Discriminating the Influence of Correlated Factors from Multivariate Observations: the Back-to-Back Regression

Jean-Rémi King[a,b,*], François Charton[b], David Lopez-Paz[b], Maxime Oquab[b]

[a] *Laboratoire des systèmes perceptifs, PSL University, CNRS*
[b] *Facebook AI*

Abstract

Identifying causes solely from observations can be particularly challenging when i) potential factors are difficult to manipulate independently and ii) observations are multi-dimensional. To address this issue, we introduce “Back-to-Back” regression (B2B), a linear method designed to efficiently estimate, from a set of correlated factors, those that most plausibly account for multidimensional observations. First, we prove the consistency of B2B, its links to other linear approaches, and show how it can provide a robust, unbiased and interpretable scalar estimate for each factor. Second, we use a variety of simulated data to show that B2B outperforms forward modeling (”encoding”), backward modeling (”decoding”) as well as cross-decomposition modeling (i.e.. canonical correlation analysis and partial least squares) on causal identification when the factors and the observations are not orthogonal. Finally, we apply B2B to a hundred magneto-encephalography recordings and to a hundred functional Magnetic Resonance Imaging recordings acquired during a one hour reading task. B2B successfully disentangles the respective contribution of collinear factors such as word length, word frequency in the early visual and late associative cortical responses respectively. These factors are better disentangled with B2B than with other standard techniques. We discuss how the speed and generality of B2B sets the foundations of an omnibus test for disentanglement of covarying factors in multidimensional observations.

Keywords: Feature Discovery, MEG, fMRI, Decoding, Encoding, Cross-Decomposition, Reading,

1. Introduction

Natural sciences are tasked to find, from a set of hypothetical factors, the minimal subset that suffices to reliably predict novel observations. This endeavor is impeded by two major challenges. First, causal and non-causal factors may be numerous and partially correlated. In neuroscience, for example, it can be challenging to identify whether word frequency modulates brain activity during reading. Indeed, the frequency of words in natural language covaries with other factors such as their length (short words are more frequent than long words) and their categories (determinants are more frequent than adverbs) [30, 38]. Instead of selecting a set of words that controls for all of these factors simultaneously, it is thus common to use a *forward* ”encoding model”, i.e. to fit a linear

[*] corresponding author: jeanremi@fb.com

regression to predict observations (e.g. brain activity) from a minimal combination of competing factors (e.g. word length, word frequency), and analytically investigate the estimated contribution of each factor from the model's coefficients [13, 35, 48, 27, 24].

The second challenge for measuring causal influence is that observations can be multidimensional. For example, brain activity is often recorded with hundreds or thousands of simultaneous measurements via functional Magnetic Resonance Imaging, magneto-encephalography (MEG) or multiple electro- physiological probes [13, 45]. The relationship between putative causes and observations is thus often done by training models in a *backward* fashion: i.e. from observations to putative causal factors. For example, it is common to fit a support vector machine across multiple brain voxels or multiple electrodes to detect the category of a stimulus [36, 5, 29, 27]. Decoding has become particularly popular in neuroscience, because brain recordings are typically corrupted by major physiological noise, such as muscle movements, eye blinks, displacements etc. As these noises sources often lie along specific components of the multidimensional recordings, the remaining neural signals can be robustly picked up by multiple decoders [27].

Both *forward* and *backward* modeling have competing benefits and drawbacks. Specifically, forward modeling disentangles the independent contribution of correlated factors but does not efficiently combine high-dimensional observations. By contrast, backward modeling combines multiple observations but does not disentangle factors that are linearly correlated [48, 20, 27]. To combine some of the benefits of forward and backward modeling, several authors have proposed to use cross-decomposition techniques such as Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA) [8, 3]. CCA and PLS aim to find, from two sets of data X and Y, the matrices H and G such that XH and YG are maximally correlated or maximally covarying respectively.

While CCA and PLS can make use of multidimensional features and observations, they are not designed for feature discovery. First, these methods are not not directional: observations and factors can be assigned to either X or Y. Second, these methods project X and Y onto a reduced but nonetheless multidimensional space. Third, because CCA and PLS are based on a generalized eigen decomposition, their resulting coefficients mix the features of X and Y in a way that makes them notoriously difficult to interpret [31].

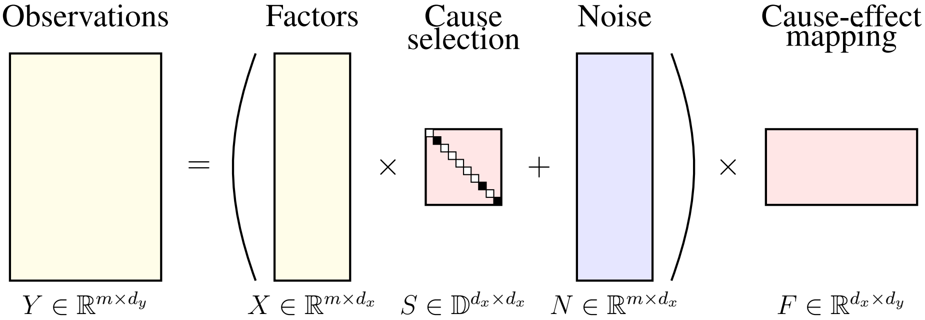
Here, we introduce the 'back-to-back regression' (B2B), which not only combines the benefits of forward and backward modeling (Section 2) but can also provide robust, interpretable, unidimensional and unbiased coefficients for each factor.

After detailing B2B and proving its convergence (Section 2.3), we show with synthetic data that it outperforms state-of-the-art forward, backward and cross-decomposition techniques in disentangling causal factors (Section 3.1). Finally, we apply B2B to large MEG and fMRI datasets acquired during a simple reading task and reveal that B2B efficiently distinguishes the respective effects of covarying word features (Section 3.3).

2. Back-to-Back regression

2.1. Problem setup

We consider the measurement of multivariate signal $Y \in \mathbb{R}^{m \times d_y}$ (the dependent variables, e.g. the neural responses), generated from a set of putative causes $X \in \mathbb{R}^{m \times d_x}$ (the independent variables, e.g. the features of a stimulus), via some unknown linear apparatus $F \in \mathbb{R}^{d_x \times d_y}$ (e.g.


Observations
Factors
Cause selection
Noise
Cause-effect mapping
=
×
+
×
$Y \in \mathbb{R}^{m \times d_y}$
$X \in \mathbb{R}^{m \times d_x}$
$S \in \mathbb{D}^{d_x \times d_x}$
$N \in \mathbb{R}^{m \times d_x}$
$F \in \mathbb{R}^{d_x \times d_y}$

$$X \sim P(X) \quad N \sim P(N) \quad \hat{S} = \mathrm{diag}(\underbrace{(X_2^\top X_2 + \Lambda_X)^{-1} X_2^\top Y_2 \overbrace{(Y_1^\top Y_1 + \Lambda_Y)^{-1} Y_1^\top X_1}^{1)\ \hat{X}\text{: regression from } Y \text{ to } X}}_{2)\ \text{regression from } X \text{ to } \hat{X}})$$

Figure 1: Back-to-back regression identifies the subset of factors $S_{ii} = 1$ in X that influence m multidimensional observations Y by 1) regressing from Y to X to obtain $\hat{X}$, and 2) returning the diagonal of the regression coefficients from X to $\hat{X}$.

the projection of neural activity onto MEG channels or fMRI voxels). Not all the variables in X exert a causal influence on Y. By considering a square binary diagonal matrix of *causal influences* $S \in \mathbb{D}^{d_x \times d_x}$, we denote by XS the causal factors of Y. In summary, the problem can be formalized as:

$$y_i = (x_i S + n_i) F \tag{1}$$

where i is a given sample, and n_i is a sample-specific additive noise drawn from a centered distribution. While X and N are independent, we allow each of them to have any form of covariance. In practice, we observe m samples (X, Y) from the model. This problem space, along with the sizes of all variables involved, is illustrated in Fig. 1. Given the model in Equation eq. (1), the goal of Back-to-Back Regression (B2B) is to estimate the matrix S, i.e. to identify the factors that reliably account for the multivariate observations.

2.2. Algorithm

Back-to-Back Regression (B2B) consists of two steps. First, we estimate the linear regression coefficients $\hat{G}$ from Y to X, and construct the predictions $\hat{X} = Y\hat{G}$. This backward regression recovers the correlations between Y and each factor of X. Second, we estimate the linear regression coefficients $\hat{H}$ from X to $\hat{X}$. The diagonal of the regression coefficients $\hat{H}$, denoted by $\hat{S} = \mathrm{diag}(\hat{H})$, is the desired estimate of the causal influence matrix S, as detailed in the Appendix A.1.

If using l2-regularized least-squares [21, 41], B2B has a closed form solution:

$$\hat{G} = (Y^\top Y + \Lambda_Y)^{-1} Y^\top X, \tag{2}$$

$$\hat{H} = (X^\top X + \Lambda_X)^{-1} X^\top Y \hat{G}, \tag{3}$$

where Λ_X and Λ_Y are two diagonal matrices of regularization parameters, useful to invert the covariance matrices of X and Y if these are ill-conditioned.

Performing two regressions over the same data sample can result in overfitting, as spurious correlations in the data absorbed by the first regression will be leveraged by the second one. To avoid this issue, we split our sample (X, Y) into two disjoint sets (X_1, Y_1) and (X_2, Y_2). The first regression is performed using (X_1, Y_1), and the second regression is performed using (X_2, Y_2). To compensate for the reduction in sample size caused by the split, the two successive regressions are repeated over many random splits, and the final estimate $\hat{S}$ of the causal influence matrix is the average over the estimates associated with each split [4]. To accelerate this ensembling procedure, we use an efficient leave-one-out (LOO) cross-validation scheme as detailed in [41] as follows:

$$\hat{Y}_{LOO} = (\Sigma_X G Y - \text{diag}(\Sigma_X G) Y) \; / \; \text{diag}(I - \Sigma_X G) \qquad \text{(element-wise division)} \tag{4}$$

where Σ_X is the X kernel matrix and where G is computed with an eigen decomposition of X:

$$\begin{aligned} \Sigma_X &= Q V Q^T \\ G &= Q(V + \lambda I)^{-1} Q^T \end{aligned} \tag{5}$$

where Q, V and λ are the eigen vectors, eigen values and regularization, respectively.

