Modeling Neuroimaging Data

– avoiding misspecification, bias and power loss

Martin Lindquist
Department of Statistics
Columbia 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

$\mathbf{V}$ is the covariance matrix whose format depends on the noise model.

The quality of the model depends on our choice of $\mathbf{X}$ and $\mathbf{V}$.

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:

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

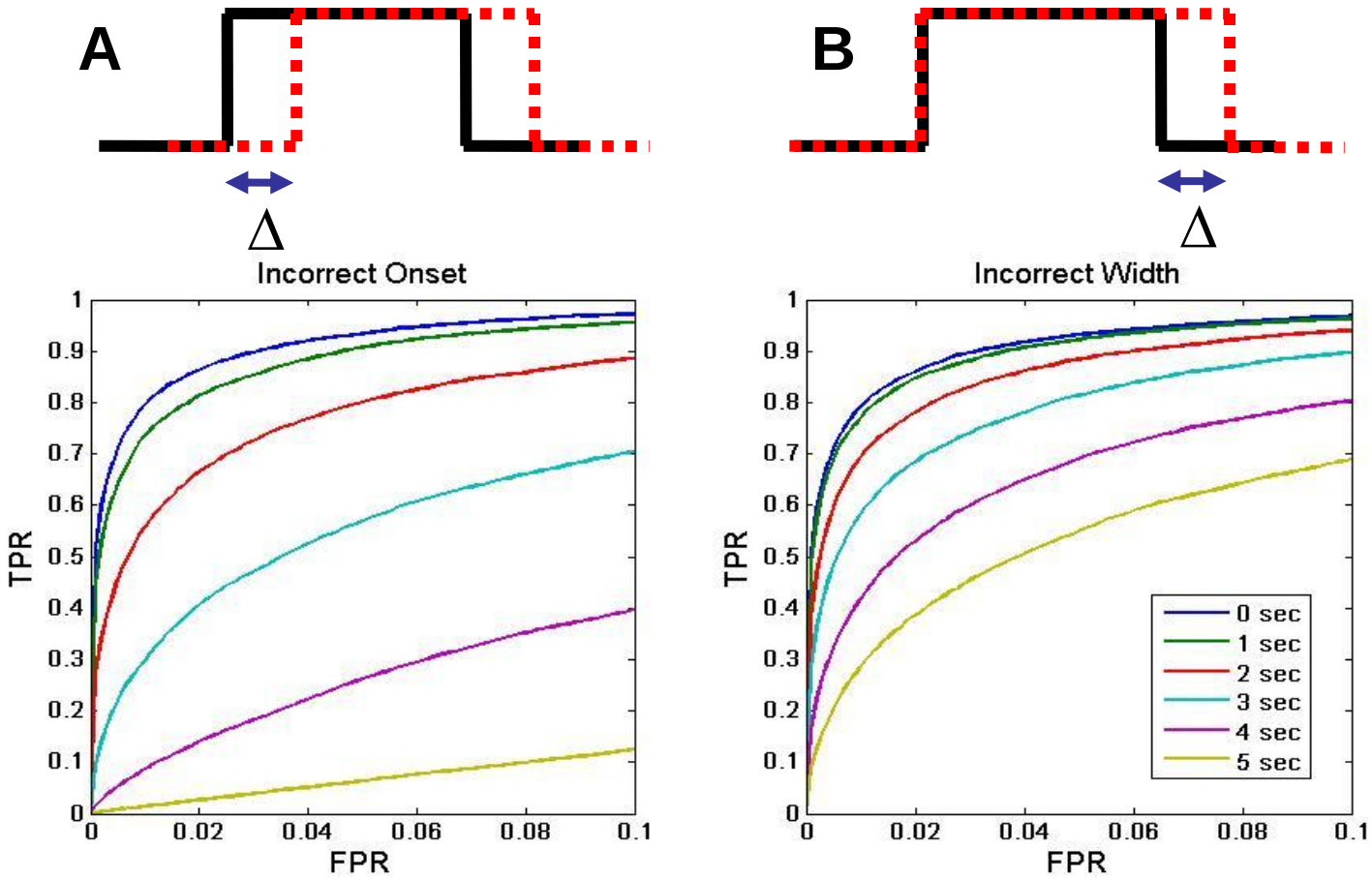
  where $\mathbf{R} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$.

