MODEL FITTING

Alexander Huth
@ UBA ECI Course
2017-7-26

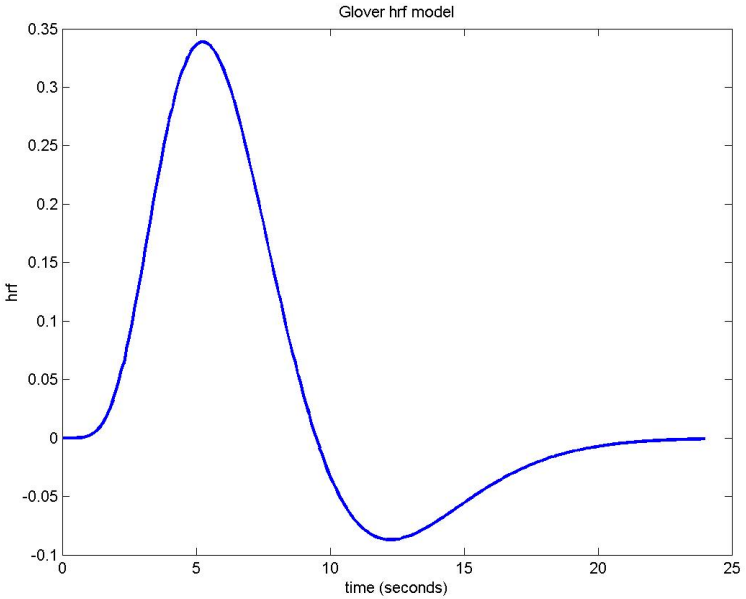
AGENDA

* Spatiotemporal models
  * HRF, separable models, inseparable models
* Regularization
  * Ridge / Tikhonov
  * LASSO

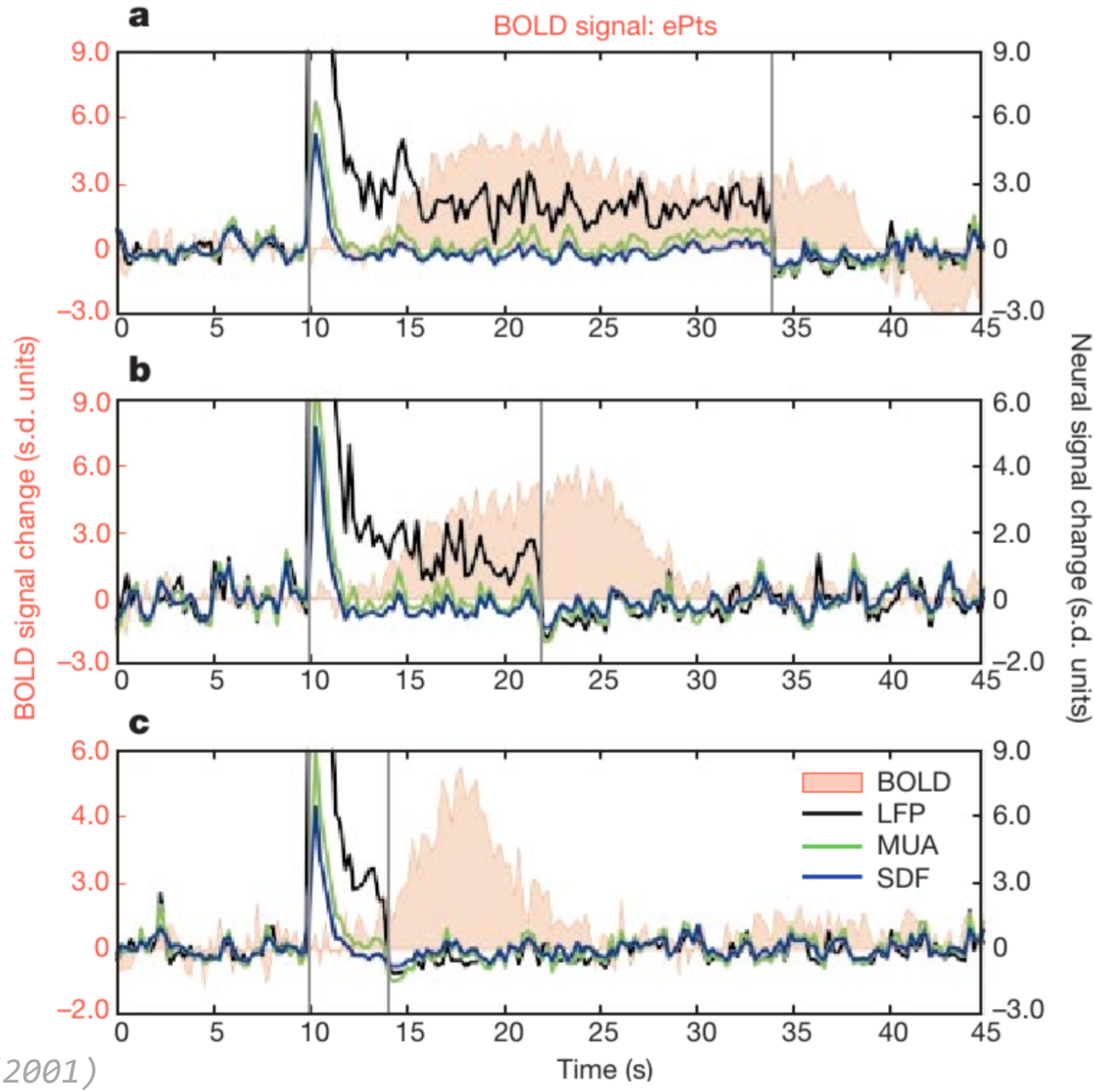
BOLD & HRF

* BOLD = blood-oxygen level dependent

* HRF = hemodynamic response function



typical model HRF



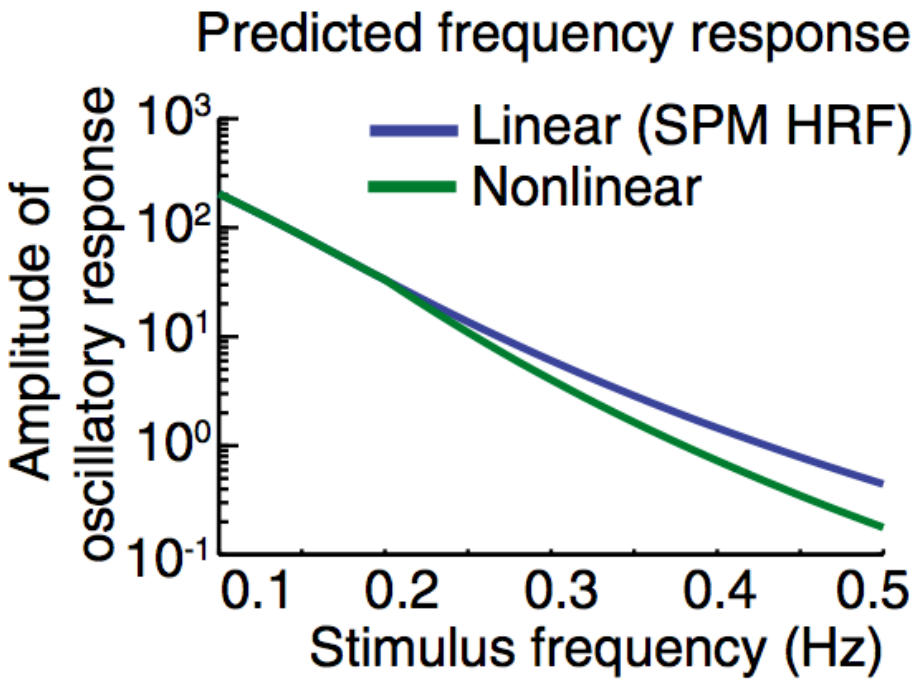
Logothetis et al. (2001)

BOLD & HRF

* HRF is approximately linear

* …but not perfectly

HRF frequency response



Lewis, Setsompop, Rosen, & Polimeni (2016)

SPATIOTEMPORAL MODEL

* HRF needs to be accounted for in model
* Three approaches:
  * assume HRF 👎 👎 👎 🙅‍♀️
  * fit **space-time separable** model
  * fit **space-time inseparable** model

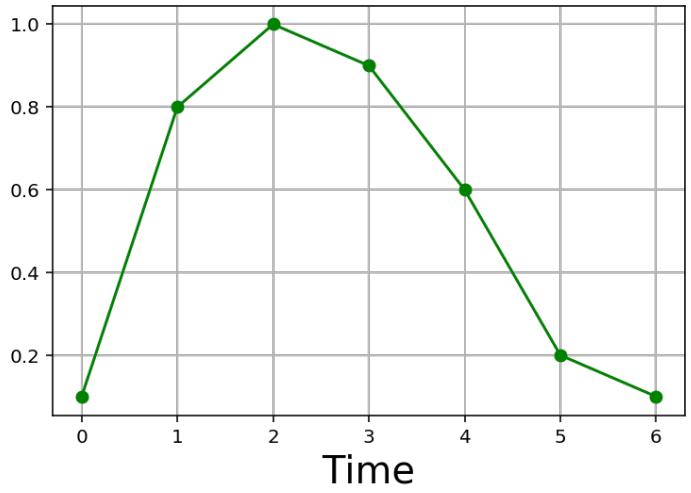
SPACE-TIME SEPARABLE

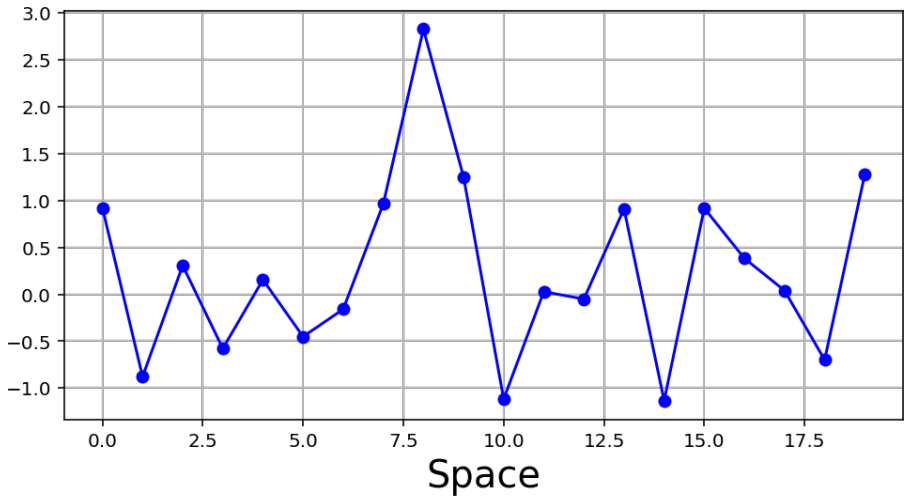
* Model is outer product of one spatial kernel & one temporal kernel

