JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

SMART JHSPH Biostatistics

Efficient Modeling of fMRI Data

– Avoiding Misspecification, Bias and Power loss

Martin Lindquist
Department of Biostatistics
Johns Hopkins University

Statistical Analysis

- Statistics is an integral part of neuroimaging research.

- Applying statistics to real-world problems is hard.
  - It requires the careful selection of appropriate data analytic techniques and verification of assumptions.

- A first step is determining an appropriate model.
  - A mathematical representation of a real-world phenomena.

Model Building

- Deciding on an appropriate model requires careful deliberation.
- In the best case we have a theoretical model laid out before proceeding with data analysis.
- In practice, we usually start with a simple model and refine it until we get it ‘right’.
  - In neuroimaging research we don’t often have this luxury due to the massive amounts of data.

General Linear Model

- The general linear model (GLM) approach has been a workhorse in the field for many years.
- It treats the data as a linear combination of model functions (predictors) plus noise (error).
- Model functions are assumed to have known shapes, with unknown amplitudes that need to be estimated.

General Linear Model

A standard GLM can be written:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{V})$$

where

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{np} & \cdots & X_{np} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

fMRI Data

Design matrix

Model parameters

Noise

V is the covariance matrix whose format depends on the noise model.

The quality of the model depends on our choice of **X** and **V**.

Model Efficiency

- Any GLM based analysis is only as good as the specified design matrix.
- Incorrect specification can lead to bias and model misfit, resulting in power loss and an inflated false positive rate.
- Problems can arise if:
  - irrelevant regressors are included, or
  - relevant regressors are omitted, or
  - certain regressors are mismodeled.

Example – Omitted Variables

- Truth:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2\right)$$

- Model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} \qquad \boldsymbol{\eta} \sim N\left(\mathbf{0}, \mathbf{I}\sigma^2\right)$$

- 'Optimal' estimates:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \qquad s^2 = \frac{\hat{\boldsymbol{\eta}}^T\hat{\boldsymbol{\eta}}}{n-p}$$

Effects of Mismodeling

- The estimate of β is biased:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \underbrace{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Z}\boldsymbol{\gamma}}_{\text{Bias}}$$

- The bias disappears if
  - the omitted variable is irrelevant, or
  - it does not correlate with the explanatory variables included in the model.

Effects of Mismodeling

- The estimate of σ^2 is biased:

$$E(s^2) = \sigma^2 + \underbrace{\frac{1}{n-p}\boldsymbol{\gamma}^T\mathbf{Z}^T\mathbf{R}\mathbf{Z}\boldsymbol{\gamma}}_{\text{Bias}}$$

where $\mathbf{R} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$.

- The variance follows a non-central χ^2 distribution with non-centrality parameter $\delta = \boldsymbol{\gamma}^T\mathbf{Z}^T\mathbf{R}\mathbf{Z}\boldsymbol{\gamma}$.

Effects of Mismodeling

- If relevant variables are omitted or mismodeled:
  - Regression coefficients and standard deviations are biased.
  - The statistic used to test significance of the regression coefficients follows a doubly non-central t-distribution rather than a standard t-distribution.

- If irrelevant variables are included:
  - Regression coefficients are still unbiased.
  - Standard error of the regression coefficients are inflated (smaller t-values).

Mismodeling in the GLM

- There are many potential issues that can give rise to mismodeling in the GLM.
  - The BOLD signal may contain low-frequency noise and artifacts related to head movement and cardiopulmonary-induced brain movement.
  - The neural response shape may not be known.
  - The hemodynamic response may vary in shape across the brain.

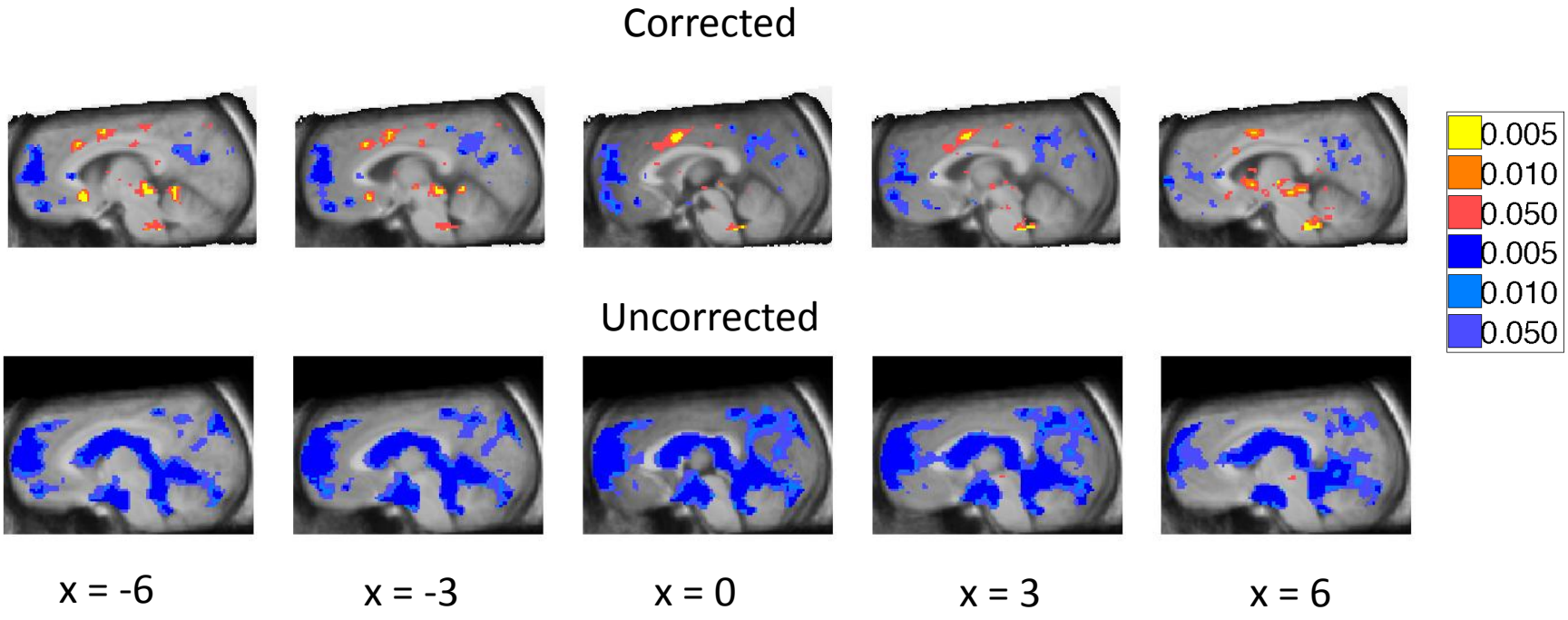
- If significant mismodeling is present it is important to perform some model refinement.

