October 22, 2017
Draft version. Intended for future publication in Neuroimage

Post-hoc modification of linear models allows for injection of prior knowledge

Marijn van Vliet[1,2*], Marc M. Van Hulle[2], and Riitta Salmelin[1]

[1]Department of Neuroscience and Biomedical Engineering, Aalto University
[2]Department of Neurosciences, KU Leuven
[*]Corresponding author: w.m.vanvliet@gmail.com

Abstract

In neuroimaging, linear models are the foundation of multivariate analysis methods, in particular linear regression and linear classifiers. We present a framework that allows us to boost the performance of any type of linear model by incorporating prior domain knowledge during the fitting of its parameters. It is based on the observation that a linear model can be disassembled into two components: 1) the identified pattern of activity that co-varies with the target variable and 2) a whitening operator. Both components can be tweaked in isolation and then reassembled. We demonstrate this technique on a linear model designed to decode the associative strength between words from electroencephalography (EEG) reading data. Our results show how the decoding accuracy of the initial linear regression model can be boosted by incorporating information about the spatio-temporal nature of the data, domain knowledge about the N400 evoked potential, data from other participants in the study, data from an earlier EEG study and finally data from an earlier magnetoencephalography (MEG) study. The framework opens up many new approaches to improve a linear model in a neuroimaging setting, which can be used in conjunction with existing general purpose methods such as feature selection, $\ell 1$ and $\ell 2$ regularization.

Keywords: multivariate analysis, linear model, beamformer, event-related potentials, N400, EEG, MEG

1 Introduction

Multivariate analysis is a good thing, because it allows us to increase the signal-to-noise ratio (SNR) of the signal. This allows us to use more intricate experimental designs and answer more detailed research questions.

Mass univariate analysis

- A target variable (stimulus property of behavioral measure) is used to reconstruct every channel (voxel, electrode, squid) and time point.
- General linear model (GLM): $\hat{\mathbf{X}} = \mathbf{A} \cdot \mathbf{y}$.
- Find $\mathbf{A}$ by regressing $\mathbf{y}$ onto each feature (=column) of $\mathbf{X}$.
- Classical neuroimaging experiments often compare the mean signal between two experimental conditions. This can be regarded as a special case of the GLM, where the target variable is: $[-1, 1]$.
- Estimates how $\mathbf{y}$ is encoded in the signal $\mathbf{X}$.
- The coefficients $\mathbf{A}$ are interpretable: the signal at (channel, time point) combi-

nations with a high weight are strongly affected by the target variable.

Multivariate analysis

- Combines information from different (channel, time point) combinations to reconstruct the target variable.
- Makes it possible to cancel out noise and extract brain signals of interest with higher sensitivity and specificity.
- $\widehat{\mathbf{y}} = \mathbf{W} \cdot \mathbf{X}$.
- Use machine learning to find $\mathbf{W}$.
- Estimates how $\mathbf{y}$ can be decoded from the signal $\mathbf{X}$.
- The weights $\mathbf{W}$ are *not* interpretable! The signal at (channel, time point) combinations with a high weight are not necessarily strongly affected by the target variable.

This study

- Challenge: decode forward association strength (FAS) from EEG data
- We expect the FAS affects mostly the N400 component, but let's see what we get.
- We will explore adding domain information about the timing of semantic processes in the brain, transfer information from other subjects, from an old EEG study and from an old MEG study.

2 Methods

2.1 The linear model

Let $\mathbf{X} \in \mathbb{R}^{N \times M}$ be a matrix where each row corresponds to one of N obervations and each column describes one of M features. Generally in neuroimaging studies, $\mathbf{X}$ contains the data recorded from the brain. For example, in our study, an observation is a single trial of EEG data, and each feature a single sample of the voltage recorded at a single electrode. Another example would be that each observation is an functional magnetic resonance imaging (fMRI) image, and each feature the blood-oxygen level dependant (BOLD) signal at a single voxel.

Let $\mathbf{Y} \in \mathbb{R}^{N \times K}$ be a matrix that contains, for each observation, the corresponding values of K target variables. The target variables represent the entities we wish to decode from the brain data. They may for example encode the experimental condition of the trial. In our study, there is only one target variable ($K = 1$): the FAS of the word-pair presented during the trial.