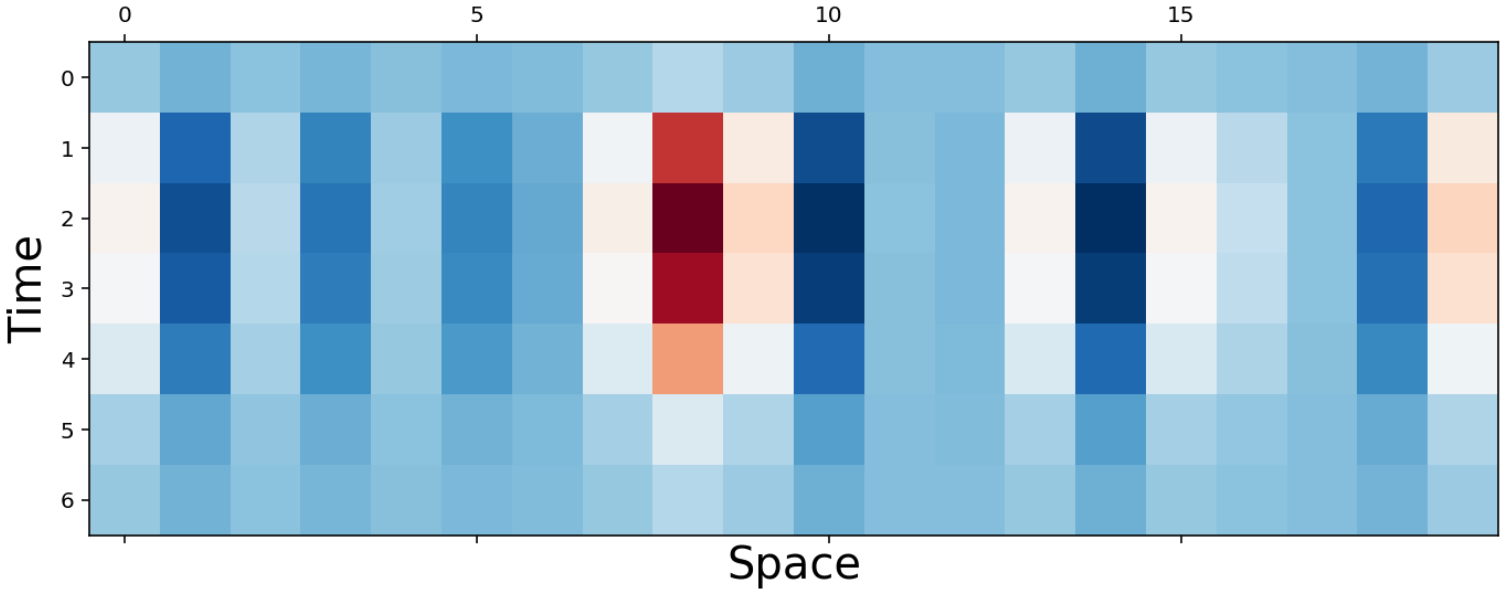
* Fit using e.g. generalized least squares (GLS)

  * Each model (i.e. each voxel) must be fit separately

SPACE-TIME SEPARABLE







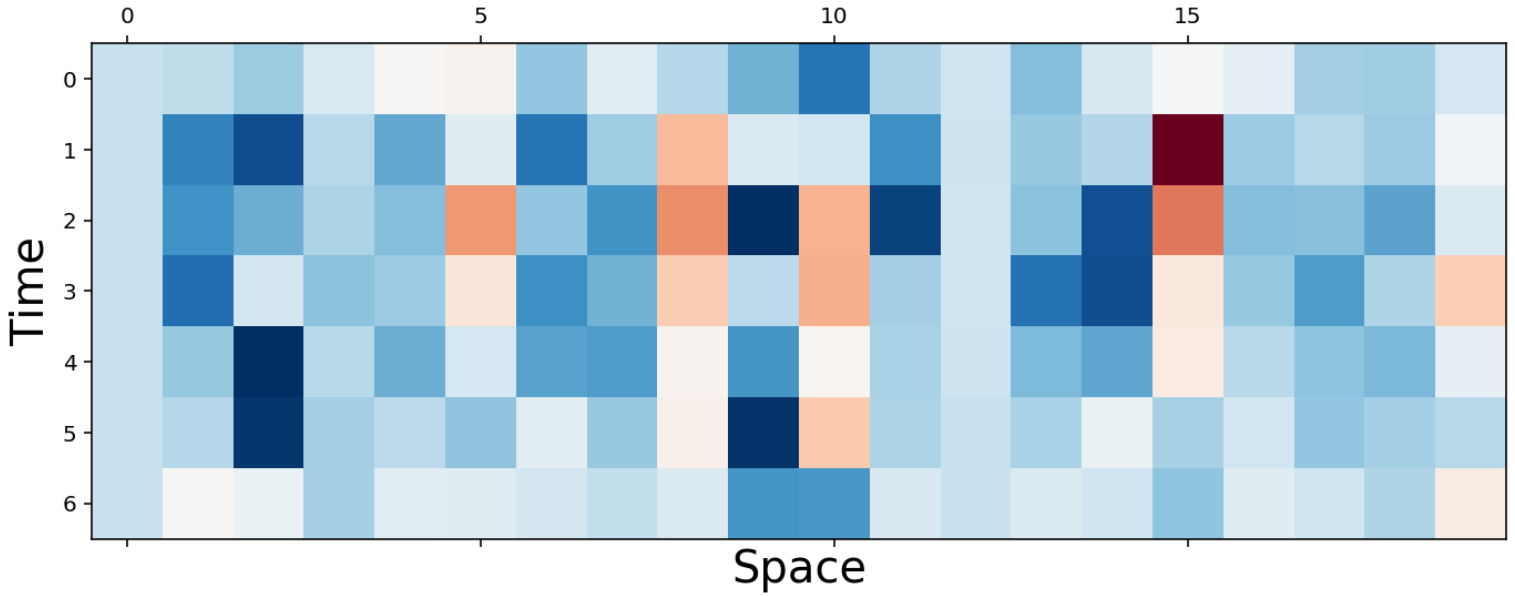
rank=1

SPACE-TIME INSEPARABLE

* Different temporal kernel for each spatial dimensions (i.e. each feature)

* Fit using ordinary methods

SPACE-TIME INSEPARABLE

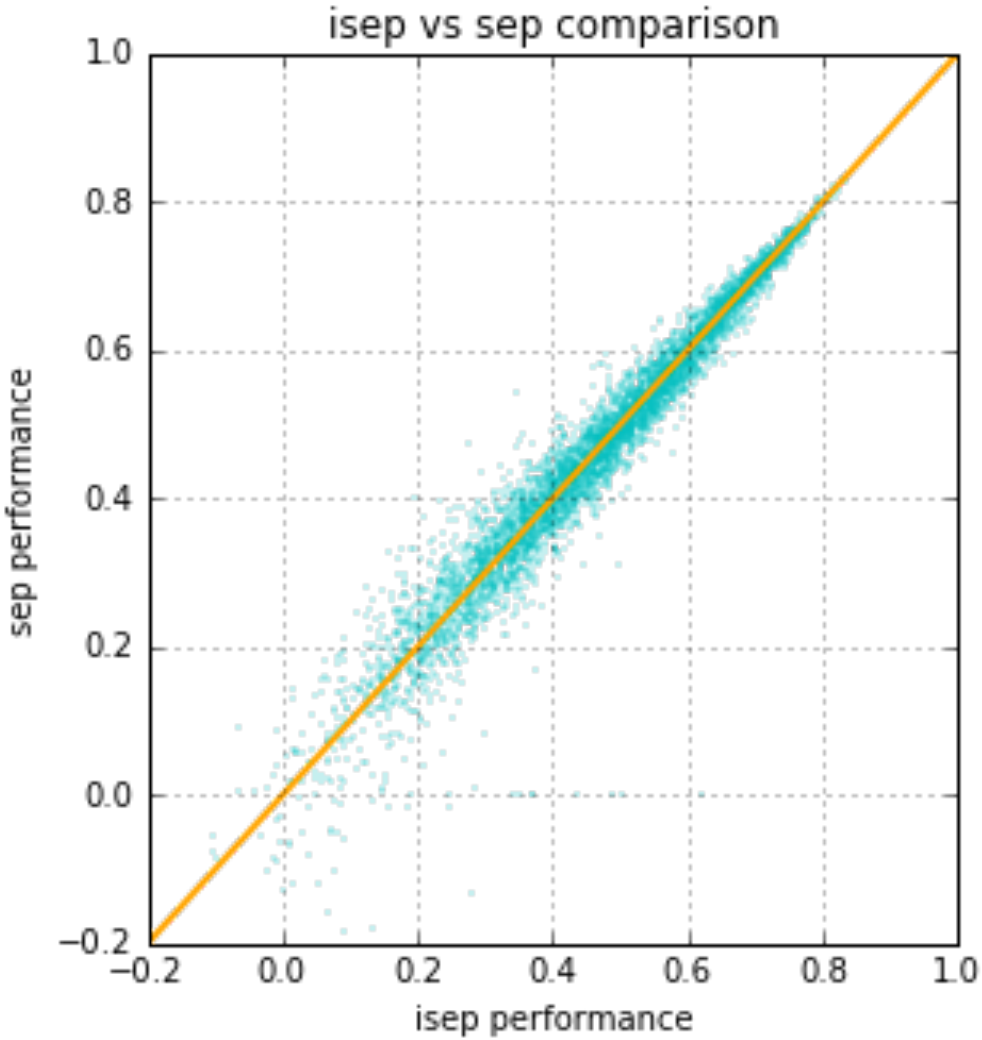


SPATIOTEMPORAL MODEL

* **Separable**
  * Expensive
  * Highly constrained
* **Inseparable**
  * Cheap
  * Unconstrained

