 simulating Go/No-Go Response. Input is connected to nodes A and V while output is measured at M. This variation of the network has 2 connections that are modulated by inputs.*

Using the simulated ROIs of 10 subjects, 16 DCMs were specified in SPM having the configurations shown in Figure 2.10 (Total 160 subject-model pairs). Following that, the DCMs were estimated using the Variational Bayes algorithm in SPM. The default SPM options were generally used in all runs (SPM 8).

For each subject with a given model, one of the four nodes was randomly removed making note of the original model label. Then the missing node was replaced with alternative versions of EM-estimated, Mean-filling, Noise-filling, and Zero-filling time series. The noise that was used for noise-substitution consisted of minimal white noise with standard deviation 0.05 % (of mean signal level).

In the case where there was a missing node, output from the missing node was first replaced by zeros. For instance if the second node was missing, then the output $y(t) = [y_1(t), y_2(t), y_3(t), y_4(t)]$ was replaced by $y(t) = [y_1(t), 0, y_3(t), y_4(t)]$. In another run, the output from the missing node was replaced by the average value of that region from other subjects where that node is available, $y(t) = [y_1(t), y_2(t), y_3(t), y_4(t)]$ would be replaced by $y(t) = [y_1(t), avg, y_3(t), y_4(t)]$. In a third run the output from the missing node was replaced by the white noise and, using the same example, $y(t) = [y_1(t), y_2(t), y_3(t), y_4(t)]$ would be replaced by $y(t) = [y_1(t), noise, y_3(t), y_4(t)]$. In a fourth run the output from the missing node was replaced by the EM algorithm estimated value taken from the true

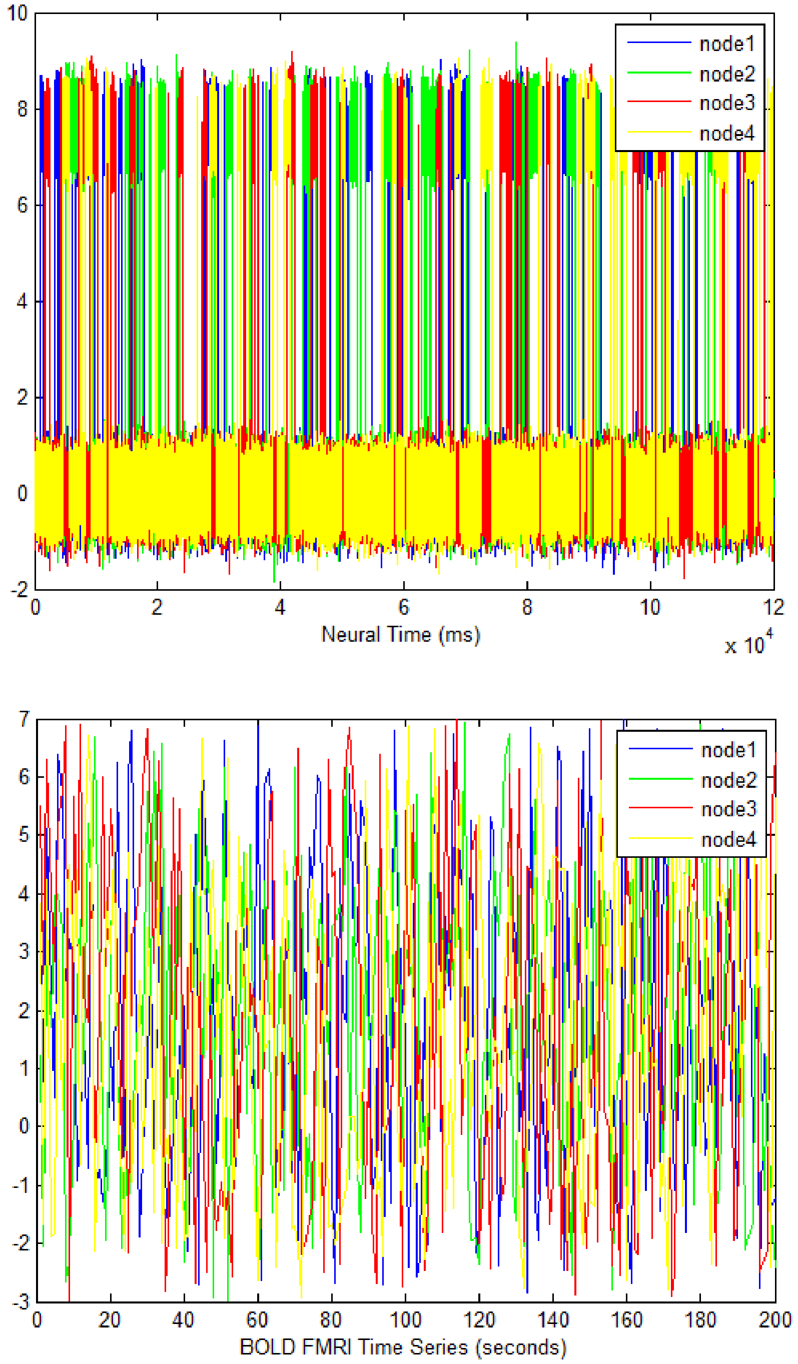

Figure 2.9: *Sample simulated neural and FMRI BOLD time series, for the 4-node network and inputs as described.*

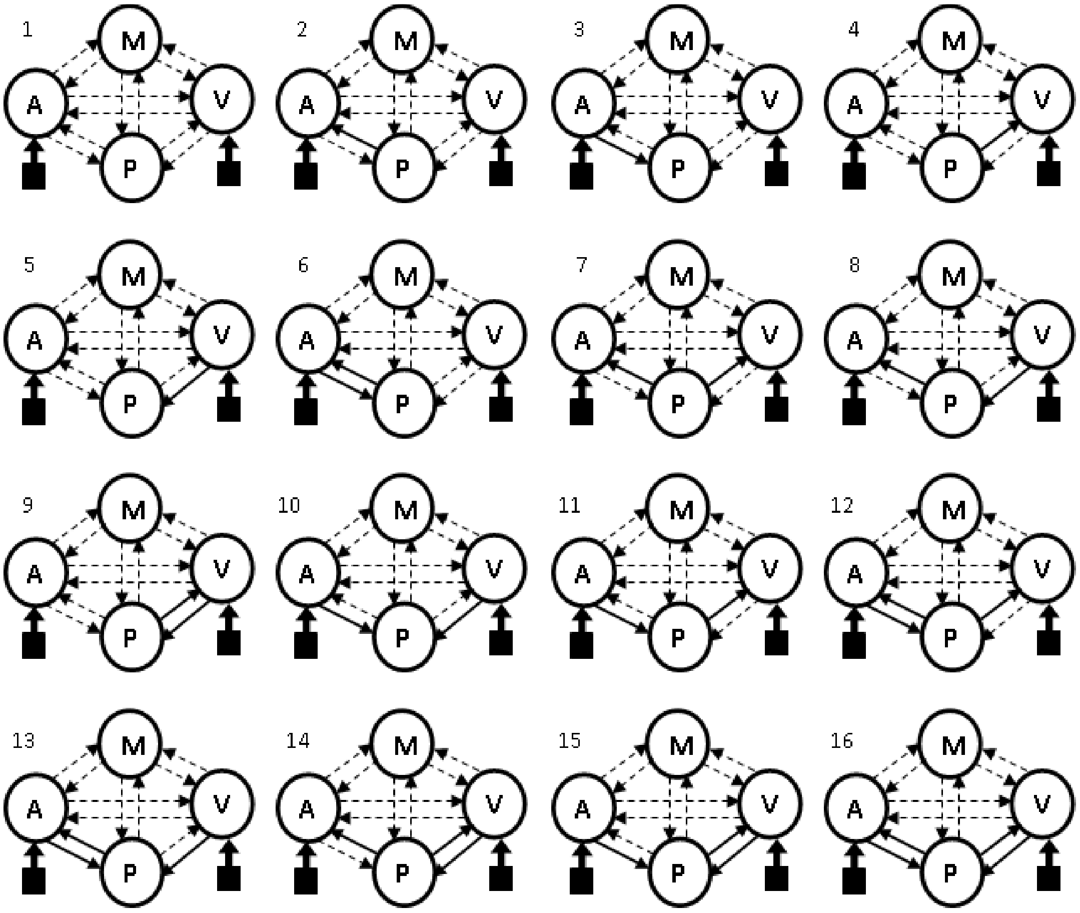

Figure 2.10: *Different models that were specified for Go/No-Go task. All DCMs are fully connected between all four regions (dotted lines). Inputs are indicated by the black boxes. The 16 models differ in their modulatory connectivity (solid lines).*

case and, using the same example, $y(t) = [y_1(t), y_2(t), y_3(t), y_4(t)]$ would be replaced by $y(t) = [y_1(t), EM_{est}, y_3(t), y_4(t)]$.

Following that, DCM model comparison was performed in SPM and the evidence was computed for the 16 models for each subject. Classification of a particular subject was then based on locating the highest model evidence. The results of the classification based on SPMs model evidence computations were compared with the known truths about the underlying models. The results are shown and discussed in Section 3.2. The same sort of analysis is also shown when randomly removing two out of the four nodes.

2.6 Real Dataset 1

2.6.1 Data Acquisition

The first fMRI dataset [73] was collected using a 3T Siemens Allegra MRI scanner. For each of two runs, 182 functional T2*-weighted echo-planar images (EPI) were acquired with the following parameters: slice thickness = 4 mm, 33 slices, time repetition TR = 2 s, time echo TE = 30 ms, flip angle = 90, matrix 64 x 64, field of view FOV = 200. Additionally, a T2-weighted matched bandwidth high-resolution anatomical scan (same slice prescription as EPI) and MPRAGE were acquired. The parameters for MPRAGE were as follows: TR = 2.3, TE = 2.1, FOV = 256, matrix = 192 x 192, saggital plane, slice thickness = 1mm, 160 slices. Stimulus presentation and timing of all stimuli and response events was done using Matlab.

This dataset [73] consisted of twenty-one healthy native English-speaking subjects (age ranged from 18 to 39). All subjects had normal or corrected-to-normal vision and were right handed. A manual stop-signal paradigm was used where a task consisted of a number of Go trials and Stop trials. On Go trials, the subject responded as fast as possible to the visual stimulus presented on a screen. For this manual task, subjects responded to the letters T or D with their right index or middle finger respectively. On Stop trials (which represented 25% of trials), the subject tried to stop his/her response when hearing a stop signal (a beep) which was played at a particular stop-signal delay (SSD) after presenting the visual stimulus. Go and Stop processes in these tasks ran independently implying that the distribution of Go processes on Stop trials (whether a response is made or not) was the same as the observed distribution of Go responses (when there is no Stop signal). When the SSD was short, the probability of inhibition [P (inhibit)] was high. Conversely, when SSD was long, P (inhibit) was low. As a result, SSD could be manipulated to achieve a certain probability of successful inhibition. Subjects participated in a behavioral test one day before the fMRI scan to estimate the SSD at which P(inhibit) = 50% under the task condition.

The basic paradigm for the pre-scan behavioral test was as follows. On Go trials, each trial started with a white fixation point (crosshair) appearing in the center of the black background screen. After 500 ms, a white letter (T or D) appeared. The letter T appeared on half the trials and the letter D appeared on the other half. The order of T and D was randomized. The letter remained on the screen until subjects made a response or after a one second delay, whichever occurred first. The next trial started after a one second interval. A Stop trial was identical to a Go trial in all respects except that a tone (900 Hz, duration 500 ms) was played at some delay after the stimulus. If the subject inhibited their response, then the stimulus remained on screen for the duration of one second. If the subject responded, then the stimulus disappeared. The next trial started after a one second delay. SSD changed dynamically throughout the experiment, depending on the subjects behavior. If the subject inhibited successfully on a Stop trial, then inhibition was made less likely on a subsequent Stop trial by increasing the SSD by 50 ms. If the subject did not successfully inhibit, then inhibition was made more likely by decreasing the SSD by 50 ms. Four step-up and step-down algorithms (staircases) were employed in this way to ensure convergence to P (inhibit) of 50% by the end of the experiment. The four staircases started with SSD values of 100, 150, 200, and 250 ms respectively. For each condition, there were 240 Go trials and 80 Stop trials. Each staircase therefore moved 20 times. The staircases were independent but

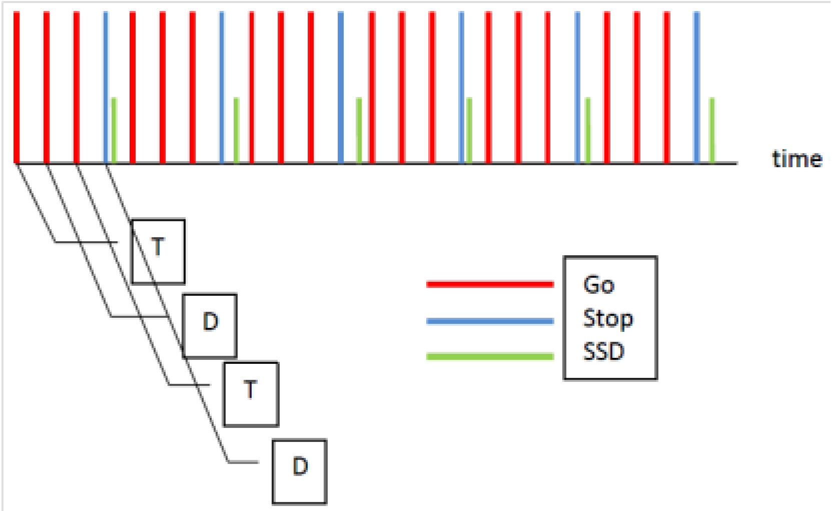

Figure 2.11: *Timing Diagram of Go/No-GoTask. For a given session, there were a series of go and stop trials. The Go trials constituted 75% while the Stop trials constituted 25%. A visual stimulus (letters T or D) was presented on a screen. A manual response to T or D consisted of a button press with either the right index or middle finger respectively. On the stop trials, the subject attempts to stop response when a stop signal (audio beep) is sounded at a stop-signal delay (SSD) subsequent to the visual stimulus.*

randomly interleaved, that is, each particular Stop trial belonged to one particular staircase, but the order of staircases was random trial by trial.

2.6.2 FMRI Preprocessing

All preprocessing was done using SPM8 [55]. The functional images were first realigned to remove any movement artifacts from the fMRI time series. This process realigns a time-series of images acquired from the same subject using a least squares approach and a six parameter (rigid body) spatial transformation consisting of translation and rotation. A representative image was used as a reference to which all subsequent scans were realigned. A mean image was also produced to be used in the following step which was co-registration. In the co-registration step, the sessions were first realigned to each other by aligning the first scan from each session to the first scan of the first session. Then the images within each session were aligned to the first image of the session. The files were also slice-timing corrected before co-registration. Differences in slice acquisition times were corrected which was necessary to make the data on each slice correspond to the same point in time. Without correction, the data on one slice could represent a point in time as far as 1/2 the TR from an adjacent slice. Co-registration was done between the structural and mean functional data, maximizing the mutual information. Following that, SPM was used to segment the structural image using the default tissue probability maps as priors. Gray matter, white matter, and bias-field corrected structural images were created. This was done to lower the impact of non-brain structural variability on the registration. Tissue classification requires the images to be registered with tissue probability maps. After registration, these maps represented the prior probability of different tissue classes being found at each location in an image. Using Bayes rule, these priors can be combined with the tissue type probabilities derived from voxel intensities, thus providing the posterior probability. Registration requires an initial tissue classification, and tissue classification requires an initial registration. So both these components are combined into a single generative model involving alternation among classification, bias correction, and registration steps. The final processes included spatial normalization of the functional data followed by smoothing of the data by an 8mm kernel.

2.6.3 FMRI Model Specification and Statistical Analysis

Categorical responses were modeled using the stimulus onset times and movement parameters from the realignment stage. Three conditions were specified, namely, “Go”, “StopFail”, and “StopSucc”. An example of the general linear model (GLM) design matrix specified for one of the subjects (including two sessions) is shown in Figure 2.12. The design matrix has one row for each scan and one column for each effect or explanatory variable (i.e. regressor or stimulus function).

Estimation of the GLM parameters was done using a Bayesian approach using a Variational

Bayes algorithm (VB) [74]. The alternative to VB is the classical approach (ReML Restricted Maximum Likelihood) [74]. The classical method assumes the error correlation structure is the same at each voxel. This correlation can be specified using either an AR(1) or an Independent and Identically Distributed (IID) error model. ReML estimation would be applied to spatially smoothed functional images. The Bayesian approach allows one to specify spatial priors for regression coefficients and regularized voxel-wise AR(P) models for fMRI noise processes. This algorithm does not require the functional images to be spatially smoothed. The VB model estimation approach takes much more time than the classical one. A model order of 0 was selected which corresponded to the assumption that the errors were IID.

After estimation, contrast vectors were applied to the results to produce statistical parametric maps (SPMs) or posterior probability maps (PPMs) and tables of statistics. Since null events were not explicitly modeled, they constituted an implicit baseline. The following t-contrasts were defined - Go-Null, StopSucc-Null, StopFail-Null, and StopSucc-Go. For the Bayesian approach, posterior inference was made using a PPM. Regions were identified where there was a high probability (level of confidence) that the response exceeded a particular size (i.e. signal change). This was quite different from classical inference, where we looked for low probabilities of the null hypothesis that the size of the response is zero.

2.7 Real Dataset 2

2.7.1 Data Acquisition

The second fMRI dataset [75] was collected using a 3T Siemens Allegra MRI scanner. For each of two runs, 151 functional T2*-weighted echo-planar images (EPI) were acquired with the following parameters: slice thickness = 4 mm, 40 slices, time repetition TR = 2 s, time echo TE = 30 ms, flip angle = 80, matrix 64 x 64, field of view FOV = 192 mm. Additionally,

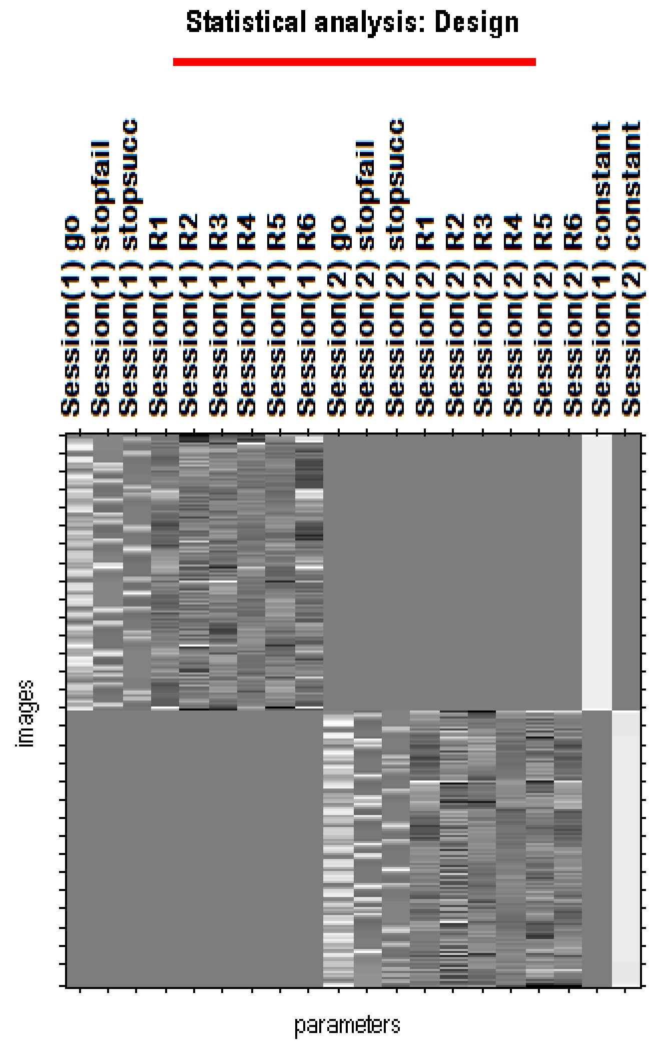

Figure 2.12: *Design matrix for a single Go/No-Go experiment. This diagram shows the design matrix for two sessions of the same experiment. Each row represents a scan. The first three columns represent the different event stimulus onsets and the rest of the columns represent the realignment parameters. The realignment parameters are used as regressors because there are residual effects correlated to the task being performed. Spin excitation history effects and non-linear distortion are due to magnetic field inhomogeneities. These effects need to be included in the temporal model so that they are explained away when making inferences about activations.*

a T1-weighted high-resolution anatomical scan was acquired. The parameters for MPRAGE were as follows: TR = 2.5 s, TE =3.93 ms, TI=900 ms, flip angle = 8, 176 slices, FOV=256 mm.

This dataset consisted of twenty-one healthy native English-speaking subjects performing a rapid event-related Simon task. In each trial (inter-trial interval (ITI) was 2.5 seconds, with null events for jitter), a red or green box appeared on the right or left side of the screen. Participants used their left index finger to respond to the presentation of a green box and their right index finger to respond to the presentation of a red box. In congruent trials the green box appeared on the left or the red box on the right, while in more demanding incongruent trials the green box appeared on the right and the red on the left. Subjects performed two blocks, each containing 48 congruent and 48 incongruent trials, presented in a pre-determined order, interspersed with 24 null trials (fixation only).

2.7.2 FMRI Preprocessing

The same preprocessing steps described in Section 2.6.2 that were applied to the first dataset were systematically applied to the second dataset as well.

2.7.3 FMRI Model Specification and Statistical Analysis

Categorical responses were modeled using the stimulus onset times and movement parameters from the realignment stage. Four conditions were specified, namely, Congruent-Correct, Congruent-Incorrect, Incongruent-Correct, and Incongruent-Incorrect. Each session had those four conditions with the appropriate stimulus onset times for each. The movement parameters were used as multiple regressors in both sessions. Six regressors/columns in the design matrix were used to model the linear rigid-body movement effects for each session. An example of the general linear model (GLM) design matrix specified for one of the subjects (including two sessions) is shown in Figure 2.13. The design matrix has one row for each scan and one column for each effect or explanatory variable (i.e. regressor or stimulus function). Estimation of the GLM parameters was done using the VB approach [74].

After estimation, contrast vectors were applied to the results to produce statistical parametric maps (SPM's) or posterior probability maps (PPM's) and tables of statistics. Since null events were not explicitly modeled, they constituted an implicit baseline. The following t-contrasts were defined: Congruent-Incongruent [1 1 -1 -1 0 0 0 0 0 0 1 1 -1 -1] and Correct-Incorrect [1 -1 1 -1 0 0 0 0 0 0 1 -1 1 -1]. For the Bayesian approach, posterior inference was made using a PPM. Regions were identified where there was a high probability (level of confidence) that the response exceeded a particular size (i.e. signal change). This was quite different from the classical inference, where we looked for low probabilities of the null hypothesis that the size of the response is zero.

2.8 Ranking of Subjects

The ability to perform model comparison is an essential goal of group studies. Model comparison can be used to rank different prospective models against subjects in a given group study. Given N subjects in a study of model comparison by the ranking of models, it is essential to determine how sensitive the ranking is to N. Also, the effect of dropping subjects on the ranking because of missing data needs to be studied. An analysis is done to show whether the usage of EM to fill in the data restores the original ranking. Another analysis shows whether the inclusion of all subjects affects the model choice or not.

To be able to rank every subject against all available models, it is first important to decide what the different available models are. The next step would be to compute the model evidence for every model and individual. One of the models needs to be chosen by penalizing incorrect model choice. Experiments were designed to rank different subjects from the two datasets by computing the evidence of each model. The evidence was computed using various techniques such as AIC/BIC and free energy approximation.

