Voxelwise encoding models with non-spherical multivariate normal priors

Anwar O. Nunez-Elizalde[1], Alexander G. Huth[*1], and Jack L. Gallant[†1,2]

[1]Helen Wills Neuroscience Institute
[2]Department of Psychology
University of California, Berkeley, CA 94720, USA

Abstract

Predictive models for neural or fMRI data are often fit using regression methods that employ priors on the model parameters. One widely used method is ridge regression, which employs a spherical multivariate normal prior that assumes equal and independent variance for all parameters. However, a spherical prior is not always optimal or appropriate. There are many cases where expert knowledge or hypotheses about the structure of the model parameters could be used to construct a better prior. In these cases, non-spherical multivariate normal priors can be employed using a generalized form of ridge known as Tikhonov regression. Yet Tikhonov regression is only rarely used in neuroscience. In this paper we discuss the theoretical basis for Tikhonov regression, demonstrate a computationally efficient method for its application, and show several examples of how Tikhonov regression can improve predictive models for fMRI data. We also show that many earlier studies have implicitly used Tikhonov regression by linearly transforming the regressors before performing ridge regression.

Keywords: encoding models, computational neuroscience, fMRI, voxelwise modeling

1 Introduction

Cognitive and systems neuroscience has in recent years become increasingly reliant on predictive encoding models. In the fMRI literature, encoding models have produced insights into the cortical representations of visual (Thirion et al., 2006, Kay et al., 2008b, Nishimoto et al., 2011, Huth et al., 2012), auditory (De Angelis et al., 2017, de Heer et al., 2017), and linguistic (Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016) information. To estimate the parameters of encoding models, many studies use L2-regularized (ridge) regression (Hoerl and Kennard, 1970). Ridge regression has several advantages over classical ordinary

[*]Present address: Departments of Computer Science and Neuroscience, The University of Texas, Austin, TX 78751, USA
[†]Corresponding author: JLG (gallant@berkeley.edu)

least squares regression: it minimizes overfitting to noise, it improves model estimates for features that are nearly collinear, and models estimated by ridge regression usually generalize better to new data. While L2 regularization is by no means optimal in all cases, it is computationally efficient. From a probabilistic perspective, L2 regularization improves regression models by imposing a multivariate normal prior on the model parameters, where the mean of the prior is zero and the covariance is spherical. However, assuming a spherical covariance is rarely optimal, and in many cases prior information or expert knowledge can be used to construct informative non-spherical priors. Indeed, several fMRI studies have used non-spherical priors in order to improve the performance of decoding models (e.g. Chu et al. 2011, Grosenick et al. 2013). In this paper we explore how non-spherical priors can be applied to several encoding model problems and we show that this can greatly improve model performance. We also show that some previously published encoding models can be reinterpreted in terms of non-spherical priors, providing new insights into why those models were successful. Finally, we offer practical advice and efficient methods for estimating encoding models with non-spherical priors.

Although encoding models have proven highly successful for modeling fMRI data, there are several complications that make them difficult to use. First, in many feature spaces it is difficult to assign a specific interpretation to the feature dimensions. This problem is especially acute for feature spaces learned using unsupervised methods, such as the word embedding space word2vec (Mikolov et al., 2013). When using these feature spaces to predict neural or BOLD responses, it is difficult to interpret what exactly a given voxel represents.

Second, it is often unclear how the regularization method used for regression interacts with the choice of feature space. For example, feature spaces that are identical up to a linear transformation (i.e. $\mathbb{L}_1(s) = \mathbb{L}_2(s)P$) can yield drastically different results even though both span the same space.

Third, although the basic shape and variability of the hemodynamic response function (HRF) are understood reasonably well (Boynton et al., 1996, Büchel et al., 1998, Friston et al., 1998, Glover, 1999), many studies do not use this prior information when estimating the HRF (Kay et al., 2008a, Nishimoto et al., 2011, Huth et al., 2012). In fact, most studies simply assume a single canonical HRF for all voxels (Penny et al., 2011).

Fourth, studies that explore high-dimensional stimulus or task spaces require methods that can recover multiple models simultaneously (Lescroart et al., 2015, de Heer et al., 2017). This is often accomplished by combining different feature spaces into one encoding model. However, this procedure assumes that all feature spaces require the same level of regularization, and this assumption is often incorrect.

Here we address all of these issues by constructing encoding models using carefully designed multivariate normal priors. In the standard encoding model formulation, complex features are extracted from the stimuli and then regularized regression is used to learn model parameters subject to simple spherical priors. In the new framework presented here, we extract simple, interpretable features from the stimuli, and then use Tikhonov regression (Tikhonov et al., 1977) to learn parameters subject to complex multivariate normal priors. This is made possible by a duality between imposing a multivariate normal prior and extracting features

from the stimuli, so the exact same model can be represented in both ways. This framework is also highly modular, making it easy to combine different feature and temporal MVN priors, and to test many different kinds of MVN priors. Note that the Tikhonov framework presented here merely increases the range of tools available to researchers. We do not argue that it is an optimal approach for every experimental design and dataset.

We evaluate each proposed application of our framework on empirical data from naturalistic experiments on vision and language. We show that non-spherical multivariate normal priors can improve prediction accuracy in a variety of settings. In order to encourage the adoption of the framework presented here, we have released an open-source Python software package that efficiently implements all the models described in this paper (`http://github.com/gallantlab/tikreg`).

2 Linearized predictive encoding models

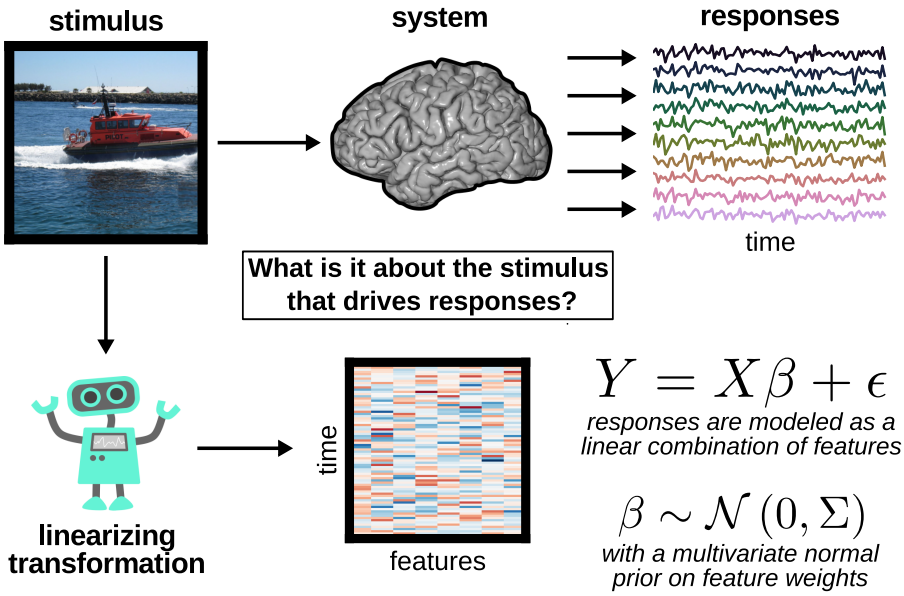


Figure 1. Modeling the stimulus-response relationship with linearized encoding models and multivariate normal priors. In a typical fMRI experiment a series of stimuli are shown to a subject while their brain activity is recorded. In the voxelwise encoding model framework, features are first extracted from the stimuli using a computational model, human labels, or by any other method. Brain activity is then modeled as a weighted, linear combination of the features. Some form of regularization is usually required when a large number of features are used or when the signal-to-noise ratio is low. The most common method of regularization is to impose a prior distribution on the feature weights. When the distribution is a spherical multivariate normal (MVN), this is called ridge regression. A spherical MVN distribution assumes equal variance in all dimensions and zero covariance between dimensions. However, a spherical MVN prior is not always appropriate. In such cases, model performance can be improved by constructing non-spherical MVN priors that incorporate expert knowledge about the problem. Encoding models with non-spherical MVN priors can be estimated using Tikhonov regression.

In a typical fMRI experiment, n brain images $y(t) \in \mathbb{R}^v$ are recorded at times $t = 0 \dots \tau$ while a subject is exposed to stimuli $s(t)$ (Figure 1). Each brain image consists of v voxels, $y_\ell(t)$ for $\ell = 1 \dots v$. The goal of the voxelwise encoding model framework is to find a function f_ℓ that maps stimuli to BOLD responses in each voxel: $f_\ell(s(1), \dots, s(t)) \approx y_\ell(t)$. Because the space of possible functions is extremely large, it is common to work under a hypothesis that limits the complexity of f. Although there often are many reasonable hypotheses that one can make about f, the only type that we shall consider here is where f is a linear combination of features that are extracted from the stimulus, usually by a nonlinear function. In this case, f is called a "linearized" model, and the function that extracts features from the stimulus, $\mathbb{L}_s(t) \in \mathbb{R}^{1\times p}$, is called the "linearizing transformation" (Wu et al., 2006). Formally, $\mathbb{L}_s(t)$ maps a u-dimensional stimulus at time t into a p-dimensional vector of stimulus features $x(t)$:

$$\mathbb{L}_s(t) : s(t) \in \mathbb{R}^{1\times u} \mapsto x(t) \in \mathbb{R}^{1\times p}$$

Under the linearized model formulation, the brain response is modeled as a linear combination of the stimulus features, usually over a fixed time window of length d in order to account for the hemodynamic delay,

$$y_\ell(t) = \left[\begin{array}{cccc} x(t) & x(t-1) & \dots & x(t-D) \end{array}\right] \beta_\ell + \epsilon_\ell(t),$$

where $\epsilon_\ell(t) \sim \mathcal{N}(0, \sigma_\ell^2)$ is stationary, zero-mean normal noise, $x(t-j) \in \mathbb{R}^{1\times p}$ is the feature vector delayed j time points, and $\beta_\ell \in \mathbb{R}^{pd\times 1}$ is a set of linear weights over the p features at each of the d delays $(0, 1, \dots, D)$.

To write the simultaneous equation for all voxels we replace $y_\ell(t)$ with a matrix $Y \in \mathbb{R}^{n\times v}$ that contains the response of each voxel at each timepoint, and we replace β_ℓ with a matrix $\beta \in \mathbb{R}^{pd\times v}$ that contains the weight vector for every voxel. We write the matrix of linearized stimulus features as $X \in \mathbb{R}^{n\times pd}$,

$$X = \left[\begin{array}{cccc} x(0) & 0_p & \cdots & 0_p \\ x(1) & x(0) & \cdots & 0_p \\ \vdots & \vdots & \vdots & \vdots \\ x(t) & x(t-1) & \cdots & x(t-D) \\ \vdots & \vdots & \vdots & \vdots \\ x(\tau) & x(\tau-1) & \cdots & x(\tau-D) \end{array}\right] = \left[\begin{array}{cccc} X_{\delta(0)} & X_{\delta(1)} & \cdots & X_{\delta(D)} \end{array}\right], \qquad (1)$$

where each row of X contains the feature vectors for the past d timepoints. Each block of p columns $X_{\delta(j)}$ contains the linearized stimulus feature matrix delayed by j time points. This is referred to as a finite impulse response model (Oppenheim et al., 1983). This allows us to rewrite the basic model as:

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

where $\epsilon_t \sim \mathcal{N}_v(0, \mathrm{diag}\{\sigma_1^2, \dots, \sigma_v^2\})$ is zero-mean, independent noise for each voxel at each time point. The only free parameter in this formula is the weight matrix $\beta \in \mathbb{R}^{pd\times v}$.