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

Example



Solid = model
Dashed = truth

The true activation paradigm is repeated 4 times (Cohen's d = 0.5).

Detecting Mismodeling

- We need a simple approach for detecting mismodeling in the massive univariate setting.
- If crucial variables are omitted, there should be signal left in the observed residuals of the GLM.
- Study the residuals to:
  - Estimate the amount of mismodeling.
  - Construct bias and power-loss maps across voxels to determine regions that are particularly effected.
  - Identify the presence of systematic mis-modeling: either periodic or as a function of the stimulus.

Loh, Lindquist & Wager (2008)

Detecting Mismodeling

- Let r be the whitened residuals and K_w a Gaussian kernel with FWHM w.

- When no mismodeling is present the statistic

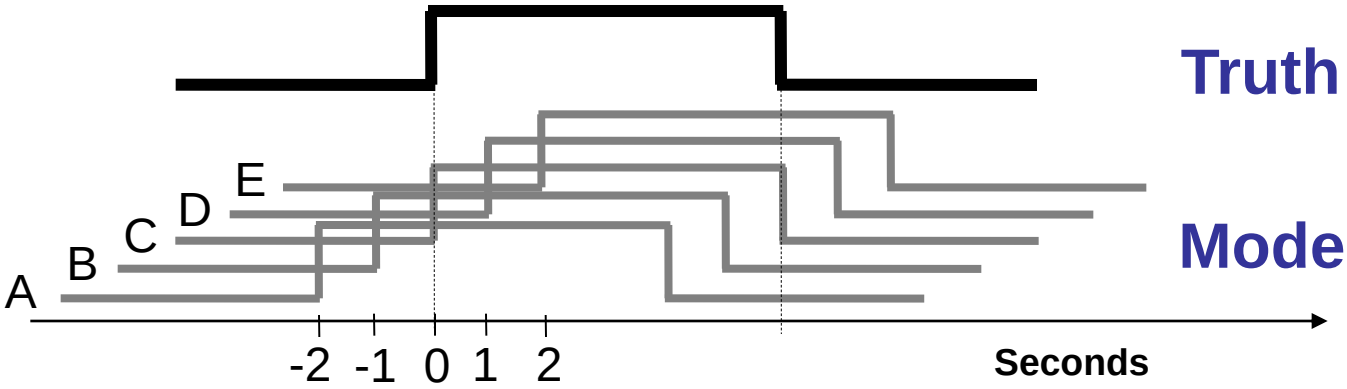
