MEASURING CAUSAL INFLUENCE WITH BACK-TO-BACK REGRESSION: THE LINEAR CASE

Anonymous authors
Paper under double-blind review

ABSTRACT

Identifying causes from observations can be particularly challenging when i) potential factors are difficult to manipulate individually and ii) observations are complex and multi-dimensional. To address this issue, we introduce "Back-to-Back" regression (B2B), a method designed to efficiently measure, from a set of co-varying factors, the causal influences that most plausibly account for multidimensional observations. After proving the consistency of B2B and its links to other linear approaches, we show that our method outperforms least-squares regression and cross-decomposition techniques (e.g. canonical correlation analysis and partial least squares) on causal identification. Finally, we apply B2B to neuroimaging recordings of 102 subjects reading word sequences. The results show that the early and late brain representations, caused by low- and high-level word features respectively, are more reliably detected with B2B than with other standard techniques.

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 collinear. In physics, for example, one may be challenged to identify whether fusion is caused by a change in temperature or a change in pressure, as these two factors may, at first, be difficult to manipulate independently. This issue becomes increasingly pronounced as the number of potential factors increases. In neuroscience, for example, identifying whether the frequency of a word presented on a subject's retina modulates brain activity can be surprisingly difficult. Indeed, the frequency of words in natural language covaries with other factors such as their length (short words are more frequent than long words), their categories (determinants are more frequent than adverbs) and so forth (Kutas & Federmeier, 2011; Pegado et al., 2014). Instead of selecting a set of words that control for all of these factors simultaneously, it is thus common to use forward modeling, i.e. to train a model to predict observations (e.g. brain activity) from a minimal combination of competing factors (e.g. word length, word frequency, e.g. (Huth et al., 2016)), and investigate, in the model, the estimated contribution of each factor (Friston et al., 1994).

The second challenge to measuring causal influence is that observations can be large and complex. 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 sensors simultaneously. As multiple sensors may be affected by common noise sources, it is common to use 'decoding' techniques, by, for example, fitting a support vector machine across multiple sensors to predict the category of a stimulus (Cichy et al., 2014; Kriegeskorte et al., 2008; Norman et al., 2006).

Both forward and backward modeling have competing benefits and drawbacks. Specifically, forward modeling disentangles the independent contribution of collinear factors, but does not combine multidimensional observations. By contrast, backward modeling combines multiple observations, but does not disentangle collinear factors Weichwald et al. (2015); Hebart & Baker (2018); King et al. (2018). To combine 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) (de Cheveigne et al., 2019). CCA and PLS aim to find, from two sets of data X and Y, the components H and G were XH and YG are maximially correlated or maximally covarying respectively. Because CCA and PLS are are based on a generalized eigen decomposition,

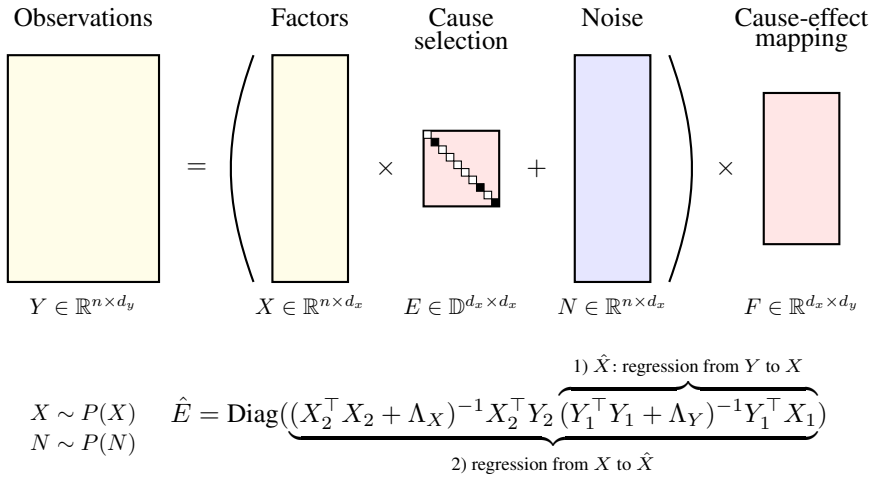

Figure 1: Back-to-back regression identifies the subset of factors $E_{ii} = 1$ in X that influence some observations Y by 1) regressing from Y to X to obtain $\hat{X}$, and ii) returning the diagonal of the regression coefficients from X to $\hat{X}$.

their resulting coefficients are mixing the features of X and Y in a way that makes them notoriously difficult to interpret (Lebart et al., 1995).