We summarize the B2B procedure in Algorithm 1. The rest of this section provides a theoretical guarantee on the correctness of B2B.

Algorithm 1: Back-to-back regression.

```
Input: input data $X \in \mathbb{R}^{m \times d_x}$, output data $Y \in \mathbb{R}^{m \times d_y}$, number of repetitions $m \in \mathbb{N}$.
Output: estimate of causal influences $\hat{S} \in \mathbb{D}^{d_x \times d_x}$.
$\hat{S} \leftarrow 0$;
for $i = 1, \ldots, m$ do
    $(X, Y) \leftarrow$ ShuffleRows($(X, Y)$);
    $(X_1, Y_1), (X_2, Y_2) \leftarrow$ SplitRowsInHalf($(X, Y)$);
    $\hat{G} =$ LinearRegression($Y_1, X_1$) ;                 ▷ $\hat{G} = (Y_1^\top Y_1 + \Lambda_Y)^{-1} Y_1^\top X_1$
    $\hat{H} =$ LinearRegression($X_2, Y_2 \hat{G}$) ;          ▷ $\hat{H} = (X_2^\top X_2 + \Lambda_X)^{-1} X_2^\top Y_2 \hat{G}$
    $\hat{S} \leftarrow \hat{S} + \text{diag}(\hat{H})$;
end
$\hat{S} \leftarrow \hat{S}/m$;
$\hat{W} \leftarrow$ LinearRegression($X\hat{S}, Y$);
return $\hat{S}$, $\hat{W}$
```

2.3. Theoretical guarantees

Theorem 1 (B2B consistency - general case)**.** *Consider the B2B model from Equation* $Y = (XS + N)F$, *N centered and full rank noise. Let $Img(M)$ refers to the image of the matrix M. If F and X are full-rank on the $Img(S)$, then, the solution of B2B, $\hat{H}$, will minimize* $\min_H \|X - XH\|^2 + \|NH\|^2$*and satisfy $S\hat{H} = \hat{H}$*

Proof. See Appendix Appendix A.1. □

Since $S\hat{H} = \hat{H}$, we have

$$\hat{H} = \arg\min_{H} \|X - XSH\|^2 + \|NSH\|^2 = (SX^\top XS + SN^\top NS)^\dagger SXX^\top. \tag{6}$$

Assuming, without loss of generality, that the active features in S are the $k \in \mathbb{Z} : k \in [0, d_x]$ first features, and rewriting $X = (X_1, X_2)$ and $N = (N_1, N_2)$ (X_1 and N_1 containing the k first features), we have:

$$X^\top X = \begin{pmatrix} \Sigma_{X_1X_1} & \Sigma_{X_1X_2} \\ \Sigma_{X_1X_2} & \Sigma_{X_2X_2} \end{pmatrix}, \qquad N^\top N = \begin{pmatrix} \Sigma_{N_1N_1} & \Sigma_{N_1N_2} \\ \Sigma_{N_1N_2} & \Sigma_{N_2N_2} \end{pmatrix}, \tag{7}$$

where Σ_{AB} is the covariance of A and B, and:

$$\hat{H} = \begin{pmatrix} (\Sigma_{X_1X_1} + \Sigma_{N_1N_1})^{-1}\Sigma_{X_1X_1} & (\Sigma_{X_1X_1} + \Sigma_{N_1N_1})^{-1}\Sigma_{X_1X_2} \\ 0 & 0 \end{pmatrix} \tag{8}$$

$$\text{diag}_k(\hat{H}) = \text{diag}((\Sigma_{X_1X_1} + \Sigma_{N_1N_1})^{-1}\Sigma_{X_1X_1}) = \text{diag}((I + \Sigma_{X_1X_1}^{-1}\Sigma_{N_1N_1})^{-1}) \tag{9}$$

In the absence of noise, we have $\Sigma_{N_1N_1} = 0$, and so $\text{diag}_k(\hat{H}) = I$, and

$$\text{diag}(\hat{H}) = \text{diag}(S)$$

Therefore, we recover S from $\hat{H}$.

In the presence of additive noise, the causal factors of S correspond to the positive elements of $\text{diag}(\hat{H})$. The methods to recover them are presented in the Appendix (Appendix A.4).

Note that hat§ is unbiased, in the sense that it is centered around zero when there is no effect, only if the second regression H is not regularized. Second-level statistics testing whether hat§ is superior to 0 are thus only valid if H is not regularized.

3. Experiments

We perform three sets of experiments to evaluate B2B: one on controlled synthetic data, a second one on a real, large-scale functional Magnetic Resonance Imaging (fMRI) dataset and a third one on a real, large-scale magneto-encephalography (MEG) dataset. We use scikit-learn's PLS and RidgeCV [37] as well as Pyrcca's regularized canonical component analysis (RegCCA, [3]) objects to compare B2B against the standard baselines, with common hyper-parameter optimizations.

3.1. Synthetic Experiment

We evaluate the performance of B2B throughout a series of experiments on controlled synthetic data. The purpose of these experiments is to evaluate the ability of B2B on its ability to 1) recover causal factors when the ground truth is known and 2) accurately predict independent and identically distributed data otherwise.

The data generating process for each experiment constructs $m = 1000$ training examples according to the model $Y = (\text{h}XS + N)F$, where h is a scalar that modulates the signal-to-noise ratio. Here, $F \in \mathbb{R}^{d_x \times d_y}$ contains entries drawn from $\mathcal{N}(0, \sigma^2)$ where σ^2 is inversely proportional to d_x, $X \in \mathbb{R}^{m \times d_x}$ contains rows drawn from $\mathcal{N}(0, \Sigma_X)$, $N \in \mathbb{R}^{m \times d_x}$ contains rows drawn from $\mathcal{N}(0, \Sigma_N)$, $S \in \mathbb{R}^{d_x \times d_x}$ is a binary diagonal matrix containing n_c ones, $\Sigma_X = AA^\top$ where $A \in \mathbb{R}^{d_x \times d_x}$ contains entries drawn from $\mathcal{N}(0, \sigma^2)$, $\Sigma_N = BB^\top$ where $B \in \mathbb{R}^{d_x \times d_x}$ contains entries drawn from $\mathcal{N}(0, \sigma^2)$, and the factor h $\in \mathcal{R}_+$.

To simulate a wide range of experimental conditions, we sample 10 values in log-space for $d_x, d_y \in [10, 100]$, $n_c \in [3, 63]$, h $\in [0.001, 10]$. We discard the cases where $n_c > d_x$, limit d_x, d_y to 100 to keep the running time under 2 hours for each condition, and average over 5 random seeds.

We compare the performance of B2B against four baseline methods.

3.1.1. Baseline models

All baseline methods were based on the implementations of scikit-learn [37] and Pyrcca [3]. For pedagogical purposes, we briefly summarize them below.

Forward regression consists of an $l2$-regularized "ridge" regression from the putative causes X to the observations Y:

$$H_{fwd} = (X^T X + \lambda I)^{-1} X^T Y \tag{10}$$

Backward regression consists of an $l2$-regularized "ridge" regression from Y to X:

$$G_{bwd} = (Y^T Y + \lambda I)^{-1} Y^T X \tag{11}$$

CCA finds $G_{cca} \in \mathbb{R}^{d_z, d_y}$ and $H_{cca} \in \mathbb{R}^{d_z, d_x}$ s.t. X and Y are maximally correlated in a latent Z space:

$$G_{cca}, H_{cca} = \underset{G,H}{\text{argmax}}\, corr(XH^T, YG^T) \tag{12}$$

PLS finds $G_{pls} \in \mathbb{R}^{d_z, d_y}$ and $H_{pls} \in \mathbb{R}^{d_z, d_x}$ s.t. X and Y are maximally covarying in a latent Z space:

$$G_{pls}, H_{pls} = \underset{G,H}{\text{argmax}}\, \text{cov}(XH^T, YG^T) \tag{13}$$

We employ five-fold nested cross-validation to select the optimal number of components for CCA and PLS. Regressions were $\ell 2$-regularized with a λ regularization parameters fitted with the efficient leave-one-out procedure implemented in scikit-learn RidgeCV [37].

3.1.2. Evaluating Causal Discovery from models' coefficients

B2B leads to *scalar* coefficients for non-causal features. The diagonal of this matrix, $\hat{S} \in \mathbb{R}^{d_x}$, can thus be directly used as a causal contribution estimate. Note that this estimate is unbiased (i.e. zeros-centered) only if the second regression H is not regularized.

In contrast, the loading coefficients of the Forward ($H_i \in \mathbb{R}^{d_y}$), Backward ($G^i \in \mathbb{R}^{d_y}$), CCA and PLS models ($H_i \in \mathbb{R}^{d_z}$) lead to a loading *vector* for each feature i. To estimate causal contribution, we must thus transform such vectors into scalars, by e.g. taking the sum of square coefficients: $\hat{S}_i = \sum_j {H_i^j}^2$ Note that in such B2B cases, the estimates are thus positive and would thus bias a second-level statistical analysis against 0.