Nuisance Covariates

- Often model factors associated with known sources of variability, but that are not related to the experimental hypothesis, need to be included in the GLM.

- Examples of possible 'nuisance regressors':
  - Physiological (e.g., respiration) artifacts
  - Head motion, e.g. six regressors comprising of three translations and three rotations.
    - Sometimes transformations of the six regressors also included.

Head Motion Example



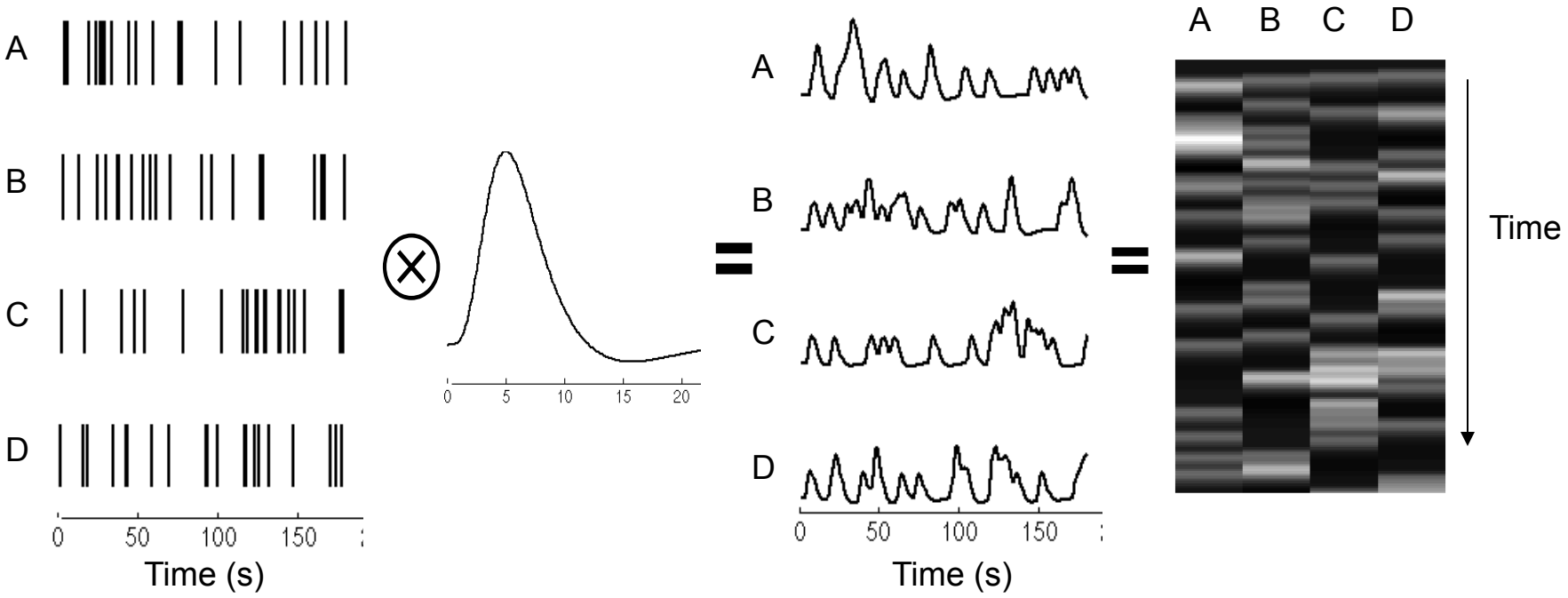
Task Related Signal

Indicator functions (Onsets)

Assumed HRF (Basis function)

Design Matrix (X^T)

Design Matrix (X)



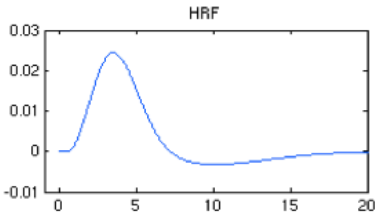
Assumptions:

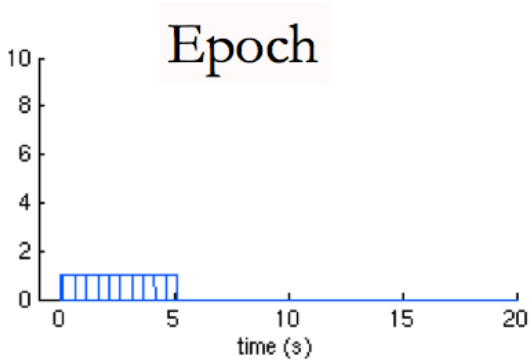
Assume neural activity function is correct