To simplify the mathematics, we are going to assume w.l.o.g. that both $\mathbf{X}$ and $\mathbf{Y}$ are centered, i.e., that $\mathbb{E}[\mathbf{X}]_n = \mathbb{E}[\mathbf{Y}]_n = 0$, where $\mathbb{E}[\cdot]_n$ denotes expectation over observations.

Then, a linear model $\mathbf{W} \in \mathbb{R}^{M \times K}$ is a matrix of weights that can be used to predict $\mathbf{Y}$

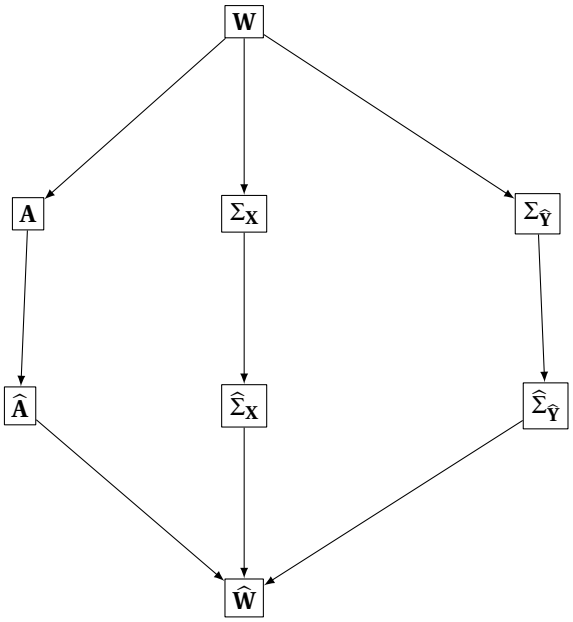

Figure 1: Overview of the "workbench" framework. A linear model $\mathbf{W}$ is disassembled into three components. The individual components can be modified and reassembled into a new model $\hat{\mathbf{W}}$.

from $\mathbf{X}$:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{W}, \tag{1}$$

where $\hat{\mathbf{Y}}$ denotes the model's approximation of $\mathbf{Y}$.

2.2 Filters and patterns

Haufe et al. (2014) showed the relationship between the linear model $\mathbf{W}$ that predicts $\mathbf{Y}$ from $\mathbf{X}$ and a linear model $\mathbf{A}$ that does the opposite and predicts $\mathbf{X}$ from $\mathbf{Y}$:

$$\mathbf{A} = \Sigma_{\mathbf{X}}\mathbf{W}\Sigma_{\hat{\mathbf{Y}}}^{-1}, \tag{2}$$

$$\hat{\mathbf{X}} = \mathbf{Y}\mathbf{A}^{\intercal}. \tag{3}$$

In the above equations, $\Sigma_{\mathbf{X}}$ is the covariance matrix of $\mathbf{X}$ and $\Sigma_{\hat{\mathbf{Y}}}^{-1}$ is the inverse of the covariance matrix of $\hat{\mathbf{Y}}$. Following the nomenclature of source separation, we refer to $\mathbf{W}$ as filter weights and $\mathbf{A} \in \mathbb{R}^{M \times K}$ as activation patterns.

To complete the set of equations, we can solve for $\mathbf{W}$ in equation 2 to compute filter weights given an activation pattern:

$$\mathbf{W} = \Sigma_{\mathbf{X}}^{-1}\mathbf{A}\Sigma_{\hat{\mathbf{Y}}}. \tag{4}$$

2.3 A workbench for linear models

Through equation 2, any linear model $\mathbf{W}$ can be disassembled into $\mathbf{A}$, $\Sigma_{\mathbf{X}}$ and $\Sigma_{\hat{\mathbf{Y}}}$, which can be reassembled into $\mathbf{W}$ through equation 4. However, we can modify $\mathbf{A}$, $\Sigma_{\mathbf{X}}$ and $\Sigma_{\hat{\mathbf{Y}}}$ before reassembly, yielding a new model $\hat{\mathbf{W}}$.

