Middle East Technical University
Biomedical Engineering Department, Bioelectric

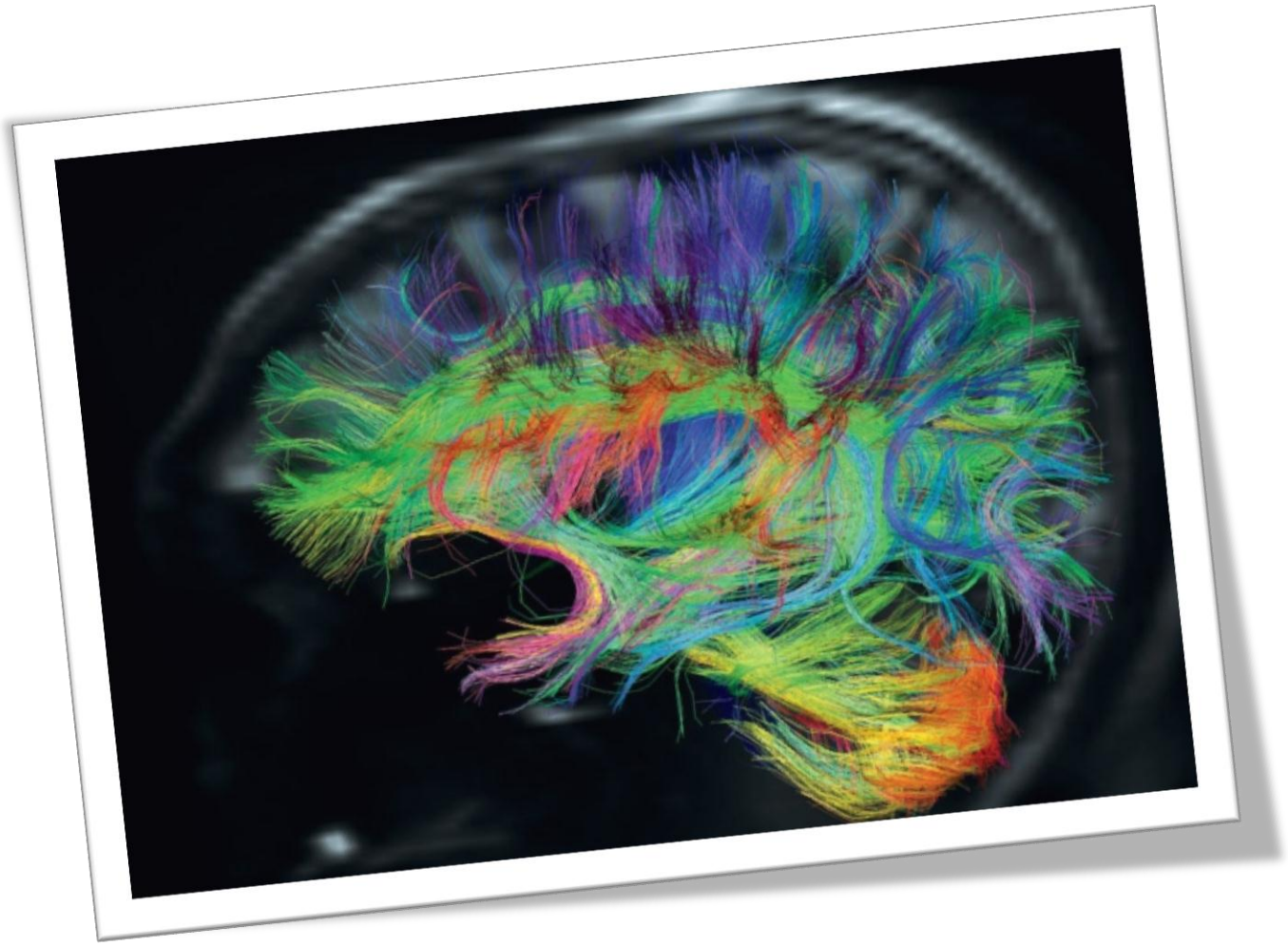
Dynamic Casual Modeling

Report

vol. III

Researcher: Arman Afrasiyabi

Adviser: Assoc. Prof. Ilkay Ulusoy

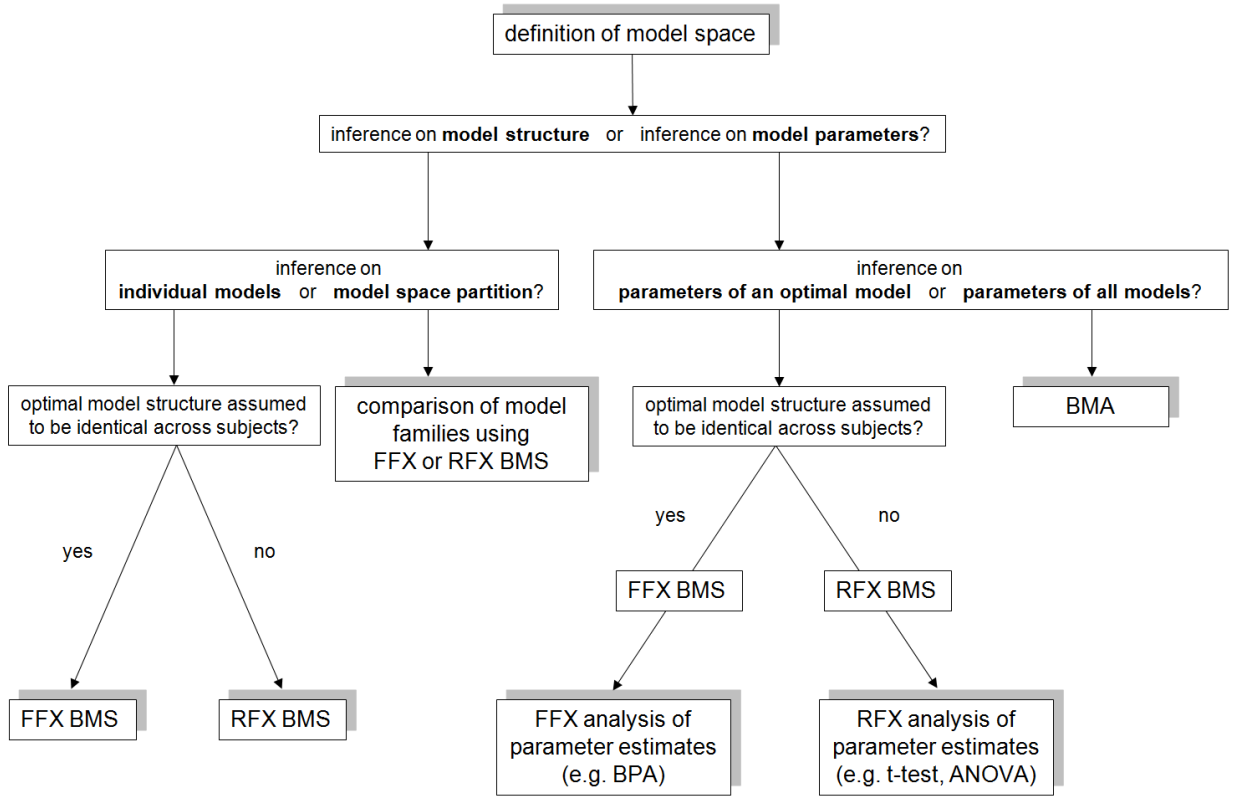


Abstraction

In this report, I will try to explain the methods used in the **model selection** of DCM. First, the recommended flowchart for applying DCM will be covered. Then, I will try to explain the **approximation methods** used in the DCM. Finally, we will see a method called **Bayes Factor** which is used for comparison purpose. In the following Reports, I will try to apply these topics in SPM8 in practical way.

DCM for fMRI: Advance Topics
Bayesian Model Selection

In this section we will try to explain the Bayesian model selection. The flow chart shown in the figure below which is strongly recommended below by **Dr. Stephan** show an approach of DCM analysis. The first step is that you should specify your model space. This must be done in such a way that reader can understand how you constructed your hypothesis. Then, after you finished this step, you have to make a choice whether you interested in expect of model structure or model parameters. If you interested in model structure, for example, is there any connection from "a" to "b"? Whether this is linear or nonlinear model? Whether there is a parallel or serial architecture in system? Then you do not need to be worry about the parameter and all the things that you should be worry have to be the skeleton of model. That is the selecting among the several computing selection models.



