On the Interpretability of Linear Multivariate Neuroimaging Analyses: Filters, Patterns and their Relationship

Felix Bießmann, Sven Dähne, Frank C. Meinecke, Benjamin Blankertz, Kai Görgen, Klaus-Robert Müller, and Stefan Haufe

Machine Learning Group, TU Berlin, Franklinstr 28/29, 10587 Berlin, Germany

Abstract. Multivariate linear methods are an important tool for the analysis of neuroimaging data. However results obtained from multivariate methods are not as easy to interpret as results from mass-univariate analyses. A common misconception is that multivariate filters, that transform measured signals into neural sources of interest, can be interpreted as activation patterns, which reflect underlying neural processes. Yet filters and patterns are not the same. Here we try to create some awareness for the difference between multivariate filters and activation patterns. We show the difference in simulations and real data examples and recapitulate a simple but efficient way of transforming one into the other.

1 Introduction

For many years mass-univariate methods [6] have been the most popular analysis method for multivariate neuroimaging data. With increasing spatial and temporal resolution of measurement devices many researchers started to use multivariate analysis methods: Support-vector machines (SVMs) [3], linear discriminant analysis (LDA) [5] and other classification methods are used to detect brain activity associated with a particular stimulus condition. Principal Component Analysis (PCA) [14], independent component analysis (ICA) [8] and other latent variable models are used to analyze neuroimaging data in an explorative fashion. There are two main reasons for the success of multivariate methods: For one, multivariate methods can detect certain neural activation patterns that univariate methods cannot detect [11]. Second, multivariate methods are more sensitive than univariate methods, see e.g. [10]. Linear\footnote{In many applications, non-linear multivariate analyses do not yield not better decoding accuracy [12], hence we will consider only linear multivariate methods here.} multivariate methods have enabled researchers to address novel research questions such as decoding of mental states using functional magnetic resonance imaging (fMRI)[9, 7]: when using fast neuroimaging devices such as the electroencephalogram (EEG) this mind-reading can be done in fractions of a second [1]. When scanning the literature and participating in discussions one often can encounter a common misconception about the results of multivariate methods. Many researchers interpret
multivariate linear filters with respect to physiological processes. This is inappropriate in most cases, which is widely known in the EEG literature [13, 2]. However in the fMRI community this knowledge seems somewhat underrepresented. Here we recapitulate why physiological interpretations of multivariate analyses can be difficult and show examples on simulated and real neuroimaging data. We also show a simple yet efficient way of transforming the results of multivariate neuroimaging analyses into a meaningful and easily interpretable representation.

2 Relationship between Filters and Patterns

In the following we will denote neuroimaging measurements as \( x(t) \in \mathbb{R}^{U \times 1} \) and neural sources as \( s(t) \in \mathbb{R}^{V \times 1} \); a given sample of measurements will be stored in a data matrix \( \{ x(t = 1), \ldots, x(t = T) \} = X \in \mathbb{R}^{U \times T} \) and \( S \in \mathbb{R}^{V \times T} \), respectively. There are two perspectives when analyzing multivariate neuroimaging data. In generative modeling, neural processes and their mappings to sensors are modeled explicitly using prior assumptions about both, while in decoding one is only interested in optimal recovery of neural sources of interest, making only implicit assumptions about the signal generation.

**Generative Modeling Perspective** Linear multivariate methods assume that the observed data is a noisy linear superpositions of the source signals. This corresponds to the generative model

\[
x(t) = As(t) + n(t),
\]

where the columns of \( A \in \mathbb{R}^{U \times V} \) are what we will refer to henceforth as patterns of neural activation. The vector \( n(t) \in \mathbb{R}^{U \times 1} \) is the noise, which has in general a non-vanishing covariance structure. In the simplest case of a univariate signal of interest, \( A \in \mathbb{R}^{U \times 1} \) would contain only one column, corresponding to the activation pattern of the corresponding source. Given a source signal \( S \) (the neural response to an external trigger, e.g. a stimulus time course) the pattern \( \hat{A} \) mapping \( S \) optimally (in the least-squares sense) to the sensors can be obtained as

\[
\hat{A} = \arg\min_A \sum_t (x(t) - As(t))^2 = XST(SST)^{-1}.
\]

Note that in practice \( S \) is unknown, and is replaced by the external trigger variable.