Finally, to estimate whether each model accurately identifies causal factors independently of their potential biases, we compute the area-under-the-curve (AUC) across factors $AUC(S, \hat{S})$. By

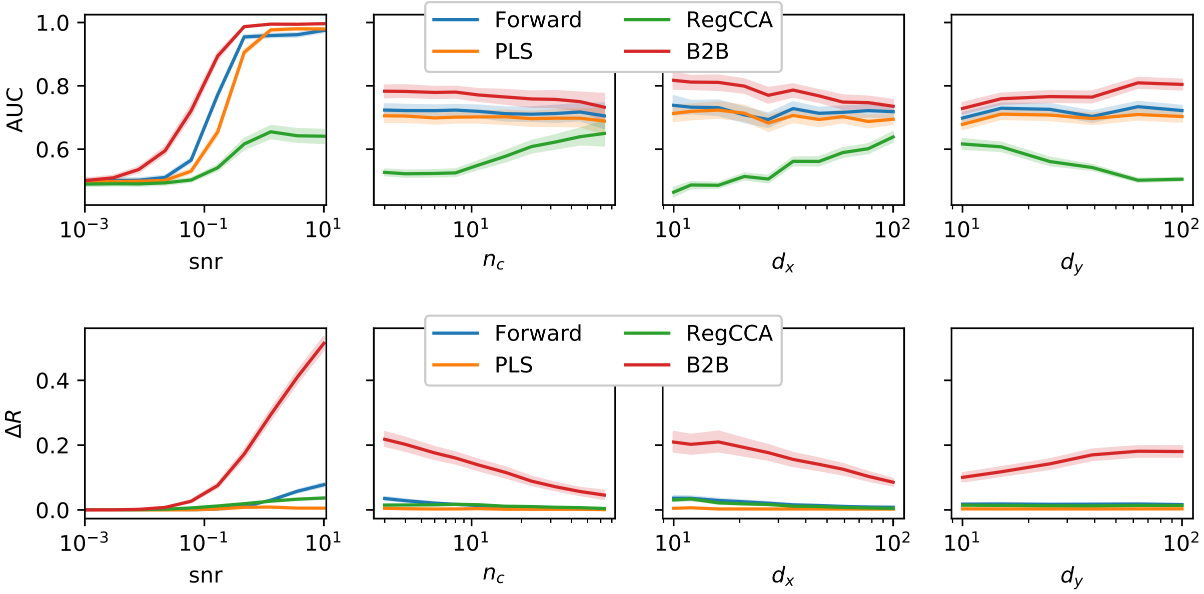


Figure 2: Synthetic experiments. Average AUC (top) and Feature Importance ΔR (bottom) when varying experimental conditions individually. Higher is better. B2B compares favorably in all cases.

definition, this AUC evaluation can only be done when ground truth labels are available, as is the case in this synthetic setup, but not in the neuroimaging experiments below.

Figures 2 (top) and B.6 (top) show the results of this AUC evaluation. Note that the figure does not display each feature separately, as they are randomly generated. The results show that B2B compares favorably to other methods on these synthetic data.

3.1.3. Evaluating Causal Discovery through the reliability of held-out prediction

In most real-world cases, S is not known. Consequently, the above AUC evaluation cannot be estimated. To address this issue, we assess the ability of each model to reliably predict independent and identically distributed data from Y, given all of the X features versus all-but-one feature X_{-i} (i.e. 'knock-out X'). This procedure results in two correlation metrics R_{full} and $R_{knockout}$ for each feature (for the B2B and Backward models) for each dimension of Y (for the Forward model) or each canonical dimension of Y (for CCA and PLS). The difference $\Delta R_i = R_{full} - R_{knockout}$ indicates how much each X_i improves the prediction of a) the target dimension (i.e. $G^i Y \in \mathbb{R}$ for B2B, b) the average across all of the dimensions j of Y ($\frac{1}{d_y}\sum_j^{d_y} \Delta R_i^j$) for the Forward model or c) the average across the canonical dimensions j of Y ($\frac{1}{d_z}\sum_j^{d_z} \Delta R_i^j$) for CCA and PLS. We show in Appendix Appendix A.3 pseudo-code to assess feature importance for each model. For the Backward Model, feature importance cannot be assessed as the X collinearity is never taken into account.

Figures 2 (bottom) and B.6 (right, in Appendix) show the results of this evaluation on held-out data. Overall both the AUC and the held-out prediction reliability evaluations show that B2B compares favorably to the baseline models.

3.2. functional Magnetic Resonance Imaging Experiment

Next, we apply our method to brain imaging data from the anonymized multimodal neuroimaging "Mother Of all Unification Studies" (MOUS) study [42]. The dataset contains functional Magnetic Resonance Imaging (fMRI) and magneto-encephalography (MEG) recordings of 102 healthy Dutch adults who performed a reading task in the scanner. Ten subjects were excluded from the analysis (9/102 MEG and 1/102 fMRI) because of technical difficulties reading the files. Subjects were exposed to a rapid serial visual presentation of Dutch words. The word lists consisted of 120 sentences, and scrambled lists of the same words. Each word was presented on the computer screen for 351ms on average (min: 300ms, max: 1400ms). Successive words were separated by a blank screen for 300ms, and successive sentences were separated by an empty screen for a few (3-4) seconds.

3.2.1. fMRI preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.0.7 ([10]; [9]; RRID:SCR_016216), which is based on *Nipype* 1.4.2 ([14]; [15]; RRID:SCR_002502).

Anatomical data preprocessing A total of two T1-weighted (T1w) images per subject were found within the input BIDS dataset. All of them were corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` [46], distributed with ANTs 2.2.0 [2, RRID:SCR_004757]. The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtractio` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` [FSL 5.0.9, RRID:SCR_002823, 49]. A T1w-reference map was computed after registration of 2 T1w images (after INU-correction) using `mri_robust_template` [FreeSurfer 6.0.1, 40]. Brain surfaces were reconstructed using `recon-all` [FreeSurfer 6.0.1, RRID:SCR_001847, 7], and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle [RRID:SCR_002438, 28]. Volume-based spatial normalization to two standard spaces (MNI152NLin2009cAsym, MNI152NLin6Asym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: *ICBM 152 Nonlinear Asymmetrical template version 2009c* [[12], RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym], *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* [[11], RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym],

Functional data preprocessing For each of the 2 BOLD runs found per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration [18]. Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` [FSL 5.0.9, 25]. BOLD runs were slice-time corrected using `3dTshift` from AFNI 20160207 [6, RRID:SCR_005927]. The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): *fsnative*, *fsaverage5*. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Automatic removal of motion artifacts using independent component analysis [ICA-AROMA, 39] was performed on the *preprocessed BOLD on MNI space* time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). Corresponding "non-aggresively" denoised runs were produced after such smoothing.

Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 [1, RRID:SCR_001362], mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

The preprocessed volumetric fMRI data was linearly projected to its closest surface using Nilearn `vol_to_surf` function with a 3 mm radius and the Freesurfer 'fsaverage5' surface. A surface searchlight analysis was then implemented by concatenating the surface fMRI data within a 8mm radius of each vertex. For each subject and vertex, we thus build an observation matrix $Y \in \mathbb{R}^{m \times d_y}$ of $m \approx 1400$ words by $d_y \approx 40$ vertices per searchlight sphere. Each of the columns of Y is normalized to have zero mean and unit variance.

These multidimensional brain observations were to be accounted for (or decoded by) four features.

3.2.2. Feature definition

We aim to identify the word features that cause a variation in brain responses. We consider four distinct but collinear features. First, 'Word Length' refers to the total number of letters. Word Length is expected to primarily cause a variation in the early evoked MEG responses (i.e. from 100 ms after stimulus onset) primarily elicited by the retinotopically-tuned visual cortices (e.g. [38].) but expected across the full visual hierarchy (e.g. [19]). Second, 'Word Frequency' indexes how frequently each word appears in Dutch and here derives from the Zipf logarithmic scale of [47] provided by the WordFreq package [44]. Word Frequency is expected to primarily cause a variation in the late evoked MEG responses (i.e. from 400 ms) in the left frontal, temporal and parietal cortices [30, 32]. Third, 'Word Function' indicates whether each word is a content word (i.e. a noun, a verb, an adjective or an adverb) or a function word (i.e. a preposition, a conjunction, a determinant, a pronoun or a numeral), and here derives from Spacy's part of speech tagger [22]. To our knowledge, this feature has not been thoroughly investigated with fMRI and MEG. While its causal contribution to reading processes in the brain thus remains unclear, this lexical feature can nonetheless be expected to present similar brain patterns to word frequency. Finally, to verify that B2B and other methods would not inadequately identify non-causal features, we added a dummy feature, constructed from a noisy combination of Word Length and Word Frequency: $dummy = z(length) + z(frequency) + \mathcal{N}$, where z normalizes features and $\mathcal{N}$ is a random vector sampling Gaussian distribution (all terms thus have a zero-mean and a unit-variance).

To account for the delay of blood oxygenation level dependent responses, these four features were convolved using the Glover hemodynamic response function of Nilearn's `compute_regressor` function an oversampling of 16 and default parameters [1].

This procedure yields an $X \in \mathbb{R}^{m \times d_x}$ matrix of $m \approx$ 840 TR (Repetition Time: 2 sec.) by $d_x = 4$ factors for each subject. Each of the columns of X is normalized to have a zero-mean and a unit variance.

3.2.3. Models and statistics

We compare B2B to four standard methods: Forward regression, Backward regression, Regularized CCA and PLS, as implemented in scikit-learn [37] and Pyrcca [3], and optimized with nested cross-validation over twenty $l2$ regularization parameters logarithmically spaced between 10^{-4} and 10^{4} (for regression methods) or 1 to 4 canonical components (for cross-decomposition methods).

We used the feature importance described in Algorithm 2 to assess the extent to which each feature X_i specifically improves the prediction of held-out Y data, using a 5-fold cross-validation (with shuffled trials to homogenize the distributions between the training and testing splits).

Each model was implemented for each subject and each time sample independently. Pairwise comparison between models were performed using a two-sided Wilcoxon signed-rank test across subjects using the average ΔR across time. Corresponding effect sizes are shown in Fig. 5, and p-values are reported below.

3.2.4. fMRI Results

We compared the ability of Forward regression, Backward regression, CCA, PLS and B2B to estimate the causal contribution of four distinct but collinear features on brain responses to words.

Supplementary Fig. Appendix B.1 shows that the Backward model decodes the dummy variable well above chance level. In addition, it decodes both Word Length and Word Frequency across large sections of the cortex, even though these features are known to primarily influence early visual cortex and associative cortices respectively. In consequence, while Backward modeling achieves high sensitivity, it is not valid to estimate the specific causal contribution of collinear features.