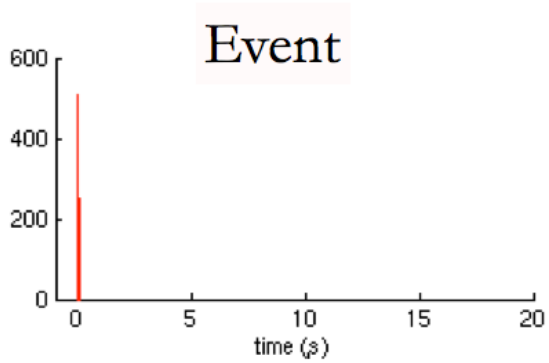
Assume HRF is correct

Assume LTI system

Stimulus Models

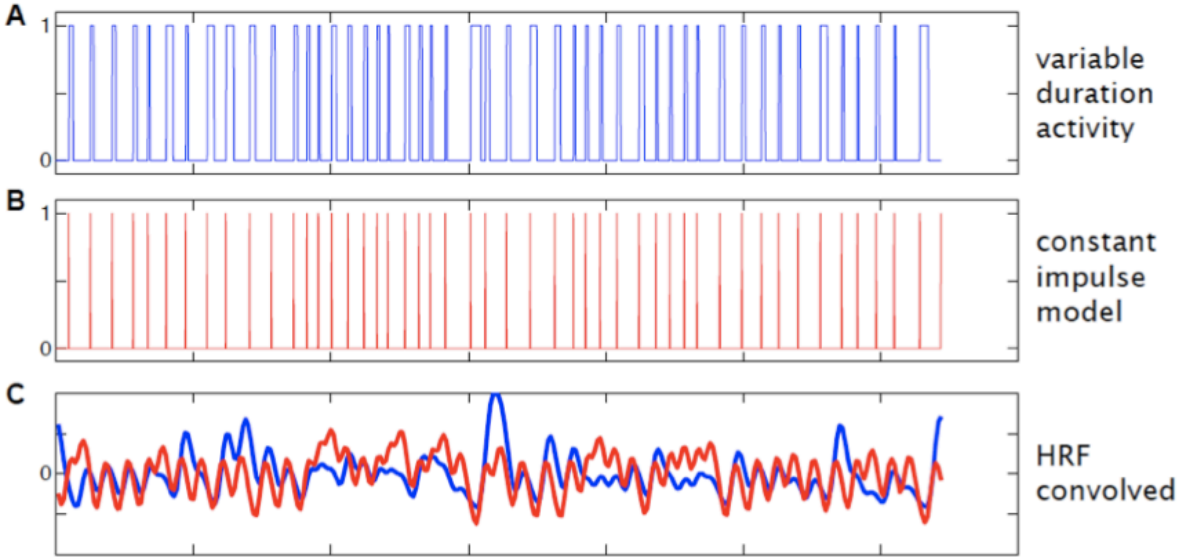






Does it matter for event-related fMRI?
* answer: More than you might think!

