1 Multi-Response Prediction in High Dimensions

Consider the problem of predicting the fMRI response to natural images for $q = 20$ different voxels in the primary visual cortex (V1). Neurons in this region have a localized receptive field and cortical columns in V1 have been well modeled using Gabor transformations. Hence it's reasonable to assume the fMRI response in a certain voxel is (roughly) a linear function of a small number of Gabor transforms of the image. Moreover, adjacent voxels will be strongly correlated, and we would like to take advantage of these correlations to improve predictive accuracy. We begin by reviewing a method for exploiting correlations in multi-response linear regression where $p < n$.

1.1 Curds & Whey & Stein

As before, X is an $n \times p$ centered data matrix. Now we have q different responses, $Y = (y^{(k)})_{k=1}^{q}$, each in $\mathbb{R}^n$, so Y is an $n \times q$ matrix. Each response is linearly related to the covariates, so

$$Y = XA + E, \tag{1}$$

where A is a $p \times q$ matrix of pq regression coefficients, and E is an $n \times q$ matrix of independent draws from $\mathcal{N}(0, \sigma^2)$. The OLS estimate is $\widehat{A} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y$. One way to assess its accuracy is via total squared prediction error (PTSE) on a new sample pair $x \sim \mathcal{N}(0, V)$ and $y = xA + \varepsilon$:

$$\mathbb{E}\left\|y - x\widehat{A}\right\|_2^2 = q\sigma^2 + \sum_{k=1}^{q} \mathbb{E}\left[(A_k - \widehat{A}_k)^\mathsf{T} V (A_k - \widehat{A}_k)\right]. \tag{2}$$

The previous display is the multi-response analogue of the bias-variance decomposition in (**??**) (note we omit an intercept), indicating that we could do better on each problem (and hence in PTSE) by applying Copas (1983) shrinkage on each problem separately. This amounts to shrinking $\widehat{A}$ by a $q \times q$ diagonal matrix:

$$\widetilde{A} = \widehat{D}\widehat{A}, \text{ where } \widehat{D}_{kk} = \left(1 - \frac{(p-2)(\widehat{\sigma}^2/n)\nu}{(\nu+2)\widehat{A}_k^\mathsf{T} S \widehat{A}_k}\right), \tag{3}$$

where $\widehat{\sigma}^2$ is a pooled estimate of variance. Perhaps we can do even better in terms of PTSE by using a *non-diagonal* shrinking $q \times q$ matrix $\widehat{B}$ and taking $\breve{A} = \widehat{B}\widehat{A}$. To leverage Stein-shrinkage with a shrinking matrix $\widehat{B}$, we require an extension of the James-Stein estimator, presented below:

Lemma 1. *(Efron & Morris, 1972) Suppose* $\widehat{\Xi} = (\widehat{\Xi}_1 \cdots \widehat{\Xi}_p)^\mathsf{T}$ *is a* $p \times q$ *matrix with* $q + 1 < p$ *and* $\widehat{\Xi}_j \overset{ind}{\sim} \mathcal{N}(\Xi_j, I_q)$. *Define* $\mathcal{S} = \widehat{\Xi}^\mathsf{T}\widehat{\Xi}$, $\nu = p - q - 1$, *and*

$$\breve{\Xi}_j := \left(I - \nu\mathcal{S}^{-1}\right)\widehat{\Xi}_j. \tag{4}$$

Then **for all** Ξ *the shrunk matrix* $\breve{\Xi}$ *dominates the MLE* $\widehat{\Xi}$ *in the Frobenius norm:*

$$\mathbb{E}\|\breve{\Xi} - \Xi\|_F^2 < \mathbb{E}\|\widehat{\Xi} - \Xi\|_F^2. \tag{5}$$