On the contrary, the H coefficients of the Forward model reveals the expected spatial specificity of Word Length and Word Frequency, peaking in the visual and temporal cortices respectively. However, such Forward coefficients may be variable across subjects, and could thus underestimate the significance of each factor at each vertex, as evaluated with a signed-rank tests across subjects. In principle, such inter-subject variability could be less impactful with multivariate observations methods such as CCA and PLS [3, 27], but they fail to provide a single and interpretable coefficient for each feature at each vertex.

In contrast, the $\hat{S}$ of an unbiased B2B (i.e. a B2B where H is not regularized) reaches higher levels of significance at each vertex than the Forward H (Supplementary Fig. Appendix B.1). (Note

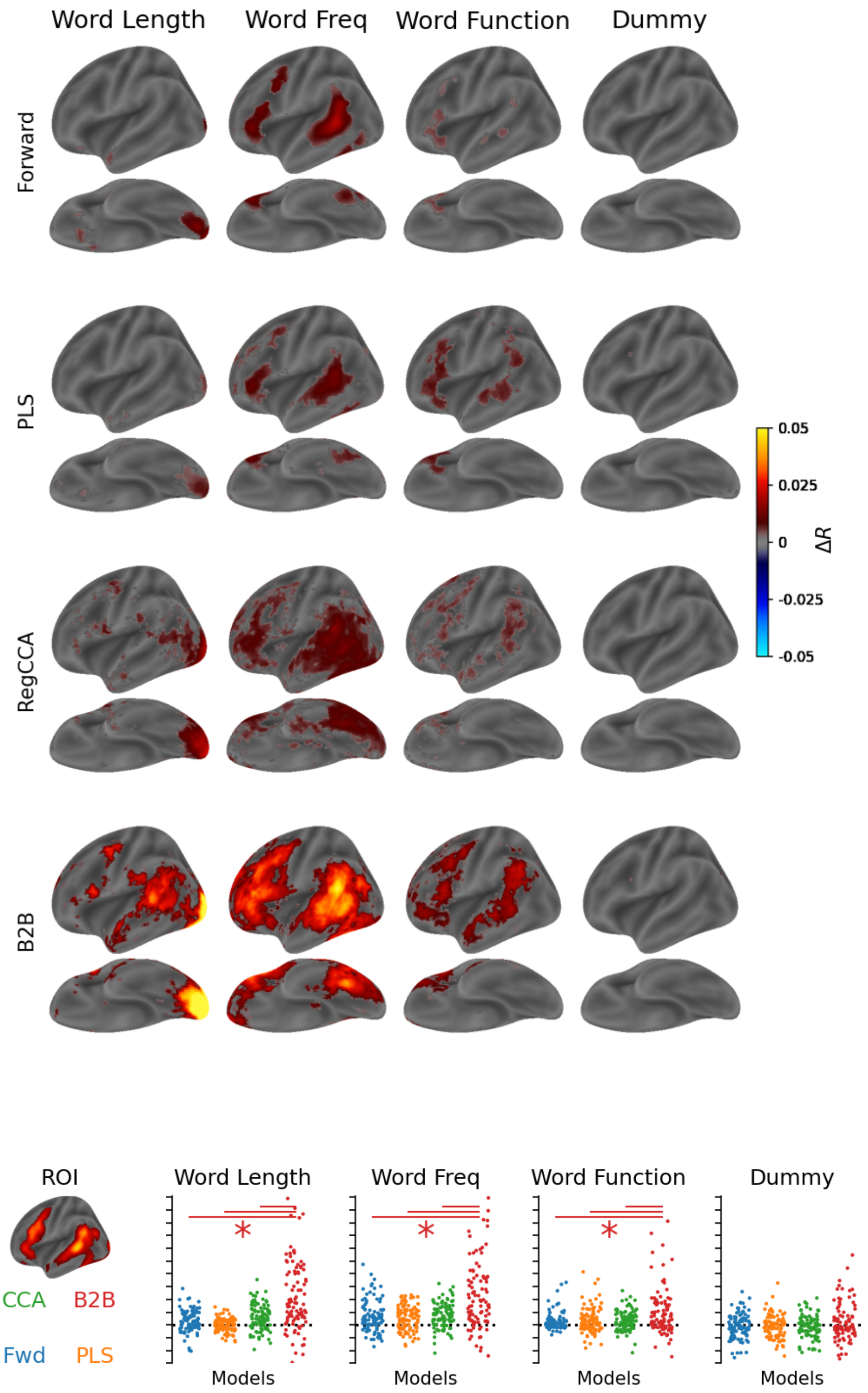


Figure 3: Top. Multiple models (rows) are compared on their ability to reliably predict out-of-sample fMRI signals evoked by words by quantifying the improvement of correlation coefficient ΔR for each of the four features (columns). Each cell displays the left hemisphere view laterally (top) þɪ ventrally (bottom). Bottom left. Region of Interest (ROI), defined by the cortex vertices that were reliably predicted for by the gold standard Forward Model using all features and thresholded with a Wilcoxon signed-rank test at $p < 0.001$ (uncorrected) across subjects. Bottom Right. Average ΔR within the ROI for each subject (dot); top horizontal lines indicate when B2B significantly outperforms other methods across subjects.

that a direct comparison between the Forward coefficients H and those of B2B ($\hat{S}$) is challenging, because these two metrics do not have the same units.)

To fairly compare the Forward, CCA, PLS and B2B methods on a common and robust evaluation metrics, robust to coefficient biases, we thus compare them on their ability to improve Y predictions of out-of-sample data when one factor is introduced into the model. This analysis leads to a ΔR as described in the synthetic experiment. As expected, the Forward, CCA, PLS and B2B method predicted that the Dummy Variable does not improve the Y prediction. This confirms that these methods accurately rule out the known non-causal factor.

The average ΔR across subjects is displayed for each model and each feature in 3.2.3. Overall, B2B favorably compares to baseline models. To quantify this assessment, we compare for each subject separately the average ΔR across vertices obtained between B2B and each baseline model. To limit the inclusion of uninformative brain regions in this summary, we restrict the analysis to vertices which can be reliably accounted for by the Forward model ($p < .001$, not corrected across vertices) The results, displayed in Fig. 3.2.3, show that B2B is significantly better than baseline models on all factors (all $p < 0.0001$) but the dummy variable (min $p = 0.0852$).

3.3. Magneto-encephalograph Experiment

3.3.1. MEG preprocessing

One hundred and two subjects performed a similar reading task to the one described above in the MEG scanner. The raw MEG data was bandpass-filtered between 0.1 and 40Hz using MNE-Python default parameters [16, 17]. Specifically, we used a zero-phase finite impulse response filter (FIR) with a Hamming window and with transition bands of 0.1Hz and 10Hz for the low and high cut-off frequencies. The raw data was then segmented 100ms before word onset and 1s after word onset ($t =$ 0ms corresponds to word onset). Finally, each resulting segment was baseline-corrected between -100ms and 0ms, and decimated by 5 and thus led a sampling frequency of 240Hz. The average responses across words is displayed in Fig. 4. For each subject and each time sample relative to word onset, we build an observation matrix $Y \in \mathbb{R}^{m \times d_y}$ of $m \approx$ 2,700 words by $d_y = 301$ MEG channels (273 magnetometers and 28 compensation channels). Each of the columns of Y is normalized to have zero mean and unit variance.

We use the same features as for the fMRI experiments, except that the features were not convolved by an hemodynamic response function. This yields an $X \in \mathbb{R}^{m \times d_x}$ matrix of $m \approx$ 2,700 words by $d_x = 4$ features for each subject. Each of the columns of X is normalized to have a mean and a standard deviation of 0 and 1 respectively.

We compare B2B against other methods following the same procedure as for the fMRI experiments, except that the searchlight swipes across time samples, and is trained across all MEG channels.

3.3.2. Results

We compared the ability of Forward regression, Backward regression, CCA, PLS and B2B to estimate the causal contribution of four distinct but linearly-correlated features on brain evoked responses to words.

As expected, the Backward model reveals a similar decoding time course for Word Length and Word Frequency, even though these features are known to specifically influence early and late

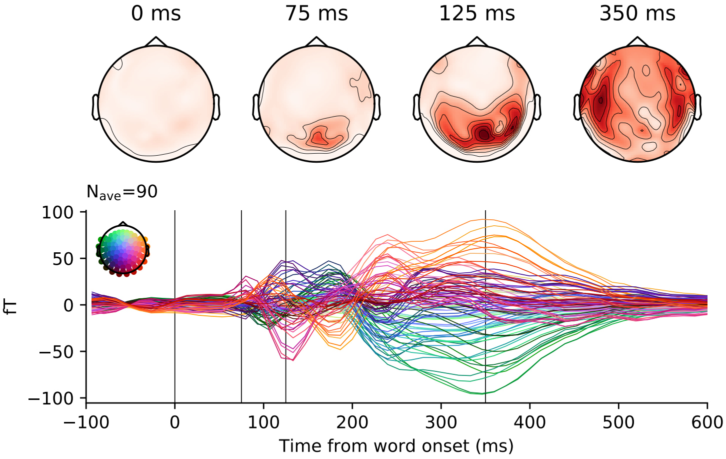


Figure 4: A hundred subjects read approximately 2,700 words while their brain activity was recorded with MEG. Top. Average brain response to words (word onset at t=0 ms), as viewed from above the head (red= higher gradient of magnetic flux). Bottom. Each line represents a magnetometer, color-coded by its spatial position. Posterior responses, typical of primary visual cortex activity, peak around 100 ms after word onset and are followed by an anterior propagation of activity typical of semantic processing in the associative cortices.

MEG responses respectively [30]. In addition, the same decoding time course was observed for the dummy variable. Once again, these results illustrate that Backward modeling cannot be used to estimate the causal contribution of correlated features.

We thus focus on the four remaining methods (i.e. Forward Regression, PLS, CCA, and B2B) and estimate their ΔR (i.e. the improvement of Y prediction induced by the introduction of a given feature into the model,as described in Algorithm 2). Contrary to the Backward Model, none of the models predicted the Dummy Variable to improve the Y prediction: all $\Delta R < 0$ (all $p > .089$).