We can find an estimate of β by maximizing the probability of the data Y given the stimulus features X

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

This estimate of β is called the maximum likelihood estimate (MLE). We can derive various analytic solutions depending on the form of the distribution we assume for $P(Y|X,\beta)$. In this paper, we assume that the responses can be modeled as multivariate normal random variables.

The likelihood of the data can be expressed as

$$P\left(Y|X,\beta\right) \propto \frac{1}{\det\left(\Sigma_\epsilon\right)}\exp\left(-\frac{1}{2}\operatorname{trace}\left(\left(Y-X\beta\right)^\top \Sigma_\epsilon^{-1}\left(Y-X\beta\right)\right)\right),$$

where $\Sigma_\epsilon = \operatorname{diag}\{\sigma_1^2,\ldots,\sigma_v^2\}$ contains the variance of the noise for each voxel. If we assume that the noise is temporally independent and that the noise variance is the same in each voxel, then we can set $\Sigma_\epsilon = \sigma^2 I$. We can also switch to using the log of the likelihood instead of the likelihood. The log-likelihood of the data can then be expressed as

$$\log P\left(Y|X,\beta\right) \propto -\frac{1}{2}\sum_{\ell}^{m}\left(\frac{1}{\sigma^2}\left\|y_\ell - X\beta_\ell\right\|_2^2\right).$$

Finally, note that maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood of the data. The β estimate for all voxels can be found simultaneously by solving

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\ \left[\frac{1}{2}\left\|Y - X\beta\right\|_2^2\right].$$

From an optimization perspective, this can be viewed as finding the β that minimizes the squared difference between predicted and actual responses (i.e. ordinary least squares regression).

However, finding the value of β that exactly minimizes this squared error function often produces results that do not generalize to new stimuli. This is due to overfitting. Overfitting occurs when model parameters capture the noise in addition to the underlying signal. This is a common problem when the data available to estimate the model parameters is small. To avoid overfitting it is common to employ regularized regression techniques (Friedman et al., 2001).

From a probabilistic perspective, regularization can be thought of as imposing a prior distribution on β. This prior distribution limits how well a model can explain the given data. Under this perspective the goal is to maximize the probability of the observed data, by finding the β that maximizes the product of the likelihood and the prior distributions

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

This estimate of β is called the maximum a posteriori (MAP) estimate. Note that in some situations, an estimate of the full posterior distribution of β might be of interest. In such

cases, Bayesian methods can be used. In this paper, however, we only consider the MAP estimate of the posterior distribution, $\hat{\beta}$. Also, while many prior distributions can be chosen, we will only consider prior distributions of multivariate normal form.

2.1 Ridge regression and spherical multivariate normal priors

Perhaps one of the simplest priors that can be imposed on the feature weights β_ℓ is a zero-mean multivariate normal distribution with spherical covariance,

$$\beta_\ell \sim \mathcal{N}_p(0, \lambda^{-2} I_p).$$

In the statistics literature, this approach is called ridge regression, and it has a long history that predates its probabilistic interpretation (Hoerl and Kennard, 1970).

Ridge regression can be implemented by adding a penalty term to the error function, where the penalty is proportional to the sum of the squared weights,

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

and the strength of the regularization is controlled by λ, the regularization parameter. The closed-form solution for the ridge regression problem is given by

$$\hat{\beta} = (X^\top X + \lambda^2 I)^{-1} X^\top Y.$$

2.2 Tikhonov regression and non-spherical multivariate normal priors

From a probabilistic perspective, ridge regression imposes a zero-mean, spherical multivariate normal prior on the feature weights. However, expert knowledge can be used to create a more sophisticated, non-spherical multivariate normal prior on the weights,

$$\beta \sim \mathcal{N}_p(0, \lambda^{-2}\Sigma),$$

where $\Sigma \in \mathbb{R}^{p \times p}$ is the positive definite prior covariance matrix. Note that λ is present as a scaling factor on Σ. This determines how much influence the prior distribution has on the estimated weights.

If we factorize the inverse of the MVN prior covariance matrix by taking its matrix square root $\Sigma^{-1} = C^\top C$, where $C \in \mathbb{R}^{p \times p}$, then we can express the prior probability of β as

$$P(\beta) \propto \exp\left(-\frac{\lambda^2}{2}\beta^\top C^\top C \beta\right) = \exp\left(-\frac{1}{2}\|\lambda C\beta\|_2^2\right)$$

The problem can then be solved by maximizing the product of the likelihood and this new MVN prior, or, as above, by minimizing the negative log likelihood,

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

This is known as Tikhonov regression and it has a long history in the statistics literature that predates its probabilistic interpretation (Tikhonov et al., 1977). Here C can be thought of as a penalty matrix that punishes β when it does not conform to the MVN prior. However, since there are many matrix square roots, C is not uniquely determined by the MVN prior covariance, and in fact any C that satisfies the given relation will produce the same $\hat{\beta}_T$. Also note that when $C = I_p$ Tikhonov regression reduces to ridge regression (Hoerl and Kennard, 1970).

The Tikhonov minimization problem has a closed form solution,

$$\hat{\beta}_T = (X^\top X + \lambda^2 C^\top C)^{-1} X^\top Y.$$

However, this formulation does not immediately admit efficient computational solutions, making it less useful for solving large-scale problems. Fortunately there is a computationally efficient method for solving Tikhonov regression problems. This method, which is often referred to as the "standard form" (Hansen, 1998), transforms a Tikhonov problem into a ridge regression problem. This transformation is accomplished in three steps.

First, a linear transformation is applied to X, giving

$$A = XC^{-1}.$$

Second, ridge regression is carried out with A, giving

$$\hat{\beta}_A = (A^\top A + \lambda^2 I_p)^{-1} A^\top Y.$$

Third, the estimated weights are projected back into the original space to give the Tikhonov estimate,

$$\hat{\beta}_T = C^{-1} \hat{\beta}_A.$$

(For a proof of this see Appendix A). Because the standard form uses ridge regression internally, it is clear that Tikhonov regression in the standard form will admit the same efficient computational solutions as ridge regression.

The standard form transformation can be used to convert any Tikhonov regression problem into a ridge regression problem by way of a linear transformation of X. By the same logic, any linear transformation of X followed by ridge regression is equivalent to some Tikhonov regression problem, and thus some non-spherical MVN prior on the model weights. This relationship has interesting implications for a number of neuroimaging studies that have applied ridge regression to linearly transformed stimuli, because the models employed by those studies can be re-interpreted as Tikhonov regression with non-spherical MVN priors. We use this technique to explore and re-interpret the models used in some previous studies.

3 Using multivariate normal priors

3.1 Feature priors

3.1.1 Word embeddings

Several earlier studies have used word embedding spaces to model how the brain represents the meaning, or semantic content, of words (Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016). In this approach, each word is converted into a vector with anywhere from 20 (Mitchell et al., 2008) to 1000 (Huth et al., 2016) embedding dimensions. These vectors are constructed using word co-occurrence statistics from large corpora of text (Turney and Pantel, 2010), and are designed such that words with similar or related meanings (such as 'month' and 'week') are assigned similar vectors, but words with dissimilar meanings (such as 'month' and 'tall') are not. After converting words to vectors, regression models are used to predict BOLD responses as a function of the embedding dimensions.

Formally, this approach starts by defining a matrix of word indicator variables $W \in \mathbb{R}^{T \times p}$, where $W_{t,i} = 1$ if word i was presented at time t and 0 otherwise. Here T is the total number of time points and p is the total number of words in the experiment. Then, in order to replace each word with its q-dimensional embedding vector, the indicator matrix is multiplied with an embedding matrix $E \in \mathbb{R}^{p \times q}$ whose rows contain the word embedding vectors. Finally, regression is performed in the embedding space, yielding the linear model

$$Y = (WE)\beta + \epsilon.$$

Interestingly, this formulation appears identical to the standard form transformation of Tikhonov regression (see Appendix A). If the model weights, β, are estimated using ridge regression (Wehbe et al., 2014, Huth et al., 2016), then this approach is equivalent to Tikhonov regression where the features are word indicators (i.e. the feature matrix is W), and the MVN prior covariance is given by dot products between the embedding vectors, $\Sigma = EE^\top$ (and $C^{-1} = E$).

Thus, the word embedding approach is equivalent to putting a multivariate normal prior on the model weights across words, such that the MVN prior covariance between weights for different words is equal to the dot product between their embedding vectors. If words that have similar meanings have similar embedding vectors, then the dot product between those vectors will be high, and the weights for those words will covary strongly. This reinterpretation of the word embedding approach seems in many ways to be more natural and intuitive than thinking of it as regression in the word embedding space, which is highly abstract and difficult to explain.

3.1.2 Evaluating MVN feature priors

To illustrate the Tikhonov approach to word embeddings we estimated two different linear models using data from a language experiment (Huth et al., 2016). Both models used

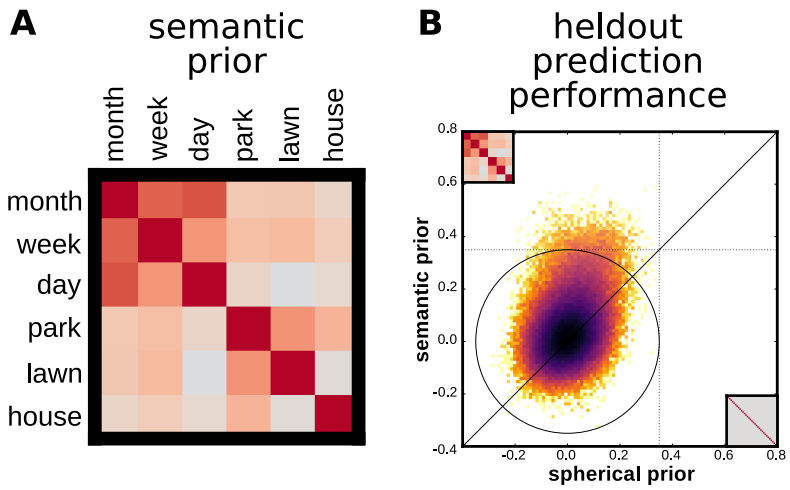


Figure 2. Comparison of non-spherical and spherical MVN feature priors in a language experiment. To evaluate the use of non-spherical versus spherical MVN feature priors, separate encoding models employing non-spherical and spherical MVN priors were estimated using data collected while two subjects listened to spoken stories and had their brain activity measured with fMRI. Brain activity was modeled as a linear combination of the words spoken in the stories. The only difference between the two encoding models is the MVN prior imposed on the covariance of features. **(A)** The non-spherical prior was constructed from word co-occurrence statistics estimated using a large corpus of narrative text. This non-spherical prior reflects the assumption that words that occur close together are likely to be semantically related. A voxelwise encoding model was estimated using this non-spherical MVN feature prior in order to explicitly model the semantic content of the words spoken in the stories. We refer to this non-spherical MVN prior as a semantic prior. A separate encoding model was also estimated using a spherical MVN prior (i.e. ridge regression), which assumes that words are equally related to one another independent of their semantic meaning. **(B)** Comparison of model prediction accuracy under the semantic prior (i.e. Tikhonov regression) and the spherical MVN prior (i.e. ridge regression). Model prediction accuracy was assessed by computing the correlation between predicted and observed voxel responses evoked by a novel, held-out story. The prediction accuracy was significantly higher ($p < 10^{-12}$, Wilcoxon signed-rank test) under the semantic prior (mean Pearson correlation r=0.037) than the spherical MVN prior (mean Pearson correlation r=0.005). This suggests that brain responses to the stories are more accurately predicted by an encoding model that includes information about the semantic meaning of words.