$$Y_w(t) = (r * K_w)(t)$$

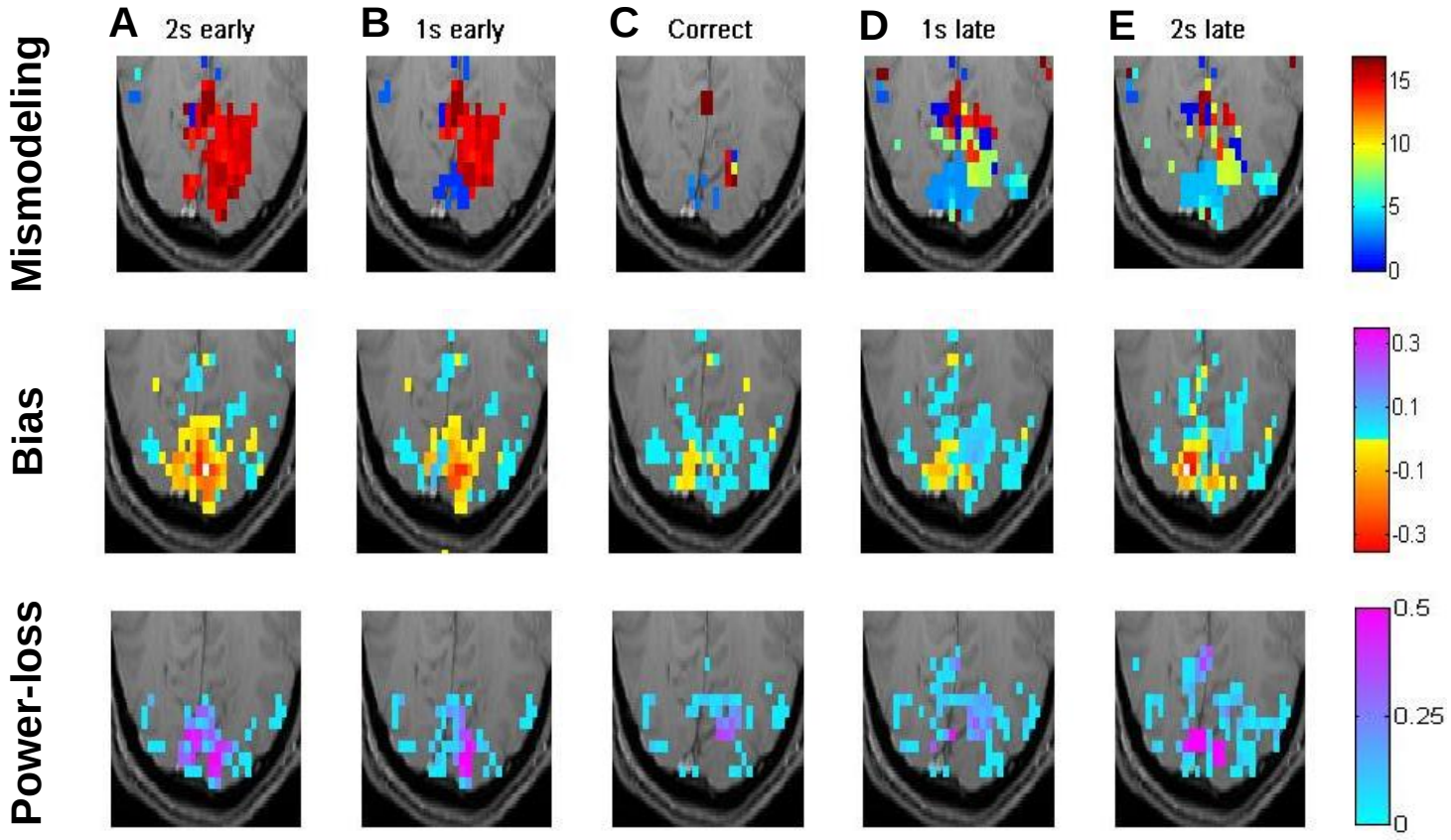
  is normally distributed with mean 0 for all w, t.

- We can calculate $P(\max_t Y_w(t) > \tau)$ using random field theory.

Loh, Lindquist & Wager (2008)

Mismodeling, Bias and Power-loss maps





Model Refinement

- If significant mismodeling is present it is important to perform some model refinement.
- One common form of mismodeling is the incorrect specification of the shape of the HRF.
  - Can result in severe power loss and inflation of the false positive rate beyond the nominal value.
- One solution is to use temporal basis functions rather than a fixed HRF.
  - Do all basis functions offer the same flexibility?

Simulation Study

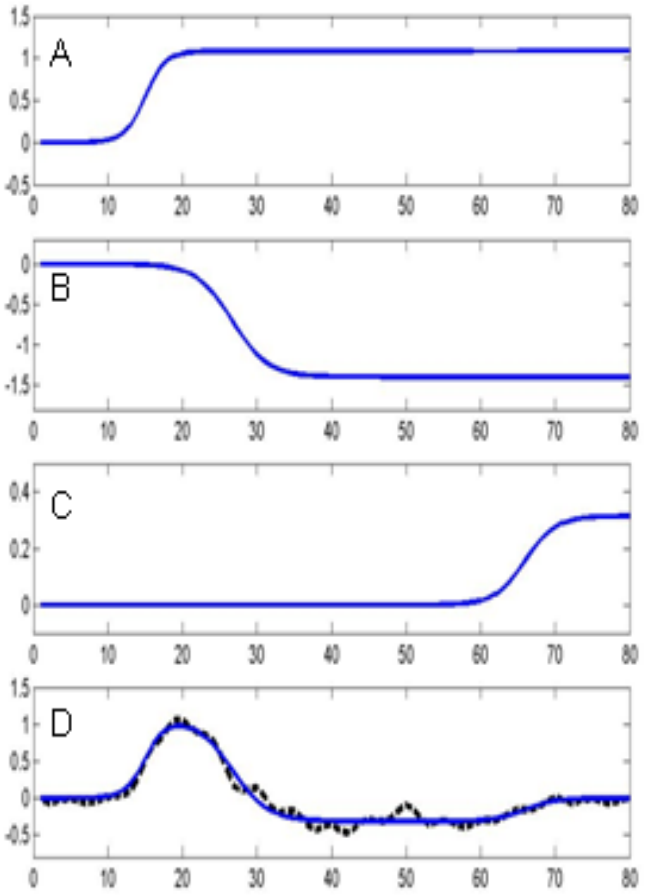
- We performed simulations to compare various models ability to handle shifts in onset and duration with respect to bias and power-loss.

