Reliability of Dynamic Causal Modeling using the Statistical Parametric Mapping Toolbox

Pegah T. Hosseini[*1], Shouyan Wang[2], Julie Brinton[3], Steven Bell[1], David M. Simpson[1]
[1]*Institute of Sound and Vibration Research, University of Southampton, Southampton, UK*
[2]*Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou, China*
[3]*Auditory Implant Service, University of Southampton, Southampton, UK*

ABSTRACT
Dynamic causal modeling (DCM) is a recently developed approach for effective connectivity measurement in the brain. It has attracted considerable attention in recent years and quite widespread use to investigate brain connectivity in response to different tasks as well as auditory, visual, and somatosensory stimulation. This method uses complex algorithms, and currently the only implementation available is the Statistical Parametric Mapping (SPM8) toolbox with functionality for use on EEG and fMRI. The objective of the current work is to test the robustness of the toolbox when applied to EEG, by comparing results obtained from various versions of the software and operating systems when using identical datasets. Contrary to expectations, it was found that estimated connectivities were not consistent between different operating systems, the version of SPM8, or the version of MATLAB being used. The exact cause of this problem is not clear, but may relate to the high number of parameters in the model. Caution is thus recommended when interpreting the results of DCM estimated with the SPM8 software.

Keywords: Dynamic Causal Modeling, Statistical Parametric Mapping, Effective Connectivity, Robustness, MATLAB

INTRODUCTION
The human brain is regarded as an ensemble of dynamic systems. Communication between neural centers is of the utmost importance in executing a mental task and many different cortical and subcortical brain areas have to coordinate for an individual to respond effectively to sensory input. Brain organization can be quantified by connectivity measures which aim to provide measures of the strength, direction, and timing information on the connections between brain areas.

Brain connectivity is sometimes broken down into three categories: structural, functional, and effective connectivity (Daunizeau et al. 2011). Structural connectivity describes the anatomical structure of the brain and the pattern of anatomical links inside the brain (Daunizeau et al. 2011). Functional and effective connectivity measurements of the brain have become very popular topics in the field of neuro-science and neuro-engineering in recent years. Functional connectivity (Chan et al. 2013; Jalili & Knyazeva, 2011; Lachaux et al. 1999) can indicate if there is a relationship between the activities of two different brain areas but it cannot reveal the

direction of the connection (information flow).The latter is referred to as the causality in the connections and can be probed through effective connectivity analyses which has attracted much attention in recent research (Baccala & Sameshima, 2001; David et al. 2006; Sakkalis, 2011).

Generally, causality refers to the relation of cause and effect. In other words, causality indicates which event/signal is the consequence of the other. It can be quantified according to different statistical measures such as Granger Causality (Granger, 1969), Partial Directed Coherence (Sameshima & Baccalá, 1999), and Directed Transfer Function (Korzeniewska et al. 2003) which all fit a linear multivariate autoregressive model to the EEG data. Nonlinear extensions of this approach, such as nonlinear Granger Causality (Ancona et al. 2004; Freiwald et al. 1999), have also been introduced allowing for a wider range of functional connectivities. The conventional linear methods have the advantage of parsimony, are generally easier to implement, need fewer assumptions, and thus are likely to be more robust than non-linear methods but can only approximate biological systems which are never entirely linear. It should also be noted that none of the above methods (either linear or non-linear ones) explicitly take prior knowledge of structural or functional connectivity into account. For extended reviews on functional and effective connectivity measurements the reader is referred to other publications (Friston, 2011; Gourévitch et al. 2006; Pereda et al. 2005).

In contrast to the above approaches, Dynamic Causal Modeling (DCM) was developed on a neurobiological basis (David et al. 2006; Friston et al. 2003). It is a biologically informed model which in principle gives it many advantages over other connectivity methods and has become popular among researchers in recent years. Whilst DCM has potential advantages over other models, a possible weakness in the approach is its large number of parameters and initial assumptions which may affect the robustness of the algorithm, reducing the reliability of results. The Statistical Parametric Mapping (SPM8) software is a freely available MATLAB®-based toolbox in which the DCM algorithm has been implemented (http://www.fil.ion.ucl.ac.uk/spm/). However, the robustness of the implementation of DCM does not appear to have been tested extensively.

In order to address this and following preliminary work that indicated potential problems with the toolbox, we tested the consistency of results between different versions of SPM8 and MATLAB® and different operating systems on identical data. If results differ widely, even on just a few examples, this would raise concern as to whether results allow robust inference on functional connectivity. To the best of our knowledge, our work is the only publication testing the robustness of the implementation of DCM in the software SPM8.

In the next sections, a review of related previous work on the reliability of DCM is presented, followed by the methods section which includes a brief explanation of DCM, and materials section which explains the data used and how this was analyzed. The results and their discussion then lead to the conclusions.