Here, we introduce the 'back-to-back regression' (B2B), which not only combines the benefits of forward and backward modeling (Section 2), but also provide robust and interpretable causal coefficients. After detailing the 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 identifying causal influence (Section 3.1). Finally, we apply B2B to a large neuroimaging dataset and reveal that distinct but collinear word features lead to distinguishable brain representations (Section 3.5).

2 BACK-TO-BACK REGRESSION

We consider the measurement of multivariate signal Y, generated from a set of putative causes X, via some unknown linear apparatus F. Not all the variables in X exert a causal influence on Y. By considering a square binary diagonal matrix of *causal influences* E, we denote by XE the causal factors of Y. In summary, the problem can be formalized as:

$$Y = (XE + N)F, \tag{1}$$

where N is a centered, unobserved noise variable. 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 (1), **the goal** of Back-to-Back Regression (B2B) is to estimate the matrix of causal influences E.

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 $\hat{E} = \text{Diag}(\hat{H})$, is the desired estimate of the causal influence matrix E.

If using regularized least-squares, arguably the most commonly employed linear regression technique (Hoerl, 1959; Rifkin & Lippert, 2007), the solution of 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 (Efron, 1992). To accelerate this ensembling procedure, we implemented an efficient leave-one-out cross-validation scheme as detailed in (Rifkin & Lippert, 2007) , as follows:

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

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 $X \in \mathbb{R}^{n \times d_x}$, 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**
  $(X, Y) \leftarrow$ 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$(X_2, Y_2 \hat{G})$; ▷ $\hat{H} = (X_2^\top X_2 + \Lambda_X)^{-1} X_2^\top Y_2 \hat{G}$
  $\hat{E} \leftarrow \hat{E} +$ Diag$(\hat{H})$;
end
$\hat{E} \leftarrow \hat{E}/m$;
$\hat{W} \leftarrow$ 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 centred and full rank noise. If F and X are full-rank on $Img(E)$, 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 A.1. □

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

Assuming, without loss of generality, that the active features are the k first, we have

$$X^\top X = \begin{pmatrix} \Sigma_{X1} & \Sigma_{X2} \\ \Sigma_{X2} & \Sigma_{X3} \end{pmatrix}, \qquad N^\top N = \begin{pmatrix} \Sigma_{N1} & \Sigma_{N2} \\ \Sigma_{N2} & \Sigma_{N3} \end{pmatrix}, \tag{7}$$

$$\hat{H} = \begin{pmatrix} (\Sigma_{X1} + \Sigma_{N1})^{-1}\Sigma_{X1} & (\Sigma_{X1} + \Sigma_{N1})^{-1}\Sigma_{X2} \\ 0 & 0 \end{pmatrix}, \tag{8}$$

$$Diag_k(\hat{H}) = Diag((\Sigma_{X1} + \Sigma_{N1})^{-1}\Sigma_{X1}) = Diag((I + \Sigma_{X1}^{-1}\Sigma_{N1})^{-1}). \tag{9}$$

In the absence of noise, we have $Diag(\hat{H}) = Diag(E)$. Otherwise the k first diagonal elements of $\hat{H}$ are all positive, and bounded by $\frac{\sigma_{X_k}}{\sigma_{X_k}+\sigma_{N_1}}$ and $\frac{\sigma_{X_1}}{\sigma_{X_1}+\sigma_{N_k}}$, where σ_{X_1}, σ_{X_k}, σ_{N_1} and σ_{N_k} denote the largest and smallest eigenvalues of Σ_{X1} and Σ_{N1}.