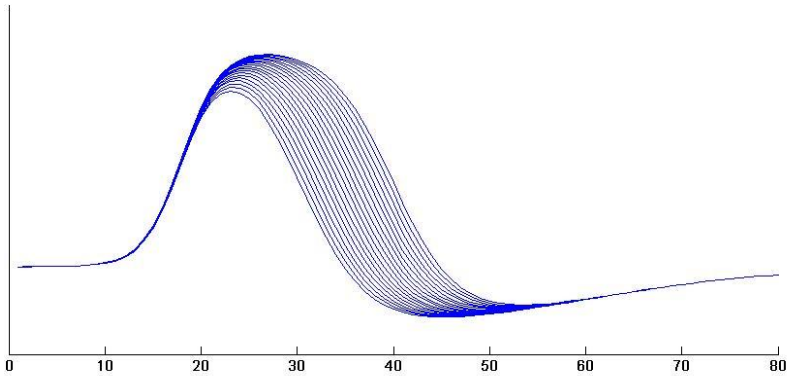
- The models we studied were:
  - The canonical HRF
  - The canonical HRF + temporal derivative
  - The canonical HRF + temporal & dispersion derivative
  - The FIR model
  - The Smooth FIR model
  - Inverse Logit model

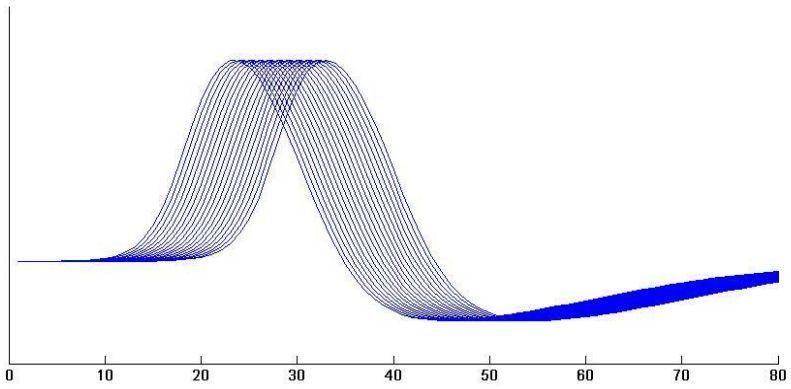
Lindquist & Wager (2007)
Lindquist, Loh, Atlas & Wager (2008)

Inverse Logit Model

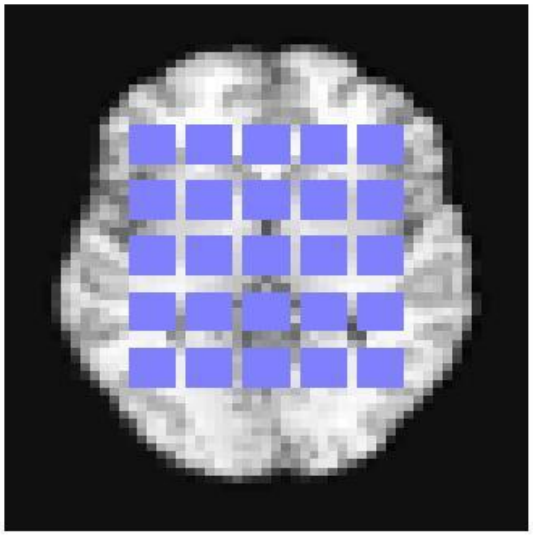
- Superposition of three inverse logit (sigmoid) functions.