SPATIOTEMPORAL MODEL

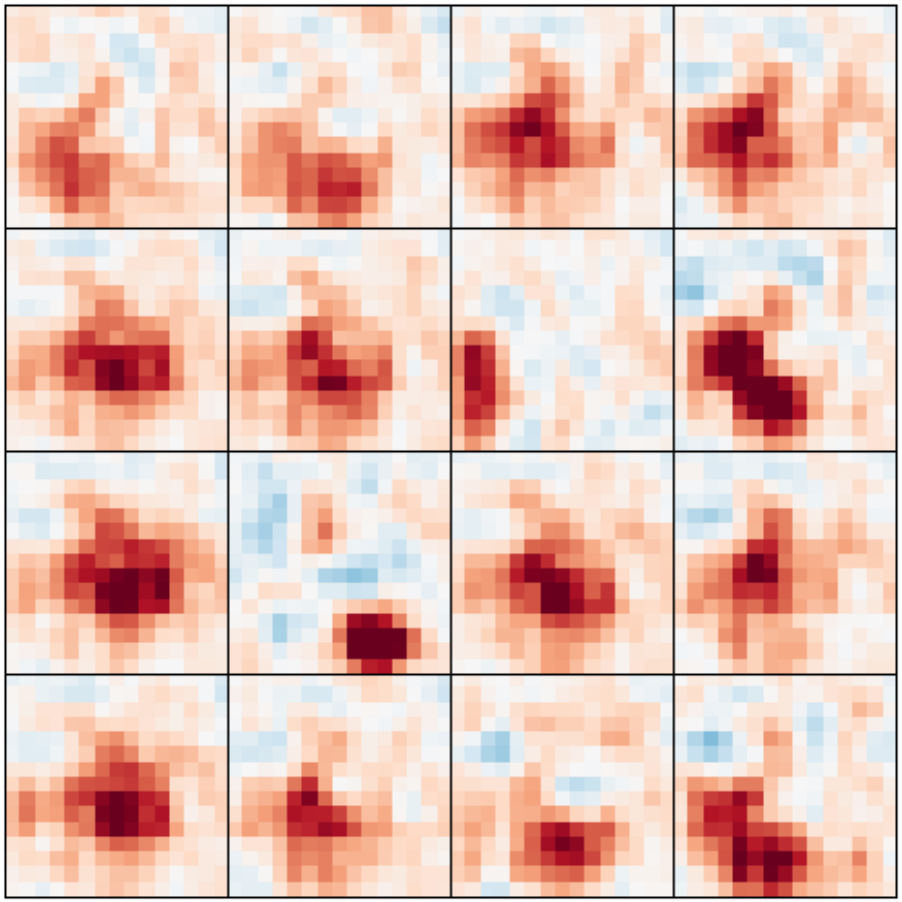
* In practice, both perform very similarly, so I prefer inseparable models



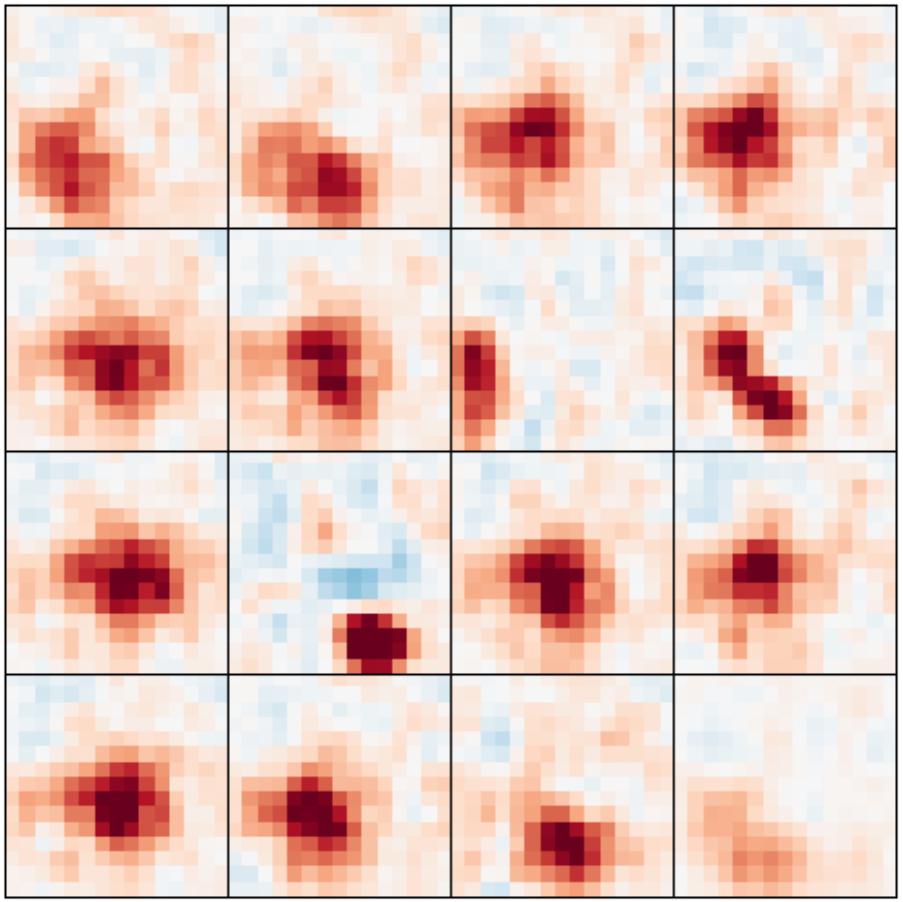
h/t Anwar Nuñez

SPATIOTEMPORAL MODEL

inseparable model weights



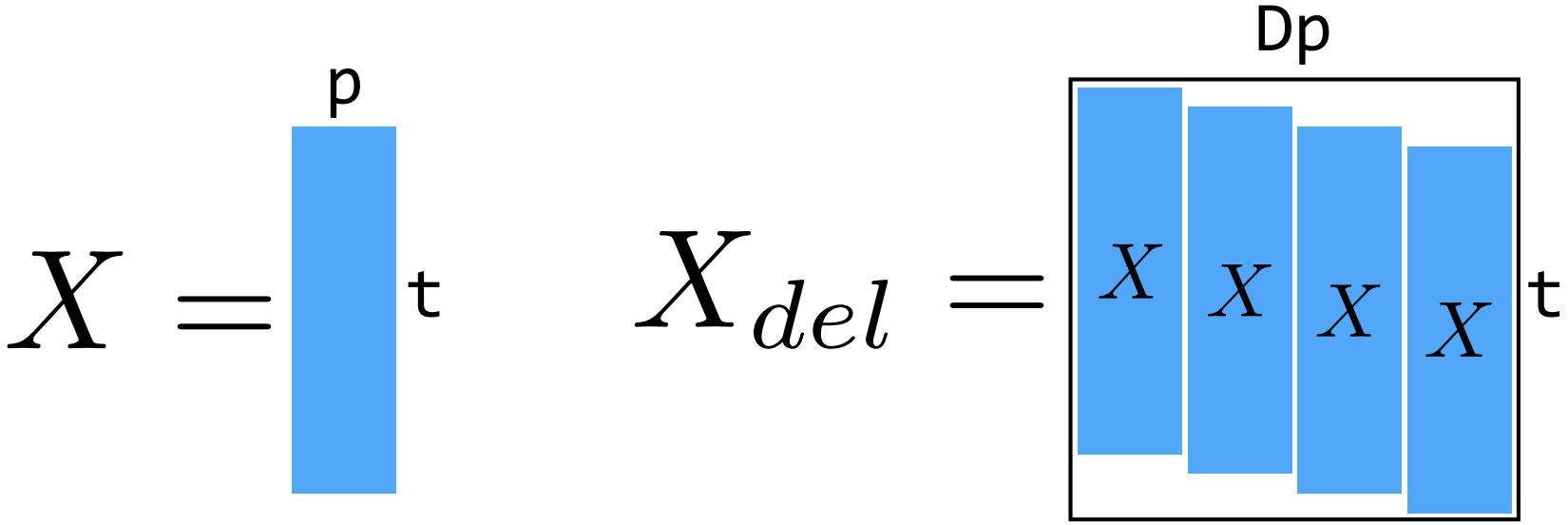
separable model weights



h/t Anwar Nuñez

FINITE IMPULSE RESPONSE (FIR)

