Curds and Whey Regression applied to fMRI

Alan Cowen and Chris Gagne

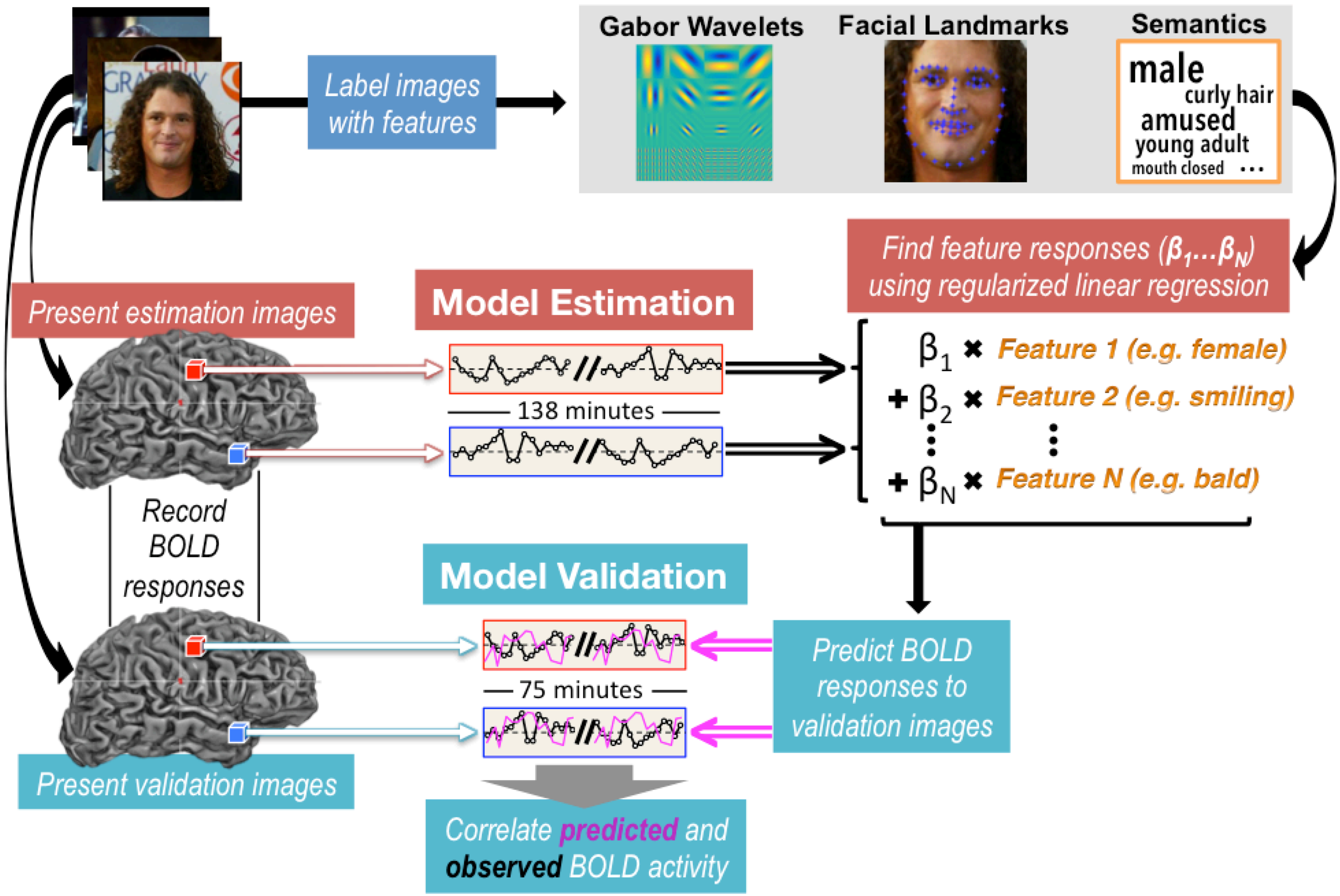
Outline

1. fMRI Dataset
2. Shrinkage and Curds and Whey
3. Simulation Performance
4. Dataset Performance

Motivation

- Predicting brain activity to validate scientific models of information processing.
- Typically, modeling done on individual voxels, either using OLS or Ridge Regression.
- In fMRI, many nearby voxels respond similarly.
- Combining prediction problems can help. (Y’s with similar B’s or shared noise).