Note that $V^{1/2}\widehat{A}_k \sim \mathcal{N}(V^{1/2}A_k, (\sigma^2/n)I_p)$, and since each error term E_{ik} is assumed independent above, $V^{1/2}\widehat{A} \sim \mathcal{N}(V^{1/2}A, (\sigma^2/n))I_{p\times q})$. To apply the lemma, let $\widehat{\Xi} = \sqrt{n/\sigma^2}V^{1/2}\widehat{A}$ and similarly for Ξ, so the estimation error term in equation (2) can be written as $\mathbb{E}\|\widehat{\Xi} - \Xi\|_F^2$. Blithely assuming $V = n^{-1}X^\mathsf{T}X$ as in Section **??**, we improve upon PTSE with

$$\begin{aligned}
\breve{A} &:= \left(I - \nu V^{-1}\left((n/\sigma^2)\widehat{A}^\mathsf{T} V \widehat{A}\right)^{-1}\right)\widehat{A} && (6)\\
&= \left(I - \sigma^2 rn(X^\mathsf{T}X)^{-1}\widehat{Q}^{-1}(Y^\mathsf{T}Y)^{-1}\right)\widehat{A} && (7)\\
&\approx \left\{(1-r)I + r\widehat{Q}^{-\mathsf{T}}\right\}^{-1}\widehat{A} && (8)
\end{aligned}$$

where $\widehat{Q} := (Y^\mathsf{T}Y)^{-1}Y^\mathsf{T}X(X^\mathsf{T}X)^{-1}X^\mathsf{T}Y$ is the CCA matrix and $r = \frac{\nu}{n} = \frac{p-q-1}{n}$. The last step follows from plugging in a residual estimate of $\sigma^2 I \approx (Y - X\widehat{A})^\mathsf{T}(Y - X\widehat{A})/(n-p)$ and applying Woodbury's formula. This gives rise to the *Curds & Whey* method of Breiman & Friedman (1997):

1. (Run CCA) Factor $\widehat{Q} = \widehat{T}^{\mathsf{T}} \widehat{C}^2 \widehat{T}^{-1}$, where $\widehat{C}^2$ is diagonal and $\widehat{T}$ forms a basis.
2. Set $\widehat{B} = \left\{(1-r)I + r\widehat{Q}^{-\mathsf{T}}\right\}^{-1} = \widehat{T}^{-1}\widehat{D}\widehat{T}$, where $\widehat{D}_{kk} = \frac{\widehat{C}_{kk}^2}{\widehat{C}_{kk}^2 + r(1-\widehat{C}_{kk}^2)}$. Define the Curds & Whey coefficients $\breve{A} = \widehat{B}\widehat{A}$. Predict $\breve{y} = x\breve{A}$ on new data.

As indicated in Figure 1 below, this closed-form method for Curds & Whey is in some cases worse than OLS, since this method uses the entire sample to estimate canonical correlations and hence is prone to overfitting. We can instead use cross-validation to select the diagonal matrix in Curds & Whey, which turns out to be the solution to a QP:

$$\widehat{D} := \text{pmax}(\text{diag}(M^{-1}u), 0) \text{ where } u = \sum_n (y_n^{\mathsf{T}} \widehat{T}_{\backslash n}^{-1}) \circ r_n, M = \sum_n \left(\widehat{T}_{\backslash n}^{-\mathsf{T}} \widehat{T}_{\backslash n}^{-1}\right) \circ r_n r_n^{\mathsf{T}} \text{ and } r_n = \widehat{T}_{\backslash n} \widehat{y}_n^{\backslash n},$$

where $\backslash n$ means the quantity is estimated with the n^{th} datapoint excluded. This works best in the top-right facet of the figure, which approaches the high dimensional case. This method is also the most extensible as it relies the least on distributional assumptions. Still, it's interesting to note how applying shrinkage à la Copas (1983) separately on each of the q problems performs comparably to Curds & Whey, as the latter is estimating the best linear predictor of y on $\widehat{y}$ and the former does not share strength across problems (beyond the pooled estimate of the error variance). Neither method improves much in relative TSE when q increases.