words as features, but one model applied a spherical MVN prior to the weights while the other applied a semantic similarity MVN prior based on a word embedding space. Data were collected while two subjects listened to approximately two hours of naturally spoken narrative stories and had their brain activity measured with fMRI. The stories were transcribed and then the transcripts were aligned to the audio to determine exactly when each word was spoken. These aligned transcripts were then used to generate the word indicator matrix, W. The word indicator matrix was then downsampled to the BOLD acquisition rate (2.0045 seconds) and delayed from 0 to 20 seconds in order to account for the HRF. This resulted in the word feature matrix X.

The first linear model was estimated separately for each voxel in the fMRI scan using a spherical MVN prior on the model weights:

$$Y = X\beta_X + \epsilon$$

$$\hat{\beta}_X = (X^\top X + \lambda^2 I)^{-1} X^\top Y$$

The second linear model was estimated using a non-spherical MVN prior based on a word embedding space. We refer to this non-spherical MVN prior as a semantic prior. This embedding space was constructed by computing the statistical co-occurrence of each word in the stories with 985 common English words (see Huth et al., 2016, for details). To apply the semantic prior, the word feature matrix, X, was projected onto the embedding matrix, E, at each delay, and then ridge regression was used to estimate the weights:

$$Y = XE\beta_E + \epsilon$$

$$\hat{\beta}_E = ((XE)^\top (XE) + \lambda^2 I)^{-1} (XE)^\top Y$$

This solution can be projected from the embedding space back to the original space of words to yield the Tikhonov solution $\hat{\beta}_T = E\hat{\beta}_E$, which is arguably easier to interpret.

Finally, we used both sets of weights to predict BOLD responses on a separate 10-minute story that had not been used for model estimation, and then computed the correlation between predicted and actual BOLD responses. This model evaluation procedure resulted in two correlation coefficients for each voxel: one for the spherical MVN prior and one for the semantic prior. To compare these values we aggregated the data from both subjects, and then computed a 2D histogram of the correlation values (Figure 2).

Figure 2 shows that model prediction accuracy is nearly always higher with the semantic prior than with the spherical MVN prior, often substantially so. Of approximately 150,000 voxels included in the analysis, about 300 were significantly predicted by the spherical MVN prior model, and about 15,000 were significantly predicted by the semantic prior model (FDR-corrected significance test of correlation coefficients, n=291, $q(FDR) < 0.05$; Benjamini and Hochberg 1995). The difference in model prediction accuracy between the semantic (mean prediction accuracy r=0.037 $\pm$ 0.0002 s.e.m.) and spherical MVN prior (mean prediction accuracy r=0.005 $\pm$ 0.0002 s.e.m.) models is large and significant ($p < 10^{-12}$, Wilcoxon signed-rank test). In the worst cases, we found that some voxels were predicted about equally well by both models.

These results suggest that the semantic prior is a much better reflection of the true underlying voxel weights than the spherical MVN prior, and thus supports the earlier conclusion that those voxels represent information about the semantic content of language (Mitchell et al., 2008, Wehbe et al., 2014, Huth et al., 2016).

3.2 Temporal priors

The temporal activation pattern of the BOLD response is referred to as the hemodynamic response function (HRF). While the neurovascular mechanisms underlying the HRF are not well understood, the shape of the HRF has been extensively studied in humans (Boynton et al., 1996, Glover, 1999). At a first approximation, the neural activation evoked by a stimulus leads to changes in blood-oxygenation that peak 4 to 6 seconds after stimulus onset. Several studies have shown that the shape of the HRF is highly variable across voxels and brain regions both within and across subjects (Aguirre et al., 1998, Handwerker et al.,

2004, Kay et al., 2008a). It is important to take this variability into account by estimating the shape of the HRF for each voxel when modeling BOLD responses.

A common method for estimating the HRF is to use a finite impulse response (FIR) model (Kay et al., 2008a). In an FIR model brain responses are modeled as a linear combination of (p) features over a fixed time window (d) before stimulus onset (see Equation 1). The number of parameters in an FIR model is much larger ($p \times d$) than the original number of features (p), and grows linearly with the length of the time window d. This increase in the number of parameters can lead to overfitting. In order to reduce overfitting, it is important to regularize FIR models.

When ridge regression is used to estimate FIR models, the implicit assumption is that feature weights are independent across time. This happens because ridge regression imposes a spherical MVN prior on the temporal covariance of each feature, $\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2} I_d\right)$, where $\beta_i \in \mathbb{R}^d$ is the vector of weights for feature i across the time window. In the Tikhonov framework, we can relax this assumption by specifying MVN temporal priors that are not spherical,

$$\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2}\Sigma^T\right).$$

This approach has been previously advocated in the context of Bayesian modeling (Woolrich et al., 2004). An insight worth highlighting is that applying a MVN temporal prior is equivalent to convolution followed by ridge regression (Appendix B). This follows from the fact that FIR models can be understood as convolution. In the context of Tikhonov regression, this means that applying a MVN temporal prior with covariance of the form $\Sigma^T = (C^\top C)^{-1}$ is equivalent to convolving each feature timecourse with a set of temporal filters given by the columns of C^{-1}. When $C^{-1} = I$ the features are convolved with Kronecker delta functions at different delays, which is identical to using delays.

3.2.1 Smoothness temporal prior

One simple and widely studied MVN temporal prior holds that feature weights are smooth across time. This type of MVN prior is typically applied by defining the penalty matrix $\mathbb{D} \in \mathbb{R}^{d \times d}$ to be a discrete difference operator that penalizes differences between neighboring weights in time,

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

In the Tikhonov framework, this corresponds to a multivariate normal prior with covariance $\mathbb{D}^{-2}$ (Wu et al., 2006):

$$\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2}\mathbb{D}^{-2}\right).$$

This and similar approaches have been used in several studies (Goutte et al., 2000, Marrelec et al., 2003, Casanova et al., 2008, Bazargani and Nosratinia, 2014).

3.2.2 HRF temporal prior

A more empirically-grounded possibility is to use published mathematical descriptions of the HRF to form a MVN temporal prior (Boynton et al., 1996, Friston et al., 1998, Glover, 1999). Previous work has resulted in the characterization of the commonly used "canonical" HRF. This canonical HRF (h_1), its temporal derivative (h_2), and its derivative with respect to time-to-peak (dispersion; h_3) together provide an informed basis set that can capture some of the empirical variation observed in HRF shapes (Friston et al., 1998). The basis set is a matrix $\mathbb{H} \in \mathbb{R}^{d \times 3}$

$$\mathbb{H} = \begin{bmatrix} | & | & | \\ h_1 & h_2 & h_3 \\ | & | & | \end{bmatrix},$$

where each h_i is a basis vector of length d. However, this basis set is not always flexible enough to capture all voxel- or region-specific variability of the HRF (Woolrich et al., 2004, Kay et al., 2008a, Pedregosa et al., 2015). In such cases, an FIR model might provide a better estimate of the the shape of the HRF. In practice, however, the FIR model might be difficult to estimate correctly because the large number of parameters ($p \times d$) can lead to overfitting.

The Tikhohnov framework offers an intermediate approach that trades off between FIR and HRF-based models. To achieve this, we compute the dot product of the HRF temporal basis set and use it as a non-spherical MVN temporal prior on the feature weights,

$$\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2}\mathbb{H}\mathbb{H}\mathbb{H}^\top\right).$$

We refer to this as the HRF temporal prior. As λ decreases, the effect of the MVN prior on the FIR weights is minimal. On the other hand, as λ increases, the MVN prior has more effect on the FIR weights.

3.2.3 Evaluating temporal priors

In order to evaluate and compare these MVN temporal priors we estimated three encoding models. The first encoding model was estimated with ridge regression, which imposes a spherical MVN temporal prior ($\Sigma^T = I_d$). In the second model, we used Tikhonov regression to impose a smoothness temporal prior on the FIR delays ($\Sigma^T = \mathbb{D}^{-2}$). Finally, in a third model we imposed a MVN temporal prior constructed from the covariance of an HRF basis set ($\Sigma^T = \mathbb{H}\mathbb{H}\mathbb{H}^\top$; Friston et al. 1998). All models had the same number of parameters and only differed in the MVN temporal prior used.

