Feature-space selection in voxelwise encoding models with banded ridge regression

2021-03-02

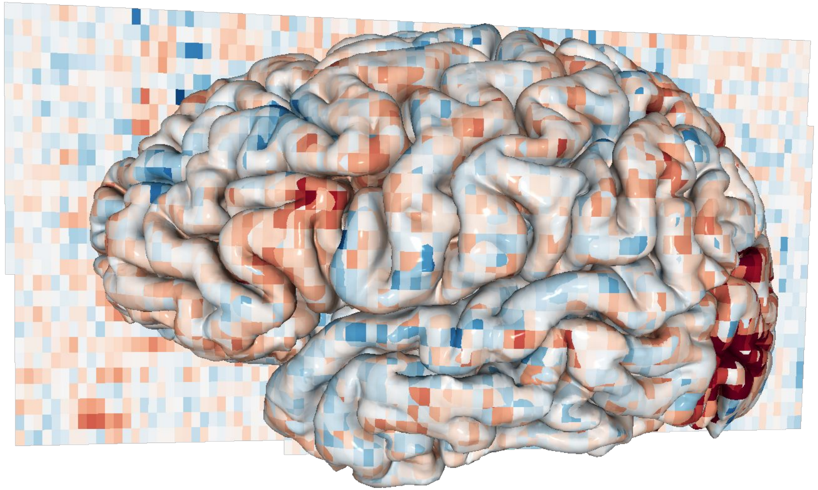
Tom Dupré la Tour
with Jack Gallant - UC Berkeley

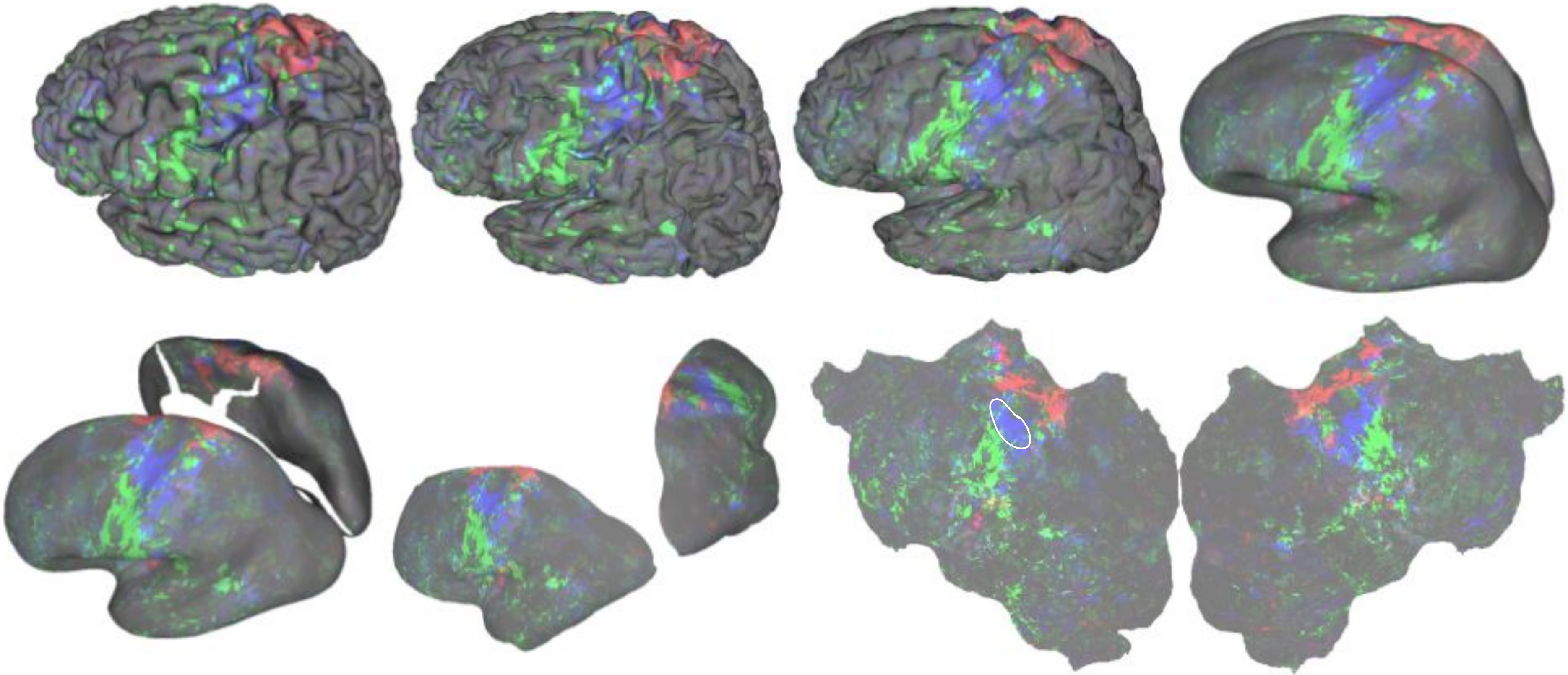
Voxelwise encoding models

Functional magnetic resonance imaging (fMRI)

Record brain activity in volume (spatial resolution: 3x3x2 mm^3, time resolution: 2 s)

example dataset: 80000 voxels, 3600 time points (1 GB)



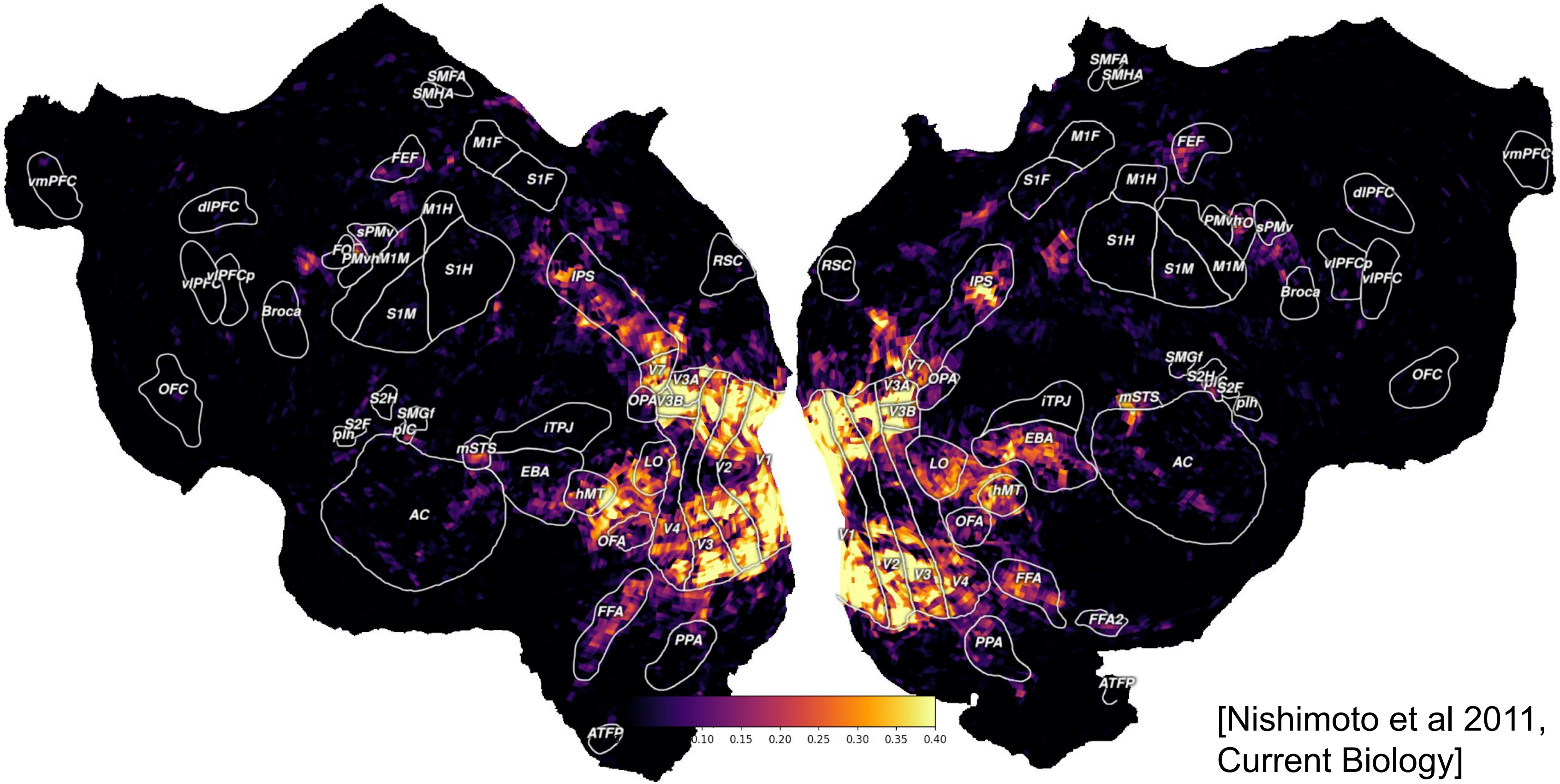


Somatotopy: foot (R), mouth (G), hand (B)