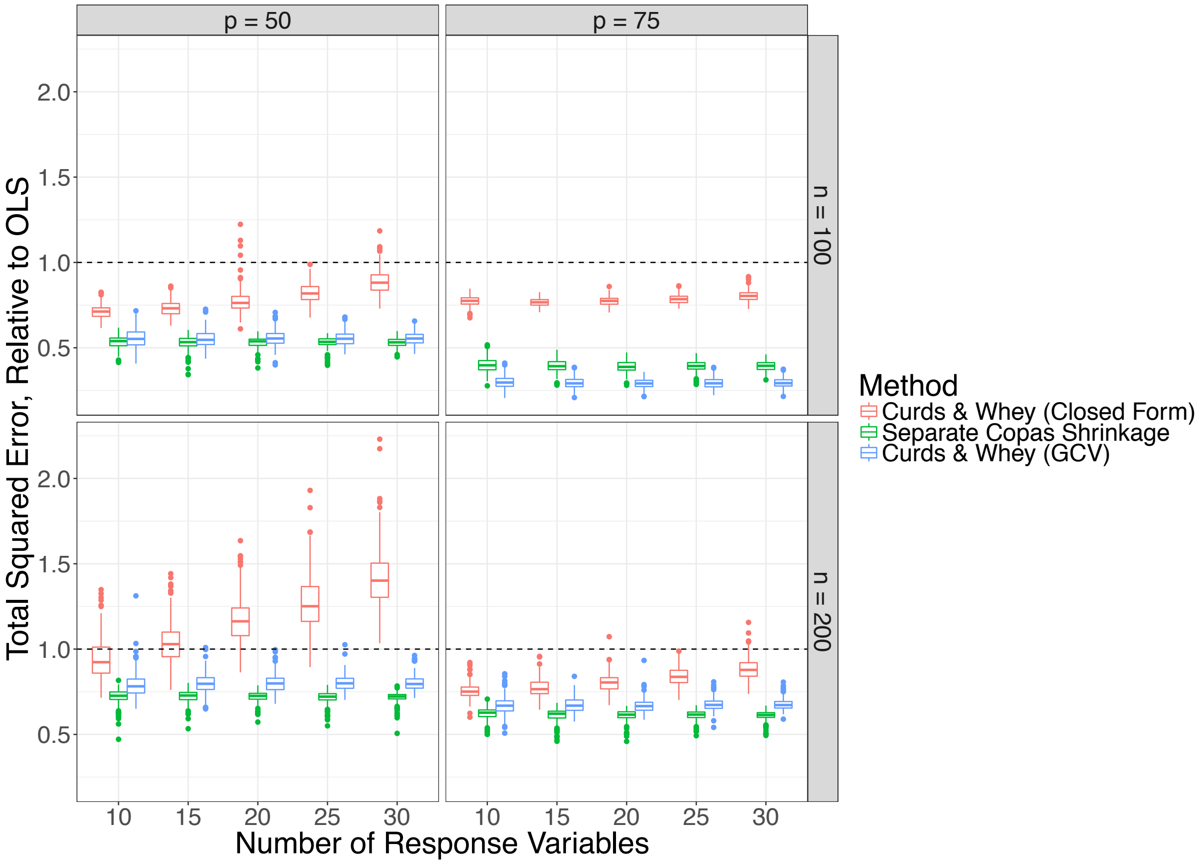

Figure 1: Simulation-based comparison of multi-response shrinkage methods in moderate dimensions p relative to the sample size n. The simulation design follows Breiman & Friedman (1997), with fixed, unit signal-to-noise ratio. The horizontal axis is the number of responses q that we are modeling, and the vertical axis is the ratio $\frac{\sum_k (A_k - \widetilde{A}_k)^{\mathsf{T}} V (A_k - \widetilde{A}_k)}{\sum_k (A_k - \widehat{A}_k)^{\mathsf{T}} V (A_k - \widehat{A}_k)}$ of PTSE (modulo constants) for each method relative to OLS. Each box plot was created using 250 runs.