* Concatenate delayed copies of the stimulus matrix



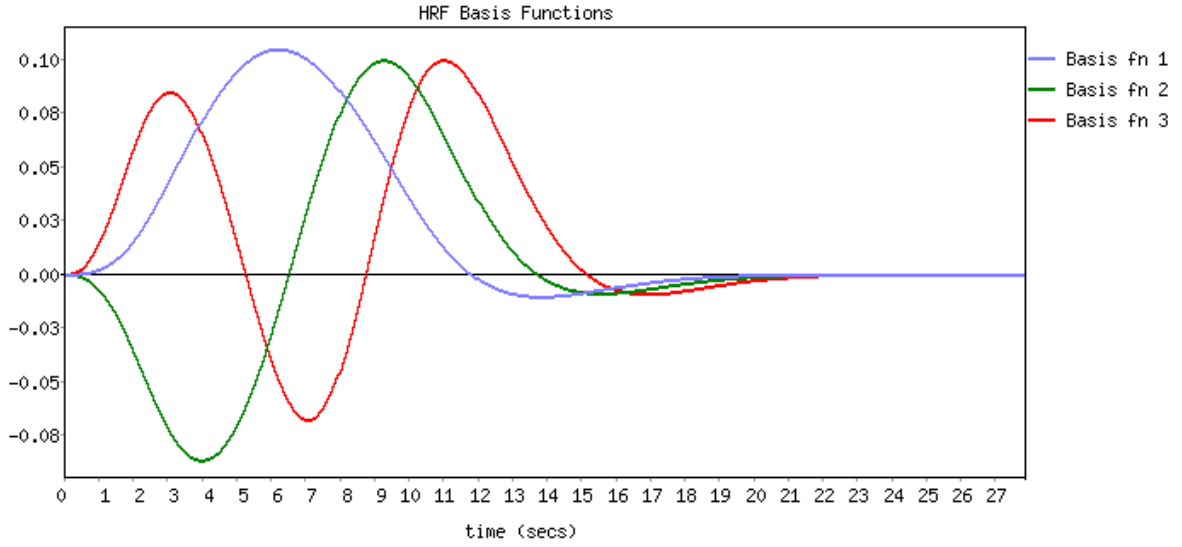
FINITE IMPULSE RESPONSE (FIR)

$$X_{del} = \text{Dp} \overbrace{\begin{bmatrix} X & X & X & X \end{bmatrix}}^{\text{t}} \qquad Y = X_{del} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

$$Y(t) = \beta_1 X(t) + \beta_2 X(t-1) + \beta_3 X(t-2) + \beta_4 X(t-3)$$

HRF BASIS

* Instead of delaying, convolve stimuli with a set of filters



* (Although delaying can be thought of as convolving with impulse delay filters!)

https://surfer.nmr.mgh.harvard.edu/

LINEAR REGRESSION

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

LINEAR REGRESSION

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

Constraining the values **beta** can take improves model performance:

REGULARIZATION

REGULARIZATION

* Regularization can be thought of in three ways:
  * **Prior**
  * **Penalty**
  * **Geometry**

REGULARIZATION AS PRIOR

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

$$Y_{t,j} \sim \mathcal{N}(X\beta, \sigma^2)$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ P(Y|X, \beta)$$

REGULARIZATION AS PRIOR

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

$$Y_{t,j} \sim \mathcal{N}(X\beta, \sigma^2)$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ P(Y|X,\beta)P(\beta)$$

REGULARIZATION AS PENALTY

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

$$E(\beta) = ||Y - X\beta||_2^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ E(\beta)$$

REGULARIZATION AS PENALTY

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$

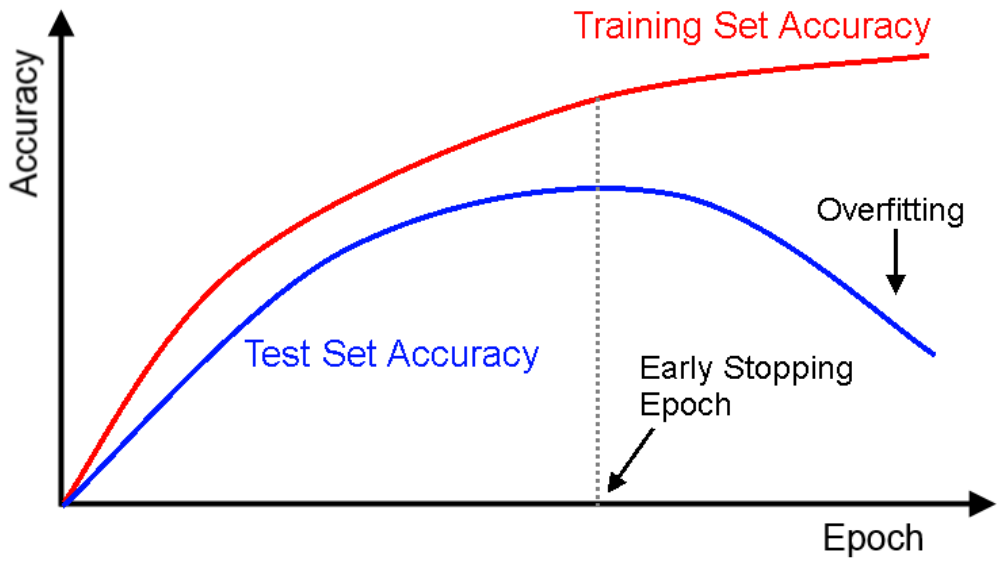
$$E_{pen}(\beta) = ||Y - X\beta||_2^2 + \lambda||\beta||_2^2$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ E_{pen}(\beta)$$

REGULARIZATION AS GEOMETRY - EARLY STOPPING

RESPONSE VARIABLES WEIGHTS NOISE

$$Y = X\beta + \epsilon$$



https://deeplearning4j.org/earlystopping

COMMON TYPES OF REGULARIZATION

* Beta is small (L2-sense) = ridge = gradient descent w/ early stopping

* Beta is small, sparse (L1-sense) = LASSO = coord. descent w/ early stopping

  * Beta is small & sparse (L1+L2 sense) = elastic net

* Beta is sparse (L0-sense) = variable selection

10 MINUTE BREAK

RIDGE REGRESSION

* Multivariate normal (MVN) prior on beta
* L2 penalty on beta
* Gradient descent w/ early stopping

RIDGE REGRESSION

$$Y = X\beta + \epsilon$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[||Y - X\beta||_2^2 + \lambda ||\beta||_2^2 \right]$$

ERROR or LOSS PENALTY

RIDGE REGRESSION

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top Y$$

$$\hat{\beta} = X^{+}_{ridge} Y$$

RIDGE REGRESSION

* Efficient solution with SVD

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top Y$$

$$\text{(SVD)} X = USV^\top \quad D = \frac{S}{S^2 + \lambda^2}$$

$$\hat{\beta} = VDU^\top Y$$

RIDGE REGRESSION

* How to choose lambda?

* GCV - Generalized Cross Validation 👎

* Block-wise cross-validation 👍

RIDGE REGRESSION

* Good implementation: scikit-learn

* Awesome implementation: http://github.com/alexhuth/ridge

TIKHONOV REGRESSION

$$Y = X\beta + \epsilon$$

* **RIDGE REGRESSION**

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[||Y - X\beta||_2^2 + \lambda ||\beta||_2^2 \right]$$

ERROR or LOSS PENALTY

* **TIKHONOV REGRESSION**

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[||Y - X\beta||_2^2 + \lambda ||C\beta||_2^2 \right]$$

PENALTY MATRIX

TIKHONOV REGRESSION

* **RIDGE REGRESSION** is a special case of **TIKHONOV REGRESSION**

* **TIKHONOV REGRESSION** puts a **ZERO-MEAN MULTIVARIATE NORMAL PRIOR** on the weights

* in **RIDGE REGRESSION** the covariance matrix of the prior has a constant diagonal

  * i.e. the prior is a **SPHERE**

* in **TIKHONOV REGRESSION** the covariance matrix can be ***ANYTHING***

TIKHONOV REGRESSION

* the multivariate normal prior given by **TIKHONOV REGRESSION**

$$\beta \sim N(0, \sigma^2 (C^T C)^{-1})$$

TIKHONOV REGRESSION

* any **TIKHONOV** problem can be converted into a **RIDGE** problem

$$A = XC^{-1}$$ ← 1. CHANGE OF BASIS

$$\hat{\beta}_A = \underset{\beta}{\operatorname{argmin}} \left[||Y - A\beta||_2^2 + \lambda ||\beta||_2^2 \right]$$

2. RIDGE REGRESSION

$$\hat{\beta} = C^{-1}\hat{\beta}_A$$ ← 3. CHANGE BASIS AGAIN

Rank-Deficient and Discrete Ill-Posed Problem: Numerical Aspects of Linear Inversion (1998; Per Christian Hansen)