The average value of the coefficients of $Diag(\hat{H})$ is the trace of $\hat{H}$ divided by k. On average, since X and N are not correlated by definition, $\Sigma_{X1}^{-1}\Sigma_{N1}$ should have a low condition number, and $\frac{Var(X)}{Var(X)+Var(N)}$ provides a good estimate of $\frac{1}{k}Tr(\hat{H})$.

3 EXPERIMENTS

We perform two 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 CCA, PLS, and RidgeCV objects to compare B2B against the standard baselines (Pedregosa et al., 2011).

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 $Y = (\text{snr} \cdot XE + N)F$, where snr is a scalar that modulates the signal-to-noise ratio. Here, $F \in \mathbb{R}^{d_x \times d_y}$ contains entries drawn from $\mathcal{N}(0, d_x^{-1})$, $X \in \mathbb{R}^{n \times d_x}$ 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 $\mathcal{N}(0, d_x^{-1})$, $\Sigma_N = BB^\top$ where $B \in \mathbb{R}^{d_x \times d_x}$ contains entries drawn from $\mathcal{N}(0, d_x^{-1})$, and the factor $\text{snr} \in (0, \infty)$.

To simulate a wide range of experimental conditions, we sample 10 values in log-space for $d_x, d_y \in [10, 100]$, $n_c \in [3, 63]$, $\text{snr} \in [0.001, 10]$. 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 (Pedregosa et al., 2011):

3.2 BASELINE MODELS

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

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

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

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

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

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

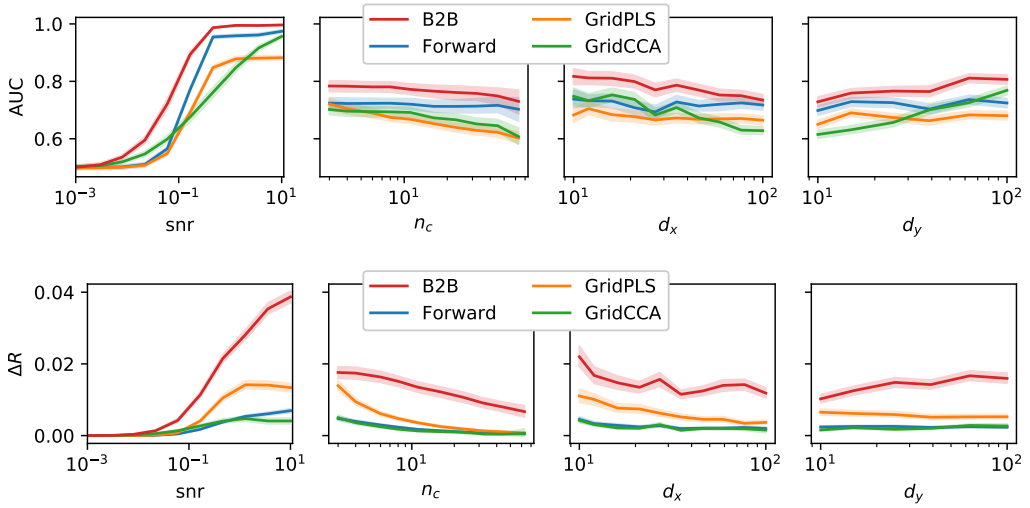

Figure 2: Synthetic experiments. Average AUC (top) and Feature Importance ΔR (bottom) when varying experimental conditions individually. Higher is better. B2B compares favorably in all cases.

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

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

We employ 5-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 (Pedregosa et al., 2011).

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 6 (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 A.2 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 colinearity is never taken into account.

We show in Figures 2 (bottom) and 6 (right, in Appendix) that our method outperforms baselines.