Experiment


Label images with features
Gabor Wavelets
Facial Landmarks
Semantics
male
curly hair
amused
young adult
mouth closed ...
Find feature responses ($\beta_1 ... \beta_N$) using regularized linear regression
Present estimation images
Model Estimation
138 minutes
β_1 × Feature 1 (e.g. female)
+ β_2 × Feature 2 (e.g. smiling)
+ β_N × Feature N (e.g. bald)
Record BOLD responses
Model Validation
75 minutes
Predict BOLD responses to validation images
Present validation images
Correlate predicted and observed BOLD activity

Single Output and Shrinkage

- Single output: Y (nx1), X (nxp)
- **OLS** predictions are orthogonal projection y into column space of X to minimize errors.

$$\hat{Y}^{ols} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- **Ridge** predictions are projections into column space, but in some directions more than others (Y^ is shrunk more in direction of smaller principle components of X)

$$\hat{Y}^{ridge} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

- And we know shrinkage is good for prediction. Adds a little bias to Y^, but reduces variance.

Multi-output Shrinkage

- Multi-output: Y (nxq), X(nxp)
- Could apply same lambda for all y's or estimate one for each y.
- Or shrink OLS predictions based on correlation between multiple y's and x's.

$$\hat{Y} = \hat{Y}^{ols} S$$

$$\hat{Y} = X(X^t X)^{-1} X^t Y S$$

Curds and Whey

- Optimal prediction in population setting related to CCA of X and Y. (Breiman and Friedman 1997, BH).
- **Canonical Correlations Analysis** (CCA):
  - "find pairs of linear combinations such that each successive pair maximizes correlation (under constraint of being uncorrelated with other pairs)"
  - Sample solution is found via eigenvalue decomposition of the following:

$$\hat{Q} = (Y^tY)^{-1}(Y^tX)(X^tX)^{-1}(X^tY) = \hat{T}^{-1}\hat{C}^2\hat{T}$$

  - Columns of T are new 'canonical' basis vectors for Y, and C^2 are squared canonical correlations. This new basis for Y highlights relationships with X.

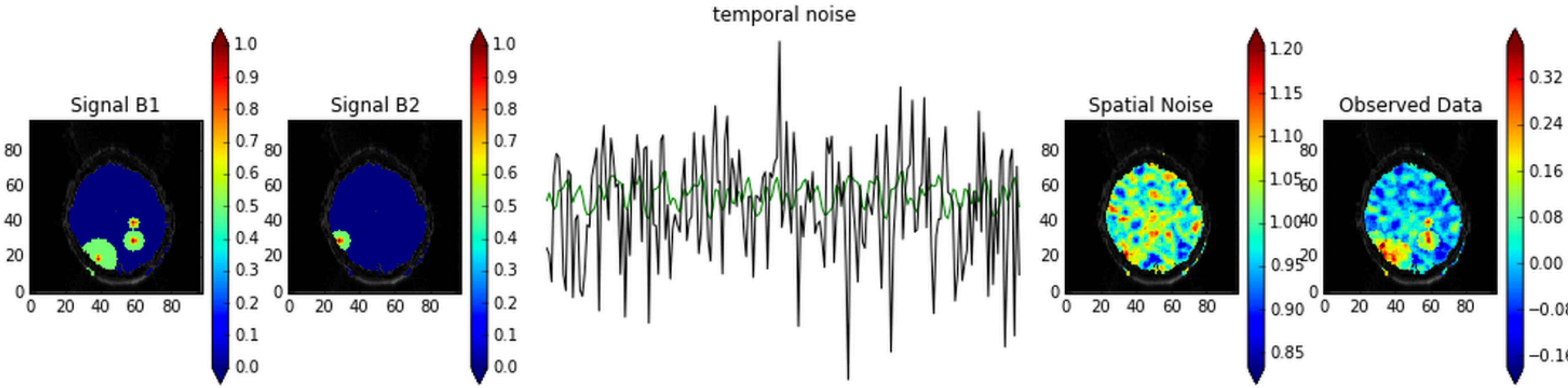
- **shrinkage matrix S** should be (BH):

$$S = \hat{T}\hat{D}\hat{T}^{-1} \qquad \hat{d}_i = \frac{(1-r)(\hat{c}_i^2 - r)}{(1-r)^2\hat{c}_i^2 + r^2(1-\hat{c}_i^2)}, \qquad i = 1, \ldots, q.$$