If we interpret the linear model as a filter, then:

1. $\widehat{\mathbf{A}}$ represents the activation patterns, i.e., the part of the signal that the filter is trying to isolate.
2. $\Sigma_{\mathbf{X}}$ represents the dependencies between the signal features
3. $\Sigma_{\widehat{\mathbf{Y}}}$ represents the dependencies between the extracted target variables

To filter the activation patterns from the rest of the signal, the linear model first whitens the input data using $\Sigma_{\mathbf{X}}^{-1}$. Then, with all inter-dependencies eliminated, the desired portion of the signal can be trivially extracted by multiplying with $\mathbf{A}$. However, the result is still whitened, so to faithfully reconstruct the target variables, any dependencies between those variables have to be re-introduced by multiplying with $\Sigma_{\widehat{\mathbf{Y}}}$.

This technique is applicable to any linear model, irregardless of how $\mathbf{W}$ was derived: linear regression, logistic regression, linear support vector machine (lSVM), with or without shrinkage, feature selection, $\ell 1$ or $\ell 2$ norms, etc.

2.4 Examples

Being able to directly modify one of the components of a linear model opens up many avenues for post-hoc enhancement of an already fitted linear model. Some examples follow.

By modifying $\mathbf{A}$, prior knowledge about the activation patterns can be straightforwardly injected into the model. For example, for datasets that include multiple recording sessions, $\mathbf{A}$ can be steered towards the grand average to achieve transfer learning.[1] Furthermore, dependencies between the features, for example temporal and spatial smoothness for EEG and MEG recordings, can be incorporated by smoothing $\mathbf{A}$ in the proper directions.

[1] van Vliet et al., 2016

An $\ell 2$ norm can be imposed post-hoc on the model by shrinking $\Sigma_{\mathbf{X}}$. Furthermore, since $\Sigma_{\mathbf{X}}$ can be estimated without labels, the initial estimation of $\Sigma_{\mathbf{X}}$ on the training set can be improved whenever new (unlabeled) data becomes available. Alternatively, one may take some non-stationarities of the signal into account by estimating $\Sigma_{\mathbf{X}}$ on a sliding window.

A linearly constrained minimum variance (LCMV) beamformer can be created by setting $\Sigma_{\widehat{\mathbf{Y}}} = (\mathbf{A}^{\mathsf{T}}\Sigma_{\mathbf{X}}^{-1}\mathbf{A})^{-1}$, which imposes the constraint that $\widehat{\mathbf{W}}\mathbf{A} = \mathbf{I}$.

3 Results

The following linear models were evaluated on their ability to decode FAS from single-trial EEG data:

ridge Ridge regression, with the optimal shrinkage parameter determined through leave-one-out crossvalidation. Model was trained and applied on each subject individually. I've run many tests and this is the absolute best I could obtain with a general purpose linear model. In the following models, I start applying my own tricks, using the "workbench" framework.