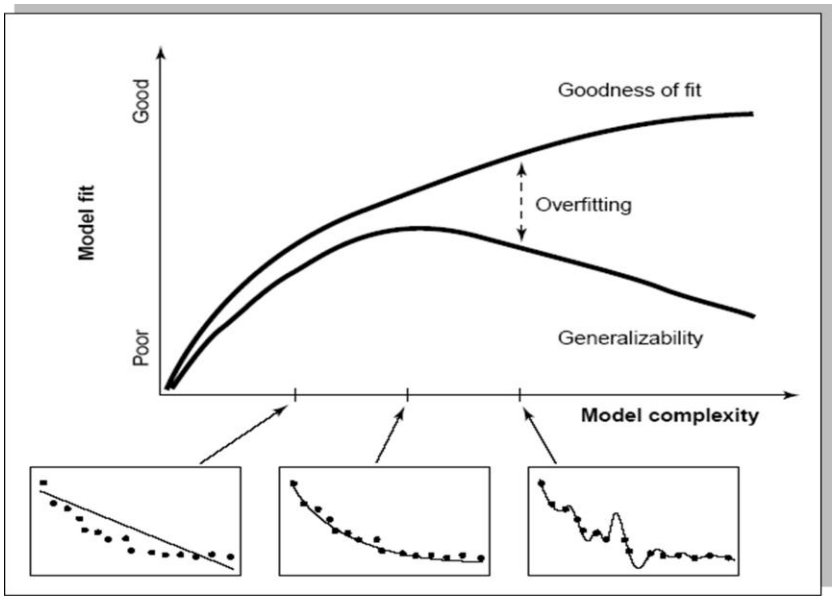
Stephan et al. 2010, *NeuroImage*

If you are interested in parameters, then you still need to model selection first because any parameter estimate is conditional on model that you used. Therefore, you first have to go to model selection step and then choose either fixed effect or random effect inference on parameter.

Model Comparison and Selection

Model comparison is mechanism of choosing the best model given the computing hypothesis on structure and function of system. It is very generic problem in science. In other words, every scientist confront with it regardless of their major. As an engineer, our duty is to decide which of several alternative hypotheses is most likely for given the measurement. This is not a trivial problem, and you cannot simply say that two computing models turns up to measured data and take the one that fit better. The reason is that in the complex models often fit the data better because they have more flexibility and more degree of freedom to explain the data.

As graph below shows, when you increase the complexity of the model and thereby increase the goodness of fit. At some point, however, you will do something that is called over fitting. In other words, you starting to fit the noise in your measurement and your parameter estimates will not be generable across the future measurements of the same process. In this stage, what is really needed is the decision about the selection of the model that represents the best balance between models fit and model complexity.



Pitt & Miyung (2002) *TICS*

Of course there are many ways of that decision making, but DCM uses the Bayesian method to find the best decision. The model which uses this method is called **evidence model**. In evidence model, you try to find the model "m" that maximizes the probability of the observing the data that you have measured given "m" [p(y|m)]. Model evidence (shown below) is the heart of the model selection if your method is Bayesian. It has a very simple expression:

$$\begin{aligned} p(y \mid m) &= \int p(y \mid \theta, m)\, p(\theta \mid m)\, d\theta \\ &= \left\langle \log p(y \mid \theta, m) \right\rangle \\ &\quad - KL\left[q(\theta), p(\theta \mid m)\right] \\ &\quad + KL\left[q(\theta), p(\theta \mid y, m)\right] \end{aligned}$$

The most advantage of this expression is that it accounts for both of accuracy and complexity of the model. Additionally, it allows us to inference about the structure of the model. It is integral over parameters for the product between likelihood and θ. The problem with this model is that it cannot evaluate the cases other than linear Gaussian models. Therefore, you cannot evaluate the complex models.

Approximation models in DCM

In most of the cases, we need an approximation to the expression. Some of the well-known approximation methods are:

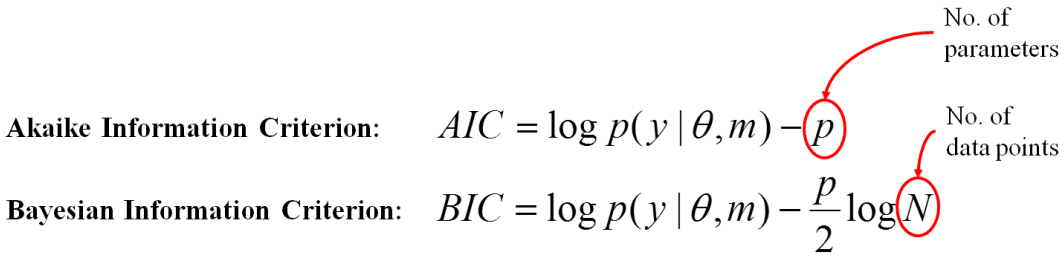
I) Negative Free Energy
II) Akaike Information Criterion (ALC)
III) Bayesian Information Criterion (BLC)