Voxelwise encoding models

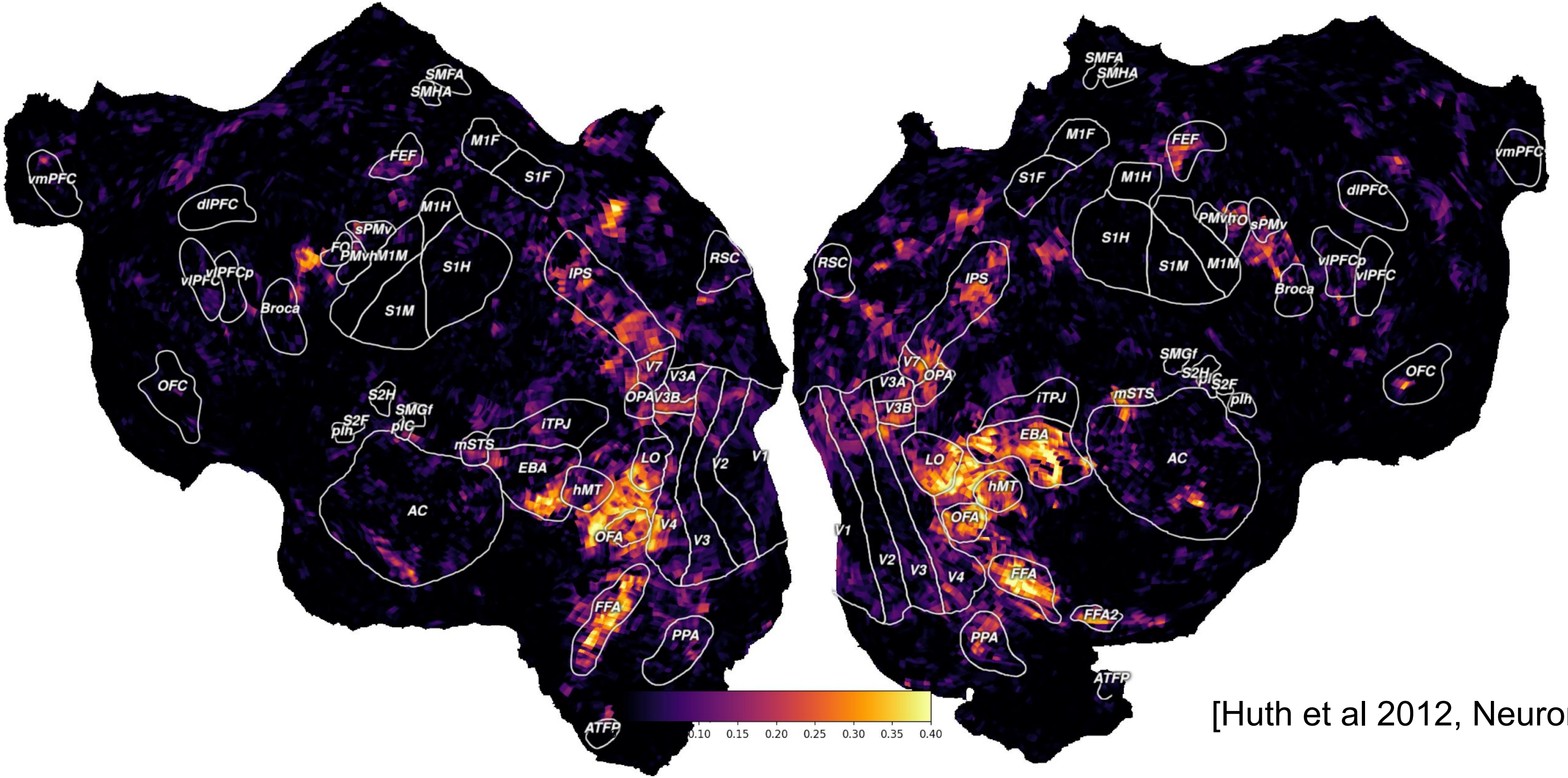
- Use naturalistic stimulus/task (movies, podcasts, driving simulator, etc.)
- Encode the stimulus/task into features X
- Linearly model each voxel activity $y = X.b$
- Quantify predictive power on a separate test set $R^2(y_{test}, X_{test}b)$
- Interpret R^2 scores and weights b

Magnitude of spatio-temporal filters on a movie (viewing)



[Nishimoto et al 2011, Current Biology]

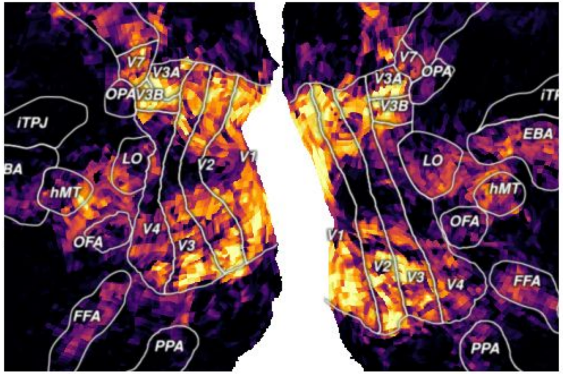
Semantic categories of objects in a movie (viewing)



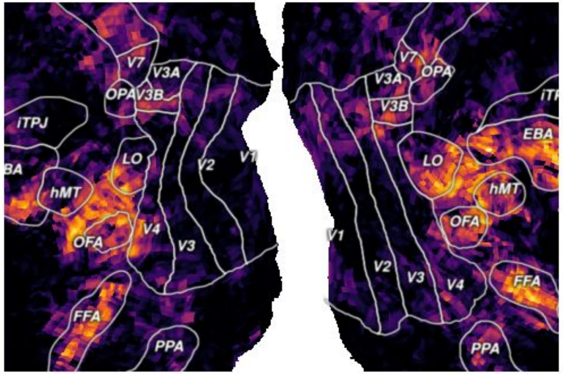
[Huth et al 2012, Neuron]

Model comparison

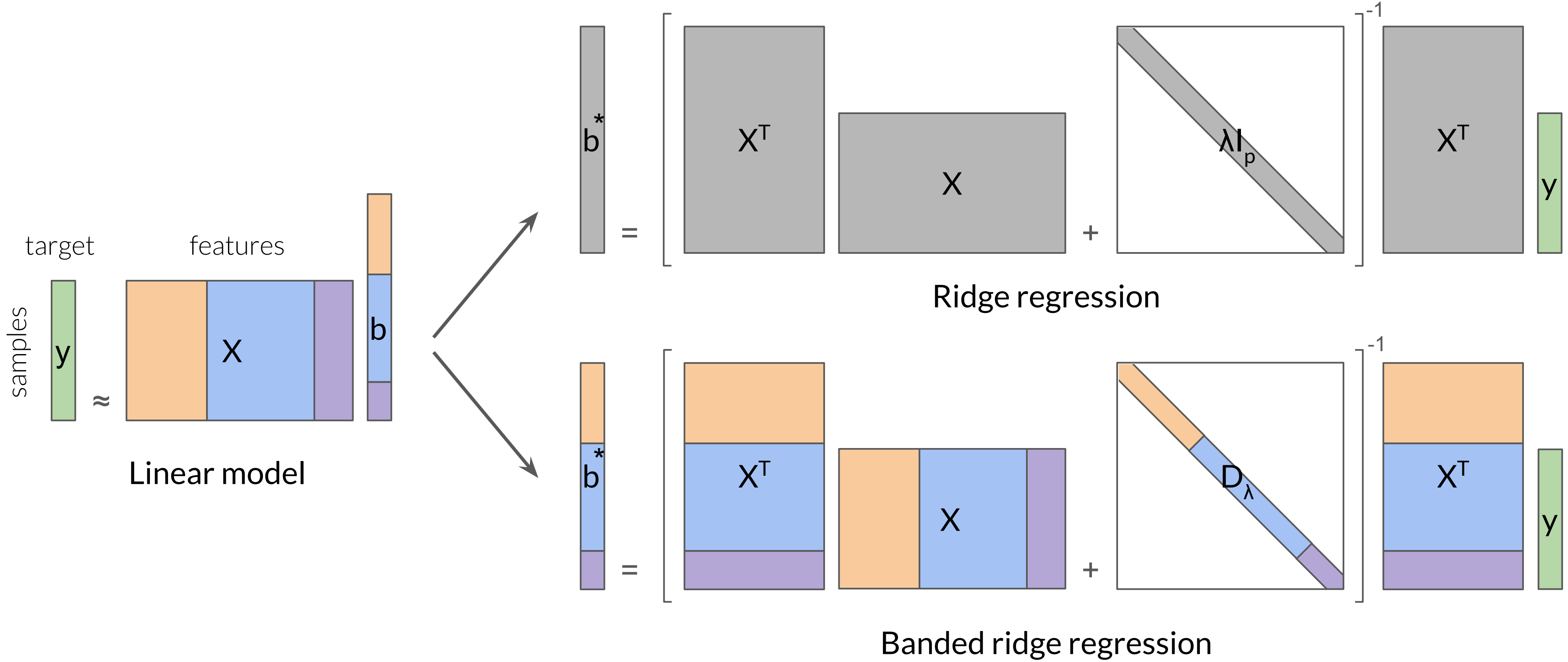
Magnitude of
spatio-temporal filters



Semantic categories
of objects



Banded ridge regression



target
features
samples
y
≈
X
b
Linear model
b*
=
XT
X
+
λIp
-1
XT
y
Ridge regression
b*
=
XT
X
+
Dλ
-1
XT
y
Banded ridge regression

Banded ridge regression

Ridge regression [Hoerl and Kennard, 1970]

$$b^* \in \mathbb{R}^p$$
$$X \in \mathbb{R}^{n \times p}$$
$$y \in \mathbb{R}^n$$
$$\lambda > 0$$

$$b^* = \underset{b}{\operatorname{argmin}} \|Xb - y\|_2^2 + \lambda \|b\|_2^2$$

Banded ridge regression [Nunez-Elizalde et al, 2019]