RELATED WORK

Since its introduction for EEG analysis, DCM has been used for a range of purposes such as understanding the neural interactions in psychological disorders (Dima et al. 2010; Dima et al. 2009; Kempton et al. 2010), in the vegetative state (Boly et al. 2011), and in response to various types of stimulation (Brown & Friston, 2012; David et al. 2006; Garrido et al. 2009; Marreiros et al. 2010). Precisely because of all the attention it has been attracting, it is very important to test its reliability, considering aspects of within and between subject repeatability, the physiological

plausibility of inferences and also the implementation of the algorithm in the toolbox that all current research uses.

DCM has been available for about a decade now but there are not many critical comments on its validity or reliability other than the ones presented by the developers of the algorithm (Daunizeau et al. 2011; David et al. 2006; Garrido et al. 2007; Kiebel et al. 2007). Using simulated data, it was shown that DCM is sensitive and specific to changes in parameters of the model but not too sensitive to the level of noise (David et al. 2006). It was also demonstrated that DCM can generate realistic evoked responses (Kiebel et al. 2007) and that it produces similar results for different subjects under similar conditions (Garrido et al. 2007). In the latter, the reproducibility of DCM was tested using the data from 13 different subjects, giving consistent (but of course not identical) results. There does not appear to have been much investigation of repeatability of the analysis in the same subjects over multiple sessions.

To the best of our knowledge, the only major questions regarding implementation of DCM have been presented by Lohmann et al. (2012), who argued that comparing different estimated DCM models and selecting the best model according to its fitness value may not be the best approach. They show that even defining a 3-area model can lead to a huge pool of possible models, not all of them plausible. They go on to discuss that it is quite possible that the winning model, even among very carefully defined set of models, may turn out to be an implausible one. This comment started a debate (Breakspear, 2013; Friston et al. 2013) about the credibility of DCM ending with Lohmann et al. (2013) claiming that the counterarguments were not quite convincing.

METHOD

In this section, DCM is explained briefly, based around the historic development and the rationale and underpinning concepts. The method is highly complex, and a detailed description is beyond the scope of this paper. The reader is referred to the paper by David et al. (2006) for more information on EEG-based DCM.

DCM considers a neuronal mass model for the brain regions of interest. The basis of the mass model used for DCM/EEG (David & Friston, 2003; David et al. 2005) is an extension of Jansen's model which was introduced in 1995 (Jansen & Rit, 1995). Jansen introduced a basic mathematical model to generate EEG, or at least a signal with a few characteristics of an EEG, from any arbitrary signal (e.g. a sensory stimulus) as the model input (Jansen & Rit, 1995). In his basic model, he assumed that an area in the model consists of one group of inhibitory and one group of excitatory neurons with the input affecting the excitatory neurons and a feedback loop from these (excitatory) neurons to both excitatory and inhibitory neurons of the same area. He then extended this basic model to a two-area model and showed that this new model can simulate even more characteristics of EEG signals than the one-area model.

This idea was further developed by David et al. (2003). They assumed one brain area to consist of two types of excitatory neurons and one type of inhibitory neurons, and extended the model to more than two brain areas, demonstrating that the new model can cover an even larger range of the brain's biological behaviours than Jansen's model (David et al. 2005) could. Using the connectivity rules which had been obtained in experimental studies, three types of extrinsic connections (between-area connections) were defined according to different layers of the brain as 1) forward or bottom-up that starts in agranular layers and ends in layer 4, 2) backward or top-down which starts in agranular layers and ends in agranular layers, and 3) lateral that starts in agranular layers and ends in both granular and agranular layers (David et al. 2005). State equations of neural states and output equations of the model were then computed according to

these rules, anatomical information in the literature, an action-potential-like input, and EEG as the output (David et al. 2006). The unknown parameters in these equations can include conduction delays, input parameters, as well as intrinsic and extrinsic coupling parameters quantifying the strength of connections within one area of the brain and between two different areas, respectively.

Employing a Bayesian framework, in which the prior is the assumed distribution of the parameters and the posterior is the probability distribution of the measured output, DCM estimates the parameters of the model. It is worth mentioning than the prior hyper-parameters are set according to knowledge about the architecture and behavioral characteristics of the brain's neural networks (David et al. 2006). The parameters are identified iteratively by minimizing the free energy of the system, which can be regarded as the estimation error, using an expectation-maximization (EM) algorithm. In each iteration, first, a posterior distribution is calculated according to the minimum free energy (E-step) and then a new set of parameters are computed according to the updated posterior distribution (M-step). The EM procedure iterates the E and M steps until the free energy reaches a minimum. The estimated conditional moments (mean and covariance) of each parameter and the log-evidence (the probability of the output given a specific model) is used to describe the connectivity and the goodness-of-fit of the model.

The inference on the connectivity can be obtained with two approaches: 1) in a specific model, by considering each connection and the probability of this connection being greater than zero (or a baseline value) thus being responsible for generating this specific output and 2) between models, by considering each with their various sets of parameters and comparing their log-evidences and selecting the best model accordingly. This best model is considered significantly better than other models if its log-evidence is at least 3 units larger than the log-evidence of other models (David et al. 2006). It is worth noting that in DCM, there is no right or wrong model. Any model can be simulated and estimated with DCM but among the models defined and estimated, the best model can be selected according to the 3-unit rule.