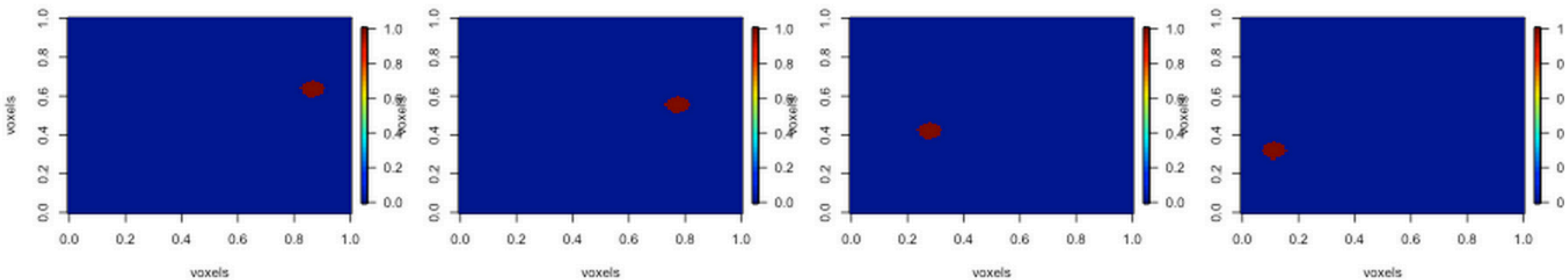
- D adjusts C^2 according to generalized cross validation. (r is p/n).
- Transform OLS predictions into new basis, shrink along the canonical coordinates in proportion to the canonical correlation (adjusted according to d), and then transform back to predict