**Decoding Perspective** Using an optimal (with respect to some property of the sources of interest) multivariate filter \( W \in \mathbb{R}^{U \times V} \) (the \( W_{ij} \)th entry is the weighting of the \( i \)th feature to obtain the \( j \)th neural source) we obtain an estimate of the neural sources \( \hat{s}(t) \in \mathbb{R}^{V \times 1} \) as

\[
\hat{s}(t) = W^T x(t).
\]
Note that the source estimate \( \hat{s}(t) \) contains both signal and noise, i.e. \( \hat{s}(t) = s(t) + W^\top n(t) \). In analogy to optimal pattern estimates in eq. 3 we can obtain estimates of the filters \( W \) as
\[
\hat{W}^\top = \arg\min_{W^\top} \sum_t (s(t) - W^\top x(t))^2
\]
\[
= SX^\top (XX^\top)^{-1}
\] (5)

(6)

If the covariance matrices \( SS^\top, XX^\top \) are ill conditioned, one often adds a ridge \( SS^\top + I \kappa_s, XX^\top + I \kappa_x \), where \( \kappa_s, \kappa_x \) are regularization coefficients [15]. The larger \( \kappa_s, \kappa_x \), the more decorrelated the rows of \( S \) and \( X \) appear to the OLS estimator (see also eqs. 8 and 9 and underlying assumptions) and the smaller will be the \( L_2 \) norm of the solution. We emphasize that all derivations hold equally true for other constraints, such as \( L_1 \) norms which lead to sparse solutions (such as Elastic Net, see section 3).

Common Misconceptions About Multivariate Filters Some common interpretations of the results of multivariate analyses and eq. 4 (i.e. the filters) are

- Weights \( W_{ij} \) show from which voxels neural process \( j \) originates.
- Weights \( W_{ij} \) show which voxels are important for neural process \( j \).
- \( W_{ij} \) is proportional to the importance of a voxel for neural process \( j \).
- The sign of \( W_{ij} \) indicates positive or negative correlation between voxel activity and stimulus/task.

All these interpretations are unjustified. Indeed, if such interpretations are desired, one must refer to the activation patterns in \( A \), not to the filter weights in \( W \). In the following we show that under mild assumptions \( A \) can easily be obtained from \( W \).

2.1 Transforming Filter into Patterns

To obtain the relationship between filters and patterns, let us right multiply eq. 6 with \( XX^\top \). We obtain \( W^\top XX^\top = SX^\top \) and thus \( XS^\top = XX^\top W \). Plugging this into eq. 3 we see that
\[
A = XX^\top W (SS^\top)^{-1},
\] (7)

where \( SS^\top \) is the covariance matrix of the neural sources. Often, it is quite reasonable and well-justified to assume the latent source signals to be uncorrelated (at least to a good approximation). If we do so, we can absorb the scaling (i.e. the variances) of the sources into the filter weights and obtain
\[
A = XX^\top W.
\] (8)

This means, that given uncorrelated source signals, we can obtain the physiologically interpretable activation patterns \( A \) from the filters \( W \) by simply left-multiplying with the empirical data covariance matrix \( C = XX^\top \). Eq. 8 also
states that each pattern is essentially the scalar product of each sensor time series with the respective estimated source signal. If the $i$th sensor (corresponding to the $i$th row of $A$) has a high correlation with the $j$th source (corresponding to the $j$th row of $S = W^\top X$) the entry $A_{ij}$ will be assigned a high value. Only if we assume that also the measurement channels are uncorrelated, we obtain

$$A = W.$$  \hspace{1cm} (9)

However, uncorrelatedness in the input space is a much less natural assumption than uncorrelatedness of source signals. For instance, different voxels in fMRI data are usually highly correlated, rendering eq. 9 false.