Figure 5 shows, for each model, the effects obtained across time (left) and subjects (right).

Word Length and Word Frequency improved the prediction performance of all methods: $\Delta R > 0$ for all models (all $p < 0.0001$). As expected, the time course associated with Word Length and Word Frequency rose from ≈ 100 ms and from ≈ 400 ms respectively. Furthermore, Word Function improved the prediction performance of all models (all $p < 0.0002$) except for PLS ($p = 0.7989$). Overall, these results confirm that Word Length, Word Frequency and Word Function specifically improve the prediction of specific periods of brain responses to words, and thus form plausible independent causal contributors.

We compare B2B to other models across subjects (Fig. 5 right). For both Word Length and Word Frequency, B2B outperforms all models (all $p < 0.0001$). For "Word Function", B2B outperforms all models (all $p < 0.0001$) but CCA ($p < 0.0001$). Overall, these results show that B2B generally compares favorably against baseline models.

4. Discussion

Here, we introduce B2B, a method to disentangle the causal contribution of collinear factors from multidimensional observations. After proving the validity of B2B, we show that it generally compares favorably against baseline models both on a wide spectrum of synthetic data and on two large neuroimaging datasets.

In addition, B2B can be very fast to compute as long as both H and G are based on l2 optimization with l2 regularization (Supplementary Fig. B.7), as is done in the above experiments. However, B2B does not need not be limited to l2 optimization and regularization: the H and G

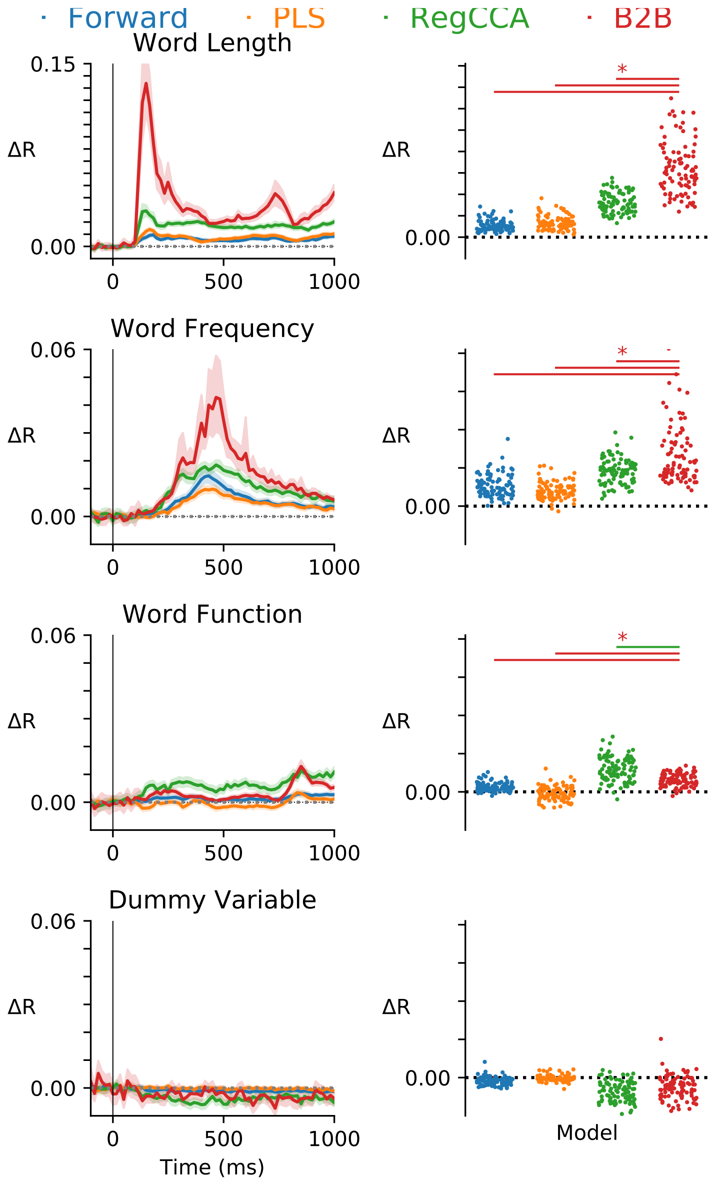


Figure 5: Multiple models (color-coded) are compared on their ability to reliably predict single-trial MEG signals evoked by words. Left. Average improvement of correlation coefficient ΔR for each of the four features (rows). Error bars indicate standard error of the mean (SEM) across subjects. Right. Average ΔR across time for each subject (dots). Top horizontal lines indicate when B2B significantly outperforms other methods (red) and vice versa.

operators can in principle be found with e.g. support vector or random forest regression. We reserve the study of these computational variants for future work.

Like forward and cross-decomposition models, B2B is limited by the correlations between factors. At the extreme, if two factors are identical, and thus fully correlated, no statistical method can disentangle their relative causal contribution, and intervention is thus mandatory. In practice, this implies that, like general linear models, B2B will best work with high signal-to-noise ratio and/or orthogonal factors, and will see its sensitivity diminish when the signal-to-noise ratio of collinear factors drops.

In the present neuroimaging context, B2B follows a long series of statistical methods designed to characterize brain representations - i.e. to identify what sensory feature causes specific brain responses [35]. In this regard, CCA and PLS have been used in electrophysiology and neuroimaging to track representations (e.g. [34]) as well as to denoise recordings as well as to align subjects [23, 8]. While CCA and PLS relates to B2B, these methods diverge in several important ways. First, they have different objectives: CCA aims to find the potentially numerous and poorly interpretable components where X and Y are maximally correlated, whereas B2B aims to recover the causal factors from X to Y. Second, B2B is not symmetric between X and Y: it aims to identify specific causal features by first optimizing over the decoders G and then over H. By contrast, CCA and PLS are symmetric between X and Y, and aims to find G and H jointly such that they project X and Y on maximally correlated dimensions. Third, CCA is based an eigen decomposition of XH and YG - the corresponding canonical components are thus mixing the X features in way that limit interpretability and potentially dilute the impact of each feature onto multiple components. In contrast B2B assesses each feature X_i on a single Y component specifically selected to maximize signal-to-noise ratio of that feature i. Fourth CCA does not separately optimize two distinct regularization parameters for G and H, whereas B2B does. Finally, CCA does not use different data splits to estimate G and H. Together, these differences may explain why B2B generally outperform CCA on estimating causal influences (Figs. 2 and B.6).

One popular method to investigate multidimensional patterns of brain activity is Representational Similarity Analysis (RSA) [29]. RSA quantifies the similarity of brain responses associated with specific categorical conditions (e.g. distinct images), by (1) fitting one-against-all classifiers on each category and (2) testing whether these classifiers discriminates all other categories. The resulting categories × categories confusion matrix is then analyzed, generally in an unsupervised manner, to reveal the categories that present similar brain activity patterns. B2B subsumes RSA in that (1) it can use regressions instead of one-hot classifications and (2) it is fully supervised. Consequently, and unlike RSA, B2B (1) provides interpretable coefficients and (2) can generalize to new items and new contexts. In practice, these elements allow B2B to apply to event-related paradigms and latent variable analyses, whereas RSA can only be applied when the same one-hot-encoded condition is repeated multiple times.

More generally, the present empirical results, together with their theoretical foundations, suggest that B2B may serve as a method of choice for causal discovery when feature-wise intervention is limited.

5. Acknowledgements

We are thankful to Gael Varoquaux and Alexandre Gramfort for the valuable feedback. This work was supported by ANR-17-EURE-0017 and the Fyssen Foundation to JRK for his work at PSL.

References

[1] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 2014.

[2] B.B. Avants, C.L. Epstein, M. Grossman, and J.C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008.

[3] Natalia Y Bilenko and Jack L Gallant. Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in neuroinformatics*, 10:49, 2016.

[4] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[5] Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature neuroscience*, 17(3):455, 2014.

[6] Robert W. Cox and James S. Hyde. Software tools for analysis and visualization of fmri data. *NMR in Biomedicine*, 10(4-5):171–178, 1997.

[7] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

[8] Alain de Cheveigne, Giovanni M Di Liberto, Dorothee Arzounian, Daniel DE Wong, Jens Hjortkjaer, Søren Fuglsang, and Lucas C Parra. Multiway canonical correlation analysis of brain data. *NeuroImage*, 186:728–740, 2019.

[9] Oscar Esteban, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, Asier Erramuzpe, Mathias Kent, James D. andGoncalves, Elizabeth DuPre, Kevin R. Sitek, Daniel E. P. Gomez, Daniel J. Lurie, Zhifang Ye, Russell A. Poldrack, and Krzysztof J. Gorgolewski. fmriprep. *Software*, 2018.

[10] Oscar Esteban, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit Ghosh, Jessey Wright, Joke Durnez, Russell Poldrack, and Krzysztof Jacek Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 2018.

[11] AC Evans, AL Janke, DL Collins, and S Baillet. Brain templates and atlases. *NeuroImage*, 62(2):911–922, 2012.

[12] VS Fonov, AC Evans, RC McKinstry, CR Almli, and DL Collins. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47, Supplement 1:S102, 2009.

[13] Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

[14] K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. Ghosh. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, 5:13, 2011.