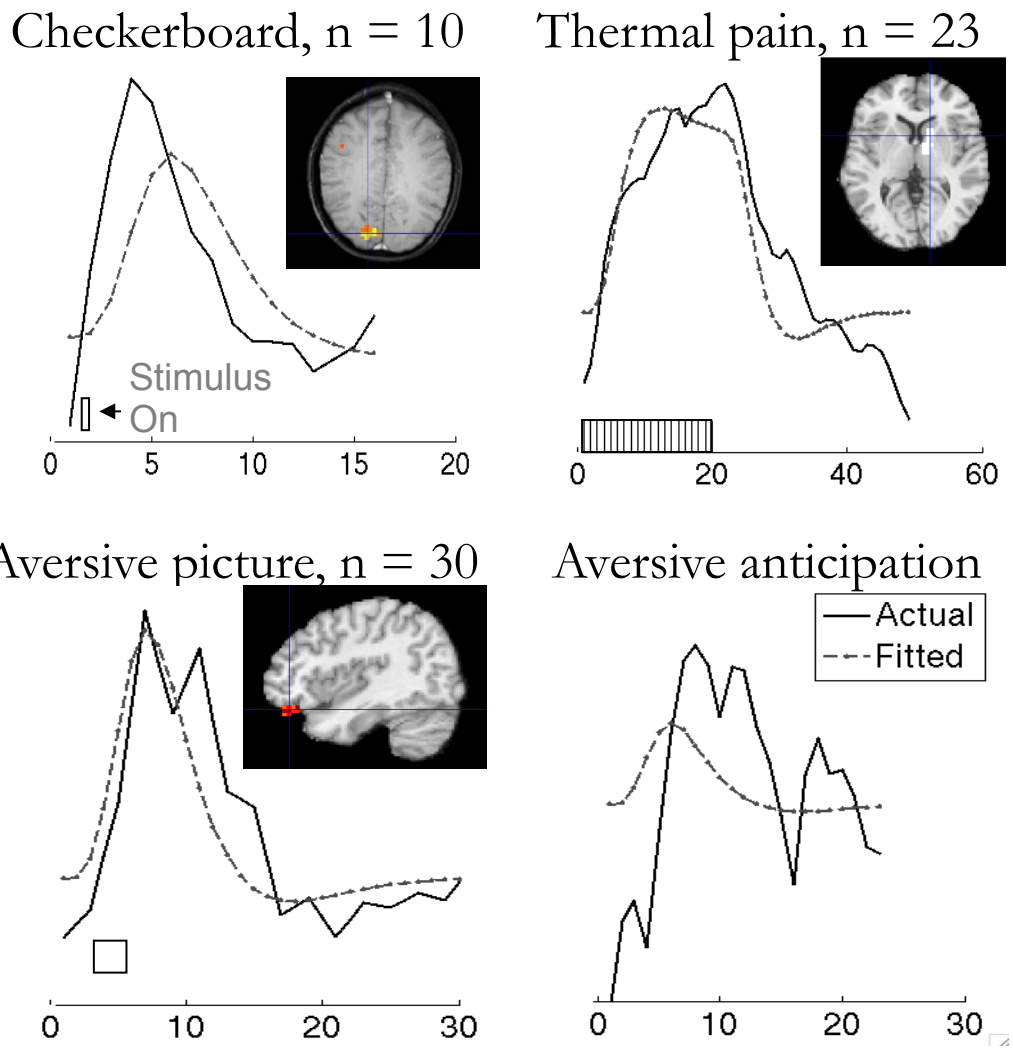
Grinband et al., 2008



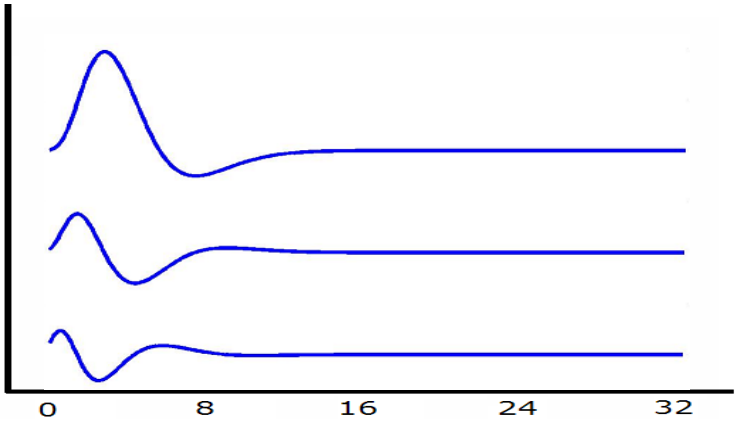
Problems

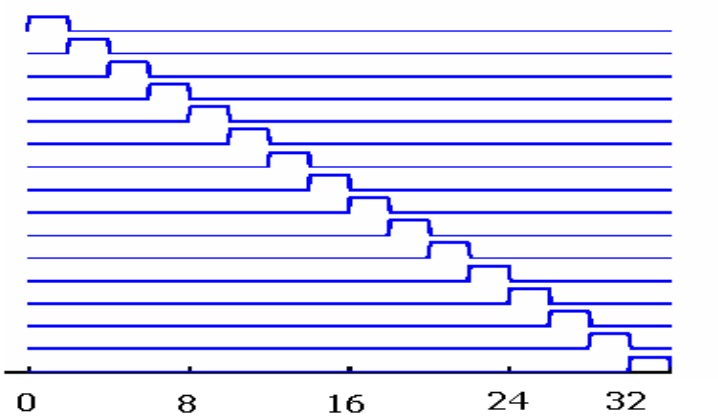
The HRF shape depends both on the vasculature and the time course of neural activity.

Assuming a fixed HRF is usually not appropriate.



Temporal Basis Sets





- Canonical HRF + Derivatives
  - Including the derivatives allows for a shift in delay and dispersion.
- Finite Impulse Response
  - The model estimates an HRF of arbitrary shape for each event type in each voxel

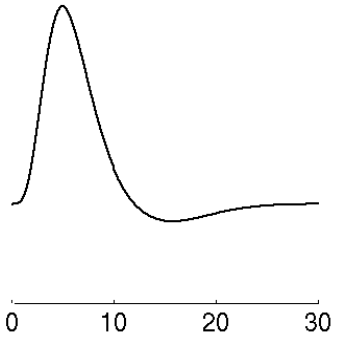
Basis sets

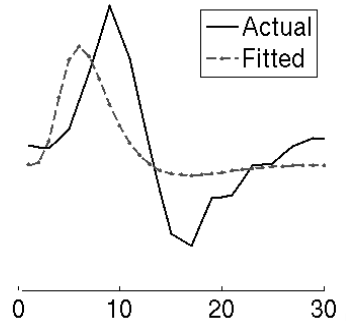
Model

Image of predictors

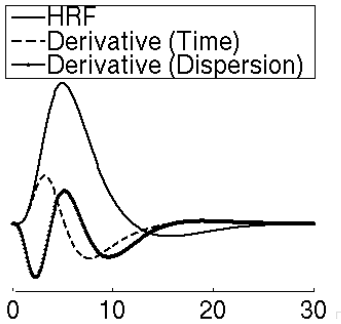
Data & Fitted

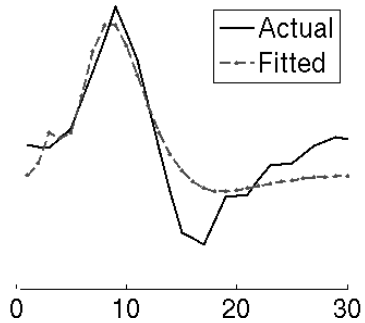
Single HRF



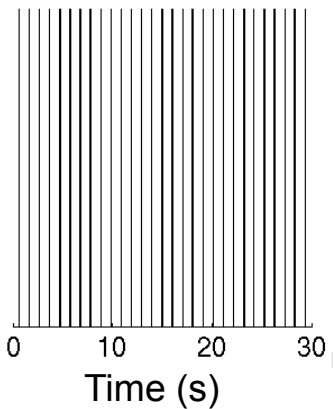


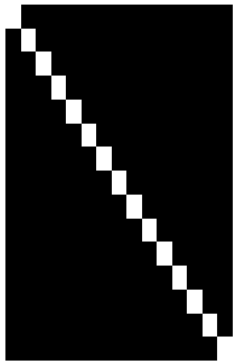
HRF + derivatives

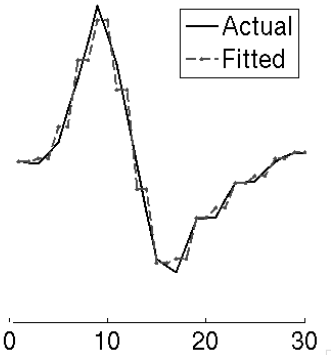




Finite Impulse Response (FIR)