DCM is thus a biophysically informed model whose computations result in not only the strength of connections between two separate areas but also estimates of the effect of one area over its own future signal values. As opposed to other techniques with unknown inputs, the couplings in a DCM model are estimated by perturbing the system with designed inputs and measuring the response. DCM accounts for the nonlinearities of the neural system by introducing nonlinear voltage transform functions to the model from which nonlinear state equations can be derived. Moreover, it can be considered as a source reconstruction technique on its own as it implicitly estimates the interactions at source level (i.e. in the brain) instead of sensor level (i.e. for the current application the EEG electrodes) so it does not need an extra source reconstruction step that most other techniques need before measuring the source level connectivity. For all these reasons, DCM has attracted much attention in the last decade and has been used to analyze responses in a range of protocols (Brown & Friston, 2012; David et al. 2006; Dima et al. 2010; Dima et al. 2009; Garrido et al. 2009; Kempton et al. 2010; Marreiros et al. 2010).

The current paper addresses the specific point of the robustness of the software implementation of DCM (SPM8), the toolbox used in (to the best of our knowledge) all recent studies involving DCM. The approach taken is to analyze identical sets of EEG data in different versions of SPM8, different versions of MATLAB® and different operating systems. One would expect results to be identical, with small differences explained by unavoidable numerical errors. Large differences, leading to clearly different inferences, even on only one set of data, would be of concern and question the reliability of the tool.

MATERIALS

In this section, the process of recording and preprocessing the EEG data is explained. The DCM models, their specification in SPM8, and the software versions used for estimating the models are also given.

The DCM method was tested on two sets of data. The first was recorded by our group and is referred to as "recorded data" throughout the paper. These data were obtained during auditory stimulation of normal-hearing subjects in experiments that aimed at elucidating changes in brain connectivity during pure-tone and speech stimulation. Only one recording from a 30 year old male subject is presented here though more data were analyzed (with similar results and inferences). University of Southampton ethical approval was received for this study and the participant consented to take part in the experiment. The second set of data was the EEG from an adult normal hearing subject downloaded from the SPM website and will be referred to as "SPM data" in this paper (accessible at http://www.fil.ion.ucl.ac.uk/spm/data/eeg_mmn).

Stimulus Characteristics

Recorded data: Two pure tones were presented binaurally at 55 dB HL approximately every 2 seconds at random intervals. The tones were 80ms-long, and either 1 kHz (120 stimulus repetitions) or 2 kHz (480 stimulus repetitions) with 5ms rise and fall times. These tones were presented in a random sequence.

SPM data: Two pure tones were presented binaurally every 2 seconds at random intervals. The tones were 70ms-long tones at 2 kHz (120 stimulus repetitions) or 1 kHz (480 stimulus repetitions) with 5ms-long rise and fall times. These tones were presented in a random sequence and the subject was asked to count the 2 kHz tones (see Garrido et al. (2007) for more details).

Data Acquisition

Recorded data: A 66-channel EEG cap with equidistant electrode positions was placed over the head. The reference of the system was the tip of the nose and the ground electrode was placed on the line passing the nose tip and the brain vertex just above the forehead. The subject listened to the randomly played tones with eyes closed while sitting in a comfortable armchair. The data were recorded with a NeuroScan system (Scan 4.3 software and SynAmps^2 amplifier) and at a sampling frequency of 1000Hz.

SPM data: A 128-channel EEG data was recorded with a Biosemi system at a sampling frequency of 512 Hz. Two extra electrodes were used to monitor eye movements.

EEG Pre-processing

Recorded data: EEG was filtered in the 0-30 Hz band and epoched around the onset of the 2 kHz stimulus (200 ms pre- and 500 ms post-stimulus). Data were visually checked and showed the expected evoked potential. Fifteen different sets of data were generated by randomly selecting 240 epochs out of 480 epochs of the 2 kHz stimulus and averaging them. DCM analysis was carried out using this dataset to assess robustness to relatively small changes in the data (within sample variability). Then one of these datasets was used for more extensive analyses using the different combinations of software systems, as well as repeatedly analyzed using the same systems, on the same computer. Electrode positions were co-registered to the template MRI map available in the SPM8 toolbox with nasion, right, and left auricular points being the fiducial points which were defined manually.

SPM data: This data was already preprocessed when it was downloaded from the website. A brief summary of the preprocessing procedure is: EEG was re-referenced, band-pass filtered in

the 0.5-30 Hz frequency band, downsampled to 200Hz, and then epoched around the onset of the auditory stimulus with a -100 ms pre- and a 400 ms post-stimulus window. Trials with amplitudes exceeding the [-80 80] μV range were removed and the remaining trials were averaged. The data already consisted of channel locations (more details can be found in the SPM8 manual accessible from http://www.fil.ion.ucl.ac.uk/spm/doc/).

