**Text S1: Bayesian regression and model parameter optimization**

 Here, we detail the formulation of 1) the Bayesian regression for the arm function and arm use model, 2) the model evidence criterion for model fit evaluation, and 3) the iterative parameter optimization techniques to tune the model parameters.

 Our state-space model in the main text Equation (1) and (2) is shown here again.

where are model parameters to be estimated for each subject ( is a “use effect” rate, a “confidence parameter” of using affected arm, and the function of the non-affected arm, as we discussed in the main text), and *t* is an time index *t=*1,…,*T* (*T*=7for the immediate group*,* and *T*=4 for the delayed group). We omit the subscript *affected* in the *F* and *U* for simplified notation, but still *F*(*t*) is arm function of stroke-affected arm at time *t*, measured by normalized negative log WMFT, and *U*(*t*) is arm use of affected arm at time *t*, measured by normalized MAL AOU. Here, we describe how to optimize these model parameters under Bayes model evidence, given *F* and *U* measurements.

 We reformulate Equations (S1) and (S2) to form equations linear in model parameters.

 

 

Then, we apply a standard linear regression form:



where (S5) and (S6) correspond to (S3) and (S4) respectively. Namely, and are the dependent (target) variables, representing left hand side of (S3) and (S4) respectively. The variables , and are the basis functions ( = and). This formulation with different basis functions and different number of model parameters is repeated to the other competitive models in Table 1 and 2. Note that we can decouple and for the purpose of model parameter estimation; hence, we use as an example in the following discussion.

 We now apply time-evolving measurements as a design matrix to the arm use regression model (S6) in a vector form.

where , , and the design matrixis given by

Since our model is derived from 1st order Markov chain, we need two consecutive measurements of U(t) and U(t+1) for the linear regression form Therefore, the number of *L* is one less than *T: L=T*-1 (*L*=6 for the immediate group, and *L=3* for the delayed group). Hence *L*×*M* matrix, where *M* is the number of model parameters (i.e., *M*=2 for arm use model (S2)).

 The regression models cannot perfectly capture the dynamics, and we assume that the target variable ***y*** varies with Gaussian distribution.

where is independent identically distributed (i.i.d) Gaussian noise with the mean and an accuracy hyper-parameter We set the initial =10-8 for model fitting and vary for sensitivity analysis (see below).

 In Bayesian regression, we treat model parameters as a probability distribution, not as single point values. We assume that the *prior* distribution of model parameters is also independent identically distributed Gaussian.

where is the mean of model parameter, and accuracy. Batch least square solutions of all subjects data from the immediate or delayed group are used for the initial mean of model parameters (except for =1 in weighted average arm function model). Since we have little knowledge of model parameters, we set very small ().

 We reflected our emphasis on individual data and lack of prior knowledge by setting the ratio of the initial values of the prior accuracy ****and the data accuracy ****to ****= 10-3 and choosing almost flat priors with ****= 10-11. We verified in simulations that when ****< 10-11, the results of model comparison are qualitatively the same as that presented. We set **** = 10-8 for model fitting and varied initial and performed a sensitivity analysis (i.e. a systematic variation) on the initial precision****.

 From (S8) and (S9), we notice that the mean of is not a deterministic function, since is also a Gaussian distribution as described in (S9). This means that we cannot evaluate the model fit using point to point distance error metric; therefore, we use a Bayesian approach. We introduce a metric called Bayesian model evidence, which is the probability density function of data, given model parameters:. Using the sum rule, and product rule,

 Substituting (S8) and (S9) into (S10), we have

where

This is a Gaussian distribution with mean and covariance .However, here, we directly evaluate integral term for derivation purpose.

We separate terms with model parameters and without them.

where
We complete the square of the model parameter terms

where
Further, we complete the square over and.

 (S15)

where is the Euclidean norm. We substitute (S15) into (S11), and use the fact that the integral of Gaussian distribution is equal to 1:

where |**A**| is the determinant of design matrix **A**. The resulting log model evidence probability is given by

 (S16)

Although (S16) looks complicated, among constant terms related with the data size and the number of model parameters, we can find a trade-off between a regularization term and a data fitting error term, which are balanced by and Roughly speaking, a complicated over-fitting model is penalized by the number of model dimension . With larger (more model parameters), can better approximate the data distribution, and the error between and , decreases. On the other hand, because the size of vector also scales up with larger number of parameters, the regularization term may increase. Similar trade-offs are found in (S13) and (S14) in the form of weighted average between prior knowledge and data. Note that, since the design matrix utilizes all data points, unlike cross-validation (e.g., leave-one-out), we do not need to spare testing data points for evaluating model fit.