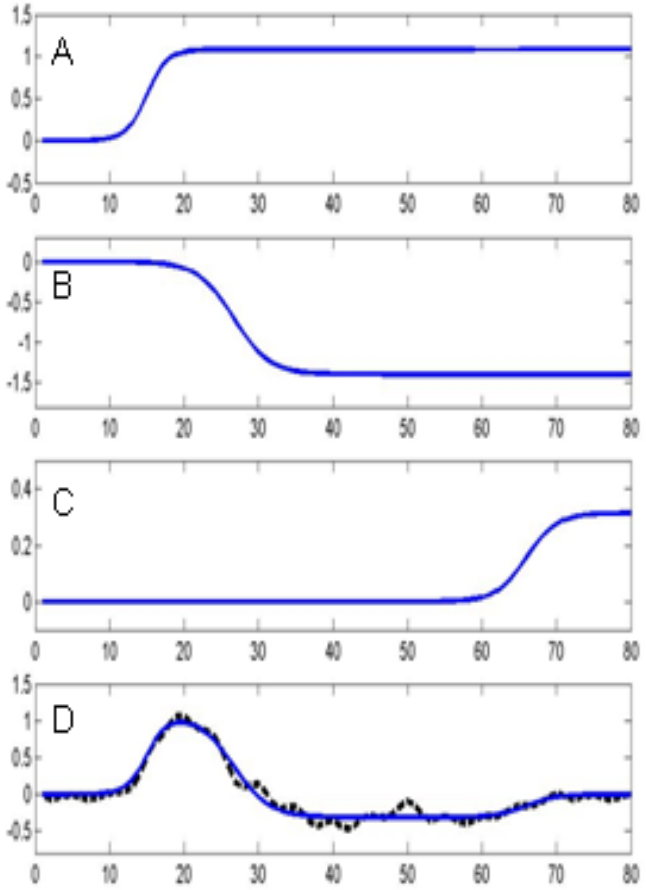


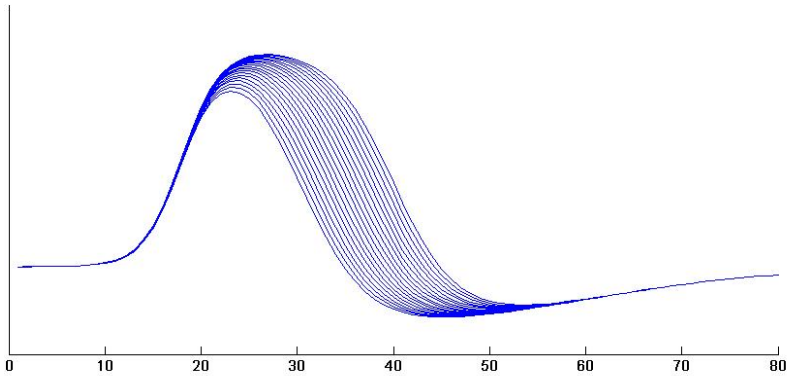


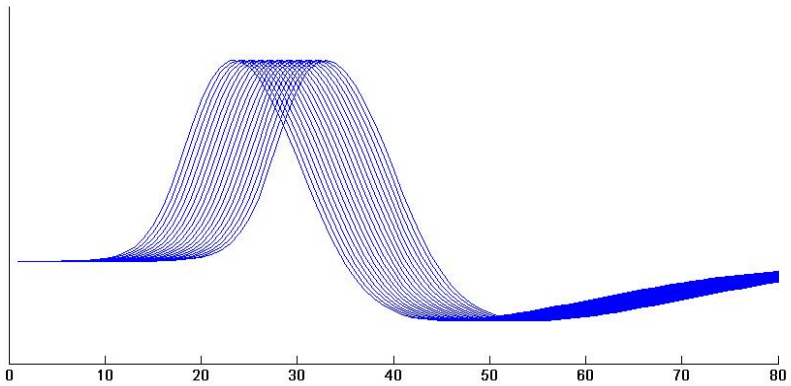


Inverse Logit Model

- Superposition of three inverse logit (sigmoid) functions.