DCM Analysis

Recorded data: The GUI interface of SPM8 DCM was used and the five different models shown in Figure 1 were defined with all forward, backward, and lateral connections present (David et al. 2006). DCM then calculates the strength of each connection within the models and allows comparison of the "fitness" of the different models. In the models, only Left and Right Primary Auditory Cortices (LA and RA), Left and Right Superior Temporal Gyri (LS and RS), and Left and Right Inferior Frontal Gyri (LI and RI) were included. These areas have been shown to be related to sound perception in the brain (David et al. 2006). The positions of these areas were taken from Garrido et al. (2007) and RI was assigned a symmetrical position to LI with respect to the sagittal line. Furthermore, the input of the system was defined to occur around 40ms after the stimulus onset and to affect both LA and RA. The distributed spatial model was set to Equivalent Current Dipole (ECD) and other parameters of DCM GUI were left as default values set in the toolbox.

SPM data: The model in Figure 2 was defined for this set of data to be consistent with the simulations in the SPM8 manual and the publications based on this data (David et al. 2006; Garrido et al. 2007). A time window from zero to 200ms and only the 1 kHz tone was considered for DCM estimations. Positions of these areas are the same as in Figure 1 and the input of the system was defined to occur around 60ms after the stimulus onset and affect both LA and RA. The distributed spatial model was set to Equivalent Current Dipole (ECD) and other parameters in the DCM GUI were left at default values, as above.

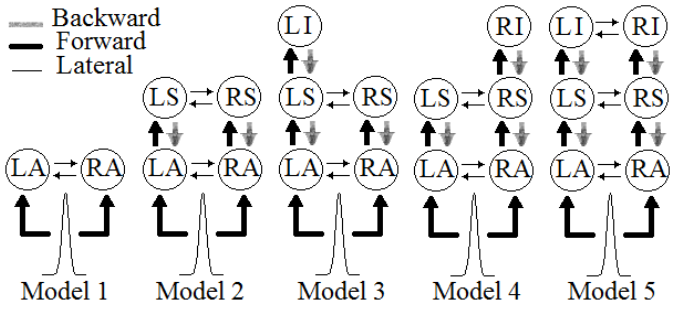

Figure 1. The five different models analyzed in DCM using the recorded data.

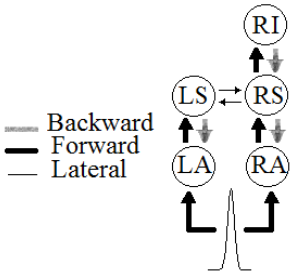

Figure 2. The DCM model studied on the SPM data

Inference from DCM

As was mentioned earlier, there are two approaches for determining the best model after DCM estimation. First, DCM estimates model parameters and reports a probability value for each parameter being greater than zero (or a baseline value). Parameters with probabilities higher than a set value can be considered responsible for the differences between the observed output and the baseline condition, the latter being assumed zero in this report. Here, this probability threshold was set to 90%.

Second, for each estimated model, a "fitness value" is calculated which depends on the goodness-of-fit and the complexity of the model. Among different models estimated for a specific set of data, the significantly better model is defined to be the one with fitness values at least three units larger than other models. It has been argued (K.J. Friston, personal communication) that a single connection should not be dealt with individually to decide if it is or is not responsible for the observed data. It was suggested that two separate models that do or do not contain an individual connection should be compared via their log-evidence and using the 3-unit rule the best model can be selected. If this best model contains a specific connection, this suggests that this connection is responsible for the specific output.

Software systems

Parameter estimations were performed using versions 4290, 4667 and 5236 of SPM8 (denoted as sv.4290, sv.4667, and sv.5236), two versions of MATLAB 64-bit (denoted as mv.2011a and mv.2012a), and three Personal Computers (PC), with two using Windows 7 64-bit and one with Red Hat Enterprise Linux 64-bit as their operating systems. Note that sv.5236 is a newer version of SPM8 than sv.4667, which in turn is newer than sv.4290, and that DCM GUI default values were the same in all the versions of SPM8 used.

RESULTS

In this section, the results of applying DCM on the data introduced in the previous section are presented. While these are not exhaustive results, inconsistencies between different versions of the same software uncovered in just a few cases are sufficient to raise concern regarding the robustness of the toolbox. It should be emphasized that the results presented here are typical of similar cases explored in developing this work.

Reproducibility of DCM

The starting point for the current work comes from exploratory work on the reproducibility of DCM on data from the same subject, and some of these results are presented first. These tests were carried out on the fifteen datasets generated by randomly selecting 240 out of 480 trials of the recorded signals.

Models 1 to 5 of Figure 1 were estimated by DCM for each dataset in the same combination of the software (with sv.4290, mv.2011a, and Windows 7). The fitness values of these models are presented in Figure 3. Model one was consistently the poorest. Using the 3-unit rule, it is clear that in all except 2 datasets (5 and 14), model 5 was significantly better than other models. This result would suggest fairly good reproducibility regarding the choices between models.