1.2 Extensions to Sparse Modeling

When the number of predictors p exceeds the sample size n, we face two major problems. First, as before, OLS is not well-defined, since $X^\top X$ is not invertible; second, CCA is also not well-defined for the same reason. Motivated by our mulitple-voxel fMRI response problem, we will assume A is sparse. The first issue can be overcome by solving LASSO separately on each problem:

$$\widehat{A}_{\text{lasso}} := \arg\min_{A'} \sum_{k=1}^{q} \left[\|y^{(k)} - XA'_k\|_2^2 + \lambda_k \|A'_k\|_1 \right], \tag{9}$$

This will be our baseline predictor, as it does not attempt to exploit any relationship between the q related problems. Each λ_k is selected via cross-validation.

1.2.1 Subset Selection then Curds & Whey

One extreme is where the sparsity patterns of the $(A_k)_{k=1}^q$ are related so that the union of their support is not much larger than the largest support. Then it makes sense to replace the separable regularization term $\sum_{k=1}^q \sum_{j=1}^p \lambda_k |A'_{jk}|$ in equation (9) with $\lambda \sum_{j=1}^p \max_{k=1:q} |A'_{jk}| = \lambda \sum_{j=1}^p \|(A')^j\|_\infty$ to select features jointly across all of the problems. Obozinski et al (2006) relax this complete feature-sharing scenario by replacing the ℓ_∞ norm with ℓ_2:

$$\widehat{A}_{\text{multitask-lasso}} := \arg\min_{A'} \sum_{k=1}^{q} \|y^{(k)} - XA'_k\|_2^2 + \lambda \sum_{j=1}^{p} \|(A')^j\|_2. \tag{10}$$

This method (which can be run in `glmnet` by setting `family = "mgaussian"`) already shares strength across the different problems, but if we are using equation (10) to select a sparse subset $\widehat{S}$ of s features jointly for all q problems, then we can run the Curds & Whey procedure on $X_{\widehat{S}} \in \mathbb{R}^{n\times s}$ and Y. This requires $q+1 < s < n$, so we need to constrain λ when doing CV. We call the resulting estimate $\widehat{A}_{\text{ML-CW}}$ for Multitask-LASSO then Curds & Whey. This method relies heavily on feature selection, which requires a much stronger signal than getting good predictions with LASSO. The benefit is that pooling effectively affords us more observations.

1.2.2 Sparse CCA

Given an $n \times p$ matrix U and an $n \times q$ matrix W with $U^\top U = I_p$ and $W^\top W = I_q$, when $n < p$ one way to make progress is to assume the canonical u-vectors are sparse and impose an ℓ_1 constraint:

$$\min_{\substack{u,w \\ \|u\|_2 \vee \|w\|_2 \le 1 \\ \|u\|_1 \le c\sqrt{p}}} \quad u^\top U^\top W w, \tag{11}$$

for $c \in [0, 1]$. Note if $c = 1$ then the ℓ_1 constraint is vacuous and this is equivalent to CCA. This assumption that the U-canonical directions are sparse makes sense in the context of our fMRI study, since the fMRI responses only 'care' about a small number of Gabor wavelets. We use the implementation of Witten et al (2009) to obtain a y-canonical basis $\widehat{T}^\top$ and then pool the LASSO predictions with $\widehat{T}^{-1}\widehat{D}\widehat{T}y$ (here $\widehat{D}$ can be obtained with CV as before but can have elements bigger than 1).