Here, the obvious question that may come to our mind is that why we use Log in the **evidence model**? In fact, we do not use the model evidence, but we use Log model. The reason is very simple. The logarithm is a monotonic function so if we maximize the model evidence. The probabilities are small numbers, so if you represent very very small number, you will get the numerical problem. The Log of a very small number is very large negative number which is very easy to do numerically. That is the reason that we take the Log of the evidence. In fact the **Log model evidence is balance between fit and complexity**

$$\begin{aligned} \log p(y \mid m) &= accuracy(m) - complexity(m) \\ &= \log p(y \mid \theta, m) - complexity(m) \end{aligned}$$

The log evidence can be written very nicely as the difference between accuracy (which is the Log of likelihood of model) and the complexity term. The different approximation that Log evidence is differ in the complexity term that all use the Log likelihood as the accuracy term, but it differs in complexity description.

Two well-known and simple approximation models are AIC and BIC. They were used in the old version of the SPM (SPM5), and they have very simple structure:



In the case of AIC the complexity is the number of parameters. In BIC, the complexity is the number of parameter divided by two times the log of the data points. The disadvantage of these models was that they only taken the account of the number of parameters, and they completely blind about the flexibility of those parameters or the prior variance. Additionally, they blind into potential independent among the parameters. This is the source of limitation for these approximations.

Due to these reasons, the SPM has updated to different approximation called the **negative free energy**. This model seems a little complicated for those who do not have advanced knowledge about the Information Theory. However, it has a very similar structure, and it decompose to accuracy term (Log likelihood) in this case the expected log likelihood under some assume posterior (q). Also, it has a complexity term (KL). For simplicity, we ignore to go in depth of these concepts.

$$\begin{aligned}
&\log p(y \mid m) \\
&\quad = \langle \log p(y \mid \theta, m) \rangle - KL[q(\theta), p(\theta \mid m)] + KL[q(\theta), p(\theta \mid y, m)] \\
&\quad = F + KL[q(\theta), p(\theta \mid y, m)]
\end{aligned}$$

F can be written as the difference between fit and complexity:

$$F = \langle \log p(y \mid \theta, m) \rangle_q - KL[q(\theta), p(\theta \mid m)]$$

Apart from the point that we discussed above, what you should know and remember is that in contrast to AIC and BIC, the advantage of the complexity term in free energy model is that it has components that express how strongly the parameters are inadequate. If you have two parameters of model and you add a parameter to that model that is inversely correlated with existent parameter, then AIC and BIC would penalize that. But, free energy would not do penalization, and it recognizes that new parameter which does not have only explanatory power, and it is not increase or decreases the degree of freedom. This is the reason that SPM8 is using this method.

Bayes Factor

Log evidence can be use to for comparison purpose in the model selection. Using one of the approximation methods that we discussed, you can approximate the log evidence of competing models. The problem is that this method is not very intuitive because you just subtract some odd numbers from another strange number. Due to this problem, people started to use the method called Bayes Factor.

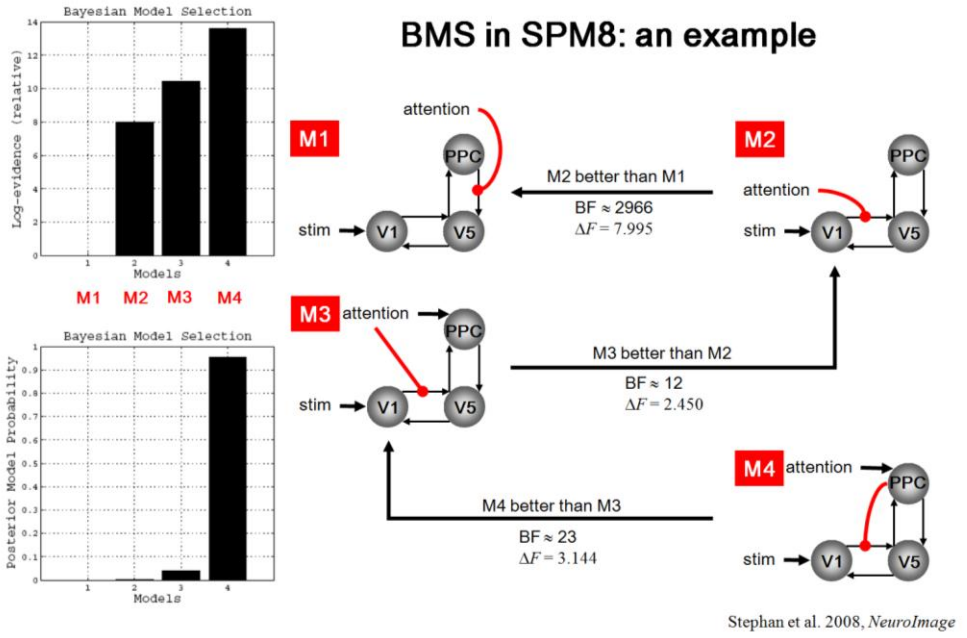
$$B_{12} = \frac{p(y \mid m_1)}{p(y \mid m_2)}$$