[15] Krzysztof J. Gorgolewski, Oscar Esteban, Christopher J. Markiewicz, Erik Ziegler, David Gage Ellis, Michael Philipp Notter, Dorota Jarecka, Hans Johnson, Christopher Burns, Alexandre Manhães-Savio, Carlo Hamalainen, Benjamin Yvernault, Taylor Salo, Kesshi Jordan, Mathias Goncalves, Michael Waskom, Daniel Clark, Jason Wong, Fred Loney, Marc Modat, Blake E Dewey, Cindee Madison, Matteo Visconti di Oleggio Castello, Michael G. Clark, Michael Dayan, Dav Clark, Anisha Keshavan, Basile Pinsard, Alexandre Gramfort, Shoshana Berleant, Dylan M. Nielson, Salma Bougacha, Gael Varoquaux, Ben Cipollini, Ross Markello, Ariel Rokem, Brendan Moloney, Yaroslav O. Halchenko, Demian Wassermann, Michael Hanke, Christian Horea, Jakub Kaczmarzyk, Gilles de Hollander, Elizabeth DuPre, Ashley Gillman, David Mordom, Colin Buchanan,

Rosalia Tungaraza, Wolfgang M. Pauli, Shariq Iqbal, Sharad Sikka, Matteo Mancini, Yannick Schwartz, Ian B. Malone, Mathieu Dubois, Caroline Frohlich, David Welch, Jessica Forbes, James Kent, Aimi Watanabe, Chad Cumba, Julia M. Huntenburg, Erik Kastman, B. Nolan Nichols, Arman Eshaghi, Daniel Ginsburg, Alexander Schaefer, Benjamin Acland, Steven Giavasis, Jens Kleesiek, Drew Erickson, René Küttner, Christian Haselgrove, Carlos Correa, Ali Ghayoor, Franz Liem, Jarrod Millman, Daniel Haehn, Jeff Lai, Dale Zhou, Ross Blair, Tristan Glatard, Mandy Renfro, Siqi Liu, Ari E. Kahn, Fernando Pérez-García, William Triplett, Leonie Lampe, Jörg Stadler, Xiang-Zhen Kong, Michael Hallquist, Andrey Chetverikov, John Salvatore, Anne Park, Russell Poldrack, R. Cameron Craddock, Souheil Inati, Oliver Hinds, Gavin Cooper, L. Nathan Perkins, Ana Marina, Aaron Mattfeld, Maxime Noel, Lukas Snoek, K Matsubara, Brian Cheung, Simon Rothmei, Sebastian Urchs, Joke Durnez, Fred Mertz, Daniel Geisler, Andrew Floren, Stephan Gerhard, Paul Sharp, Miguel Molina-Romero, Alejandro Weinstein, William Broderick, Victor Saase, Sami Kristian Andberg, Robbert Harms, Kai Schlamp, Jaime Arias, Dimitri Papadopoulos Orfanos, Claire Tarbert, Arielle Tambini, Alejandro De La Vega, Thomas Nickson, Matthew Brett, Marcel Falkiewicz, Kornelius Podranski, Janosch Linkersdörfer, Guillaume Flandin, Eduard Ort, Dmitry Shachnev, Daniel McNamee, Andrew Davison, Jan Varada, Isaac Schwabacher, John Pellman, Martin Perez-Guevara, Ranjeet Khanuja, Nicolas Pannetier, Conor McDermottroe, and Satrajit Ghosh. Nipype. *Software*, 2018.

[16] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.

[17] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *Neuroimage*, 86:446–460, 2014.

[18] Douglas N Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1):63–72, 2009.

[19] Donald J Hagler Jr and Martin I Sereno. Spatial maps in frontal and prefrontal cortex. *Neuroimage*, 29(2):567–577, 2006.

[20] Martin N Hebart and Chris I Baker. Deconstructing multivariate decoding for the study of brain function. *Neuroimage*, 180:4–18, 2018.

[21] Arthur E Hoerl. Optimum solution of many variables equations. *Chemical Engineering Progress*, 55(11):69–78, 1959.

[22] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 2017.

[23] H. Hotelling. Relations between two sets of variables. *Biometrika*, (28):129–149, 1936.

[24] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453, 2016.

[25] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002.

[26] G. V. Kass. Significance testing in automatic interaction detection (a.i.d.). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):178–189, 1975.

[27] Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis Engemann, Eric Larson, and Alexandre Gramfort. Encoding and decoding neuronal dynamics: Methodological framework to uncover the algorithms of cognition, 2018.

[28] Arno Klein, Satrajit S. Ghosh, Forrest S. Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, Brian Rossa, Martin Reuter, Elias Chaibub Neto, and Anisha Keshavan. Mindboggling morphometry of human brains. *PLOS Computational Biology*, 13(2):e1005350, 2017.

[29] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[30] Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647, 2011.

[31] Ludovic Lebart, Alain Morineau, and Marie Piron. *Statistique exploratoire multidimensionnelle*, volume 3. Dunod Paris, 1995.

[32] Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends in cognitive sciences*, 7(7):293–299, 2003.

[33] J. A. Morgan and J. N. Sonquist. Problems in the analysis of survey data: and a proposal. *J. Amer. Statist. Ass.*, (58):415–434, 1963.

[34] Kathrin Müsch, Kevin Himberger, Kean Ming Tan, Taufik A Valiante, and Christopher J Honey. Transformation of speech sequences in human sensorimotor circuits. *Proceedings of the National Academy of Sciences*, 117(6):3203–3213, 2020.

[35] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[36] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[37] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[38] Felipe Pegado, Enio Comerlato, Fabricio Ventura, Antoinette Jobert, Kimihiro Nakamura, Marco Buiatti, Paulo Ventura, Ghislaine Dehaene-Lambertz, Régine Kolinsky, José Morais, et al. Timing the impact of literacy on visual processing. *Proceedings of the National Academy of Sciences*, 111(49):E5233–E5242, 2014.

[39] Raimon H. R. Pruim, Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. Ica-AROMA: A robust ICA-based strategy for removing motion artifacts from fmri data. *NeuroImage*, 112(Supplement C):267–277, 2015.

[40] Martin Reuter, Herminia Diana Rosas, and Bruce Fischl. Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, 53(4):1181–1196, 2010.

[41] Ryan M Rifkin and Ross A Lippert. Notes on regularized least squares. Technical report, MIT, 2007.

[42] Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche HL Lam, Julia Uddén, Annika Hultén, and Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific data*, 6(1):17, 2019.

[43] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Scipy, 2010.

[44] Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. Luminosoinsight/wordfreq: v2.2, October 2018.

[45] Nicholas A Steinmetz, Christof Koch, Kenneth D Harris, and Matteo Carandini. Challenges and opportunities for large-scale electrophysiology with neuropixels probes. *Current opinion in neurobiology*, 50:92–100, 2018.

[46] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

[47] Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. Subtlex-uk: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology*, 67(6):1176–1190, 2014.

[48] Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015.

[49] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, 2001.

6. Appendices

Appendix A. Appendix

Appendix A.1. Proof of consistency theorem

Proof of the theorem in 2.3:

Theorem 2 (B2B consistency - general case)**.** *Consider the B2B model from equation 1*

$$Y = (XS + N)F$$

with N centered and full rank noise.

If F and X are full-rank on $Img(S)$, then, the solution of B2B, $\hat{H}$ minimizes

$$\min_H \|X - XH\|^2 + \|NH\|^2$$

and satisfies

$$S\hat{H} = \hat{H}$$

Proof. Let $\hat{G}$ and $\hat{H}$ be the solutions of the first and second regressions of B2B.

Since $\hat{G}$ is the least square estimator of X from Y

$$\hat{G} = \arg\min_G \mathbb{S}[\|YG - X\|^2]$$

Replacing Y by its model definition $Y = (XS + N)F$, we have

$$\hat{G} = \arg\min_G \mathbb{S}[\|X - (XS + N)FG\|^2] = \arg\min_G \mathbb{S}[\|X - XSFG + NFG\|^2]$$

Since N is centered and independent of X, we have

$$\hat{G} = \arg\min_G \|X - XSFG\|^2 + \|NFG\|^2 \tag{A.1}$$

In the same way, for $\hat{H}$, we have

$$\begin{aligned}\hat{H} = \arg\min_H \mathbb{S}[\|XH - Y\hat{G}\|^2] &= \arg\min_H \mathbb{S}[\|XH - (XS + N)F\hat{G}\|^2] \\ &= \arg\min_H \mathbb{S}[\|X(H - SF\hat{G})\|^2] + \mathbb{S}[\|NF\hat{G}\|^2] \\ &= \arg\min_H \mathbb{S}[\|X(H - SF\hat{G})\|^2]\end{aligned}$$

a positive quantity which reaches a minimum (zero) for

$$\hat{H} = SF\hat{G} \tag{A.2}$$

Let us now prove that $SF\hat{G} = F\hat{G}$.

Let $F^\dagger$ be the pseudo inverse of F, and $Z = F^\dagger SF\hat{G}$, we have $FZ = FF^\dagger SF\hat{G}$

Since F is full rank on $Img(S)$, we have $FF^\dagger S = S$, and $FZ = SF\hat{G}$

As S is a binary diagonal matrix, it is an orthogonal projection and therefore a contraction, thus

$$\|NSF\hat{G}\|^2 \leq \|NF\hat{G}\|^2$$

and

$$\|X - XSFZ\|^2 + \|NFZ\|^2 = \|X - XSF\hat{G}\|^2 + \|NSF\hat{G}\|^2 \leq \|X - XSF\hat{G}\|^2 + \|NF\hat{G}\|^2$$

But since $\hat{G} = \arg\min_G \|X - XSFG\|^2 + \|NFG\|^2$, we also have

$$\left\|X - XSF\hat{G}\right\|^2 + \left\|NF\hat{G}\right\|^2 \leq \|X - XSFZ\|^2 + \|NFZ\|^2$$

Summarizing the above,

$$\left\|X - XSF\hat{G}\right\|^2 + \left\|NF\hat{G}\right\|^2 \leq \|X - XSF\hat{G}\|^2 + \|NSF\hat{G}\|^2 \leq \|X - XSF\hat{G}\|^2 + \|NF\hat{G}\|^2$$

$$\left\|X - XSF\hat{G}\right\|^2 + \left\|NF\hat{G}\right\|^2 = \|X - XSF\hat{G}\|^2 + \|NSF\hat{G}\|^2$$

$$\left\|NF\hat{G}\right\|^2 = \|NSF\hat{G}\|^2$$

N being full rank, this yields $SF\hat{G} = F\hat{G}$.