1.2.3 Sparse Precision Estimation

The CCA matrix $\widehat{Q} := (Y^\top Y)^{-1} Y^\top X (X^\top X)^{-1} X^\top Y$ can be thought of as estimating $\Sigma_{yy}^{-1}\Sigma_{yx}\Theta\Sigma_{xy}$, where $\Theta = \Sigma_{xx}^{-1}$. In the simulations of Breiman & Friedman (1997), the covariates are drawn from an AR$_p$(1) Gaussian graphical model $(\Sigma_{xx})_{ij} = r^{|i-j|}$, which is a Markov model, and hence the inverse covariance or *precision* matrix Θ is tridiagonal (by Hammersley-Clifford). If we knew of this sparse structure a-priori we could bring high-dimensional sparse precision matrix estimation to bear on constructing a plug-in estimate for $\widehat{Q}$. This seems relevant to the Gabor wavelet features of natural images, but maybe not the most direct way to take advantage of structure.

1.2.4 Best Linear Predictor via Cross-Validation

The goal of each of the previous extensions of the Curds & Whey procedure was to take weighted combinations of the q separate LASSO predictions $\widehat{y}$ to share strength across problems. That is, we take $\widetilde{y} = \widehat{B}\widehat{y}$ where $\widehat{B}$ is a linear predictor for y based on $\widehat{y}$. Ideally, we would use the best linear predictor $B^* = \arg\min_B \mathbb{E}\|y - B\widehat{y}\|_2^2$, and we can construct an unbiased estimate of this using cross-validation. The recipe is straightforward:

1. Split the training data into $\mathcal{V} = 5$ equally sized subsets. For each $v = 1 : 5$,
2. Fit q LASSO problems separately using all the data except group v. Let $\widehat{y}(v)$ denote the predictions on the held-out data, and $y(v)$ the true values on that subset.
3. Fit q OLS problems separately. For $k = 1 : q$, regress $y_k(v)$ on $\widehat{y}(v)$. Write the regression coefficients in a $q \times q$ matrix $\widehat{B}(v)$. Take $\widehat{B} = \frac{1}{\mathcal{V}} \sum_{v=1}^{\mathcal{V}} \widehat{B}(v)$.

In small datasets, this will likely be highly variable, which can be mitigated to an extent with the choice of $\mathcal{V}$. We refer to this method as Best Linear Predictor via Cross-Validation (BLP-CV).

1.2.5 Simulations

In Figure 2 below, we compare each of these extensions to LASSO in terms of PTSE. The simulation design again follows Breiman & Friedman (1997); specifically, we use their method for generating a block of s non-zero coefficients in each problem, and then append $p - s$ zero coefficients.