We used data from a previously published vision experiment (Huth et al., 2012). Data were collected while three subjects watched several hours of natural movies and had their brain activity measured with fMRI. A total of 6,555 motion energy features were extracted from these movies using a three-dimensional Gabor filter pyramid (Adelson and Bergen, 1985, Watson and Ahumada, 1985, Nishimoto et al., 2011). We used 10 temporal delays in order to account for the HRF (0-20 seconds). This resulted in an FIR model with a total of 65,550

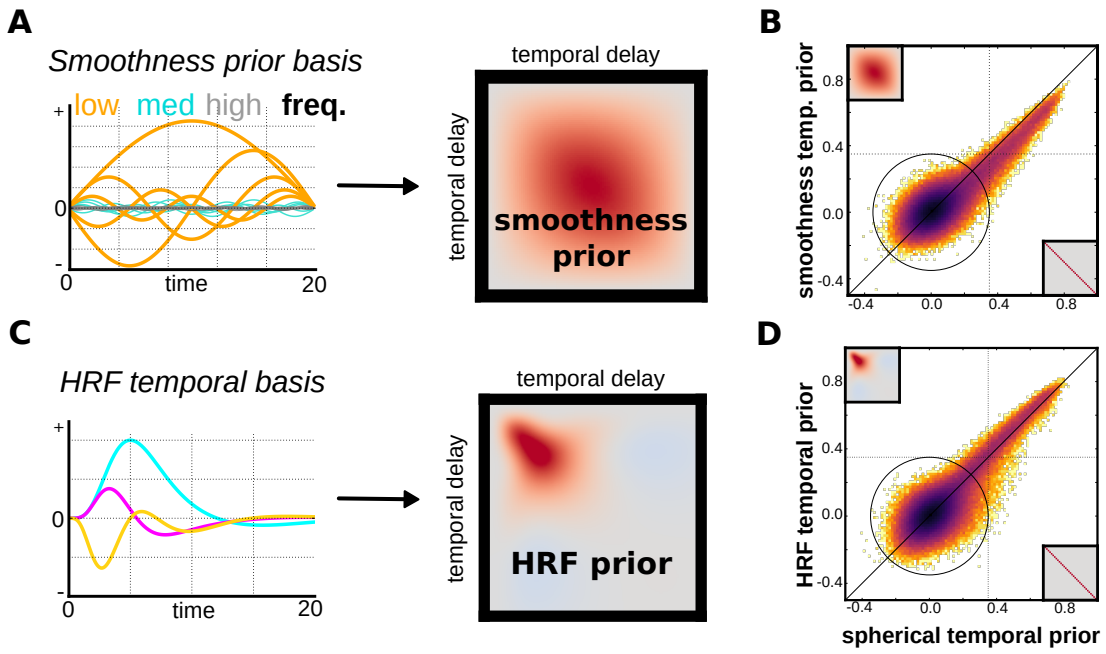


Figure 3. Comparison of non-spherical and spherical MVN temporal priors in a vision experiment. To evaluate the use of non-spherical versus spherical MVN temporal priors, three separate encoding models were estimated using data collected while three subjects watched several hours of natural movies and had their brain activity measured with fMRI. Brain activity was modeled as a linear combination of motion energy features using ten temporal delays in order to account for the HRF. The only difference between the three encoding models is the MVN prior imposed on the covariance of temporal delays. **(A)** A non-spherical MVN prior that enforces similarity between neighboring temporal delays was constructed using a second order difference operator. This non-spherical MVN temporal prior captures the belief that BOLD responses are smoothly varying in time and we therefore refer to it as a temporal smoothness prior. **(B)** Comparison of model prediction accuracy under the temporal smoothness prior and the spherical MVN temporal prior (i.e. ridge regression). Model prediction accuracy was assessed by computing the correlation between predicted and observed voxel responses evoked by a novel, held-out stimulus. Because the temporal smoothness prior enforces high covariance in the middle of the HRF time-course, it does not improve prediction accuracy relative to a spherical MVN temporal prior. **(C)** An additional non-spherical MVN prior was constructed using the temporal covariance of three HRF basis functions. This non-spherical MVN temporal prior reflects the assumption that the HRF can be characterized by a linear combination of the three basis functions. We refer to this prior as an HRF temporal prior. **(D)** Comparison of model prediction accuracy under the HRF temporal prior and the spherical MVN temporal prior (i.e. ridge regression). The HRF temporal prior improves prediction accuracy in well-predicted voxels relative to the spherical MVN prior. This suggests that voxel responses can be better modeled by including information about the shape of the HRF. However, the improvement in prediction accuracy is small.

channels and 3,600 time points. We selected the regularization parameter, λ, using a cross-validation procedure (5-fold cross-validation repeated 20 times). This was done separately per voxel for each of the three encoding models estimated. We evaluated model performance for each model by computing the correlation coefficient between predicted and actual BOLD responses on a held-out dataset, which was not used for estimation. The held-out dataset consisted of 270 samples and was constructed by taking the mean temporal BOLD signal across 10 repetitions of a 540 second movie (Schoppe et al., 2016)

A total of approximately 230,000 voxels from three subjects were used in the analyses (Figure 3). We found that the HRF temporal prior provided better prediction accuracy than either the spherical MVN temporal prior or the smoothness temporal prior for the best voxels in the population ($p's < 10^{-12}$, Wilcoxon signed-rank test; top 10,000 voxels). The differences in mean prediction accuracy in the top 10,000 voxels for the models estimated with the HRF (mean prediction accuracy r=0.55 $\pm$ 0.001 s.e.m.), spherical (mean prediction accuracy r=0.54 $\pm$ 0.001 s.e.m.) and smoothness (mean prediction accuracy r=0.45 $\pm$ 0.001 s.e.m.) MVN priors were small but consistent. However, across the total population of voxels the spherical MVN prior yielded better prediction accuracy ($p's < 10^{-12}$, Wilcoxon signed-rank tests). These results suggest that the HRF temporal prior has a small but consistent advantage for well-predicted voxels.

4 Combining feature and temporal priors

When both feature and temporal MVN priors are available, they can be used to construct a single feature-temporal multivariate normal prior. Feature-temporal MVN priors allow us to incorporate prior information about the covariance of the feature weights and the covariance of the temporal delays when estimating predictive encoding models. However, as the number of features (p) and temporal delays (d) increase, the covariance of the feature-temporal MVN prior becomes large ($(p \times d)^2$). This makes the estimation of encoding models with non-spherical MVN feature-temporal priors impractical for neuroimaging. In this section, we present a solution to that makes the estimation of these models tractable when $n < p$.

The feature-temporal MVN prior is constructed by computing the Kronecker product ($\otimes$) between the feature prior $\Sigma^X \in \mathbb{R}^{p \times p}$ and the temporal prior $\Sigma^T \in \mathbb{R}^{d \times d}$,

$$\Sigma = \Sigma^T \otimes \Sigma^X = \begin{bmatrix} \Sigma^T_{1,1}\Sigma^X & \cdots & \Sigma^T_{1,d}\Sigma^X \\ \vdots & \ddots & \vdots \\ \Sigma^T_{d,1}\Sigma^X & \cdots & \Sigma^T_{d,d}\Sigma^X \end{bmatrix}.$$

The resulting feature-temporal prior is $\Sigma \in \mathbb{R}^{pd \times pd}$. Notice that when both the feature and the temporal priors are spherical, the feature-temporal prior is also spherical.

The Tikhonov solution to an encoding model with a feature-temporal multivariate normal prior $\Sigma^T \otimes \Sigma^X$ can be expressed as (see Appendix C):

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \left(X \left(\Sigma^T \otimes \Sigma^X\right) X^\top + \lambda^2 I\right)^{-1} Y.$$

This is equivalent to the ridge regression solution when both MVN priors are spherical ($I_d \otimes I_p = I_{pd}$). However, computing this solution involves constructing an extremely large $(p \times d)^2$ feature-temporal MVN prior covariance matrix, which we would like to avoid. Luckily, the properties of the Kroenecker product allow us to derive a computationally efficient solution in cases where $n < p$ (Appendix D). This formulation makes it tractable to fit large encoding models with non-spherical MVN feature-temporal priors.

4.1 Evaluating feature-temporal MVN priors

To illustrate the power of feature-temporal MVN priors, we estimated four different encoding models using data from a language experiment (Huth et al., 2016). Data were collected while two subjects listened to spoken stories and had their brain activity measured with fMRI. We modeled voxel responses to the stimulus as a linear combination of words, and estimated models that differed only in the feature-temporal MVN prior used:

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

$$\beta \sim \mathcal{N}_{pd}\left(0, \lambda^{-2}\Sigma^T \otimes \Sigma^X\right).$$

The first and simplest model we evaluated was ridge regression. Ridge regression corresponds to a feature-temporal MVN prior where both feature and temporal MVN priors are spherical ($I_d \otimes I_p$). The second model used a word embedding MVN prior Σ^X constructed from word co-occurrence statistics estimated from a large text corpus (described above), and a spherical MVN temporal prior ($I_d \otimes \Sigma^X$). The third model was constructed using a spherical MVN feature prior and an HRF temporal prior constructed from a set of HRF basis functions ($\Sigma^T \otimes I_p$). Finally, the fourth model evaluated used a feature-temporal prior that combines both the word embedding feature prior and the HRF temporal prior ($\Sigma^T \otimes \Sigma^X$). We refer to this last non-spherical MVN feature-temporal prior as the semantic-temporal prior.

The models were constructed using 10 TR temporal delays (20 seconds) in order to account for the hemodynamic lag. A MVN temporal prior $\Sigma^T \in \mathbb{R}^{10\times10}$ was constructed from the temporal covariance of an HRF basis set during the same time period. The FIR matrix X was built using 10 temporal delays for each of the 3,000 channels. This resulted in an FIR feature matrix with a total of 30,000 features and 3,737 time points.

We found that the model estimated with the semantic-temporal prior performed better than the same model estimated with either the semantic or the HRF temporal prior on their own (Figure 4). From a total of about 150,000 voxels, approximately 22,500 were significantly predicted ($n = 291$, $q(FDR) < 0.05$) when using the semantic-temporal prior (mean prediction accuracy r=0.045 $\pm$ 0.0003 s.e.m.), approximately 5,500 with the HRF temporal prior (mean prediction accuracy r=0.019$\pm$0.0002 s.e.m.), and approximately 15,000 with the semantic prior (mean prediction accuracy r=0.037 $\pm$ 0.0002 s.e.m.). The semantic-temporal prior performed much better than the HRF temporal prior model alone ($p < 10^{-12}$, Wilcoxon signed-rank test). This is not surprising since the semantic-temporal prior includes the semantic prior and that on its own improves prediction accuracy (see Figure 2). However, we found that the semantic-temporal prior improved prediction accuracy over and above the semantic prior alone ($p < 10^{-12}$, Wilcoxon signed-rank test). In sum, we can gain the best from both worlds by combining feature and temporal priors into a single feature-temporal prior and thereby improve the prediction accuracy of encoding models.