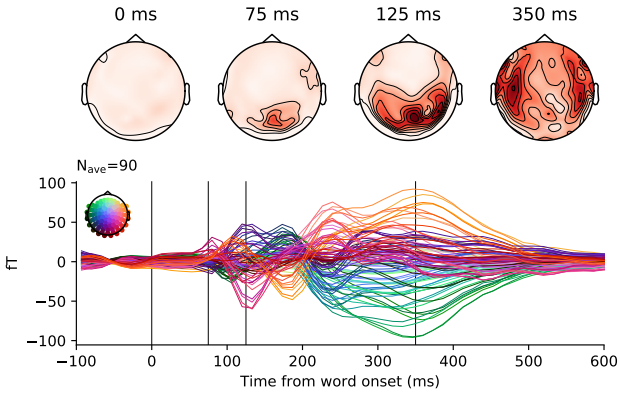

Figure 3: Ninety subjects each read approximately 2,700 words while their brain activity was recorded with MEG. Top. Average brain response to words (word onset at t=0 ms), as viewed from above the head (red= higher gradient of magnetic flux). Bottom. Each line represents a magnetometer, color-coded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

3.5 MAGNETOENCEPHALOGRAPHY DATA

Next, we apply our method to brain imaging data from the anonymized multimodal neuroimaging "Mother Of all Unification Studies" (MOUS) dataset (Schoffelen et al., 2019). The dataset contains magneto-encephalography (MEG) recordings of 102 healthy native-Dutch adults who participated in a 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.

3.5.1 MEG PREPROCESSING

The raw MEG data was bandpass-filtered between 0.1 and 40Hz using MNE-Python default parameters (Gramfort et al., 2013; 2014). 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 $n \approx 2700$ 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.

3.5.2 FEATURE DEFINITION

We aim to identify the word features that cause a variation in brain responses. We consider four distinct but colinear 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 (i.e. from 100 ms after stimulus onset) elicited by the retinotopically-tuned visual cortices (e.g. (Pegado et al., 2014).). Second, 'Word Frequency' indexes how frequently each word appears in Dutch and was derived with the the Zipf logarithmic scale of (Van Heuven et al., 2014) provided by the WordFreq package (Speer et al., 2018). 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 (Kutas & Federmeier, 2011). 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 (Honnibal & Montani, 2017). To our knowledge, this feature has not been thouroughly 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 $n \approx 2700$ 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.

3.5.3 Models and statistics

We compare B2B to four standard methods: Forward regression, Backward regression, CCA and PLS, as implemented in scikit-learn (Pedregosa et al., 2011), 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 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 3.5.3, and p-values are reported below.

3.5.4 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 evoked responses to words.

Supplementary Figure 5 shows that the Backward model decodes the dummy variable well above chance level. In addition, 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 (Kutas & Federmeier, 2011). These results illustrate that backward modelling cannot be used to estimate the causal contribution of collinear 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 3.5.3 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 $\approx$ 100 ms and from $\approx$ 400 ms respectively. Furthermore, Word Function improved the prediction performance of all models (all $p < 0.0002$) except for

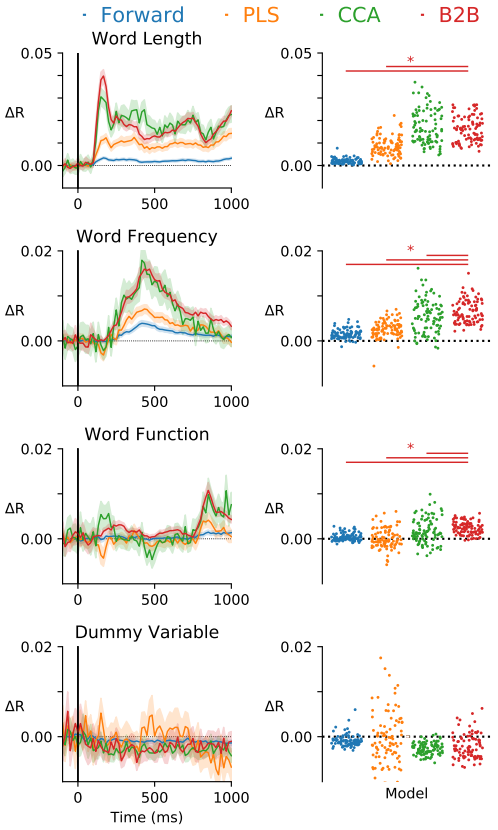

Figure 4: Multiple models (color-coded) are compared on their ability to reliably predict single-trial MEG signals evoked by words. Left. Average improvement of correlation coefficient ΔR for each of the four features (rows). Error bars indicate standard error of the mean (SEM) across subjects. Right.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

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 3.5.3 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.