$$\hat{Y}^{cw} = X(X^tX)^{-1}X^tYS$$

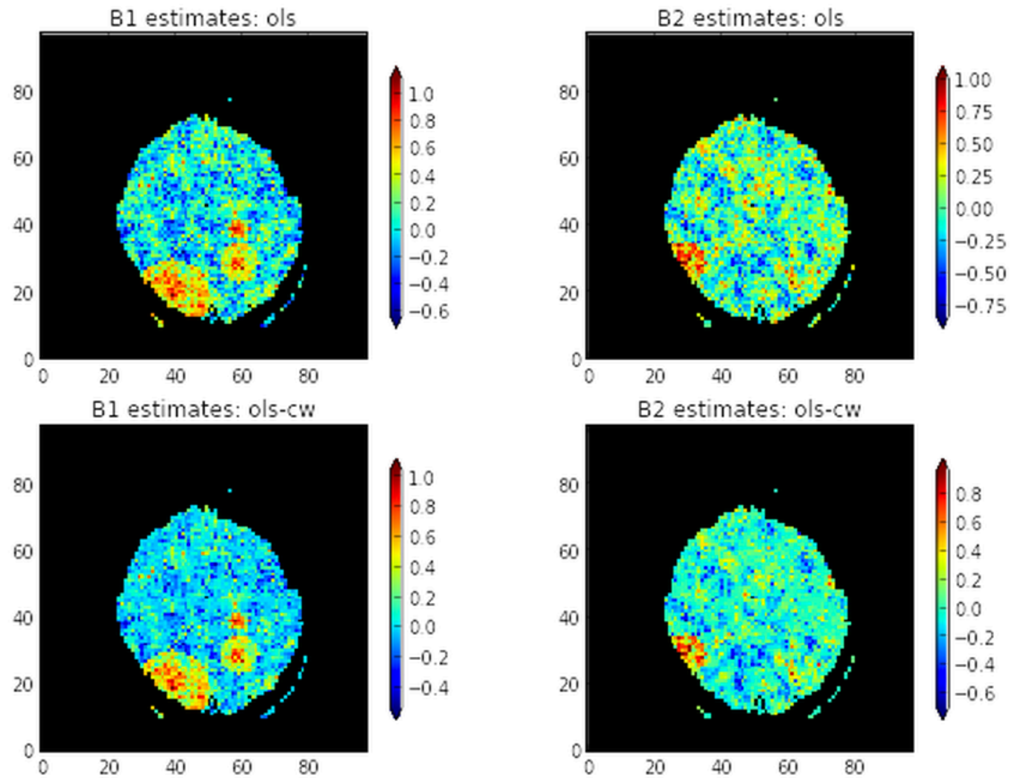
Simulated Data

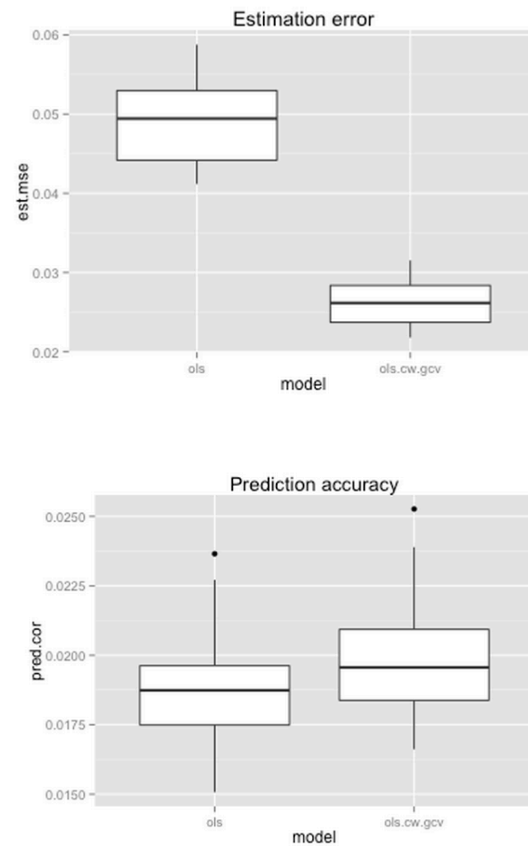


Search Light / Nearest Neighbors



Simulation Performance





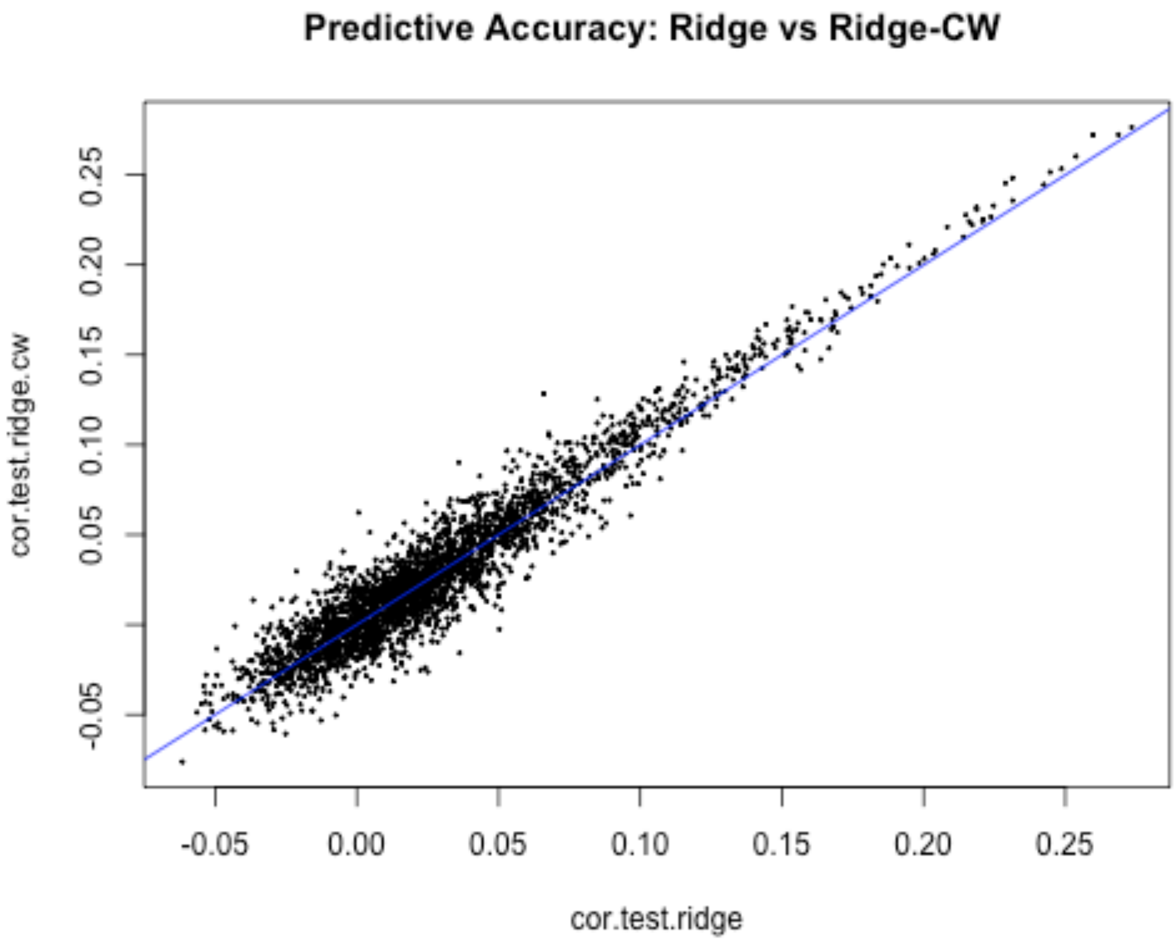
Ridge Curds and Whey for fMRI

- Our model has too many regressors for OLS
- Can shrink in both Y space (cw) and X space (ridge)

$$\hat{Y}^{ridgecw} = X(X^t X + \lambda I)^{-1} X^t Y S$$

- CCA done on Y and Y^ ridge (not X)