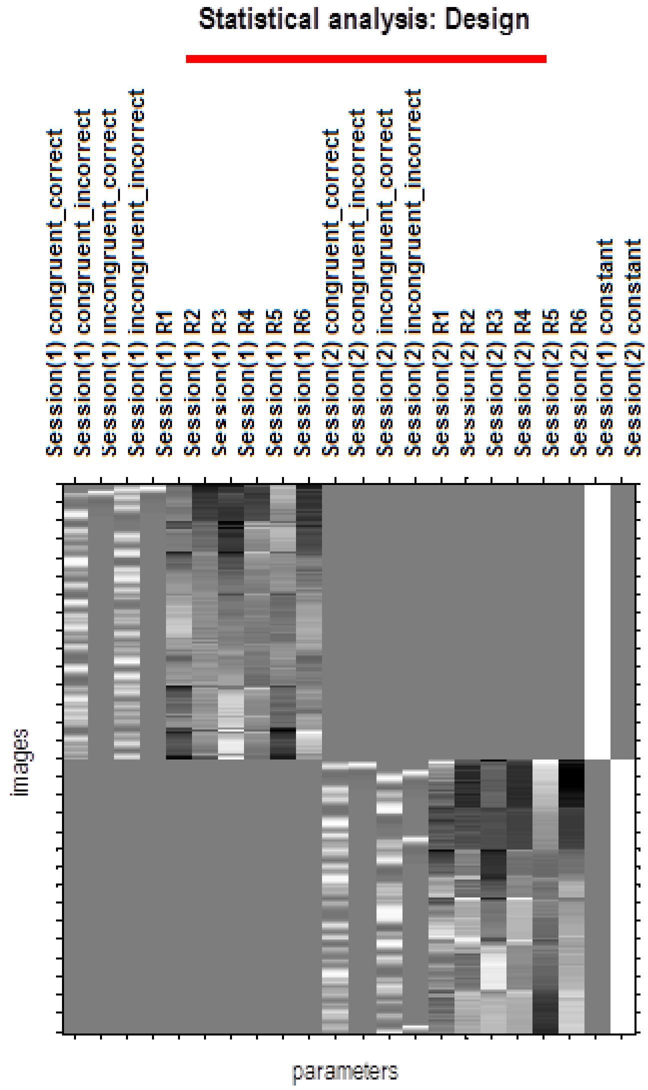

Figure 2.13: *Design matrix for a single Simon task experiment. This diagram shows the design matrix for two sessions of the same experiment. Each row represents a scan. The first four columns represent the different event stimulus onsets and the rest of the columns represent the realignment parameters.*

2.8.1 Computation of Model Evidence using AIC and BIC

The Akaike information criterion (AIC) [76] is a measure of the relative goodness of fit of a statistical model. It provides a means for model selection but does not provide a test of a model in the sense of testing a null hypothesis. The Bayesian information criterion (BIC) [76] is another criterion for model selection among a set of models. Both are partly based on the likelihood function and solves the problem of over-fitting by introducing a penalty term for the number of parameters in the model.

The different models were defined based on variations in [15], namely, which regions received direct input (entries in C matrix) and which connections were modulated by factors (entries in B matrix). There are four brain regions involved for each subject. Models having only symmetric full intrinsic connectivity were selected due to the large number of combinations possible. Given that each connection (considering both directions) can be either modulated or not, there are 2^{12} possible patterns of modulatory connections. Given that the auditory and visual stimuli are either direct inputs to a region or not, there are $2^8 - 1$ possible patterns of input connectivity (excluding the model without any input). The modulatory patterns are then crossed with the input patterns producing a total of $(2^{12})(2^8 - 1)$ different models. The selected models were fitted using regression to the data collected from N simulated subjects.

In order to measure how sensitive the ranking is to the number of subjects N, different values of N were tested ranging from $10, 15, 20, ..., 100$. All subjects had the same number of nodes. The measure of sensitivity to the ranking was computed as a percentage of the number of subjects whose highest evidence model changed when increasing N by 5 to the total N:

$$\frac{N_{affected}}{N} * 100\% \tag{2.27}$$

The effect on the rankings of dropping subjects because of missing data was also studied. Using the same 10, 15, 20, ... , 100 subjects, and randomly removing nodes from about a third of the subjects, the percentage change of the subjects' labels was also noted. The number of subjects whose label was affected due to dropping of subjects as a percentage of the population are shown in Chapter 3.

EM was then used to fill in the missing data to test if the original ranking was restored. Using EM to estimate the missing nodes allowed the inclusion of those subjects who did not have all nodes. The comparison was done with the original 2/3 of subjects that were included in the previous round.

2.8.2 Computation of Model Evidence using the Free Energy Approximation

Previous works [15] and [77] state that using AIC/BIC criterion can lead to overly simple models being selected. Therefore, the evidence of DCM models was computed using the free energy approximation as well, and the different models were again fitted to data collected from the N simulated subjects using VB. To measure the sensitivity of the ranking to the number of subjects N, the same values of N ranging from 10, 15, 20, ... , 100 were used.

2.8.3 Fixed Effects and Random Effects Analysis

Neuro-imaging datasets usually include data from several subjects. The group model inference is where models $m = m_1,\ m_2,\ \dots\ m_n$ are fit to data from subjects $s = s_1,\ s_2,\ \dots\ s_n$. Every model is fitted to every subject's data. In Fixed Effects (FFX) Analysis it is assumed that every subject uses the same model, where as Random Effects (RFX) Analysis allows for the possibility that different subjects use different models.

Given that the complete dataset, Y, which includes data for each subject, Y_s, is independent over subjects, the overall model evidence can be written as

$$p(Y|m) = \prod{}_{s=1}^{S} p(y_s|m) \tag{2.28}$$

$$logp(Y|m) = \sum{}_{s=1}^{S} logp(y_s|m) \tag{2.29}$$

Bayesian inference at the model level can be implemented using Bayes rule

$$p(m|Y) = \frac{p(Y|m)p(m)}{\sum_{m=1}^{M} p(Y|m)p(m)} \tag{2.30}$$

Using uniform model priors $p(m)$, the comparison of a pair of models, m_i and m_j, can be performed using the Bayes Factor. When comparing two models across a group of subjects, the individual Bayes factors can be multiplied to obtain the Group Bayes Factor (GBF) [15]. If the physiological mechanism is unlikely to vary across subjects then it can be assumed that all subjects have an identical model structure.

RFX is an alternative procedure for group level model inference that allows the possibility of having different models across subjects. This could be used in disease cases or in dealing with cognitive tasks that can be performed with different strategies.

Chapter 3

Results

Although several extensions exist, a DCM analysis requires the number of nodes in the models to be consistent and have the same topological connectedness, even if the specific connections vary, in order to create model families that can be compared and used for experimental design [5]. The number of nodes and their locations are specified experimentally by a first level analysis (activation detection). Subjects without active regions mapping to all model nodes must be dropped from the analysis. Also, subject data with extraneous activation is ignored, using only data from the subset of regions consistent with the model.

Inter-subject variability in fMRI activation patterns has been studied extensively and attributed to numerous sources [7]. In group level analysis of activation, it is assumed these sources of variability result in identically distributed additive noise. However, because DCM is a second-level analysis, this variability leads to missing data since there is no way to specify the nodes location, and hence to extract the principal eigenvector.

For example, Den Ouden et al. [64] compared models at the group level. To accommodate for inter-subject variability in the location of activation maxima in brain regions, subject-specific anatomical and functional constraints were set. A regional time series was extracted only if it passed a prior set threshold and was located within a certain distance from the group maximum. Otherwise, it was excluded from the analysis.

While the dropping of subjects does result in decreased power in group-level studies, its impact is negligible if the number of excluded subjects is small. A more difficult problem arises when individual-level analysis is desired, for example to use the DCM for single subject classification and prediction.

This thesis presents an approach to enable the inclusion of all subjects in a DCM analysis displaying at least some overlap with the hypothesized model by treating the absence of activation in a given node as missing data. The research problem was restated as a missing feature problem. A mismatch in topology was translated into an unavailable data case. The experiments performed here are used in selecting the most appropriate method to estimate missing nodes in order to compensate for certain mismatches in network topology. Once the missing data problem has been solved, the standard computational methods for group analyses can be applied as inter-subject variability will only be governed by parameter variations.

To validate the effectiveness of the solution, a variety of classification tests were performed. These classification tests labeled a given unknown subject with the correct DCM label. Experiments were set up having a group of subjects with unknown model labels where the aim was to label each subject with a known DCM model. The patterns from classification experiments mirrored the computed values in the evidence matrix supporting the solution.

Data was collected from both the simulation of theoretical models and subjects using Matlab [78] and from several real subject fMRI datasets. The missing data approaches were applied to the simulated subjects to evaluate effectiveness. The data that was collected from real experiments produced variability on the network topology level. Concepts taken from the simulation were then reapplied to the experimental data and analyzed to determine the efficiency of the solution.

Table 3.1: *Average Percentage of Surviving Voxels across Subjects from Real Dataset 1. There is a significant drop in the surviving voxels when correction is applied versus the uncorrected subjects.*

Average % Voxels ($p < 0.05$)			
Contrast	Uncorrected %	FWE %	FDR %
Go-No Go	8.36	0.82	1.24
Go-Null	6.52	0.39	1.02
StopSuc-Null	6.21	0.37	0.97
StopFail-Null	7.48	0.75	1.13
StopSuc-Go	6.14	0.31	0.94

3.1 Results of FP/FN Simulations

As mentioned in Section 2.4 two approaches were used for the correction of multiple comparisons - control of FWE rate and control of FDR. Table 3.1 shows the percentage of voxels that survived the threshold averaged over the population for the uncorrected versus the corrected images. From the values in this table, it is shown that correction leads to a considerable drop in the surviving voxels with respect to the whole volume. This indicates that correction leads to significantly varying results when compared to not performing any correction.

Table 3.2 indirectly shows the effect of correction on the false positives and false negatives. There is no actual ground truth to accurately compute the change in false positives and false negatives against. Thus measuring the percentage of appearing or disappearing voxels after correction relative to before correction is used as an indication of the existence of these errors. The conclusion is that no matter whether correction was performed or not, there is still the possibility of occurrence of both false positives and false negatives.

The occurrence of false positives and false negatives inevitably leads to the problem of different topologies occurring across subjects performing the same experimental task. A possible solution to handling the difference in topologies is by taking the union of all the activation maps from all the individual subjects to form a new comprehensive and common subject space. The union can then be projected into the individual subject spaces. By taking the union and projecting each subject back into the common space, all subjects would have the same topology potentially solving the problem of varying topologies across subjects in the same study.

The projected voxels can then be masked out thus restricting the number of voxels that are to be tested and ultimately affecting the values of false positives and false negatives. A 3 mm voxel neighborhood around the projected voxels was selected. The images were originally smoothed with a full width at half maximum (FWHM) Gaussian 8 mm kernel in the x, y, z directions and the voxel size was 3 mm. A neighborhood size of 3 voxels was selected and the projected voxels were then masked out with this 3 voxel neighborhood around them. The SPM was then done on the masked out volume and the number of appearing and disappearing voxels after the projection compared to before projection were measured.

The total number of voxels in the projected mask was computed and was used to normalize the number of appearing and disappearing voxels by the total number of voxels in the projected mask to obtain percentages. The brain voxels in the common space covered approximately 17% of the total space and there were approximately 15 formed clusters, for example, in the common space for Contrast 1. Table 3.3 shows the volume analyses for all subjects in terms of appearing and disappearing voxel percentages. Also shown in Table 3.3 are the cluster analyses for the same subjects. Based on the data combined from both analyses, some subjects show loss of volume which can be translated into either the shrinkage of some cluster(s) or the disappearance of that/those cluster(s) altogether. The data also indicates that some subjects show the gain of volume which can also be translated as expansion of some already existing cluster(s) or the appearance of newly formed cluster(s). Therefore, the data suggests that using the group union and projecting into the individual subject space not only possibly changes the shape of activation (by shrinkage/expansion) but can possibly result in the addition/deletion of new areas or nodes. This can be considered

Table 3.2: *Effect of Correction on Activation Volume. This evidence indicates that correction leads to various changes in the activation characteristics (in the form of appearing and disappearing voxels).*

Average % of activated voxels with FWE and FDR ($p < 0.05$)				
	FWE		FDR	
Contrast	Appear%	Disappear%	Appear%	Disappear%
Go-No Go	2.25	4.75	0.87	5.65
Go-Null	2.98	4.81	1.51	5.50
StopSuc-Null	3.54	5.78	2.33	6.25
StopFail-Null	1.26	4.95	0.75	5.40
StopSuc-Go	1.45	4.24	1.26	4.75

as evidence that there is a substantial number of false positives and negatives.

Using the union method could be a potential solution that is much simpler than using the EM method to fill in the missing data. The problem, however, is that the common space contains a lot of voxels and one can not design a DCM that is based on the number of nodes given by the number of clusters in the common space because it is relatively large (15 for instance). Having 15 clusters in the common space, there are potentially 15 nodes for each individual after the projection. In the literature, a Go/No-Go task presumably involves somewhere between 3-4 nodes. The assumed areas where there would be activation would be the visual node, the auditory node, the motor node, and the pre-frontal cortex node. However, if all 15 nodes were to be included, it would be difficult to determine how they would be connected.

A possible approach to reducing the number of clusters is to take those 15 clusters after projecting them into the individual space and generate the principal eigenvector. Then for each of those 15, correlation can be done and the cross-correlation matrix of the principal eigenvector across those 15 nodes can be generated. This would give some indication as

Table 3.3: *Volume and Cluster Analysis across Subjects. The changes in activation volume are seen in both the individually activate voxels and whole activate clusters.*

	Average % change of active voxels		Average Number of Clusters	
Contrast	Appear%	Disappear%	Appear%	Disappear%
Go-No Go	5.18	2.49	2.2	1.4
Go-Null	4.5	2.44	1.8	1.0
StopSuc-Null	5.63	3.17	3.1	1.5
StopFail-Null	4.03	1.95	2.0	1.0
StopSuc-Go	5.6	3.69	2.5	0.9

to what the parameters should be for that individual (regarding the A matrix of the DCM which specifies which regions are connected). The actual correlation values contained a lot of close-to-zero values suggesting that the DCM A matrix would be rather sparse and could be reduced. However, the values did include small figures (even if they are considered negligible) suggesting that there is a functional relationship (or at least a linear one) between most nodes. Having a 15x15 matrix would be problematic since that would translate into a large number of parameters.

To increase the sparseness of the A matrix, relatively small values (less than 0.1) were considered zeros while the larger values were compared across the matrix to form groups of most correlated clusters. Still, based on the relative values of the individual cross-correlation matrices, the 21 subjects did not show consistency in the cluster grouping of nodes that were formed. Seven of the subjects formed two groups of linearly related nodes, eight of the subjects formed three groups, and six of the subjects formed four groups. Moreover, the formed groups of nodes were not consistent in the specific included nodes. Figures 3.1, 3.2, and 3.3 illustrate the variation for 6 different subjects for a given contrast. The groups are color coded indicating which nodes are linearly related. The conclusion is that even if we consider linear correlations to try to merge nodes in order to decrease their number, using

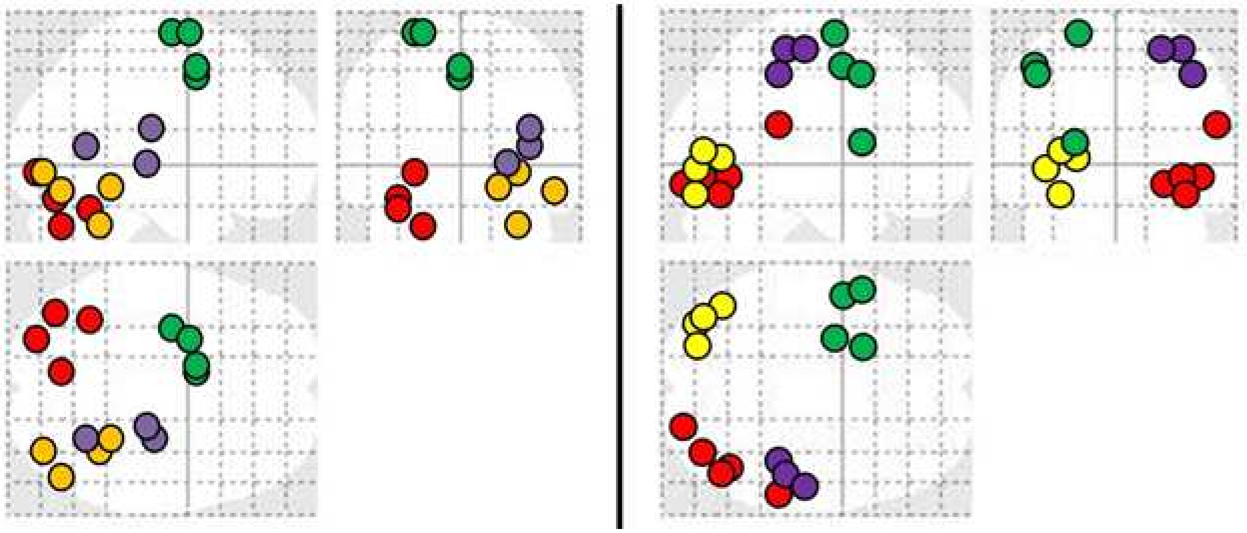

Figure 3.1: *Two sample subjects with 4 groups of correlated clusters. Each group of nodes belonging to the same group is identified by a different color.*

the union of all the activations within a group still leads to different regions per subject. Thus there is still a difference in the network topologies.

3.2 Results of Simulated Go/No-Go Task

Figure 3.4 shows the estimation steps of the DCMs using the Variational Bayes algorithm in SPM. Figure 3.5 shows the classification accuracy based on highest model evidence versus ground truth using the 160 simulated subject-model pairs described in Section 2.5. This figure shows the percentage at which the true underlying model could be recovered for the different techniques. Taking a subset of the 16 DCMs (8 instead of 16 including the true model), the same kind of analysis was performed and the results are shown in the same Figure. Again, taking an even smaller subset of the original 16 DCMs (4 instead of 16 including the true model), the same analysis was performed and those results are shown.

As shown in Figure 3.5, using the EM-algorithm for filling in missing data yields the highest classification accuracies relative to the other methods. This holds for various dataset sizes

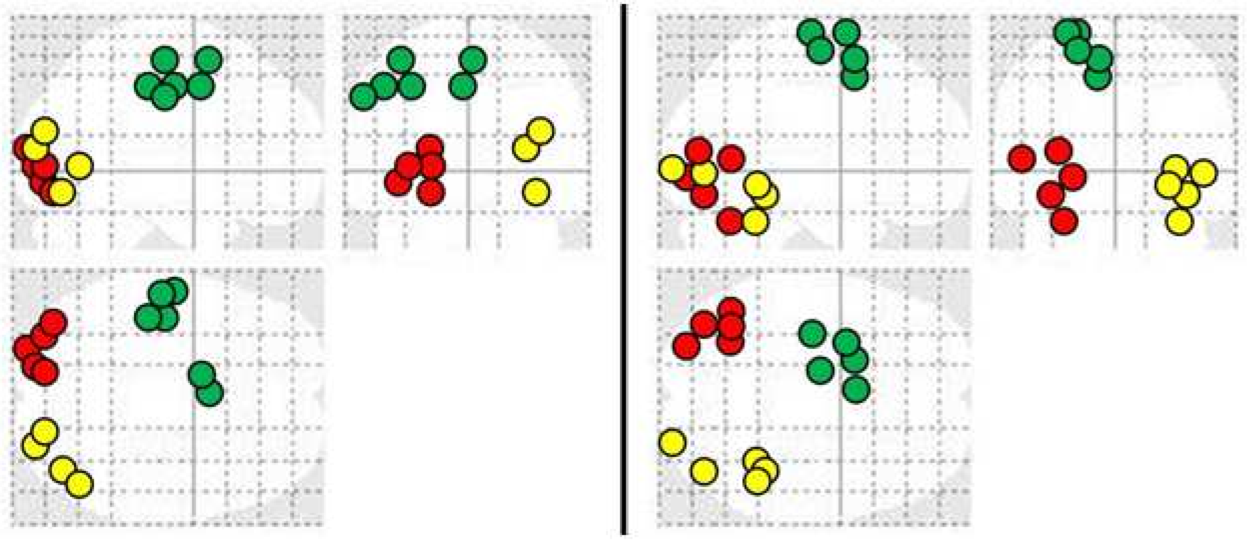

Figure 3.2: *Two sample subjects with 3 groups of correlated clusters. Here 3 colors are used to distinguish between the 3 groups each considered a separate cluster.*

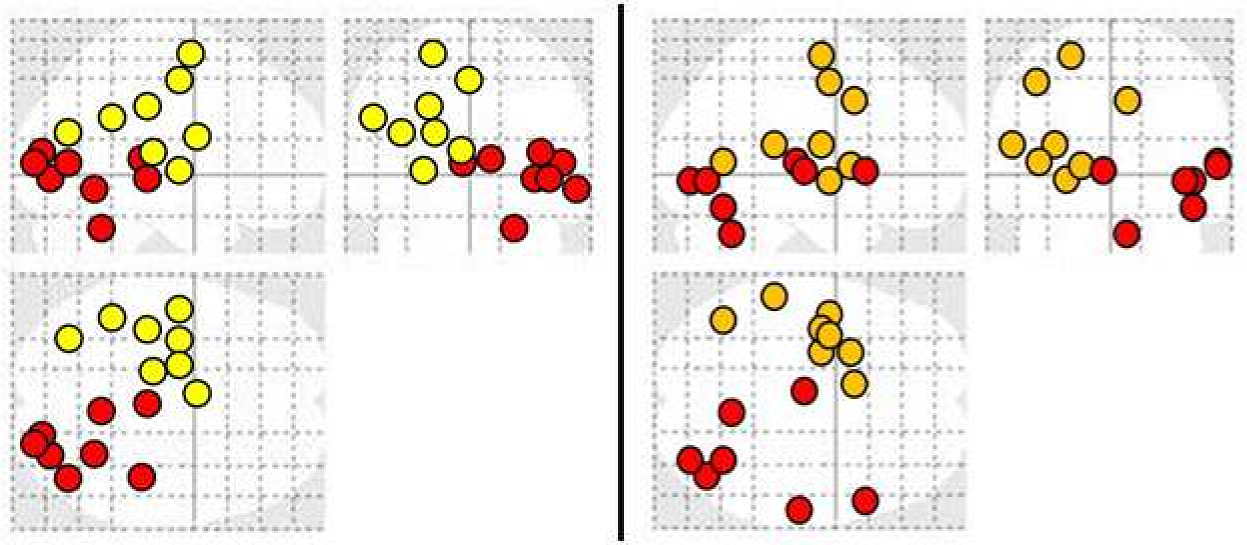

Figure 3.3: *Two sample subjects with 2 groups of correlated clusters. In subjects were 2 clusters were identified, the clusters were more spatially scattered, but are identified by a color scheme.*

Table 3.4: *Comparison of estimation methods runtime. There is a trade off between the run time and complexity of the selected estimation method.*

Runtime (seconds)			
Zero-Substitution	*Noise-Substitution*	*Mean-Substitution*	*EM-Substitution*
0.000311	0.000882	0.000529	0.722002

and varying numbers of model choice. Having two missing nodes instead of one causes a reduction in the performance, indicating that a greater proportion of missing data negatively affects the estimation process. This is intuitive since there is a severe decrease from mean value of 83.125% to 63.75% when dropping a second node in the case of EM substitution, and a similar trend for other methods. Also, when comparing having the full simulation data to having one or two missing nodes, it is noted that classification accuracy increases, given more available data.

The effect of increasing the noise added to the signals on the model selection was also studied. Two levels of additive noise, relative to the signal to noise ratio, were examined - low noise (standard deviation of 0.001) and high noise (standard deviation of 0.01) results in Figure 3.6. There is a general decrease in the percentage of correctly identified models corresponding to systematically increasing the additive noise.