4 Related work

Forward and cross-decomposition (CCA, PLS) models are regularly used to identify the causal contribution of collinear features onto multi-dimensional observations (e.g. (Naselaris et al., 2011)). 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. 7) 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) (Kriegeskorte et al., 2008), 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.

Furthermore, B2B closely relates to CCA (Hotelling, 1936), but departs from it in four main aspects. First, 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. Second, 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. Third, CCA does not typically optimize two distinct regularization parameters. 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 6). Extensions of CCA have been explored in neuroimaging (see (de Cheveigne et al., 2019) for brief review), but typically do not address the issue of causal discovery.

Finally, B2B is also related to, but different from causal discovery algorithms used outside the neuroimaging field. These methods are often nonlinear (Peters et al., 2017) and base their causal estimates on temporal delays (Granger, 1969; Janzing et al., 2013). For example, B2B relates to tthe half-sibling regression (Schölkopf et al., 2016), in the sense that we observe a large number of variables corrupted by the same source of noise; we leave the extension of B2B to nonlinear time-series for future work.

5 Conclusion

In this work, we proposed Back-to-Back (B2B) regression, a linear method to measure the causal influence of a potential set of variables generating multidimensional observations. B2B performs two successive multidimensional regressions: one from the output domain, and another one from the input domain. 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 efficacy of B2B for real data analysis.

REFERENCES

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

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.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593. Springer, 1992.

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.

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.

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.

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

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

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

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

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

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.

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

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.

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

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.

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

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

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.

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.

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.

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

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

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.

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.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. Luminosoinsight/wordfreq: v2.2, October 2018. URL `https://doi.org/10.5281/zenodo.1443582`.

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.

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.

A APPENDIX

A.1 PROOF OF CONSISTENCY THEOREM

Proof of the theorem in 2.2:

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

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

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

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

Let us prove that $EF\hat{G} = F\hat{G}$, that is, that the d_x bottom rows of $F\hat{G}$ are zero. Let $Z = F^\dagger EF\hat{G}$. We have $FZ = FF^\dagger EF\hat{G} = EF\hat{G}$ ($FF^\dagger E = E$ since F is full rank on $Img(E)$). Since E is a contraction, we have $\|NEF\hat{G}\|^2 \leq \|NF\hat{G}\|^2$. Therefore,

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

But as $\hat{G} = \arg\min_G \|X - XEFG\|^2 + \|NFG\|^2$, the above inequality is an equality, $Z = \hat{G}$ and $EF\hat{G} = F\hat{G}$. Therefore, $\|X - XEFG\|^2 + \|NFG\|^2 = \|X - XEFG\|^2 + \|NEFG\|^2 = \|X - XH\|^2 + \|NH\|^2$ is minimised by $\hat{G}$ and $\hat{H}$. Finally, $E\hat{H} = EEF\hat{G} = EF\hat{G} = \hat{H}$. This completes the proof. □

A.2 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**
  $K = Id$;
  $K[i] \leftarrow 0$;
  $H_k = \text{LinearRegression}(X_{train}K, Y_{train}G_i)$;
  $R_k = \text{corr}(X_{test}H_k, Y_{test}G_i)$;
  $\Delta R_i = R_{full} - R_k$;
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 $\Delta R \in \mathbb{D}^{d_x, d_y}$.