Performance on Dataset Ridge v Ridge-CW



Double Ridge

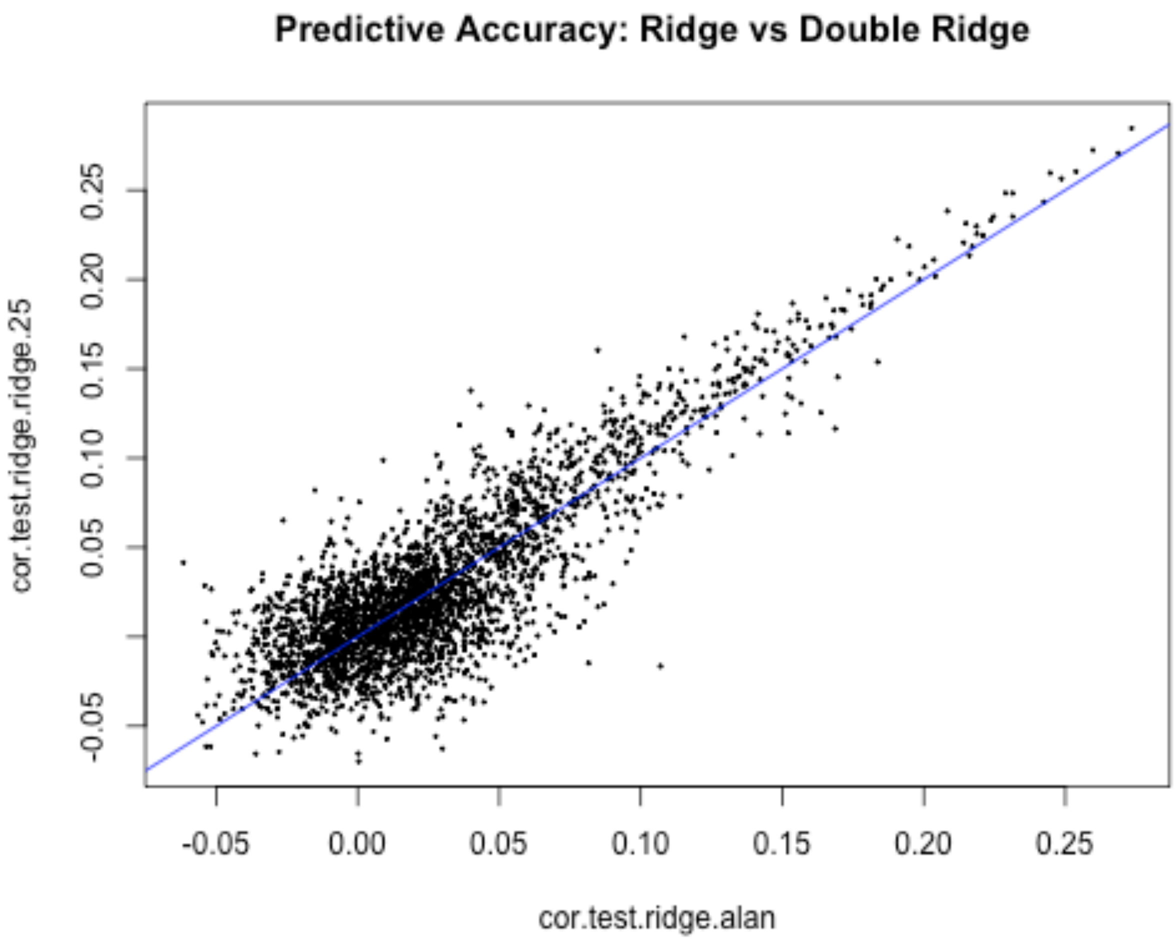
Estimate the optimal shrinkage using another stage of ridge.

- After estimation of the model using ridge regression, Y is regressed onto out-of-sample Y^ using another ridge regression. The model is trained with an optimal regularization parameter (gamma) selected using cross-validation.