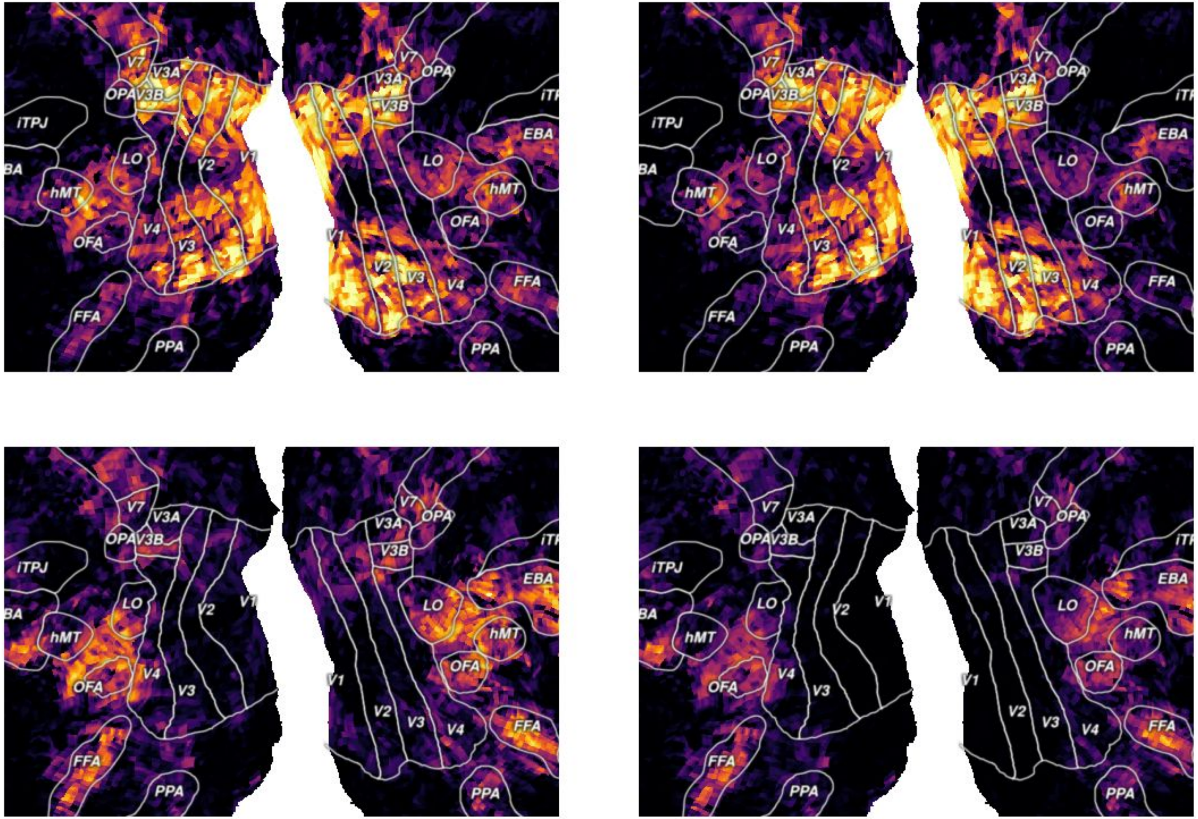
$$b^* = \underset{b}{\operatorname{argmin}} \left\| \sum_{i=1}^{m} X_i b_i - y \right\|_2^2 + \sum_{i=1}^{m} \lambda_i \|b_i\|_2^2$$

Model comparison

Magnitude of spatio-temporal filters

Semantic categories of objects

Ridge (separate models) | Banded ridge (joint model)



A feature-space selection mechanism

$$b^* = \underset{b}{\operatorname{argmin}} \left\| \sum_{i=1}^{m} X_i b_i - y \right\|_2^2 + \sum_{i=1}^{m} \lambda_i \|b_i\|_2^2$$

A large regularization λ_i leads to the feature space i being unused.
The cross-validation can discard the non-predictive or redundant feature space.

Banded ridge regression leads to (soft) sparsity at the feature-space level.
(see related models later)

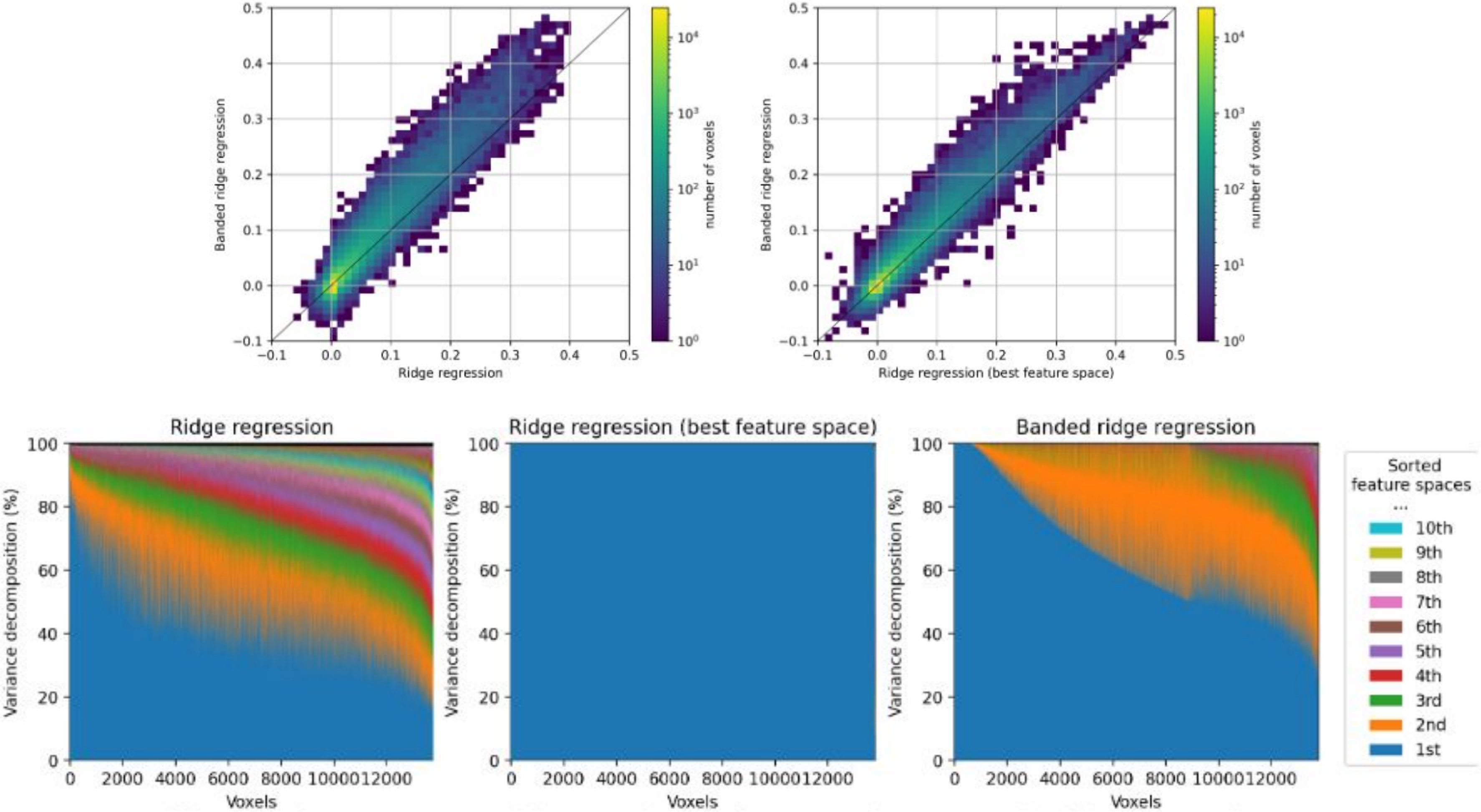
Benchmark on an actual Gallant-lab use case

Dataset: narrative short films watched without fixation

- n_samples = 3572
- n_voxels = 85483
- n_feature_spaces = 22
  - motion energy, visual/speech semantic, spectrogram, phonemes, ...

Models

- Banded ridge regression
- Ridge regression
- Ridge regression (best feature-space)



Banded ridge regression
Ridge regression
number of voxels
Ridge regression (best feature space)
Ridge regression
Ridge regression (best feature space)
Banded ridge regression
Variance decomposition (%)
Voxels
Sorted feature spaces
...
10th
9th
8th
7th
6th
5th
4th
3rd
2nd
1st

Related models

$$K = XX^\top \in \mathbb{R}^{n \times n}$$

Kernel formulation

Ridge

[Hoerl and Kennard, 1970]

$$b^* = \underset{b}{\operatorname{argmin}} \|Xb - y\|_2^2 + \lambda \|b\|_2^2$$

Kernel ridge

[Saunders et al., 1998]

$$w^* = \underset{w}{\operatorname{argmin}} \|Kw - y\|_2^2 + \lambda w^\top K w$$

Banded ridge

[Nunez-Elizalde et al, 2019]

$$b^* = \underset{b}{\operatorname{argmin}} \left\| \sum_{i=1}^{m} X_i b_i - y \right\|_2^2 + \sum_{i=1}^{m} \lambda_i \|b_i\|_2^2$$

Multiple-kernel ridge

[Bach 2004]

$$w^* = \underset{w}{\operatorname{argmin}} \left\| \sum_{i=1}^{m} \gamma_i K_i w - y \right\|_2^2 + \mu w^\top \sum_{i=1}^{m} \gamma_i K_i w$$

Related models with group-sparsity

Multiple-kernel learning [Lanckriet et al., 2004, Bach et al., 2004]

Group lasso [Yuan & Lin 2006] $\|b\|_{2,1} = \sum_{i=1}^{m} \|b_i\|_2$

Squared group lasso [Bach 2004]

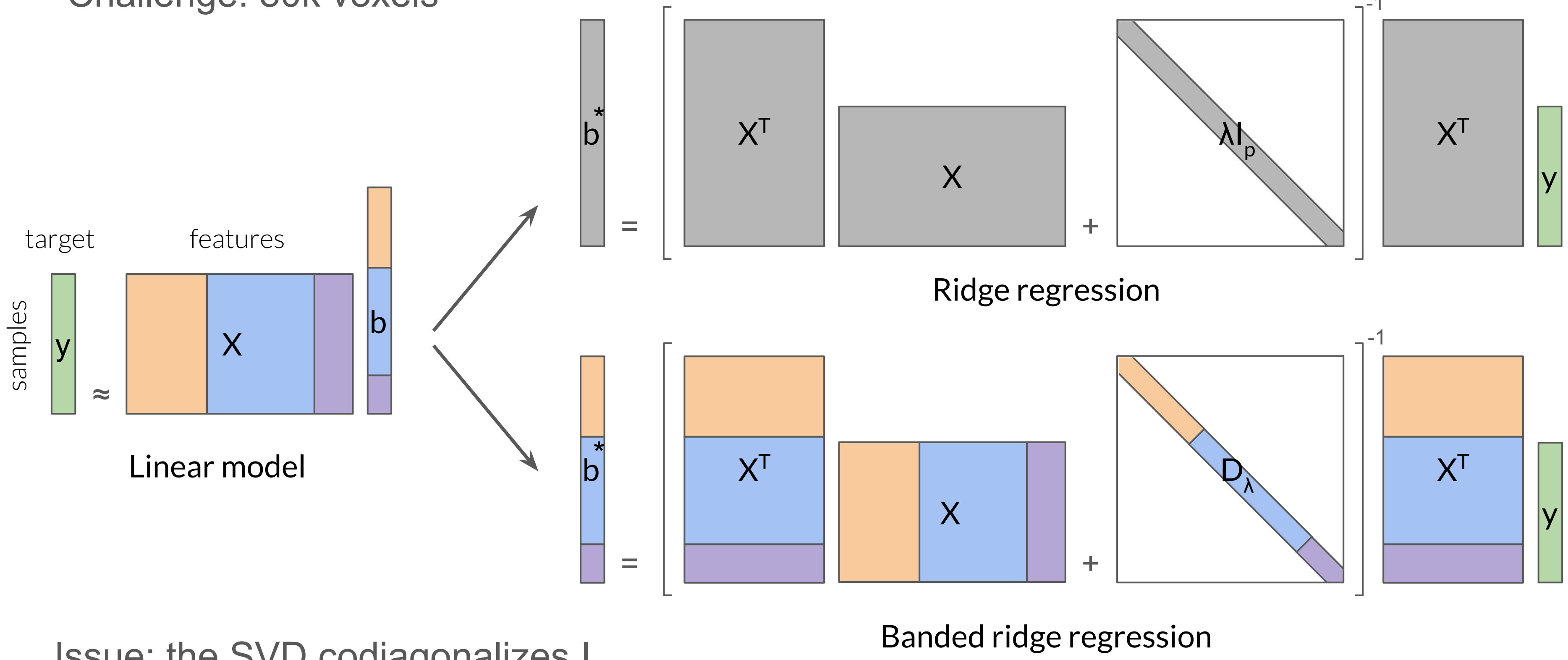
Equivalent to multiple-kernel ridge [Bach 2008, Rakotomamonjy et al. 2008]