The motivation behind using synthetic data was the existence of the problem of missing data in real studies. When producing SPM's for real subjects, statistics showed differences across the subjects in the number and size of active clusters. The data used to produce Figure 2.4 was taken from [79]. Both volume and cluster analyses for subjects in terms of appearing and disappearing voxels and clusters showed discrepancy in the number of active nodes. Some real subjects showed loss of volume i.e. shrinkage of some cluster(s) or the disappearance of that/those cluster(s) altogether. Also, some subjects showed the gain of volume i.e. expansion of some already existing cluster(s) or the appearance of newly formed cluster(s). Thus, not only are there differences in the shape of activation (shrinkage/expansion) but

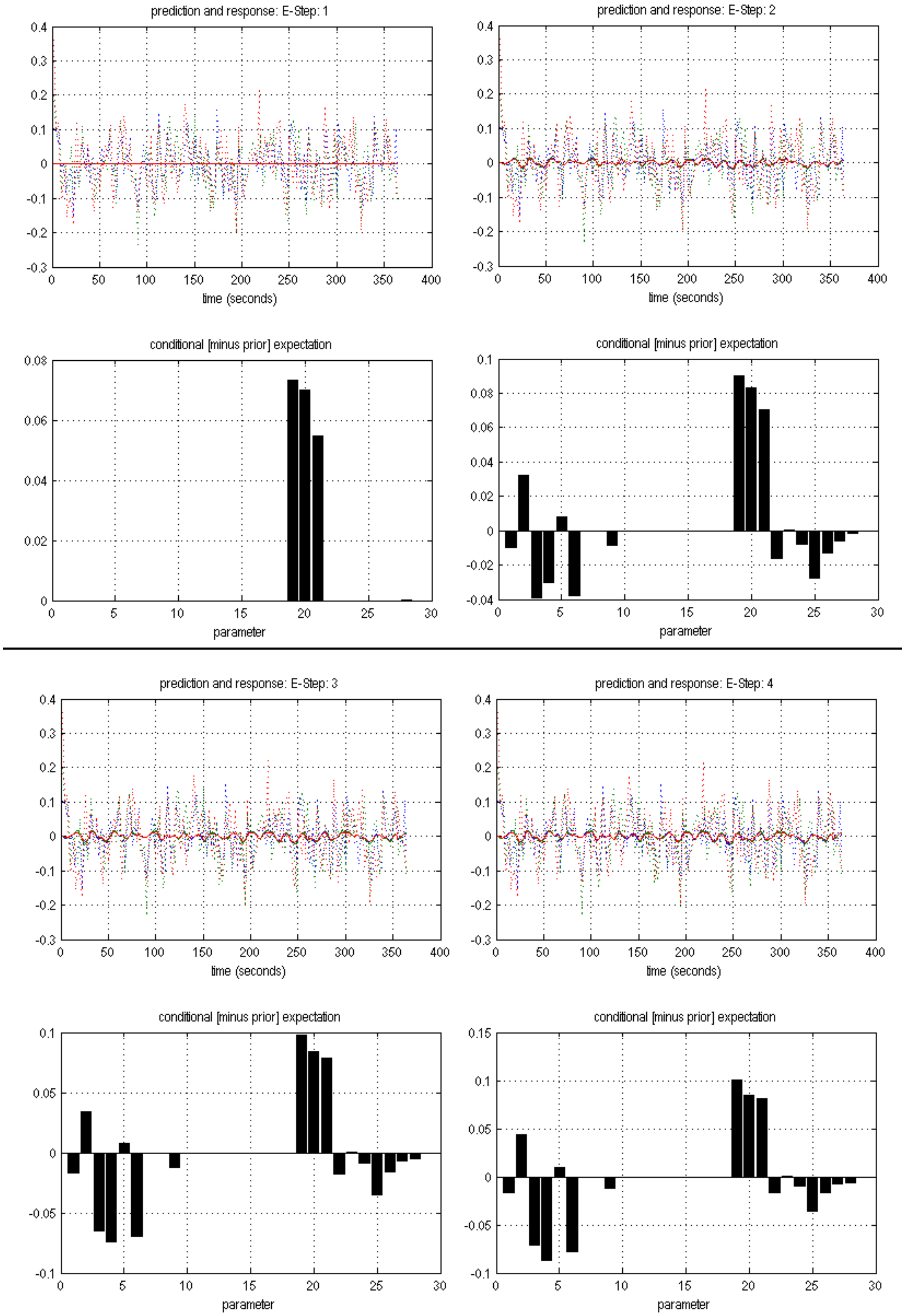

Figure 3.4: *DCM Estimation steps using the Variational Bayes algorithm in SPM.*

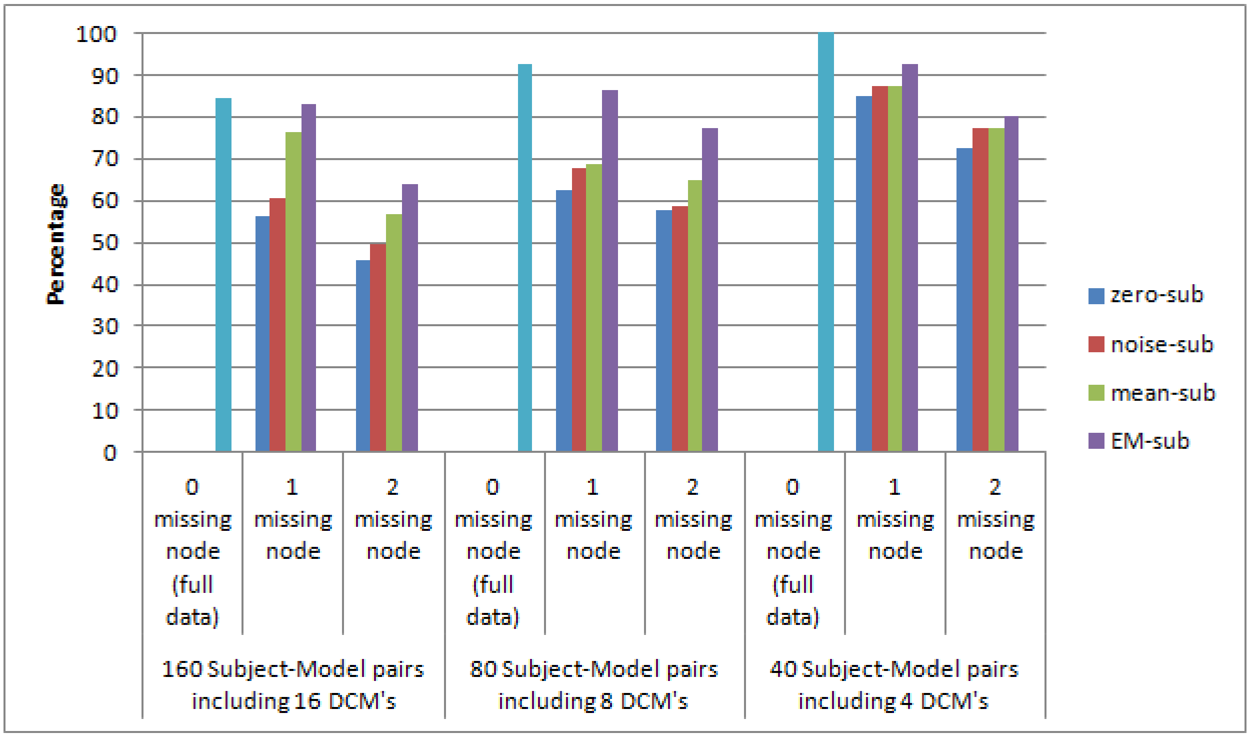

Figure 3.5: *Accuracy of classification based on highest model evidence versus ground truth. Zero missing nodes is the baseline for comparison.*

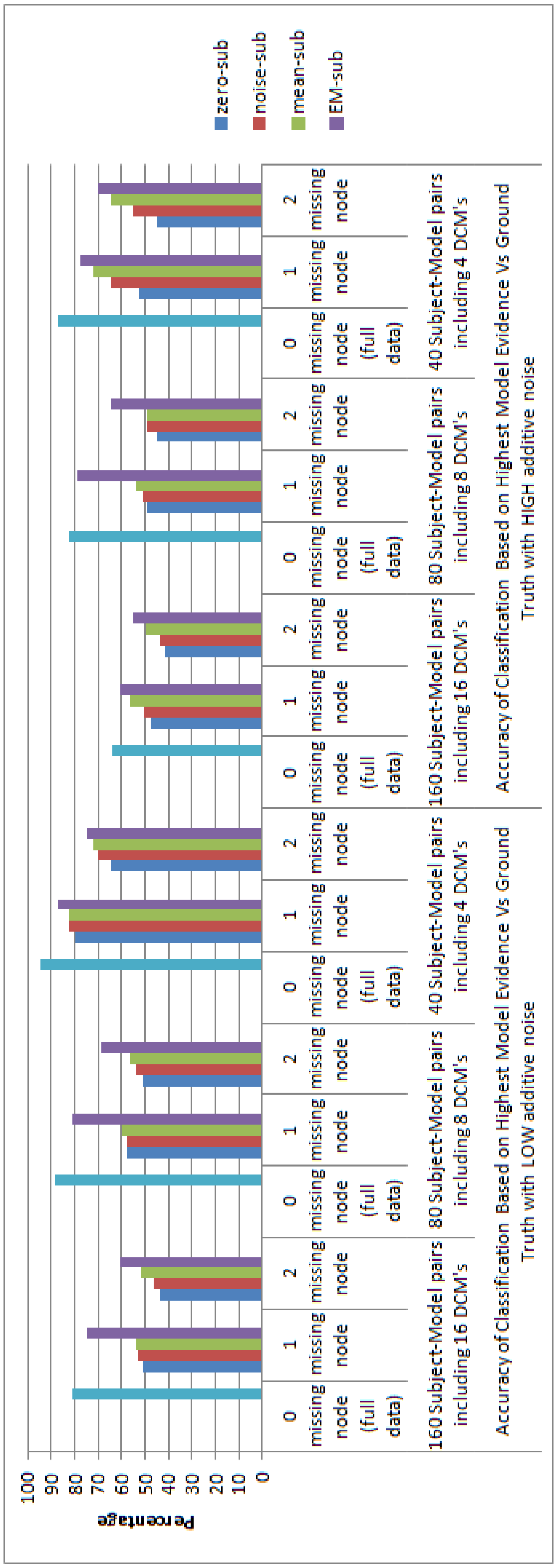

Figure 3.6: *Accuracy of classification based on highest model evidence versus ground truth with varying levels of noise. Zero missing nodes is again the baseline for comparison.*

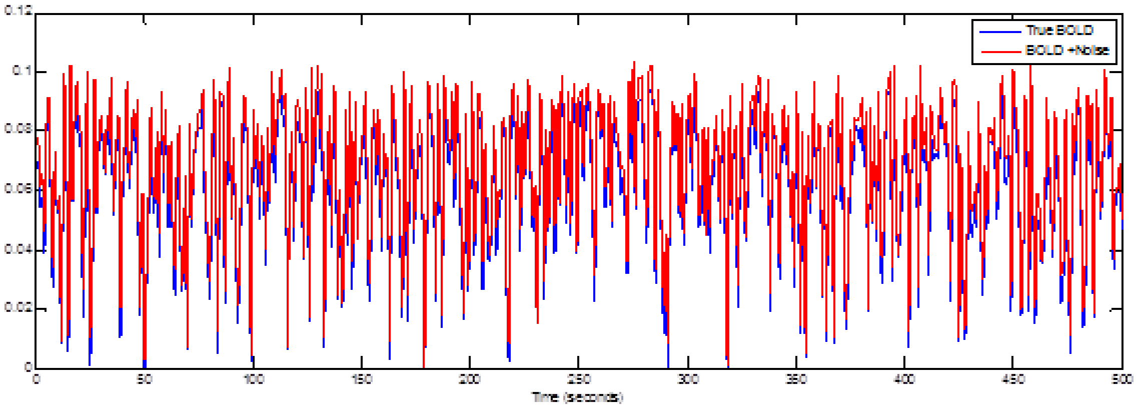

Figure 3.7: *Sample simulated true BOLD signal before and after noise addition.*

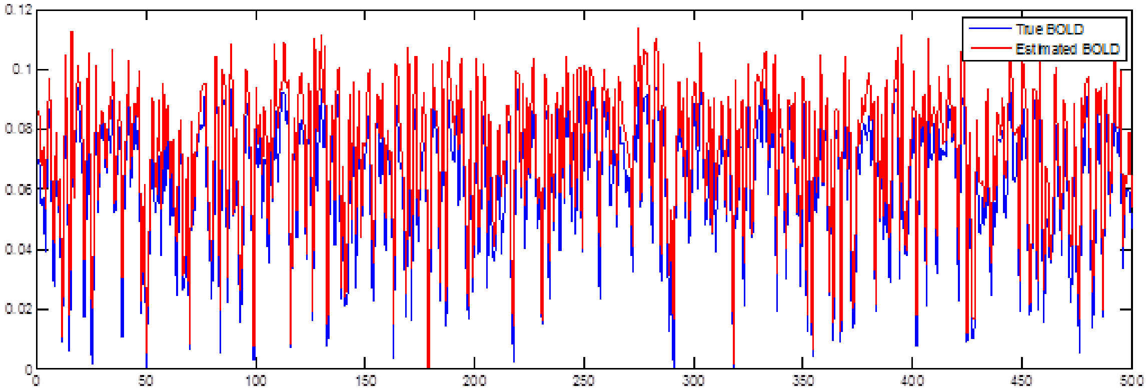

Figure 3.8: *Sample estimated BOLD signal vs. true BOLD signal.*

possibly the addition/deletion of new nodes. This is evidence of FP's and FN's and inevitably leads to the inconsistency in the number of nodes per subject network within a group analysis hindering the computation of the model evidence.

Solutions to the missing data problem included the simple zero-substitution, mean-substitution, and noise substitution methods (which corresponds to having less conservative p-value), and the more complex EM-estimation method. The highest classification rate was obtained when using the EM algorithm as opposed to the other three. The complexity of EM, however is much greater than using those simpler methods. The trade off for the added complexity of EM includes longer runtime, although it is considered trivial in this case, but a better classification rate of subject DCMs was achieved (Table 3.4).

3.2.1 Choice of Loss Function in Classification

Bayesian model selection based on Bayes factor requires the person to choose one particular model out of two or more models, and there could be a 0-1 loss on the decision. This means that making the choice does not depend on how close the choice is and the Bayes factor is sufficient only if a 0-1 loss is obtained. There are other losses available such as conjugate utilities for exponential families which are more natural than step function losses. These lead to other criteria for model selection. Selection of an estimator needs to be consistent with the actual loss experienced in the context of a particular problem. Some common loss functions are the squared loss and the absolute loss.

In these simulations, model selection was based on the relative comparison of evidence values for each model. However, by comparing the evidences of different models and performing model selection/classification by choosing the highest evidence, an assumption of a 0-1 loss is being made. A value of 0 is given if the predicted output is the same as the actual output, and a value of 1 if the predicted output is different from the actual output. This means that although a certain model might have higher evidence than another one, it might not necessarily be the most rational one. There may be occasions when one model clearly dominates the others and other occasions when the choice is misleading. Choosing the correct model would require the consideration of costs which can be included using a certain utility function.

A 0-1 loss function loss function may not be a good choice in a medical scenario because it should be expected for the cost of an incorrect decision to be different from a correct one, i.e. an incorrect decision should have a higher penalty than a correct one. One does not necessarily have to choose a single best model according to some criterion. Models that are poor can be deselected, keeping a subset for further consideration. The subset can consist of a single model or more. In the latter case the model choice can be driven by considering the costs via utility functions.

Model selection was used here to perform classification. Given data from a subject and a number of possible models (m_1, m_2, ... ,m_N), DCM can be used to estimate a bound on the evidence for each model as $\{\ p(data|m_1), p(data|m_2), ..., p(data|m_N)\ \}$. The classification decision involves choosing which model to assign to the subject. However, the cost/utility of a correct versus incorrect assignment is different. For example, m_1 might correspond to a disease state, and m_2 could correspond to a normal state. The cost of labeling the person as sick when they are not is not supposed to be symmetrical to labeling them as normal when they are sick.

The subjective expected utility combines two subjective concepts, a personal utility function, and a personal probability distribution. If an uncertain event has possible outcomes x_i , each with a utility $u(x_i)$ then the choices can be explained as arising from a function in which there is belief that there is a subjective probability of each outcome $P(x_i)$. The subjective expected utility is the expected value of the utility:

$$\sum_i u(x_i)P(x_i) \tag{3.1}$$

The decision that is preferred depends on which subjective expected utility is higher. Dif-

Table 3.5: *Accuracy of Classification based on highest model evidence using a 0/1 loss.*

	Percentage of Correctly Identified Models			
0 missing nodes (full data)	86/100=86%			
	Zero-Sub	**Noise-Sub**	**Mean-Sub**	**EM-Sub**
1 missing node	61/100=61%	68/100=68%	70/100=70%	84/100= 84%
2 missing nodes	55/100=55%	63/100=63%	61/100=61%	72/100=72%

ferent decisions can be made using different utility functions or different beliefs about the probabilities of different outcomes. It is possible to take convex combinations of decisions and preserve preferences. This is different from having a 0/1 loss where there is no penalty for an incorrect diagnosis.

The classification decision in the simulations was evaluated under different utilities that could be possible in a clinical case. Two different models were generated (representing diseased/healthy states) with prior probabilities for each model taken from a training population. Given a new subject, the model evidence was estimated for each model, using the priors and the utilities. The prior probabilities were made realistic using asymmetric utilities (the cost of correct versus incorrect diagnosis was different). The sensitivity of the improvement using EM filling was thus tested. This allows the impact/significance of improvement to be more emphasized when using utilities and priors that are relevant to the real world.

Tables 3.5 and 3.6 compare the accuracy of classification when it is being based on the highest relative model evidence and when it is based on using asymmetric utilities respectively. Also, Figures 3.9 and 3.10 show plots for various asymmetries in the utility for one missing node and two missing nodes respectively.

To the best of our knowledge, using zero-substitution, mean-substitution and EM-estimation to estimate the missing nodes in these simulated subjects has not been performed before in

Table 3.6: *Accuracy of Classification based on using asymmetric utilities.*

	Percentage of Correctly Identified Models			
0 missing nodes (full data)	92/100 = 92%			
	Zero-Sub	**Noise-Sub**	**Mean-Sub**	**EM-Sub**
1 missing node	67/100=67%	73/100=73%	75/100=75%	88/100=88%
2 missing nodes	61/100=61%	72/100=72%	71/100=71%	80/100=80%

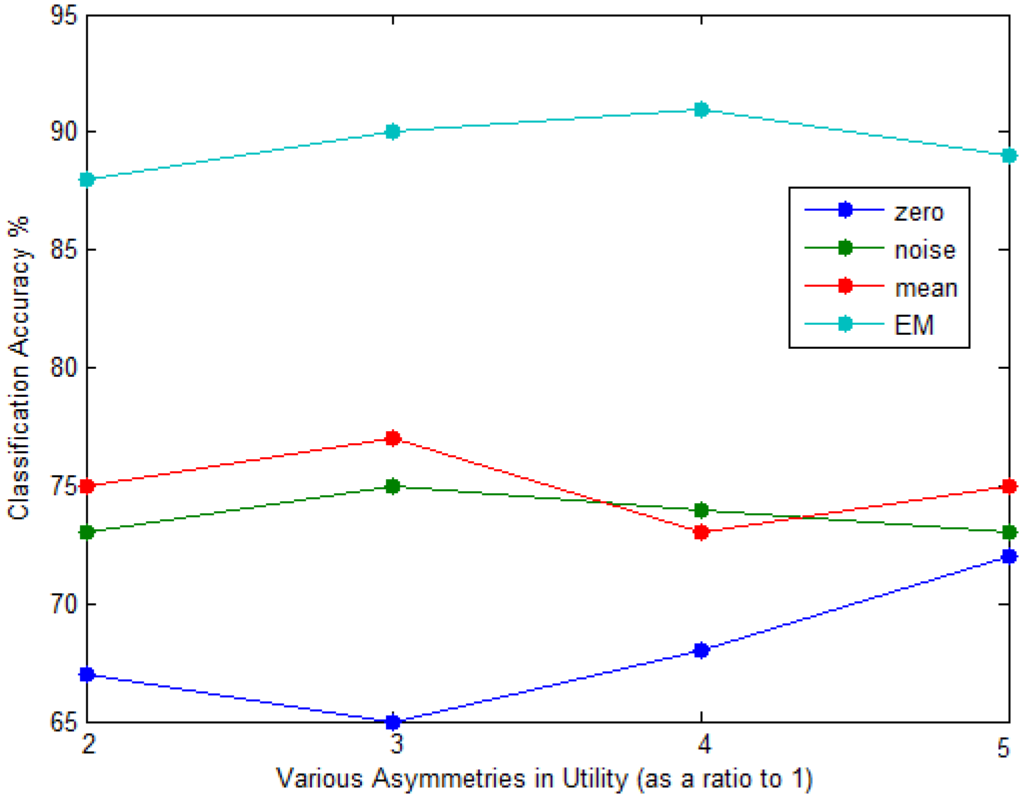

Figure 3.9: *Plot of Asymmetric Utilities for 1 Missing Node.*

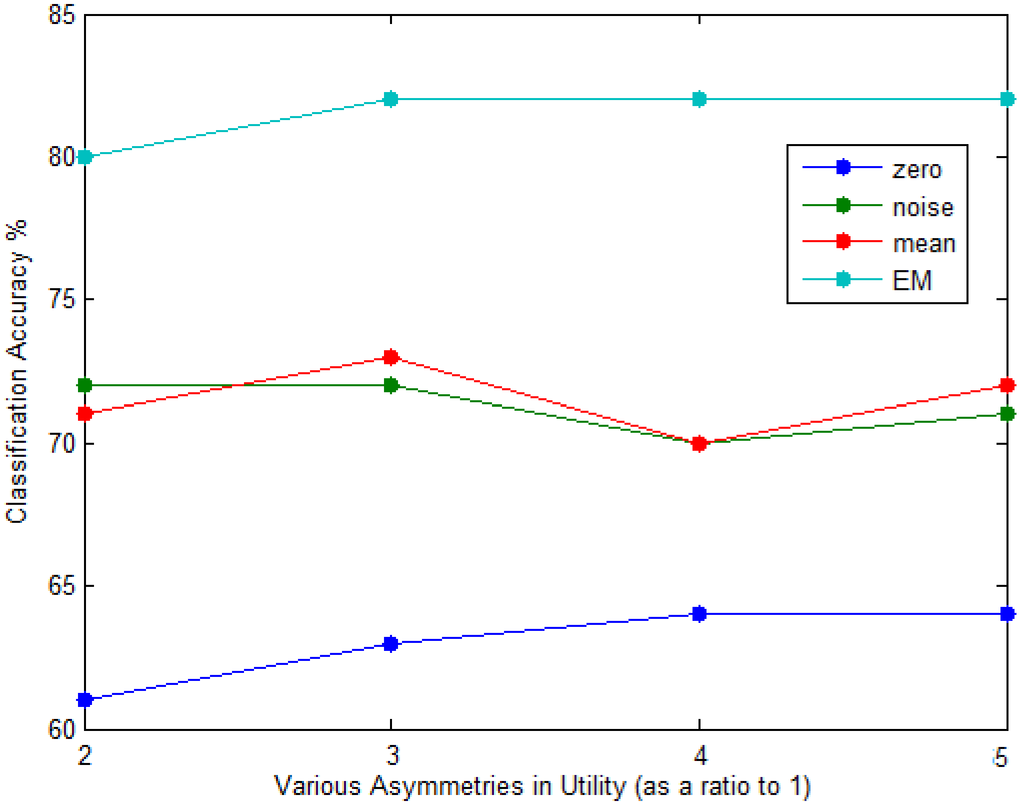

Figure 3.10: *Plot of Asymmetric Utilities for 2 Missing Nodes.*

the context of DCM. In fact, the usage of missing data approaches in general hasn't been used to overcome this problem in DCM. The results presented in this section can be used as preliminary indications of the validity of the proposed solution. The next two sections take this a step further by testing these methods on real fMRI data.

3.3 Results of Estimating Missing Data in Real Datasets

3.3.1 Comparison of Estimated and Actual Time-series and Parameters

Linear regression and Variational Bayes were used for the estimation of parameters for each DCM in the real subjects. To measure the difference between the computed parameters before and after estimation of missing data, for each time series where the full model could be estimated, the Root Mean Square Error (RMSE) was used. RMSE is a quantitative measure of the difference between the SPM computed parameters with the full data, and the computed SPM parameters vector after estimating the missing data. The initial estimated parameter vector is defined as: $\theta = \{A, B, C, h\}$ where the BOLD parameters are $h = \{\tau, \alpha, E_0, V_0, \tau_s, \tau_f, \epsilon\}$. $\hat{\theta}$ is the estimated parameters vector after computing the missing data. The RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum(\hat{\theta}-\theta)^2} \qquad (3.2)$$

Also, the Mutual Information (MI) was computed between the true BOLD signal and the estimated BOLD signal. The MI is a method of measuring the interdependence of two random variables. If two signals are truly independent, then the mutual information will be zero. The mutual information of two discrete random variables x and y can be defined as:

$$MI(x,y) = \sum_{y}\sum_{x} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \tag{3.3}$$

The VOI time series from this dataset were based on centers of peak activation. VOIs were sphere shaped with a radius of 8 mm. The peak activation locations were used to place spheres (8mm) to extract the principal eigenvectors as the nodes' signals. Four nodes were considered ideal for each subject. The p-value was manually tweaked to drop nodes (using a more conservative p-value). Any extra nodes were ignored for all subjects. Missing nodes were estimated using zero-substitution, mean-substitution, and EM-estimation. The parameters were estimated for all subjects. Estimation of parameters was done using both Regression and VB. The MI was computed between the predicted and measured response. The RMSE was measured between the real and estimated parameters.

