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 measure, 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 provides a robust, unbiased and interpretable scalar estimate for each factor. Second, we use a variety of simulated data to show that B2B outperforms least-squares regression and cross-decomposition techniques (e.g. canonical correlation analysis and partial least squares) on causal identification when the factors and the observations are partially collinear. Finally, we apply B2B to magneto-encephalography of 102 subjects recorded during a reading task to test whether our method appropriately disentangles the respective contribution of word length and word frequency - two correlated factors known to cause early and late brain responses respectively. The results show that these two factors are better disentangled with B2B than with other standard techniques.

Keywords: Cross-Decomposition, Feature Discovery, Magnetoencephalography, Decoding, Encoding, Reading, N400

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) [18, 24]. 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 regression to predict observations (e.g. brain activity) from a minimal combination of

*corresponding author: jeanremi@fb.com

competing factors (e.g. word length, word frequency), and analytically investigate, the estimated contribution of each factor from the model's coefficients [5, 21, 32, 16, 13].

The second challenge to measuring causal influence is that observations can be multidimensional. The relationship between causes and effects is thus often considered in a *backward* manner, by training models to maximally predict causes from multidimensional observations. For example, brain activity is often recorded with hundreds or thousands of simultaneous measurements via functional Magnetic Resonance Imaging, magneto-encephalography (MEG) or multiple electrophysiological probes [5, 30]. As simultaneous measurements may be affected by common noise sources, it is common to use backward modeling, by, for example, fitting a support vector machine across multiple sensors to decode the category of a stimulus [22, 3, 17].

Both *forward* and *backward* modeling have competing benefits and drawbacks. Specifically, forward modeling disentangles the independent contribution of correlated factors, but does not combine multidimensional observations. By contrast, backward modeling combines multiple observations, but does not disentangle factors that are linearly correlated [32, 9, 16]. 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) [4]. CCA and PLS aim to find, from two sets of data X and Y, the components H and G were XH and YG are maximally correlated or maximally covarying respectively. Because CCA and PLS are based on a generalized eigen decomposition, their resulting coefficients are mixing the features of X and Y in a way that makes them notoriously difficult to interpret [19].

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

The present paper focuses on the restricted issue of disentangling the influence of linearly correlated predictors (X) onto noisy multivariate observations (Y). The present approach thus differs from other causal discovery algorithms based on temporal-delays and/or nonlinear interactions in systems where the directionality of causation (from X to Y or vice versa) is unknown (e.g. [25, 8, 14, 28].

After detailing B2B method and proving its convergence (Section 2.2), 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 a large neuroimaging dataset and reveal that distinct but linearly-correlated word features lead to distinguishable brain representations (Section 3.5).

2. Back-to-Back regression

We consider the measurement of multivariate signal $Y \in \mathbb{R}^{n \times d_y}$ (the dependent variables), generated from a set of putative causes $X \in \mathbb{R}^{n \times d_x}$ (the independent variables), via some unknown linear apparatus $F \in \mathbb{R}^{d_x \times d_y}$. Not all the variables in X exert a causal influence on Y. By considering a square binary diagonal matrix of *causal influences* $E \in \mathbb{D}^{d_x \times d_x}$, we denote by XE the causal factors of Y. In summary, the problem can be formalized as:

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

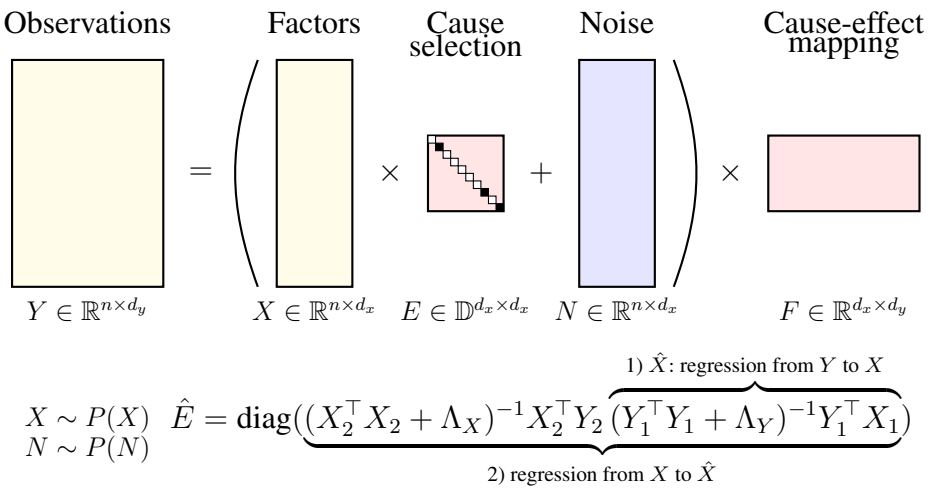

Figure 1: Back-to-back regression identifies the subset of factors $E_{ii} = 1$ in X that influence some observations Y by 1) regressing from Y to X to obtain $\hat{X}$, and 2) returning the diagonal of the regression coefficients from X to $\hat{X}$.