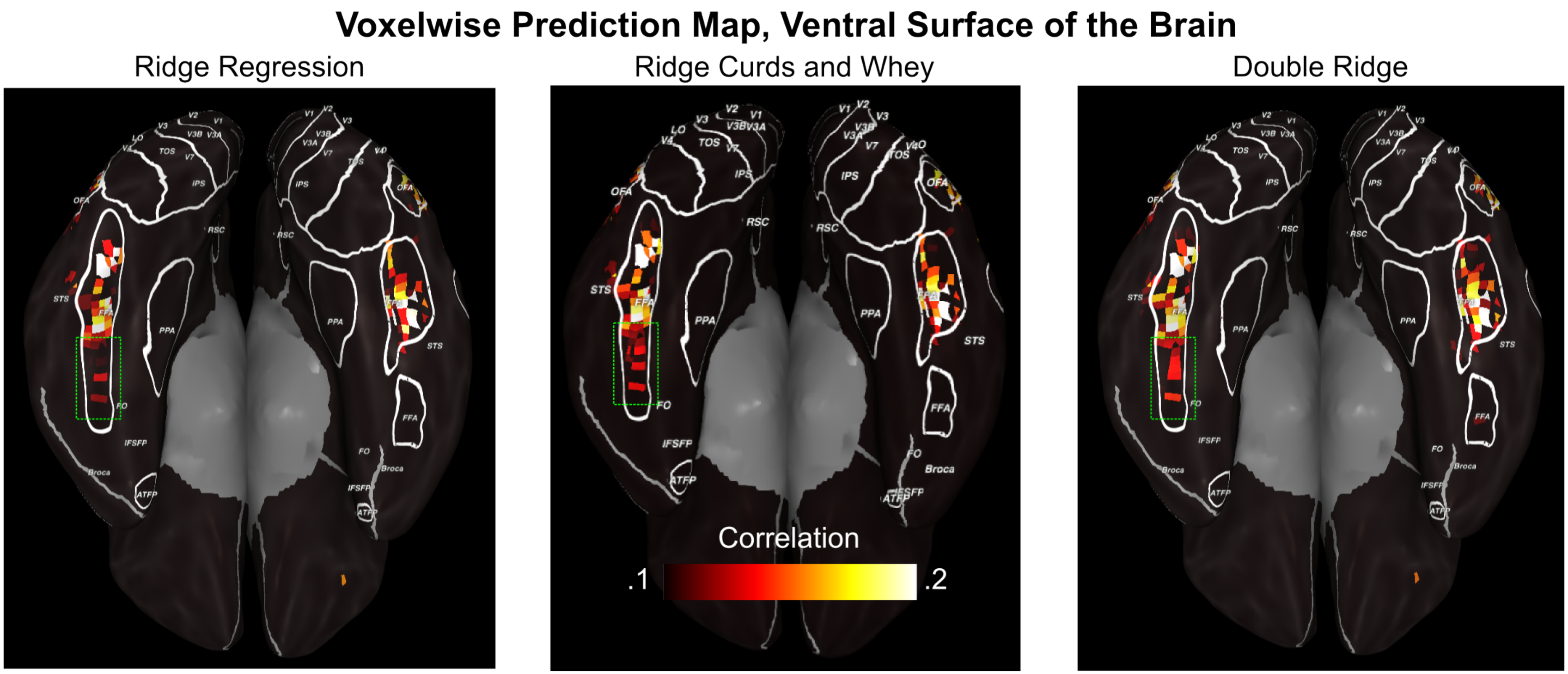
$$\hat{S} = argmin_S[\Sigma(Y_i - S\hat{Y_i})^2 + \gamma\Sigma(S_j)^2]$$

- For validation, Y^ is multiplied by S^.

Performance on Dataset Ridge v Double Ridge



Performance Maps



Extensions

- Generalized cross-validatoin approximates loocv. High variance.
- Continue to use good basis for Y.
- Use k-fold cross validation to estimate a good amount of shrinkage (D).

$$S = \hat{T}\hat{D}\hat{T}^{-1}$$

The END

Mscl

$$
\begin{aligned}
\hat{Y}^{ridge} &= \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{U}\,\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\,\mathbf{U}^T\mathbf{y} \\
&= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^T\mathbf{y},
\end{aligned}
$$

$$
\begin{aligned}
\hat{Y}^{ols} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\
&= \mathbf{U}\mathbf{U}^T\mathbf{y},
\end{aligned}
$$