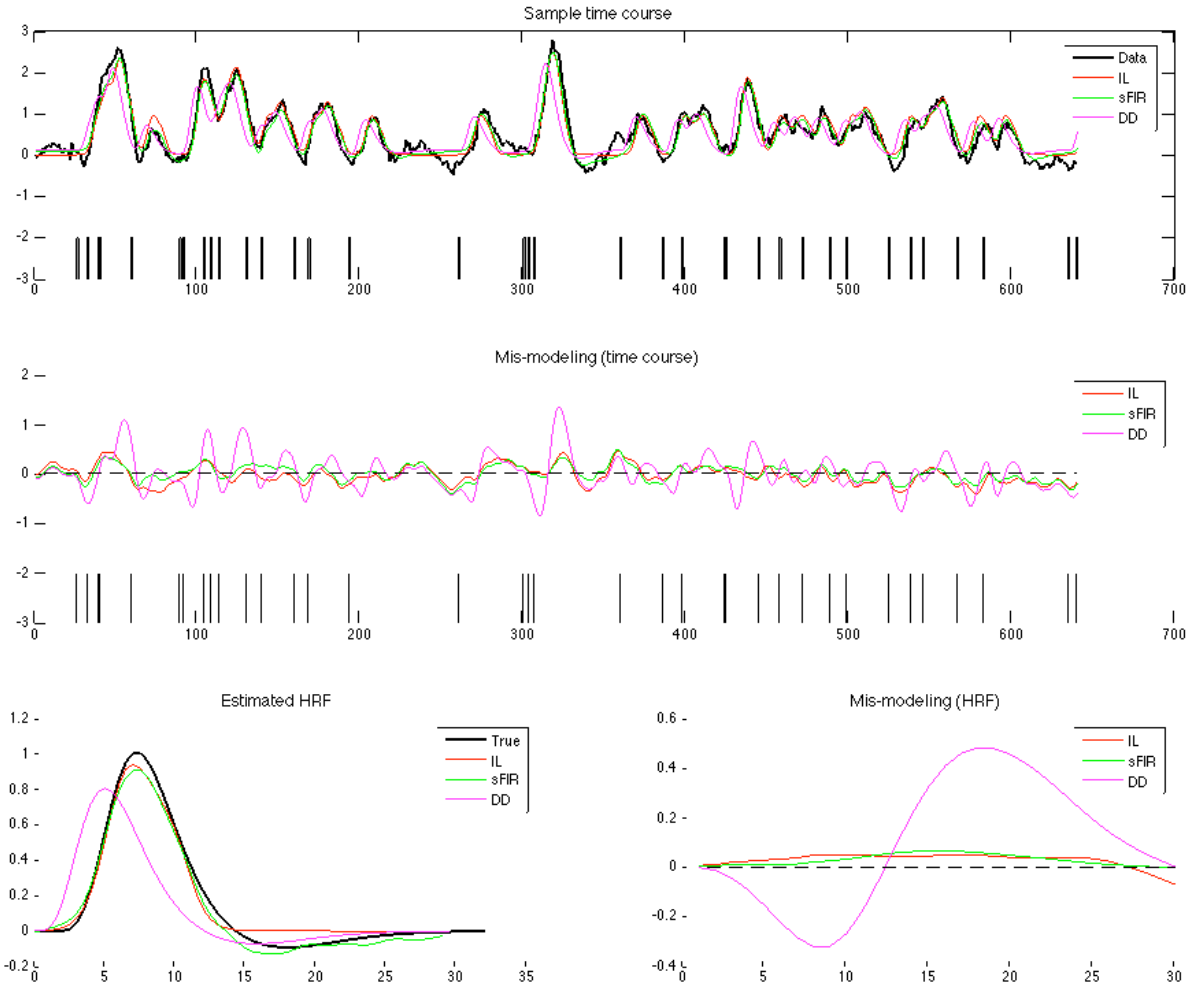


Detecting Mismodeling

- The residuals of the GLM provide important clues about possible mismodeling.
  - If crucial variables are omitted, there should be signal left in the observed residuals.

- Study the residuals to:
  - Estimate the amount of mismodeling.
  - Construct bias and power-loss maps across voxels to determine regions that are particularly effected (Loh et al. 2008).
  - Identify the presence of systematic mismodeling: either periodic or as a function of the stimulus.

Example

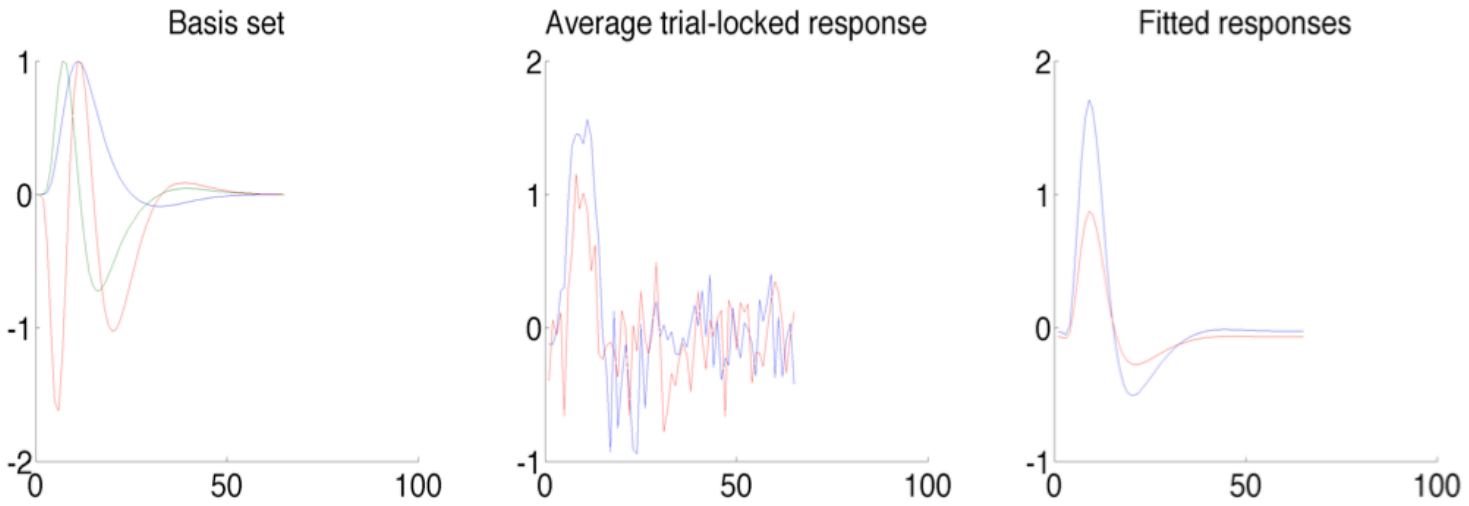

Sample time course
Data
IL
sFIR
DD
Mis-modeling (time course)
IL
sFIR
DD
Estimated HRF
True
IL
sFIR
DD
Mis-modeling (HRF)
IL
sFIR
DD

Group Analysis

- When using temporal basis sets at the first level it can be difficult to summarize the response magnitude with a single number, making multi-subject inference difficult.

- In this setting we can perform group analysis using
  - the “main” basis function,
  - all basis functions, or
  - re-parameterized fitted responses (Calhoun et al. (2004); Lindquist et al. (2009)).
    - Recreate the HRF and estimate the magnitude.
    - Use this information at the second level.

Example



- Suppose we want to estimate A-B
  - Difference between amplitude of fitted responses: 0.84
  - Difference between canonical HRF betas: 0.43

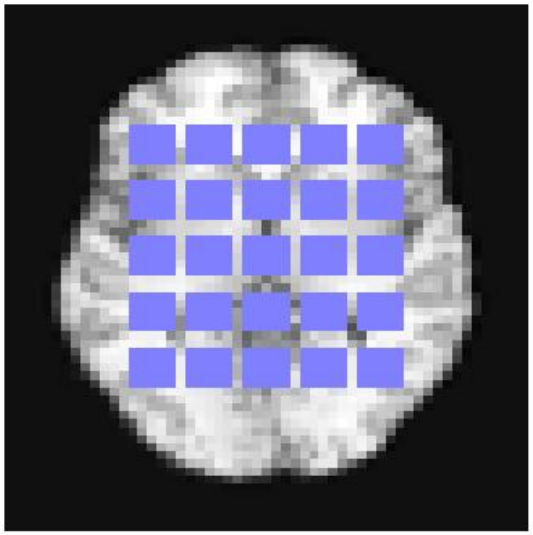
Simulation Study

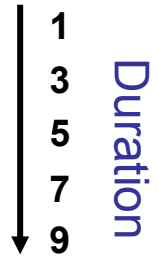
- We performed simulations to compare various models ability to handle shifts in onset and duration with respect to bias and power-loss.