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

2.1. 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 $\hat{E} = \mathrm{diag}(\hat{H})$, is the desired estimate of the causal influence matrix E, as detailed in the Appendix A.1.

If using l2-regularized least-squares [10, 26], 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 splits (X_1, Y_1) and (X_2, Y_2). Then, 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, B2B is repeated over many random splits, and the final estimate $\hat{E}$ of the causal influence matrix is the average over the estimates

associated to each split [2]. To accelerate this ensembling procedure, we implemented an efficient leave-one-out cross-validation scheme as detailed in [26] as follows:

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

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

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

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 $X \in \mathbb{R}^{n \times d_x}$, output data $Y \in \mathbb{R}^{n \times d_y}$, number of repetitions $m \in \mathbb{N}$.
Output: estimate of causal influences $\hat{E} \in \mathbb{D}^{d_x \times d_x}$.

$\hat{E} \leftarrow 0$;
for $i = 1, \ldots, m$ **do**
  $(X, Y) \leftarrow \text{ShuffleRows}((X, Y))$;
  $(X_1, Y_1), (X_2, Y_2) \leftarrow \text{SplitRowsInHalf}((X, Y))$;
  $\hat{G} = \text{LinearRegression}(Y_1, X_1)$; ▷ $\hat{G} = (Y_1^\top Y_1 + \Lambda_Y)^{-1} Y_1^\top X_1$
  $\hat{H} = \text{LinearRegression}(X_2, Y_2 \hat{G})$; ▷ $\hat{H} = (X_2^\top X_2 + \Lambda_X)^{-1} X_2^\top Y_2 \hat{G}$
  $\hat{E} \leftarrow \hat{E} + \text{diag}(\hat{H})$;
end
$\hat{E} \leftarrow \hat{E}/m$;
$\hat{W} \leftarrow \text{LinearRegression}(X\hat{E}, Y)$;
return $\hat{E}$, $\hat{W}$

2.2. Theoretical guarantees

Theorem 1 (B2B consistency - general case)**.** *Consider the B2B model from Equation* $Y = (XE + N)F$*, 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* $Img(E)$*, then, the solution of B2B,* $\hat{H}$*, will minimize* $\min_H \|X - XH\|^2 + \|NH\|^2$ *and satisfy* $E\hat{H} = \hat{H}$

Proof. See Appendix Appendix A.1. □

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

$$\hat{H} = \arg\min_H \|X - XEH\|^2 + \|NEH\|^2 = (EX^\top XE + EN^\top NE)^\dagger EXX^\top. \tag{6}$$

Assuming, without loss of generality, that the active features in E 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:

$$X^\top X = \begin{pmatrix} \Sigma_{X_1 X_1} & \Sigma_{X_1 X_2} \\ \Sigma_{X_1 X_2} & \Sigma_{X_2 X_2} \end{pmatrix}, \qquad N^\top N = \begin{pmatrix} \Sigma_{N_1 N_1} & \Sigma_{N_1 N_2} \\ \Sigma_{N_1 N_2} & \Sigma_{N_2 N_2} \end{pmatrix}, \tag{7}$$

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

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

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

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

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

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

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

3. Experiments

We perform two sets of experiments to evaluate B2B: one on controlled synthetic data, and a second one on a real, large-scale magneto-encephalography (MEG) dataset. We use scikit-learn's PLS and RidgeCV [23] as well as Pyrcca's regularized canonical component analysis (RegCCA, [1]) objects to compare B2B against the standard baselines.

3.1. Synthetic data

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 in terms of prediction of independent and identically distributed data, as well as a method to recover causal factors.

The data generating process for each experiment constructs $n = 1000$ training examples according to the model $Y = (\text{h}XE + N)F$, 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, $X \in \mathbb{R}^{n \times d_x}$ contains rows drawn from $\mathcal{N}(0, \Sigma_X)$, $N \in \mathbb{R}^{n \times d_x}$ contains rows drawn from $\mathcal{N}(0, \Sigma_N)$, $E \in \mathbb{R}^{d_x \times d_x}$ 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 $\text{h} \in \mathcal{R}_+$.