3 Examples on simulated data

In the following we demonstrate that the distinction between filters and patterns is important in all multivariate linear analyses, including classification, regression (also with sparsifying norms) and unsupervised methods such as ICA. We created toy data according to eq. 1. The signal pattern $A_s \in \mathbb{R}^{100 \times 1}$ is depicted in the top panel in the leftmost column, the bottom panel shows a noise pattern $A_n \in \mathbb{R}^{100 \times 1}$. Noise was generated as

$$\mathbf{n}(t) = A_n n(t) + A_\epsilon \epsilon(t),$$  \hspace{1cm} (10)

where the noise component $n(t) \in \mathbb{R}^1$ (corresponding to the pattern $A_n$) and the noise $\epsilon(t) \in \mathbb{R}^{100 \times 1}$ (corresponding to the random orthonormal basis $A_\epsilon \in \mathbb{R}^{100 \times 100}$, not plotted), were both drawn from a Gaussian distribution $\mathcal{N}(0,1)$ with zero mean and unit variance. We generated a classification setting and an unsupervised analysis setting. For the classification scenario signals $s(t) \in \mathbb{R}^1$ were drawn from $\mathcal{N}(-1,1)$ (negative class) and $\mathcal{N}(0,1)$ (positive class). For the unsupervised scenario the signal was a sine wave. In both scenarios we have $T = 5000$. All filters, independent of the method that was used to obtain them (i.e. ordinary least squares (OLS, equivalent to LDA), Elastic Net, or ICA), show a superposition of both noise and signal pattern. When applying eq. 8 and projecting the filter matrix $W$ onto the data covariance matrix we obtain the true pattern $A_s$.

4 Examples on experimental data

The effects illustrated in the simulations are very prominent in real data, too: Filters always try to suppress the noise and enhance the signal, hence they are a superposition of noise and signal patterns. Figure 2 shows an example of a spatial filter optimized for a classical P300 speller experiment [4]; a subject has to concentrate on a target stimulus and count the number of target presentations. The electrical potentials associated with targets can be extracted using an LDA classification between EEG features (in a $U$-dimensional feature space, where $U$
is the number of EEG sensors) after target presentation and non-target presentation. Figure 2 shows results extracted from the $N_2$ response, a scalp potential at $\approx 200$ms after stimulus presentation. The filter $W \in \mathbb{R}^{U \times 1}$ is depicted in the left panel and the corresponding pattern $A_s \in \mathbb{R}^{U \times 1}$ is depicted in the right panel. The filter is a rather irregular superposition of signal and noise pattern, while the pattern $A_s$ clearly reflects occipital and central dipolar sources.

5 Conclusion

Multivariate methods are powerful tools and increasingly used in neuroimaging studies. When interpreting results obtained with multivariate methods, classification, regression or latent variable models such as ICA, it is important to keep in mind that filters (the results of multivariate analyses) are not the same as the patterns that reflect the underlying neural sources. In many cases researchers are somewhat aware of these problems and only report correlations between stimulus and brain activity or classification accuracies. But there are few studies that refrain from interpreting the filters in some way. When doing so, they implicitly assume that the measurements are uncorrelated (see eq. 9), which is almost never the case in real neuroimaging data. Here we showed that the distinction between filters and patterns is highly relevant in all multivariate neuroimaging analyses. We also re-stated a simple and effective means of transforming one into the other. We firmly believe that this distinction is crucial for meaningful interpretations of multivariate analysis results and hope that this manuscript helps to make researchers aware of this issue.

Fig. 1. Filters and patterns on simulated data; data was generated according to eq. 1 with true activation pattern $A_s$ (top left) and noise activation pattern $A_n$ (bottom left); we plot filters (bottom row) and patterns (top row) obtained by ordinary least-squares solution (OLS), elastic net (EL) and ICA. All methods estimate a filter which reflects both a noise-suppressing part and the true activation pattern; estimates of the true patterns $A_s$ were obtained using eq. 8.
Fig. 2. *Left:* Filter \( W \) for optimal classification of \( N2 \) response into target/non-target presentation does not show clear structure. *Right:* Corresponding pattern \( A_s \) shows clearly occipital and central neural dipoles.

**References**