MI increases when using better estimation methods, i.e. mean-substitution vs. zero-substitution and EM-substitution vs. mean-substitution. RMSE error decreases when using better estimation methods, i.e. mean-substitution vs. zero-substitution and EM-substitution vs. mean-substitution. The more missing nodes there are that need to be estimated, the greater the error RMSE and the lower the MI.

3.3.2 Cluster Analysis

In order to perform the same sort of classification tests in real data, some kind of ground truth must be available. Because the classification of real subjects from any dataset would essentially have no ground truth, the parameter space was explored by looking at the variation/groupings in the estimated parameters for the population. Groupings were compared by looking at the distance between cluster centers by using different values of k in k-means clustering. This gave an idea about the potential variation that would exist in a population if a classifier were to be used. The aim of clustering the parameters was to validate the pattern of results obtained from the simulation experiments. It was expected that there would be

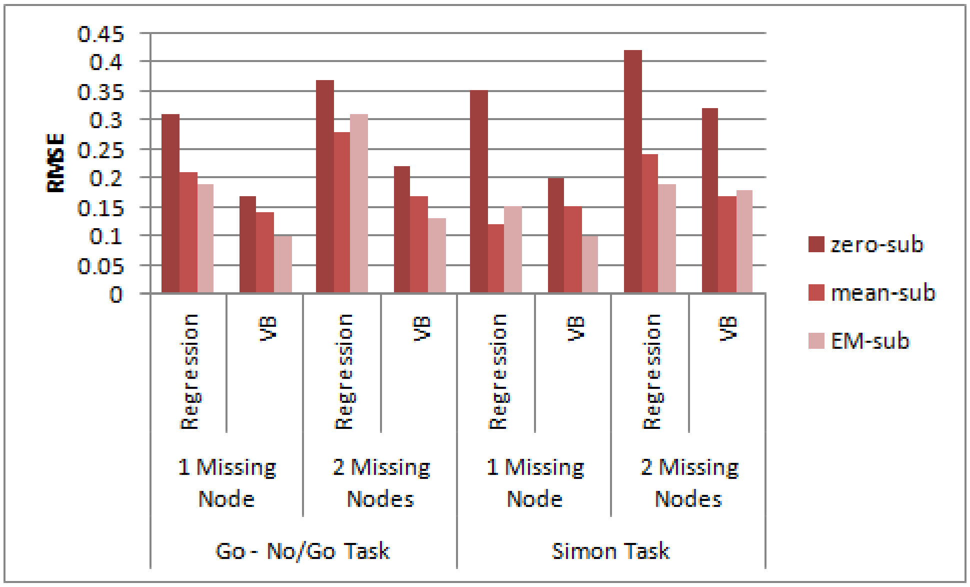

Figure 3.11: *Average RMSE between initial and final parameters for both datasets.*

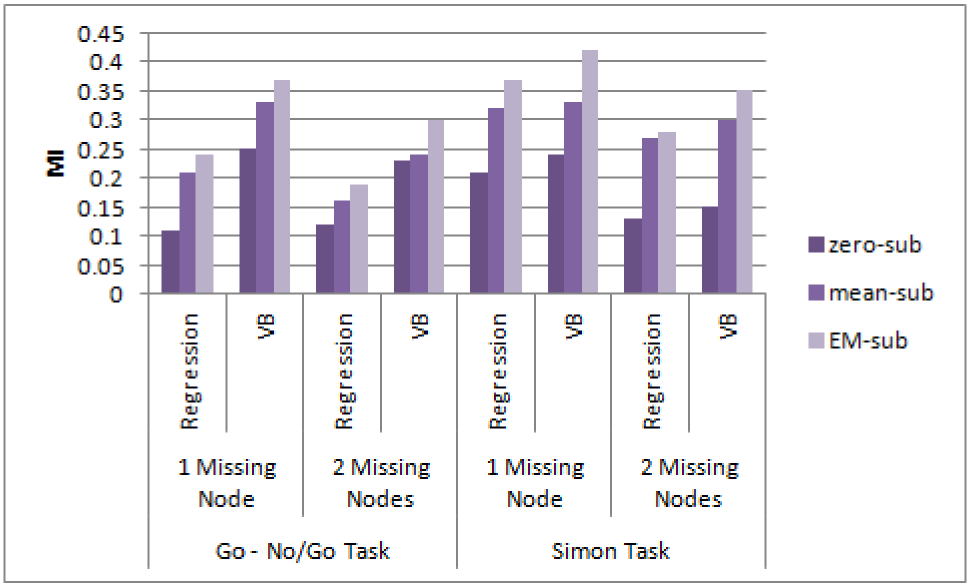

Figure 3.12: *Average MI between predicted and measured response for both datasets.*

an increase in the between cluster scatter and a decrease in the within scatter of clusters if there was an improvement.

For all the subjects from the Go/No-Go dataset, the peak activation locations were used to place spheres (8mm) to extract the principal eigenvectors as the nodes' signals. Four nodes were considered for each subject. Any extra nodes were ignored for all subjects. Missing nodes were estimated using zero-substitution, mean-substitution, and EM-estimation. The parameters were estimated for all subjects. Clustering was performed on the estimated parameters from all the subjects. That was proceeded with some statistical analysis on the clustered parameters including computations of the mean, mode, and standard deviation.

The within cluster sum of squared errors was reduced for all methods (zero-substitution, mean-substitution, and EM-estimation) when the number of clusters k was increased from 2 to 3 to 4. The within cluster sum of squared errors was also higher for zero-substitution in comparison with mean-substitution and higher for mean-substitution with respect to EM-estimation. The error magnitude was reduced when using VB versus regression.

Also, a comparison of distance between cluster centers for different values of k (averaged over all parameters and all clusters) was performed. The distance between cluster centers is based on several factors. An increase in the number of clusters spreads them out, decreasing the distances between them. The distance between cluster centers was smaller for zero-substitution in comparison with mean-substitution and smaller for mean-substitution with respect to EM-estimation. These patterns can be deduced from Figure 3.13.

Details of the cluster composition are shown in Figure 3.14. Parameter estimation for all subjects was done using both regression and VB. The specific cluster compositions might not indicate more than that changing the number of clusters obviously changes each cluster composition and the determination of population distribution as a whole. However, the formed cluster groupings for the various k values are indications for variations existing among these populations. Since these variations can be grouped, this would allow the usage of classification to identify or label an unknown subject based on the formed groups/clusters.

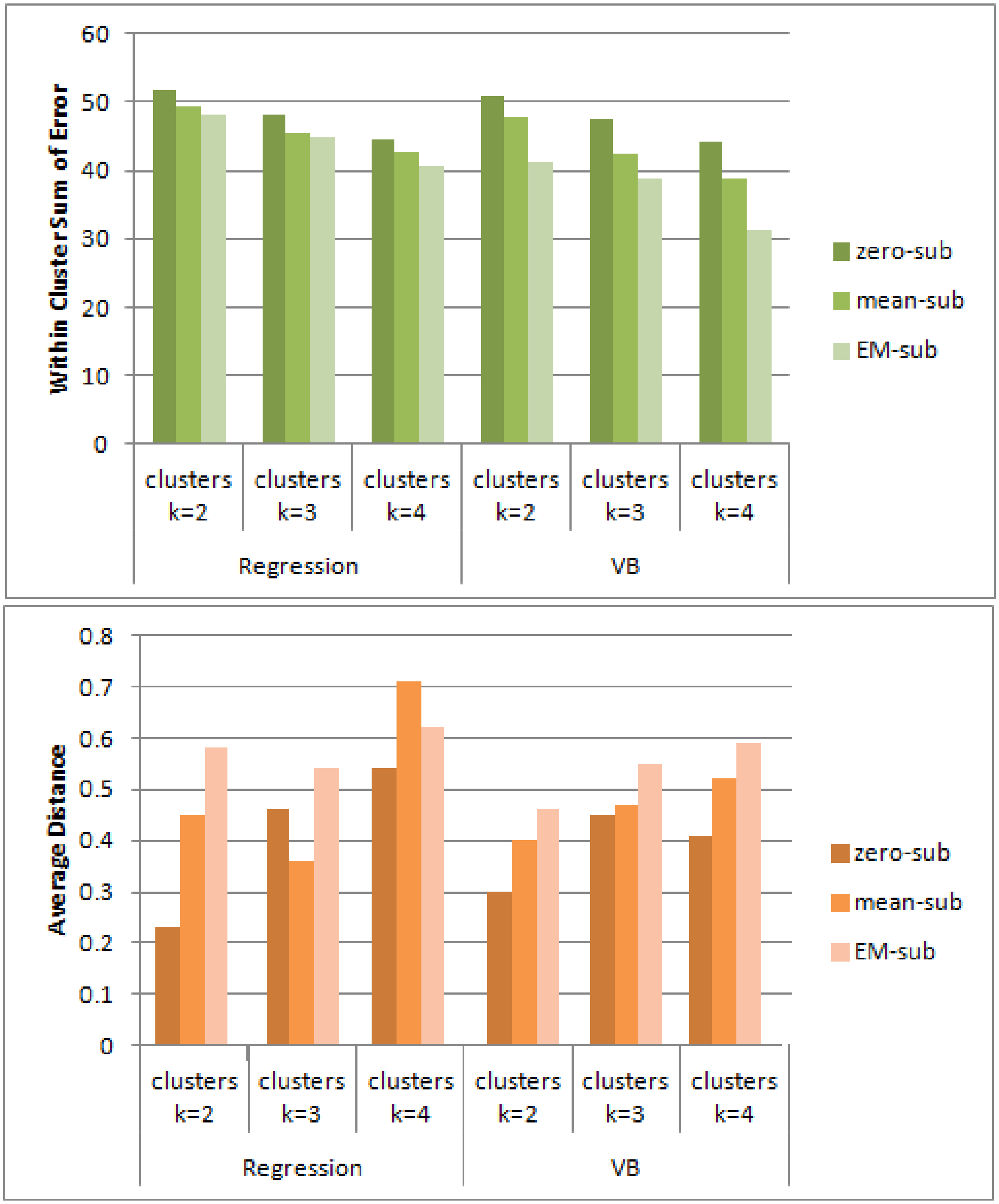

Figure 3.13: *The within cluster sum of squared errors (top) and average distance between cluster centers for the Go No/Go dataset (bottom).*

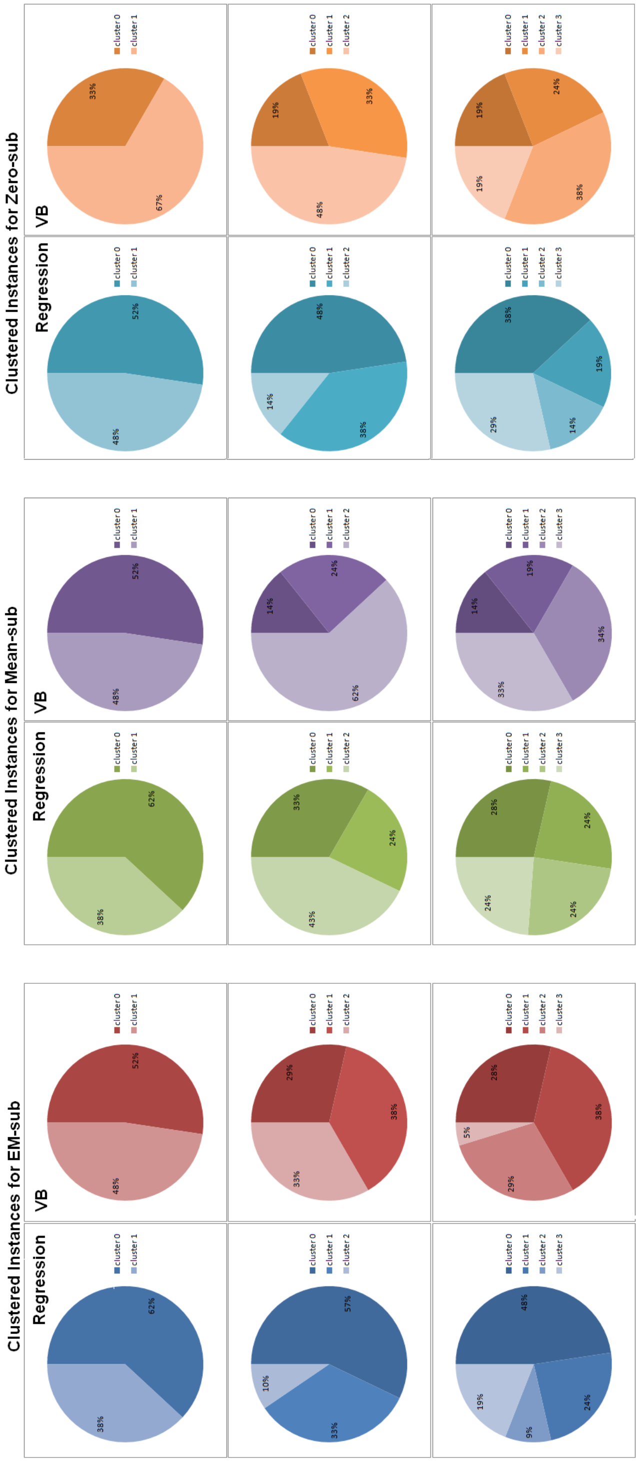

Figure 3.14: *The percentage of population belonging to each cluster for varying values of k, using all three methods (EM-sub, Mean-sub, and Zero-sub). This is shown using both Regression and VB for parameter estimation.*

3.3.3 Anatomical Regions Involved in Simon Task

Using the BRAINMAP software [80] to search for studies associated with the Simon task paradigm, the expected areas of activations and potential model designs were studied. This was used to determine the anatomical areas involved in the regions of activation and to determine common regions among subjects.

In [81], activations observed during the Simon task included anterior cingulate, supplementary motor, visual association, inferior temporal, inferior parietal, inferior frontal, and dorsolateral prefrontal cortices, as well as the caudate nuclei.

In [82], the Simon task activated brain regions that serve as a source of attentional control (dorsolateral prefrontal cortex) and posterior regions that are sites of attentional control (the visual processing stream-middle occipital and inferior temporal cortices). Other brain regions activated by the Simon task were those sensitive to detection of response conflict, response selection, and planning (anterior cingulate cortex, supplementary motor areas, and precuneus), and visuospatial-motor association areas.

A sample of the locations involved in a Simon task using the Incongruent versus Congruent contrast can be shown in Table 3.7. The Automated Anatomical Labeling (AAL) in SPM was used for the labeling of clusters in subjects. Table 3.8 shows a detailed description of cluster contents for a particular subject in terms of the number of voxels and the anatomical structure components. The common regions were analyzed across the dataset resulting in common regions in only a subset of subjects (discarding about 1/3 of the dataset). Figure 3.15 shows the common regions of activations. Common regions included:

Right Cerebrum

Parietal Lobe

Precuneus

Supramarginal Gyrus

Angular-R

Limbic Lobe

Cingulate Gyrus

brodmann area 32

Cingulum-Mid-R

Medial Frontal Gyru (aal)

Left Cerebrum

Temporal Lobe

Supramarginal Gyrus

Angular-L

Frontal Lobe

Middle Frontal Gyrus

Frontal-Mid-L (aal)

3.4 Results of Ranking Experiments

The different models described in Section 2.5 were used to measure how sensitive the ranking is to the number of subjects N. The model evidence was calculated using AIC/BIC and using the free energy approximation. The different models were again fitted to data collected from the N simulated subjects using VB. Different values of N were tested ranging from $10, 15, 20, ..., 100$. Figure 3.16 shows this computed sensitivity measure as the population size is increased. There is an obvious trend of decreasing sensitivity as the population grows.

The effect on the rankings of having to drop subjects because of missing data is also shown. Using the same 10, 15, 20, ... , 100 subjects, and randomly removing nodes from about a third of the subjects the percentage change of the subjects' labels was also noted. Figure 3.17 shows the number of subjects whose label became affected due to the dropping of subjects as a percentage of the population.

EM was then used to fill in the missing data to test if the original ranking was restored. The comparison was done with the original 2/3 of subjects that were included in the previous

Table 3.7: *Locations involved in Simon Task (Incongruent vs. Congruent Contrast).*

X (mm)	Y (mm)	Z (mm)	Talairach X (mm)	Talairach Y (mm)	Talairach Labels
31	0	45	27.14	-5.84	Right Cerebrum,Frontal Lobe, Middle Frontal Gyrus,Gray Matter, Brodmann area 6
-32	-3	43	-31.15	-8.15	Left Cerebrum,Frontal Lobe, Sub-Gyral,White Matter
29	-75	10	25.52	-72.65	Right Cerebrum,Occipital Lobe, Cuneus.White Matter
-28	-77	9	-27.22	-74.15	Left Cerebrum,Occipital Lobe, Middle Occipital Gyrus.,White Matter
11	17	21	8.95	12.19	Right Cerebrum,Limbic Lobe, Anterior Cingulate,White Matter
-10	13	24	-10.53	8.3	Left Cerebrum,Limbic Lobe, Cingulate Gyrus,White Matter
43	-49	-10	38.77	-46.75	Right Cerebrum,Occipital Lobe, Fusiform Gyrus,White Matter
-46	-52	-10	-43.6	-49.12	Left Cerebrum,Temporal Lobe, Sub-Gyral,White Matter
45	-38	32	40.15	-40.17	Right Cerebrum,Parietal Lobe, Inferior Parietal Lobule,White Matter
-45	-37	31	-43.12	-38.72	Left Cerebrum,Parietal Lobe, Inferior Parietal Lobule,White Matter
44	8	13	39.56	4.34	Right Cerebrum,Sub-lobar,Insula, White Matter
-44	5	14	-41.89	1.88	Left Cerebrum,Sub-lobar,Insula, White Matter

Table 3.8: *Cluster Composition for a particular subject.*

Number of Voxels	Structure	Number of Voxels	Structure
2669	Right Cerebrum	119	Supp-Motor-Area-L (aal)
1461	Gray Matter	117	Inter-Hemispheric
1370	White Matter	116	Brodmann area 7
1227	Parietal Lobe	115	Anterior Cingulate
1142	Frontal Lobe	114	Brodmann area 32
592	Postcentral Gyrus	113	Supp-Motor-Area-R (aal)
512	Postcentral-R (aal)	104	Precuneus-L (aal)
486	Left Cerebrum	101	Brodmann area 24
371	SupraMarginal-R (aal)	101	Sub-lobar
357	Limbic Lobe	96	Frontal-Sup-Medial-L (aal)
327	Temporal Lobe	94	Brodmann area 2
322	Middle Frontal Gyrus	93	Parietal-Sup-R (aal)
295	Temporal-Sup-R (aal)	88	Brodmann area 3
278	Cingulate Gyrus	86	Cingulum-Ant-L (aal)
259	Inferior Parietal Lobule	84	Supramarginal Gyrus
255	Medial Frontal Gyrus	83	Brodmann area 5
253	Precuneus	82	Insula
251	Superior Temporal Gyrus	82	Sub-Gyral
236	brodmann area 6	75	Cingulum-Ant-R (aal)
230	Cingulum-Mid-R (aal)	70	Rolandic-Oper-R (aal)
230	Frontal-Mid-R (aal)	61	Temporal-Mid-R (aal)
218	Precentral-R (aal)	60	Brodmann area 9
208	Precentral Gyrus	57	Brodmann area 4
193	Paracentral Lobule	55	Brodmann area 41
179	Brodmann area 40	51	Brodmann area 31
148	Precuneus-R (aal)	46	Brodmann area 8
141	Cingulum-Mid-L (aal)	43	Parietal-Inf-R (aal)

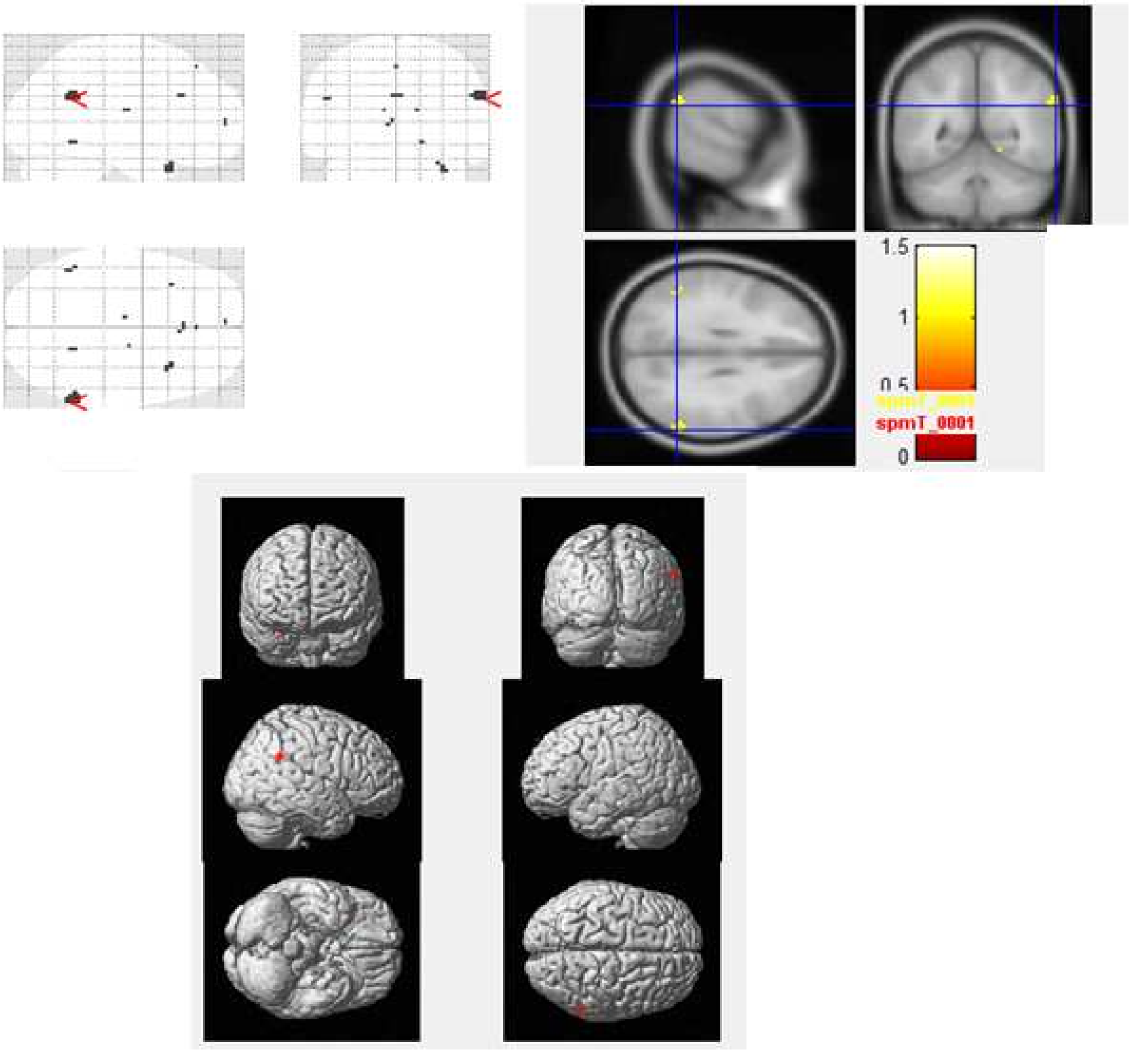

Figure 3.15: *Common regions of activations for Simon Task. The overlap is quite small as this is computed taking into account all the subjects in the analysis.*