To simulate a wide range of experimental conditions, we sample 10 values in log-space for $d_x, d_y \in [10, 100]$, $n_c \in [3, 63]$, $\text{h} \in [0.001, 10]$. 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 competing methods, all implemented in scikit-learn [23] and pyrcca [1]:

3.2. Baseline models

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

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

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:

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

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:

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

We employ five-fold 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 [23].

3.3. Evaluating Causal Discovery from models' coefficients

B2B leads to unbiased (i.e. zeros-centered) scalar coefficients for non-causal features. In contrast, the Forward, Backward, CCA and PLS models lead to a loading vector H_i per feature i (or one vector G^i for the backward model). To transform such vector into an estimated causal contribution $\hat{E}$, we take the sum of square coefficients: $\hat{E}_i = \sum_j H_i^{j\,2}$

To estimate whether models accurately identify causal factors, we compute the area-under-the-curve (AUC) across factors $AUC(E, \hat{E})$. The AUC allows evaluating the capacity of models at detecting the causal importance of factors when ground truth labels are available, as is the case in this setup.

We report AUC results in Figures 2 (top) and B.5 (left, in Appendix), and compare favorably to all baselines.

3.4. Evaluating Causal Discovery with held-out prediction reliability

In most cases, E is not known and AUC can thus not 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-ones feature X_{-i} (i.e. 'knock-out X'). This procedure results in two correlation metrics R_{full} and $R_{knockout}$, whose difference $\Delta R_i = R_{full} - R_{knockout}$ indicates how much each X_i improves the prediction of Y. In our figures, ΔR is the average of ΔR_i. A higher score means that for prediction, the model relies on individual features rather than combinations of features.

We show in Appendix Appendix A.3 pseudo-code to assess feature importance for our algorithm as well as baselines. For the Backward Model, feature importance cannot be assessed as the X collinearity is never taken into account.

We show in Figures 2 (bottom) and B.5 (right, in Appendix) that our method outperforms baselines.

Next, we apply our method to brain imaging data from the anonymized multimodal neuroimaging "Mother Of all Unification Studies" (MOUS) dataset [27]. The dataset contains magneto-encephalography (MEG) recordings of 102 healthy native-Dutch adults who participated in a

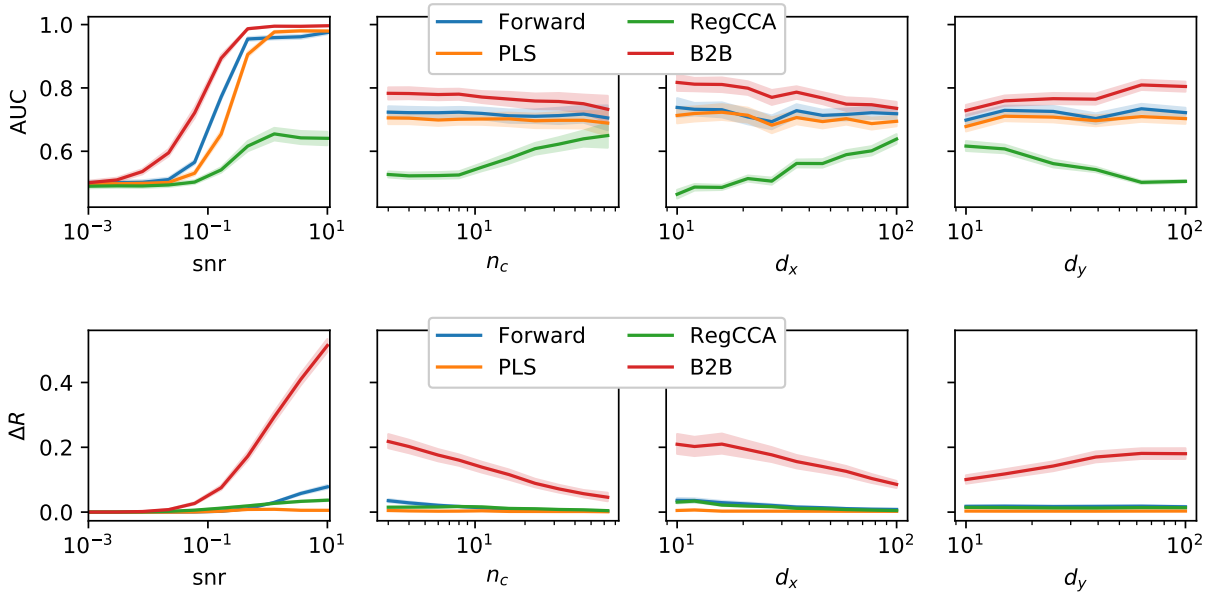

Figure 2: Synthetic experiments. Average AUC (top) and Feature Importance ΔR (bottom) when varying experimental conditions individually. Higher is better. B2B compares favorably in all cases.

reading task. Twelve subjects were excluded from the analysis because of corrupted file headers. 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.