Simulation

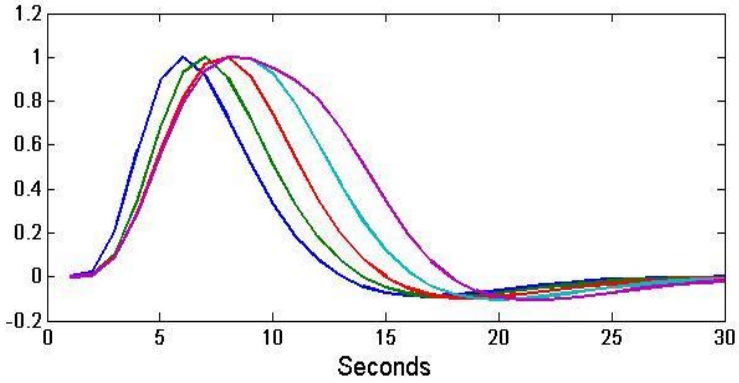


1 3 5 7 9

Duration

1 2 3 4 5

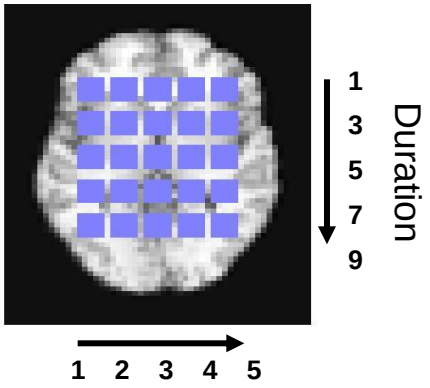
Onset shift

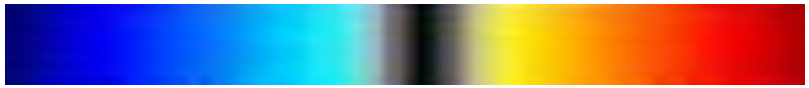


25 unique activations

- 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

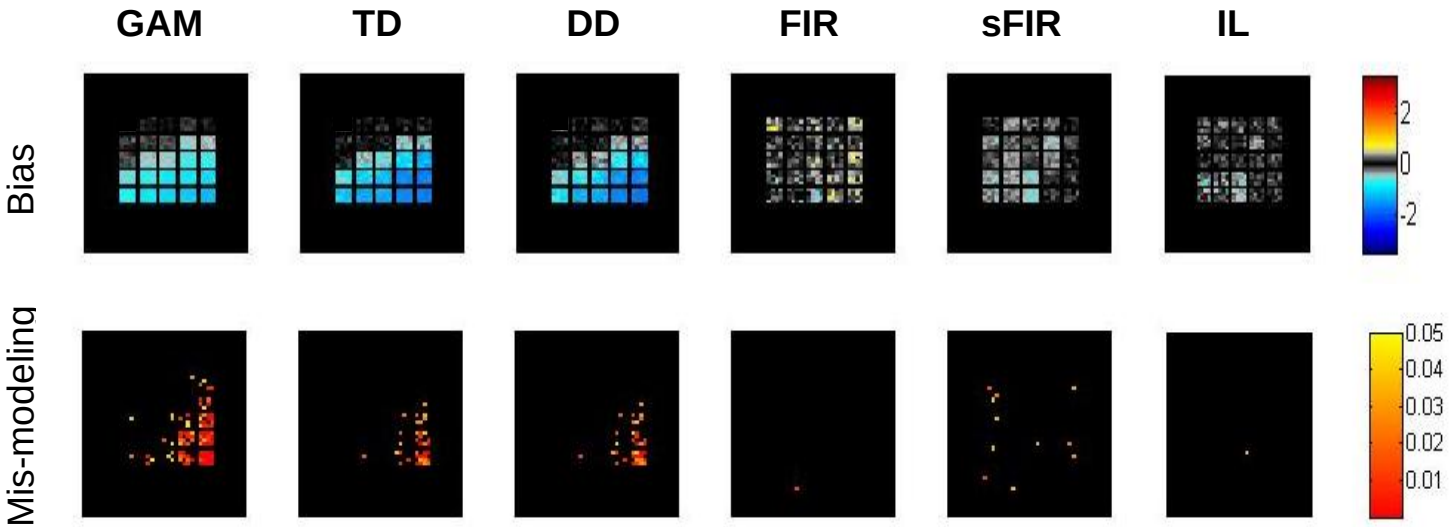




Negative Bias

No Significant Bias

Positive Bias



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.

Thank You!

- Collaborators:
  - Lauren Atlas (Columbia University)
  - Ji Meng Loh (AT&T Labs)
  - Tor Wager (University of Colorado)
- HRF estimation and mismodeling software available in MATLAB.

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