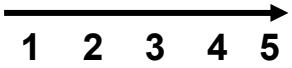
- The models we studied were:
  - The canonical HRF
  - The canonical HRF + temporal derivative
  - The canonical HRF + temporal & dispersion derivative
  - The FIR model
  - The Smooth FIR model
  - Inverse Logit model

Lindquist & Wager (2007)
Lindquist, Loh, Atlas & Wager (2008)

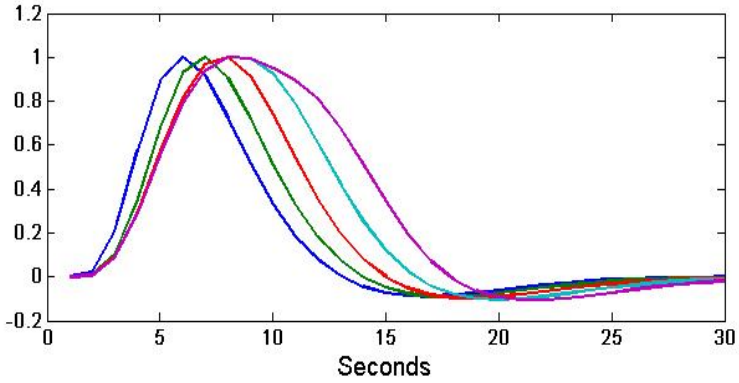
Simulation







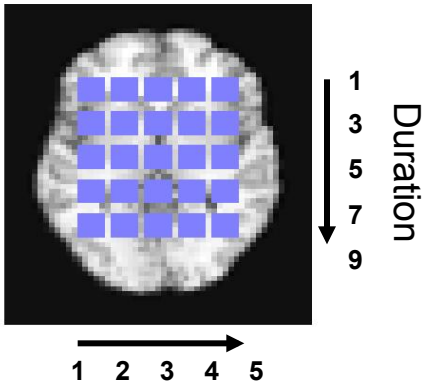
Onset shift



25 unique activations

- TR=1, ISI = 30, 10 epochs, 15 “subjects”, Cohen’s d = 0.5
- Estimates of amplitude were obtained and averaged across the 15 subjects.