Banded ridge ⇔ multiple-kernel ridge (with cross-validation)

Squared group lasso ⇔ multiple-kernel ridge (within set)

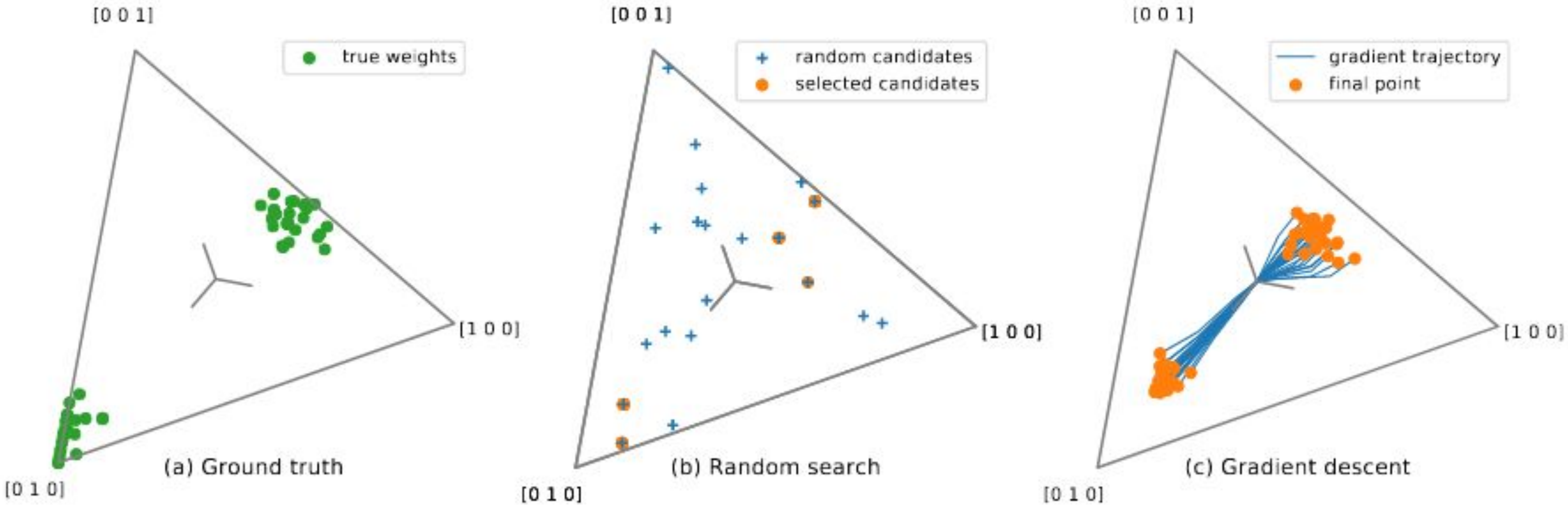
Banded ridge solvers

Challenge: 80k voxels



Issue: the SVD codiagonalizes I_p but not D_λ

Banded ridge solvers



[Bengio, 2000] [Bergstra and Bengio, 2012]

Hyperparameter random search

Multiple-kernel ridge

$$w^* = \underset{w}{\operatorname{argmin}} \left\| \sum_{i=1}^{m} \gamma_i K_i w - y \right\|_2^2 + \mu w^\top \sum_{i=1}^{m} \gamma_i K_i w$$

Dirichlet distribution

$$p(\gamma) = \frac{1}{B(\alpha)} \prod_i \gamma_i^{\alpha_i - 1},$$

Log-spaced grid

$$\mu > 0$$

Efficient for multiple-targets (80k), multiple mu (20)

Hyperparameter gradient descent

Reparametrization

$$\mathcal{L}_{\text{train}}(w, \delta) = \left\| \sum_{i=1}^{m} e^{\delta_i} K_{\text{train},i} w - y_{\text{train}} \right\|_2^2 + w^\top \sum_{i=1}^{m} e^{\delta_i} K_{\text{train},i} w,$$

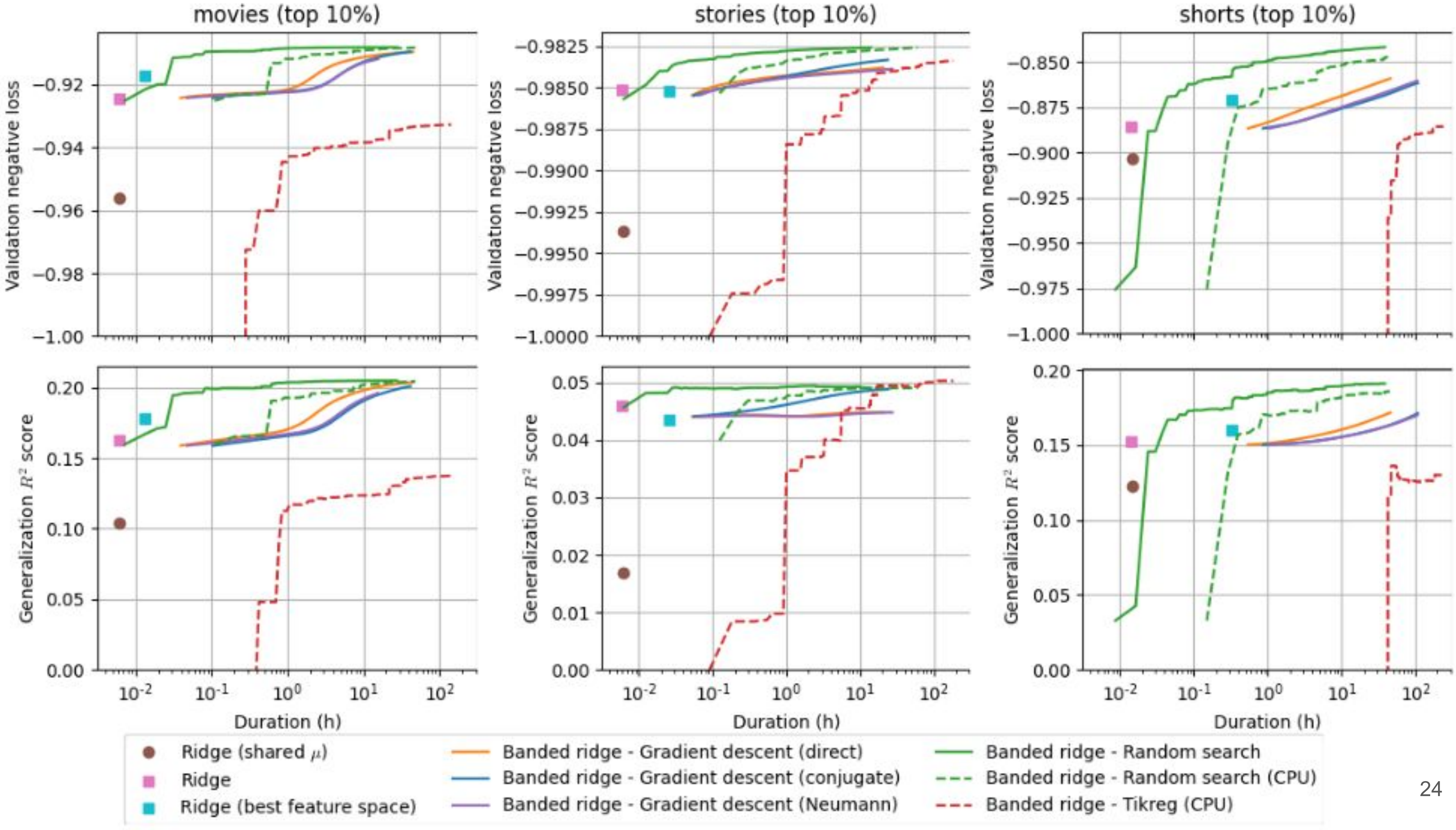
$$\mathcal{L}_{\text{val}}(w^*(\delta), \delta) = \left\| \sum_{i=1}^{m} e^{\delta_i} K_{\text{val},i} w^*(\delta) - y_{\text{val}} \right\|_2^2,$$

Gradient, using implicit differentiation [Larsen et al., 1996, Chapelle et al., 2002]

$$\frac{\partial \mathcal{L}^*_{\text{val}}}{\partial \delta} = \frac{\partial \mathcal{L}_{\text{val}}}{\partial \delta} - \frac{\partial \mathcal{L}_{\text{val}}}{\partial w^*} \left(\frac{\partial^2 \mathcal{L}_{\text{train}}}{\partial w \partial w^\top} \right)^{-1} \frac{\partial^2 \mathcal{L}_{\text{train}}}{\partial w \partial \delta^\top}.$$