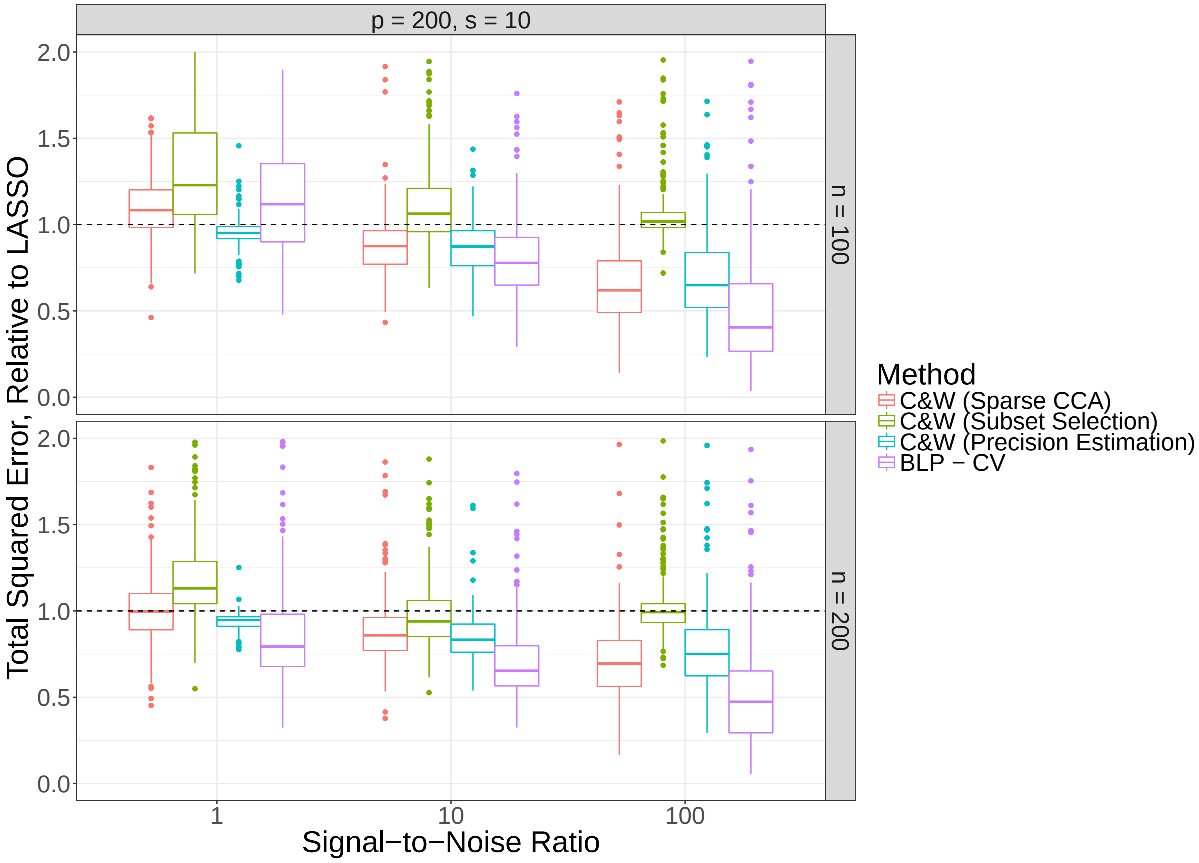

Figure 2: Simulation-based comparison of multi-response ($q = 5$) shrinkage methods in high dimensions p relative to the sample size n. The setup is the same as Figure 1 with additional $p - s$ zeros added. Again, each box plot was created using 250 runs. $< 1\%$ of runs went above 2 on the vertical axis.

Note since each method in Figure 2 returns a $q \times p$ matrix of the form $\widehat{B}\widehat{A}_{\text{lasso}}$, each is strictly slower to compute than LASSO, but the computational cost of estimating $\widehat{B}$ also varies. The Sparse Precision Estimation method is by far the fastest, since it assumes the model $(\Sigma_{xx})_{ij} = r^{|i-j|}$, estimates r via autocorrelation, and then computes the tri-diagonal inverse in closed form. Since this model is correct in our simulations, this yields a fairly reliable plugin estimate of the CCA matrix, even in high dimensions.

The next fastest method is running Curds & Whey on a subset of the covariates. This in part because, rather than running (10), we estimate the support directly from $\widehat{A}_{\text{lasso}}$. As implemented, subset selection does not do consistently better than LASSO, even when the SNR is **obscenely high**. This is surprising in light of Figure 1, since if we knew the support exactly, this method would even beat OLS run on the support (which definitely beats LASSO in prediction error). This result appears to underscore the difficulties of subset selection.

The matrix algebra in computing the ordinary CCA matrix ($q < p < n$) should be $\mathcal{O}(qp^2)$. The cost of this implementation of sparse CCA is unclear. The algorithm runs for a fixed number of iterations (15), and each update has a single matrix-vector product $\mathcal{O}(qs)$ where $s \ll p$ is the sparsity level as well as a grid search. In terms of wall-clock time, the Sparse CCA approach is the third fastest. As in the GCV approach to Curds & Whey, we run Sparse CCA $\mathcal{V} + 1 = 6$ times, and that's the bulk of the computational cost.