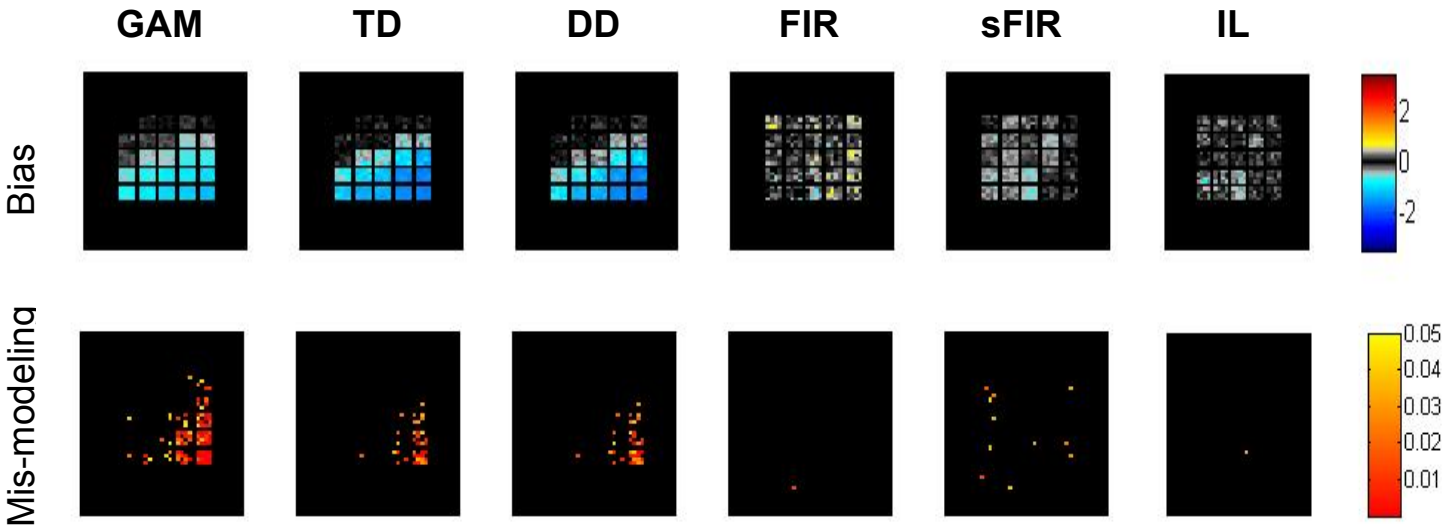
Results



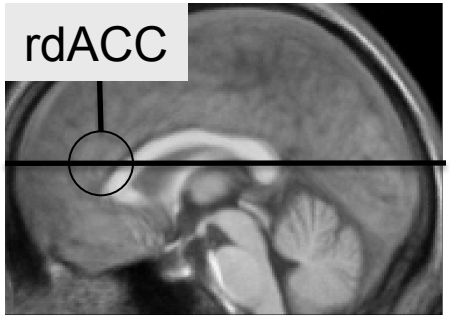
Negative Bias

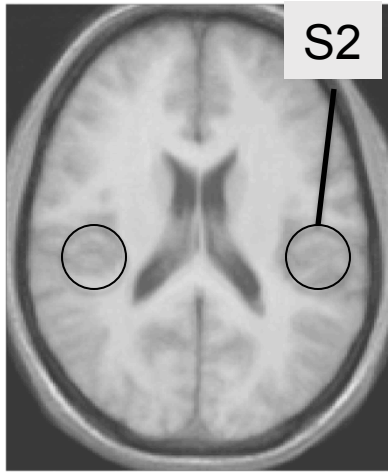
No Significant Bias

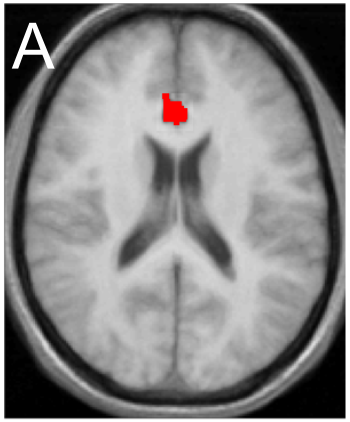
Positive Bias



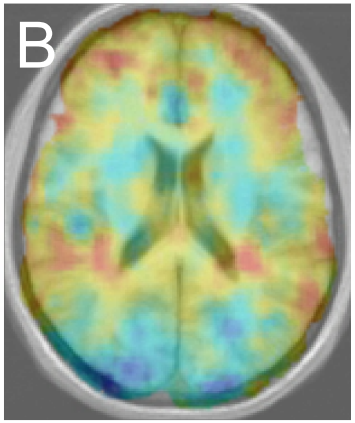
Pain Study



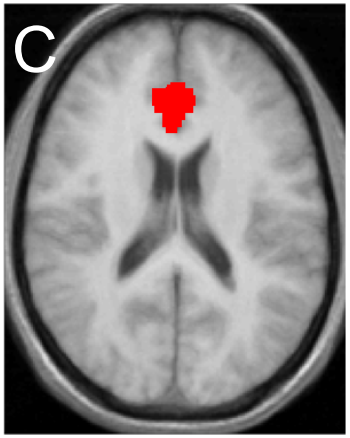




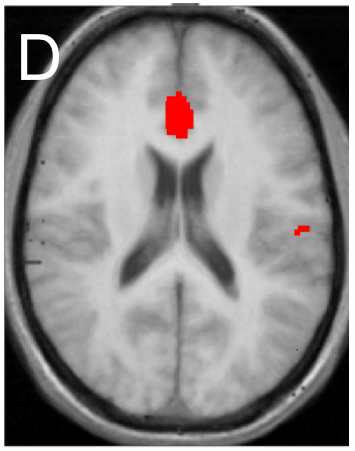
DD



DD- mismodel

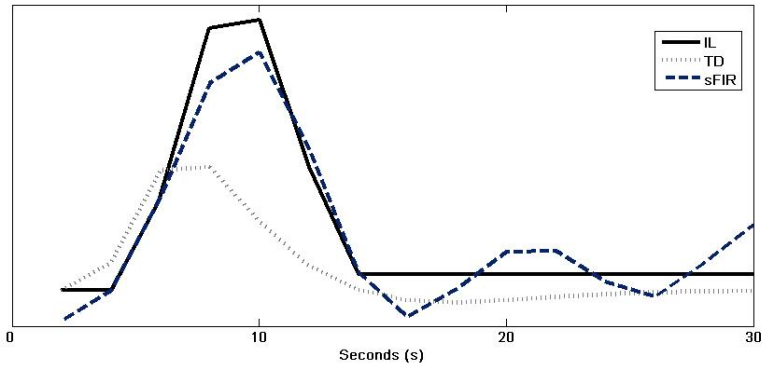


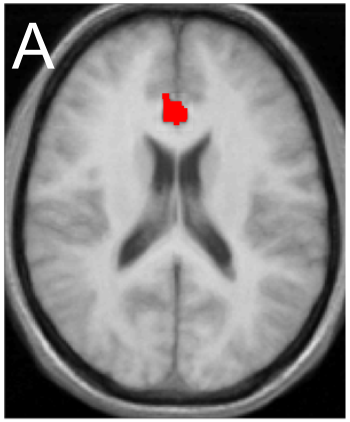
sFIR



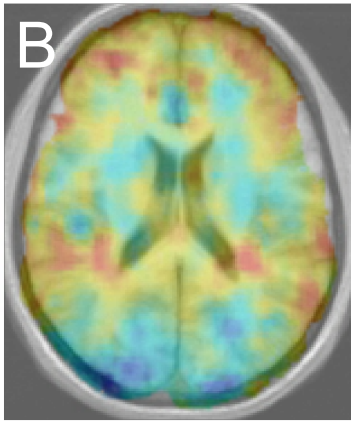
IL

Pain Study

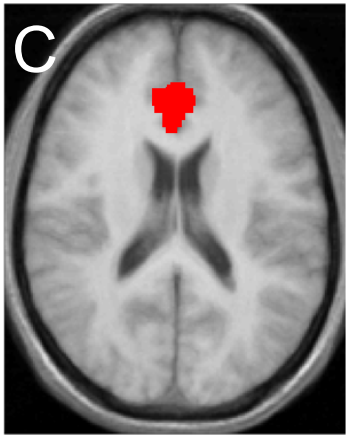




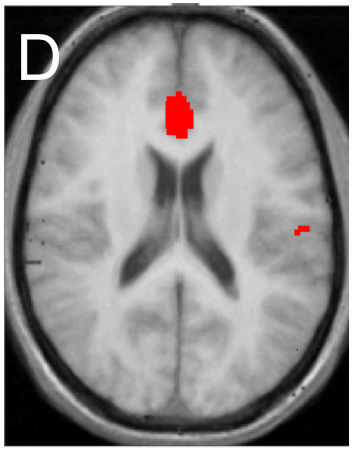
DD



DD- mismodel



sFIR



IL

Comments

- Model building is difficult in neuroimaging.
- Always be skeptical about your models.
  - If model assumptions don't hold, be careful about the conclusions you are willing to make.
  - Present results together with the assumptions made.
  - Try to check all verifiable assumptions.
  - Critically evaluate non-verifiable assumptions.
- Connectivity studies are even more complicated.
  - Different assumptions provide different conclusions.

References

- Martin Lindquist and Tor Wager (2006). Validity and Power in Hemodynamic Response Modeling: A comparison study and a new approach. *Human Brain Mapping, 28(8) 764-784.*

- Ji-Meng Loh, Martin Lindquist and Tor Wager (2008). Residual Analysis for Detecting Mis-modeling in fMRI. *Statistica Sinica, 18, 1421-1448.*

- Martin Lindquist, Ji Meng Loh, Lauren Atlas, and Tor Wager (2008). Modeling the Hemodynamic Response Function in fMRI: Efficiency, Bias and Mis-modeling. *NeuroImage, 45, S187-S198.*

Thank You!

- HRF estimation and mismodeling software available in MATLAB.

  www.stat.columbia.edu/~martin/Software.html