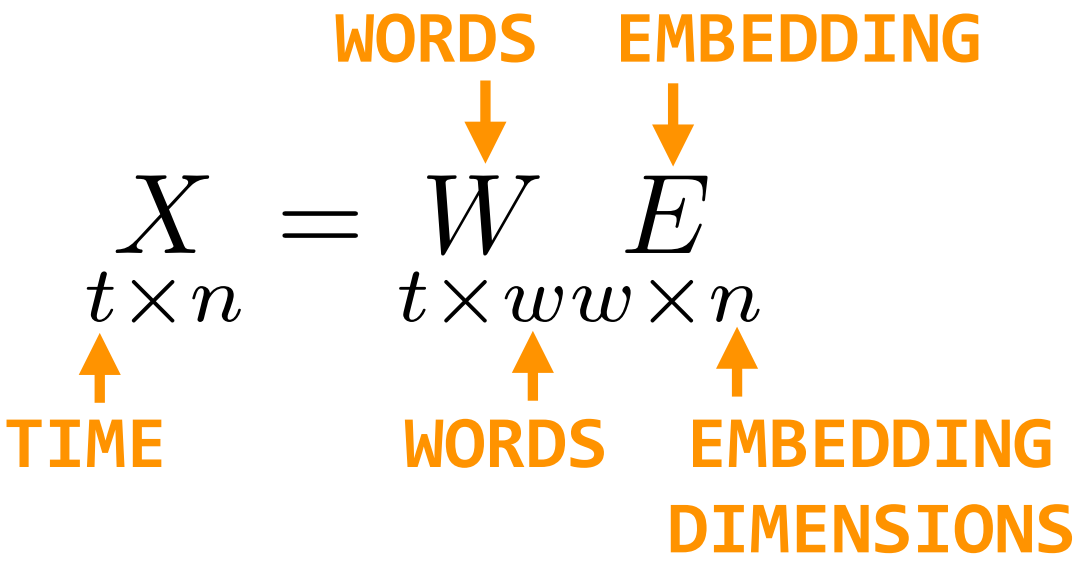
TIKHONOV REGRESSION

* conversely, **ANY LINEAR TRANSFORMATION** of X followed by **RIDGE REGRESSION** is equivalent to some **TIKHONOV REGRESSION** problem

TIKHONOV REGRESSION

* **WORD EMBEDDING MODELS**

* think of stimulus matrix as **WORDS** over time projected onto **WORD EMBEDDING**



TIKHONOV REGRESSION

* this is equivalent to **TIKHONOV REGRESSION** on the **WORDS** with a prior determined by the **WORD EMBEDDING**

$$\frac{1}{\sigma^2}\Sigma_\beta = (C^T C)^{-1} = E^T E$$

PRIOR COVARIANCE

INVERSE OF PENALTY INNER PRODUCT

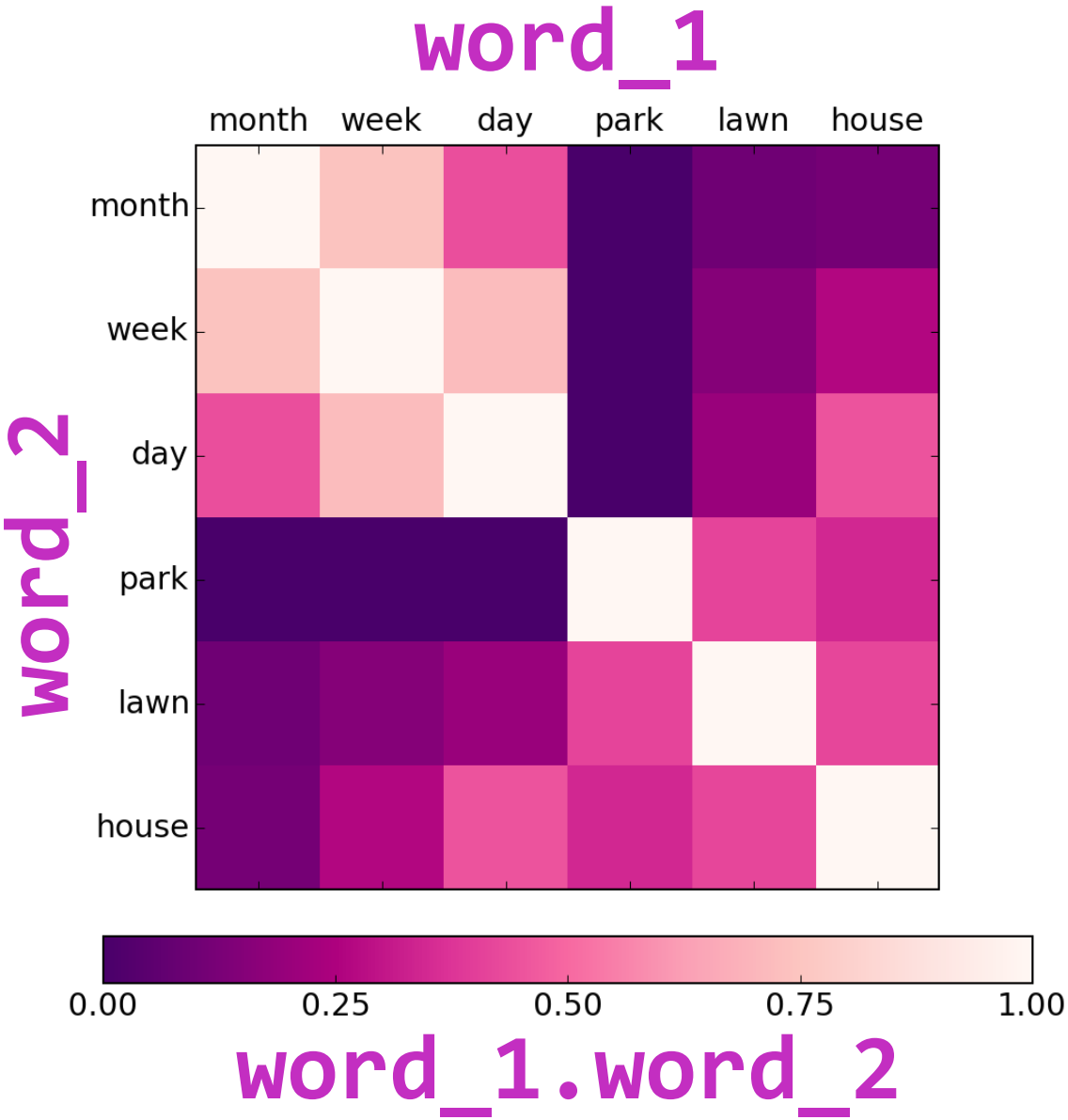
EMBEDDING INNER PRODUCT

* **i.e. the prior covariance between two words' weights is equal to the dot product of their embedding vectors**

TIKHONOV REGRESSION

$$E^T E =$$

EMBEDDING INNER PRODUCT, ***english1000***



TIKHONOV REGRESSION

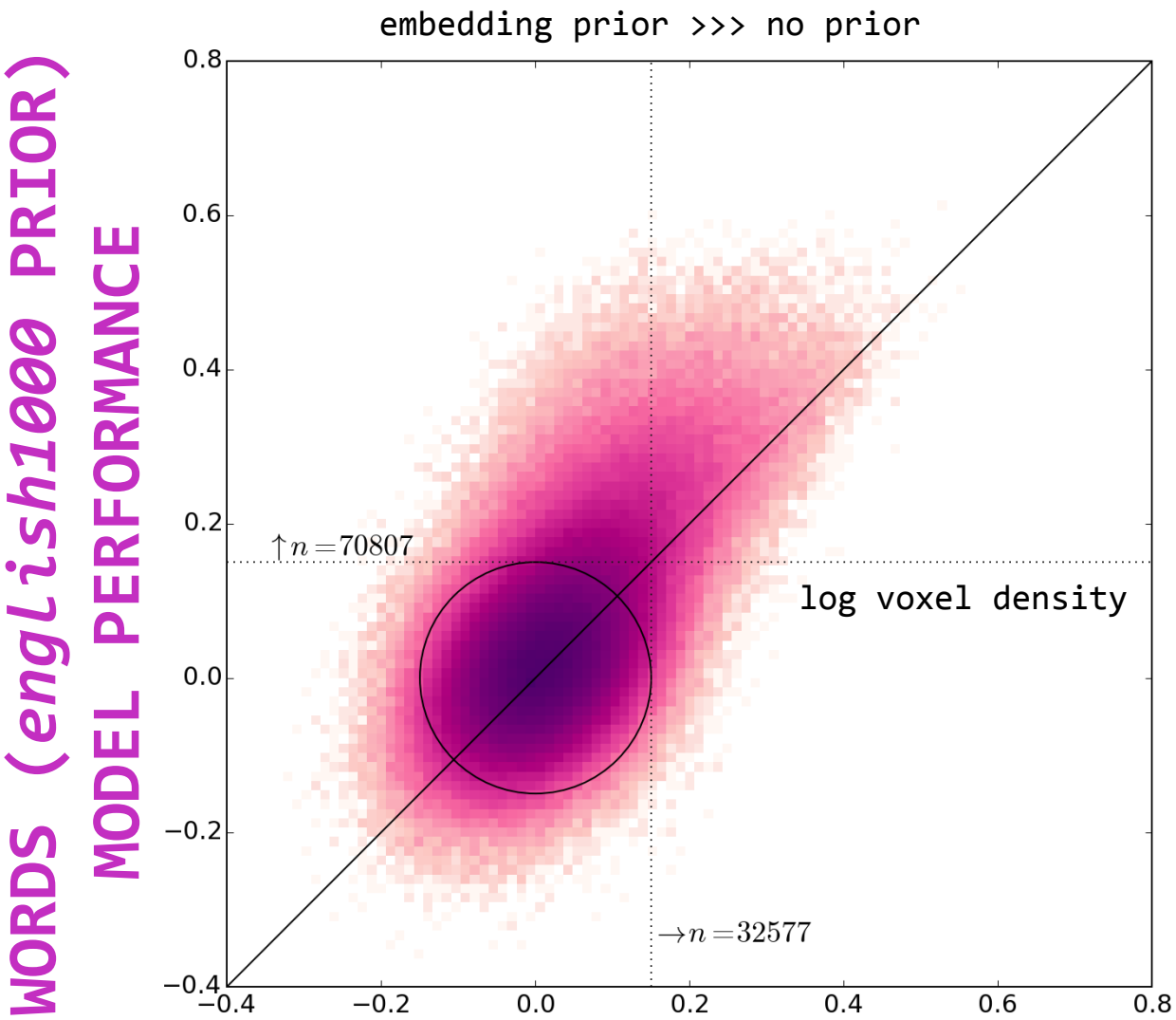
* to get **WEIGHTS ON WORDS** we just project onto the **EMBEDDING**