The BLP-CV approach is a clear front-runner in terms of relative improvement over LASSO but also computational cost. The algorithm takes roughly $\mathcal{V} + 1 = 6$ times as long to run as LASSO. In the top left of Figure 2 (relatively low signal and sample size), LASSO beats BLP-CV most of the time, so it's not always true that the added computational cost will be worth it.

1.3 Application to fMRI Study

A single subject observed a total of n = 1750 images, each of shape 128 × 128. After pre-processing each image using a Gabor wavelet pyramid, p = 10921 covariates remain.

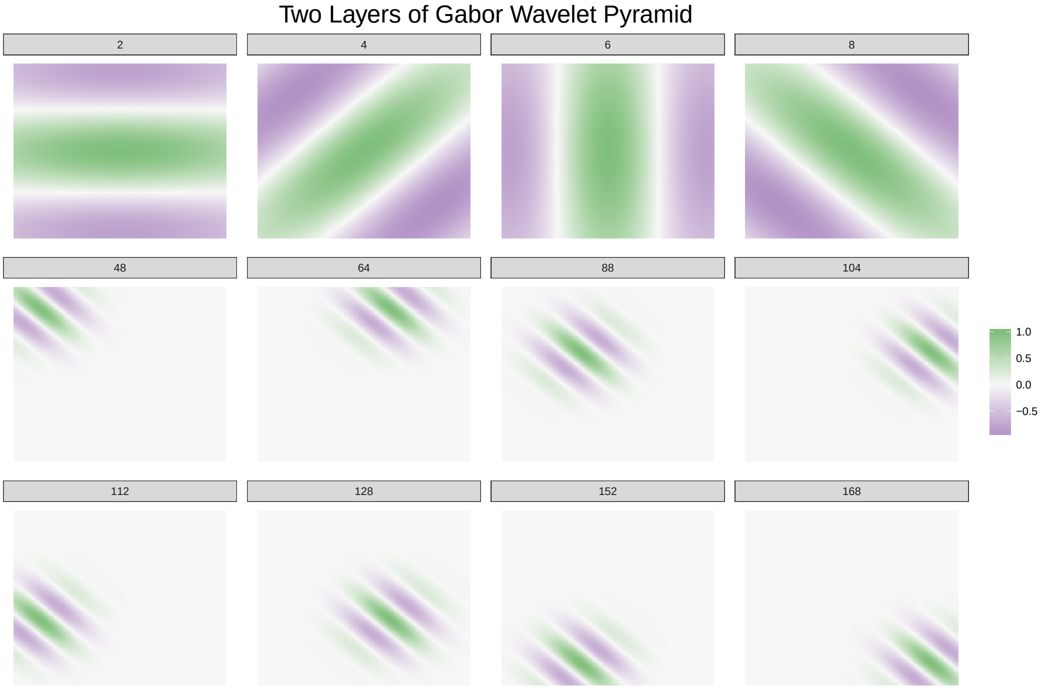

Figure 3: Understanding the features by the transforms they represent. Each feature is a coefficient from the image convolved with a wavelet basis element. Up in the pyramid, the wavelets depend on the entire 128 × 128 image; toward the bottom wavelets are localized to a small patch.