3.4.1. MEG preprocessing

The raw MEG data was bandpass-filtered between 0.1 and 40Hz using MNE-Python default parameters [6, 7]. 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 ($t = 0$ms 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 Figure 3. For each subject and each time sample relative to word onset, we build an observation matrix $Y \in \mathbb{R}^{n \times d_y}$ of $n \approx 2{,}700$ 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.

3.4.2. Feature definition

We aim to identify the word features that cause a variation in brain responses. We consider four distinct but linearly-correlated features. First, 'Word Length' refers to the total number of letters. Word Length is expected to specifically cause a variation in the early evoked MEG responses

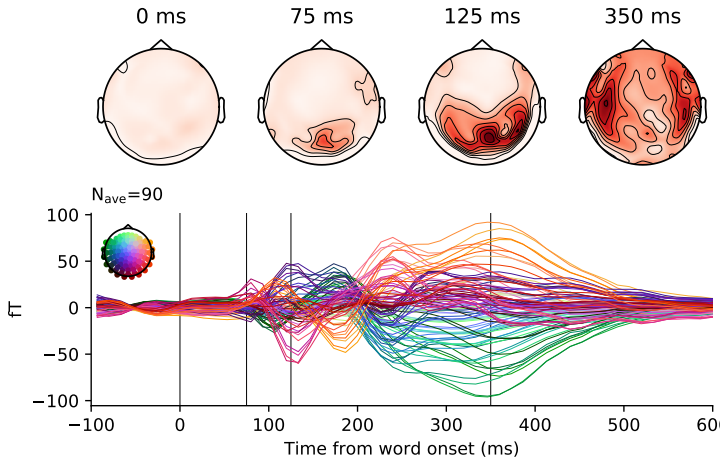

Figure 3: A hundred subjects read approximately 2,700 words while their brain activity was recorded with MEG. Top. Average brain response to words (word onset at t=0 ms), as viewed from above the head (red= higher gradient of magnetic flux). Bottom. Each line represents a magnetometer, color-coded by 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.

(i.e. from 100 ms after stimulus onset) elicited by the retinotopically-tuned visual cortices (e.g. [24].). Second, 'Word Frequency' indexes how frequently each word appears in Dutch and was derived with the the Zipf logarithmic scale of [31] provided by the WordFreq package [29]. Word Frequency is expected to specifically cause a variation in the late evoked MEG responses (i.e. from 400 ms), because it variably engages semantic processes in the temporal cortices [18]. 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 was derived from Spacy's part of speech tagger [11]. To our knowledge, this feature has not been thoroughly investigated with MEG. Its causal contribution to reading processes in the brain thus remains unclear. 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). This procedure yields an $X \in \mathbb{R}^{n \times d_x}$ matrix of $n \approx 2{,}700$ 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.

3.4.3. Models and statistics

We compare B2B to four standard methods: Forward regression, Backward regression, CCA and PLS, as implemented in scikit-learn [23] and [1], and optimized with nested cross-validation over twenty $l2$ regularization parameters logarithmically spaced between 10^{-4} and 10^{4} (for regression and CCA methods) or 1 to 4 canonical components (for PLS).

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 five-fold cross-validation (with shuffled trials to homogeneize 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 test across subjects (n=90) using the average ΔR across time. Corresponding effect sizes are shown in Figure 4, and p-values are reported below.