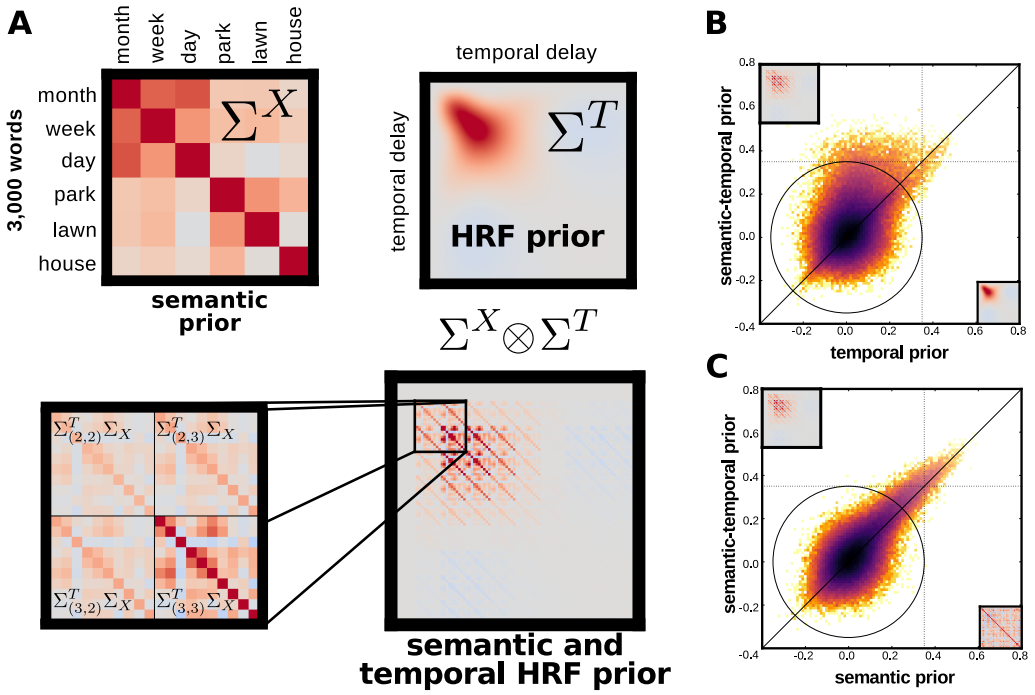


Figure 4. Evaluation of a feature-temporal MVN prior constructed by combining non-spherical feature and temporal priors in a language experiment. To evaluate the use of feature-temporal MVN priors, three separate encoding models were estimated using data collected while two subjects listened to spoken stories and had their brain activity measured with fMRI. Brain activity was modeled as a linear combination of the words spoken in the stories using ten temporal delays in order to account for the HRF. **(A)** A non-spherical feature prior and a non-spherical temporal prior were used to create a single non-spherical MVN feature-temporal prior. The covariance of the feature-temporal MVN prior was constructed by computing the Kroenecker product between the semantic feature prior (Figure 2A) and the HRF temporal prior (Figure 3C). The Kroenecker product can be understood as a scaling of the semantic feature prior by each element of the HRF temporal prior and then concatenating the result. We refer to this non-spherical MVN feature-temporal prior as the semantic-temporal prior. **(B)** Comparison of model prediction accuracy under the combined semantic-temporal prior and the HRF temporal prior alone. Model prediction accuracy was assessed by computing the correlation between predicted and observed voxel responses evoked by a novel, held-out story. The encoding model estimated using the semantic-temporal prior consistently yields better prediction accuracy than a model using the HRF temporal prior alone ($p < 10^{-12}$, Wilcoxon signed-rank test). **(C)** Comparison of model prediction accuracy under the combined semantic-temporal prior versus the semantic prior alone. The encoding model estimated using the semantic-temporal prior consistently yields more accurate predictions than the model using the semantic prior alone ($p < 10^{-12}$, Wilcoxon signed-rank test).

5 Combining multiple feature-temporal priors

An important goal of computational neuroscience is to identify what specific information is represented in each region of the brain. In practice, several different hypotheses often exist concerning the specific type of information that is represented in a brain region. These hypotheses are often instantiated as computational models that extract features from stimulus or task conditions. To compare among candidate models, brain activity is modeled as a function of the feature spaces and the variance explained by each feature space is computed.

However, when the models are estimated on their own, the total variance explained by each feature space is inflated by the variance it shares with the other feature spaces. This shared variance component can bias standard model comparison analyses towards models with large total variance whose unique explained variance is nevertheless low. To avoid inflation due to shared variance, one solution is to simultaneously estimate a joint model containing all feature spaces and then conduct further analysis (e.g. variance partitioning; Borcard et al. 1992, Lescroart et al. 2015, de Heer et al. 2017). However, estimating joint models with ordinary regularization techniques (e.g. ridge, LASSO, elastic net) assumes that all feature spaces require the same level of regularization. This is often an incorrect assumption. In practice, the level of regularization for a feature space depends on factors such as the feature space covariance, the number of features, and the fraction of variance explained by that feature space. The choice of regularization level for each feature space is critically important to the prediction accuracy of models that combine multiple feature spaces.

Suppose we have two feature spaces $X_1 \in \mathbb{R}^{n \times p}$ and $X_2 \in \mathbb{R}^{n \times q}$ that are combined into a single encoding model:

$$Y = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon$$

Using ridge regression to estimate such a model is equivalent to choosing the same level of regularization on each feature space. The MVN feature prior imposed by ridge regression on the joint feature weights can be expressed as:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \sim \mathcal{N}_{p+q} \left(0, \begin{bmatrix} \lambda^{-2} I_p & 0 \\ 0 & \lambda^{-2} I_q \end{bmatrix} \right),$$

where the regularization level λ is selected via cross-validation or other methods. It is clear that the prior on each feature space is the same ($\lambda^{-2} I$). However, feature spaces X_1 and X_2 might require different levels of regularization due, for example, to differences in the number of features. Because the globally optimal regularization level (λ) will often be suboptimal for the individual feature spaces, ridge regression can lead to poor prediction performance when used for joint model estimation. This issue also applies to LASSO (Tibshirani, 1996) and elastic-net (Zou and Hastie, 2005) models.

5.1 Banded ridge regression

Instead of using a single spherical MVN prior across all feature spaces, we can impose separate spherical MVN feature priors on the weights of each feature space in the joint model:

$$\beta_1 \sim \mathcal{N}_p(0, \lambda_1^{-2} I_p)$$

$$\beta_2 \sim \mathcal{N}_q(0, \lambda_2^{-2} I_q).$$

The Tikhonov framework allows us estimate the joint model with a separate spherical MVN prior on each feature space,

$$Y = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon$$

$$\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} \sim \mathcal{N}_{pq}\left(0, \Sigma^T \otimes \begin{bmatrix}\lambda_1^{-2} I_p & 0\\ 0 & \lambda_2^{-2} I_q\end{bmatrix}\right),$$

where λ_1 and λ_2 can take different values. For the sake of clarity, assume a spherical MVN temporal prior ($\Sigma^T = I_d$). Estimating this model is equivalent to solving:

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \underset{\beta_1,\beta_2}{\operatorname{argmin}} \left[\|Y - X_1\beta_1 - X_2\beta_2\|_2^2 + \|\lambda_1\beta_1\|_2^2 + \|\lambda_2\beta_2\|_2^2\right]$$

The solution is:

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \left(\begin{bmatrix}X_1^\top X_1 & X_1^\top X_2\\ X_2^\top X_1 & X_2^\top X_2\end{bmatrix} + \underbrace{\begin{bmatrix}\lambda_1^2 I_p & 0\\ 0 & \lambda_2^2 I_q\end{bmatrix}}_{C^\top C}\right)^{-1} \begin{bmatrix}X_1^\top\\ X_2^\top\end{bmatrix} Y$$

Notice that the penalty ($C^\top C$) becomes the ridge penalty when $\lambda_1 = \lambda_2$. However, when $\lambda_1 \neq \lambda_2$ the structure of the penalty becomes "banded" with the first p values along the diagonal equal to λ_1 and the next q values equal to λ_2.

We can also transform the Tikhonov problem into standard form:

$$A = XC^{-1} = \begin{bmatrix}X_1 & X_2\end{bmatrix} \begin{bmatrix}\lambda_1^{-1} I_p & 0\\ 0 & \lambda_2^{-1} I_q\end{bmatrix} = \begin{bmatrix}\frac{X_1}{\lambda_1} & \frac{X_2}{\lambda_2}\end{bmatrix}.$$

This is a very simple expression. It says that scaling the features is equivalent to adjusting the strength of the prior. This is due to the inverse relationship between feature scaling and feature weights. All else being equal, dividing the features by a constant is equivalent to multiplying the weights by that constant. Note that this simple modification can be easily incorporated into other methods (e.g. banded LASSO, banded elastic net, etc).

Finally, the kernelized standard form solution becomes

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}\lambda_1^{-2} X_1^\top\\ \lambda_2^{-2} X_2^\top\end{bmatrix} \left(\sum_{i\in\{1,2\}} \lambda_i^{-2} X_i X_i^\top + I\right)^{-1} Y.$$

5.2 Evaluating banded ridge regression

To evaluate the use of banded ridge regression versus ridge regression, we estimated two separate encoding models using data from a vision experiment (Huth et al., 2012). Data were collected while three subjects watched several hours of natural movies and had their brain activity measured with fMRI. We constructed a single joint encoding model that combined two previously published feature spaces. The first feature space X_1 captured low-level visual properties from the stimulus (Nishimoto et al., 2011). The second feature space X_2 captured high-level visual properties consisting of object and action categories (Huth et al., 2012). We

modeled brain responses as a linear combination of these low- and high-level visual feature spaces:

$$Y = \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \epsilon$$

We estimated the joint encoding model using banded ridge regression and ridge regression separately. The only difference between these models is the MVN feature prior used: banded ridge regression uses a non-spherical MVN prior,

$$\Sigma_T^X = \begin{bmatrix} \lambda_1^{-2} I_p & 0 \\ 0 & \lambda_2^{-2} I_q \end{bmatrix},$$

whereas ridge regression uses a spherical MVN prior ($\Sigma_R^X = \lambda^{-2} I_{p+q}$).

Low-level motion energy features were extracted from the natural movies using a three-dimensional Gabor pyramid (Adelson and Bergen, 1985, Watson and Ahumada, 1985, Nishimoto et al., 2011). This yielded a total of 6,555 features which differed in orientation, spatial and temporal frequency, location, size, and direction of motion. The high-level object and action category features were tagged by hand from each one second segment of the movies and labeled using WordNet synsets (Miller, 1995, Huth et al., 2012). The hyponyms for each synset were inferred from the WordNet graph and also included. This process yielded a total of 1,705 object and action category features. An FIR model was then constructed by including 10 TR temporal delays for each feature in order to account for the hemodynamic response function. The resulting model consisted of 8,260 stimulus features times 10 delays (82,600 total features) and 3,600 time points. For simplicity, we used a spherical MVN temporal prior. The MVN feature prior hyperparameters for both banded ridge (λ_1 and λ_2) and ridge (λ) models were selected per voxel via 5-fold cross-validation. The performance of each model was assessed by computing the correlation between model predictions and actual responses using a held-out dataset not used for model estimation.

Banded ridge regression provided far better joint model prediction performance than standard ridge regression (Figure 5; $p < 10^{-12}$, Wilcoxon signed-rank test). Of the approximately 230,000 voxels, about 40,000 were significantly predicted ($n = 270$, $q(FDR) < 0.05$) with banded ridge (mean prediction accuracy r=0.06 $\pm$ 0.0003 s.e.m.). In contrast, approximately 20,000 voxels were significantly predicted ($n = 270$, $q(FDR) < 0.05$) with ridge regression (mean prediction accuracy r=0.03 $\pm$ 0.0002 s.e.m.). We used the weights estimated with banded ridge regression to compute the prediction accuracy of each feature space on its own. This gave us a separate prediction accuracy value per voxel for each the motion energy features and for the object category features. These prediction accuracy values are plotted on the cortical surface for one subject in Figure 5D. There is a strong separation in prediction performance between early visual cortex being best predicted by motion energy features, and higher visual cortex better predicted by object category features (Figure 5E).

A benefit of banded ridge regression is that it helps reduce spurious correlations between feature spaces. When estimating the prediction accuracy of a model fit on its own, the estimate may be inflated due to correlations with other feature spaces. To illustrate, note that when the object category model is estimated by itself using ridge regression, responses of many voxels in early visual cortex are predicted accurately (Figure 6A, green; mean

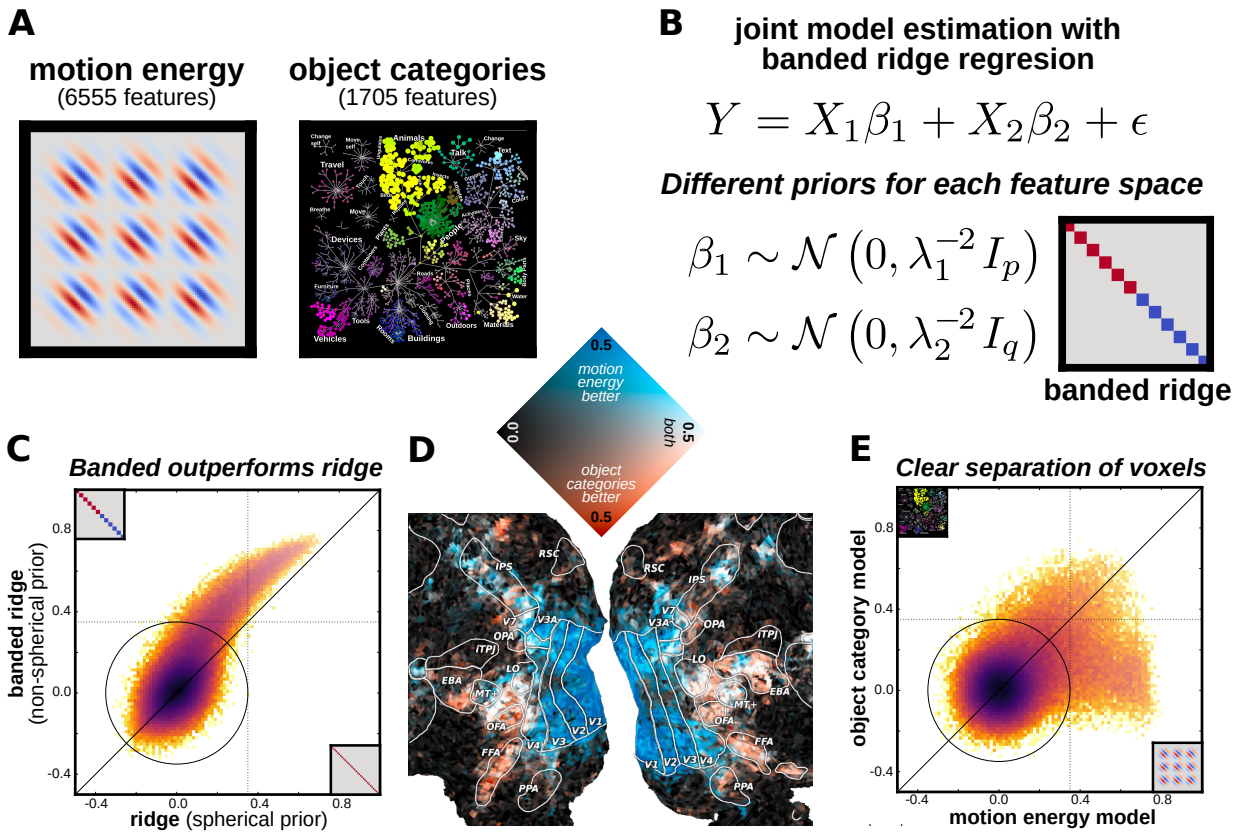


Figure 5. Comparison of banded ridge regression and ridge regression for joint model estimation in a vision experiment. **(A)** To evaluate the use of non-spherical (i.e. banded ridge) and spherical (i.e. ridge) MVN priors in encoding models that incorporate multiple feature spaces, a single joint model was constructed by concatenating motion energy and object category features. The joint model was estimated using data collected while three subjects watched several hours of natural movies. Brain activity was measured using fMRI. Brain activity was modeled as a linear combination of motion energy and object category features. An encoding model was estimated using a spherical MVN prior (i.e. ridge regression). **(B)** An additional encoding model was estimated using a non-spherical MVN feature prior constructed by combining two spherical MVN feature priors with different variance, one for each feature space. Because the covariance of the non-spherical MVN feature prior has a banded appearance, we refer to this approach as banded ridge regression. **(C)** Comparison of model prediction accuracy under banded ridge regression and ridge regression. Model prediction accuracy was assessed by computing the correlation between predicted and observed voxel responses evoked by a novel, held-out stimulus for each model separately. The joint encoding model estimated with banded ridge regression (i.e. a non-spherical MVN prior) yields better prediction accuracy than the joint model estimated with ridge regression (i.e. a spherical MVN prior; $p < 10^{-12}$, Wilcoxon signed-rank test). **(D)** Prediction accuracy of motion energy and object category features plotted on the cortical surface of one subject. The voxelwise predictions for each individual feature set (e.g. motion energy) were computed using the weights for those features obtained from the joint model estimated with banded ridge regression. Prediction accuracy for each feature set was then assessed by computing the correlation between predicted and observed voxel responses. Voxels located in early visual cortex are accurately predicted by motion energy features (blue) and voxels located in higher visual cortex are accurately predicted by object category features (red). A subset of voxels are accurately and similarly predicted by both feature spaces (white). **(E)** Comparison of prediction accuracy obtained from motion energy and object category features using the banded ridge regression estimate of the joint encoding model (as in D; all subjects).

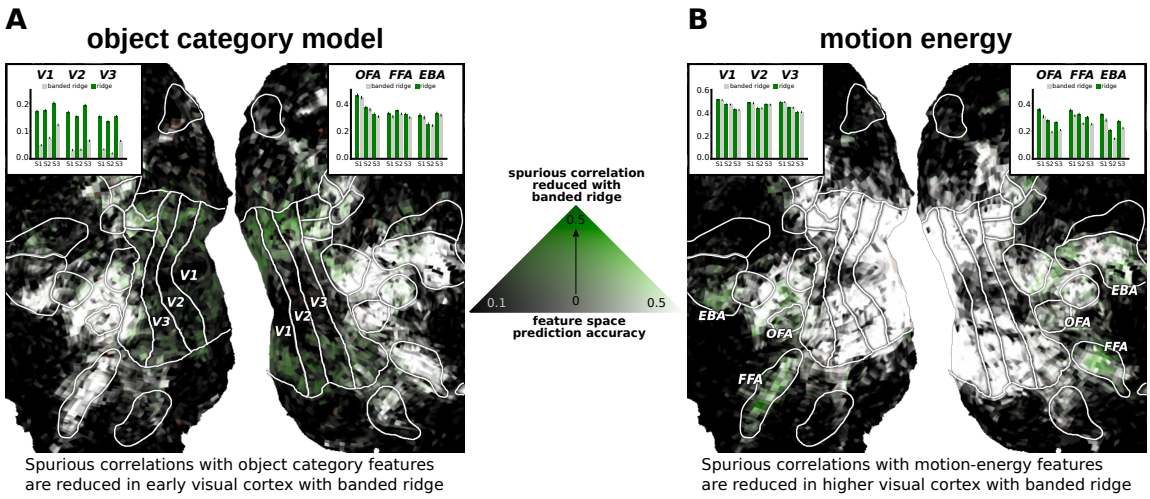


Figure 6. Effect of banded ridge regression on the prediction accuracy of object category and motion energy features. Three separate encoding models were estimated using data collected while three subjects watched several hours of natural movies. Brain activity was measured using fMRI. Ridge regression was used to model brain activity as a linear combination of object category features. A separate ridge regression model was estimated using motion energy features. In addition, a joint model was constructed by concatenating motion energy and object category features. This joint model was estimated using banded ridge regression (i.e. a non-spherical MVN feature prior; see Figure 5). **(A)** The prediction accuracy of object category features was computed for both banded ridge and ridge regression models separately. The prediction accuracy of each model is plotted simultaneously on the cortical surface of one subject using a 2D colormap. Voxels located in anterior visual cortex are accurately predicted by object category features under both banded ridge regression and standard ridge regression (voxels colored in white). However, voxels located in early visual cortex (EVC) appear to be more accurately predicted by object category features under ridge regression than under banded ridge regression. This difference occurs because ridge regression attributes all explainable variance to object category features, while under banded ridge regression the variance is split correctly between object category and motion energy features. Because banded ridge regression allows motion energy features to explain some of the variance away from object category features, it reduces the spurious prediction accuracy of object category features observed under ridge regression in EVC voxels (colored in green). **(B)** The prediction accuracy of motion energy features was computed for both the banded ridge and the ridge regression models separately (as in A). Voxels located in EVC are accurately predicted by motion energy features under both banded ridge regression and standard ridge regression (voxels colored in white). Banded ridge regression allows object category features to explain some of the variance away from the motion energy features in anterior visual cortex voxels (colored in green). Inset bar plots show the mean prediction accuracy under banded ridge regression (light grey) and ridge regression (green) in regions of interest for each of the three individual subjects.

prediction accuracy r=0.167, 5,250 voxels located in V1-V3 across three subjects). However, early visual cortex does not explicitly represent visual object categories. It is therefore likely that this is an artifact of stimulus correlations. For example, while the presence of vehicles is explicitly encoded as an object category feature, the motion energy model may implicitly encode the presence of vehicles using direction selective filters located in the lower visual field. Because of these correlations, the object category model can accurately predict responses in early visual cortex when estimated on its own (and the motion energy model can make accurate predictions in higher visual cortex). Joint model estimation with banded ridge regression ameliorates this problem by effectively decorrelating the model features to an amount determined by their covariance and the regularization parameters. When evaluating

the prediction accuracy of each feature space using the weights estimated with banded ridge regression, the amount of variance explained by stimulus correlations can decrease relative to ridge regression. Indeed, we find that object category model prediction accuracy decreased by 67% for voxels located in V1-V3 when using the weights from banded ridge regression relative to ridge regression (Figure 6A; mean prediction accuracy for banded ridge r=0.055 and ridge r=0.167). Similarly, the prediction accuracy of motion energy features decreased by 20% for voxels located in OFA, FFA, and EBA (Figure 6B; mean prediction accuracy for banded ridge r=0.229 and ridge r=0.286). Thus, banded ridge regression yields estimates of the variance that can be explained by any one feature space that are less affected by stimulus correlations.

6 Discussion

This paper introduces a powerful new Tikhonov framework that can improve estimation and interpretation of models fit to fMRI data. The Tikhonov framework can be used to incorporate prior information about how features in the model covary, how the measured signals vary in time, and how multiple feature spaces can be used to build a single predictive model. This framework may be particularly valuable for modeling data acquired in rich, naturalistic experiments where multiple feature spaces are probed simultaneously.

The Tikhonov framework is merely another method available to researchers and is by no means optimal in every experimental setting. Indeed, simpler methods (e.g. stimulus-triggered average) can produce estimates that are at least as good as our approach when the assumptions made by those methods are met by the data. Similarly, when a lot of data are available, more complex methods such as artificial neural networks (ANNs) can produce estimates that are at least as good as our approach. In general, the researcher should treat the choice of fitting procedure (e.g. FIR, grouped L1, OLS, ANN, etc) as a hyperparameter on its own right and use statistical learning theory to make a decision. The Tikhonov framework is presented as another method in the toolkit available to researchers. The banded ridge model proposed is of particular utility when estimating joint models that combine multiple feature spaces to predict brain activity. We have shown a computationally efficient framework for incorporating non-spherical multivariate normal priors into voxelwise encoding models. And that this framework is flexible enough to work well in a variety of cases. The software used to estimate all the models presented in this paper is publicly available (`http://github.com/gallantlab/tikreg`). We hope this facilitates the adoption of this framework.

Acknowledgments

This work was supported by grants from the National Science Foundation (NSF; IIS1208203), the National Eye Institute (EY019684 and EY022454) and the Office of Naval Research (N00014-15-1-2861). A.G.H. was also supported by a Burroughs-Wellcome fellowship. We

thank Leila Wehbe for useful technical discussions, and Brittany Griffin for segmenting and flattening cortical surfaces. The authors declare no conflict of interest.

References

E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284, 1985.

G. K. Aguirre, E. Zarahn, and M. D'esposito. The variability of human, BOLD hemodynamic responses. *NeuroImage*, 8(4):360–9, Nov. 1998.

N. Bazargani and A. Nosratinia. Joint maximum likelihood estimation of activation and Hemodynamic Response Function for fMRI. *Medical image analysis*, 18(5):711–24, jul 2014.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

D. Borcard, P. Legendre, and P. Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73(3):1045–1055, 1992.

G. M. Boynton, S. A. Engel, G. H. Glover, and D. J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, 16(13): 4207–4221, 1996.

C. Büchel, A. Holmes, G. Rees, and K. Friston. Characterizing stimulus–response functions using nonlinear regressors in parametric fmri experiments. *NeuroImage*, 8(2):140–148, 1998.

R. Casanova, S. Ryali, J. Serences, L. Yang, R. Kraft, P. J. Laurienti, and J. A. Maldjian. The impact of temporal regularization on estimates of the bold hemodynamic response function: a comparative analysis. *NeuroImage*, 40(4):1606–1618, 2008.

C. Chu, Y. Ni, G. Tan, C. J. Saunders, and J. Ashburner. Kernel regression for fmri pattern prediction. *NeuroImage*, 56(2):662–673, 2011.

V. De Angelis, F. De Martino, M. Moerel, R. Santoro, L. Hausfeld, and E. Formisano. Cortical processing of pitch: Model-based encoding and decoding of auditory fmri responses to real-life sounds. *NeuroImage*, 2017.

W. A. de Heer, A. G. Huth, T. L. Griffiths, J. L. Gallant, and F. E. Theunissen. The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, 37(27):6539–6557, 2017.

A. Doicu, T. Trautmann, and F. Schreier. *Numerical regularization for atmospheric inverse problems*. Springer Science & Business Media, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Springer, New York, NY, 2001.

K. J. Friston, O. Josephs, G. Rees, and R. Turner. Nonlinear event-related responses in fmri. *Magnetic resonance in medicine*, 39(1):41–52, 1998.

G. H. Glover. Deconvolution of impulse response in event-related bold fmri. *NeuroImage*, 9 (4):416–429, 1999.

C. Goutte, F. A. Nielsen, and K. Hansen. Modeling the hemodynamic response in fmri using smooth fir filters. *IEEE transactions on medical imaging*, 19(12):1188–1201, 2000.

L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor. Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, 72:304–321, 2013.

D. A. Handwerker, J. M. Ollinger, and M. D'Esposito. Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21(4):1639–1651, 2004.

P. C. Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600): 453–458, 2016.

K. N. Kay, S. V. David, R. J. Prenger, K. A. Hansen, and J. L. Gallant. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fmri. *Human Brain Mapping*, 29(2):142–156, 2008a.

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008b.

M. D. Lescroart, D. E. Stansbury, and J. L. Gallant. Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Frontiers in Computational Neuroscience*, 9, 2015.

G. Marrelec, H. Benali, P. Ciuciu, M. Pelegrini-Issac, and J.-B. Poline. Robust bayesian estimation of the hemodynamic response function in event-related bold fmri using basic physiological information. *Human Brain Mapping*, 19(1):1–17, 2003.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19): 1641–1646, 2011.

A. V. Oppenheim, A. Willsky, and I. Young. *Signals and systems.* Englewood Cliffs, NJ: Prentice-Hall, 1983.

F. Pedregosa, M. Eickenberg, P. Ciuciu, B. Thirion, and A. Gramfort. Data-driven hrf estimation for encoding and decoding models. *NeuroImage*, 104:209–220, 2015.

W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

O. Schoppe, N. S. Harper, B. D. Willmore, A. J. King, and J. W. Schnupp. Measuring the performance of neural models. *Frontiers in Computational Neuroscience*, 10, 2016.

B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J.-B. Poline, D. Lebihan, and S. Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4):1104–16, Dec. 2006.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. N. Tikhonov, V. I. Arsenin, and F. John. *Solutions of ill-posed problems*, volume 14. Winston Washington, DC, 1977.

P. D. Turney and P. Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of the Optical Society of America*, 2(2):322–342, 1985.

L. Wehbe, B. Murphy, P. Talukdar, A. Fyshe, A. Ramdas, and T. Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS ONE*, 9(11):e112575, 2014.

M. W. Woolrich, T. E. Behrens, and S. M. Smith. Constrained linear basis sets for hrf modelling using variational bayes. *NeuroImage*, 21(4):1748–1761, 2004.

M. C.-K. Wu, S. V. David, and J. L. Gallant. Complete Functional Characterization of Sensory Neurons By System Identification. *Annual Review of Neuroscience*, 29(1):477–505, Jan. 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix A Standard form derivation

$$\begin{aligned}
\hat{\beta}_T &= \left(X^\top X + \lambda^2 C^\top C\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= C\left(X^\top X + \lambda^2 C^\top C\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= \left(X^\top X C^{-1} + \lambda^2 C^\top C C^{-1}\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= \left(X^\top X C^{-1} + \lambda^2 C^\top\right)^{-1} X^\top Y \\
C\hat{\beta}_T &= \left(X^\top X C^{-1} + \lambda^2 C^\top\right)^{-1} C^\top C^{-1\top} X^\top Y \\
C\hat{\beta}_T &= \left(C^{-1\top} X^\top X C^{-1} + \lambda^2 I_p\right)^{-1} C^{-1\top} X^\top Y
\end{aligned}$$

Define $A = XC^{-1}$ and $\hat{\beta}_A = C\hat{\beta}_T$. The solution becomes

$$\hat{\beta}_A = \left(A^\top A + \lambda^2 I_p\right)^{-1} A^\top Y,$$

and one can recover the original weights with

$$\hat{\beta}_T = C^{-1}\hat{\beta}_A.$$

There exists an interesting relationship between the prior covariance matrix, Σ, and the Tikhonov penalty matrix, C. When the penalty Gram matrix, $C^\top C$, is full-rank, it is invertible and there exists a corresponding prior,

$$\Sigma = \left(C^\top C\right)^{-1}.$$

However, the standard form decouples the two concepts. There exist well-defined Tikhonov penalties for which a prior cannot be expressed. In particular, if the penalty Gram matrix is not positive semi-definite, no inverse exists and therefore the prior cannot be formally expressed. The converse is also true. A rank-deficient prior can be used if the problem is in standard form, yet there is no corresponding penalty matrix. See Doicu et al. (2010) for a full treatment.

Appendix B Equivalence of FIR models with temporal priors and convolution followed by ridge

Estimating an FIR model with a temporal prior $\Sigma^T = \mathbb{B}\mathbb{B}^\top$

$$Y = X\beta + \epsilon$$
$$\beta_i \sim \mathcal{N}_d\left(0, \lambda^{-2}\mathbb{B}\mathbb{B}^T\right)$$

is equivalent to convolving the features x_i with the columns of $\mathbb{B}$ and estimating the model using ridge regression:

$$Y = \begin{bmatrix} | & | & | & | & | & | & | \\ (x_1 * b_1) & \dots & (x_1 * b_k) & \dots & (x_p * b_1) & \dots & (x_p * b_k) \\ | & | & | & | & | & | & | \end{bmatrix} \beta + \epsilon \tag{3}$$
$$\beta \sim \mathcal{N}_{pk}\left(0, \lambda^{-2} I_{pk}\right)$$

Recall the definition of the standard form transform:

$$\begin{aligned} \Sigma^T &= \left(C^\top C\right)^{-1} \\ A &= XC^{-1}, \end{aligned}$$

where $C^{-1} = \mathbb{B}$ for a temporal prior $\Sigma^T = \mathbb{B}\mathbb{B}^\top$. The standard transform of the FIR model can be written as

$$A = \overbrace{\begin{bmatrix} | & | & | & | & | & | \\ X^{(1)} & X^{(2)} & \dots & X^{(i)} & \dots & X^{(p)} \\ | & | & | & | & | & | \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} b_1(0) & b_2(0) & \dots & b_k(0) \\ b_1(1) & b_2(1) & \dots & b_k(1) \\ \vdots & \vdots & \dots & \vdots \\ b_1(d) & b_2(d) & \dots & b_k(d) \end{bmatrix}}^{\mathbb{B}}$$

where each $X^{(i)} \in \mathbb{R}^{n \times d}$ is a matrix that contains every feature $x_i \in \mathbb{R}^{n \times 1}$ at delays 0 through d, and every row of $X^{(i)}$ corresponds to a particular time point t:

$$X^{(i)}(t) = \begin{bmatrix} x_i(t) & x_i(t-1) & \dots & x_i(t-d) \end{bmatrix}$$

We can express every entry of the matrix A as the dot product between $X^{(i)}(t)$ and each column of the temporal basis set, b_j :

$$\begin{aligned} a_i^{b_j}(t) &= \begin{bmatrix} x_i(t) & x_i(t-1) & \dots & x_i(t-d) \end{bmatrix} \begin{bmatrix} b_j(0) \\ b_j(1) \\ \vdots \\ b_j(d) \end{bmatrix} \\ a_i^{b_j}(t) &= \left\langle \begin{matrix} [x_i(t), & x_i(t-1), & x_i(t-2), & \dots, & x_i(t-d)], \\ [b_j(0), & b_j(1), & b_j(2), & \dots, & b_j(d)] \end{matrix} \right\rangle \\ a_i^{b_j}(t) &= \sum_{\delta=0}^{d} x_i(t-\delta)\, b_j(\delta) \end{aligned}$$

which is the definition of discrete convolution

$$\begin{aligned}(x_i * b_j)[t] &\equiv \sum_{\delta=0}^{d} x_i(t-\delta) b_j(\delta) \\ a_i^{b_j}(t) &= (x_i * b_j)[t]\end{aligned}$$

Finally, we rewrite $A = X\mathbb{B}$ as the convolution of each feature i with each temporal basis j

$$\begin{aligned}a_i &= \begin{bmatrix} | & | & | & | \\ (x_i * b_1) & (x_i * b_2) & \dots & (x_i * b_k) \\ | & | & | & | \end{bmatrix} \\ A &= \begin{bmatrix} | & | & | & | & | & | \\ a_1 & a_2 & \dots & a_i & \dots & a_p \\ | & | & | & | & | & | \end{bmatrix} \\ A &= \begin{bmatrix} | & | & | & | & | & | \\ (x_1 * b_1) & \dots & (x_1 * b_k) & \dots & (x_p * b_1) & \dots & (x_p * b_k) \\ | & | & | & | & | & | \end{bmatrix}\end{aligned}$$

This is exactly Equation 3.

Appendix C Kernel solution to encoding models with feature-temporal MVN priors

The standard form solution is

$$\hat{\beta}_A = \left(A^\top A + \lambda^2 I\right)^{-1} A^\top Y$$

The kernel solution to the standard form problem becomes

$$\hat{\beta}_A = A^\top \left(AA^\top + \lambda^2 I\right)^{-1} Y$$

Expanding this out using the fact that $A = XC^{-1}$

$$\hat{\beta}_A = C^{-1\top} X^\top \left(XC^{-1}C^{-1\top}X^\top + \lambda^2 I_p\right)^{-1} Y$$

We know $\Sigma = C^{-1}C^{-1\top} = (CC^\top)^{-1}$. Replacing this in

$$\hat{\beta}_A = C^{-1\top} X^\top \left(X\left(C^{-1}C^{-1\top}\right)X^\top + \lambda^2 I_p\right)^{-1} Y$$

To recover the Tikhonov solution, recall that $\hat{\beta}_T = C^{-1}\hat{\beta}_A$. Substituting this in

$$\hat{\beta}_T = C^{-1}C^{-1\top} X^\top \left(X\left(C^{-1}C^{-1\top}\right)X^\top + \lambda^2 I_p\right)^{-1} Y$$

We know $\Sigma = C^{-1}C^{-1\top}$. Replacing this in

$$\hat{\beta}_T = \Sigma X^\top \left(X\Sigma X^\top + \lambda^2 I_p\right)^{-1} Y$$

In the case of feature-temporal kernels $\Sigma = \Sigma^T \otimes \Sigma^X$. The full solution becomes

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \left(X\left(\Sigma^T \otimes \Sigma^X\right)X^\top + \lambda^2 I_p\right)^{-1} Y$$

Appendix D Efficient kernel solution for models with feature-temporal MVN priors

We now derive a computationally efficient solution for the kernel solution for an encoding model with non-spherical feature-temporal multivariate normal priors. This formulation makes the estimation of these models computationally tractable.

The feature-temporal prior is constructed by computing the Kronecker product ($\otimes$) between the temporal prior $\Sigma^T \in \mathbb{R}^{d \times d}$ and the feature prior $\Sigma^X \in \mathbb{R}^{p \times p}$,

$$\Sigma = \Sigma^T \otimes \Sigma^X = \begin{bmatrix} \Sigma^T_{1,1}\Sigma^X & \cdots & \Sigma^T_{1,d}\Sigma^X \\ \vdots & \ddots & \vdots \\ \Sigma^T_{d,1}\Sigma^X & \cdots & \Sigma^T_{d,d}\Sigma^X \end{bmatrix}.$$

The resulting feature-temporal prior is $\Sigma \in \mathbb{R}^{pd \times pd}$. Notice that when both the feature and the temporal priors are spherical, the feature-temporal prior is also spherical.

The Tikhonov solution to an encoding model with a feature-temporal multivariate normal prior $\Sigma^T \otimes \Sigma^X$ can be expressed as (see Appendix C):

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \underbrace{\overbrace{\left(X\left(\Sigma^T \otimes \Sigma^X\right)X^\top\right.}^{K} \left. + \lambda^2 I\right)^{-1} Y}_{\hat{\alpha}}.$$

A computationally efficient solution can be derived by re-arranging terms. First, notice that the kernel regression solution to the standard form problem is embedded within the Tikhonov solution above:

$$\hat{\alpha} = \left(X\left(\Sigma^T \otimes \Sigma^X\right)X^\top + \lambda^2 I\right)^{-1} Y.$$

The term inside the parenthesis is the regularized $n \times n$ kernel matrix K of the standard form transformation:

$$\left(K + \lambda^2 I\right) = \left(X\left(\Sigma^T \otimes \Sigma^X\right)X^\top + \lambda^2 I\right).$$

Recall that X is an $n \times pd$ FIR matrix which includes delayed copies of the linearized stimulus feature matrix. Computing the kernel matrix thus requires the following matrix multiplication

$$K = \begin{bmatrix} X_{\delta(1)} \cdots X_{\delta(d)} \end{bmatrix} \begin{bmatrix} \Sigma^T_{1,1}\Sigma^X & \cdots & \Sigma^T_{1,d}\Sigma^X \\ \vdots & \ddots & \vdots \\ \Sigma^T_{d,1}\Sigma^X & \cdots & \Sigma^T_{d,d}\Sigma^X \end{bmatrix} \begin{bmatrix} X_{\delta(1)} \\ \vdots \\ X_{\delta(d)} \end{bmatrix}$$

Finally, this matrix multiplication can be expressed as a sum of matrix products,

$$K = \sum_j^d \sum_i^d \Sigma^T_{(i,j)} \left(X_{\delta(i)} \Sigma^X X^\top_{\delta(j)}\right).$$

This formulation makes the problem of estimating encoding models with feature-temporal multivariate normal priors tractable in contexts where $n < p$.

$$\hat{\beta}_T = \left(\Sigma^T \otimes \Sigma^X\right) X^\top \left(\sum_j^d \sum_i^d \Sigma^T_{(i,j)} \left(X_{\delta(i)} \Sigma^X X^\top_{\delta(j)} \right) + \lambda^2 I \right)^{-1} Y$$

Appendix E Extension to priors on priors: hyper-priors

We have shown the usefulness of imposing various temporal and feature priors on feature weights to improve predictive models. There exist situations, however, when the expert prior itself needs to be regularized. This can be the case when the expert prior is derived empirically and is noisy, or when the prior can be modified to match the data better. In such cases, we can apply the same principle and impose a prior on the prior—a hyper-prior. We next show an example on how to incorporate hyper-priors to our framework.

In Section 3.2, we described the smoothness prior. Our results show that imposing a smoothness prior on the temporal delays does not improve the prediction performance of the motion energy model. This is surprising. We expect the hemodynamic response function to be temporally smooth, and so imposing a smoothness prior should improve prediction performance. This intuition, however, ignores the structure of the smoothness prior.

The smoothness prior imposes a strong covariance to delays in the middle of the temporal filter (see Figure 3). This is problematic because the goodness of the prior will depend on the number of delays. This is a bad assumption in many cases. In order to avoid this issue, we can impose a spherical prior on the smoothness prior. This can be thought of as trading off between a spherical prior and the smoothness prior, where the trade-off is controlled by the hyper-prior hyper-parameter.

In general, hyper-priors can be expressed as

$$\beta \sim N_p\left(0, \lambda^{-2}\Sigma\right)$$

$$\Sigma \sim W_p\left(\gamma^{-2}\Lambda_p\right),$$

where W is a Wishart distribution. In the case of the smoothness prior, this results in

$$\Sigma^* = \lambda^{-2}\left(\mathbb{D}^2 + \gamma^2 I_p\right)^{-1},$$

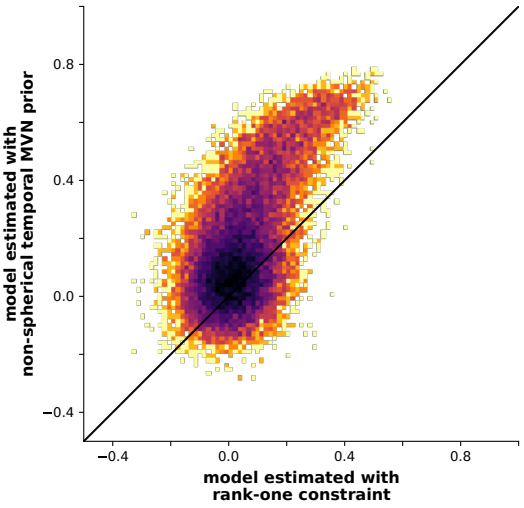
where λ and γ are hyper-parameters.

Estimating models that include both a prior and a hyper-prior is feasible under our framework (and implemented in the accompanying software). However, this flexibility comes at the cost of computational resources because the hyper-prior hyper-parameter (γ) needs to be estimated via cross-validation.

Appendix F Comparison to time-feature separable rank-one models

Another approach to modeling the hemodynamic response function (HRF) is to assume a single HRF shape for each voxel. This approach in effect separates the temporal weights from the feature weights when solving the regression problem. A powerful method for estimating time-feature separable encoding models is rank-one regression (Pedregosa et al., 2015). We proceeded to compare the rank-one model approach against the approach presented in Section 3.2.2.

Functional MRI data were collected while one subject watched several hours of natural movies (Huth et al., 2012). BOLD responses were modeled as a linear combination of motion energy features using rank-one regression with an HRF-basis set (Pedregosa et al., 2015). In order to reduce computational requirements, we only estimated rank-one regression models for the best 20,000 voxels in terms of functional SNR. Prediction accuracy was estimated as the correlation coefficient between predicted and actual responses to a held-out stimulus (n=270). The prediction accuracy of the rank-one method was compared against the prediction accuracy of an FIR model estimated with a non-spherical temporal prior constructed from an HRF-basis set (see Section 3.2.2 for more details). The prediction accuracy of the rank-one method is consistently lower than our approach (Supplemental Figure 1). The rank-one method performs poorly because it does not include additional regularization on the feature weights. We suspect that adding a regularization term on the feature weights in addition to the rank-one constraint would improve the prediction accuracy of rank-one regression. However, rank-one regression is more computationally expensive than our Tikhonov approach.



Supplemental Figure 1. Comparison between our approach and a rank-one regularized model. One subject watched hours of natural movies while their brain activity was measured with fMRI. The top 20,000 voxels with the highest explainable variance were modeled as a linear combination of motion energy features. Two models were estimated. One model was estimated with a non-spherical temporal MVN prior built from an HRF basis set (our approach; Section 3.2.2). A second model was estimated using a rank-one constraint which enforces separate HRF and feature weight estimates (Pedregosa et al., 2015). The prediction accuracy of each model was assessed by computing the correlation coefficient between predicted and actual responses in a held-out dataset. The prediction accuracy of the model estimated with rank-one regression is lower than that of the model estimated using our approach. This result is likely because the rank-one model does not regularize the feature weights during estimation.