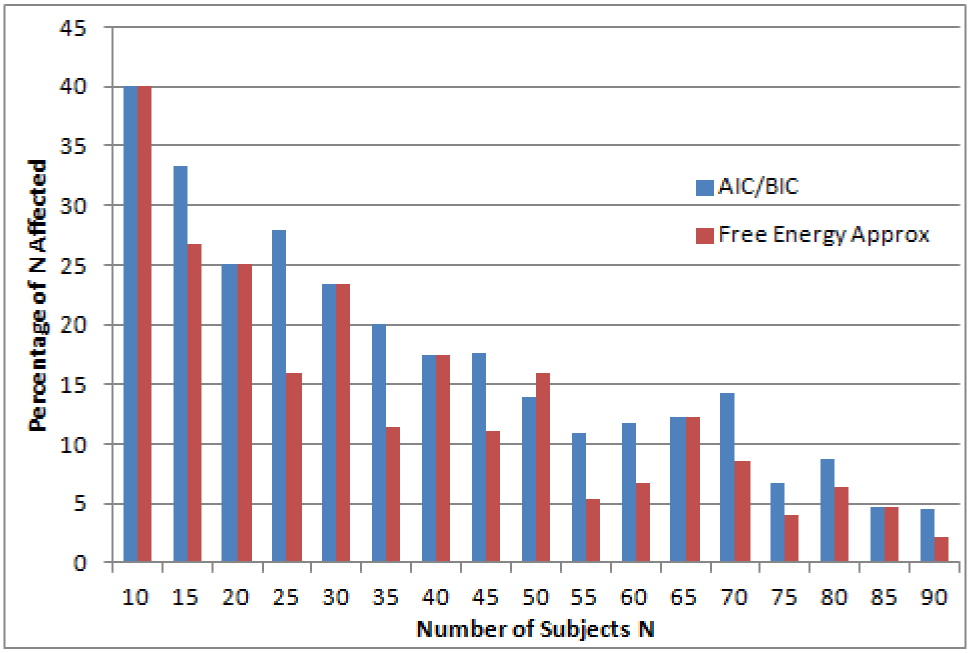

Figure 3.16: *Sensitivity of Ranking of Models on the Population Size. The trend shows a decline in the sensitivity of model ranking as the number of subjects in the experiment is increased.*

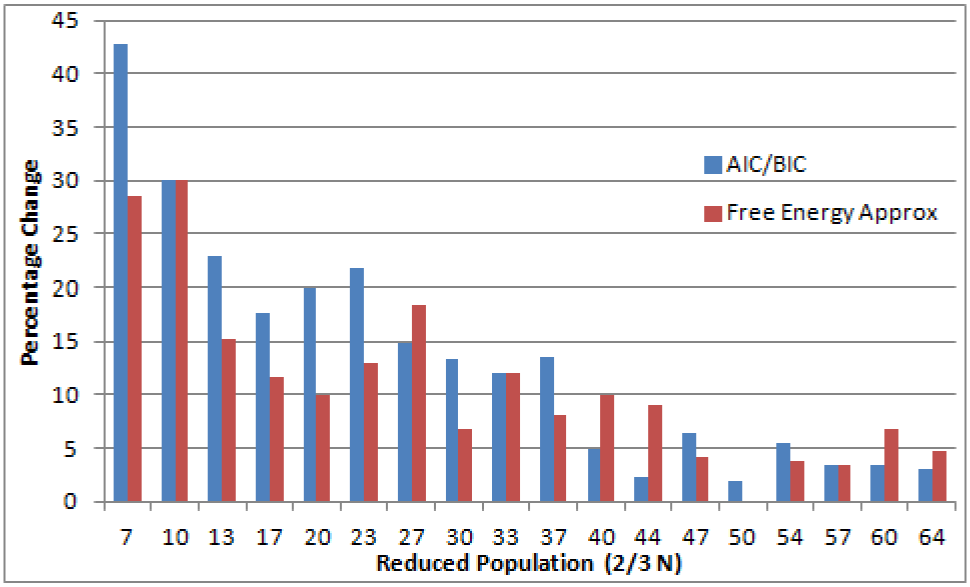

Figure 3.17: *Effect on the labeling process of dropping subjects.*

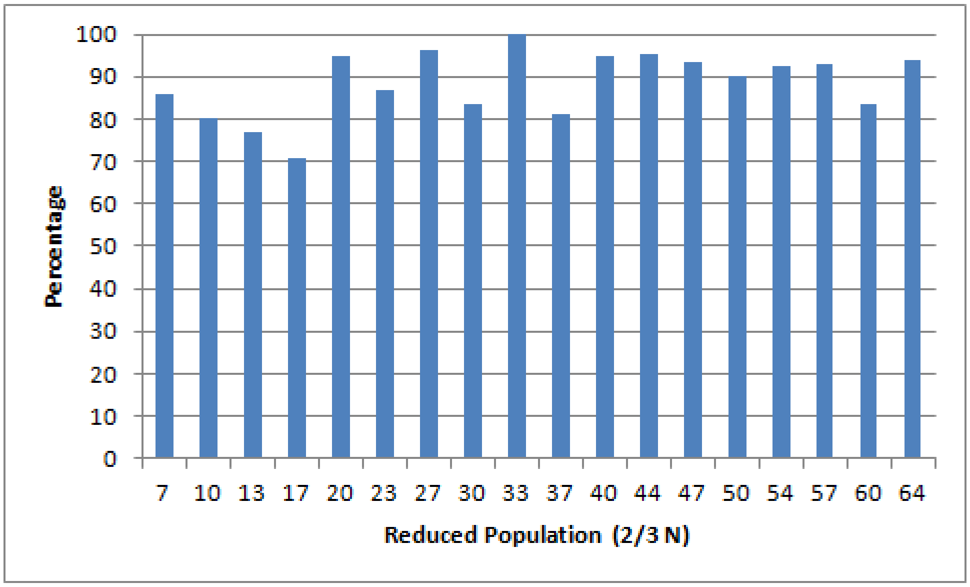

Figure 3.18: *Effect of EM on correct labeling of subjects.*

round. Figure 3.18 shows the percentage of subjects that were correctly re-labeled after EM was used to fill in the missing nodes.

According to the numbers, the sensitivity of the ranking of subjects decreased as N increased. This could be because as N is increased in the denominator, the effect of a single subject on the entire group decreases. Also, it appears that the degree of effect on the rankings due to the dropping of subjects because of missing data decreased as N increased. This could be because of the same reason. EM estimation to fill in the missing data allows the re-labeling of original labels to a great extent but not entirely. This could be due to the differences between the ground truth and the estimated data.

In the study of the effect of using EM to restore missing data on the correct labeling of subjects, the differences between Figure 3.17 and Figure 3.18 are shown in Figure 3.19 using regression for the estimation of parameters. Also, shown in Figure 3.20, is the same sort of post-analysis but using VB instead of regression.

In general, the sensitivity of the ranking of subjects decreases as N increases. This could be because as N increases in the denominator, the effect of a single subject on the group

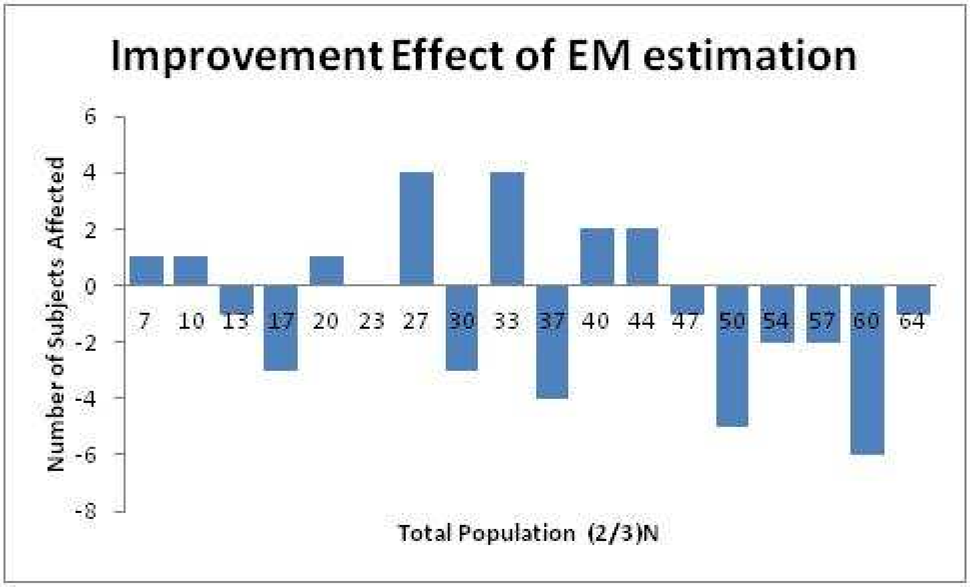

Figure 3.19: *Effect of using EM to restore missing data on the correct labeling of subjects (regression)*

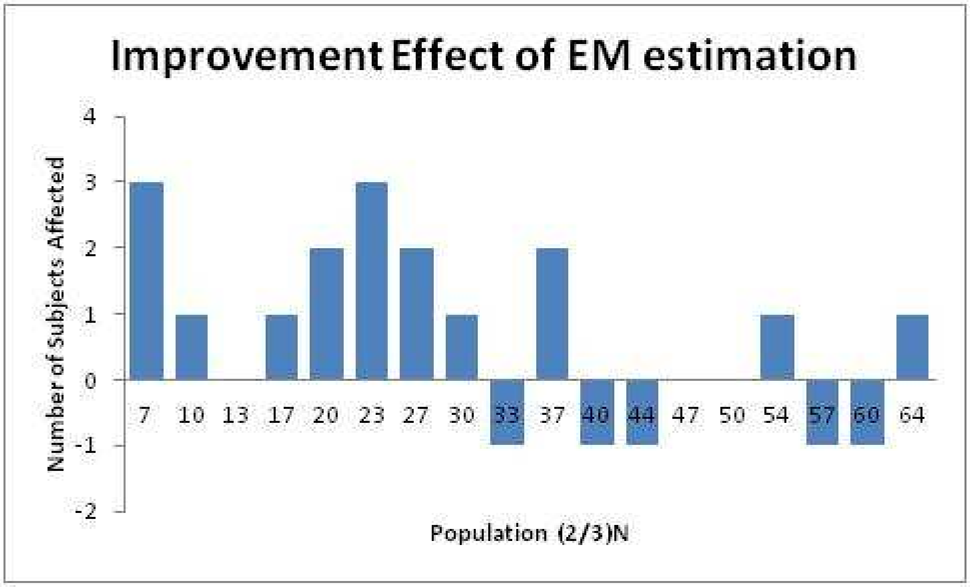

Figure 3.20: *Effect of using EM to restore missing data on the correct labeling of subjects (VB)*

decreases. Also for the same reason, the degree of effect on the rankings due to the dropping of subjects because of missing data decreases as N increases.

EM estimation to fill in the missing data leads to changes in the original ranking. This could be due to the differences between the ground truth and the estimated data. In order to compare the differences caused by using regression versus VB the improvement factor is also studied. It appears that for regression, there is mostly a positive improvement until the population size expands to a certain limit after which the improvement is negative. As for VB, the improvement is mostly positive and although it is slightly reduced as the population size is expanded. Negative improvement is considered negligible and only for the larger population sizes.

Using the Go/No-Go dataset, comparing the different models was done by first specifying alternative models and comparing against the others that were defined and estimated. The basic DCM, from which all other models were derived, is shown in Figure 3.21. Figure 3.23 shows the log evidence and model posterior probability for one of subjects, indicating that Model 10 was the best for that particular subject. Table 3.10 shows the best model computed for each of the other subjects. Model 10 had the highest recurrence among the subjects. Out of the subjects that had one or more missing nodes, 5/8 or 63% selected Model 10. Looking at the entire group, 6/21 or 29% selected Model 10, which seems to be the most commonly selected model.

Using the same data but changing the inference method to RFX instead of FFX, which uses Gibbs sampling, Figure 3.24 shows that again model 10 is the best model, but not by much. The expected posterior probability and the exceedance probability of each model (i.e. the probability that this model is more likely than any other model) are plotted. Table 3.11 shows the best model computed for each of the other subjects. With respect to the group, Model 10 had the highest selection rate of 7/21 or 33%. Considering only the subjects that had one or more missing nodes, 7/8 or 87.5% selected Model 10.

Doing FFX Bayesian model averaging on all models, and selecting the Winning Family

Table 3.9: *11 models in Occam's window.*

Model 3	$p(m\mid Y)$=0.03
Model 6	$p(m\mid Y)$=0.03
Model 7	$p(m\mid Y)$=0.05
Model 8	$p(m\mid Y)$=0.03
Model 9	$p(m\mid Y)$=0.15
Model 10	$p(m\mid Y)$=0.31
Model 11	$p(m\mid Y)$=0.18
Model 13	$p(m\mid Y)$=0.04
Model 14	$p(m\mid Y)$=0.05
Model 15	$p(m\mid Y)$=0.07
Model 16	$p(m\mid Y)$=0.04

option, a weighted average of the model parameters where the weights are given by the evidences for each model was implemented. After averaging, the number of models in Occam's window was 11 (models 3,6,7,8,9,10,11,13,14,15,16) shown in Table 3.9. Figure 3.25 shows the Bayesian Model Averaging (BMA) results.

The same type of analysis was repeated using the Simon Task Dataset as well. The basic DCM, from which all other models were derived, is shown in Figure 3.26. Visual input enters region V. The three regions have full intrinsic connectivity. The modulatory input was then used to modulate a subset of connections in the model. These are the forward and backward connections between M and V, and the forward and backward connections between P and V. Since these are either present or absent this results in $2^4 = 16$ different DCMs.

Figure 3.27 shows the log evidence and model posterior probability for one of subjects, indicating that Model 7 was the best for that particular subject. Table 3.12 shows the best model computed for each of the other subjects. Model 7 had the highest recurrence among

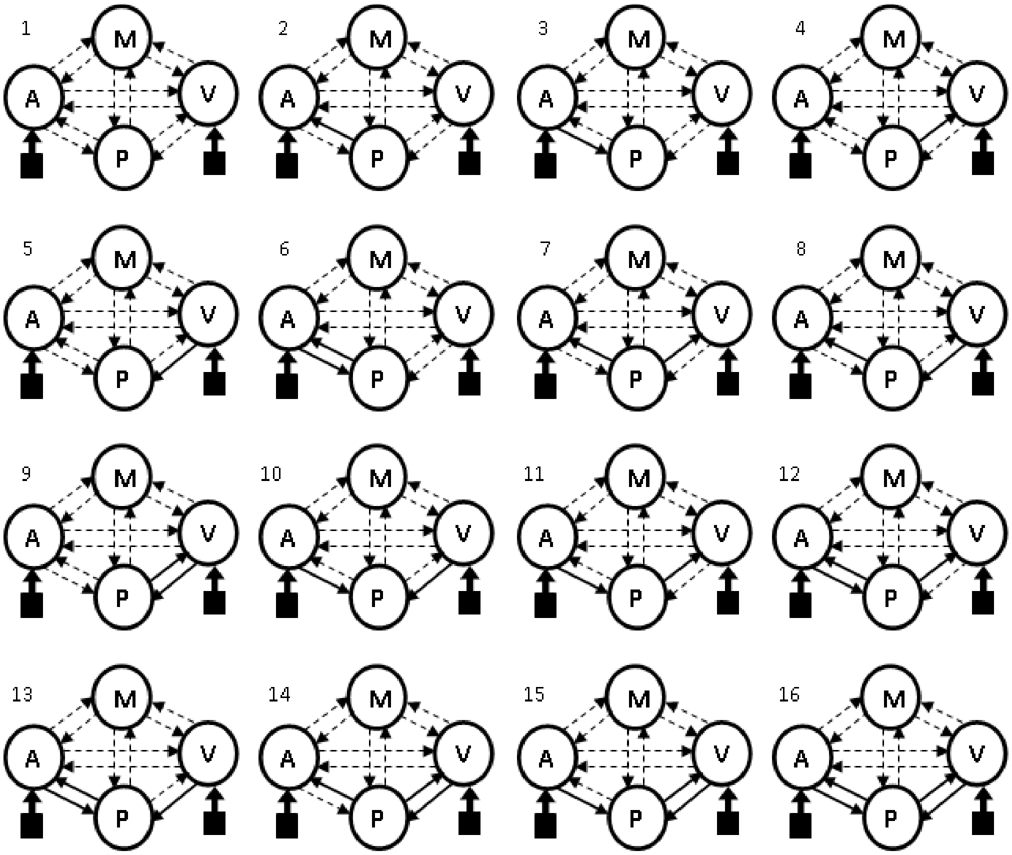

Figure 3.21: *Different models used for the Go/No-Go task. All DCMs are fully connected between all four regions (dotted lines). Inputs are indicated by the black boxes. The 16 models differ in their modulatory connectivity (solid lines).*

Table 3.10: *Best model for each subject (Go/No Go Task) using FFX.*

Subject Number	Had Missing Node(s)	Best Model
1	yes	10
2	yes	10
3	no	6
4	no	8
5	yes	11
6	no	9
7	no	7
8	no	10
9	yes	12
10	yes	10
11	no	13
12	no	6
13	no	11
14	no	8
15	yes	10
16	no	7
17	no	6
18	yes	10
19	yes	9
20	no	7
21	no	8

Table 3.11: *Best model for each subject (Go/No Go Task) using RFX.*

Subject Number	Had Missing Node(s)	Best Model
1	yes	10
2	yes	10
3	no	6
4	no	9
5	yes	10
6	no	9
7	no	7
8	no	8
9	yes	10
10	yes	10
11	no	13
12	no	6
13	no	11
14	no	8
15	yes	10
16	no	7
17	no	6
18	yes	10
19	yes	9
20	no	7
21	no	9

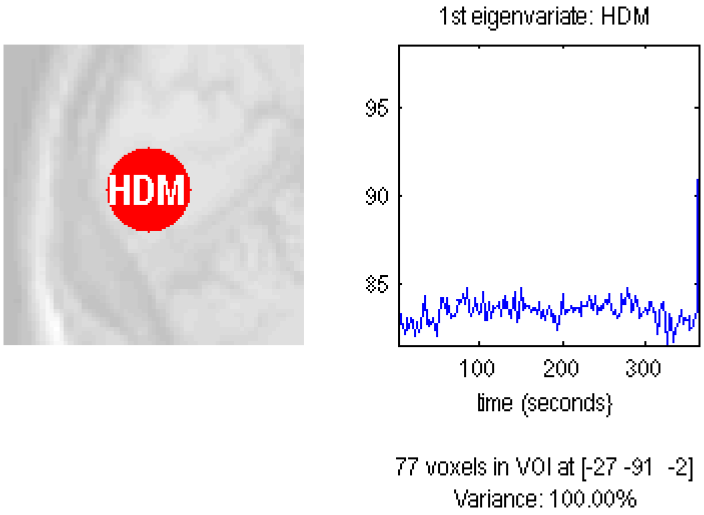

Figure 3.22: *Sample response of given node. This response was measured from the center of a specified given node i.e. VOI (consisting of 77 voxels) for a given subject. A plot of the 1st eigenvariate is shown.*

the subjects. Out of the subjects that had one or more missing nodes, 7/11 or 64% selected Model 7. Looking at the entire group, 8/21 or 38% selected Model 7, which seems to be the most commonly selected model.

Using the same data but changing the inference method to RFX instead of FFX, Figure 3.28 shows that again, model 7 was the best model. The expected posterior probability and the exceedance probability of each model (i.e. the probability that this model is more likely than any other model) are plotted. Table 3.13 shows the best model computed for each of the other subjects. With respect to the group, Model 7 had the highest selection rate of 8/21 or 38%. Considering only the subjects that had one or more missing nodes, 7/11 or 64% selected Model 7.

Doing FFX Bayesian model averaging on all models, and selecting the Winning Family option, a weighted average of the model parameters where the weights are given by the evidences for each model was implemented. After averaging, the number of models in Occam's window was 5 (models 6,7,8,9,10) as shown in Table 3.14. Figure 3.29 shows the Bayesian

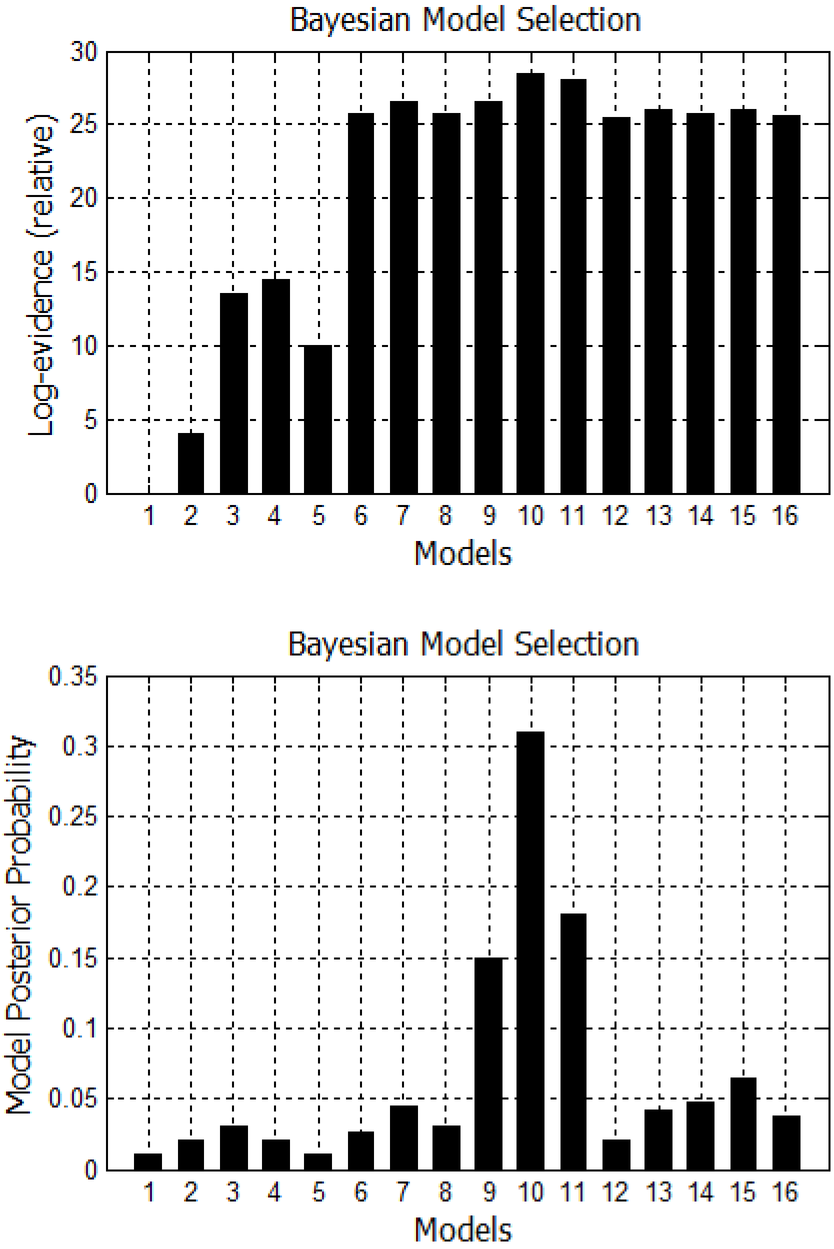

Figure 3.23: *Fixed Effects Analysis for Go/No-Go Dataset. The relative log-evidence and model posterior probability are highest for the 10th model.*