Approximations

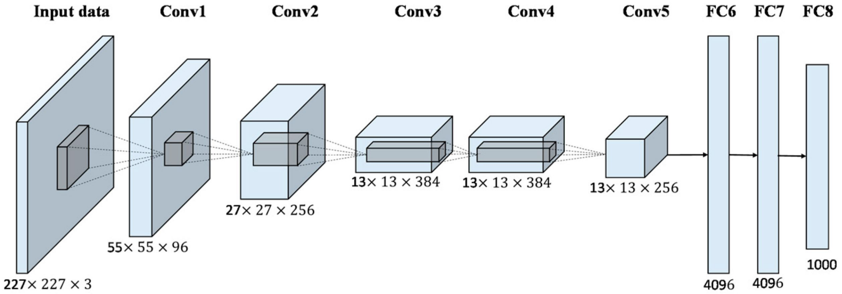
- conjugate gradient [Pedregosa, 2016], Neumann series [Lorraine et al., 2019]
- direct gradient's Lipschitz constant



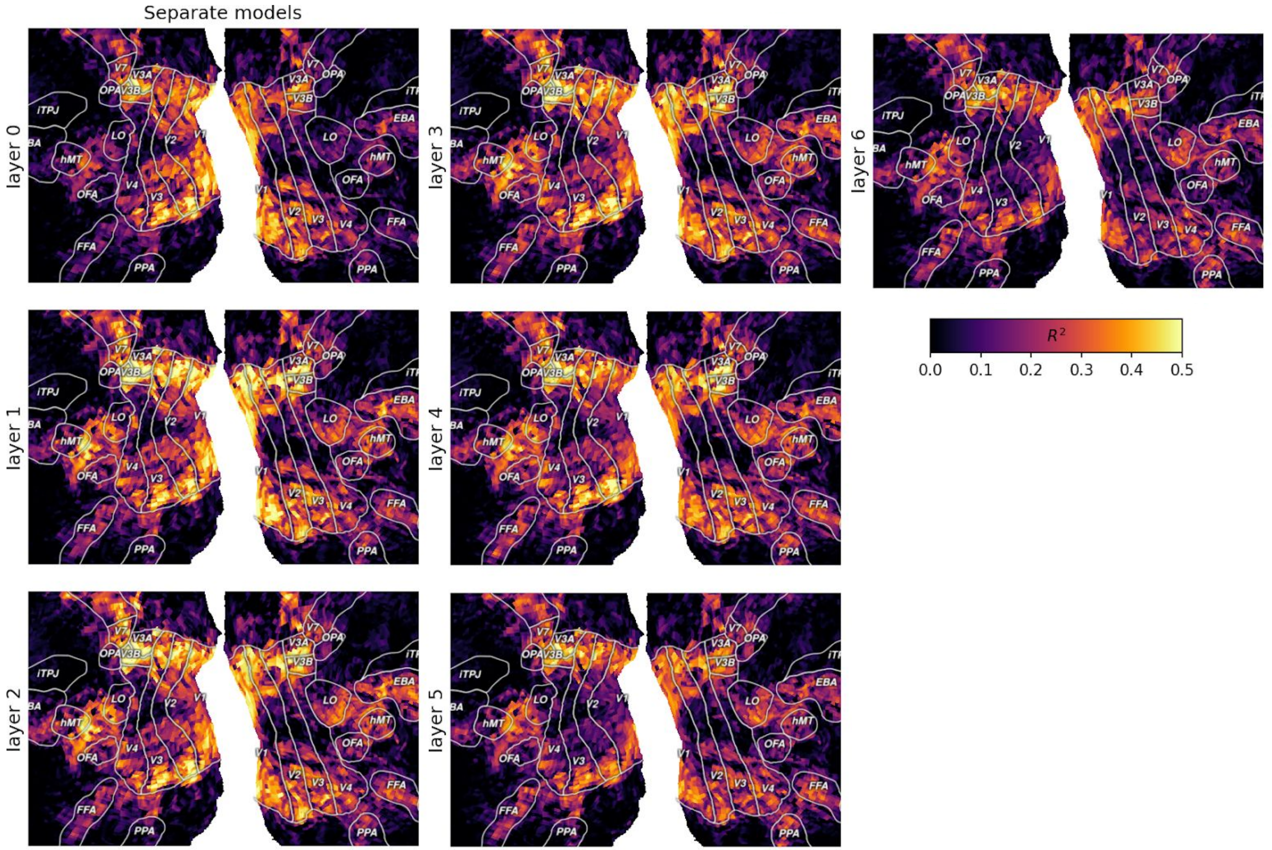
movies (top 10%)
stories (top 10%)
shorts (top 10%)
Validation negative loss
Generalization R^2 score
Duration (h)
Ridge (shared μ)
Ridge
Ridge (best feature space)
Banded ridge - Gradient descent (direct)
Banded ridge - Gradient descent (conjugate)
Banded ridge - Gradient descent (Neumann)
Banded ridge - Random search
Banded ridge - Random search (CPU)
Banded ridge - Tikreg (CPU)

Applications

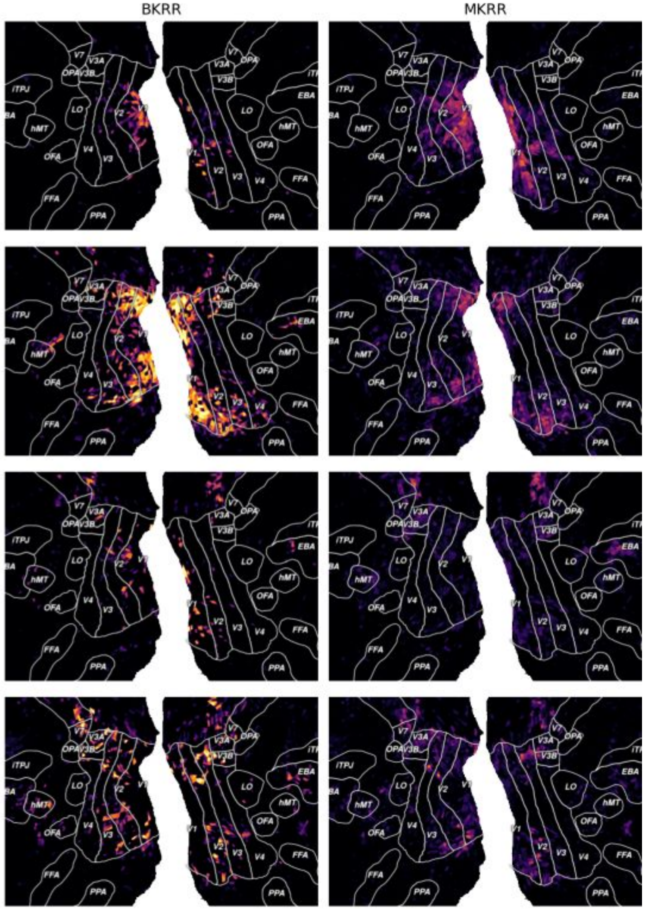
Extracting features from Alexnet

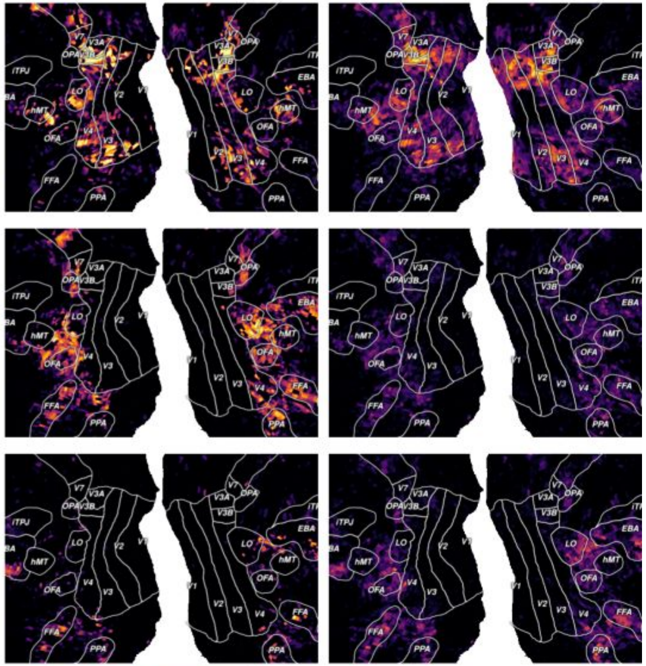


Extracting features from Alexnet



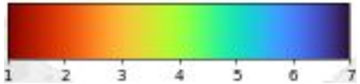
(Best-layer ridge) vs (banded ridge)