WEIGHTS IN WORD SPACE — EMBEDDING — WEIGHTS IN EMBEDDING SPACE

$$\underset{w\times v}{\hat{\beta}_W} = \underset{w\times n}{E}\ \underset{n\times v}{\hat{\beta}_X}$$

* (this is equivalent to simulating responses to single words)

TIKHONOV REGRESSION



LASSO

* Laplacian prior on beta_i
* L1 penalty on beta
* Coordinate descent w/ early stopping

LASSO

$$Y = X\beta + \epsilon$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right]$$

ERROR or LOSS PENALTY

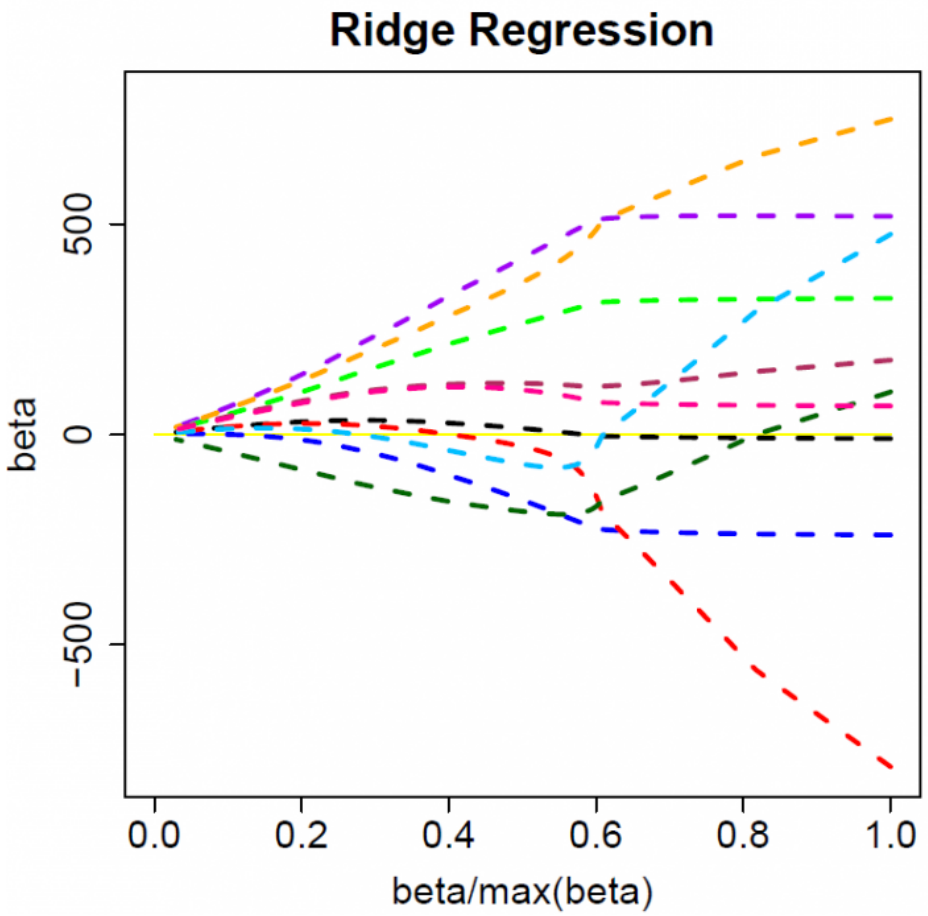
LASSO

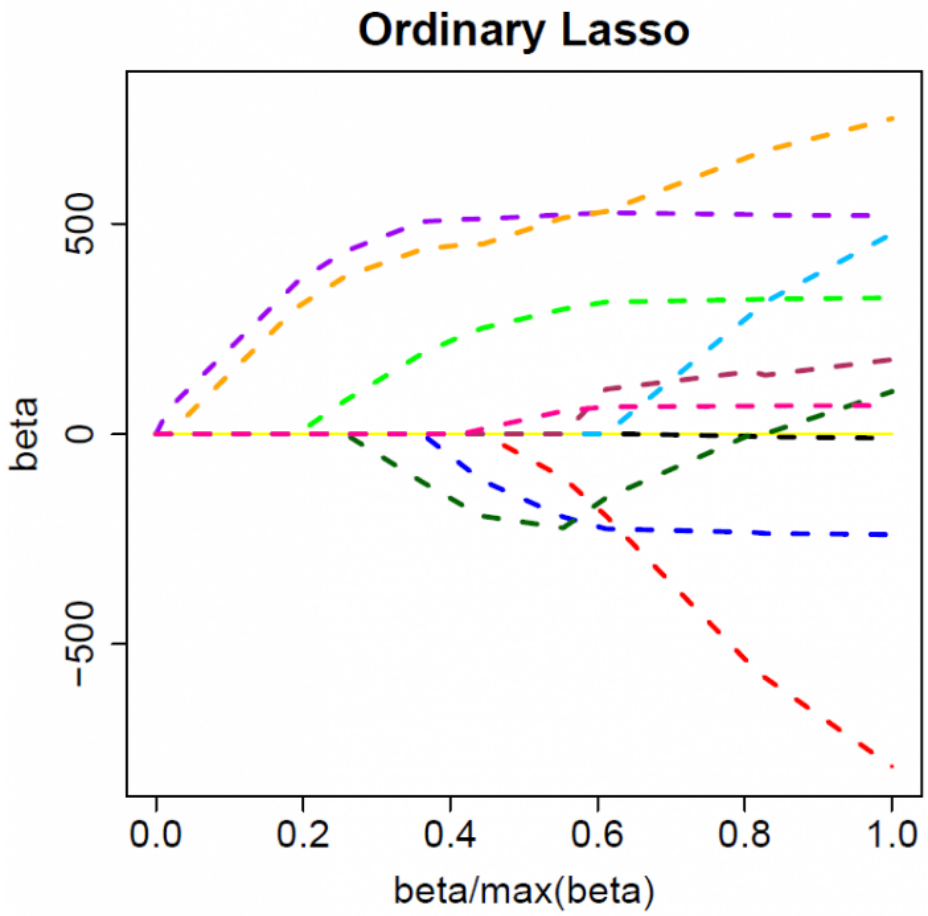
* No closed form solution

* Solved via coordinate descent, LARS (least-angle regression) or other methods

* *Sssssllllloooooowwwww……*

LASSO





h/t https://onlinecourses.science.psu.edu/stat857/node/158

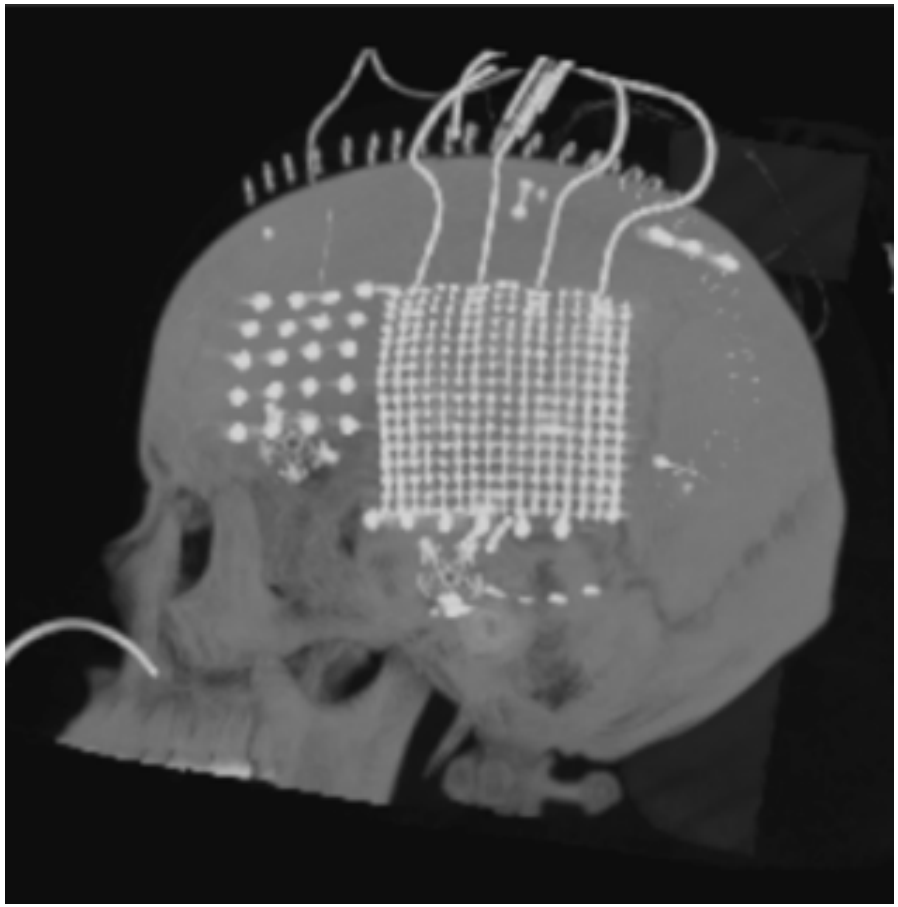
OTHER METHODS

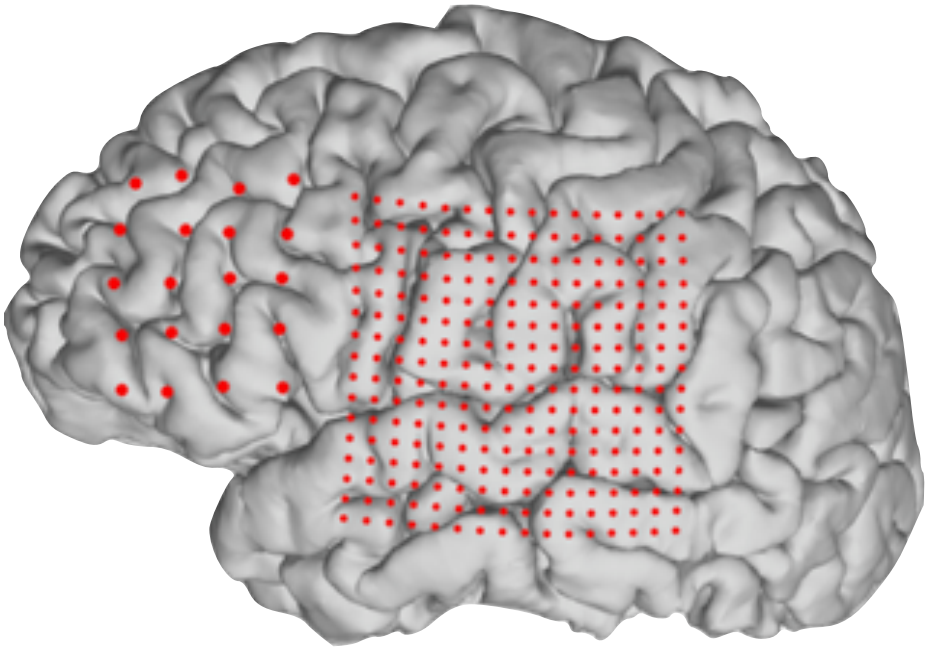
* Neural networks

* Random forests

* Feature selection (~L0-norm)