3.4.4. 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 MEG responses respectively [18]. In addition, the same decoding time course was observed for the dummy variable. 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 4 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

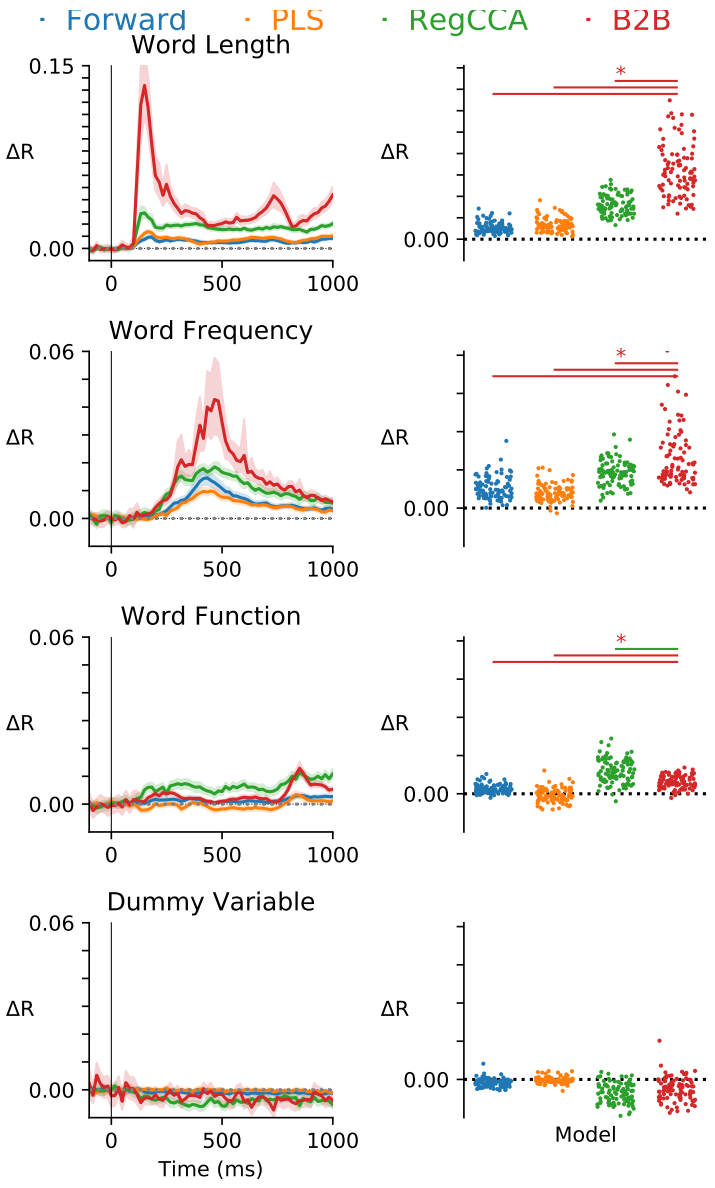

Figure 4: Multiple models (color-coded) are compared on their ability to reliably predict single-trial MEG signals evoked by words. Left. Average improvement of correlation coefficient ΔR for each of the four features (rows). Error bars indicate standard error of the mean (SEM) across subjects. Right. Average ΔR across time for each subject (dots). Top horizontal lines indicate when B2B significantly outperforms other methods (red) and vice versa.

Frequency rose from $\approx$ 100 ms and from $\approx$ 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 causally influence specific periods of brain responses to words.

To assess which model would be most sensitive to these causal discoveries, we compared B2B to other models across subjects (Figure 4 right). For Word Length B2B outperforms all models (all $p < 0.00001$) but CCA ($p = 0.0678$). For Word Frequency, B2B outperforms all models (all $p < 0.0006$). For "Word Function", B2B outperforms all models (all $p < 0.0015$). Overall, these results show that B2B reliably outperforms standard methods, especially when the effects are difficult to detect.

3.5. Magnetoencephalography data

4. Related work

Forward and cross-decomposition models have been used to identify the causal contribution of correlated factors onto multi-dimensional observations (e.g. [21]). These approaches typically lead to multiple coefficients for each features (i.e. one per dimension of Y or one per component respectively). Furthermore, these coefficients can be difficult to summarize into a single causal estimate. By contrast, B2B quickly (Fig. B.6) leads to a single unbiased scalar values $\hat{E}$ tending towards 1 and 0 for causal and non-causal features respectively.

A variety of other statistical methods applied to neuroimaging data have been proposed to clarify what is being represented in brain responses - i.e. what feature causes specific brain activity. One of the popular linear method is Representational Similarity Analysis (RSA) [17], and consists in analyzing the similarity of brain responses associated with specific categorical conditions (e.g. distinct images), by (1) fitting one-against-all classifiers on each condition and (2) testing whether these classifiers can discriminate all other conditions. The resulting confusion matrix is then analyzed in an unsupervised manner to reveal which conditions lead to similar brain activity patterns. B2B differs from RSA in that (1) it uses regressions instead of classifications, and can thus generalize to new items and new contexts and (2) it is fully supervised.