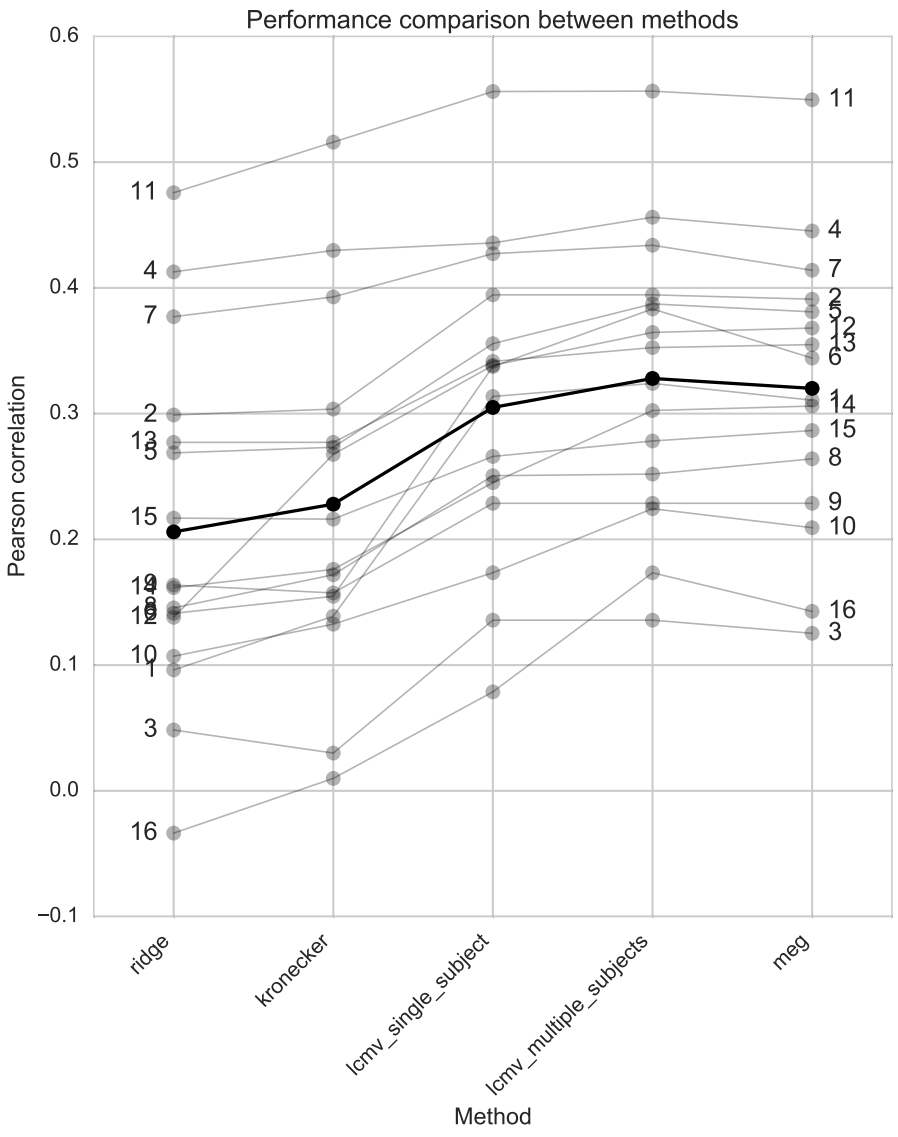

Figure 2: Results of attempting to decode FAS from single-trial EEG data. Performance is measured through Pearson correlation between the true and predicted FAS. Each thin line prepresents a subject and the thick line is the mean across subjects. Along the columns are various linear models: see the main text for a description of each model.