Replacing into (A.1), and setting $H = SFG$, we have

$$\begin{aligned}\hat{G} &= \arg\min_G \|X - XSFG\|^2 + \|NFG\|^2 \\ &= \arg\min_G \|X - XSFG\|^2 + \|NSFG\|^2 \\ \hat{H} &= \arg\min_H \|X - XH\|^2 + \|NH\|^2\end{aligned}$$

Finally, $S\hat{H} = SSF\hat{G} = SF\hat{G} = \hat{H}$, since S, a binary diagonal matrix, is involutive. This completes the proof. □

Appendix A.2. Modeling measurement noise

Equation 1 does not explicitly contain a measurement noise term. Yet, in most experimental cases, the problem is best described as:

$$Y = (XS + N)F + M \tag{A.3}$$

with $M \in \mathbb{R}^{m \times d_y}$.

This equation is actually equivalent to Equation 1 given our hypotheses. Indeed, we can rewrite $M = MF^{-1}F$ over $Img(F)$, which leads to:

$$Y = (XS + N)F + M = (XS + N + MF^{-1})F = (XS + N')F$$

Consequently, assuming that F is full rank on $Img(XS)$, B2B yields the same solutions to equations 1 and A.3.

Appendix A.3. Feature importance

For B2B, feature importance is assessed as follows:

Algorithm 2: B2B feature importance.

```
Input: $X_{train} \in \mathbb{R}^{m \times d_x}$, $X_{test} \in \mathbb{R}^{m' \times d_x}$, $Y_{train} \in \mathbb{R}^{m \times d_y}$, $Y_{test} \in \mathbb{R}^{m' \times d_y}$,
Output: estimate of prediction improvement $\Delta R \in \mathbb{D}^{d_x}$.
$H, G =$ B2B$(X_{train}, Y_{train})$;
$R_{full} =$ corr$(X_{test}H, Y_{test}G)$;
for $i = 1, \ldots, d_x$ do
    $K = Id$;
    $K[i] \leftarrow 0$;
    $R_k =$ corr$(X_{test}KH, Y_{test}G_i)$;
    $\Delta R_i = R_{full} - R_k$;
end
return $\Delta R$
```

For the Forward Model, the feature importance is assessed as follows:

Algorithm 3: Forward feature importance.

```
Input: $X_{train} \in \mathbb{R}^{m \times d_x}$, $X_{test} \in \mathbb{R}^{m' \times d_x}$, $Y_{train} \in \mathbb{R}^{m \times d_y}$, $Y_{test} \in \mathbb{R}^{m' \times d_y}$,
Output: estimate of prediction improvement $\Delta R \in \mathbb{D}^{d_x, d_y}$.
$H =$ LinearRegression$(X_{train}, Y_{train})$ $R_{full} =$ corr$(X_{test}K, Y_{test})$;
for $i = 1, \ldots, d_x$ do
    $K = Id$;
    $K[i] \leftarrow 0$;
    $R_k =$ corr$(X_{test}KH, Y_{test})$;
    $\Delta R_i = R_{full} - R_k$;
end
return $\Delta R$
```

For the CCA and PLS models, the feature importance is assessed as follows:

Algorithm 4: CCA and PLS feature importance.

```
Input: $X_{train} \in \mathbb{R}^{m \times d_x}$, $X_{test} \in \mathbb{R}^{m' \times d_x}$, $Y_{train} \in \mathbb{R}^{m \times d_y}$, $Y_{test} \in \mathbb{R}^{m' \times d_y}$,
Output: estimate of prediction improvement $\Delta R \in \mathbb{D}^{d_x, d_z}$.
$H, G = \text{CCA}(X_{train}, Y_{train})$;
$R_{full} = \text{corr}(X_{test}H, Y_{test}G)$;
for $i = 1, \ldots, d_x$ do
    $K = Id$;
    $K[i] \leftarrow 0$;
    $R_k = \text{corr}(X_{test}KH, Y_{test}G)$;
    $\Delta R_i = R_{full} - R_k$;
end
return $\Delta R$
```

For the Backward Model, feature importance cannot be assessed because there is no prediction combining multiple factors.

Appendix A.4. Recovering S

In case of noise, B2B yields non binary $\hat{S}$. Three thresholding rules can be used to binarize its values thus explicitly recover ”causal” features.

First, given known signal-to-noise ratio, the threshold above which a feature should considered to be ”causal” can be derived analytically. Indeed, Equation 9 implies that the k first diagonal elements of $\hat{H}$ are bounded:

$$0 \leq \frac{\sigma_{X_k}}{\sigma_{X_k} + \sigma_{N_1}} \leq \text{diag}_k(\hat{H}) \leq \frac{\sigma_{X_1}}{\sigma_{X_1} + \sigma_{N_k}}$$

where σ_{X_1}, σ_{X_k}, σ_{N_1} and σ_{N_k} denote the largest and smallest eigenvalues of $\Sigma_{X_1 X_1}$ and $\Sigma_{N_1 N_1}$.

The average value μ of non-zero coefficients of diag$(\hat{H})$ is the trace of $\hat{H}$ divided by k, and can be computed as

$$\mu = \frac{Var(X)}{Var(X) + Var(N)} \tag{A.4}$$

The decision threshold between ”causal” and ”non-causal” elements is thus a fraction μ, whose proportion arbitrarily depends on the necessity to favor type I and type II errors. In practice, we cannot use this procedure for our fMRI and MEG experiments, because signal-to-noise ratio is unknown.

Second, diag$(\hat{H})$ can be binarized with the Sonquist-Morgan criterion [33], a non-parametric clustering procedure separating small and large values in a given set. This procedure maximizes the ratio of inter-group variance while minimizing the intra-group variance, over all possible splits of the diagonal into p largest values and $d_x - p$ smallest values. Let m_0 and m_1 be the average values of the two clusters, p and $d_x - p$ their size, and v the total variance of the sample, Sonquist-Morgan criterion maximizes [26]:

$$\frac{p(d_x - p)}{d_x} \frac{(m_1 - m_0)^2}{v} \tag{A.5}$$

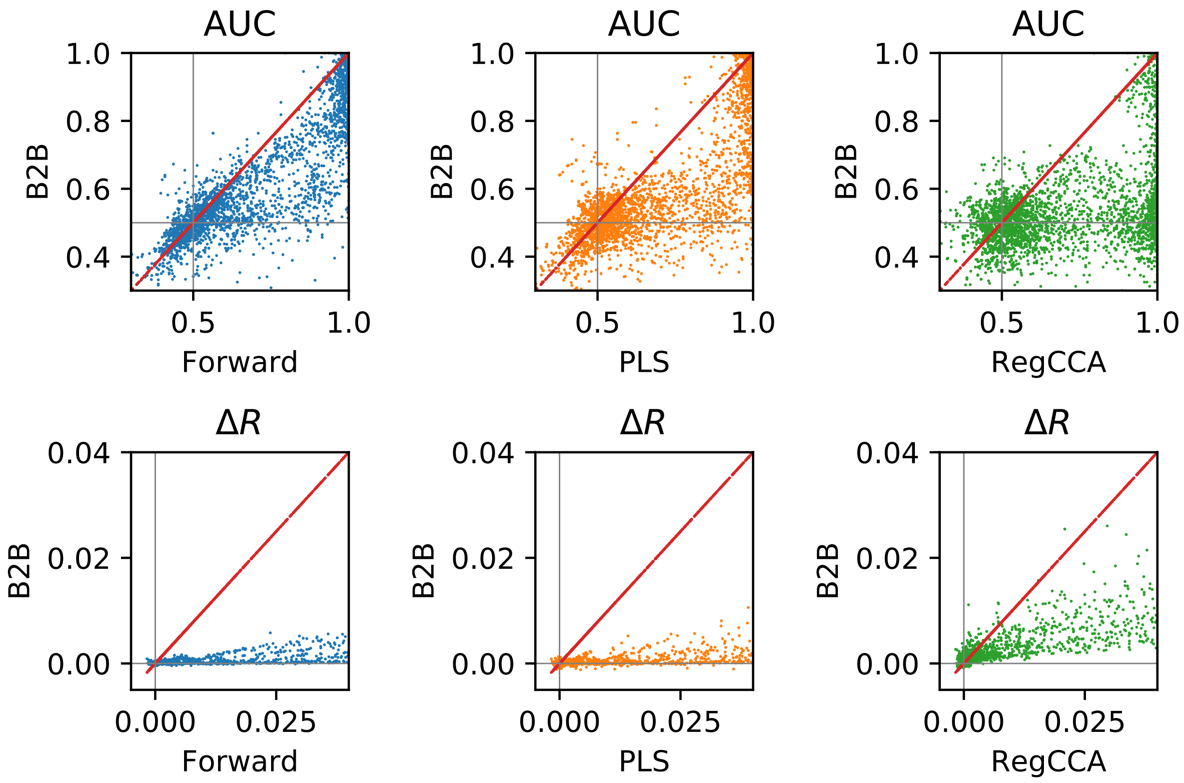


Figure B.6: Synthetic experiments. Distribution (over conditions) of AUC (top) and Feature Importance ΔR (bottom) metrics between our method (y-axis) and the baselines (x-axis). Each dot is a distinct synthetic experiment. Dots below the diagonal indicates that B2B outperform the tested model.

This procedure assumes that there exists at least one causal and at least one non-causal feature. Third, second-order statistics across multiple datasets can be used to identify the elements of $\text{diag}(\hat{H})$ that are significantly different from 0. This procedure is detailed in the method section of our MEG experiment.

Overall, these three procedures thus vary in their additional assumptions: i.e. (1) a known signal-to-noise ratio, (2) the existence of both causal and non-causal factors or (3) independent repetitions of the experiment.