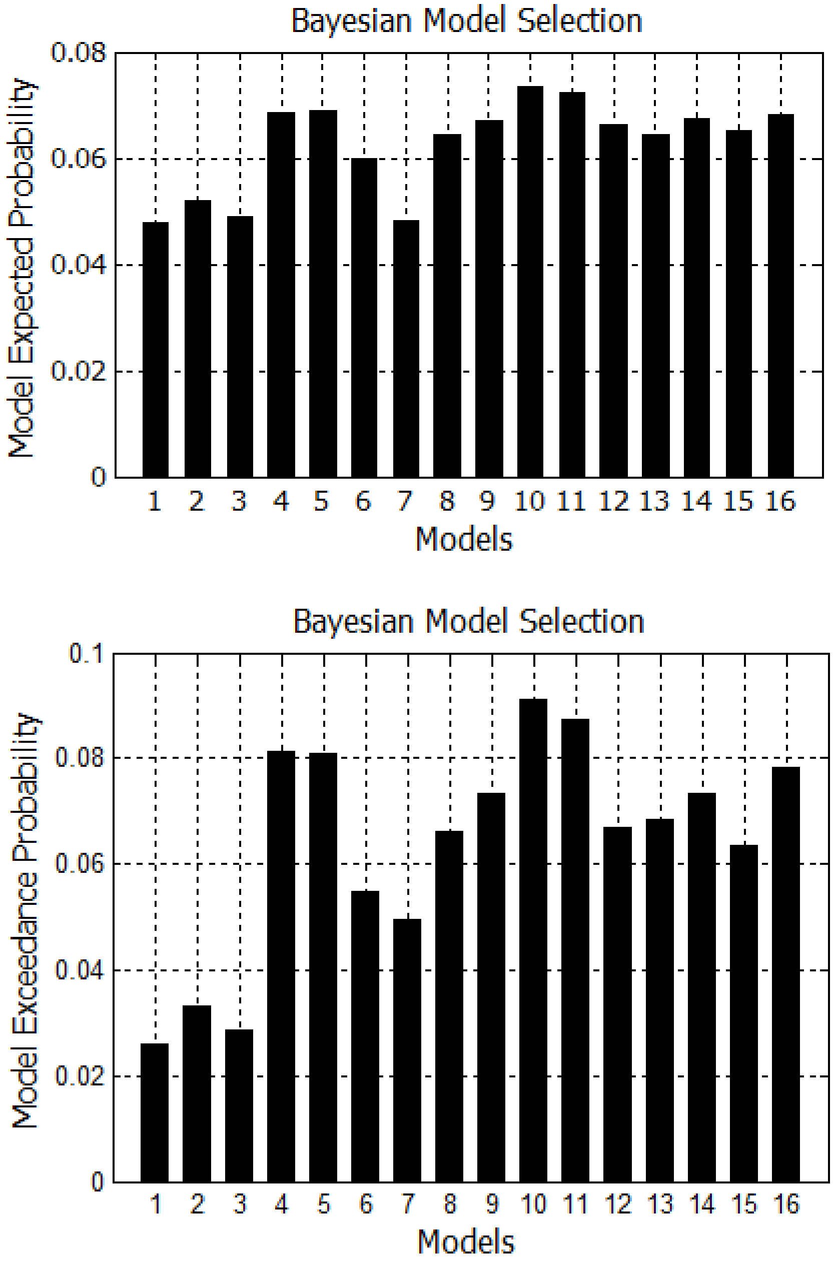

Figure 3.24: *Random Effects Analysis for Go/No-Go Dataset. The model expected probability and model exceedance probabilities are again highest for Model 10.*

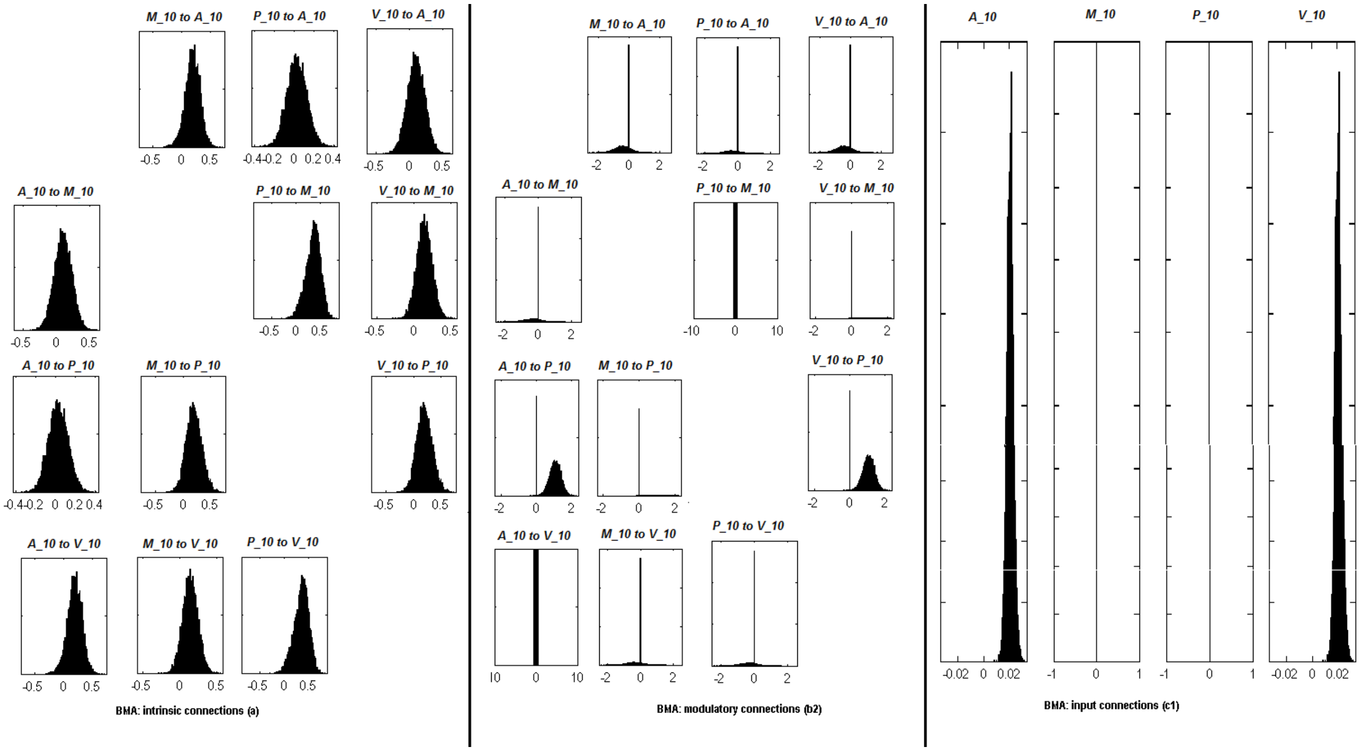

Figure 3.25: *Bayesian model averaging over all 16 models for Go/No-Go dataset. The left figure shows the connectivity between nodes.The modulatory action is distinguished between nodes A and P, and also between nodes V and P as shown in the middle figure. The right figure shows where the inputs are connected.*

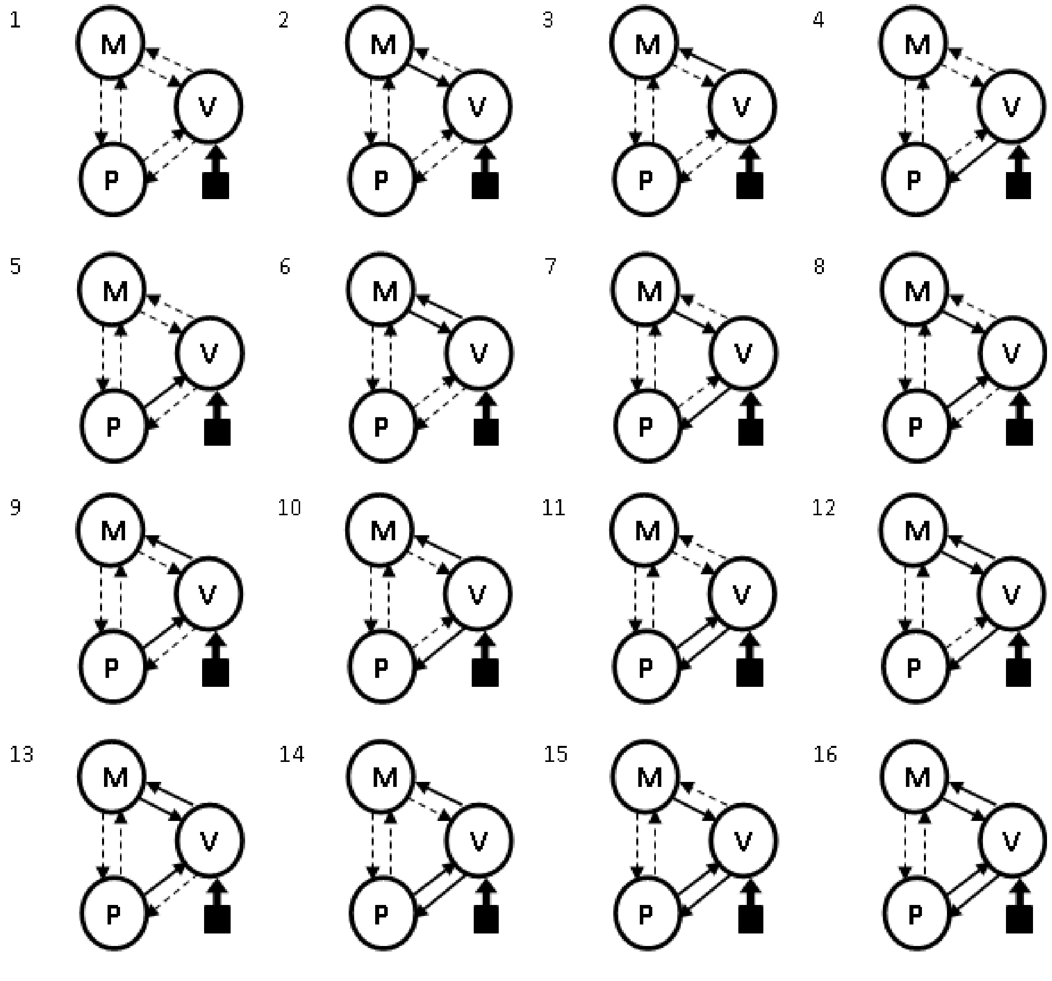

Figure 3.26: *Different models used for the Simon task. All DCMs are fully connected between all three regions (dotted lines). The input to each model is indicated by the black boxes. The 16 models differ in their modulatory connectivity (solid lines).*

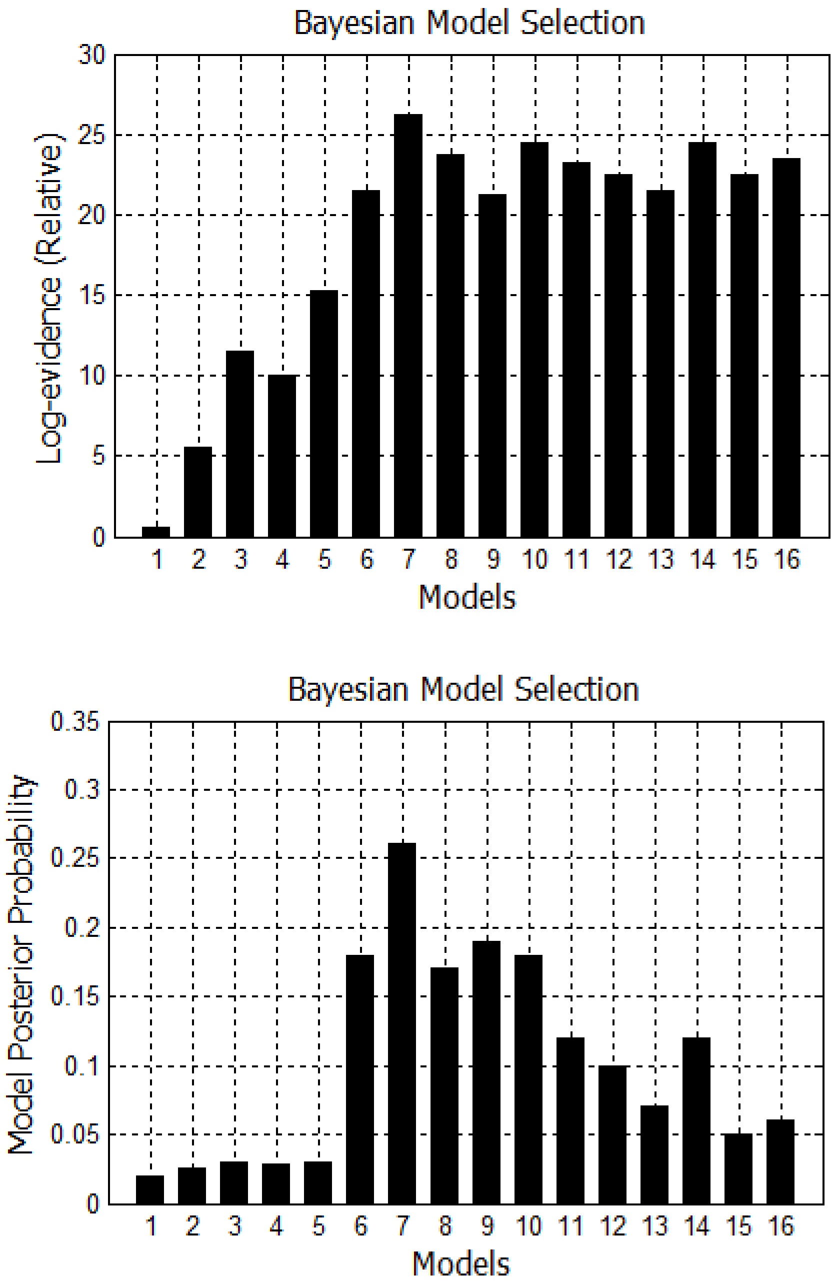

Figure 3.27: *Fixed Effects Analysis for Simon Dataset. Both the relative log-evidence and model posterior probability are highest for the 7th model.*

Table 3.12: *Best model for each subject (Simon Task) using FFX.*

Subject Number	Had Missing Node(s)	Best Model
1	yes	6
2	no	8
3	yes	7
4	yes	9
5	no	9
6	no	7
7	no	8
8	yes	7
9	no	10
10	yes	9
11	no	7
12	no	6
13	yes	10
14	yes	7
15	yes	7
16	yes	7
17	yes	10
18	no	8
19	no	6
20	yes	6
21	no	7

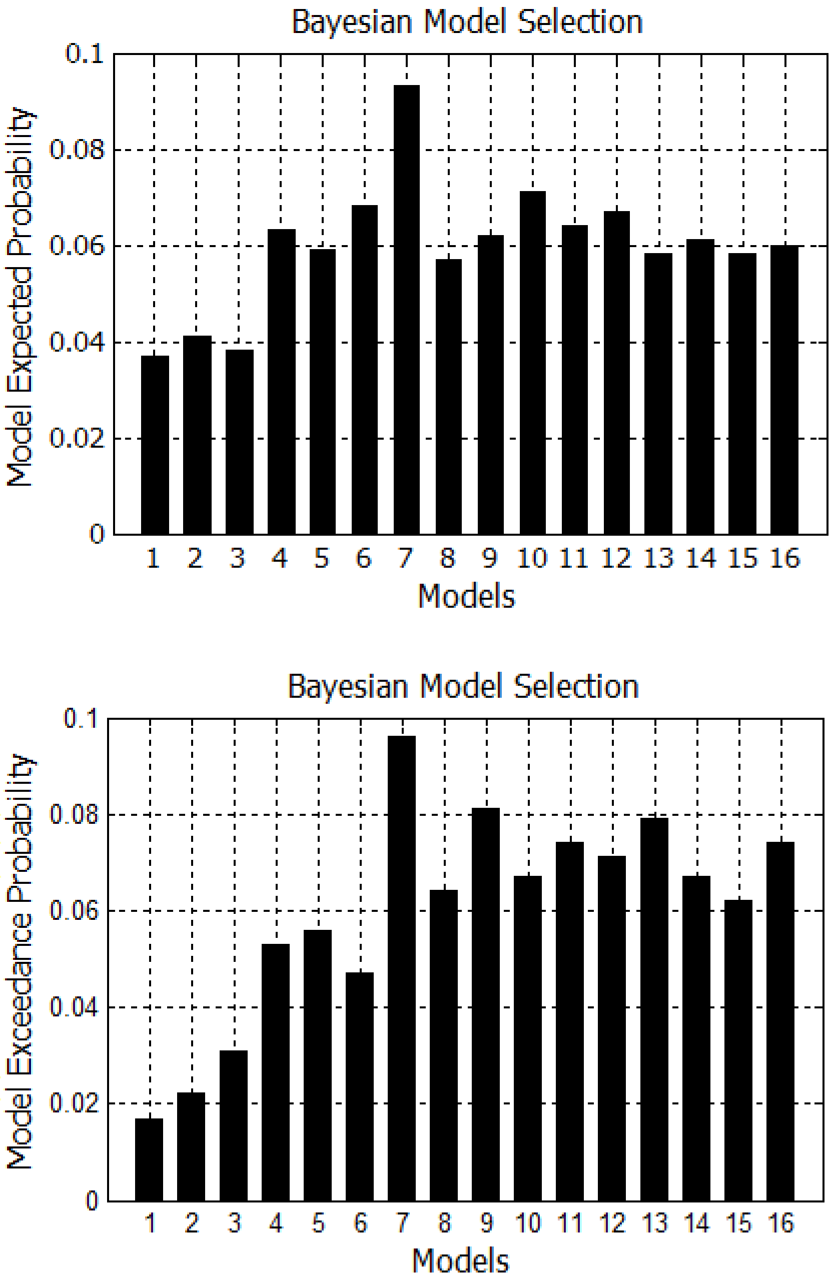

Figure 3.28: *Random Effects Analysis for Simon Dataset. Again both the model expected probability and model exceedance probability are highest for the 7th model.*

Table 3.13: *Best model for each subject (Simon Task) using RFX.*

Subject Number	Had Missing Node(s)	Best Model
1	yes	6
2	no	8
3	yes	7
4	yes	7
5	no	9
6	no	6
7	no	8
8	yes	7
9	no	10
10	yes	9
11	no	8
12	no	6
13	yes	7
14	yes	7
15	yes	7
16	yes	7
17	yes	10
18	no	8
19	no	6
20	yes	6
21	no	7

Table 3.14: *5 models in Occam's window.*

Model 6	$p(m\|Y)$=0.18
Model 7	$p(m\|Y)$=0.26
Model 8	$p(m\|Y)$=0.17
Model 9	$p(m\|Y)$=0.19
Model 10	$p(m\|Y)$=0.18

Model Averaging (BMA) results.

3.5 Reducing FP's using a Conservative P-value

A first level analysis is susceptible to the occurrence of false positives because of the multiple comparisons problem. One method to reduce false positives is to be conservative in the choice of p-value. However, a conservative p-value can lead to another issue. If the p-value is too conservative, the first level analysis can actually miss areas of activation i.e. nodes. This would lead to the missing node problem addressed in this thesis. Following the same logic of needing to compensate for missing nodes, it makes sense to reverse the process causing the missing the nodes in the first place (using a conservative p-value), which is to oppose using a conservative p-value i.e. or using a higher p-value. This technique is not explicitly stated, but appears to be common practice in the DCM community when nodes are missing and they do not want to drop subjects. When an expected node does not appear, the p-value can be increased in stages until activation is produced at or close to the expected location.

It is important to determine how increasing the p-value to locate missing nodes compares to using EM to estimate those missing nodes in real data. We explore whether the evidence is higher (in model comparison) for a missing node that was EM estimated or for that same node to be located by systematically increasing the p-value until a noise point eventually emerges nearby (nearest Euclidean distance). This helps in determining at what point this

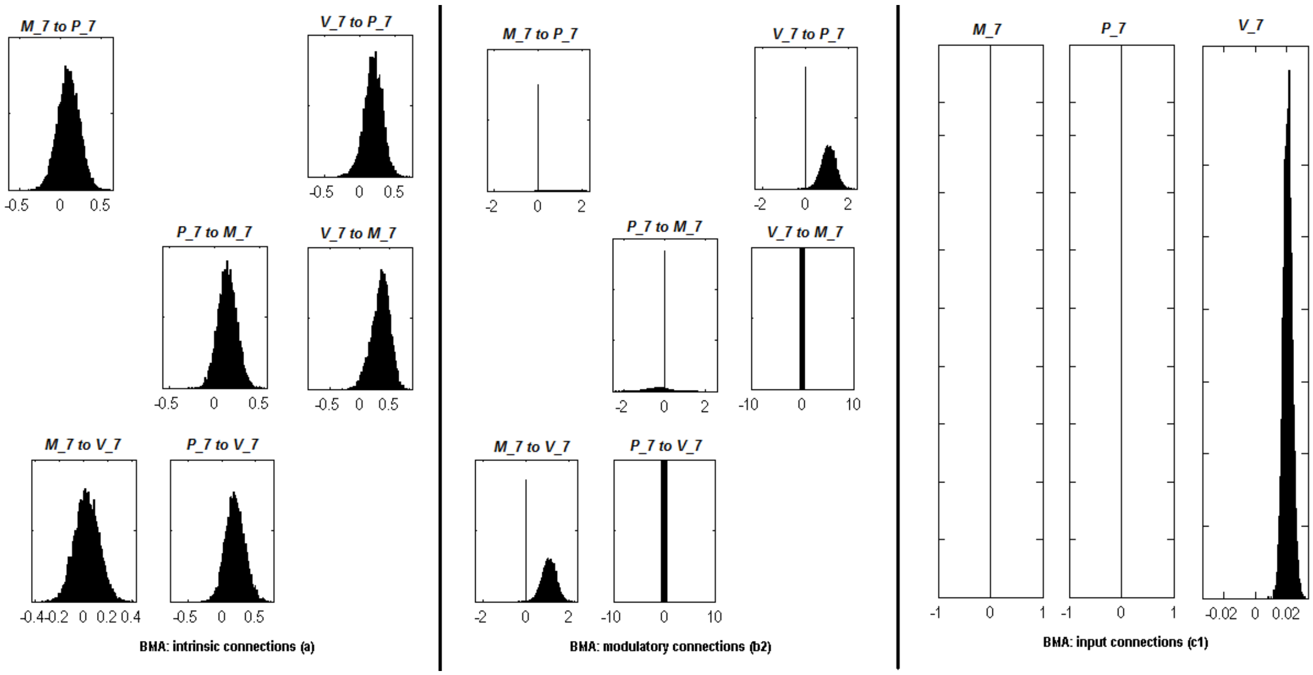

Figure 3.29: *Bayesian model averaging over all 16 models for Simon dataset. The leftmost captured window indicates the intrinsic connectivity relative to the nodes. The middle window identifies the connection links that were modulated, namely V to P and M to V. The rightmost window indicates which node the input is fed through to propagate into the network.*

technique becomes noise.

For the Go/No-Go Task Dataset a p-value of 0.001 was set in the beginning. At this conservative value, 8 out of the 21 total subjects had one or more missing nodes. The p-value was increased in increments of 0.005 up to 0.1 Family-wise corrected. Table 3.15 shows the subjects that had missing nodes and the p-values that were increased until activation appeared in the missing node locations.

Repeating the same experiment shown in Figure 3.23 , but replacing the EM estimated nodes with the nodes obtained after increasing the p-value the evidence is computed for the 16 models shown in Figure 3.21. The Fixed Effects Analysis for Go/No-Go Dataset for Subject 1 is shown in Figure 3.30 where the relative log-evidence and maximum of the posterior (MAP) are plotted. The model posterior probability are highest for the 10th model when using either technique. However, both the relative log-evidence and the model posterior probability for the winning model (model 10) in model comparison are lower when using a higher p-value than when using EM-estimation. Also, the other competing models posterior probabilities are increased when using a higher p-value compared to EM-estimation. This makes the distinction of the best model among other models more subtle and could eventually lead to selection of another model if the noise is high enough.

For Subject 10, the p-value had to be increased to 0.1 for the missing nodes to be located. From the numbers shown in Table 3.10, Subject 10 had the highest evidence for model 10 when using EM. However, when using a high p-value of 0.1, the highest evidence was no longer for the same 10th model. That was also the case for subject 18.

Figure 3.31 shows the variance in evidence for the 16 selected models across the 21 subjects. The 10th model, which was most frequently the winner among the group, has the least variance and highest mean relative log-evidence.