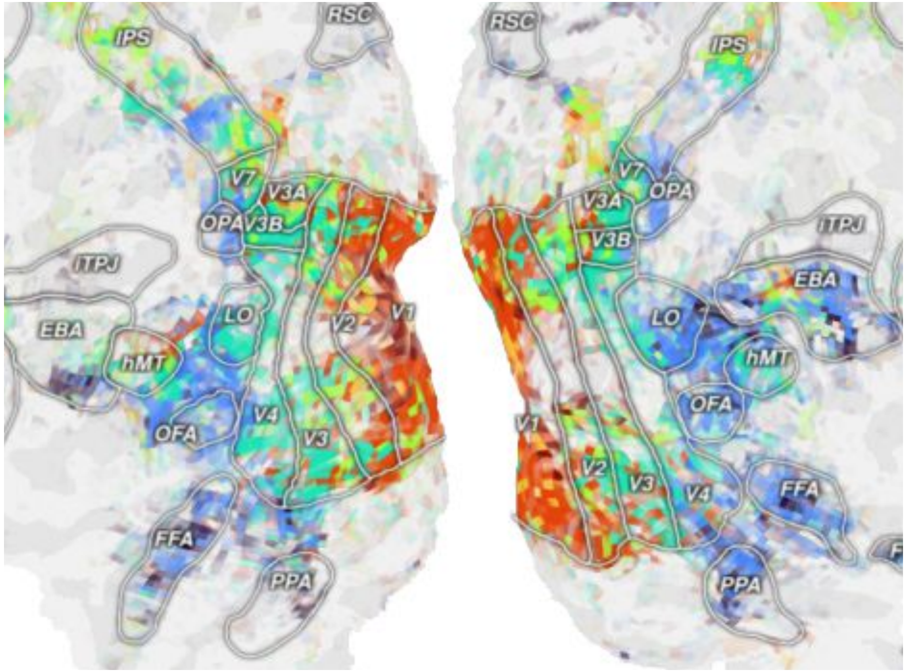


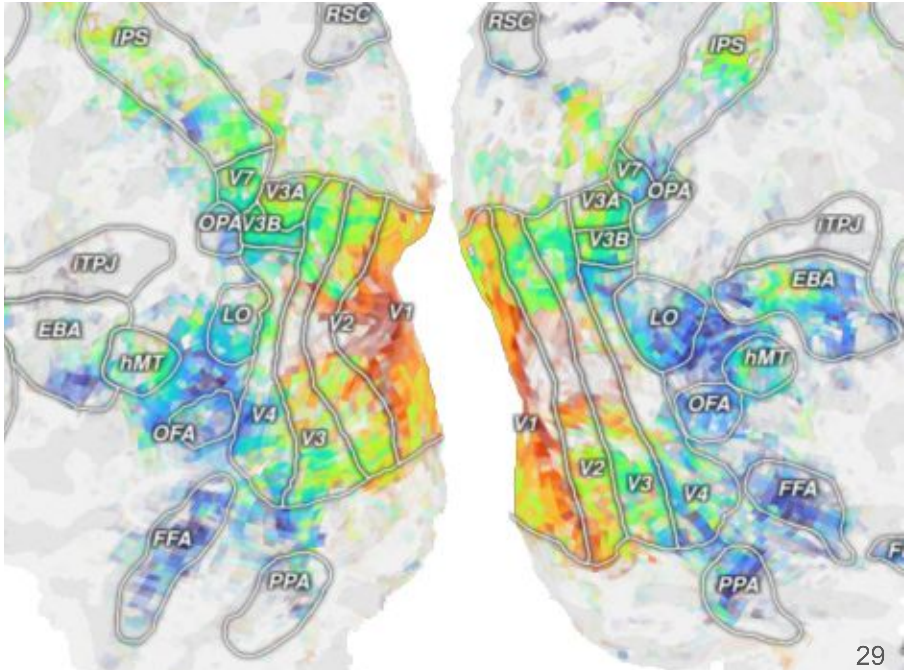
Weighted-average layer

Best-layer ridge

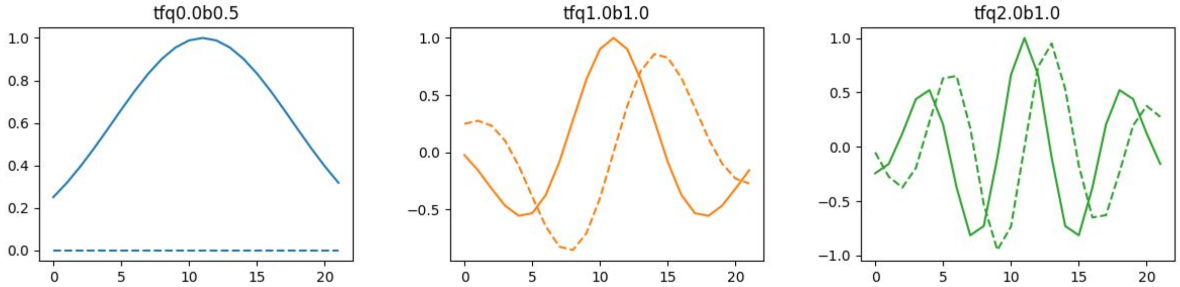


Banded ridge

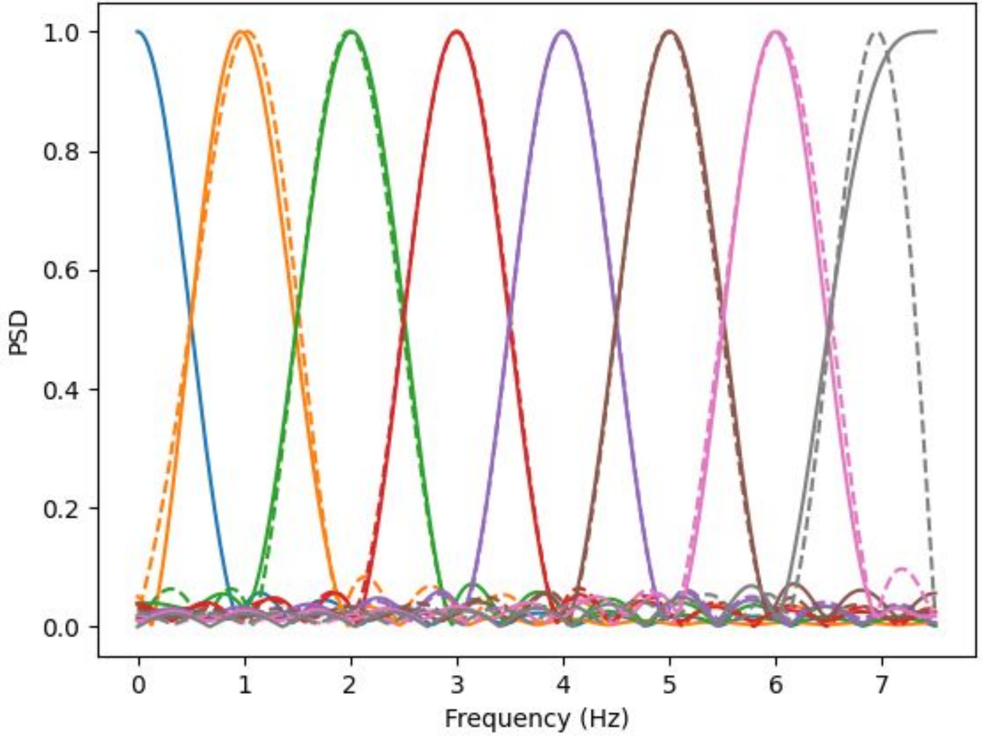




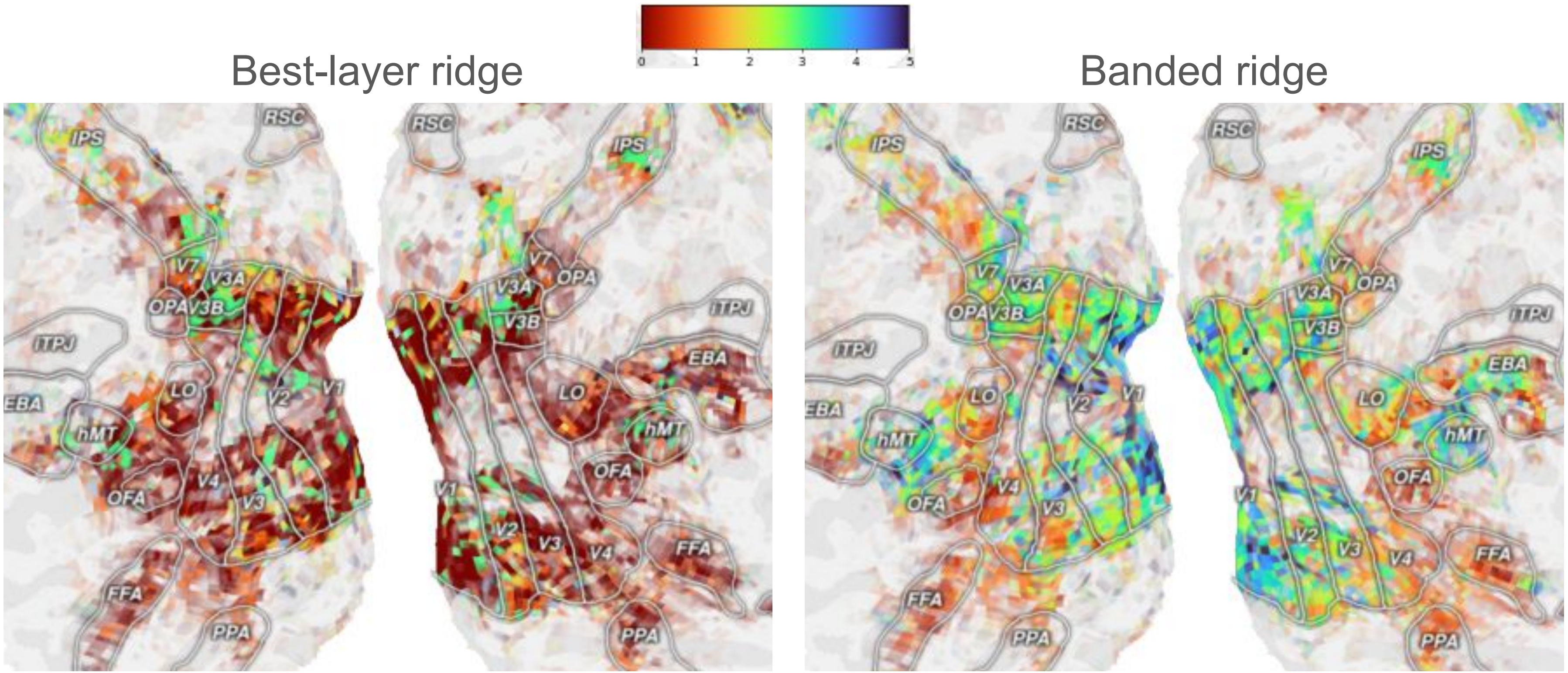
Temporal filters

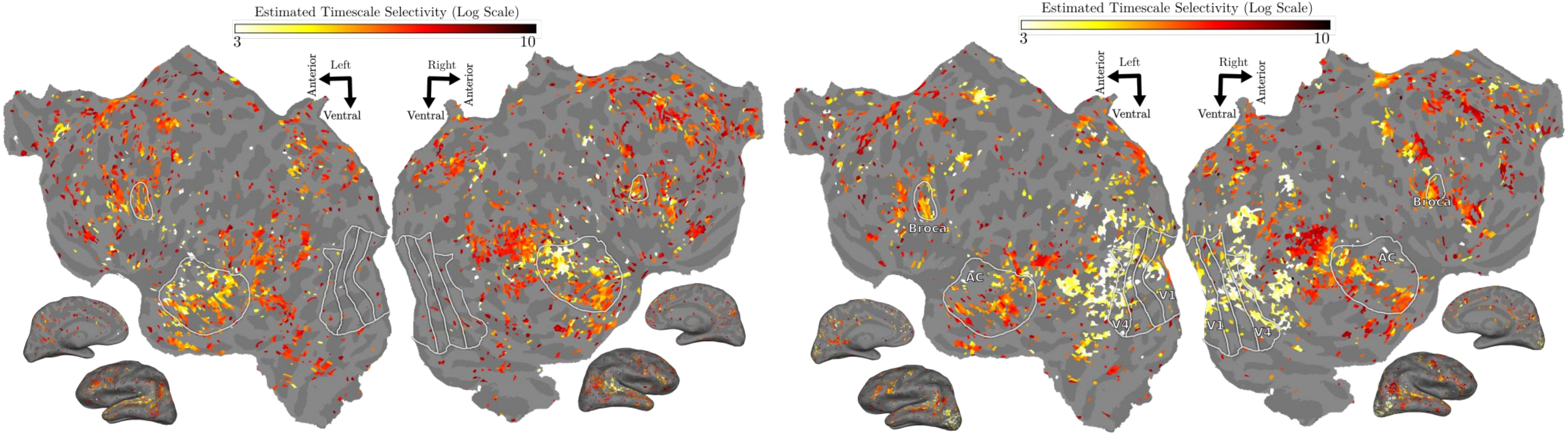


a) Layer activations
b) Temporal filters
c) Quadrature
d) Decimation to 0.5 Hz
e) Non-linearity: log(1 + x)