 Now, we maximize this model evidence probability in terms of which controls model parameter distribution (S9), and which controls data distribution (S8). Note that (S13)and(S14)are also functions of and We use an iterative method, where we fix and in the first step and optimizeand and update and with the new and in the second step (Chap. 3.5.2, Bishop, 2006). Note that this iterative method can be formulated equivalently by EM algorithm (Chap. 9.3.4, Bishop, 2006). Taking derivative of (S16) with respect to is trivial in most terms but term. The determinant of the matrix can be expressed by the product of eigenvalues of **A**: **=**. Since ,the eigenvalue of **A** is given by where is *i*-th eigenvalue of Hence,
Using this, we take the derivative of (S16) with respect to , and set the derivative = 0.

Arranging this equation, and defining with

Alternatively, we can express as follows.

From (S19),

Similarly, taking derivative of (S16) with respect to ,

where we used the following equation of the derivative of term with respect to

It is similar to (S17), but is a linear function of in , where **,** andis an eigenvector associated with Taking derivative: **,** and multiplying to the left hand side, we have  Hence, the derivative of an eigenvalue with respect to is the ratio of them: . Solving (S22) for , we have

 Our iterative method computes and using (S21) and (S23) in the first step, and then use the updated and to re-compute and using (S13) and (S14) in the second step. We repeat these steps until and converge. After several iterations, suppose that we have the converged values of and denoted by and . This means that the model evidence log probabilityis maximized to .

 Now, we can use this model evidence probability to compute the posterior probability of model parameters. Remember that the prior distribution of model parameters is given by Gaussian , and the data distribution is also Gaussian , in which the mean is a linear combination of . Applying a linear Gaussian formula (Roweis & Ghahramani, 1999 and Chap. 2.3.3, Bishop, 2006) of given we have

with the mean and covariance

where are actually given by (S13) and (S14) and are converged values after optimizationand . From this, we find that used in the model evidence computation above is the mean and accuracy (inverse of covariance) of the model parameter probability distribution. Note that like EM algorithm, is sub-optimal and hence model evidence too.

 Here is a summary of the algorithm to compute Bayesian evidence with and optimization.

|  |
| --- |
| 1. Set andinitialand 2. Compute and using (S21) and (S23)  where and is *i*-th eigenvalue of 3. Update using (S13) and (S14) 3.1  3.2 4. Repeat 2 and 3 until convergence5. Set and , and compute model evidence (S16), and posterior model parameter distribution (S24)5.2  |

 This discussion is based on (Chap. 3.5, Bishop, 2006 and Chap. 28, D. J. C. MacKay, 2002). For more detail, please refer to the books. It should be noted that Bayesian regression fully utilizes the data set for both model fitting and evaluation, without partitioning data for cross-validation.

 We used the model evidence log probability as a metric of model fitting optimization, and applied this algorithm for individual subject data (48 subjects for immediate group, and 45 for delayed group) for all different models (9 models for arm function, and 4 models for arm use). The evidence for each model was used to compare the models by computing the Bayes factor (BF). Given the model evidence probabilities for our reference model and for a competitive model, the Bayes factor is given by /. This Bayes factor was computed for each subject.

 We evaluated Positive Evidence Ratio (PER) (Stephan, et. al., 2007), which compares the number of subjects, where the model evidence is positive (BF>=3) versus negative (BF=<1/3). Note that this criterion excludes the “barely worth mentioned” evidence (1/3< BF < 3). Since it simply compares each subject Bayes factor, this metric is robust against outliers. The result of PER was shown in Table 3 and 4 of the main text.

**References**

Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer.

MacKay, D. J. C. (2002). Information Theory, Inference & Learning Algorithms. Cambridge University Press.

Roweis, S., & Ghahramani, Z. (1999). A Unifying Review of Linear Gaussian Models. Neural Computation, 345(1995), 305-345.

Stephan, K. E., Marshall, J. C., Penny, W. D., Friston, K. J., & Fink, G. R. (2007).Interhemispheric integration of visual processing during task-driven lateralization. The Journal of neuroscience, 27(13), 3512-22.