Appendix B. Additional Figures

Appendix B.1. Supplementary comparison of models' coefficients

Following the recommendations of one of our reviewers, we implemented a multivariate variant of the forward model, i.e. a MANOVA, using the statsmodels implementation [43]. MANOVA is primarily used as a inferential statistics, and does not trivially convert to a predicting method. Consequently, we did not find a way to compare MANOVA against B2B with the ΔR evaluation. However, the effects of MANOVA are generally summarized with the Wilk's Lambda stastistics or its transformation into an F-value. For each searchlight, we thus use the F-values of the Wilk's

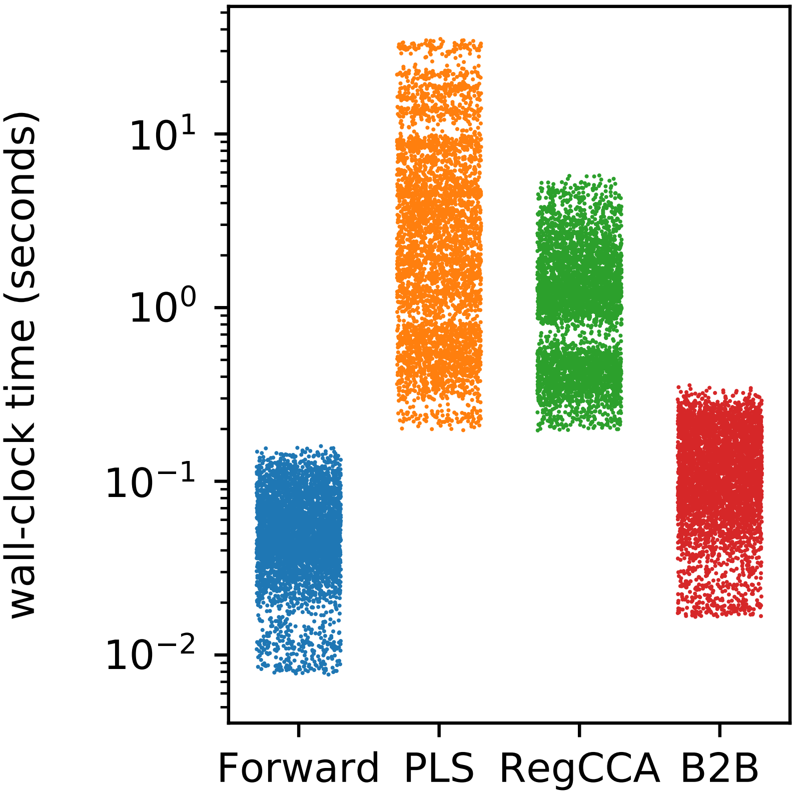


Figure B.7: Wall-clock run-time for our method B2B and for the baselines. Each dot is a distinct synthetic experiment. B2B runs much faster than cross-decomposition baselines.

Lambda statistics as proxy for $\hat{S}$ and feeds it to a second-level Wilcoxon signed-rank tests across subjects, like we did for the other models.

The results show that all factors, including the Dummy variable, are systematically above chance level in all recorded brain regions (Fig. Appendix B.1). This result can be explained by the large dimensionality of Y. Indeed, limiting the searchlight to a 1mm radius did not lead to these spurious effects but provided results similar to the Forward model.

Nonetheless, MANOVA does appear to capture some plausible effects. Indeed, the F-values obtained for both Word Length and Word Frequency were weakly but significantly higher than those obtained with the Dummy variable in the occipital and temporal brain areas (Fig. Appendix B.1). This results suggests that the effect size of the MANOVA can be biased and, thus, is not valid for second-level statistics.

Overall, this suggests that MANOVA (1) can lead to positively biased estimates of $\hat{S}$ (2) appears weaker than B2B in terms of second-level analysis across subjects (3) misses the effect of Word Function detected with the Forward model, and (4) does not trivially translate into a prediction tool. Together these elements thus suggest that MANOVA is less suitable to the present objective than B2B.

Appendix B.2. Robustness to increasing number of factors

To test whether each of the methods robustly scales to an increasingly large number of potential causes X, we enhanced the four ad-hoc features (word length, word frequency, word function, dummy variable) with another ten features. These additional features corresponds to the first dimensions of word embedding as provided by Spacy [22]. The MEG results shown in Fig. B.9, show that the feature importance of ad-hoc features as derived by B2B remain unchanged and are actually improved.

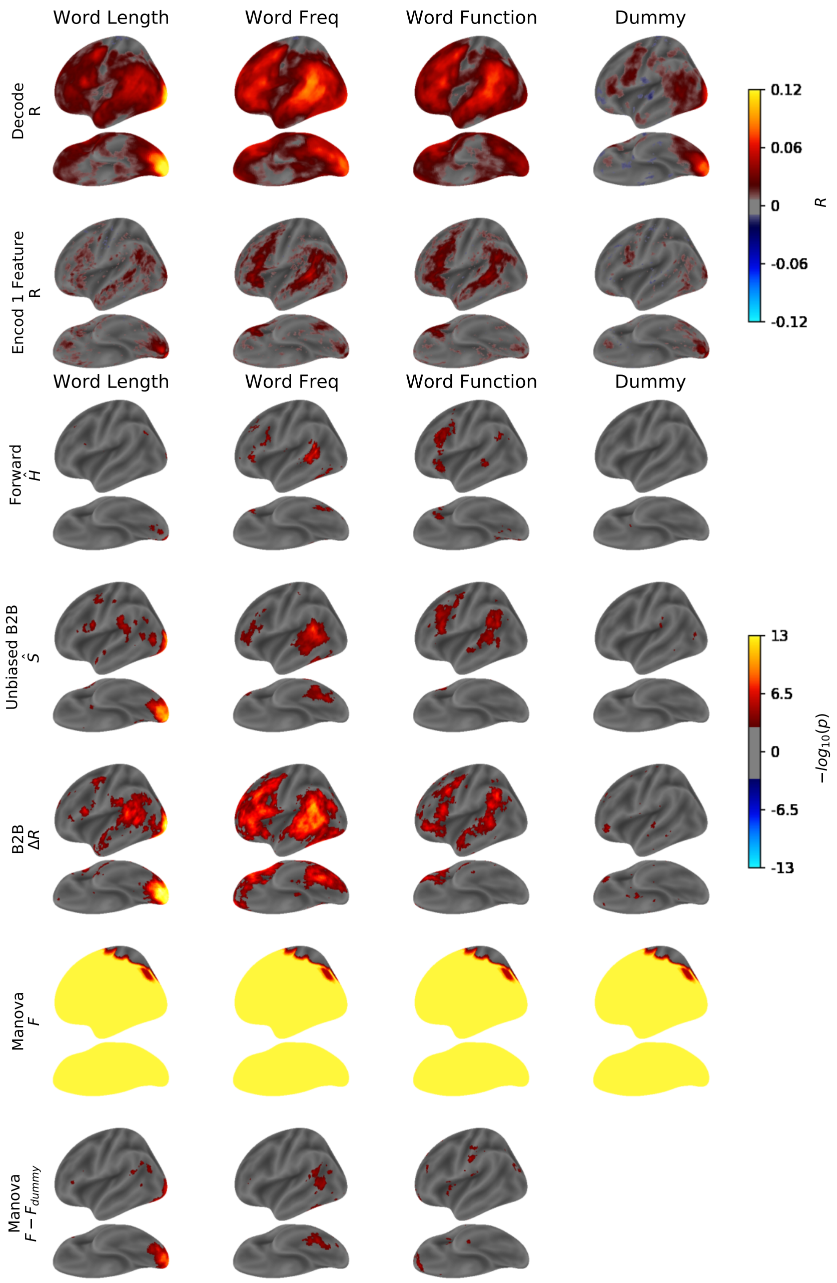


Figure B.8: Top two rows. Pearson R correlation obtained for a Backward decoding model and a Forward encoding model trained with one factor at a time. These models can not take into account the factor covariance, and thus lead to spurious effects (e.g. visual cortex effect for the dummy variable). Bottom five rows. Second- level p-values across subjects for the coefficients of the Forward, a B2B and a Manova trained with all factors, as well as for the ΔR of the B2B, for comparison. B2B achieves better p-values, without leading to spurious effects for the Dummy variable. The Manova leads to biased estimates due to its inability to deal with overfitting. The last row shows where the Manova's F-values of each factor differs significantly from the F-values estimated for the Dummy variable.

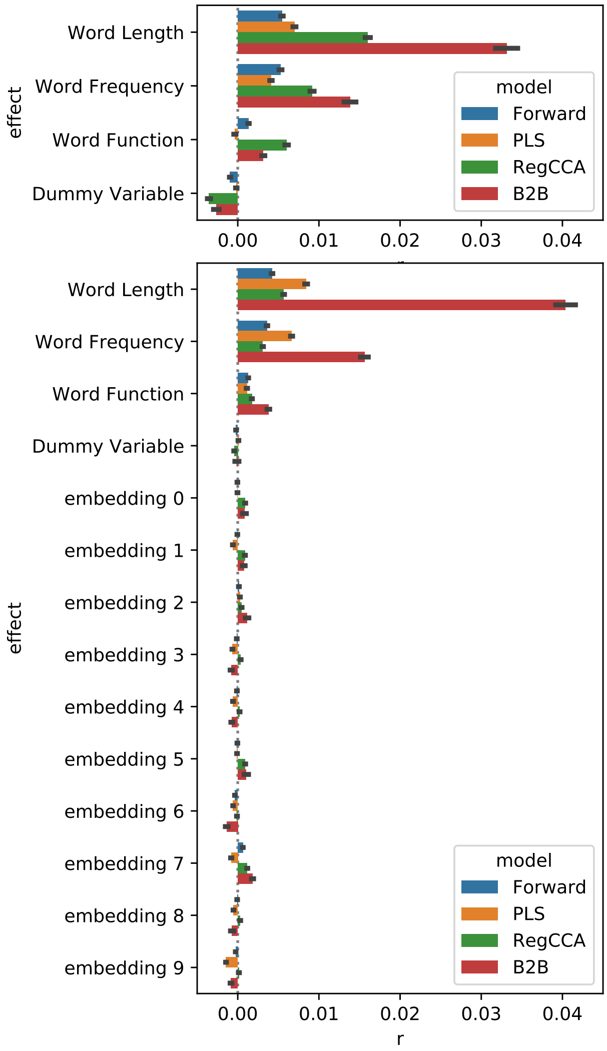


Figure B.9: Comparison of ΔR when the models are tested on four variables (top) and when the models are tested on an these four variables as well as another 10 word-embedding features (bottom). These results illustrate that, unlike Regularized CCA, B2B remains robust even when the number of tested factors increases.