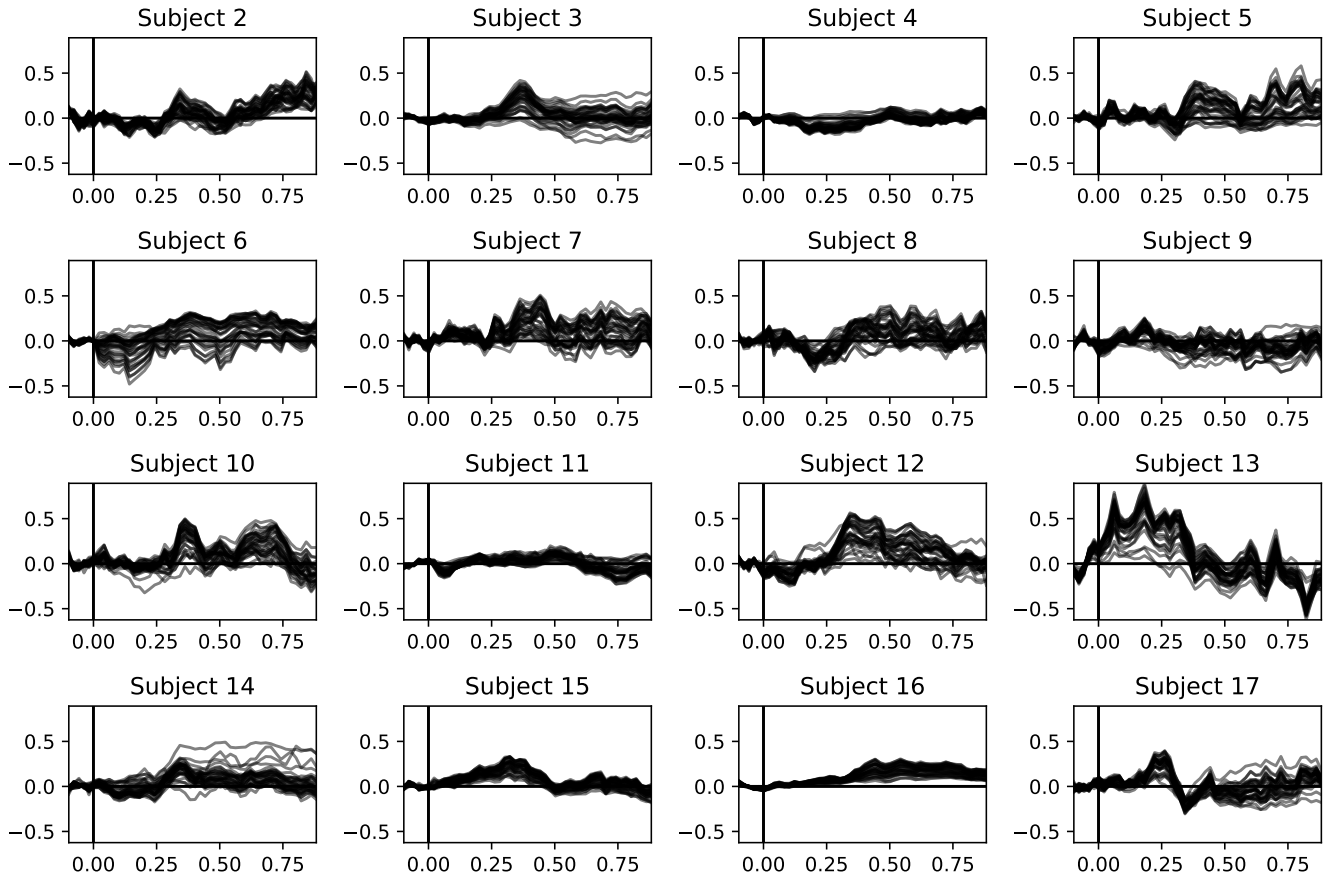

Figure 3: The activation patterns for each subject, estimated by fitting a ridge regressor to the data and applying equation 2 to obtain **A**. In constrast to the regression weights **W**, these patterns are interpretable. The patterns are plotted as "butterly" plots: each line is a single EEG channel. In this case, the N400 potential is clearly visible for some subjects (e.g., 3, 5, and 7). We expect the performance of a ridge regressor to be good for such subjects. NOTE: in this version of the manuscript, the subjects numbers of this figure and figure 2 do not match!! For others, the patterns are clearly very noisy (e.g. 6, 13 and 14) and we expect poor performance of a ridge regressor. However, we can boost the performance for the latter subjects by transplanting the grand-average patterns into their corresponding linear models!

kronecker A version of ridge regression where different shrinkage parameters are used for the spatial and temporal sub-matrices of the covariance matrix. Optimal values were determined through leave-one-out crossvalidation. I call this "Kronecker" shrinkage.

lcmv_single_subject The "workbench" framework was applied to the "kronecker" model to inject prior information about the timing of the N400. The activation patterns **A** were multiplied with a Gaussian kernel centered around 400 ms. The width of the kernel was determined through leave-one-out crossvalidation.

lcmv_multiple_subjects Like the previous model, only $\widehat{\mathbf{A}}$ is now a weighted average of **A** and the grand average $\mathbf{A}_{\text{GA}}$ across all other subjects. The weighting of **A** and $\mathbf{A}_{\text{GA}}$ is determined through leave-one-out crossvalidation.

meg Getting a little crazy now. $\widehat{\mathbf{A}}$ is determined through an earlier MEG study on semantic priming. The idea is that a grand-average MEG scan may yield a more precise "template" of the N400 than a grand-average EEG scan. Not sure if this is actually the case.

References

Haufe, S., Meinecke, F. C., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. doi:10.1016/j.neuroimage.2013.10.067

van Vliet, M., Chumerin, N., De Deyne, S., Wiersema, J. R., Fias, W., Storms, G., & Van Hulle, M. M. (2016). Single-trial ERP component analysis using a spatiotemporal LCMV beamformer. *IEEE Transactions on Biomedical Engineering*, *63*(1), 55–66. doi:10.1109/TBME.2015.2468588