EXAMPLE - ECOG

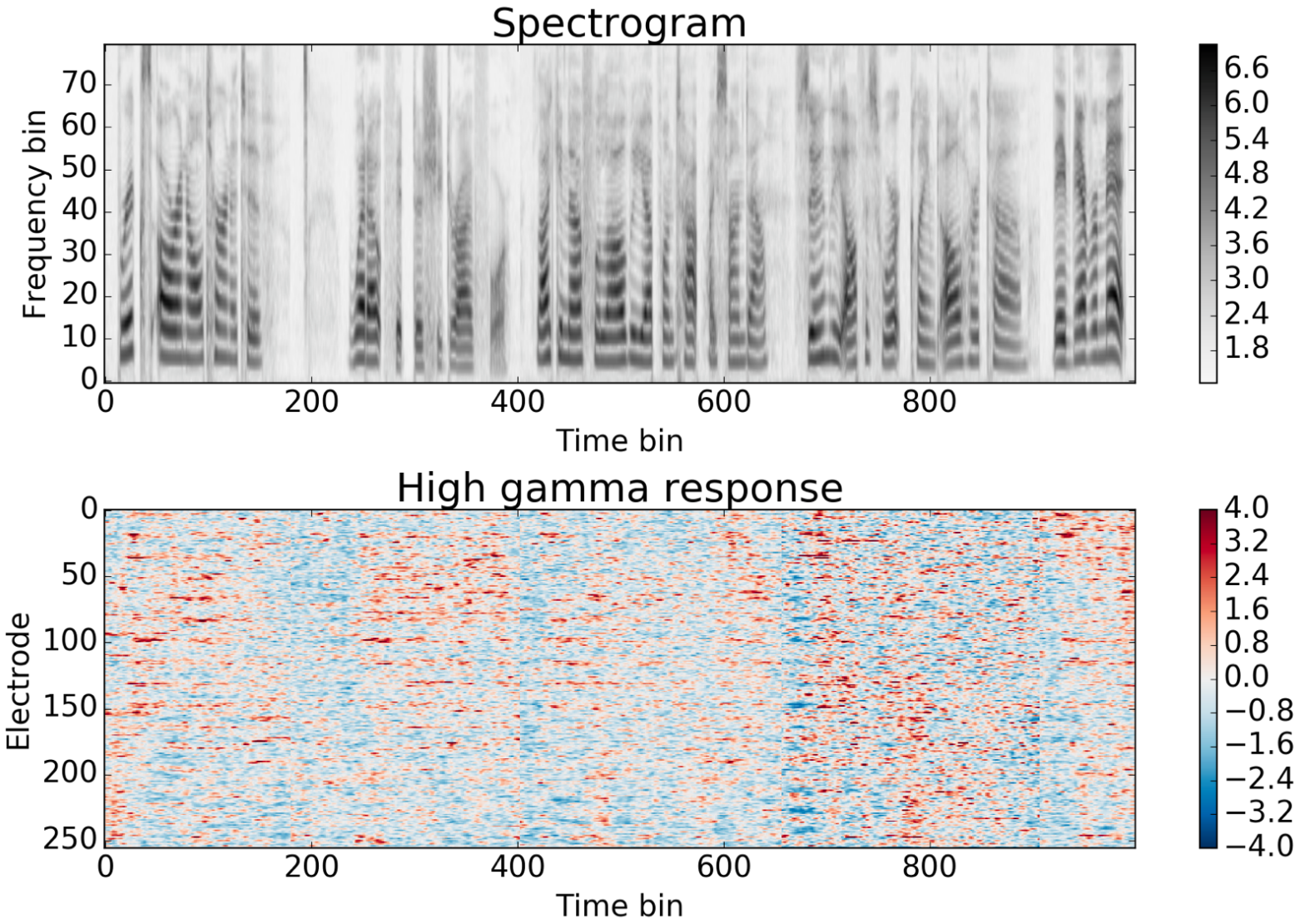




data from Hamilton et al. (2017) biorXiv

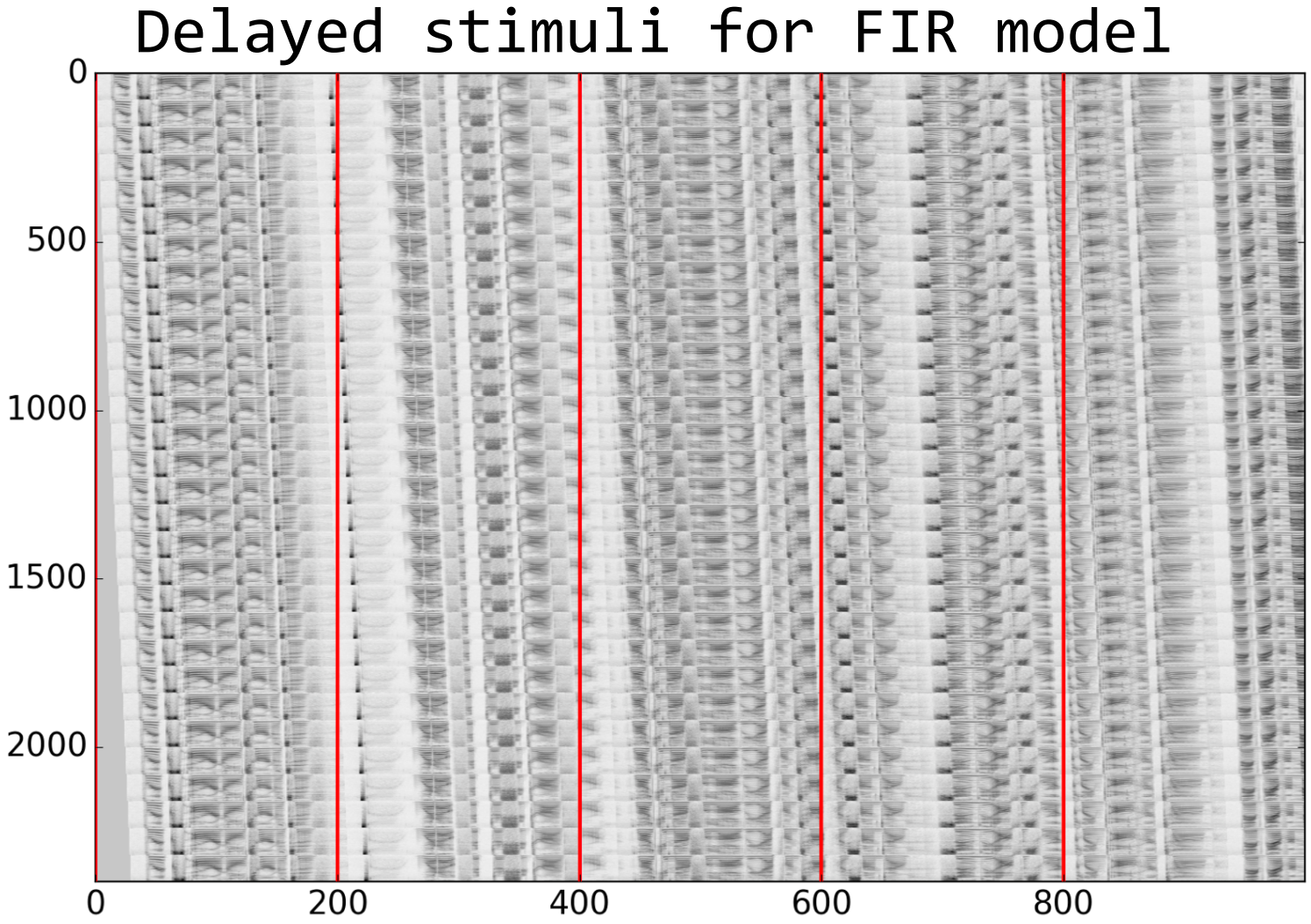
h/t Liberty Hamilton

EXAMPLE - ECOG



h/t Liberty Hamilton

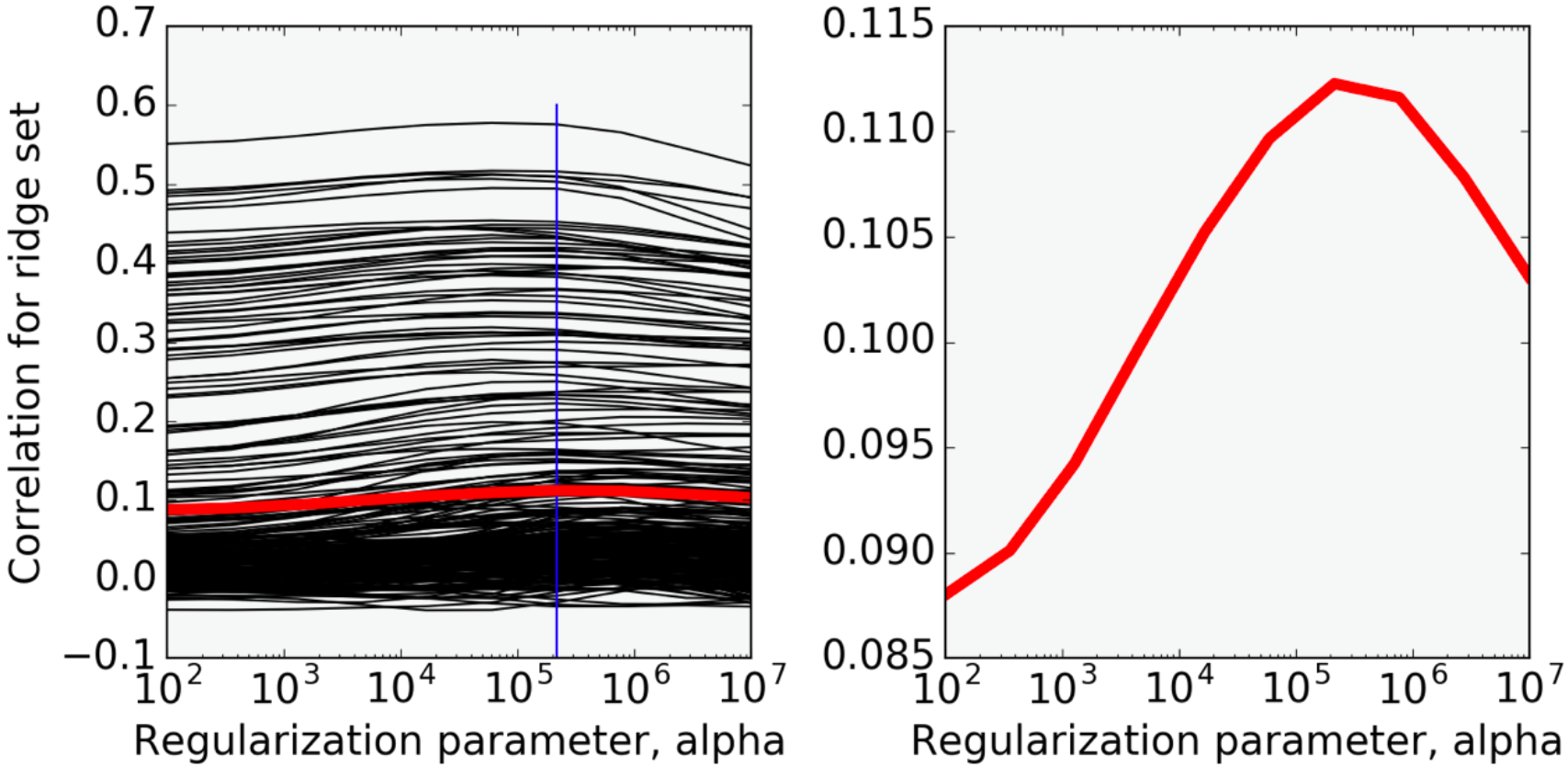
EXAMPLE - ECOG



h/t Liberty Hamilton

EXAMPLE - ECOG

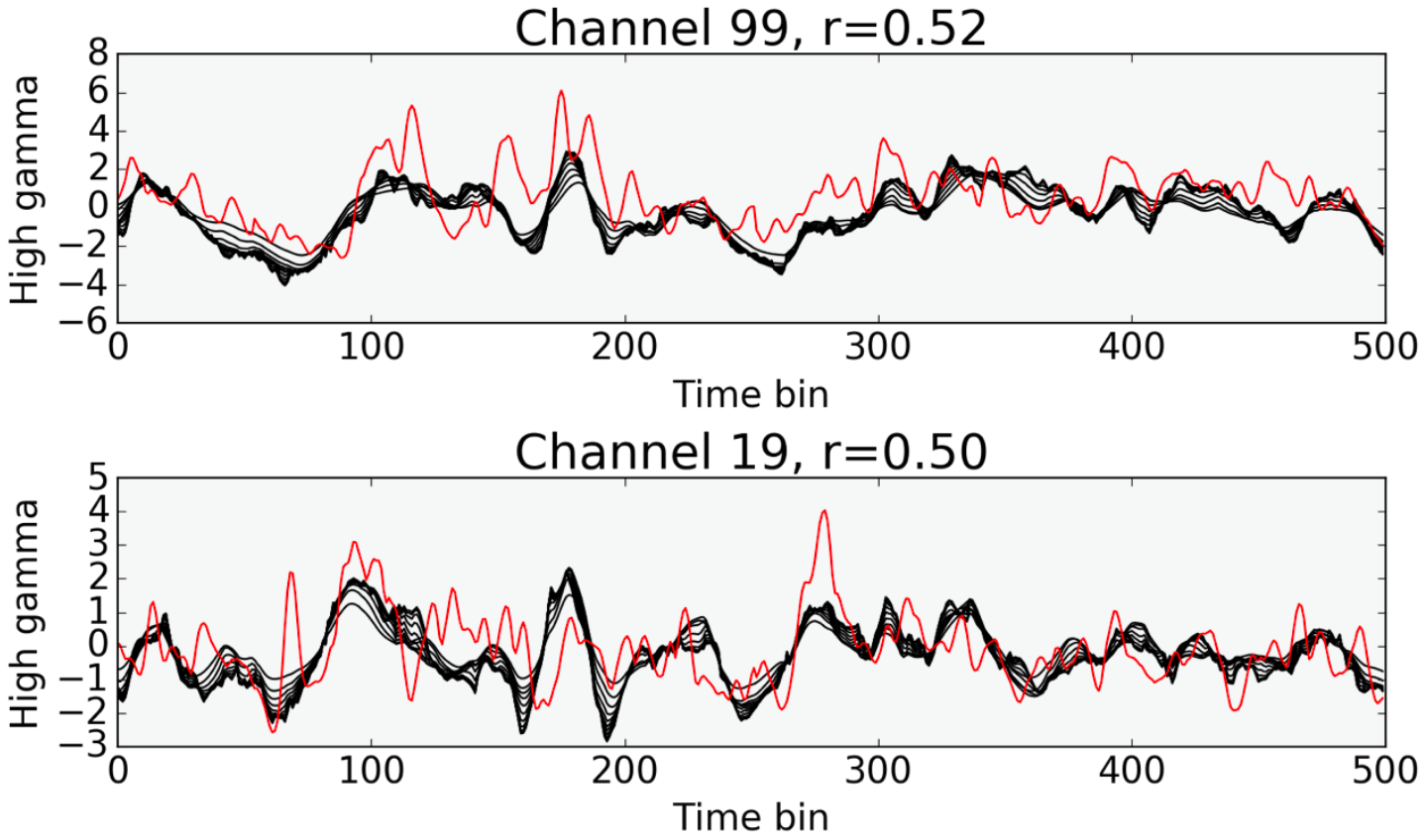