Finally, CCA has been used in neuroimaging for a variety of purposes such as denoising and subject alignment [12, 4]. While CCA relates to B2B, these two methods diverge in several ways. First, CCA and B2B 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 is symmetric between X and Y, and aims to find G and H 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^j on a single Y component specifically selected to maximize signal-to-noise ratio of that feature j. Fourth, and unlike B2B, CCA does not separately optimize two distinct regularization parameters for G and H. Finally, CCA does not use different data splits to estimate G and H. Together, these differences may explain why B2B reliably outperform CCA on estimating causal influences (Figs. 2 and B.5).

5. Conclusion

In this work, we proposed Back-to-Back (B2B) regression, a linear method to disentangle confounded factors from multidimensional observations. B2B repeatedly performs two successive multidimensional regressions on independent subsets of the data: the first regression is applied on the output domain (as in backward decoding), whereas the second regression is applied on the input domain (as in forward encoding). We provided a theoretical guarantee about the consistency of B2B, and compared it to several baselines in controlled synthetic experiments. We also applied B2B to a recent brain imaging dataset, analyzing the timing of brain responses and their connection to word features. We obtained results consistent with prior work in neuroscience literature, confirming the reliability of B2B for real data analysis.

6. 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.

References

[1] 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.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

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

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

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

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

[13] 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.

[14] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, Bernhard Schölkopf, et al. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

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

[27] 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.

[28] Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016.

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

[30] 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.

[31] 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.

[32] 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.

7. Appendices

Appendix A. Appendix

Appendix A.1. Proof of consistency theorem

Proof of the theorem in 2.2:

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

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

with N centered and full rank noise.

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

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

and satisfies

$$E\hat{H} = \hat{H}$$

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

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

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

$$\hat{G} = \arg\min_G \mathbb{E}[\|X - (XE + N)FG\|^2] = \arg\min_G \mathbb{E}[\|X - XEFG + NFG\|^2]$$

Since N is centered and independent of X, we have

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

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

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

a positive quantity which reaches a minimum (zero) for

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

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

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

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

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

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

and

$$\|X - XEFZ\|^2 + \|NFZ\|^2 = \|X - XEF\hat{G}\|^2 + \|NEF\hat{G}\|^2 \leq \|X - XEF\hat{G}\|^2 + \|NF\hat{G}\|^2$$

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

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

Summarizing the above,

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

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

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

N being full rank, this yields $EF\hat{G} = F\hat{G}$.
Replacing into (A.1), and setting $H = EFG$, we have

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

Finally, $E\hat{H} = EEF\hat{G} = EF\hat{G} = \hat{H}$, since E, 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:

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

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

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:

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

Consequently, assuming that F is full rank on $Img(XE)$, 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: $X_{train} \in \mathbb{R}^{n \times d_x}$, $X_{test} \in \mathbb{R}^{n' \times d_x}$, $Y_{train} \in \mathbb{R}^{n \times d_y}$, $Y_{test} \in \mathbb{R}^{n' \times d_y}$,
Output: estimate of prediction improvement $\Delta R \in \mathbb{D}^{d_x}$.

$H, G = \text{B2B}(X_{train}, Y_{train})$;
$R_{full} = \text{corr}(X_{test}H, Y_{test}G)$;
for $i = 1, \ldots, d_x$ **do**
  $K = Id$;
  $K[i] \leftarrow 0$;
  $R_k = \text{corr}(X_{test}KH, Y_{test}G_i)$;
  $\Delta R_i = R_{full} - R_k$;
end
return ΔR

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

Algorithm 3: Forward feature importance.

Input: $X_{train} \in \mathbb{R}^{n \times d_x}$, $X_{test} \in \mathbb{R}^{n' \times d_x}$, $Y_{train} \in \mathbb{R}^{n \times d_y}$, $Y_{test} \in \mathbb{R}^{n' \times d_y}$,
Output: estimate of prediction improvement $\Delta R \in \mathbb{D}^{d_x, d_y}$.

$H = \text{LinearRegression}(X_{train}, Y_{train})$ $R_{full} = \text{corr}(X_{test}K, Y_{test})$;
for $i = 1, \ldots, d_x$ **do**
  $K = Id$;
  $K[i] \leftarrow 0$;
  $R_k = \text{corr}(X_{test}KH, Y_{test})$;
  $\Delta R_i = R_{full} - R_k$;
end
return ΔR

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

Algorithm 4: CCA and PLS feature importance.