In this method, you go back from the Log space to real probabilities (to the evidence) and perform the ration of the two models. If the Bayes Factor is "1", it means that both of the models are equally look like. If Bayes Factor is "10", then the first model ten times more likely given the measurement. Additionally, there are different names for the range of Bayes Factor which are shown in the table below.

B_{12}	$p(m_1 \mid y)$	Evidence
1 to 3	50-75%	weak
3 to 20	75-95%	positive
20 to 150	95-99%	strong
≥ 150	≥ 99%	Very strong

Simple Example

To clarify the subject, again suppose the simple example of the first report-attention to motion using bilinear-nonlinear components. We want to model how visual stimulation enters V1 and V5 and PPC (posterior prior context). We want to answer to the question that when these simple hierarchy three regions do three regions does the attention enter?



In M1, we assume it to be in the backward connection from PPC to V5. We compare M1 with M2 in which attention enters the forward connection to V5. After examination, it is found that M2 is much better that M1. As you can see from the figure above, the Log evidence difference is almost 8, and the Bayes Factor is about 3000. This information is shown in the form of bar in the SPM. In M3, we can now ask whether attention stimuli PPC and found it better than M2. In this case, as figure shows, the Bayes Factor is 12. This effect is better in the M4 which is a nonlinear model.

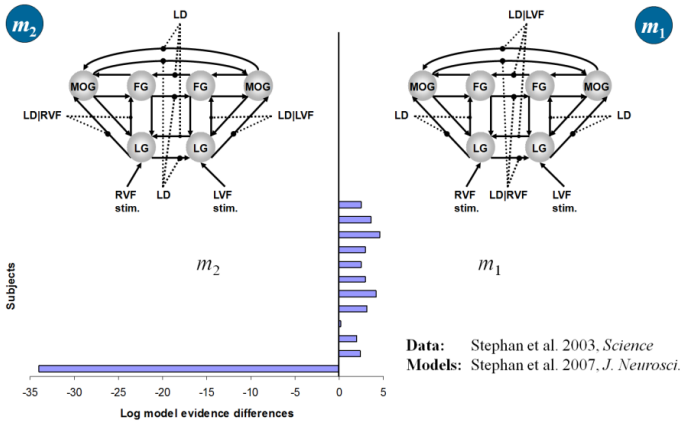
The example that we just saw was a simple way of hierarchical testing models. However, we usually have more than one subject. In this case the question arises to adopt fixed effects or random effects. As we discussed, fixed effects are very simple. You simply multiply the Bayes Factors to get the group Bayes Factor. Therefore for k subject we would have,

$$GBF_{ij} = \prod_{k} BF_{ij}^{(k)}$$

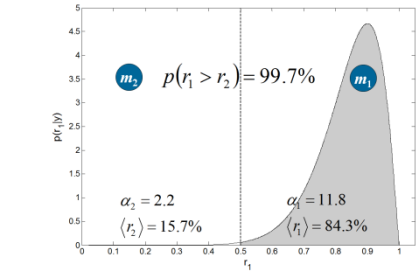
The problem with this method is that you have heterogeneous groups. You will not take that into account and singly outlines are completely distorts the results. One solution to this problem is to adapt to the **random variable approach**. In this case, you treat the model as random variable in the population and you can implement it by using a hierarchy of Bayesian Approach. This approach is developed by Dr. Stephan at 2010.

Where given the data, you can estimate the distribution of model probabilities in the population in several ways. In other words, you can ask how likely is it if I draw someone from the population whose data are generated by this model.

For example, as you can see from the figure below, we added two other regions to our previously discussed example. The models M1 and M2 are different from each other in that they switch around some of the inputs. If you apply these models to the twelve subjects, then plot the Log-model evidence difference. You can see that M1 is superior in eleven of them. The exception one is the subject who has a strong preference for the M2.



It is important to note that if we pooled the Log evidences in fixed fashion, the second model (M2) will win the competition. This is not the thing which we want. However, if you apply the random effects model which is shown in figure below, then the first model M1 would win.



Stephan et al. 2009a, *NeuroImage*

This is a new way of looking at the question that what is the probability that one model is more likely that another and this is something called expedience probability.