Cross-validation to choose ridge parameter



h/t Liberty Hamilton

EXAMPLE - ECOG

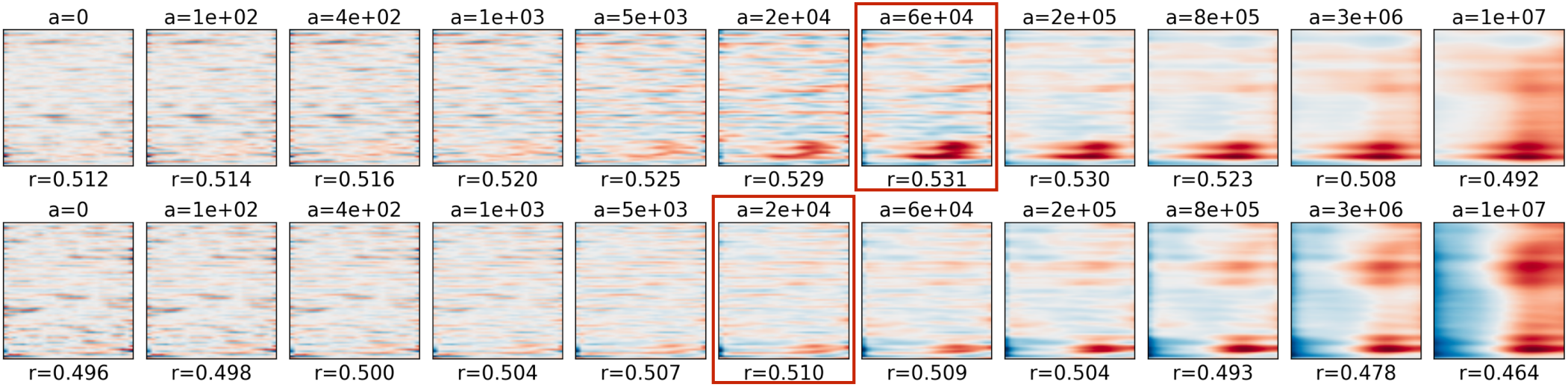
Predicted vs. actual responses
as function of ridge parameter



h/t Liberty Hamilton

EXAMPLE - ECOG

Receptive fields as function of ridge parameter



h/t Liberty Hamilton

THANKS