Input: $X_{train} \in \mathbb{R}^{n \times d_x}$, $X_{test} \in \mathbb{R}^{n' \times d_x}$, $Y_{train} \in \mathbb{R}^{n \times d_y}$, $Y_{test} \in \mathbb{R}^{n' \times d_y}$,
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 ΔR

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

Appendix A.4. Recovering E

In case of noise, B2B yields non binary $\hat{E}$. 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 $\Sigma_{X_1X_1}$ and $\Sigma_{N_1N_1}$.

The average value μ of non-zero coefficients of $\text{diag}(\hat{H})$ 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 MEG study, because signal-to-noise ratio is unknown.

Second, $\text{diag}(\hat{H})$ can be binarized with the Sonquist-Morgan criterion [20], 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 [15]:

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

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

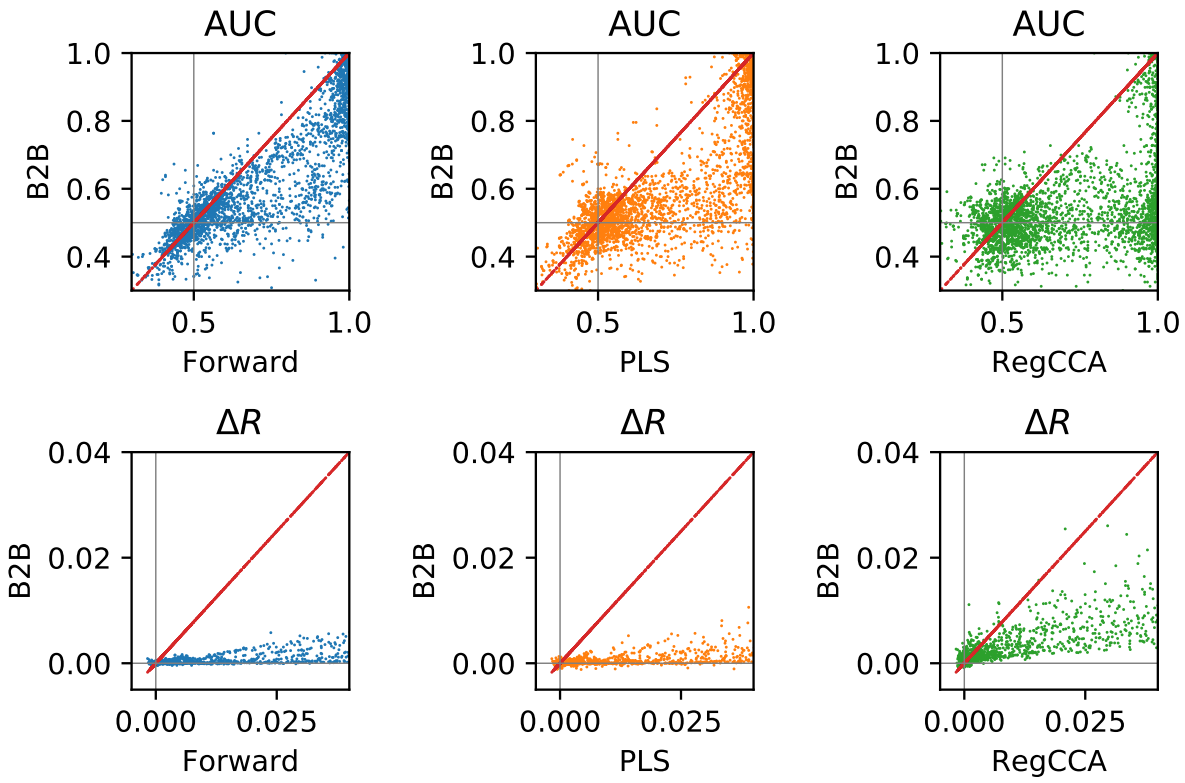

Figure B.5: Synthetic experiments. Distribution (over conditions) of AUC (top) and Feature Importance ΔR (bottom) metrics between our method (y-axis) and the baselines (x-axis). Each dot is a distinct synthetic experiment. Dots below the diagonal indicates that B2B outperform the tested model.

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. 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 [11]. The results shown in Figure B.7, show that the feature importance of ad-hoc features as derived by B2B remain unchanged and are actually improved.