However, closer investigation of the estimated strengths of connections of model 5 showed that for each dataset, different connections were held responsible for the output (probability >90%). Thus, even though model 5 was selected as the best model most of the time, not much could be inferred about the individual parameters of the model, raising some concern regarding inferences on connectivity patterns.

Tests on different computer systems to increase speed of processing led to the concern about consistency in analyzing identical datasets with different versions of the software, which will be described in greater detail below.

Reliability of SPM

As was discussed in the Method section, one can derive inferences from the results of DCM using firstly, the confidence intervals and probability of individual parameters or, secondly, comparing a set of models using their fitness values (log-evidence of the model). Both procedures are approached in the following subsections.

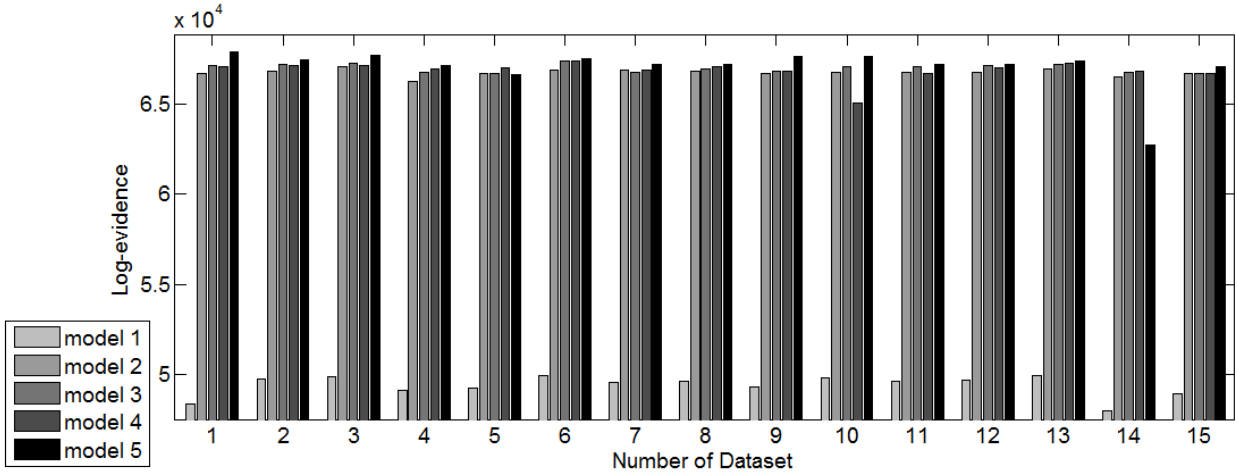

Figure 3. Fitness values of models 1-5 when applied on 15 generated datasets with sv.4290, mv.2011a, and Windows 7.

Individual parameters

Here the first approach is followed, with an assessment of reliability of parameter estimates with repeated analysis of the same models.

Recorded data: In this work models 2 and 5 of Figure 1 (see also Figure 3) were selected for further analyses. These two models were selected because they had an average (model 2) and a large (model 5) log-evidence value compared to others (see Figure 3). The estimation of model parameters was carried out using different versions of SPM8 and MATLAB, as explained in the Materials section, for only one of the fifteen generated datasets. Results are presented in Figure 4, where only the connections with conditional probabilities higher than 90% are plotted. These connections are assumed to affect the EEG most. Figure 4 shows very different connectivity patterns for this dataset when the version of SPM8 or MATLAB or the operating system changes. This discrepancy was observed in both model 2 and model 5. For example in Figure 4.A.i, both lateral connections between LA and RA are responsible connections whereas in Figure 4.A.ii only the connection from RA to LA seems responsible and in Figure 4.A.iii no connection between RA and LA is reported responsible. In another example, Figure 4.B.i shows that the input enters both primary auditory cortices but Figure 4.B.ii indicates that the input enters RA only. On the other hand, Figure 4.B.iii shows that for the same model, no input (stimulation) is responsible for the observed evoked response.

It should be emphasized that all the pre-processing steps taken and all parameters entered the estimation algorithm for different versions of SPM8 were the same, only software versions differed.

In another test, the versions of SPM8 and MATLAB were kept unchanged but the analyses were run on two different PC's with the same operating system (Windows 7 64-bit). In this case the responsible connections did not vary for either of the two models (models 2 and 5). However, when the operating system of one of the PC's changed from Windows 7 to Linux, even with the same SPM8 and MATLAB version, different results were obtained. As an example, the implementation of this condition for model 2 is presented in Figure 5. As before, only connections with probability higher than 90% are plotted and it is clear that the two operating systems produce different results. For example in Windows 7, the connection between LA and RA seems responsible whereas in Linux that connection does not seem to be affecting the output (observed evoked response).

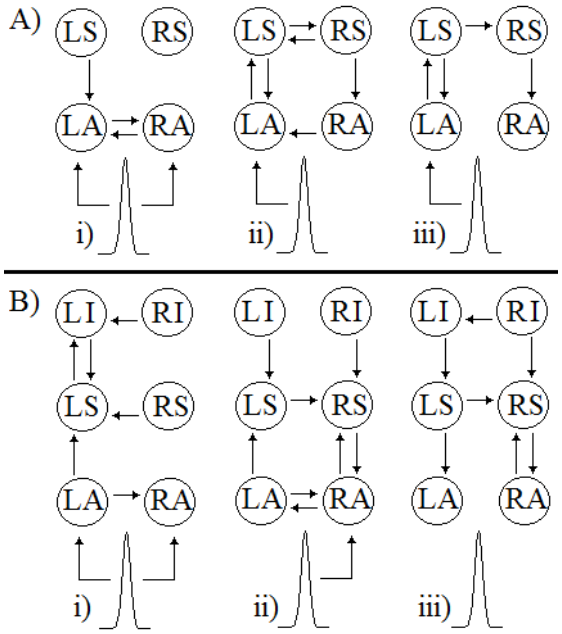

Figure 4. Responsible connections in generating the evoked response (probability >90%) in A) model 2 and B) model 5 for the same set of data. i) mv.2012a & sv.4667, ii) mv.2012a & sv.5236, iii) mv.2011a & sv.5236. A pulse acts as the input to the model.

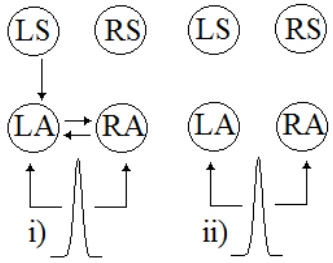

Figure 5. Responsible connections in generating the evoked response (probability >90%) in model 2. In both i and ii, sv.4667 and mv.2011a were used for the same set of data but the operating system was different: i) Windows 7 and ii) Linux. A pulse acts as the input to the model.

SPM data: Similar analyses were performed on the data downloaded from the SPM website. The model presented in Figure 2 was used to initialize the DCM algorithm. To analyze the data with DCM, the SPM8 version and the operating system were kept the same but the MATLAB version was changed. Once again, different responsible connections were obtained for the same model and the same set of data. The results of these analyses can be seen in Figure 6. In this figure, only connections with probability higher than 90% are plotted. For example, Figure 6.i suggests that only the connection from LA to LS is causing the observed evoked response whereas Figure 6.ii indicates that many of the connections in the model are responsible for the observed output.

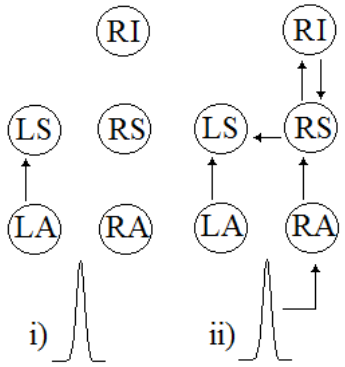

Figure 6. Responsible connections in generating the evoked response in model 4 for SPM data. In both i and ii, sv.4667 and Windows 7 were used but the MATLAB version was different: i) mv.2011a and ii) mv.2012a. A pulse acts as the model input.

It should also be noted that when the same model was estimated more than once in the same combination of software versions and on the same set of data, results were identical. Also, when the software combination (SPM8/MATLAB/operating system) was kept consistent on different computers, the results were again identical, as was observed in the recorded data reported above.

Model comparison

The second approach to the evaluation of DCM was used next, in which different models are compared.

This approach was also tested with the 5 models of Figure 1 on the same set of data (one of the generated datasets) with two combinations of the software. In Figure 7, the log-evidence of these estimated models are presented when the estimation was performed in two different combinations of MATLAB and SPM8 software. Note that in this figure, the scale is not the same, but for both, a 3-unit difference should be considered significant. It is clear that the log-evidence of the estimated models did not vary in a consistent way across software versions. Although model 5 (the most complex model) was identified as the best model in both combinations, the ranking of the remaining models is not consistent. According to Figure 7.i, the model order will be 5 ➔ 4 ➔ 2 ➔ 3 ➔ 1 but in Figure 7.ii, it will be 5 ➔ 3 ➔ 4 ➔ 2➔ 1. Thus, if model 5 is excluded, within the remaining ones, model 4 would be selected in one software combination and model 3 in the other. Therefore it cannot be said if the connection to the right or the left Inferior Frontal Gyrus is responsible for the output EEG if the estimations are performed in these two software combinations. The inferences are thus quite different but in both cases deemed significant (greater than 3-unit difference).

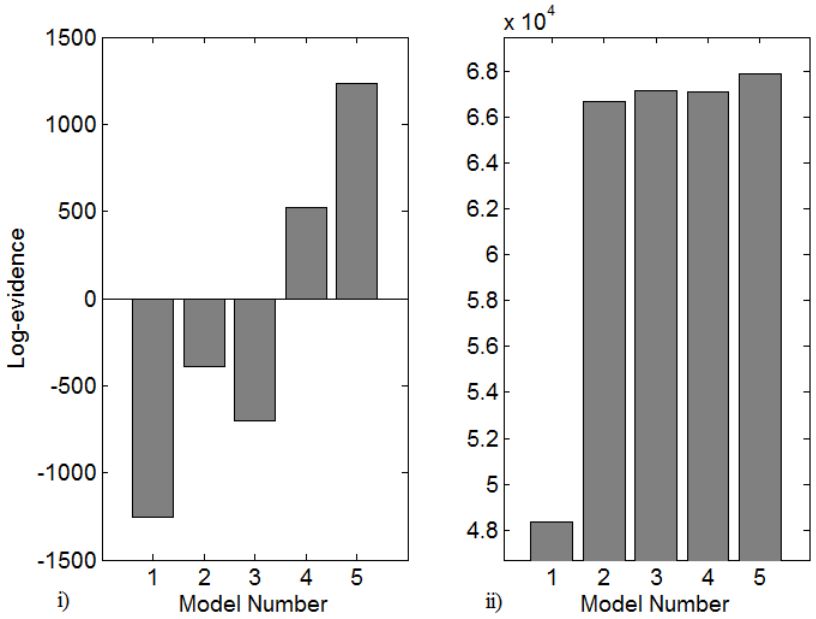

Figure 7. Fitness values of models 1-5 for the same set of data and operating system (Windows 7) but different version of the software: i) mv.2012a & sv.5236 and ii) mv.2011a & sv.4290

DISCUSSION

Estimating DCM for the same set of models over different datasets from the same subject and with the same software systems showed some inconsistencies. The fact that 13 out of 15 datasets reported the same model (model 5) as the best model was both promising and worrying. It was promising because one might expect some variability within a biological sample, but worrying because it may indicate that DCM can give inconsistent results. Discrepant results in 2 out of 15 datasets randomly selected from within the same recording acquired over a period of a few minutes in the same subject would seem rather high. Furthermore, if model 5 is eliminated from the comparison, the inconsistency continues, with almost half of the datasets reporting model 3 as the best model and the remainder favoring model 4. These exploratory results on reproducibility of DCM analysis within the same session were insightful, but highlighted the need for further investigation which should consider multiple subjects, and larger datasets acquired within the same session and then also on different days.

The key result of this study was that DCM/SPM8 showed discrepancies when using different combination of the software (SPM8, MATLAB, and operating system). These discrepancies were observed both in the estimates of the strength of individual connections and in model comparisons using the log-evidence. Within the same model, depending on the software combination used, different connections showed higher probability of affecting the output. Likewise, when comparing models according to fitness values, the order of preference for model selection was not the same for the various combinations of the software, even on identical datasets.

The current results clearly do not prove that connectivity measures derived from DCM are always unreliable. However, the examples presented (from a larger set of results that provide a consistent message) show clear evidence of lack of robustness and raise concern that results can be misleading. Possible reasons behind the variation in results could include different numerical precision of the MATLAB versions used with different operating systems, or perhaps different numbers of computational loops in iterative estimates, or slightly different estimation algorithms

in various versions of SPM8. Though it is not clear what the exact reasons are, it does raise questions regarding the robustness of the DCM algorithm or the SPM8 software implementation when applied to the EEG and caution should be employed in the interpretation of results of DCM using the SPM8 toolbox.

CONCLUSION

As an initial step towards understanding DCM, a reproducibility study was carried out using various model structures and it was observed that DCM may not give reliable results. However, a wide range of other evaluations should be performed before any final decisions can be made.

It is also shown for the first time in this paper that the results of estimating DCM parameters using SPM8 toolbox can vary greatly depending on the operating system or the version of MATLAB or SPM8 being used. This variation occurs both at the level of individual parameters and model comparison using log-evidence. This discrepancy was tested with both auditory evoked potentials recorded in our lab and sample EEG data available on the SPM website. For both datasets, it was observed that the connections reported having high probability of affecting the observed data may differ considerably if the version of MATLAB or SPM8, or the operating system changes. Moreover, depending on the pool of models defined for DCM estimation, different models may be selected as the best models in various software combinations. The reason behind this problem is not yet known, but clearly indicates that DCM/SPM8 should be used cautiously.

ACKNOWLEDGEMENTS

This research is funded by the Institute of Sound and Vibration Research (ISVR), the University of Southampton, and its Auditory Implant Service (USAIS).

REFERENCES

Ancona, N., Marinazzo, D., & Stramaglia, S. (2004). Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E, 70*(5), 056221.

Baccala, L., & Sameshima, K. (2001). Partial directed coherence: a new concept in neural structure determination. *Biological Cybernetics, 84*(6), 463-474.

Boly, M., Garrido, M. I., Gosseries, O., Bruno, M.-A., Boveroux, P., Schnakers, C., et al. (2011). Preserved Feedforward But Impaired Top-Down Processes in the Vegetative State. *Science, 332*(6031), 858-862.

Breakspear, M. (2013). Dynamic and stochastic models of neuroimaging data: A comment on Lohmann et al. *NeuroImage, 75*(0), 270-274.

Brown, H. R., & Friston, K. J. (2012). Dynamic causal modelling of precision and synaptic gain in visual perception - an EEG study. *NeuroImage, 63*(1), 223-231.

Chan, H.-L., Chu, J.-H., Fung, H.-C., Tsai, Y.-T., Meng, L.-F., Huang, C.-C., et al. (2013). Brain connectivity of patients with Alzheimer's disease by coherence and cross mutual information of electroencephalograms during photic stimulation. *Medical Engineering & Physics, 35*(2), 241-252.

Daunizeau, J., David, O., & Stephan, K. E. (2011). Dynamic causal modelling: A critical review of the biophysical and statistical foundations. *NeuroImage, 58*(2), 312-322.

David, O., & Friston, K. J. (2003). A neural mass model for MEG/EEG: coupling and neuronal dynamics. *NeuroImage, 20*(3), 1743-1755.

David, O., Harrison, L., & Friston, K. J. (2005). Modelling event-related responses in the brain. *NeuroImage, 25*(3), 756-770.

David, O., Kiebel, S. J., Harrison, L. M., Mattout, J., Kilner, J. M., & Friston, K. J. (2006). Dynamic causal modeling of evoked responses in EEG and MEG. *NeuroImage, 30*(4), 1255-1272.
Dima, D., Dietrich, D. E., Dillo, W., & Emrich, H. M. (2010). Impaired top-down processes in schizophrenia: A DCM study of ERPs. *NeuroImage, 52*(3), 824-832.
Dima, D., Roiser, J. P., Dietrich, D. E., Bonnemann, C., Lanfermann, H., Emrich, H. M., et al. (2009). Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. *NeuroImage, 46*(4), 1180-1186.
Freiwald, W. A., Valdes, P., Bosch, J., Biscay, R., Jimenez, J. C., Rodriguez, L. M., et al. (1999). Testing non-linearity and directedness of interactions between neural groups in the macaque inferotemporal cortex. *Journal of Neuroscience Methods, 94*(1), 105-119.
Friston, K., Daunizeau, J., & Stephan, K. E. (2013). Model selection and gobbledygook: Response to Lohmann et al. *NeuroImage, 75*(0), 275-278.
Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain Connect, 1*(1), 13-36.
Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage, 19*(4), 1273-1302.
Garrido, M. I., Kilner, J. M., Kiebel, S. J., Stephan, K. E., Baldeweg, T., & Friston, K. J. (2009). Repetition suppression and plasticity in the human brain. *NeuroImage, 48*(1), 269-279.
Garrido, M. I., Kilner, J. M., Kiebel, S. J., Stephan, K. E., & Friston, K. J. (2007). Dynamic causal modelling of evoked potentials: A reproducibility study. *NeuroImage, 36*(3), 571-580.
Gourévitch, B., Bouquin-Jeannès, R., & Faucon, G. (2006). Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biological Cybernetics, 95*(4), 349-369.
Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica, 37*(3), 424-438.
Jalili, M., & Knyazeva, M. G. (2011). EEG-based functional networks in schizophrenia. *Computers in Biology and Medicine, 41*(12), 1178-1186.
Jansen, B., & Rit, V. (1995). Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological Cybernetics, 73*(4), 357-366.
Kempton, M., Dima, D., Roiser, J., Stephan, K., Friston, K., & Frangou, S. (2010). PW01-148 - Effective connectivity within the network of fearful facial affect recognition in patients with bipolar disorder compared to healthy controls. *European Psychiatry, 25*(Supplement 1), 1564-1564.
Kiebel, S. J., Garrido, M. I., & Friston, K. J. (2007). Dynamic causal modelling of evoked responses: The role of intrinsic connections. *NeuroImage, 36*(2), 332-345.
Korzeniewska, A., Mańczak, M., Kamiński, M., Blinowska, K. J., & Kasicki, S. (2003). Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method. *Journal of Neuroscience Methods, 125*(1-2), 195-207.
Lachaux, J.-P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping, 8*(4), 194-208.
Lohmann, G., Erfurth, K., Müller, K., & Turner, R. (2012). Critical comments on dynamic causal modelling. *NeuroImage, 59*(3), 2322-2329.
Lohmann, G., Müller, K., & Turner, R. (2013). Response to commentaries on our paper: Critical comments on dynamic causal modelling. *NeuroImage, 75*(0), 279-281.

Marreiros, A. C., Kiebel, S. J., & Friston, K. J. (2010). A dynamic causal model study of neuronal population dynamics. *NeuroImage, 51*(1), 91-101.
Pereda, E., Quiroga, R. Q., & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology, 77*(1–2), 1-37.
Sakkalis, V. (2011). Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG. *Computers in Biology and Medicine, 41*(12), 1110-1117.
Sameshima, K., & Baccalá, L. A. (1999). Using partial directed coherence to describe neuronal ensemble interactions. *Journal of Neuroscience Methods, 94*(1), 93-103.