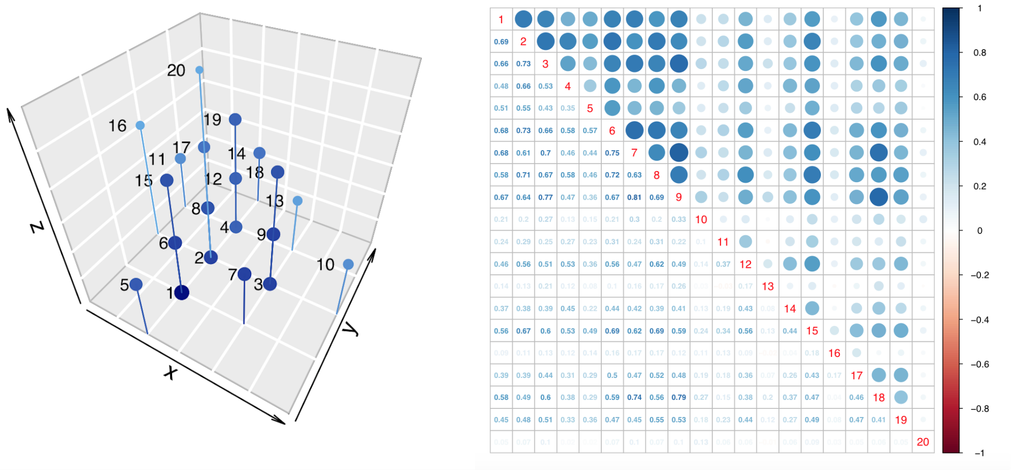

Figure 4: Left: Each voxel in relative spatial coordinates, colored (and sized) proportional to the correlation between that voxel's fMRI responses and those of voxel 1. Right: pairwise correlations between voxels. The first row of this thus gives the intensities for the 3D-plot on the left.

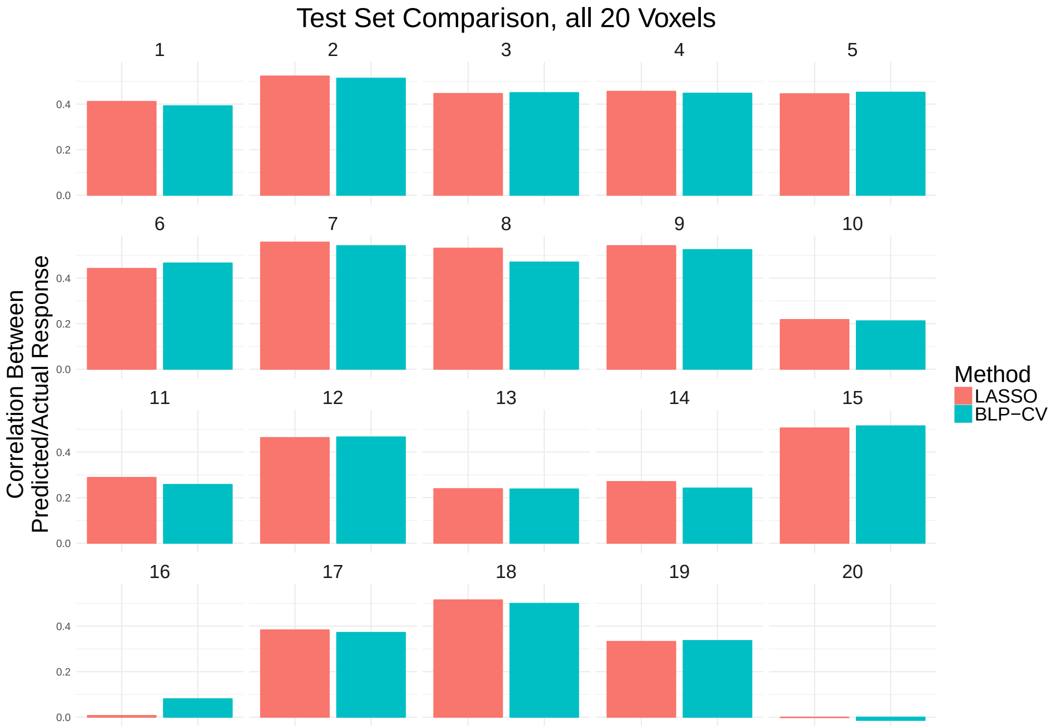

Figure 5: Model comparison using the test set. We report the correlation between fitted and observed fMRI response values for all voxels.

TO DO: add citations for BF 97, EM 72, Copas 83, and
Obozinski, G., Taskar, B., and Jordan, M. Multi-task feature selection. Technical report, Department of Statistics, University of California, Berkeley, 2006,
witten hastie tibs A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis