Exploring an EM-algorithm for banded regression in computational neuroscience

Søren A. Fuglsang[1,2], Kristoffer H. Madsen[1,3], Oula Puonti[1,4], Hartwig R. Siebner[1,5,6], and Jens Hjortkjær[1,2]

[1] *Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital - Amager and Hvidovre, Copenhagen, Denmark.*
[2] *Center for Auditory Neuroscience, Department of Health Technology, Technical University of Denmark, Kgs. Lyngby, Denmark.*
[3] *Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark, Kgs. Lyngby, Denmark.*
[4] *Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, USA.*
[5] *Department of Neurology, Copenhagen University Hospital Bispebjerg, Copenhagen, Denmark.*
[6] *Institute for Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.*

September 22, 2023

Abstract

Regression is a principal tool for relating brain responses to stimuli or tasks in computational neuroscience. This often involves fitting linear models with predictors that can be divided into groups, such as distinct stimulus feature subsets in encoding models or features of different neural response channels in decoding models. When fitting such models, it can be relevant to impose differential shrinkage of the different groups of regression weights. Here, we explore a framework that allow for straightforward definition and estimation of such models. We present an expectation-maximization algorithm for tuning hyperparameters that control shrinkage of groups of weights. We highlight properties, limitations, and potential use-cases of the model using simulated data. Next, we explore the model in the context of a BOLD fMRI encoding analysis and an EEG decoding analysis. Finally, we discuss cases where the model can be useful and scenarios where regularization procedures complicate model interpretation.

1 Introduction

Understanding how the brain responds to different stimuli or tasks is a key objective in systems neuroscience. In recent years, significant progress has been made in developing models to analyze neural responses to complex, naturalistic stimuli [1]. One potent approach is encoding models that combine stimulus features to predict channel-wise neural responses from fMRI or electrophysiology through linear regression [2, 3, 4, 5, 6, 7]. Regularization is often needed in such models to minimize overfitting and improve model prediction on held-out data. However, multiple - potentially co-linear - stimulus feature sets can complicate the assessment of the relative importance of the individual features. In this case, it can be relevant to allow for differential shrinkage of different groups of regression weights. For instance, neural responses to natural speech depend on both low-level acoustic features and higher-level phonetic or semantic features [8, 5, 6, 9]. Such feature groups vary on different time scales, have different dimensionality, and some features may not even be reflected in the neural response.

Several regularization strategies could have relevance for this situation [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. Recently, "banded" regularization [23, 24] proposed to assign different levels of regularization to different feature bands (groups of predictors) in encoding analyses. Such banded regression models can yield higher out-of-sample predictive power compared to Ridge regression models in fMRI encoding analyses [23, 24]. However, estimating banded regression models can be practically challenging when several feature spaces are considered.

Similar regression problems arise in decoding models, i.e., models that combine neural channels (voxels, electrodes) to predict a continuous stimulus- or task feature [25, 26]. One pertinent example is decoding models for EEG or MEG responses to continuous stimuli, like speech or music [27, 28, 29]. Here, regularized regression can be used to identify multidimensional finite impulse response (FIR) filters that predict stimulus features by combining different electrode signals at different time lags and/or different frequency bands [30, 31, 32]. As with encoding models, it can be useful to individually shrink groups of weights corresponding to groups of predictors.

Here, we explore a general framework for estimating "banded"-type [23, 24] regression models. The framework is closely related to automatic relevance determination [33, 21, 22, 34, 35, 36] and is an extension of work in [35]. The framework can be used to formulate regularized regression estimators with different regularization properties. We use an expectation-maximization (EM) algorithm to tune hyperparameters that control variances of prior distributions. The model can be used to formulate priors that allow for differential shrinkage of sets of regression weights. The model can additionally be used to incorporate smoothness constraints on sets of regression weights by enforcing correlation between related regression weights.

In this paper, we describe the modeling framework and provide simulations to illustrate the properties, limitations and potential use-cases of the described algorithm. Additionally, we illustrate properties of the model by using it to fit stimulus-response models on two publicly available datasets [5, 37] and compare these model fits with the widely used Ridge estimator [11]. In the first dataset, we consider encoding models that use a set of audio features to predict BOLD fMRI responses to music [37]. In the second, we fit decoding models that use features of multi-channel time-lagged EEG data to predict an audio stimulus feature during a speech listening task [5]. Software implementations of the model in both MATLAB and Python are available on at https://github.com/safugl/embanded.

2 Regression model

Encoding and decoding analyses often involve linear regression models in the following form:

$$y = Xw + \epsilon, \tag{1}$$

where ϵ is a zero-mean Gaussian variable.

Here, encoding models would treat stimulus- or task features as predictors, X, and a (preprocessed) neural response as target variable, y. Decoding models treat features extracted from neural activity measurements as predictors, X, and a stimulus- or task feature as target variable, y. The differences in direction of inference have large implications for model interpretation and model properties [38, 25] as well as how potential stimulus- or task confounders can be dealt with.

Despite their differences, such encoding and decoding models lead to regression problems that can be encompassed in a general formulation. Let $y \in \mathrm{R}^{M\times 1}$ denote a target variable where M indicates the number of observations. Moreover, let X denote a matrix with D predictors such that $X \in \mathrm{R}^{M\times D}$. Suppose that X can be divided into a set of meaningful groups such that $X = [F_1, F_2, ..., F_J]$, $j = 1, 2, ..., j, ...J$. Here, F_j denotes a group of D_j predictors, i.e., $F_j \in \mathrm{R}^{M\times D_j}$. This group of predictors could, for example, be a stimulus feature subset in an encoding model, or response features of different voxels or electrodes in a decoding model. We thus consider a regression model in the following form [23, 24]:

$$y = F_1\beta_1 + F_2\beta_2 + ... + F_J\beta_J + \epsilon, \tag{2}$$

where β_j denotes regression weights for a given group of predictors, F_j. Further, let w denote a column-vector collecting all the corresponding weights $w = \left[\beta_1^\top, \beta_2^\top, ..., \beta_J^\top\right]^\top$, with $w \in \mathrm{R}^{D\times 1}$. We will assume that input variables and target variables have been centered and that no intercept term is needed in the regression model (since the intercept typically carries no meaningful information and hence shrinkage is typically not desired). The conditional distribution of the target variable y given a set of observed predictors X will be assumed to follow a Gaussian distribution in the following form:

$$p(y|X, w, \nu) = \mathcal{N}(y|Xw, \nu I_M), \tag{3}$$

where $\mathcal{N}(y|Xw, \nu I_M)$ is used to indicate a multivariate Gaussian with mean Xw and a diagonal covariance matrix νI_M of size $M \times M$. A zero-mean Gaussian prior distribution will be placed over the weights:

$$p(w|\Lambda, \eta, \tau) = \mathcal{N}(w|0, \Lambda), \tag{4}$$

where Λ is a $D \times D$ block-diagonal covariance matrix. Such a prior distribution over the weights can be used to "shrink" the estimated regression weights. Here, we specifically define that Λ has a block structure defined as follows

$$\Lambda \equiv \begin{bmatrix} \lambda_1 I_{D_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_J I_{D_j} \end{bmatrix} \begin{bmatrix} \Omega_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Omega_J \end{bmatrix}. \tag{5}$$

By introducing such a prior distribution over the weights, it is now possible to allow for differential shrinkage of different groups of weights. We specifically define that each of the $j = 1, ..., D_j$ blocks in Λ relates to a single group of predictors, F_j. The column indices in X unique to predictor group F_j will thus index rows and columns in Λ for that block. Separate hyperparameters λ_j are thus considered for each predictor group as opposed to defining one hyperparameter that is shared across weights (e.g., as in Ridge regression). The terms Ω_j denote matrices of size $D_j \times D_j$ that are assumed to be fixed a priori. For now, we will assume that Ω_j, $j = 1, ..., J$ are identity matrices with D_j rows. The following parametrization [39, 40] of these terms will later be used for imposing local smoothness constraints via a Matérn covariance function with order 3/2 and length scale h_j:

$$\Omega_j = \left(1 + \frac{\sqrt{3}|k-i|}{h_j}\right) \exp\left(-\frac{\sqrt{3}|k-i|}{h_j}\right), \quad k = 1, ..., D_j, \;\; i = 1, ..., D_j. \tag{6}$$

Note that rather than attempting to estimate h_j and impose hyperpriors on the h_j terms, we will fix these parameters a priori. We complete the model by placing Inverse-Gamma prior distributions over the λ_j terms and over the ν terms

$$p(\nu|\phi, \kappa) = \text{Inv-Gamma}\left(\nu\,|\phi, \kappa\right), \tag{7}$$

$$p(\lambda_j|\eta, \tau) = \text{Inv-Gamma}\left(\lambda_j\,|\eta, \tau\right), \tag{8}$$

where $\text{Inv-Gamma}\left(x\,|a, b\right)$ denotes an inverse-gamma distribution with shape a, scale b and mode $b/(a+1)$:

$$\text{Inv-Gamma}\left(x\,|a, b\right) = \frac{b^a}{\Gamma(a)} \frac{1}{x^{a+1}} \exp\left(\frac{-b}{x}\right), \;\; a > 0, \;\; b > 0 \tag{9}$$

To simplify notation, we will henceforth let λ denote a set of $\{\lambda_j\}_{j=1}^J$ hyperparameters and not highlight dependence on terms that remain fixed. We now need to regard λ and ν as random variables and by doing so the posterior distribution over all variables takes the form:

$$p(w, \lambda, \nu | y, X) = \frac{p(y|X, w, \nu)p(w|\lambda)p(\lambda)p(\nu)}{\int\int\int p(y|X, w, \nu)p(w|\lambda)p(\lambda)p(\nu)\ \mathrm{d}w\ \mathrm{d}\lambda\ \mathrm{d}\nu} \tag{10}$$

This expression cannot be evaluated analytically. In what follows, we will resort to an empirical Bayesian estimation procedure and assume that $p(w|y, X) = \int\int p(w, \nu, \lambda|X, y)\ \mathrm{d}\nu\ \mathrm{d}\lambda$ can be approximated by $p(w|y, X, \tilde{\nu}, \tilde{\lambda})$ where $\tilde{\nu}$ and $\tilde{\lambda}$ are fixed values obtained by maximizing the marginal posterior distribution:

$$p(\lambda, \nu | y, X) = \int p(w, \lambda, \nu | y, X)\ \mathrm{d}w \tag{11}$$

We will return to assumptions entailed by this procedure later. We use an expectation-maximization (EM) framework [41, 42, 43] to find a (local) mode in this marginal distribution. In brief, this involves lower-bounding $\ln p(\lambda, \nu|y, X)$ with a function $\mathcal{F}(\lambda, \nu, q)$ that is a functional of a valid probability distribution $q(w)$ over w and a function of the parameters λ and ν [36]. We thus recast the problem into finding a set of hyperparameters as well as a probability distribution $q(w)$ that maximizes $\mathcal{F}(\lambda, \nu, q)$

$$\mathcal{F}(\lambda, \nu, q) = \int q(w) \ln \frac{p(w, \lambda, \nu|y, X)}{q(w)}\ \mathrm{d}w \tag{12}$$

Suppose we initialize the algorithm with a set of hyperparameters $\lambda^{(k)}$ and $\nu^{(k)}$ where $k = 0$. In the expectation step (E-step), we fix the hyperparameters to these values and maximize $\mathcal{F}(\lambda^{(k)}, \nu^{(k)}, q)$ with respect to $q(w)$. It can be shown that the lower bound $\mathcal{F}$ will touch the objective when $q(w) = p(w|y, X, \lambda^{(k)}, \nu^{(k)}) \equiv q^{(k)}$ where the posterior distribution of w given a fixed set of parameters λ and ν takes the form

$$p(w|y, X, \lambda, \nu) = \frac{p(y|X, w, \nu)p(w|\lambda)}{p(y|X, \nu, \lambda)} \tag{13}$$

$$= \mathcal{N}(w|\mu, \Sigma), \tag{14}$$

$$\Sigma = \left(\Lambda^{-1} + \nu^{-1}X^\top X\right)^{-1} \tag{15}$$

$$\mu = \nu^{-1}\Sigma X^\top y. \tag{16}$$

We will henceforth let $\tilde{\mu}$ and $\tilde{\Sigma}$ denote mean and covariance of $p(w|y, X, \lambda^{(k)}, \nu^{(k)})$. Defining $q(w)$ given $\lambda^{(k)}$ and $\nu^{(k)}$ is the E-step. In the maximization step (M-step) we keep this distribution fixed and now maximize $\mathcal{F}(\lambda, \nu, q^{(k)})$ with respect to λ and ν. Recall here that the denominator in Eq. 10 does not involve λ and ν. Ignoring irrelevant terms, one can show that we need to find a set of parameters $\lambda^{(k+1)}$ and $\nu^{(k+1)}$ that maximizes the following expression:

$$\{\lambda^{(k+1)}, \nu^{(k+1)}\} = \arg\max_{\lambda,\nu} \mathbb{E}_{q^{(k)}}\left[\ln\left(p(y|X, w, \nu)p(w|\lambda)p(\lambda)p(\nu)\right)\right] \tag{17}$$

where $\mathbb{E}_{q^{(k)}}[\cdot]$ denotes expectation under $q^{(k)}$ and where

$$
\begin{aligned}
\ln p(y, w, \lambda, \nu|X) = &-\frac{1}{2\nu}(y - Xw)^\top (y - Xw) - \frac{M}{2}\ln\nu - \frac{1}{2}w^\top \Lambda^{-1} w \\
&- \frac{1}{2}\ln|\Lambda| + \sum_{j=1}^{J} \ln\ \frac{1}{\lambda_j^{1+\eta}} - \sum_{j=1}^{J}\frac{\tau}{\lambda_j} \\
&+ \ln\ \frac{1}{\nu^{1+\phi}} - \frac{\kappa}{\nu} + \text{constant}
\end{aligned}
$$

Recall that w is a column vector that contains all β_j weights stacked column-wise and that Λ is a block diagonal matrix. We know that $p(w|y, X, \lambda^{(k)}, \nu^{(k)})$ is Gaussian with mean $\tilde{\mu}$ and covariance $\tilde{\Sigma}$. Let $\tilde{\mu}_j$ and $\tilde{\Sigma}_{jj}$ denote blocks in $\tilde{\mu}$ and $\tilde{\Sigma}$, respectively, associated with predictor group j. We again discard irrelevant terms, and now see that the M-step amounts to maximizing the following expression with respect to ν and λ:

$$
\mathcal{L} = -\frac{1}{2\nu}\left((y - X\tilde{\mu})^\top (y - X\tilde{\mu}) + \mathrm{Tr}(X^\top X\tilde{\Sigma})\right) - \frac{M}{2}\ln\nu \tag{18}
$$

$$
- \sum_{j=1}^{J}\frac{1}{2\lambda_j}\left(\tilde{\mu}_j^\top \Omega_j^{-1}\tilde{\mu}_j + \mathrm{Tr}(\Omega_j^{-1}\tilde{\Sigma}_{jj})\right) - \frac{1}{2}\sum_{j=1}^{J} D_j \ln\lambda_j \tag{19}
$$

$$
+ \sum_{j=1}^{J}\ln\ \frac{1}{\lambda_j^{1+\eta}} - \sum_{j=1}^{J}\frac{\tau}{\lambda_j} + \ln\ \frac{1}{\nu^{1+\phi}} - \frac{\kappa}{\nu} \tag{20}
$$

Finding a set of λ and ν parameters that fullfil $\partial\mathcal{L}/\partial\lambda_j = 0$ and $\partial\mathcal{L}/\partial\nu = 0$ yields the following closed-form update rules:

$$
\lambda_j^{(k+1)} = \frac{\tilde{\mu}_j^\top \Omega_j^{-1}\tilde{\mu}_j + \mathrm{Tr}(\Omega_j^{-1}\tilde{\Sigma}_{jj}) + 2\tau}{D_j + 2\eta + 2} \tag{21}
$$

$$
\nu^{(k+1)} = \frac{(y - X\tilde{\mu})^\top (y - X\tilde{\mu}) + \mathrm{Tr}(X^\top X\tilde{\Sigma}) + 2\kappa}{M + 2 + 2\phi} \tag{22}
$$

The procedure of fitting the model thus involves (I) specifying hyperprior parameters ϕ, κ, η and τ, (II) starting with an initial guess for $\lambda_1, \lambda_2, ...\lambda_{D_j}$ and ν, (III) estimating $\tilde{\Sigma}$ and $\tilde{\mu}$, (IV) updating λ_j and ν according to Eq. 21 and 22, and subsequently iterating between step (III) and (IV) until convergence to a local maximum.

2.1 Predictions

The above procedure yields a set of hyperparameters $\tilde{\nu}$ and $\tilde{\lambda}$ from the training data that may be sensible in several circumstances, and we will consequently focus on $p(w|y, X, \tilde{\nu}, \tilde{\lambda})$. For predictive purposes, we assume that the following approximation holds:

$$
p(\hat{y}|y, X, \hat{X}) \approx p(\hat{y}|y, X, \hat{X}, \tilde{\nu}, \tilde{\lambda}) = \int p(\hat{y}|w, \hat{X}, \tilde{\nu})p(w|y, X, \tilde{\nu}, \tilde{\lambda})\mathrm{d}w, \tag{23}
$$

That is, to make predictions about a novel response variable, $\hat{y}$, given a new set of predictors $\hat{X}$ and inferred model parameters, it suffices to focus on $p(\hat{y}|y, X, \hat{X})$ and let $p(\lambda, \nu|X, y)$ be represented by a delta function $\delta(\nu - \tilde{\nu}, \lambda - \tilde{\lambda})$ [35]. The resulting distribution is Gaussian with mean $\hat{X}\tilde{\mu}$ and covariance $\tilde{\nu}I + \hat{X}\tilde{\Sigma}\hat{X}^\top$. The terms $\tilde{\mu}$ and $\tilde{\Sigma}$ here refer to the estimates in Eq. 15 and Eq. 16

respectively. To make point predictions, we may now focus on the estimated model weights $w = \tilde{\mu}$ and simply compute $\hat{X}w$. Our focus will be on such point predictions.

For the interested reader and illustrative purposes, we visualize variability in $\tilde{\mu}$ in Figure 1-3 based on Eq. 14. It is important to stress that $p(w|y, X, \tilde{\lambda}, \tilde{\nu})$ is conditional on the point estimates of $\tilde{\lambda}$ and $\tilde{\nu}$ and that it may fail to be even close to a reasonable approximation of $\int\int p(w, \nu, \lambda|X, y)\ \mathrm{d}\nu\ \mathrm{d}\lambda$ [44].

3 Simulations

In addition to evaluating the predictive accuracy of encoding- or decoding models on held-out data, neuroimaging studies often inspect and interpret point estimates of model parameters. In linearized encoding models, model weights are often assumed to be directly interpretable in terms of the underlying brain activity generating a measured response [38]. Weights of general linear models (GLMs) in fMRI analysis [45] or temporal response functions in M/EEG analysis [27, 46] are prominent examples. However, the choice of regularization can significantly impact the relative magnitude of the estimated weights, in particular in problems with many, potentially correlated, sets of predictors. It can therefore be relevant to simulate data with known target model weights in order to better understand the properties of regularized regression estimators and identify situations where they can be useful or can complicate interpretation.

Below, we consider different simulations to illustrate the properties of the described model. We compare the proposed EM-based banded regression estimator (henceforth EM-banded) with a standard Ridge regression estimator: $w = \left(X^T X + \alpha I\right)^{-1} X^T y$ [11]. To simplify readability we consider simulations where we fix the parameters related to the Inverse-Gamma priors to the same value, γ, such that $\tau = \eta = \kappa = \omega = \gamma$.

3.1 Simulation 1: suppressing contributions from a group of stimulus-irrelevant predictors

In models with groups of predictors representing different stimulus feature spaces, it is sometimes needed to identify entire feature spaces that can be excluded from an analysis without compromising predictive accuracy. To simulate such a situation, we here define three stimulus feature spaces, F_1, F_2 and F_3, and one response variable y which contains mixed versions of F_1 and F_3, but not F_2 (F_2 does not contribute to the response). We thus simulate the response as $y = F_1 w_1 + F_3 w_3 + \epsilon$ where ϵ is a Gaussian noise term. In this situation, we thus need the estimator to identify (close to) zero weights for F_2. Each feature space has 64 predictors and a total of $M = 1024$ observations are simulated. The weights, w_1 and w_3, as well as the feature sets F_1 and F_3 are drawn from Gaussian distributions. The mixed target signal and the noise term have equal power. We compare weights estimated by the EM-banded model with weights estimated by Ridge regression models. To exemplify how hyperparameters controlling inverse-gamma hyperpriors can impact the estimated weights, we fit models for γ fixed to 10^{-4}, 10^{-3}, 10^{-2} and 10^{-1}. The Ridge α hyperparameter is set to 10^4, 10^3, 10^2 and 10^1. The terms Ω_j are identity matrices.

Figure 1 shows results from Simulation 1. The Ridge regression model yields "dense" solutions for all considered Ridge hyperparameters, α, with no dramatic overall scale differences between weights associated with F_1, F_2 or F_3. In contrast, the EM-banded model applies excessive regularization to feature set F_2 for smaller values of γ. As expected, the EM-banded model tends to show signs of under-shrinkage when the inverse-gamma prior distributions have little mass near zero and in these cases, the overall scale differences between the different sets of weights associated with F_1, F_2 and F_3 are smaller.

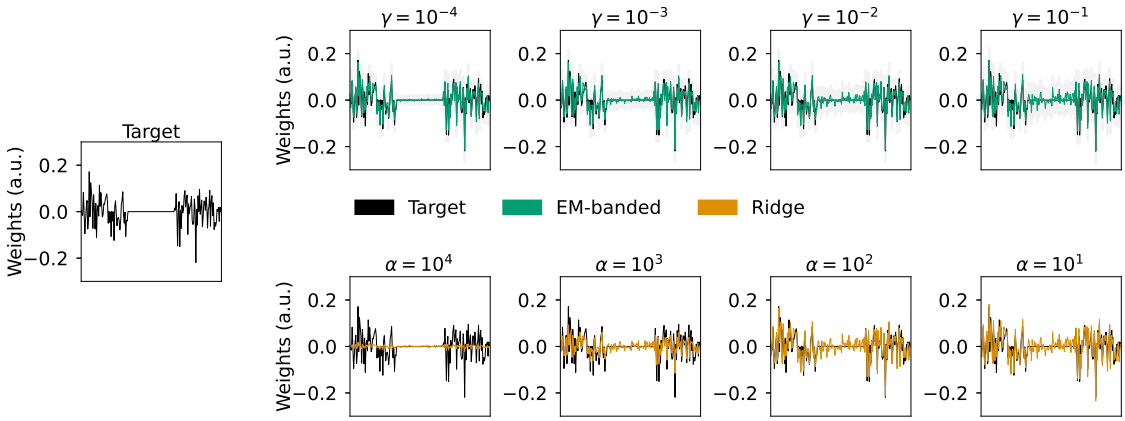

Figure 1: Simulation 1 illustrating regularization properties in a regression problem with a group of stimulus-irrelevant predictors. Left middle panel: simulated target weights (black). The predictors are divided into three distinct predictor groups, F_1, F_2 and F_3 where only target weights associated with F_1 and F_3 are non-zero ("relevant"). Top row: weights estimated with EM-banded model (red) for different values of $\eta = \phi = \tau = \kappa = \gamma$. Shaded gray errorbars depict 1% respectively 99% percentile of samples from a multivariate Gaussian with mean and covariance defined according to Eq. 14. Bottom: weights estimated with Ridge regression model (yellow) for different Ridge regularization parameters α. The target weights are shown in black.

3.2 Simulation 2: encouraging smoothness for a feature subset

It is sometimes relevant to encourage smoothness on sets of weights, for instance, in analyses that approximate smooth FIR filters. An example hereof is fMRI encoding analyses that model the relation between a regressor augmented with multiple time-lags and the BOLD voxel response. In this case, it can *a priori* be expected that the approximated FIR filter should be smooth due to hemodynamic properties of the BOLD signal [47]. It can similarly be relevant to assume smoothness across spectral channels in analyses that model relations between average stimulus power spectra and BOLD fMRI responses.

Here, we explore one way of encouraging smoothness using the definition of Ω_j in Eq. 6. To illustrate the effect of Ω_j, we consider a simulated regression problem with two feature spaces, F_1 and F_2. Each feature space has 64 predictors and is drawn from a Gaussian distribution. Data is simulated as $y = F_1 w_1 + F_2 w_2 + \epsilon$ where ϵ is drawn from a Gaussian. A total of $M = 1024$ observations are simulated. The weights of the first feature space w_1 are simulated to be sinusoidal (i.e., smooth), while w_2 weights are Gaussian. The signal-to-noise ratio (SNR) is -5 decibels (dB) and we fit the EM-banded models with $\gamma = 10^{-4}$. The Ω_1 term associated with F_1 is parameterized according to Eq. 6 with h_1 set to 1, 5, and 10 to illustrate its effect.

Figure 2 shows weights estimated with Ridge estimators for different α values (below) and weights estimated by the EM-banded estimators (above). At higher levels of h_1, the EM-banded approximates the weight distribution of the first feature space F_1 with high accuracy. The Ridge estimator, on the other hand, yields weights that tend to show more rapid (high frequency) fluctuations across neighboring weights associated with F_1 compared to the target weights.

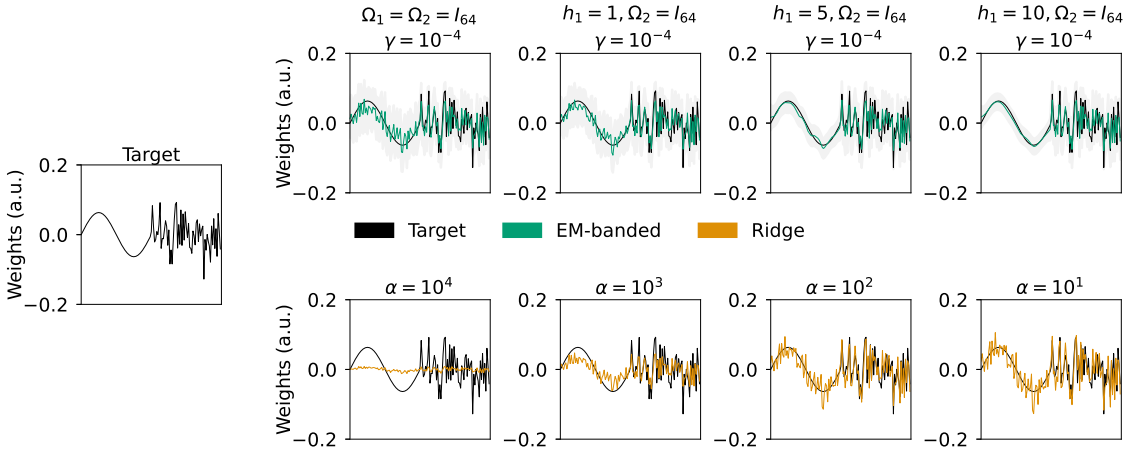

Figure 2: Simulation 2 illustrating effects of smoothness. Left middle panel: simulated target weights (black) divided into two distinct predictor groups, F_1 and F_2. Weights associated with F_1 are sinusoidal, while weights associated with F_2 are drawn from a Gaussian. Top row: weights estimated with the EM-banded model (red) for $\gamma = 10^{-4}$ but with different parameterizations for Ω_1 (defined in Eq. 6). Specifically, we change the h_1 term in this equation in order to encourage smoothness constraints. Shaded gray errorbars depict 1% respectively 99% percentile of samples from a multivariate Gaussian with mean and covariance defined according to Eq. 14. Bottom: weights estimated with Ridge regression model (yellow) for different Ridge regularization parameters α. The target weights are shown in black.

3.3 Simulation 3: encouraging differential shrinkage of individual weights

In simulation 1, the EM-banded model was used to shrink weights of an entire feature space (i.e., groups of predictors). The model can similarly encourage differential shrinkage of individual weights. To illustrate this, we consider a simulation in which we assign distinct λ_j parameters to each individual predictor in a simulated regression problem. We simulate 512 predictors predictors, F_1, F_2, ..., F_{512}, and a response variable y which contains mixed versions of some of these predictors. A total of $M = 1024$ observations are simulated. There is a high proportion of irrelevant predictors, i.e. only a small subset of the "true" regression weights w_j are non-zero. It is assumed that $y = \sum_j F_j w_j + \epsilon$ where ϵ is Gaussian. Each row in $[F_1, F_2, ..., F_{512}]$ is drawn from a multivariate Gaussian with zero mean and a covariance defined as $C = \exp(-0.3|k - i|^2), k = 0, ..., 511, i = 0, ..., 511$. The SNR is equal to 0 dB. We again compare Ridge regression model fits with EM-banded model fits using the same set of fixed hyperparameters as in simulation 1.

The estimated weights for the Ridge and EM-banded model are shown in 3. It is evident that the EM-banded estimator model where $\gamma = 10^{-4}$ excessively shrinks several of the individual weights. In effect, the EM-banded model recovers the target weights with relatively high accuracy. In comparison, the Ridge estimator tends to dilute effects across neighboring (correlated) predictors when α attains high values and pull all weights towards zero.

3.4 Simulation 4: Interpretation of weights with correlated predictors

Until now, we have not explored how strong correlations among groups of predictors can impact results. Models involving correlated groups of predictors occur often in EEG or fMRI encoding models that model responses to naturalistic stimuli [1]. To exemplify this situation, this simulation considers the example of predicting responses to natural speech. Speech signals can be characterized by their

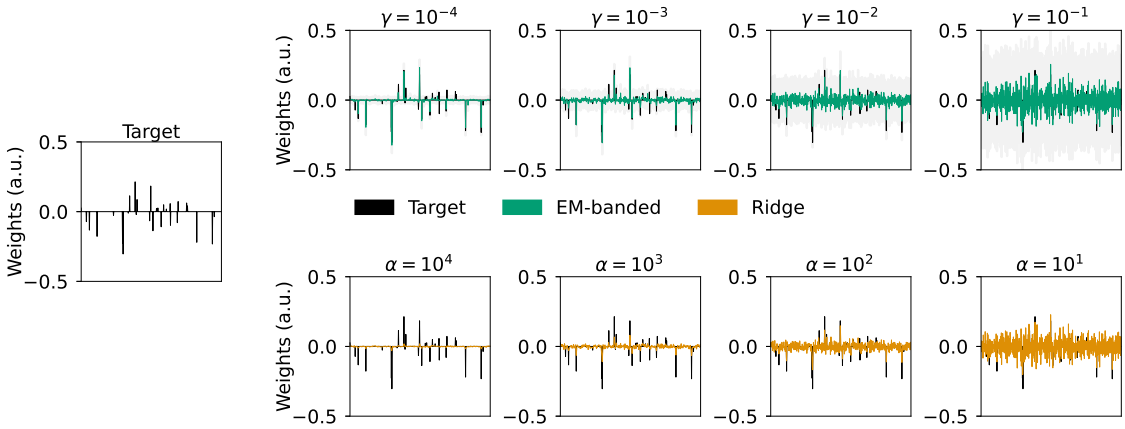

Figure 3: Simulation 3 illustrating differential shrinkage of individual weights. Left middle panel: simulated target weights (black). Top row: weights estimated with EM-banded model (red) for different values of $\eta = \phi = \tau = \kappa = \gamma$. Shaded gray errorbars depict 1% respectively 99% percentile of samples from a multivariate Gaussian with mean and covariance defined according to Eq. 14. Bottom: weights estimated with Ridge regression model (yellow) for different Ridge regularization parameters α. The target weights are shown in black.

lower-level acoustic features as well as higher-level phonetic or semantic features [8, 48]. The different hierarchical levels can each be modeled by multidimensional features, but the statistics of natural speech imply a degree of correlation between such feature groups (see e.g. panel B in Fig. 4). This can complicate inferences about neural responses unique to a given group.

To illustrate this, we simulate responses to speech acoustic and phonetic features extracted from natural speech audio [49] (see section C.1 for details on feature extraction). A first feature set F_1 contains time-lagged spectrogram features and F_2 contains time-lagged regressors created from phonetic annotations. We simulate a response as $y = F_1 w_1 + \epsilon$, i.e., only the feature set associated with the spectrogram features F_1 affects the response variable, y. The SNR is equal to 0 dB and the target filter weights, w_1, are shown in Figure 4A. We again compare Ridge and EM-banded regression model fits using a set of fixed hyperparameters.

Figure 4 shows the results of this simulation. The EM-banded estimator accurately recovers the weights associated with the speech spectrogram F_1 while excessively shrinking the weights associated with the phonetic features F_2, as desired. The Ridge estimator, on the other hand, pulls weights associated with correlated predictors in the two feature groups towards each other at higher levels of regularization. This has the undesired consequence that a temporally located response function associated with the ''irrelevant" phonetic feature group F_2 emerges. This can complicate the interpretation of the weights associated with F_2, e.g. if such point estimates of weights were interpreted or further analysed in second-level inferences. Simulating the opposite case, i.e., simulating only responses to phonetic features (and no contribution from spectrogram features), leads to the equivalent result: weights associated with the "irrelevant" spectrogram feature group show temporally located response functions (see section C.2). The described effect depends on the degree of correlation among predictors and on the amount of data available for fitting the models. We note that high levels of regularization often occur in practice for Ridge estimators, for instance, when tuning the regularization term by maximizing Pearson's correlation that is insensitive to scaling and shifts between target and predicted variables. For a related, but more stylized simulation, see section C.3.

This simulation highlights a potential source of pitfalls for the interpretation of model weights in

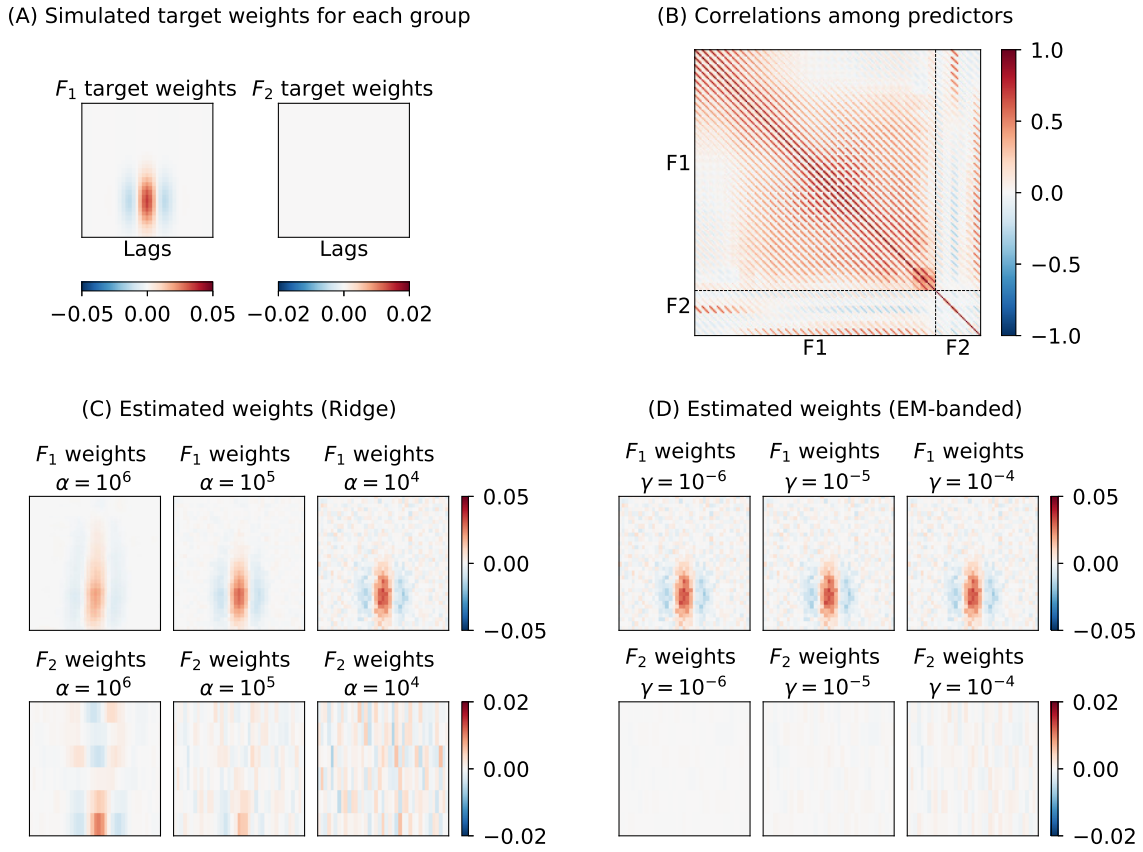

Figure 4: Simulation 4 illustrating behavior of the Ridge estimator and the EM-banded estimators when there are strong correlations among groups of predictors, F_1 and F_2. The target variable is simulated as $y = F_1 w_1 + \epsilon$. (A) Target weights associated with each of the two feature sets. (B) Pearson product-moment correlation coefficients among all predictors. Dashed lines are used to highlight predictor groups, F_1 and F_2. (C) Weights estimated with Ridge estimator for each of the two groups. (D) Weights estimated with the EM-banded estimator for each of the two groups. Ridge estimators with α set to 10^6, 10^5 and 10^4 are shown. EM-banded estimators were considered with γ set to 10^{-6}, 10^{-5} and 10^{-4}

situations with co-varying features affecting the response. To illustrate this further, consider again the simulation where the response is driven by the first feature set F_1, but now only the correlated set F_2 is included in the model. Figure 5 depicts the estimated weights by Ridge models in this case. The model with only F_2 predictors now misleadingly indicates a temporally located response for all of the considered α parameters, despite the fact that the target regression weights associated with F_2 are zero. Interpretation of the reduced model is clearly different from the model that models both groups of predictors.

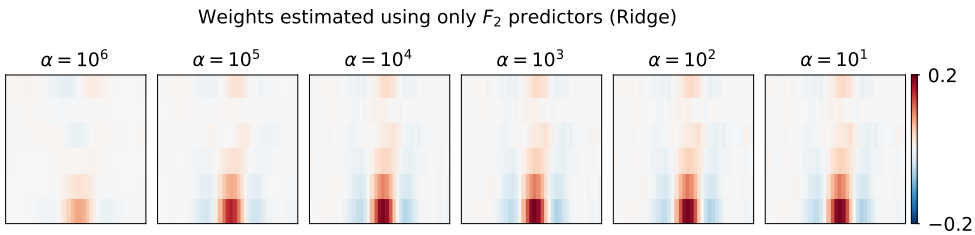

Figure 5: Overlooking important features in a Ridge regularized encoding model. The outcome variable is again simulated as $y = F_1 w_1 + \epsilon$, i.e., the simulated target contains only F_1 and not F_2, but only F_2 is included in the prediction model. Panels reflect estimates with different Ridge regularization strengths.

3.5 Simulation 5: visualizing behavior under unfavorable signal-to-noise ratios

As illustrated in Simulation 3, the EM-banded estimator can be used to encourage differential shrinkage of weights associated with individual predictors. The behavior of such an estimator will depend on several factors, including the amount of data available to support the identification of λ, correlations among predictors, as well as the SNR. Here, we illustrate the behavior of the EM-banded estimator in a simulation with poor SNR. We simulate two predictors F_1 and F_2 and a response variable $y = F_1 w_1 + F_2 w_2 + \epsilon$. The predictors as well as the noise term ϵ are drawn from Gaussian distributions. The target weights are fixed to non-zero values. The SNR is approximately -20 dB. We consider simulations where the number of observations is 512, 1024 and 8192, and we fit EM-banded models to the data from each simulation. We focus on estimators with $\gamma = 10^{-4}$ respectively $\gamma = 10^{-2}$ to illustrate how differences in these prior choices affect the estimated parameters. This procedure is repeated many times to elucidate the effect.

Figure 6 shows 2D histogram visualizations of the estimated weights w_1 and w_2 from these simulations. Notably, the estimator with the lower $\gamma = 10^{-4}$ tends to exhibit excessive shrinkage to either of the two weights when the number of observations are low ($M = 512$). Such over-shrinkage is undesirable. The estimator even collapses to $w_1 \approx w_2 \approx 0$ occasionally, which again is undesirable. Increasing γ to 10^{-2} here yields dramatically different estimates and less differential shrinkage of the two weights. These simulations underline subtleties that can yield dramatic effects on the estimated weights, which is important to keep in mind when focusing on "default" priors for λ_j.

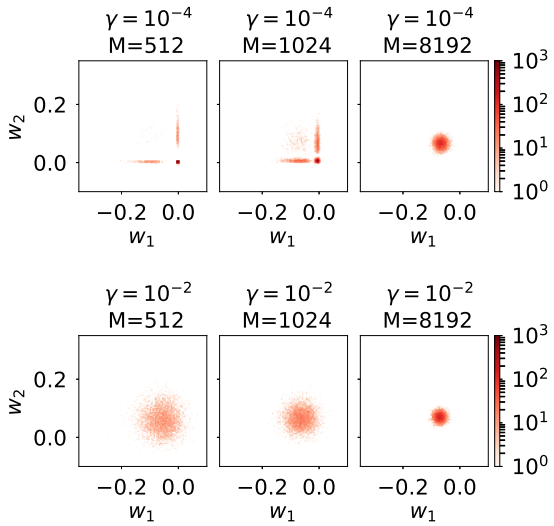

Figure 6: Simulation 5 illustrating regularization properties with only two predictors and a poor SNR ratio. 2D histogram visualizations of estimated regression weights, w_1 and w_2. Columns show simulations with increasing number of observations. Top row: weights estimated with the EM-banded estimator with $\gamma = 10^{-4}$. Bottom row: weights estimated with the EM-banded estimator with $\gamma = 10^{-2}$.

4 fMRI encoding analysis with "stimulus-irrelevant" predictors

Building encoding models often involve deciding on which of multiple potentially relevant feature sets to include as predictors. Choosing features both implies a risk of overlooking feature sets that are potentially relevant, as well as including feature sets that are irrelevant for predicting the target variables. In this section, we investigate how the inclusion of stimulus-irrelevant predictors - here defined as simulated noise predictors that are unrelated to the task or stimuli - can impact results in an fMRI encoding analysis and how the EM-banded estimator may help suppress their contribution.

The example uses a publicly available BOLD fMRI dataset [50] described in [37, 51]. The dataset contains BOLD fMRI data acquired from five participants listening to excerpts from music in 12 training runs and 6 test runs. Each functional run lasted 10 min, each consisting of 40 excerpts from musical clips from [52]. We focused on encoding models with three stimulus feature spaces as predictors of the BOLD fMRI responses. The first two feature sets were extracted from the music stimuli: a 10-dimensional genre-label feature given by [52] indicating which of 10 musical genres a given stimulus is assigned to (F_1), and a time-averaged 32-dimensional spectrogram of the audio stimulus (F_2). As a the third feature set (F_3) unrelated to the stimuli, we included a 400-dimensional multivariate Gaussian noise. These feature spaces each constitutes a group of predictors with individual λ_j parameters for the EM-banded estimator, i.e., we allowed for differential shrinkage of weights associated with each of these three feature spaces. For simplicity, we always initialize the EM algorithms with λ_j and ν parameters set to 1 and define that no more than 200 iterations should be considered. Note that the algorithm is susceptible to local maxima, and it may sometimes be advisable to use different initialization strategies. More details related to fMRI preprocessing, audio feature extraction, and

model fitting are described in section A.

Figure 7 shows the predictive accuracies obtained with the Ridge estimator and the EM-banded estimator. Predictive accuracy is here evaluated both as Pearson's correlation coefficient between the predicted and measured BOLD fMRI response data, and as explained variance ratio. Compared to the Ridge estimator, the EM-banded estimator yielded higher predictive accuracy for all subjects in both goodness of fit metrics (all $p < 0.05$, Wilcoxon signed-rank tests across all voxels used for fitting models). The improvement in terms of Pearson's correlation coefficient among the top 10000 voxels was (mean $\pm$ standard deviation): 0.0335 ± 0.0515 (sub-001), 0.0411 ± 0.0629 (sub-002), 0.0568 ± 0.0699 (sub-003), 0.0391 ± 0.0637 (sub-004), 0.0298 ± 0.0522 (sub-005).

The regularization term α of the Ridge estimator varies substantially across voxels. This can occur, for instance, if voxels with higher signal-to-noise ratio require less overall regularization. Changes in α impacts the overall scale of all weights. This can lead to misleading spatial weight maps indicating effects of predictors that are stimulus-irrelevant. This is illustrated in figure 8 showing the group-mean standard deviation of predictions obtained with stimulus-irrelevant predictors. The predictions were obtained using only the task-irrelevant predictors to predict voxel responses (discarding the spectrogram and genre features) with the Ridge estimator. As can be seen, the standard deviation of these predictions is high specifically in auditory cortical regions. This may reflect that the Ridge estimator declares less overall shrinkage in voxels in regions with higher SNR. In effect, the map can misleadingly be interpreted to suggest regional auditory-evoked activation to a random feature *because of* the regularization. Encouraging differential shrinkage of each feature space with the EM-banded estimator avoids this problem, showing no clear spatial patterns in these maps.

5 EEG-based decoding analysis

In this example, we consider a decoding analysis of single-trial EEG responses to continuous natural speech. The example follows a popular approach of predicting the continuous speech amplitude envelope using linear combinations of time-lagged EEG scalp electrode responses at lower frequencies [27, 53, 54, 55, 5]. In this approach, the goal is to find a multi-dimensional filter w that linearly maps from (features in) the multi-channel EEG response to the envelope of the speech stimulus y. The decoding analysis can be formulated as $\hat{y}(t) = \sum_c \sum_\tau x(t-\tau, c) w(c, \tau)$ where $x(t-\tau, c)$ denotes preprocessed electrode response in channel c at time point $t-\tau$ and where τ indicates the time delay between stimulus and response. The model can be specified by concatenating multiple time-lagged versions of each EEG electrode response in a design matrix X in Eq. 1 [53]. The regression problem now amounts to estimating coefficients of a multi-dimensional FIR filter [30]. In this context, it is natural to consider each time-lagged electrode response as a feature group F_j in Eq. 2 and allow for individual weight shrinkage for each group. For instance, it can be relevant to allow for excessive shrinkage of a noisy or irrelevant electrode for all its time-lag copies.

While EEG speech decoding studies often focus on low-frequency activity (< 15 Hz), it is known from intracranial recordings that the power of higher frequency neural activity in the gamma range also track features of speech stimuli [31, 56, 57]. These frequencies are typically highly attenuated in scalp EEG. Nonetheless, including high gamma power features has been shown to improve predictive accuracy of EEG decoding models for some subjects [58]. However, including more EEG features, some of which may have very low SNR, is also be associated with a risk of adding predictors to the decoding model that introduce noise to the predictions. It may thus be desirable to allow for differential shrinkage of different EEG features. In this example, we therefore explore predictive accuracy of the banded EM model when both lower and higher frequency EEG features are included in such a decoding analysis.

The example uses publicly available EEG data (https://doi.org/10.5061/dryad.070jc) described in [5]. In brief, 19 subjects listened to an audiobook while continuous EEG data from 128 scalp electrodes were recorded. Data were acquired in 20 trials each approximately 180 s in length. We fit decoding models again using the EM-banded framework compared to standard Ridge regres-

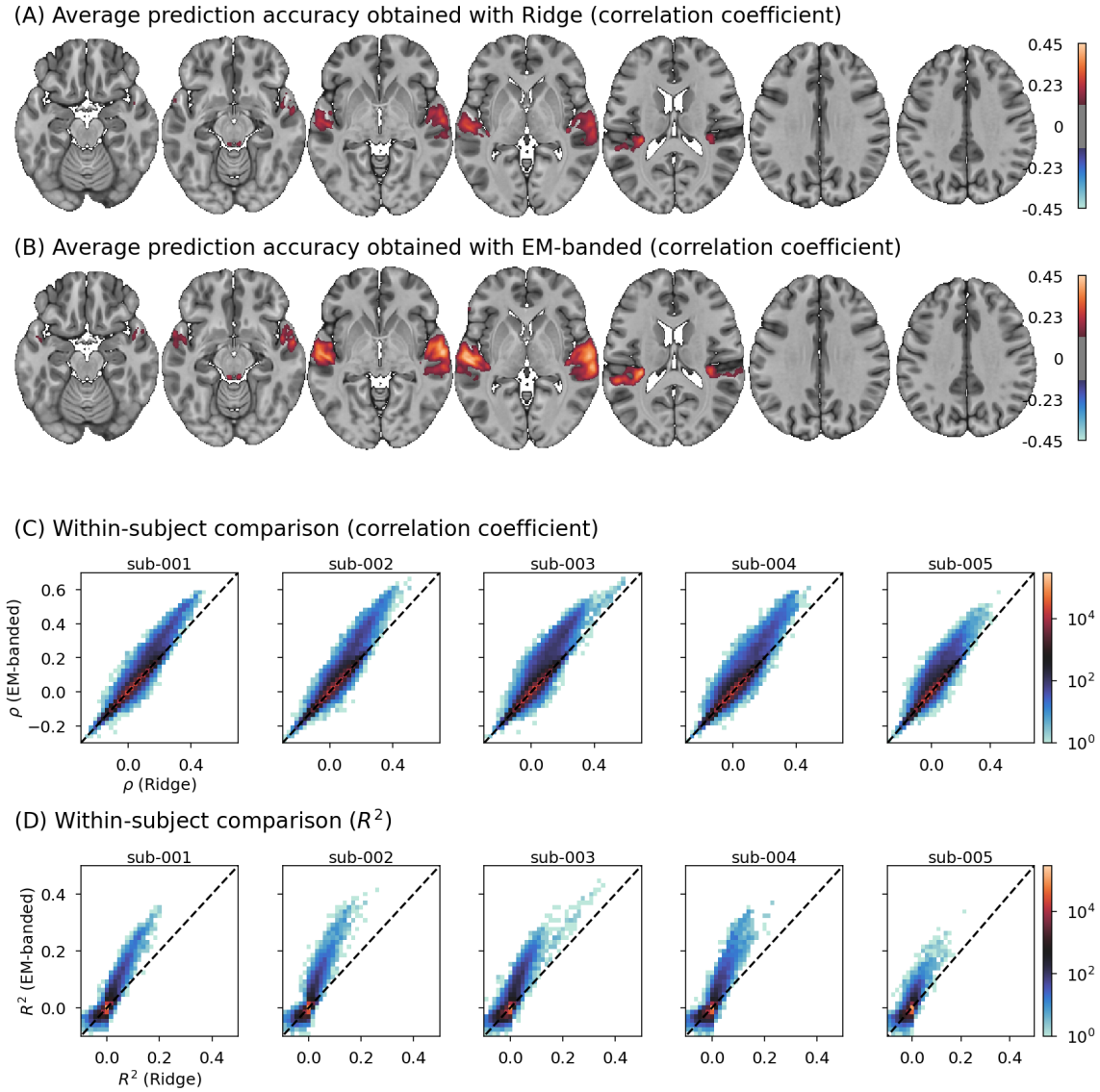

Figure 7: A BOLD fMRI encoding analysis where stimulus-irrelevant predictors are included in the analysis. (A): group-mean prediction accuracies for models trained with the EM-banded framework. (B): group-mean prediction accuracies for cross-validated Ridge regression models. Maps in (A) and (B) have been arbitrarily thresholded for visualization purposes. (C) 2D histogram visualization of prediction accuracies (Pearson's ρ) for the EM-banded and Ridge estimators for each subject. The dashed line indicates similar performance across models. (D) Similar visualizations as in (C) but using R^2 as performance metric.

sion. For both models, we used the time-lagged low-frequency (1-8 Hz) EEG activity as well as the envelope of EEG power fluctuations in a higher frequency band (25-35 Hz) [59] to predict the speech envelope. Further details related to EEG preprocessing, audio feature extraction, and model fitting

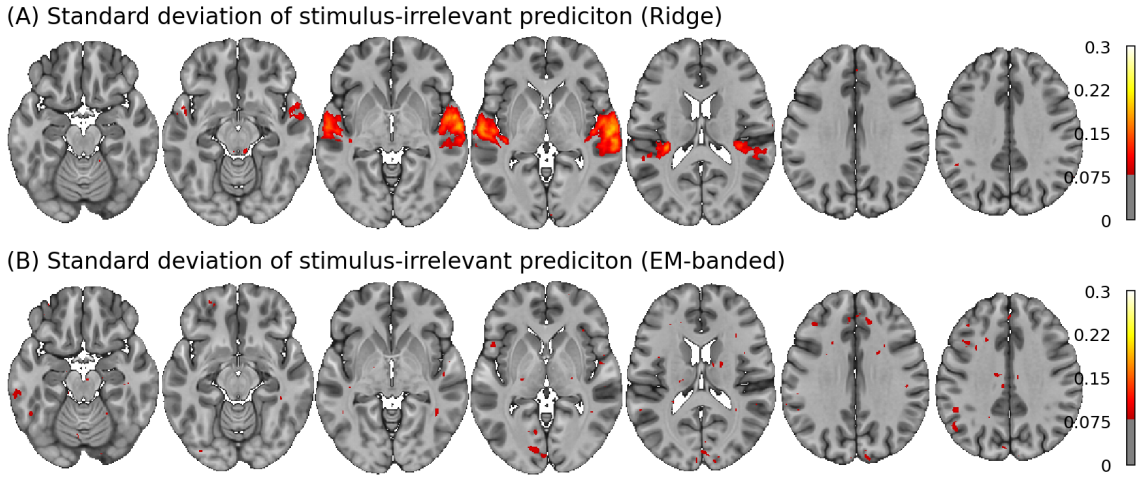

Figure 8: Group-mean standard deviations of predictions with a stimulus-unrelated noise feature estimated with Ridge regression (above) and the EM-banded framework (below).

are described in section B. Note that cross-validation was only used for hyperparameter tuning in the Ridge model.

The results from the analysis are shown in figure 9. The prediction accuracies obtained with the EM-banded estimator are modestly but consistently higher than the Ridge estimates ($p < 0.01$, Wilcoxon signed rank test). The average correlation between the predicted speech envelope and the target envelope is 0.1368 ± 0.0383 for the EM-banded model and $0.1267 \pm 0.0.0377$ for the Ridge model. The EM-banded approach excessively shrinks groups of weights, whereas the Ridge estimator yields a denser distribution of weights. Figure 9 (right) shows an example of weights estimated by the two models for one subject. This illustrates how the overall scale of weights associated with the HF features and LF features are comparable for the Ridge estimator whereas the EM-banded estimator differentially shrinks both weights associated with HF features and groups of LF features associated with single electrodes. Given that high-frequency EEG features are weak and in some subjects not even measurable by scalp electrodes, including HF features can negatively model impact predictions, in particular with a Ridge estimator. Differential shrinkage of groups of weights may in such cases allow for better exploitation of potentially informative HF features.

6 Discussion

We explored a framework for estimating "banded"-type [24, 23] encoding- and decoding models. The framework can be used to specify hyperparameters that control shrinkage of groups of regression weights. Specifically, the model can be used to 1) specify one hyperparameter that is shared across all weights, 2) specify individual hyperparameters for each weight (e.g., as in [35]), 3) specify individual hyperparameters to groups of weights, and 4) encourage smoothness constraints on groups of weights. In the following, we discuss modeling considerations and model limitations.

6.1 Encoding- and decoding models with correlated predictors

Regression models with highly correlated predictors frequently occur in encoding- and decoding analyses. Features of continuous naturalistic stimuli (e.g., video or speech) used as predictors in encoding

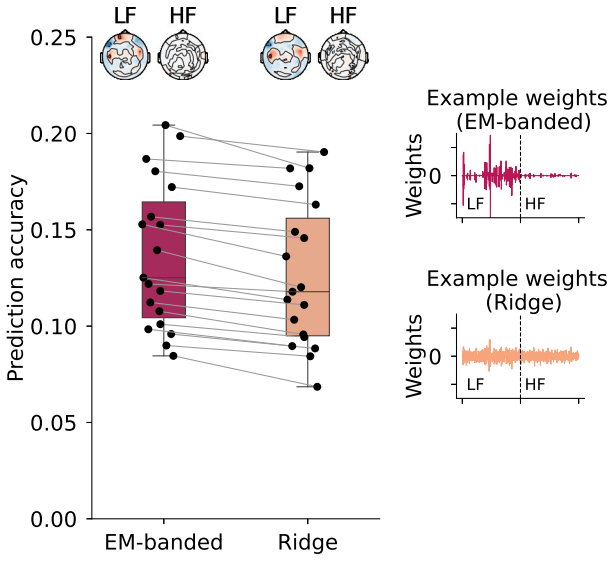

Figure 9: Results from the EEG decoding analysis. Left: boxplot of prediction accuracies obtained with each model. Each point reflects data from a given subject. Prediction accuracy was defined as the average correlation coefficient between the target envelope and decoded envelope in held out trials. Topographies depict weights of each model at a single time lag for illustrative purposes. The weights are here first normalized by their standard deviation and averaged across participants. Topographies are shown both for weights associated with low-frequency (LF) features and higher-frequency (HF) features. Right: Examples of regression weights for one participant.

models are often highly correlated. Continuous scalp EEG channels used as predictors in decoding models are typically also highly correlated, and pairs of channels can commonly exceed $\rho = 0.8$ with high-density EEG recordings.

The described estimator may encourage solutions where subsets of correlated predictors are excessively shrunken. However, the true posterior of weights in analyses with many correlated predictors may show complex multimodal distributions, and point estimates of model parameters (or, approximate conditional distributions) should be interpreted with caution (see e.g., Simulation 5). Excessive shrinkage of a given feature set could, e.g., be misinterpreted as indicating that only one of multiple feature sets is encoded in a response. Such issues can be particularly difficult to identify when there are many outcome variables (e.g. different voxels with different noise profiles), and little training data to support the identification of model parameters.

Shrinkage procedures can complicate interpretation of weights in encoding analyses with correlated predictors, as illustrated in our simulations. In decoding models, model prediction accuracy on held-out data is often a primary measure. If the main goal is to optimize model performance for some application (e.g., a decoding model optimized for BCI performance), then the choice of shrinkage procedure may be less critical for interpretation as long as it does not compromise performance or lead to exclusion of important features. However, another goal may be to correlate prediction accuracies with behavior or clinical outcome measures. Here, the choice of shrinkage procedure can affect group-level inferences. For instance, multi-channel EEG data may show distinct noise characteristics in different individuals (e.g., noisy electrodes or attenuation of higher-frequency activity due to head anatomy). A given shrinkage procedure might differentially impact decoding accuracies across individuals, potentially complicating the interpretation of group-level correlations with behavior.

With highly correlated predictors, it can sometimes be relevant to transform the predictors using data dimensionality reduction techniques [60, 61, 62, 63, 64, 65] and then fit model using the transformed data. This can potentially reduce computational burden without compromising predictive accuracy. Dimensionality reduction can be combined with banded regularization procedures to allow differential shrinkage of regression weights associated with each component. The described framework can also be used in the context of feature selection [66, 67, 68, 69], for instance, to inform about which stimulus-response lags to consider. Whether the assumptions entailed by such procedures are reasonable again depends on the specific research goals.

6.2 Computational burden

Encoding- and decoding analyses often involve high-dimensional regression problems. In encoding models, the stimulus or task feature selection is often motivated by some pre-defined hypotheses. Yet, even for well-defined hypotheses, feature sets often translate to a high number of predictors. This, in turn, implies a high computational burden. The proposed EM model involves matrix inversions (often with complexity $\mathcal{O}(D^3)$ unless specific structure of the matrix can be exploited) which may limit the relevance of the described model, in particular, if the number of outcome variables is high. Alternative algorithms described in [24, 23] can be particularly relevant for fMRI encoding analyses with many predictors and a large number of response variables (voxels). For completeness, section D considers an alternative formulation of the EM-banded model where it is assumed that the λ_j and ν parameters are shared across outcome variables to allow for a more efficient implementation in scenarios where there are many outcome variables.

6.3 Risks of under-shrinkage and over-shrinkage

The described model applies independent inverse-gamma priors on λ_j terms associated with different groups of predictors. The choice of hyperpriors, or lack hereof, can have a large impact on the results and misspecified hyperpriors can be associated with over-shrinkage or under-shrinkage, as exemplified in our simulations. In both cases, this may compromise both predictive accuracy and interpretability of model weights. In the encoding and decoding analysis examples, we focused on inverse-gamma hyperpriors with hyperparameters set to $\tau = \eta = \kappa = \omega = 10^{-4}$. This pragmatic choice was found to work well in several analyses, allowing for excessive shrinkage of (groups of) predictors while also putting prior mass on larger values for λ_j and ν. This choice could work well in various types of encoding or decoding analyses and fixed hyperparameters would allow for determination of λ_j and ν terms from training data alone without resorting to cross-validation. However, there may be many cases where this choice of priors is not optimal. For example, in regression problems with extremely low signal-to-noise ratios the estimator may excessively shrink relevant (groups of) predictor(s) (see e.g., Section 6). Again, whether this behavior is desirable will depend on the research question, and the choice of prior should be informed about the expected signal-to-noise ratio.

6.4 Other perspectives

We used the described modeling framework to fit encoding- and decoding models trained on data from single subjects (i.e., single-subject analyses). The described framework could also be incorporated in "searchlight" decoding analyses [70] where different subsets of features from multi-dimensional neural data recordings are used as predictors in separate decoding models. Such procedures could be relevant in both single-subject analyses and group-level [71] analyses (e.g., analyses that explore associations between structural MRI measures and clinical scores) where there are many neural channels / voxels.

One appealing aspect of group shrinkage procedures is that they can allow for ranking importance of feature groups. This may aid interpretation of encoding analyses [24, 23]. In recent years, encoding models have been used increasingly to model neural responses with features extracted from task-optimized neural networks [7, 72, 73, 74]. This typically involves very high-dimensional feature sets and

hence encoding models where the number of predictors is much larger than the number of observations, i.e., $D >> M$. Surprisingly, over-parameterized models may show good generalization performance even with vanishing explicit regularization in minimum-norm least squares estimators [75, 76, 77]. This behavior is clearly different from analyses where explicit regularization is critical for performance. Yet, the ability to establish importance of sets of features can similarly be useful in these settings, and group shrinkage procedures may be one way of achieving this.

7 Conclusion

In this paper, we explored a framework for estimating banded-type regression models in encoding- and decoding analyses. It offers a tool that can have relevance for specific modelling problems in computational neuroscience, complimenting other related estimators [24, 23, 21, 22]. We used simulations and data examples to illustrate properties of the described model in comparison with Ridge estimators. Several of the examples highlight scenarios where a focus on maximum a posteriori (MAP) estimates of model parameters can be associated with an appreciable risk of misinterpretation.

Declaration of Competing Interest

HRS has received honoraria as speaker and consultant from Lundbeck AS, Denmark, and as editor (Neuroimage Clinical) from Elsevier Publishers, Amsterdam, The Netherlands. He has received royalties as book editor from Springer Publishers, Stuttgart, Germany, Oxford University Press, Oxford, UK, and from Gyldendal Publishers, Copenhagen, Denmark.

Funding

This work was supported by the Center for Auditory Neuroscience grant from the William Demant foundation and by a synergy grant from Novo Nordisk foundation to the UHEAL project Uncovering Hidden Hearing Loss (grant number NNF17OC0027872). HRS was supported by a grand solutions grant "Precision Brain-Circuit Therapy - Precision-BCT)" from Innovation Fund Denmark (grant number 9068-00025B) and a collaborative project grant "ADAptive and Precise Targeting of cortex-basal ganglia circuits in Parkinson's Disease - ADAPT-PD" from Lundbeckfonden (grant number R336-2020-1035). OP was supported by a grant from Lundbeckfonden (grant number R360–2021–395). KHM was supported by the Pioneer Centre for AI, DNRF grant number P1.

Acknowledgements

The authors would like to thank Edmund C. Lalor (University of Rochester) for helpful discussions.

Data and Code Availability

Code implementations in both Matlab (The MathWorks, Inc., Natick, Massachusetts, United States) and Python are available on https://github.com/safugl/embanded. The Python code relies on *numpy* [78], *scipy* [79] and *numba* [80]. EEG data and BOLD fMRI data used in this study have been presented previously [5, 37, 50] and is available on https://doi.org/10.5061/dryad.070jc and https://openneuro.org/datasets/ds003720, respectively.

A fMRI encoding analysis: preprocessing, feature extraction and model fitting

A.1 fMRI preprocessing

We used a publicly available BOLD fMRI dataset [50] described in [37, 51] which contains data acquired from five participants (procedures were approved by local ethics and safety committees [51, 37]). We considered a minimal BOLD fMRI preprocessing pipeline which involved motion correction [81], spatial smoothing (with a 3 mm FWHM Gaussian kernel) and skullstripping. These processing steps relied on SPM 12.6685 [82] and FSL 6.0.1 [83, 84]. Part of the motivation for incorporating spatial smoothing was to improve signal-to-noise ratio, but also that susceptibility-induced distortions rendered it difficult to adequately align volumes across functional runs. Six motion regressors (translations in x, y, and z directions and rotations about x, y, and z) were extracted for each subject and functional run. These were used to create a set of nuisance regressors that further included a cosine basis set from a one-dimensional discrete cosine transform. Time courses from each voxel were deflated using this set of regressors. We defined the average response to a given sound stimulus block as the average of these residualized voxel time courses from 6-s after block onset to 16.5-s after block onset (except in the last block where we averaged from 6-s after block onset to the last acquired volume). This was a pragmatic choice and effectively means that only a subset of the block will contribute to the estimated "average" response (while minimizing contamination from the previous block). The first block of each functional run (both test- and training runs) was discarded from all further analyses. All analyses were conducted in native subject space. For visualization purposes, we mapped the resulting goodness of fit maps to a Montreal Neurological Institute (MNI) template. This was achieved using tools available with Advanced Normalization Tools (ANTs) [85]. Visualization of results relied on scientific computing software in Python, and included use of tools *nilearn* [86], *seaborn* [87], *numpy* [78], and *matplotlib* [88].

A.2 Stimulus-related and stimulus-irrelevant features

We extracted two features from the music stimuli in each experimental block. The first, F_1, was a genre-label dummy coded feature that indicated which of 10 genres the stimulus excerpt was assigned to by [52]. The second, F_2, was a time-averaged "spectrogram" representation of the speech waveform. The audio stimuli were first passed through a 32-channel "gammatone" filterbank (as implemented by [89]) with filters spaced from 50 to 11050 Hz [90]. Envelopes were extracted at the output of each filter via the absolute value of the Hilbert transform, compressed by a power-law exponent ($c = 0.3$), and averaged from 2-s after block onset to 11-s after block onset. Finally, we considered a third feature space, F_3, which contained 400 stimulus-irrelevant noise predictors. These predictors were simulated for each stimulus block as a random vector from a multivariate Gaussian distribution with zero mean and covariance $C = \exp(-0.3|k-i|^2), k = 0, ..., 399, i = 0...399$.

A.3 Model fitting

Separate regression models were fit to data from each voxel. Predictors and the voxel response time courses were standardized. For each voxel, we fit models to data from the 12 training runs and evaluated predictive accuracy on data from test runs. For the EM-banded estimator, we defined distinct λ_j parameters to each of the three feature sets and defined that $\tau = \eta = \kappa = \omega = 10^{-4}$. For the Ridge estimator, we considered a 5-fold cross-validation procedure to inform the selection of α from data from the 12 training runs. A range of possible α values over a grid of 50 logarithmically spaced values between 10^{-5} and 10^{10} were considered. The chosen α parameter was the parameter that minimized the average mean square error between model predictions and target responses in held-out folds. Prediction accuracies on the test set were quantified in two ways. First, we computed

Pearson's correlation coefficient between predicted responses in test runs and the target responses. Second, for each voxel time course, we computed R^2 defined as follows:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}, \tag{24}$$

where $\hat{y}_i$ is the predicted response in block i, where y_i is the target response in block i, and where $\overline{y}$ is the mean response across stimulus blocks.

B EEG decoding analysis: preprocessing, feature extraction and model fitting

We used a publicly available EEG dataset https://doi.org/10.5061/dryad.070jc which includes data from 20 trials from each of the 19 participants [48] (procedures were approved by local ethics committees). The dataset includes a 'broadband' envelope representation of each speech audio stimulus. The first-order temporal difference of the envelope representations was computed and the output was then half-wave rectified [58, 91, 92]. The envelope features were centered and band-pass filtered between 1 Hz and 8 Hz using a second-order Butterworth filter. Zero-phase filtering was achieved by filtering in the forward and reverse directions. For the EEG, the dataset includes data from each experimental trial sampled at 128 Hz. These data were re-referenced to the average of two mastoid channels, detrended using a second-order polynomial, and low pass filtered at 40 Hz using a fourth-order Butterworth filter. For the low-frequency (LF) features, the data were filtered between 1 Hz and 8 Hz using the same filters as those applied to the audio envelope representations. For the higher-frequency (HF) EEG features we filtered the data between 25 Hz and 35 Hz using zero-phase second-order Butterworth filters and then computed the absolute value of the Hilbert transform from each electrode time course. The EEG power time courses were subsequently filtered between 1 Hz and 8 Hz using again the same filters as those applied to the audio envelope representations. Finally, the EEG and audio features were downsampled to 40 Hz to reduce computation time. We augmented EEG features from each channel with a set of time lags by concatenating the time-lagged features side-by-side so as to absorb temporal shifts between stimulus and EEG features [93]. The time lags ranged between -250 ms and 0 ms. We used data from 10-s post-trial onset to approximately 160-s post-trial onset for both EEG and audio features.

Models were fit and tested using a leave-one-trial-out procedure. Prediction accuracies were defined as the Pearson's correlation between model predictions and target envelope in held-out test trials. This procedure was repeated for all trials and test accuracies was averaged across all splits. The correlation coefficients were here averaged across trials after first applying the Fisher-Z transform and the average was inverse-transformed using the inverse of the Fisher-Z transform. Hyperparameter tuning was based on data from the training trials. For each split, we standardized audio and EEG features to their mean and standard deviation estimated from the training data. For the Ridge estimator, the training trials were further split using a 5-fold cross-validation procedure to tune the α parameter. α values between 10^{-5} and 10^{10} in 50 logarithmically spaced steps were considered as in section A. The α parameter was tuned to maximize the correlation between predictions and target variables in held-out inner folds. For the EM-banded model, we assigned separated λ_j parameters to each time lag group from each $1, ..., 128$ electrode for both the LF features and the HF features (i.e., a given group contained multiple time-lagged versions of either a LF feature or a HF feature). Smoothness constraints were imposed on temporal filters for each feature by defining that $h_j = 0.5$ for all $j = 1, ..., 256$ groups with h_j defined according to Eq. 6. We fixed the hyperparameters to $\gamma = 10^{-4}$. Visualization of results made use of *MNE-Python* [94], *matplotlib* [88] and *seaborn* [87].

C Interpretation of weights with correlated predictors

C.1 Simulation 4: Acoustic and phonetic features

Data from the *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus* (TIMIT) [49] were used for this simulation example. Ten sentences each spoken by 190 speakers were used. Audio waveforms were root mean square normalized. Auditory spectrogram features were extracted from each audio waveform using a bank of 32 filters with centre frequencies spaced between 8000 Hz and 50 Hz [89, 95] and then computing the envelope at the output of each filter as the absolute value of the Hilbert transform. The envelopes were power-law compressed ($c = 0.3$), lowpass filtered, and downsampled to 128 Hz. For the phonetic features, we defined boxcar regressors based on time-aligned phonetic labels available from each sentence [49, 96]. Specifically, we constructed six boxcar regressors flagging time periods where the annotation indicated the occurrence of the following categories: stops, affricatives, fricatives, nasals, semivowels and glides, and vowels. We define these boxcar regressors as phonetic features in this simulation. It should be stressed that our goal is not to make statements about the coding of phonetic features, but rather to illustrate properties of regularized estimators when considering such stimulus feature sets. Spectrogram features and phonetic features were extracted from each sentence. Next, we created multiple time-lagged versions of each predictor in these feature sets for the time lags: 0, 1, ..., 38. The first 39 samples of features extracted from each sentence were subsequently discarded. This procedure was repeated for all sentences and the features were stacked column-wise and the predictors were subsequently standardized. This resulted in a total of 661633 samples, corresponding to approximately 86 minutes of data. The feature group F_1 has 1248 columns and the feature group F_2 has 234 columns

C.2 Alternative version of simulation 4

Simulation 4 assumes that the response variable only contains a mixed version of F_1, but not of F_2. For completeness, Figure 10 depicts results of the "reverse" simulation where the response is simulated as $y = F_2 w_2 + \epsilon$, such that only the phonetic features affects the response. Ridge models and EM-banded models are again fit to the data using both feature sets. The simulated "target" weights w_2 are shown in Figure 10A.

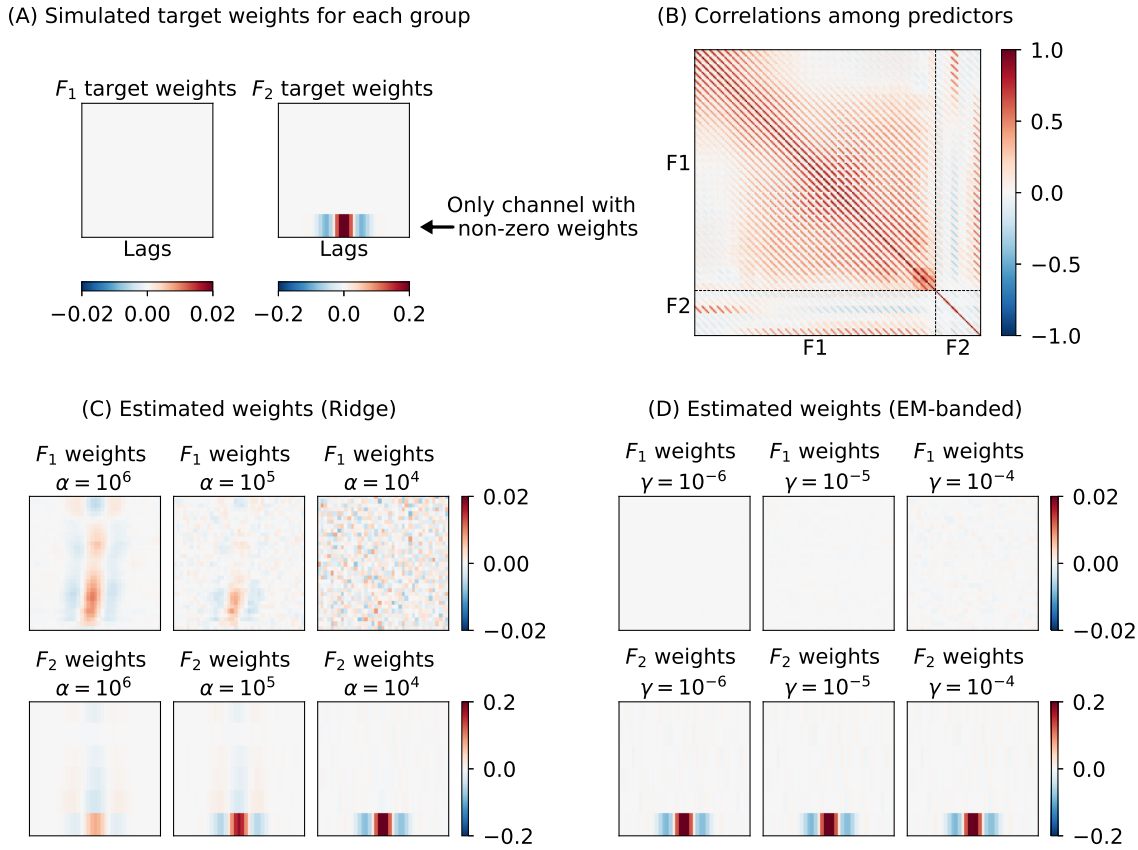

Figure 10: The "reverse" simulation of Simulation 4. (A) Target weights associated with each of the two feature sets. Here, only phonetic features affects the response. (B) Pearson product-moment correlation coefficients among all predictors. Dashed lines are used to visualize predictor groups, F_1 and F_2. (C) Weights estimated with Ridge estimator for each of the two groups. (D) Weights estimated with the EM-banded estimator for each of the two groups. Ridge estimators with α set to 10^6, 10^5 and 10^4 are shown. EM-banded estimators were considered with γ set to 10^{-6}, 10^{-5} and 10^{-4}

C.3 Simulation 6

This simulation extends on Simulation 4. Here, we consider a simulation in which we simulate two groups of predictors, F_1 and F_2. Each of these two feature sets has 128 dimensions and 2048 rows. We assume $y = F_1 w_1 + \epsilon$ and that only feature set F_1 affects the outcome variable, y. The target weights w_1 are shown in 11. As previously, we assume that ϵ is drawn from a Gaussian distribution. Each row in $[F_1, F_2]$ is drawn from a multivariate Gaussian with zero mean and a covariance C defined as follows:

$$C = \begin{bmatrix} I_{128} & 0.3 \cdot I_{128} \\ 0.3 \cdot I_{128} & I_{128} \end{bmatrix}$$

This simulates correlations among predictors associated with the different groups. The SNR is 0 dB. We again compare Ridge regression model fits with EM-banded model fits using a set of fixed

hyperparameters. The estimated weights for the Ridge and EM-banded model are shown in 11. The EM-banded estimator excessively shrinks the weights associated with F_2 (as desired) when γ attains a low value. In all cases, the EM-banded model accurately recovers the weights associated with F_1. The Ridge estimator, on the other hand, tends to pull correlated weights towards each other when α attains a high value (in this case when $\alpha = 1000$ and $\alpha = 10000$). This has the undesired consequence that the weights associated with the stimulus-irrelevant predictors, F_2, tend to resemble the weights associated with F_1 more. This could potentially lead to misinterpretation of the weights associated with F_2.

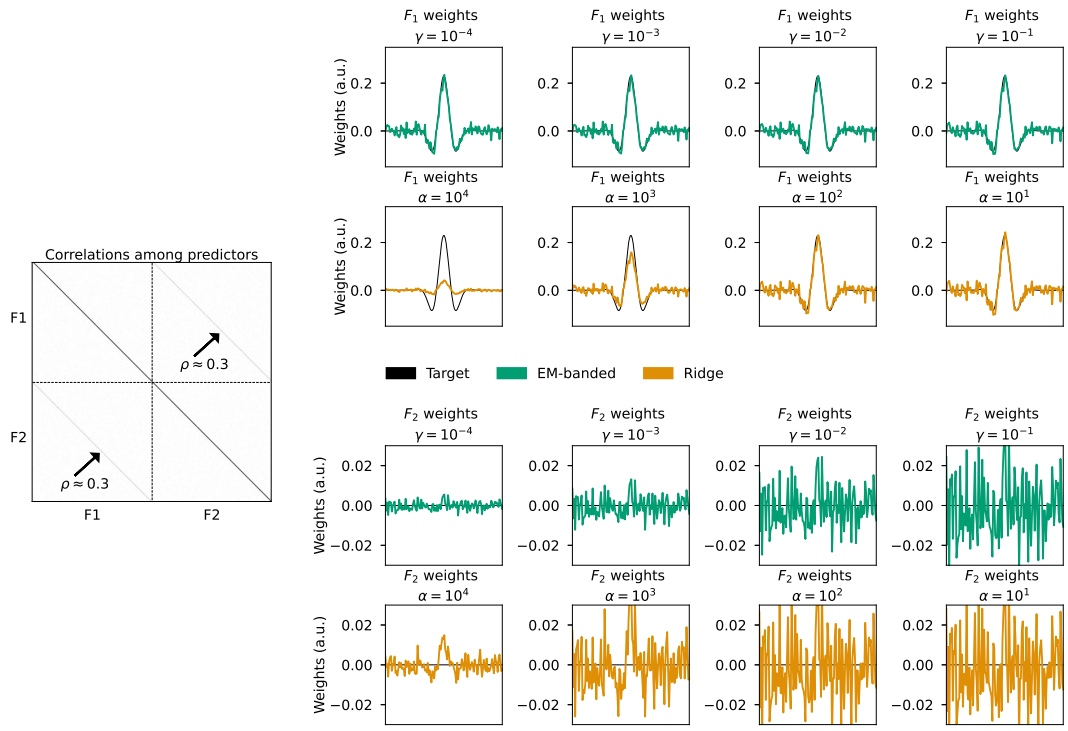

Figure 11: Simulation 6 illustrating behavior of the Ridge estimator and the EM-banded estimators when there are correlations among stimulus-relevant F_1 and stimulus-irrelevant F_2 predictors. Left middle panel: Pearson product-moment correlation coefficients among all predictors. Dashed lines are used to visualize predictor groups, F_1 and F_2. Right: Weights estimated with Ridge estimator (red) and with the EM-banded estimator. The top two rows indicate weights associated with the stimulus-relevant predictors F_1. The bottom two rows indicate weights associated with the stimulus-irrelevant predictors F_2. The target weights are shown in black.

D Simplifying assumptions with many outcome variables

We will now assume a scenario with multiple outcome variables, $p = 1, ..., P$, and a design matrix X that is shared for all outcome variables. Let y_p denote one of a total of P outcome variables and let w_p denote weights associated with outcome variable y_p. We will assume that each outcome variable has been scaled in a way that justifies assuming that ϵ follows a normal distribution with one global variance parameter ν. Consider the following

$$Y = HW + \epsilon, \tag{25}$$

$$Y \equiv \left[y_1^\top, \quad \ldots, \quad y_P^\top\right]^\top \tag{26}$$

$$W \equiv \left[w_1^\top, \quad \ldots, \quad w_P^\top\right]^\top, \tag{27}$$

$$H \equiv I_P \otimes X, \tag{28}$$

where $\otimes$ denotes Kronecker product and where X denote the same design matrix as in the main text. We assume that the weights associated with outcome variable y_p can be divided into J meaningful groups $w_p = \left[\beta_{1,p}^\top, \beta_{2,p}^\top, ..., \beta_{J,p}^\top\right]^\top$ as in the main text, and that separate λ_j hyperparameter are assigned to each of these groups, but that these are shared for all p outcome variables. We assume the following model:

$$p(Y|H, W, \Lambda, \eta, \tau, \phi, \kappa, \nu) = \mathcal{N}(Y|HW, \nu I_{M \cdot P}), \tag{29}$$

$$p(W|\Lambda, \eta, \tau) = \mathcal{N}(W|0, I_P \otimes \Lambda), \tag{30}$$

$$p(\lambda_j|\eta, \tau) = \text{Inv-Gamma}\left(\lambda_j\,|\eta, \tau\right), \quad j = 1, ..., J \tag{31}$$

$$p(\nu|\phi, \kappa) = \text{Inv-Gamma}\left(\nu\,|\phi, \kappa\right), \tag{32}$$

Here, we have:

$$p(W|Y, H, \lambda, \nu) = \mathcal{N}(W|\mu, S), \tag{33}$$

$$S = \left((I_P \otimes \Lambda)^{-1} + \nu^{-1} H^T H\right)^{-1} \tag{34}$$

$$\mu = \nu^{-1} S H^\top Y \tag{35}$$

where the mean μ can be divided into separate blocks associated with each of the P outcome variables such that $\mu \equiv \left[\mu_1^\top, \quad \ldots, \quad \mu_P^\top\right]^\top$. Let $\Sigma \equiv \left(\Lambda^{-1} + \nu^{-1} X^\top X\right)^{-1}$ and let

$$S = I_P \otimes \Sigma \tag{36}$$

$$\mu_p \equiv \nu^{-1} \Sigma X^\top y_p \tag{37}$$

We let $\tilde{\mu_p}$ and $\tilde{\Sigma}$ denote these terms estimated for a given set of λ and ν parameters and we let $\tilde{\Sigma}_{jj}$ denote block in $\tilde{\Sigma}$ associated with predictor group j and similarly let $\tilde{\mu}_{j,p}$ denote a block in $\tilde{\mu}_p$ associated with predictor group j for outcome variable p. We take the same procedure as in the main text for estimating a set of hyperparameters λ_j and ν. This yields the following update rules:

$$\lambda_j^{(k+1)} = \frac{\sum_{p=1}^{P} \left[\tilde{\mu}_{j,p}^\top \Omega_j^{-1} \tilde{\mu}_{j,p} + \text{Tr}(\Omega_j^{-1} \tilde{\Sigma}_{jj})\right] + 2\tau}{P \cdot D_j + 2\eta + 2} \tag{38}$$

$$\nu^{(k+1)} = \frac{\sum_{p=1}^{P} \left[(y_p - X\tilde{\mu_p})^\top (y_p - X\tilde{\mu_p}) + \text{Tr}(X^\top X \tilde{\Sigma})\right] + 2\kappa}{P \cdot M + 2 + 2\phi} \tag{39}$$

Note that it can be useful to pre-compute $X^\top X$ and avoid computing this term multiple times. It can similarly be relevant to make use of the Woodbury matrix identity [97] and define $\left(\Lambda^{-1} + \nu^{-1} X^\top X\right)^{-1} = \Lambda - \Lambda X^\top \left(\nu I_M + X \Lambda X^\top\right)^{-1} X \Lambda$ when $D >> M$.

References

[1] L. S. Hamilton, A. G. Huth, The revolution will not be controlled: natural stimuli in speech neuroscience, Language, cognition and neuroscience 35 (5) (2020) 573–582.

[2] S. O. Dumoulin, B. A. Wandell, Population receptive field estimates in human visual cortex, Neuroimage 39 (2) (2008) 647–660.

[3] K. N. Kay, J. Winawer, A. Rokem, A. Mezer, B. A. Wandell, A two-stage cascade model of BOLD responses in human visual cortex, PLoS computational biology 9 (5) (2013) e1003079.

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

[5] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, E. C. Lalor, Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech, Curr. Biol. 28 (5) (2018) 803—-809.e3. doi:10.1016/j.cub.2018.01.080.

[6] G. M. Di Liberto, J. A. O'Sullivan, E. C. Lalor, Low-frequency cortical entrainment to speech reflects phoneme-level processing, Current Biology 25 (19) (2015) 2457–2465.

[7] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy, Neuron 98 (3) (2018) 630–644.

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

[9] S. Jain, A. Huth, Incorporating context into language encoding models for fmri, Advances in neural information processing systems 31 (2018).

[10] W. F. Massy, Principal components regression in exploratory statistical research, Journal of the American Statistical Association 60 (309) (1965) 234–256.

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

[12] A. N. Tikhonov, V. I. Arsenin, Solutions of Ill-posed Problems: Andrey N. Tikhonov and Vasiliy Y. Arsenin. Translation Editor Fritz John, Wiley, 1977.

[13] G. H. Golub, P. C. Hansen, D. P. O'Leary, Tikhonov regularization and total least squares, SIAM journal on matrix analysis and applications 21 (1) (1999) 185–194.

[14] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse-group lasso, Journal of computational and graphical statistics 22 (2) (2013) 231–245.

[15] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the royal statistical society: series B (statistical methodology) 67 (2) (2005) 301–320.

[16] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[17] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (1) (2005) 91–108.

[18] M. A. Van De Wiel, T. G. Lien, W. Verlaat, W. N. van Wieringen, S. M. Wilting, Better prediction by use of co-data: adaptive group-regularized ridge regression, Statistics in medicine 35 (3) (2016) 368–381.

[19] P. Zhao, G. Rocha, B. Yu, The composite absolute penalties family for grouped and hierarchical variable selection, The Annals of statistics 37 (6A) (2009) 3468–3497.

[20] S. V. David, N. Mesgarani, S. A. Shamma, Estimating sparse spectro-temporal receptive fields with natural stimuli, Network: Computation in neural systems 18 (3) (2007) 191–212.

[21] M. Sahani, J. Linden, Evidence optimization techniques for estimating stimulus-response functions, Advances in neural information processing systems 15 (2002).

[22] M. Park, J. W. Pillow, Receptive field inference with localized priors, PLoS computational biology 7 (10) (2011) e1002219.

[23] A. O. Nunez-Elizalde, A. G. Huth, J. L. Gallant, Voxelwise encoding models with non-spherical multivariate normal priors, Neuroimage 197 (2019) 482–492.

[24] T. D. la Tour, M. Eickenberg, A. O. Nunez-Elizalde, J. L. Gallant, Feature-space selection with banded ridge regression, NeuroImage 264 (2022) 119728.

[25] T. Naselaris, K. N. Kay, S. Nishimoto, J. L. Gallant, Encoding and decoding in fmri, Neuroimage 56 (2) (2011) 400–410.

[26] M. C. Wu, S. V. David, J. L. Gallant, Complete functional characterization of sensory neurons by system identification, Annu. Rev. Neurosci. 29 (2006) 477–505.

[27] N. Ding, J. Z. Simon, Neural coding of continuous speech in auditory cortex during monaural and dichotic listening, Journal of neurophysiology 107 (1) (2012) 78–89.

[28] J. A. O'sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, E. C. Lalor, Attentional selection in a cocktail party environment can be decoded from single-trial eeg, Cerebral cortex 25 (7) (2015) 1697–1706.

[29] G. M. Di Liberto, G. Marion, S. A. Shamma, Accurate decoding of imagined and heard melodies, Frontiers in Neuroscience 15 (2021) 673401.

[30] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, E. Lalor, Decoding the auditory brain with canonical component analysis, Neuroimage 172 (2018) 206–216.

[31] N. Mesgarani, E. F. Chang, Selective cortical representation of attended speaker in multi-talker speech perception, Nature 485 (7397) (2012) 233–236.

[32] B. Mirkovic, S. Debener, M. Jaeger, M. De Vos, Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications, Journal of neural engineering 12 (4) (2015) 46007.

[33] D. J. MacKay, et al., Bayesian nonlinear modeling for the prediction competition, ASHRAE transactions 100 (2) (1994) 1053–1062.

[34] D. Wipf, S. Nagarajan, A new view of automatic relevance determination, Advances in neural information processing systems 20 (2007).

[35] M. E. Tipping, Sparse bayesian learning and the relevance vector machine, Journal of machine learning research 1 (Jun) (2001) 211–244.

[36] C. M. Bishop, Pattern recognition and machine learning, Information science and statistics., Springer, New York, 2006.

[37] T. Nakai, N. Koide-Majima, S. Nishimoto, Music genre neuroimaging dataset, Data in Brief 40 (2022) 107675.

[38] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, F. Bießmann, On the interpretation of weight vectors of linear models in multivariate neuroimaging, Neuroimage 87 (2014) 96–110.

[39] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning series, MIT Press, Cambridge, 2005.

[40] B. Matérn, Spatial Variation: Stochastic Models and Their Application to Some Problems in Forst Survey and Other Sampling Investigations, Esselte, 1960.

[41] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the royal statistical society: series B (methodological) 39 (1) (1977) 1–22.

[42] T. Minka, Expectation-maximization as lower bound maximization 7 (2) (1998).

[43] R. M. Neal, G. E. Hinton, A view of the em algorithm that justifies incremental, sparse, and other variants, Learning in graphical models (1998) 355–368.

[44] D. H. Wolpert, C. E. Strauss, What bayes has to say about the evidence procedure, in: Maximum Entropy and Bayesian Methods: Santa Barbara, California, USA, 1993, Springer, 1996, pp. 61–78.

[45] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, R. S. Frackowiak, Statistical parametric maps in functional imaging: a general linear approach, Human brain mapping 2 (4) (1994) 189–210.

[46] E. C. Lalor, A. J. Power, R. B. Reilly, J. J. Foxe, Resolving precise temporal processing properties of the auditory system using continuous stimuli, Journal of neurophysiology 102 (1) (2009) 349–359.

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

[48] M. P. Broderick, A. J. Anderson, E. C. Lalor, Semantic context enhances the early auditory encoding of natural speech, Journal of Neuroscience 39 (38) (2019) 7564–7575.

[49] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1, NASA STI/Recon technical report n 93 (1993) 27403.

[50] T. Nakai, N. Koide-Majima, S. Nishimoto, "music genre fmri dataset" (2023). doi:doi:10.18112/openneuro.ds003720.v1.0.1.

[51] T. Nakai, N. Koide-Majima, S. Nishimoto, Correspondence of categorical and feature-based representations of music in the human brain, Brain and Behavior 11 (1) (2021) e01936.

[52] G. Tzanetakis, P. Cook, Musical genre classification of audio signals, IEEE Transactions on speech and audio processing 10 (5) (2002) 293–302.

[53] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, E. C. Lalor, Attentional selection in a cocktail party environment can be decoded from single-trial EEG, Cerebral Cortex 25 (7) (2014) 1697–1706.

[54] D. D. E. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, A. de Cheveigné, A comparison of regularization methods in forward and backward models for auditory attention decoding, Frontiers in neuroscience 12 (2018) 531.

[55] M. J. Crosse, G. M. Di Liberto, E. C. Lalor, Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration, J. Neurosci. 36 (38) (2016) 9888–9895.

[56] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, E. F. Chang, Reconstructing speech from human auditory cortex, PLoS biology 10 (1) (2012) e1001251.

[57] N. E. Crone, D. Boatman, B. Gordon, L. Hao, Induced electrocorticographic gamma activity during auditory perception, Clinical neurophysiology 112 (4) (2001) 565–582.

[58] S. R. Synigal, E. S. Teoh, E. C. Lalor, Including measures of high gamma power can improve the decoding of natural speech from eeg, Frontiers in human neuroscience 14 (2020) 130.

[59] A.-L. Giraud, D. Poeppel, Cortical oscillations and speech processing: emerging computational principles and operations, Nature neuroscience 15 (4) (2012) 511–517.

[60] A. de Cheveigné, J. Z. Simon, Denoising based on spatial filtering, Journal of neuroscience methods 171 (2) (2008) 331–339.

[61] A. de Cheveigné, L. C. Parra, Joint decorrelation, a versatile tool for multichannel data analysis, Neuroimage 98 (2014) 487–505.

[62] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural networks 13 (4-5) (2000) 411–430.

[63] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, Journal of computational and graphical statistics 15 (2) (2006) 265–286.

[64] A. de Cheveigné, G. M. Di Liberto, D. Arzounian, D. D. Wong, J. Hjortkjær, S. Fuglsang, L. C. Parra, Multiway canonical correlation analysis of brain data, Neuroimage 186 (2019) 728–740.

[65] A. De Cheveigné, Shared component analysis, Neuroimage 226 (2021) 117614.

[66] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (5) (2008) 849–911.

[67] E. Bair, T. Hastie, D. Paul, R. Tibshirani, Prediction by supervised principal components, Journal of the American Statistical Association 101 (473) (2006) 119–137.

[68] R. M. Neal, J. Zhang, High dimensional classification with bayesian neural networks and dirichlet diffusion trees, Feature extraction: Foundations and applications (2006) 265–296.

[69] I. Guyon, A. Elisseeff, An introduction to feature extraction, Feature extraction: foundations and applications (2006) 1–25.

[70] N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping, Proceedings of the National Academy of Sciences 103 (10) (2006) 3863–3868.

[71] C.-W. Woo, L. J. Chang, M. A. Lindquist, T. D. Wager, Building better biomarkers: brain models in translational neuroimaging, Nature neuroscience 20 (3) (2017) 365–377.

[72] G. Tuckute, J. Feather, D. Boebinger, J. H. McDermott, Many but not all deep neural network audio models capture brain responses and exhibit hierarchical region correspondence, bioRxiv (2022) 2022–09.

[73] D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex, Nature neuroscience 19 (3) (2016) 356.

[74] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex, Proceedings of the national academy of sciences 111 (23) (2014) 8619–8624.

[75] M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proceedings of the National Academy of Sciences 116 (32) (2019) 15849–15854.

[76] D. Kobak, J. Lomond, B. Sanchez, The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization, The Journal of Machine Learning Research 21 (1) (2020) 6863–6878.

[77] T. Hastie, A. Montanari, S. Rosset, R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation, Annals of statistics 50 (2) (2022) 949.

[78] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy, Nature 585 (7825) (2020) 357–362. doi:10.1038/s41586-020-2649-2.
URL https://doi.org/10.1038/s41586-020-2649-2

[79] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272. doi:10.1038/s41592-019-0686-2.

[80] S. K. Lam, A. Pitrou, S. Seibert, Numba: A llvm-based python jit compiler, in: Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, LLVM '15, Association for Computing Machinery, New York, NY, USA, 2015. doi:10.1145/2833157.2833162.
URL https://doi.org/10.1145/2833157.2833162

[81] K. J. Friston, S. Williams, R. Howard, R. S. Frackowiak, R. Turner, Movement-related effects in fmri time-series, Magnetic resonance in medicine 35 (3) (1996) 346–355.

[82] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, Statistical parametric mapping: the analysis of functional brain images, Elsevier, 2011.

[83] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith, Fsl, Neuroimage 62 (2) (2012) 782–790.

[84] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, et al., Advances in functional and structural mr image analysis and implementation as fsl, Neuroimage 23 (2004) S208–S219.

[85] B. B. Avants, N. Tustison, G. Song, et al., Advanced normalization tools (ants), Insight j 2 (365) (2009) 1–35.

[86] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, G. Varoquaux, Machine learning for neuroimaging with scikit-learn, Frontiers in neuroinformatics (2014) 14.

[87] M. L. Waskom, Seaborn: statistical data visualization, Journal of Open Source Software 6 (60) (2021) 3021.

[88] J. D. Hunter, Matplotlib: A 2d graphics environment, Computing in science & engineering 9 (03) (2007) 90–95.

[89] M. Slaney, et al., An efficient implementation of the patterson-holdsworth auditory filter bank, Apple Computer, Perception Group, Tech. Rep 35 (8) (1993).

[90] B. R. Glasberg, B. C. Moore, Derivation of auditory filter shapes from notched-noise data, Hearing research 47 (1-2) (1990) 103–138.

[91] I. Hertrich, S. Dietrich, J. Trouvain, A. Moos, H. Ackermann, Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal, Psychophysiology 49 (3) (2012) 322–334.

[92] C. Daube, R. A. Ince, J. Gross, Simple acoustic features can explain phoneme-based predictions of cortical responses to speech, Current Biology 29 (12) (2019) 1924–1937.

[93] A. De Cheveigné, M. Slaney, S. A. Fuglsang, J. Hjortkjaer, Auditory stimulus-response modeling with a match-mismatch task, Journal of Neural Engineering 18 (4) (2021) 046040.

[94] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, et al., Meg and eeg data analysis with mne-python, Frontiers in neuroscience (2013) 267.

[95] M. Slaney, Auditory toolbox, Interval Research Corporation, Tech. Rep 10 (1998) (1998) 1194.

[96] V. W. Zue, S. Seneff, Transcription and alignment of the timit database, in: Recent Research Towards Advanced Man-Machine Interface Through Spoken Language, Elsevier, 1996, pp. 515–525.

[97] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.