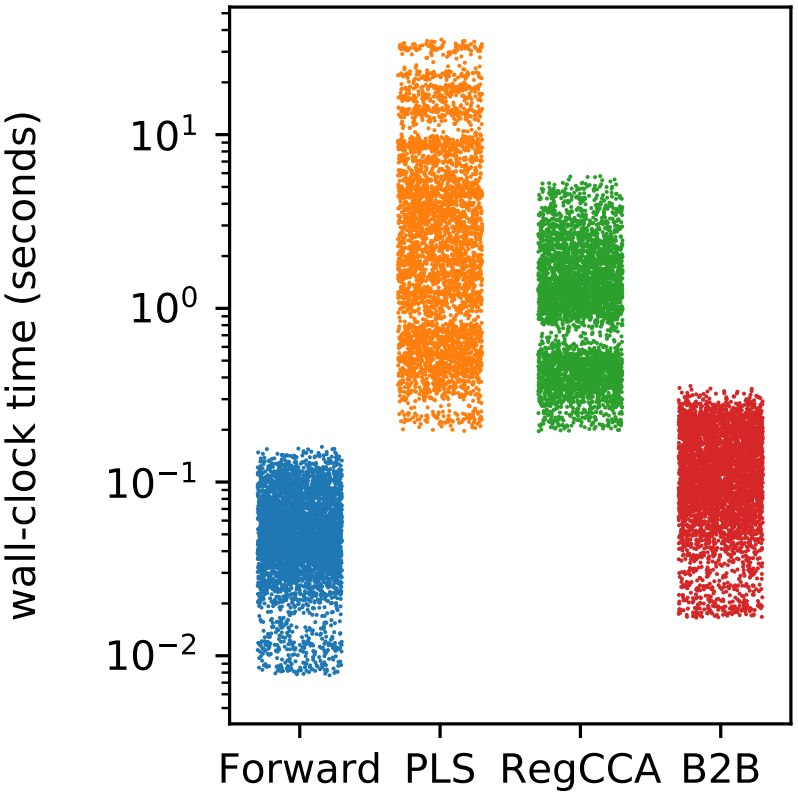

Figure B.6: Wall-clock run-time for our method B2B and for the baselines. Each dot is a distinct synthetic experiment. B2B runs much faster than cross-decomposition baselines.

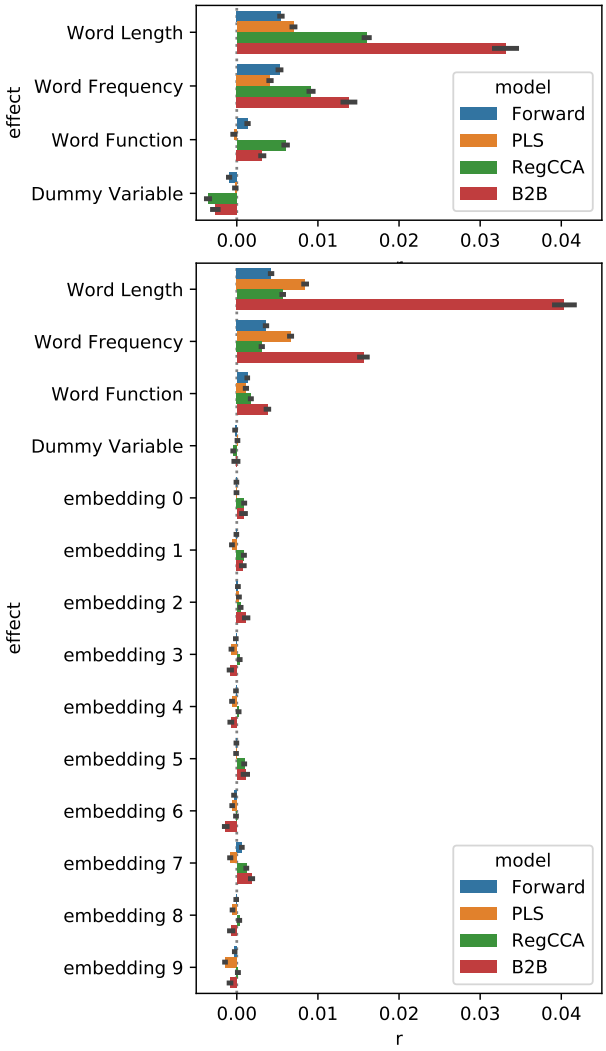

Figure B.7: Comparison of ΔR when the models are tested on four variables (top) and when the models are tested on an these four variables as well as another 10 word-embedding features (bottom). These results illustrate that, unlike Regularized CCA, B2B remains robust even when the number of tested factors increases.