Weighted-average time-frequency





a. Listening **b. Reading**

Overview

Intro: Voxelwise encoding models

Banded ridge regression
- Feature-space selection
- Efficient GPU solvers

Applications on DNN features
- Layer selectivity
- Timescale selectivity

Thanks for your attention

Thanks for your attention

$$\hat{y} = \sum_{i=1}^{m} \hat{y}_i$$

Can we decompose the variance ?

Standard R^2 score

$$R^2(\hat{y}) = 1 - \frac{\sum_t (y[t] - \hat{y}[t])^2}{\sum_t y[t]^2}$$

$$\hat{y} = \sum_{i=1}^{m} \hat{y}_i$$

Can we decompose the variance ?

Standard R^2 score

$$R^2(\hat{y}) = 1 - \frac{\sum_t (y[t] - \hat{y}[t])^2}{\sum_t y[t]^2} = \frac{\sum_t \hat{y}[t]\,(2y[t] - \hat{y}[t])}{\sum_t y[t]^2}$$

"Split" R^2 score (définition)

$$\tilde{R}^2(\hat{y}_i) = \frac{\sum_t \hat{y}_i[t](2y[t] - \hat{y}[t])}{\sum_t y[t]^2}$$

(similar to Pratt's measure [Hoffman 1960, Pratt 1967])

Property

$$R^2(\hat{y}) = \sum_{i=1}^{m} \tilde{R}^2(\hat{y}_i[t])$$

Can we quantify soft sparsity ?

Starting from a variance decomposition $\rho \in \mathbb{R}^m$; (e.g. the split R² score)

we sort them $(\rho_0 \geq \rho_1 \geq \ldots \geq \rho_{m-1})$

then we compute the "effective sparsity":

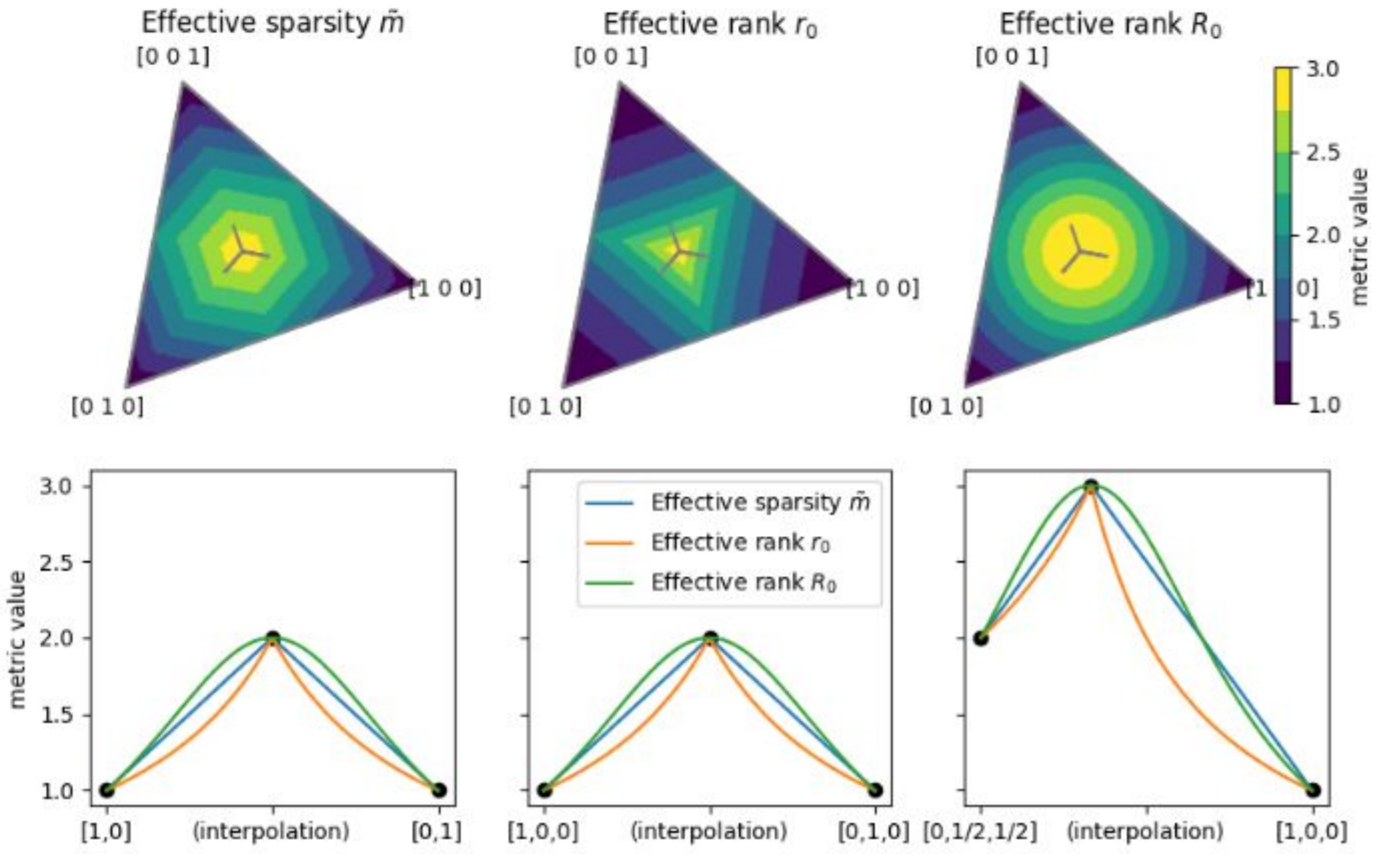
$$\tilde{m} = 1 + 2\frac{\sum_{i=0}^{m-1} i\rho_i}{\sum_{i=0}^{m-1} \rho_i}$$

Properties

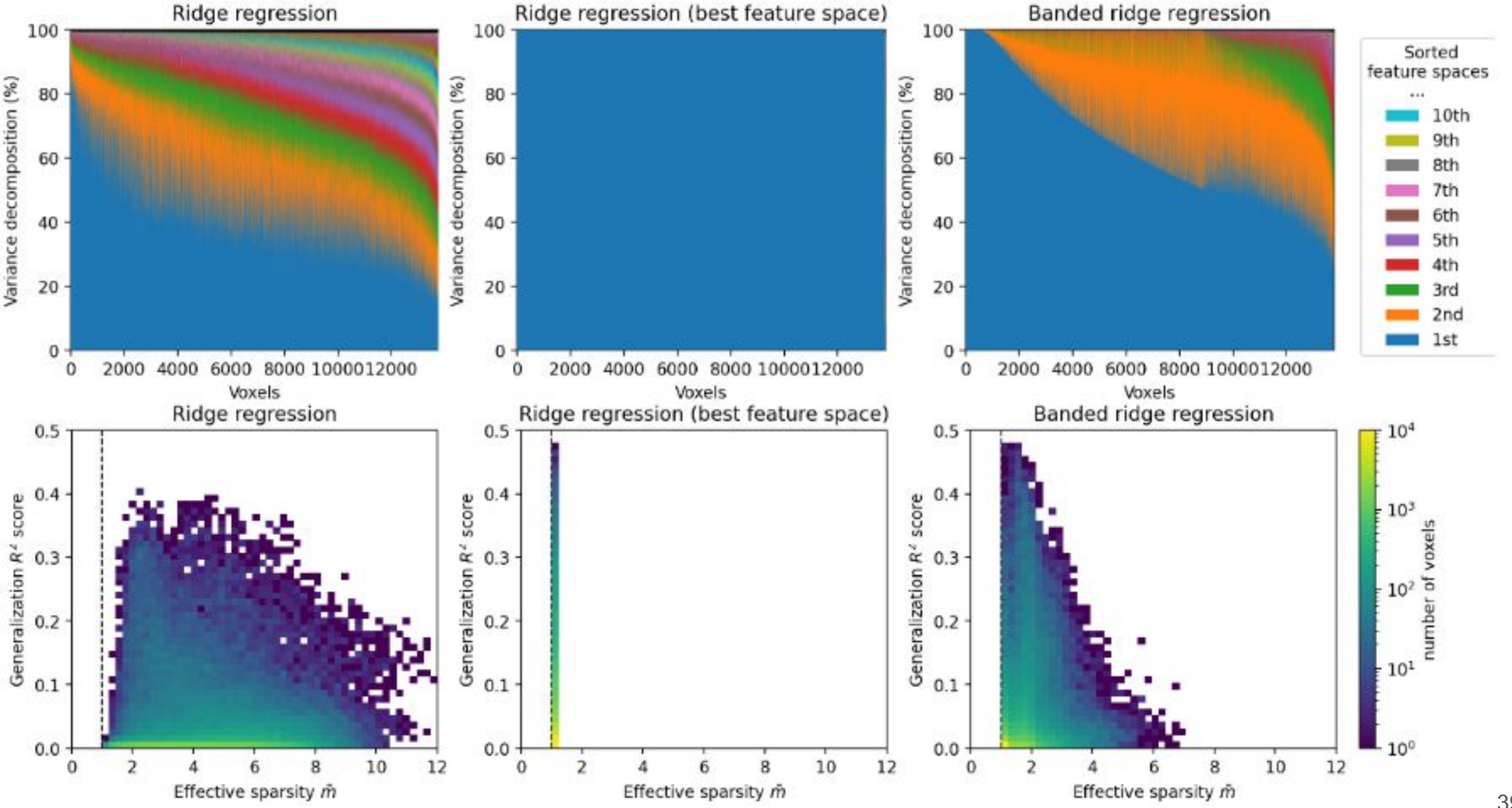
- continuous, with values in [1, m]
- equal to k when the variance is equally distributed between k groups