For the Simon task dataset a p-value of 0.001 was set in the beginning. At this conservative value, 11 out of the 21 total subjects had one or more missing nodes. The p-value was increased in increments of 0.005 up to 0.1 Family-wise corrected. Table 3.16 shows the sub-

Table 3.15: *Increased P-values to get missing nodes in Go/No-Go Task Dataset.*

Subject	Missing nodes at p-value 0.001	Highest p-value to force node to emerge
1	Yes	0.025
2	Yes	0.045
3	No	
4	No	
5	Yes	0.05
6	No	
7	No	
8	No	
9	Yes	0.035
10	Yes	0.1
11	No	
12	No	
13	No	
14	No	
15	Yes	0.01
16	No	
17	No	
18	Yes	0.1
19	Yes	0.05
20	No	
21	No	

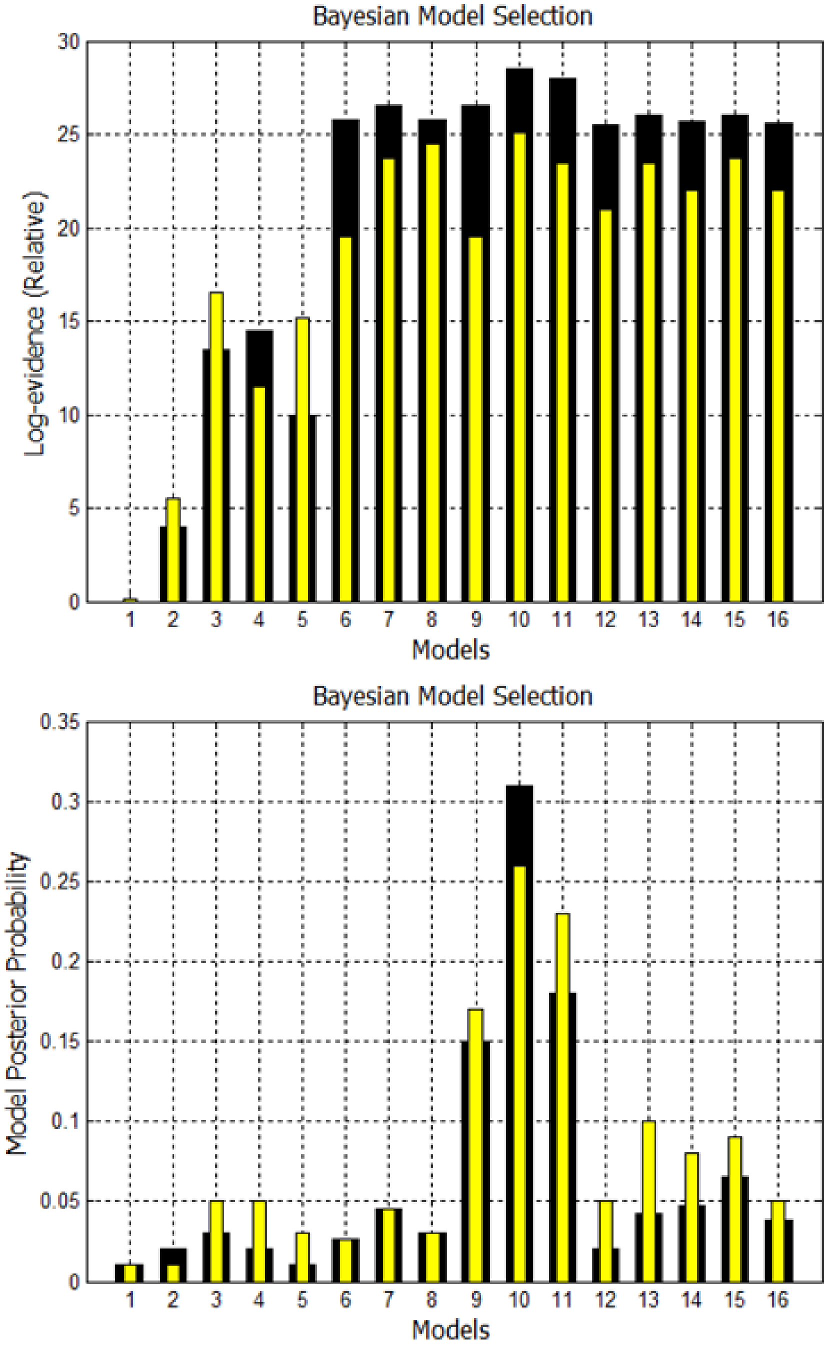

Figure 3.30: *Comparison of FFX Analysis for Go/No-Go Dataset for two techniques. Black bars are for estimation of missing nodes with EM. Yellow bars are for using the higher p-values to get the missing nodes. Model 10 has highest evidence in both cases, but other competing models are getting closer.*

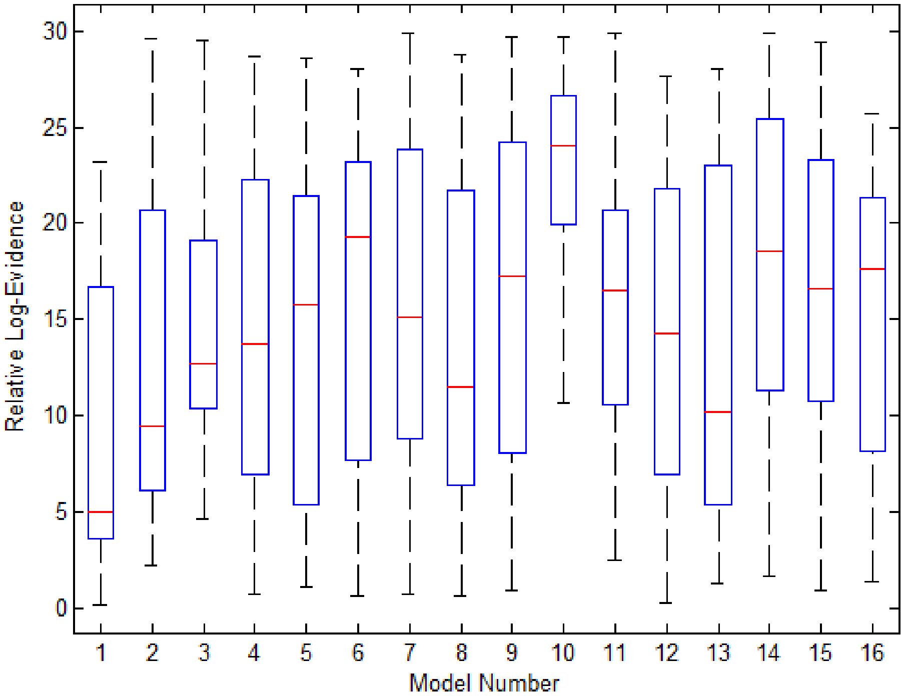

Figure 3.31: *The variance in the 16 models' evidence for the Go/No-Go dataset when using EM for estimation of missing nodes.*

jects that had missing nodes and the p-values that were increased until activation appeared in the missing node locations.

Repeating the same experiment shown in Figure 3.27, but replacing the EM estimated nodes with the nodes obtained after increasing the p-value the evidence is computed for the 16 models shown in Figure 3.26. The Fixed Effects Analysis for Simon Task Dataset for Subject 3 is shown in Figure 3.32. The relative log-evidence and model posterior probability are highest for the 7th model when using either technique. However, both the relative log-evidence and the model posterior probability for the winning model (model 7) in model comparison are lower when using a higher p-value than when using EM-estimation. Also, the other competing models posterior probabilities are increased when using a higher p-value compared to EM-estimation.

For Subject 8, the p-value had to be increased to 0.1 for the missing nodes to be located. From the numbers shown in Table 3.12, Subject 8 had the highest evidence for model 7 when using EM. However, when using a higher p-value of 0.1, the highest evidence was no longer for the same 7th model. That was also the case for subjects 10 and 17. Further discussion of these experiments is in Section 4.1.5.

Table 3.16: *Increased P-values to get missing nodes in Simon Task Dataset.*

Subject	Missing nodes at p-value 0.001	Highest p-value to force node to emerge
1	Yes	0.035
2	No	
3	Yes	0.05
4	Yes	0.02
5	No	
6	No	
7	No	
8	Yes	0.1
9	No	
10	Yes	0.1
11	No	
12	No	
13	Yes	0.01
14	Yes	0.06
15	Yes	0.01
16	Yes	0.075
17	Yes	0.1
18	No	
19	No	
20	Yes	0.05
21	No	

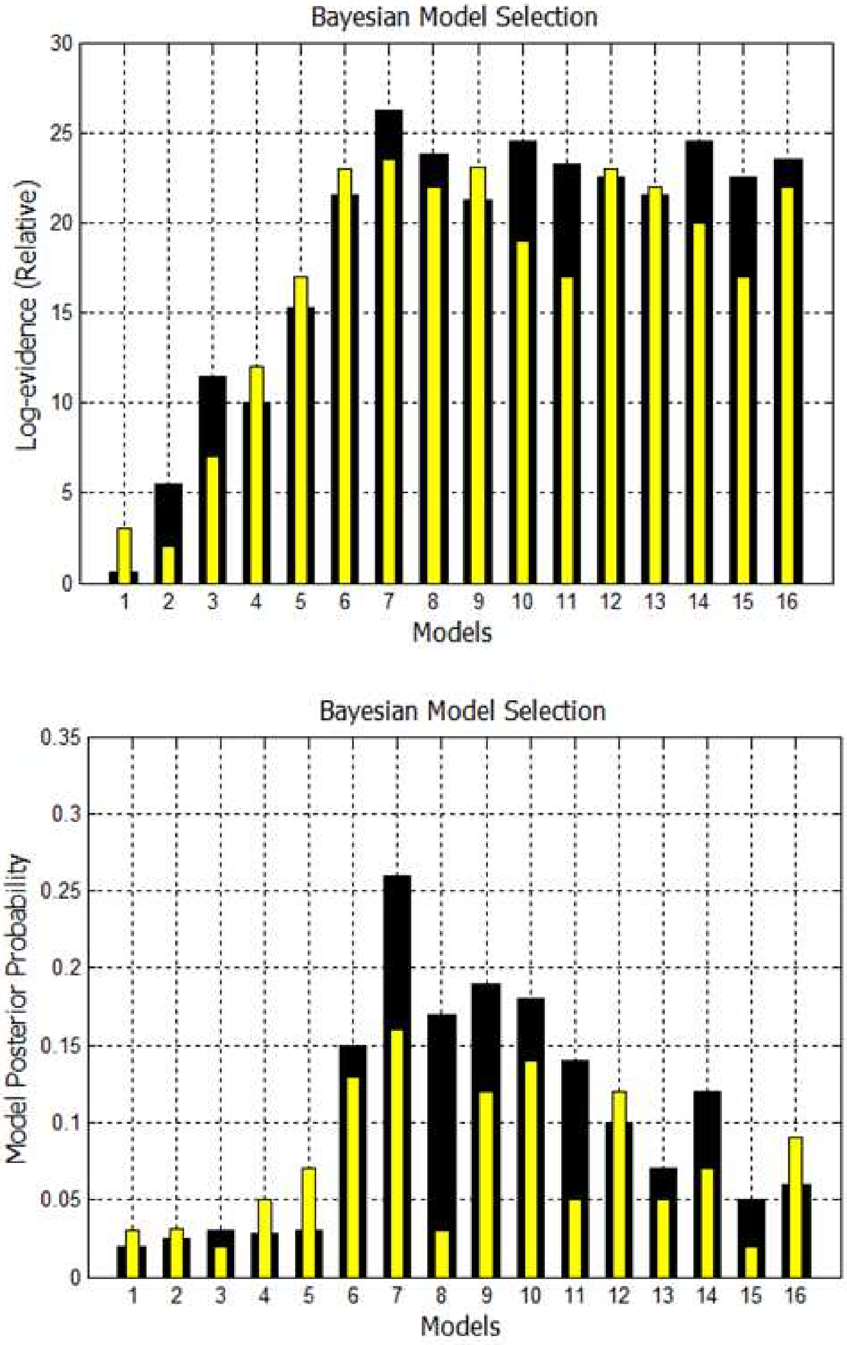

Figure 3.32: *Comparison of FFX Analysis for Simon Task Dataset for two techniques. Black bars are for estimation of missing nodes with EM. Yellow bars are for using the higher p-values to get the missing nodes. Model 7 has the highest evidence in both cases.*

Chapter 4

Discussion and Conclusion

4.1 Discussion

This work presents a solution to overcome the computational issues faced when group studies include variations in network topology. This will improve scientific knowledge by broadening the range of group studies that can be performed by including a higher level of inter-subject variability. By including subjects with differences in topology in addition to the differences in dynamics in the same study, disease-related changes in brain processing can hopefully be understood. A difference in topology might not necessarily be the flag indicating the presence of a disease, but can be studied with regard to diseased and healthy subjects along with other factors causing topological differences. Questions like whether recruiting additional/different neural regions can boost or accelerate neural processing can also be addressed.

Reasons one may be interested in carrying out group studies include, but are not limited to, identifying individual differences and similarities among the members of the group [25]. Group studies usually involve some form of system identification where a set of observations is collected during a task and the goal is to find the parameters of the network that best describes the data. Mathematical models are built in any system identification problem for purposes such as analysis or prediction on groups. Analysis is very important in neuro-science because it attempts to use those models to explain the underlying system. It can also be used to establish structure-function relationships. Other potential benefits to neuro-science of using models in group studies might be in the detection and treatment of disease.

Using models in group studies can also allow predictions regarding the relationship between behavioral and structural changes. In essence, model studies can potentially be shaped for the use in prognosis or in predicting future cognitive functions. They can also be used in the evaluation or personalization of cognitive and pharmacological therapies for individuals based on the group study. So the prospective impact of performing group studies is broad and pertains to both clinical disease and neuro-science.

In current research, there is a limit to the range of subjects that can be incorporated in group studies. By solving the computational problem of performing group analyses on subjects with different topologies, this expands the spectrum of subjects and studies that can be performed in future research. Current research compares families of DCMs by applying Bayesian model averaging within families to provide inferences about parameters that are independent of model structure [15]. All the DCMs used in that work had 3 fully connected regions. Differences included which regions received direct input (affecting entries in the C matrix) and which connections were modulated by factors (affecting entries in the B matrix). The families were grouped according to different input patterns. The focus was also on model selection that best fits the subjects within families, but no reference of any sort was made to including subjects with varying DCM topologies.

[64] also used Bayesian model selection (BMS) for comparing competing models but using a hierarchical method. The only accommodation they made for inter-subject variability was in the exact location of the involved regions, and not in the underlying topology. The connectivity layout of their models was guided by prior anatomical and physiological data as opposed to experimental data and the regional time series were constrained to be located within a certain distance from the group maximum. Subjects that were far off from that

pre-assigned threshold were specifically excluded from the analysis, instead of dealing with topological difference.

Obtaining similar/reproducible fMRI patterns across different subjects or even within the same subject does not necessarily occur. In [83], the inter- and intra-subject reproducibility of fMRI activation was evaluated for three memory encoding tasks. Inter-subject reproducibility was evaluated by examining the percent of subjects who showed activation within a given region of interest and the extent to which individual laterality indices varied from the mean. As for intra-subject reproducibility, that was measured by examining the correlation between a session and another, the average difference between the two sessions, and concordance ratios of both the complete volume and the overlap of the activation volumes. Considerable differences in fMRI reproducibility were noted to occur and those differences appeared to be mainly dependent on the task being performed. Factors such as health and handedness were also noted to affect the degree of inter-subject reproducibility. Considerable intra-subject variability in fMRI activation volume has also been shown in several other studies for a variety of tasks [84][85][86].

Several studies [23] [17] [24] [25] use Bayesian model selection in their group studies neglecting any reference to including DCMs with different regional structure. Although not outwardly stated, all appear to only use the data that yielded activation in the same regions not taking into consideration variations in network topology. The fixed effects approach was used to assign a model to be used by all members of the group. A drawback of this approach is that it does not account for between-subject variability and can thus make the resulting inferences over-confident. It is not robust to the presence of outliers since it does not account for the inclusion of topological differences.

4.1.1 Discussion of FP/FN Simulations

When it comes to the first level statistical analysis in fMRI images, various factors can affect the patterns of brain activations. Experimental analysis has shown that varying simple factors can lead to major differences in the output. Whether errors such as FP's or FN's, differences in cognitive processing, subject-specific related factors, or a combination of all, are the cause of activation differences, the activation areas will never be consistent across an entire group of different subjects. Even a subject performing the same task under the same experimental setup and conditions a second round through can produce variations in brain activation patterns. Correction using various methods can give slight assurance about the presence of a node, but still does not solve the problem of missing or extra nodes.

4.1.2 Discussion of Simulated Go/No-Go

Most available fMRI datasets range in subject counts from tens to hundreds. Simulated fMRI data is much easier to obtain in abundance and the time consumed in data collection can be drastically reduced. Also, varying experimental factors can be done more efficiently and the data can be resynthesized, recollected, and reanalyzed. However, the problem of using simulated data is that since one can not confidently determine the cause behind node variation, the simulated models might not necessarily follow the same natural paths of real data, and may contain certain experimental biases as a result of imposing certain properties in the simulation. Thus the simulated dataset was only used as a preliminary test of the proposed solution to missing nodes in DCM. The simulations have shown a significant improvement in subject classification when using the EM algorithm as a method to estimate missing nodes. The ability to perform subject classification presents an advantage of the solution presented in this thesis in the light of previous work where such types of analyses were avoided.

Model selection based on the highest evidence is subject to several factors such as the number of models being compared, how similar each of the models is to the other models being considered in the same comparison, and the signal-noise ratio of the data. When performing model selection, reducing the number of models being compared can improve the accuracy of model selection. Also, including models that are less similar (based on lower evidence) can improve the accuracy of model selection. In this context, model selection was based on Bayesian Model Comparison which is a relative measure. Since the complete model space can be rather large, model comparison including the entire model space would be impractical.

4.1.3 Discussion of Real Datasets

The ability to obtain results similar to the simulation experiments in the real data experiments is very much desired. However, since classification experiments by nature need the ground truth to be predetermined, and such information is not available, unsupervised methods such as clustering are the only option. Clustering allows the formation of certain groupings and the properties/parameters of the clusters can be used to study the effect of missing node estimation.

4.1.4 Discussion of Ranking Experiments

The ability to perform ranking experiments is one of the main goals of the presented missing node estimation method(s). As described in Table 1.3, computing the evidence is a necessary step to the ranking of different models. Previously published results have avoided the ranking of subjects unless the node structure was the same and variations only existed in the intrinsic connectivity or modulatory effects. The ability to compute a numerical value for the evidence is shown by obtaining physical values in these experiments, and moving a step forward, to sorting these values descendingly, also known as model ranking.

The technique to determine which anatomical regions are involved in active nodes has been based on the literature. Instead of basing the choice of region selection strictly on prior knowledge, the ability to perform Bayesian Model Averaging can be another source of determining the involved anatomical regions. This could incorporate prior knowledge to experimental results, possibly improving the power of such analyses.

4.1.5 Discussion of Reducing FP's using a Conservative P-value Experiments

Although using a less conservative p-value appears to be a reasonable solution to the missing node problem, it likely introduces non-task related information to the data. The previous experiments have shown that both the relative log-evidence and the model posterior probability for the winning model in model comparison are lower when using a higher p-value than when using EM-estimation. This direct comparison suggests that the introduction of unnecessary information can affect the computation of the evidence in a model comparison problem. Moreover, competing models of posterior probabilities are actually shown to increase when using a higher p-value compared to EM-estimation which could eventually lead to selection of an incorrect model. For this reason we would recommend using EM estimation as opposed to a relaxed p-value to replace missing nodes in fMRI-DCM.

4.2 Conclusions

This thesis presents a solution to the problem of missing nodes in fMRI DCM subject groups contributing to the capacity of individual analyses. It was first shown that this problem is valid using a real dataset and that the occurrence of missing nodes is demonstrable.

Based on sufficient evidence in the literature, the problem of missing nodes/regions in group studies is quite common. Experiments performed in this thesis showed different statistics across the subjects in the number and size of active clusters when producing the SPMs for a real dataset. Both volume and cluster analyses for subjects in terms of appearing and disappearing voxels and clusters showed the existence of the varying network node structure. This is evidence of FP's and FN's and inevitably leads to the network size problem within a group analysis hindering the computation of the model evidence.

The main contribution of this thesis was the introduction of missing data approaches as a prior step in dynamic causal modeling and dealing with the inconsistency in network structure, specifically the number of involved nodes. Missing data approaches have handled differences in node topology allowing computation of the model evidence of individuals in group studies. This approach is particularly useful when it comes to group studies, as the previously required manual filtering of the useless subjects and discarding of subjects who did not follow a given DCM model is no longer needed. Previous DCM work has resorted to the approach of discarding subjects with missing regions and discarding extra regions in some of the subjects. As an alternative to rejecting subjects and regions, this thesis has presented a solution(s) to the problem of missing data (nodes) when modeling DCM networks in fMRI group analyses. A direct effect on clinical use, as presented, is the ability to classify individuals or patients. The limitation of not being able to test a certain subject in the context of classifying individual subjects' DCMs is no longer a setback. This increases the practical and clinical impact of a given study.

Simulated experiments were designed and implemented to imitate a real dataset where the multiple topology problem existed. Solutions to the missing data problem included the simple zero-substitution and mean-substitution methods, and the more complex EM-estimation method. Given prior set ground truth, the highest classification rate was obtained when using the EM algorithm as opposed to the other methods. The computational complexity of EM, however is much greater, adding to the overall analysis run time (approximately 1000 times longer). This is considered trivial however as it is only fractions of a second and resulted in a better classification rate.

Feasibility was shown first using simulated subjects to model the variability in network size. Then the various missing data approaches were investigated. To assess the efficiency of the selected algorithms, classification tests were designed and carried out. These classification tests showed added ability in individual classification/labeling.

For practical purposes, the proposed usage of methods for data estimation in the context of DCM, real datasets were then used to verify effectiveness of the solution. The rationale behind using two real datasets is to show that simulation results can be supported by real evidence. Comparable results were indeed achieved and this was considered additional support for the validity of the solution to the problem.

Similar patterns recurred for the two datasets justifying the efficiency of using missing data estimation methods as a means of forcing topological correspondence. This enabled certain errors as first level statistical analysis errors to be overcome and subjects with different topologies due to various reasons to be included within a single group analysis.

The usage of missing data approaches, to estimate missing nodes in DCM, also showed a direct effect on the model ranking. The ability to now include all subjects increased the model ranking capacity by far, when compared to previously not being able to include all subjects. The ability to do model comparison is an important benefit in both individual and group analyses. Model comparison was not even possible when the model and subject being compared had different topologies. It was shown that using missing data approaches on the model selection process and the eventual ranking of models to subjects had quite a positive effect. This was a vital contribution to individual analyses allowing the computation of the complete evidence matrix rather than having to assume no evidence or hindering the computation of evidence all together. We recommend that EM-estimation be used rather than a relaxed statistical criteria to correct for missing nodes.

It was also shown that the sensitivity of the model ranking is dependent on the population size. This sensitivity decreases as the population size increases. Moreover, it was shown that certain changes in model ranking are bound to occur as a result of having to drop those subjects not conforming to the correct network topology. It was concluded that the usage of EM to fill in the missing nodes somewhat restores the original model ranking and has some positive impact on the model selection and ranking processes. This was all first shown using simulated subjects and models followed by experiments that included real subjects.

Both fixed effects and random effects analyses were carried out on the real datasets. These experiments using real data showed feasibility in using the missing data approaches as a preprocessing method allowing the computation of evidence and the model posterior probabilities for all subjects and all models. Model selection and ranking were both feasible for a complete group of subjects and models without having to discard any subjects (specifically those who did not follow a predefined topology). This is considered a valuable advantage to DCM studies. In addition, Bayesian model averaging is also possible since the inclusion of subjects with missing nodes is now acceptable.

4.3 Future Work

The current solution to the missing nodes in fMRI brain activation maps is to compensate for missing nodes by estimation methods using other available data. This forces topological correspondence regardless of the reasons behind the differences which do not necessarily have to be due to first level statistical analysis errors. A direct extension to this work would be to be able to incorporate subjects having missing/extra nodes due to other factors such as disease cases. The models might have to be different in terms of node structure due to the physiological nature of the disease. A question that arises is whether estimating an actually non-existing node could lead to an improper classification (or diagnosis) and how this could be dealt with in the shadow of estimating missing nodes in DCM.

Bibliography

[1] P. R. Montague, R. J. Dolan, K. J. Friston, and P. Dayan, "Computational psychiatry.," *Trends in cognitive sciences*, vol. 16, pp. 72–80, Jan. 2012.

[2] G. Miller, "Psychiatry. Beyond DSM: seeking a brain-based classification of mental illness.," *Science (New York, N.Y.)*, vol. 327, p. 1437, Mar. 2010.

[3] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems.," *Nature reviews. Neuroscience*, vol. 10, pp. 186–98, Mar. 2009.

[4] K. Friston, L. Harrison, and W. Penny, "Dynamic causal modeling," *NeuroImage*, vol. 19, no. 4, pp. 1273–1302, 2003.

[5] J. Daunizeau, O. David, and K. Stephan, "Dynamic causal modelling: A critical review of the biophysical and statistical foundations," *Neuroimage*, vol. 58, no. 2, pp. 312–322, 2011.

[6] C. Bennett and M. Miller, "How reliable are the results from functional magnetic resonance imaging?," *Annals of the New York Academy of Sciences*, vol. 1191, pp. 133–55, Mar. 2010.