(Similar to the effective ranks r0 and R0 [Bartlett et al 2020])

$$r_0 = \frac{\sum_{i=0}^{m-1} \rho_i}{\rho_0} \qquad R_0 = \frac{(\sum_{i=0}^{m-1} \rho_i)^2}{\sum_{i=0}^{m-1} \rho_i^2}$$

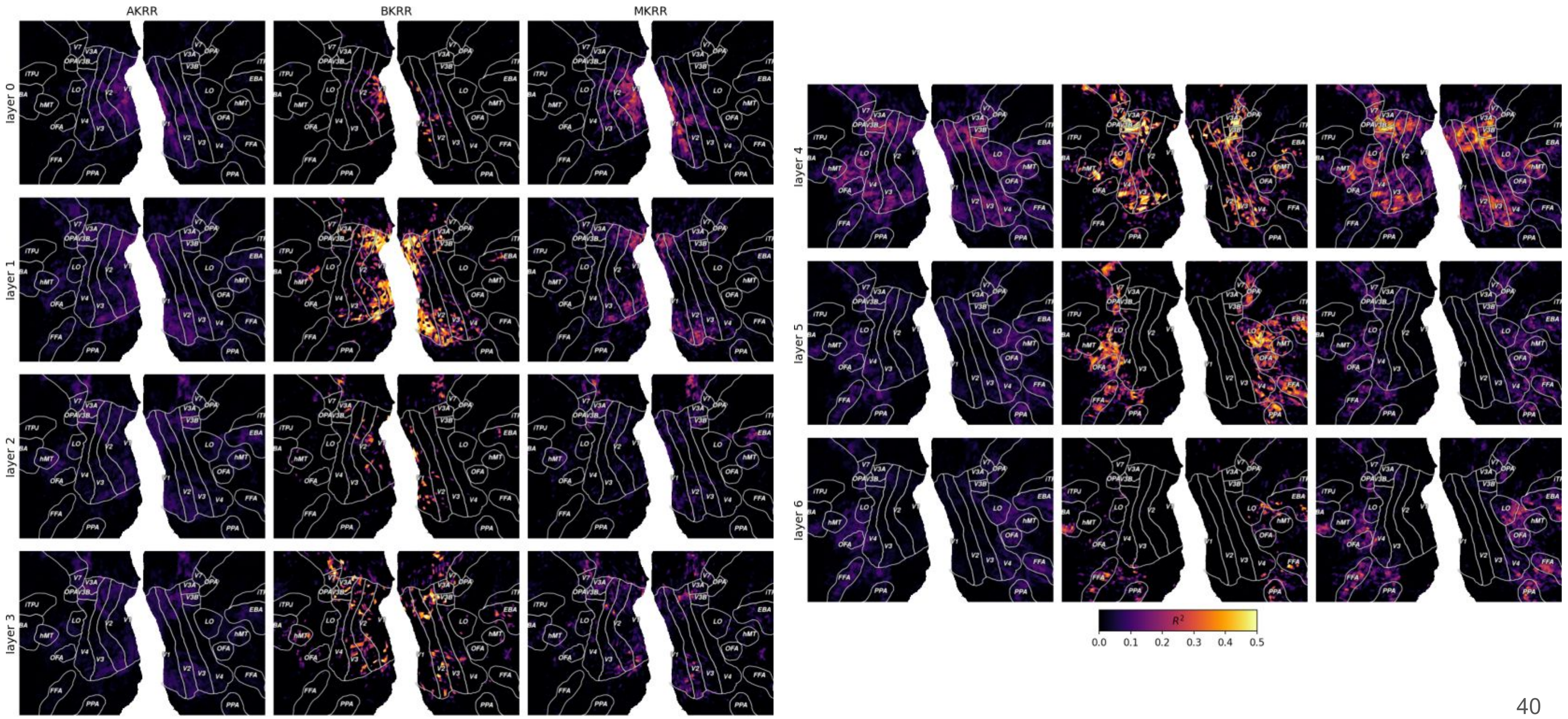


Effective sparsity $\tilde{m}$
Effective rank r_0
Effective rank R_0
[0 0 1]
[1 0 0]
[0 1 0]
metric value
3.0
2.5
2.0
1.5
1.0
Effective sparsity $\tilde{m}$
Effective rank r_0
Effective rank R_0
[1,0]
(interpolation)
[0,1]
[1,0,0]
[0,1,0]
[0,1/2,1/2]

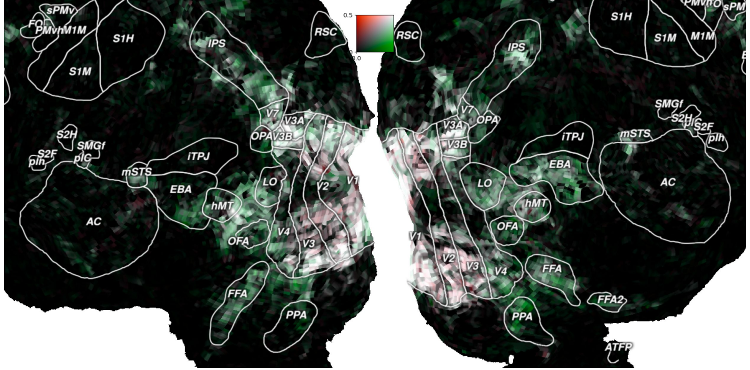


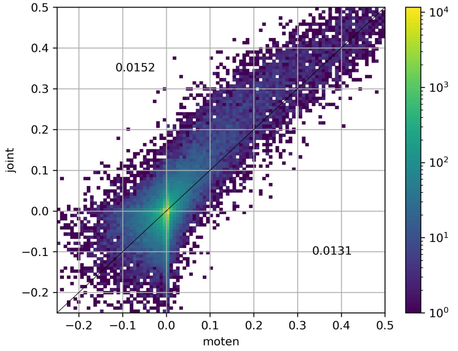
Ridge regression
Ridge regression (best feature space)
Banded ridge regression
Variance decomposition (%)
Voxels
Sorted feature spaces
...
10th
9th
8th
7th
6th
5th
4th
3rd
2nd
1st
Generalization R^2 score
Effective sparsity $\tilde{m}$
number of voxels

(Ridge) vs (best-layer ridge) vs (banded ridge)

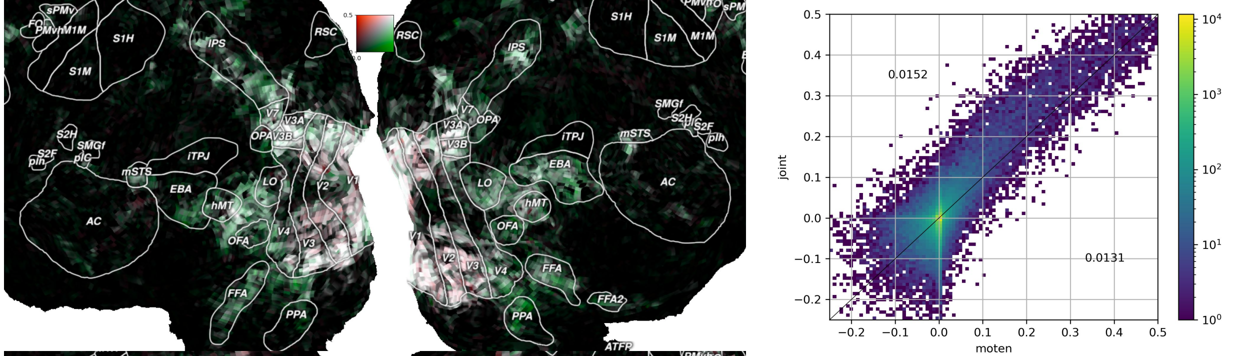


vs Magnitude of spatio-temporal filters

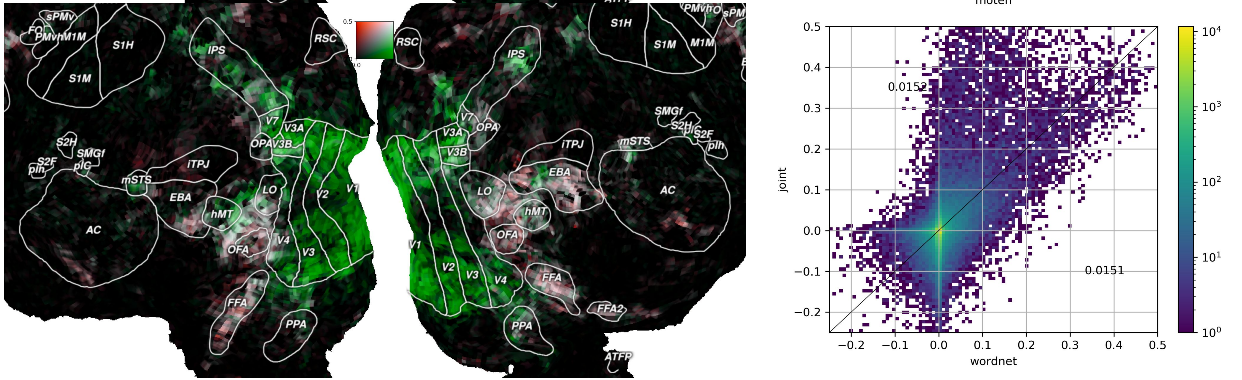




vs Magnitude of spatio-temporal filters



vs Semantic categories of objects



vs Magnitude of spatio-temporal filters

vs Semantic categories of objects

vs best of the two