[7] J. D. Van Horn, S. T. Grafton, and M. B. Miller, "Individual Variability in Brain Activity: A Nuisance or an Opportunity?," *Brain imaging and behavior*, vol. 2, pp. 327–334, Dec. 2008.

[8] R. Cabeza, “Hemispheric asymmetry reduction in older adults: The harold model,” *Psychology and Aging*, vol. 17, no. 1, p. 85100, 2002.

[9] G. Wagner, K. Koch, C. Schachtzabel, C. Schultz, C. Gaser, J. Reichenbach, H. Sauer, K. Bar, and R. Schlosser, “Structural basis of the fronto-thalamic dysconnectivity in schizophrenia: A combined dcm-vbm study,” *Neuroimage: Clinical*, vol. 3, pp. 95–105, 2013.

[10] K. Friston, “Functional and effective connectivity in neuroimaging: A synthesis,” *Human Brain Mapping*, vol. 2, pp. 56–78, 1994.

[11] B. Siesjo, “Brain energy metabolism,” *New York: Wiley*, p. 612, 1978.

[12] K. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann, “Neurophysiological investigation of the basis of the fmri signal,” *Nature*, vol. 412, no. 5843, pp. 150–157, 2001.

[13] M. Schulz, W. Chau, S. Graham, A. McIntosh, B. Ross, R. Ishii, and C. Pantev, “An integrative megfmri study of the primary somatosensory cortex using cross-modal correspondence analysis,” *NeuroImage*, vol. 22, no. 1, pp. 120–133, 2004.

[14] M. Czisch, R. Wehrle, A. Stiegler, H. Peters, K. Anadrade, F. Holsboer, and P. Samann, “Acoustic oddball during nrem sleep: A combined eeg/fmri study,” *PLoS ONE*, vol. 4, no. 8, 2009.

[15] W. Penny, K. Stephan, J. Daunizeau, M. R. K. Friston, T. Schofield, and A. Leff, “Comparing families of dynamic causal models,” *PLoS Computational Biology*, vol. 6, no. 3, p. e1000709, 2010.

[16] R. Duda, P. Hart, and D. Stork, *Pattern Classification Second Edition*. 2000.

[17] K. Stephan, W. Penny, J. Daunizeau, R. Moran, and K. Friston, “Bayesian model selection for group studies,” *NeuroImage*, vol. 46, p. 10041017, 2009.

[18] R. Cabeza, S. Daselaar, F. Dolcos, S. Prince, M. Budde, and L. Nyberg, "Task-independent and task-specific age effects on brain activity during working memory, visual attention and episodic retrieval," *Cerebral Cortex*, vol. 14, no. 4, p. 364375, 2004.

[19] W. Penny, K. Stephan, A. Mechelli, and K. Friston, "Comparing dynamic causal models," *NeuroImage*, vol. 22, pp. 1157–1172, 2004.

[20] R. Kass and A. Raftery, "Bayes factors and model uncertainty," *Technical Report*, no. 254, 1993.

[21] R. Kass and A. Raftery, "Bayes factors," *Journal of the American Statistical AssociationJournal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[22] A. Raftery, "Bayesian model selection in social research," *Sociological Methodology*, vol. 25, pp. 111–163, 1995.

[23] M. Rosaa, S. Bestmann, L. Harrison, and W. Penny, "Bayesian model selection maps for group studies," *Neuroimage*, vol. 1, no. 49, p. 217224, 2010.

[24] W. Penny, K. Stephan, A. Mechelli, and K. Friston, "Comparing dynamic causal models," *PLoS Computional Biology*, vol. 6, no. 3, pp. 1157–1172, 2003.

[25] W. Penny and A. Holmes, "Random-effects analysis," *Human Brain Interaction*, p. 843850, 2003.

[26] J. Bushberg, J. Seibert, E. Leidholdt, and J. Boone, *The Essential Physics of Medical Imaging Section edition.* 2002.

[27] T. Obata, "Discrepancies between bold and flow dynamics in primary and supplementary motor areas: application of the balloon model to the interpretation of bold transients," *NeuroImage*, vol. 21, no. 1, pp. 144–153, 2004.

[28] R. Buxton, “Introduction to functional magnetic resonance imaging, principles and techniques,” *Cambridge University Press*, 2009.

[29] H. Abdi and L. Williams, “Priniciple component analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, pp. 433–459, 2010.

[30] P. Comon, “Independent component analysis: a new concept?,” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[31] K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, “Statistical parametric mapping: The analysis of functional brain images first edition,” *Elsevier Ltd*, 2007.

[32] L. Harrison, W. Penny, and K. Friston, “Multivariate autoregressive modeling of fmri time series,” *Neuroimage*, vol. 19, pp. 1477–1491, 2003.

[33] S. Bressler and A. Seth, “Wiener-granger causality: A well established methodology,” *NeuroImage*, 2010.

[34] J. Pearl, *Causality: Models, Reasoning, and Inference*. 2000.

[35] A. McIntosh and F. Gonzalez-Lima, “Structural equation modelling and its application to network analysis in functional brain imaging,” *Human Brain Mapping*, vol. 2, pp. 2–22, 1994.

[36] D. Lashkari, E. Vul, N. Kanwisher, and P. Golland, “Discovering structure in the space of fmri selectivity profiles,” *NeuroImage*, vol. 50, no. 3, pp. 1085–1098, 2010.

[37] S. Smith, K. Miller, G. Salimi-Khorshidi, M. Webster, C. Beckmann, T. Nichols, J. Ramsey, and M. Woolrich, “Network modelling methods for fmri,” *NeuroImage*, vol. 54, no. 2, pp. 875–891, 2011.

[38] M. Greicus, K. Supekar, V. Menon, and R. Dougherty, “Resting-state functional connectivity reflects structural connectivity in the default mode network,” *Cerebral Cortex*, vol. 19, no. 1, pp. 72–78, 2009.

[39] D. Bihan, J. Mangin, C. Poupon, C. Clark, S. Pappata, N. Molko, and H. Chabriat, "Diffusion tensor imaging: Concepts and applications," *Journal of Magnetic Resonance Imaging*, vol. 13, pp. 534–546, 2001.

[40] R. Cabeza, J. Locantore, and N. Anderson, "Lateralization of prefrontal activity during episodic memory retrieval: Evidence for the production-monitoring hypothesis," *Journal of Cognitive Neuroscience*, vol. 15, no. 2, p. 249259, 2003.

[41] R. Cabeza, C. Grady, L. Nyberg, A. McIntosh, E. Tulving, S. Kapur, J. Jennings, S. Houle, and F. Craik, "Age-related differences in neural activity during memory encoding and retrieval: A positron emission tomography study," *The Journal of Neuroscience*, vol. 17, no. 1, p. 391400, 1997.

[42] P. Reuter-Lorenz, J. Jonides, E. Smith, A. Hartley, A. Miller, C. Marshuetz, and R. Koeppe, "Age differences in the frontal lateralization of verbal and spatial working memory revealed by pet," *Journal of Cognitive Neuroscience*, vol. 12, no. 1, pp. 174–187, 2000.

[43] C. Grady, "Age related differences in face processing: A meta-analysis of three functional neuro imaging experiments," *Canadian Journal of Experimental Psychology*, vol. 56, no. 3, pp. 208–220, 2002.

[44] D. Madden, T. Turkington, J. Provenzale, L. Denny, T. Hawk, L. Gottlob, and R. Coleman, "Adult age differences in the functional neuroanatomy of verbal recognition memory," *Human Brain Mapping*, vol. 7, no. 2, pp. 115–135, 1999.

[45] M. Stevens, K. Kiehl, G. Pearlson, and V. Calhoun, "Brain network dynamics during error commission," *Human Brain Mapping*, vol. 30, p. 2437, 2009.

[46] M. C. Stevens, K. A. Kiehl, G. D. Pearlson, and V. D. Calhoun, "Functional neural networks underlying response inhibition in adolescents and adults," *Behavioural Brain Research*, vol. 181, no. 1, pp. 12–22, 2007.

[47] J. Netz, T. Lammers, and V. Homberg, "Reorganization of motor output in the non-affected hemisphere after stroke," *Brain*, vol. 120, no. 9, p. 15791586, 1997.

[48] Y. Cao, E. Vikingstad, K. George, A. Johnson, and K. Welch, "Cortical language activation in stroke patients recovering from aphasia with functional mri," *Stroke*, vol. 30, no. 11, pp. 2331–2340, 1999.

[49] C. Thomas, E. Altenmuller, G. Marckmann, J. Kahrs, and J. Dichgans, "Language processing in aphasia: changes in lateralization patterns during recovery reflect cerebral plasticity in adults," *Electroencephalography and clinical Neurophysiology*, vol. 102, pp. 86–97, 1997.

[50] S. Levitan and J. Reggia, "Interhemispheric effects on map organization following simulated cortical lesions," *Artificial Intelligence in Medicine*, vol. 17, no. 1, pp. 59–85, 1999.

[51] J. Frost, J. Binder, J. Springer, T. Hammeke, P. Bellgowan, S. Rao, and R. Cox, "Language processing is strongly left lateralized in both sexes evidence from functional mri," *Brain*, vol. 122, no. 2, p. 199208, 1999.

[52] http://www.fmrib.ox.ac.uk/fsl/.

[53] http://afni.nimh.nih.gov/.

[54] http://www.brainvoyager.com/.

[55] http://www.fil.ion.ucl.ac.uk/spm/.

[56] K. Friston, "Statistical parametric mapping," *Functional neuroimaging: Technical foundations.*

[57] A. Mechelli, C. Price, U. Noppeney, and K. Friston, "A dynamic causal modeling study on category effects: Bottomup or topdown mediation?," *Journal of Cognitive Neuroscience*, vol. 15, no. 7, pp. 925–934, 2003.

[58] R. Buxton, E. Wong, and L. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: The balloon model," *MRM*, vol. 39, pp. 855–864, 1998.

[59] S. Zeki and S. Shipp, "The functional logic of cortical connections," *Nature*, vol. 335, pp. 440–442, 1988.

[60] M. Greicus, B. Karsnow, A. Reiss, and V. Menon, "Functional connectivity in he resting brain: a network analysis of the default mode hypothesis," *Proc. Natl. Acad. Sci*, vol. 100, pp. 253–528, 2002.

[61] K. Friston, A. Mechelli, RTurner, and C. Price, "Nonlinear responses in fmri: The balloon model, volterra kernels and other hemodynamics," *NeuroImage*, vol. 12, no. 4, pp. 466–477, 2000.

[62] T. Deneux and O. Faugeras, "Using nonlinear models in fmri data analysis: model selection and activation detection," *NeuroImage*, vol. 32, no. 4, pp. 1669–1689, 2006.

[63] J. Pedro, G. Laencina, J. Sancho-Gomez, and A. Vidal, "Pattern classification with missing data: a review," *Neural Comput and Applic*, vol. 19, pp. 263–282, 2010.

[64] H. E. M. den Ouden, J. Daunizeau, J. Roiser, K. J. Friston, and K. E. Stephan, "Striatal prediction error modulates cortical coupling.," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 30, pp. 3210–9, Mar. 2010.

[65] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky, "Bayesian model averaging: A tutorial," *Statistical Science*, vol. 14, no. 4, p. 382417, 1999.

[66] D. Glaser, W. Penny, R. Henson, M. Rugg, and K. Friston, "Correcting for non-sphericity in imaging data using classical and bayesian approaches," *Neuroimage*, vol. 13, no. 6, 2001.

[67] R. Little and D. Rubin, *Statistical analysis with missing data.* 1987.

[68] J. Schafer, "Multiple imputation: a primer," *Stat Methods Med Res*, vol. 8, no. 1, pp. 3–15, 1999.

[69] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain," *Neuroimage*, vol. 15, no. 1, p. 273289, 2002.

[70] http://www.fmrib.ox.ac.uk/analysis/netsim/.

[71] D. Handwerker, J. Ollinger, and M. D'Esposito, "Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *Neuroimage*, vol. 21, pp. 1639–1651, 2004.

[72] C. Chang, M. Thomason, and G. Glover, "Mapping and correction of vascular hemodynamic latency in the bold signal.," *Neuroimage*, vol. 43, pp. 90–102, 2008.

[73] http://www.openfmri.org/dspages/ds000007.html.

[74] K. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and bayesian inference in neuroimaging: Theory," *NeuroImage*, vol. 16, pp. 465–483, 2002.

[75] http://www.openfmri.org/dspages/ds000101.html.

[76] A. Leff, T. Schofield, K. Stephan, J. Crinion, and K. Friston, "The cortical dynamics of intelligible speech," *Neuroscience*, vol. 28, pp. 13209–13215, 2008.

[77] M. Beal, Z. Ghahramani, J. Bernardo, M. Bayarri, J. Berger, and A. Dawid, "The variational bayesian em algorithms for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics 7*, 2003.

[78] http://www.mathworks.com/.

[79] G. Xue, A. Aron, and R. Poldrack, “Common neural substrates for inhibition of spoken and manual responses,” *Cerebral Cortex*, vol. 18, pp. 1923–1932, 2008.

[80] http://www.brainmap.org/index.html.

[81] B. Peterson, M. Kane, G. Alexander, M. Lacadie, P. Skudlarski, C. Leung, J. May, and C. Gore, “An event-related functional mri study comparing interference effects in the simon and stroop tasks,” *Cognitive Brain Research*, vol. 13, pp. 427–440, 2002.

[82] X. Liu, M. Banich, B. Jacobson, and J. Tanabe, “Common and distinct neural substrates of attentional control in an integrated simon and spatial stroop task as assessed by event-related fmri,” *NeuroImage*, vol. 22, pp. 1097–1106, 2004.

[83] G. Harrington, S. Farias, M. Buonocore, and A. Yonelinas, “The intersubject and intrasubject reproducibility of fmri activation during three encoding tasks: implications for clinical applications,” *Functional Neuroradiology*, vol. 48, p. 495505, 2006.

[84] W. Machielsen, S. Rombouts, F. Barkhof, P. Scheltens, and M. Witter, “Fmri of visual encoding: reproducibility of activation,” *Human Brain Mapping*, vol. 9, no. 3, p. 156164, 2000.

[85] A. Miki, J. Raz, T. van Erp, C. Liu, J. Haselgrove, and G. Liu, “Reproducibility of visual activation in functional mr imaging and effects of postprocessing,” *AJNR*, vol. 21, no. 5, p. 910915, 2000.

[86] G. Fernandez, K. Specht, S. Weis, I. Tendolkar, M. Reuber, and J. Fell, “Intrasubject reproducibility of presurgical language lateralization and mapping using fmri,” *Neurology*, vol. 60, no. 6, p. 969975, 2003.

Appendix

Variational Bayes (VB)

A VB framework can be used to obtain the posterior distributions of the unknown parameters and latent variables. Let $\theta = \{A, C_1, ..., C_j, D, Q, R, B\}$ represent the unknown parameters and $S = \{s(t), t = 1, 2, ..., T\}$ be the latent variables of the model. Given the observations $Y = \{y(t), t = 1, 2, ..., T\}$ and the probabilistic model, the Bayesian approach aims to find the joint posterior $p(S, \theta|Y)$. The problem here is that obtaining the posterior distribution using a fully Bayesian approach has no analytical solution for most models. Thus an analytical approximation of $p(S, \theta|Y)$ is made in VB. Let $q(S, \theta|Y)$ be an arbitrary probability distribution, then the log of the marginal distribution of the observations Y can be written as

$$logP(Y) = L(q) + KL(q||p) \tag{1}$$

where

$$L(q) = \int dSd\theta q(S, \theta|Y) log \frac{p(Y, S, \theta)}{q(S, \theta|Y)} \tag{2}$$

$$KL(q||p) = -\int dSd\theta q(S, \theta|Y) log \frac{p(S, \theta|Y)}{q(S, \theta|Y)} \tag{3}$$

$KL(q||p)$ is the Kullback-Leibler divergence between $q(S, \theta|Y)$ and $p(S, \theta|Y)$. $KL(q||p) \geq 0$ if and only if, $q(S, \theta|Y) = p(S, \theta|Y)$. Thus, $L(q)$ represents the lower bound on the log of the evidence ($logP(Y)$). The maximum of this lower bound occurs when KL divergence is zero when the optimal choice of $q(S, \theta|Y)$ is $p(S, \theta|Y)$. Since $p(S, \theta|Y)$ is not tractable, certain assumptions on the form of $q(S, \theta|Y)$ must be made and then the optimal distribution is found by maximizing the lower bound $L(q)$. Such assumptions could be that the posterior distribution $q(S, \theta|Y)$ factorizes over S and θ such that

$$q(S, \theta|Y) = q_s(S|Y)q_\theta(\theta|Y) \tag{4}$$

These quantities are obtained by taking functional derivatives of $L(q)$ with respect to $q_s(S|Y)$ and $q_\theta(\theta|Y)$. It can be shown that

$$logq_s(S|Y) \propto E_\theta(logp(Y, S, \theta)) \tag{5}$$

$$logq_s(\theta|Y) \propto E_S(logp(Y, S, \theta)) \tag{6}$$

Equation 5 is the VB- Expectation step and Equation 6 is the VB-Maximization step. Expectations are computed with respect to $q_\theta(\theta|Y)$ in Equation 5 and with respect to $q_S(S|Y)$ in Equation 6. In the VB-Expectation step, the distribution of the latent signal $s(t)$, for each t, is updated given the current distribution of the parameters θ. The latent signal $s(t)$ has a Gaussian distribution and in the expectation step updating the distribution amounts to updating the mean and variance of the Gaussian distribution. Thus, in the VB-E step, estimating the means of $s(t)$ at every t is equivalent to estimating the latent signals. In the VB-Maximization step, the distributions for model parameters θ are updated given the update distributions for latent signal $s(t)$. These VB-E and VB-M steps are repeated until convergence. The details of the derivation of the posterior probabilities are described as follows.

VB-Expectation step

In this step, the posterior distributions of latent variables $q_S(S|Y)$ estimated given the current posterior probability of model parameters $q_\theta(\theta|Y)$. The posteriors of the embedded latent signals x(t) are then computed from which the posterior of s(t) can be obtained. The distribution over these latent variables is obtained using a sequential algorithm where the point estimates of the parameters are replaced by their expectations of the type $E(ZWZ')$ where Z is some parameter of the model and W is a matrix. These expectations are computationally expensive for high order models, so an approximation of $E(AWA') = E(A)WE(A')$ is used to give qualitatively similar results but is computationally more efficient. The mean and covariance of $x(t)$ are given by x_t^T and $\sum_t^T$ respectively.

VB-Maximization step

In this step, the posterior distributions of the model parameter $q_\theta(\theta|Y)$ are estimated given the current posterior probability of the latent variables $q_S(S|Y)$. The joint posterior distribution of parameters $q_\theta(\theta|Y)$ further factorizes as

$$q_\theta(\theta|Y) = q(A, C_1, ..., C_j, D, Q)q(B, R) \tag{7}$$

The state and noise covariance matrices (Q and R) are assumed to be diagonal. Thus, the distribution of the elements in the rows of $A, C_1, ..., C_j, D\&B$ can be inferred separately. Given the state equation for the m^{th} node

$$s_m(t) = (a_m + \sum_{j=1}^{J} \nu_j(t)c_{j,m}s_m(t-1) + d_m u(t) + w_m(t)), w_m(t) \sim N(0, \beta_m) \tag{8}$$

where a_m and $c_{j,m}$ are the m-th rows of A and C_j respectively and $\beta_m = \frac{1}{Q(m,m)}$. In terms of the embedded signal $x(t)$, the above equation can be written as:

$$s_m(t) = \theta'_m[F(t)x(t); u(t)] + w_m(t) \tag{9}$$

where $\theta'_m = [a_m, c_{1,m}, ..., c_{J,m}, d_m]F(t) = [l_M\nu_1(t)I_M...\nu_J(t)I_M]'F$.

The following Gaussian-Gamma conjugate priors are assumed for θ_m and β_m

$$p(\theta_m, \beta_m|\alpha) \sim N(0, (\beta_m A_\alpha)^{-1}Ga(a_0, b_0) \tag{10}$$

Here $\alpha = [\alpha_1, \alpha_2, ..., \alpha_{2M+1}]$ are the hyper-priors on each element of θ_m and $A_\alpha = diag(\alpha)$. Let the prior on α be

$$p(\alpha) = \prod_{i=1}^{2M+1} Ga(c_0, d_0) \tag{11}$$

By applying Equation 6, the joint posterior for θ_m and β_m is given by

$$q(\theta_m, \beta_m|Y) = N(\overline{\theta_m}, \beta_m^{-1}\sum_m)G_a(a_{m,N}, b_{mN}) \tag{12}$$

where

$$A = \begin{pmatrix} \sum_{t=2}^T F(t)P(t-1)F(t)' & \sum_{t=2}^T F(t)x_{t-1}^T u(t) \\ \sum_{t=2}^T x_{t-1}^T u(t)F(t)' & \sum_{t=2}^T u(t)^2 \end{pmatrix} \tag{13}$$

$$\overline{\theta_m} = \sum_m \begin{pmatrix} \sum_{t=2}^T F(t)E(s_m(t)x(t-1)) \\ \sum_{t=2}^T P_s(t, t-1)F(t)' \end{pmatrix} \tag{14}$$

$$a_{m,M} = a_0 + \frac{T+2M}{2} \tag{15}$$

$$b_{m,M} = b_0 + 0.5\left(\sum_{t=2}^T E(s_m^2(t)) - \overline{\theta'_m}\sum_m^{-1}\overline{\theta_m}\right) \tag{16}$$

The posterior for hyper parameters α is given by

$$q(\alpha|Y) = \prod_{i=1}^{2M+1} G_a(c_N, d_{N_i}) \tag{17}$$

where

$$c_N = c_0 + \frac{1}{2} \tag{18}$$

$$d_{N_i} = d_0 + \frac{1}{2}\left(\overline{\theta}_m^2(i)\frac{a_{m,N}}{b_{m,N}} + \sum_m(i,i)\right) \tag{19}$$

The posteriors for θ_m, β_m and α are estimated for each m= 1, 2,..., M from which the posteriors for $A, C_1, ..., C_j, D\&Q$ are computed.

The posterior distribution for the model parameters in the output equation is similarly computed. Since R is assumed to be diagonal, the observation equation in the m^{th} node is given by

$$y_m(t) = b_m\phi x_m(t) + e_m(t), e_m(t) \sim N(0, \lambda_m) \tag{20}$$

where $\lambda_m = \frac{1}{R(m,m)}$. Again assuming Gaussian-Gamma conjugate priors for b_m and λ_m

$$p(b_m, \lambda_m|\alpha) \sim N(0, (\lambda_m A_\alpha)^{-1})Ga(a_0, b_0) \tag{21}$$

Where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_p]$ are the hyper priors on each element of b_m and $A_\alpha = diag(\alpha)$. Let the prior on α be

$$p(\alpha) = \prod_{i=1}^{p} Ga(a_0, b_0) \tag{22}$$

By applying Equation 6, the joint posterior for b_m and λ_m is given by

$$q(b_m, \lambda_m|Y) = N(\overline{b_m}, \lambda_m^{-1}V_m)Ga(a_{m,N}, b_{mN}) \tag{23}$$

$$V_m^{-1} = \phi \sum_{t=1}^{T} P_m(t)\phi' + \overline{A_\alpha} \tag{24}$$

$$\overline{b}_m = V_m\phi \sum_{t=1}^{T} y_m(t)x_t^T(m) \tag{25}$$

$$a_{m,N} = a_0 + \frac{T+P-1}{2} \tag{26}$$

$$b_{m,N} = b_0 + 0.5\left(\sum_{t=2}^{T} E(y_m^2(t)) - \overline{b'}_m V_m^{-1}\overline{b}_m\right) \tag{27}$$

$$c_N = c_0 + \frac{1}{2} \tag{28}$$

$$d_{Ni} = d_0 + \frac{1}{2}\left(\overline{b_m^2}(i)\frac{a_{m,N}}{b_{m,N}} + V_m(i,i)\right) \tag{29}$$

$$\overline{A_\alpha}(i,i) = \frac{C_N}{d_{Ni}} \tag{30}$$

The posteriors for b_m, λ_m and α are estimated for each $m = 1, 2, ..., M$. The VB-E and VB-M steps are repeated until convergence.