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

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

B ADDITIONAL FIGURES

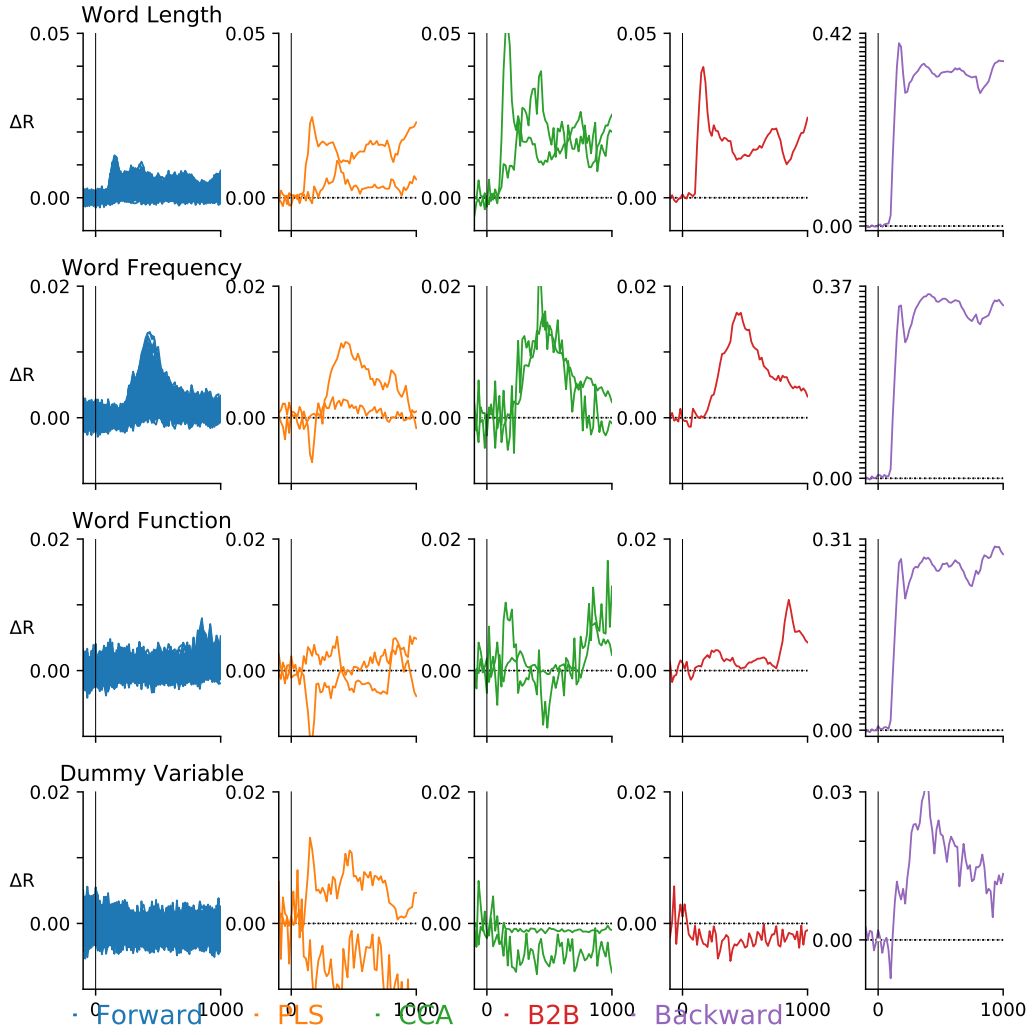

Figure 5: Average ΔR for each Y dimension (Forward model), canonical components (PLS and CCA) or feature (B2B) across subjects. Formally, Backward model cannot have a ΔR because it never unmixes the multiple X features. In such cases, we thus simply report the decoding score.

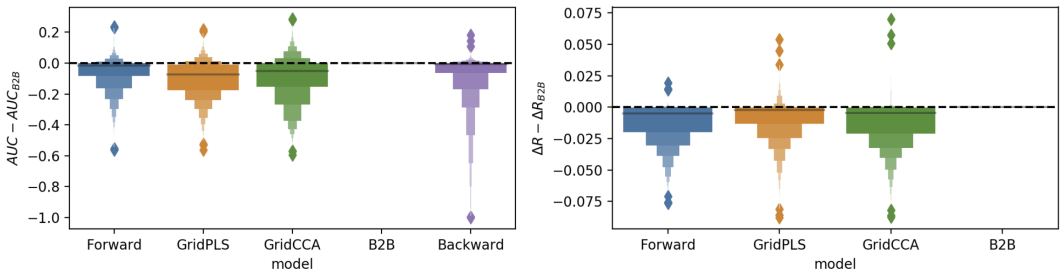

Figure 6: Synthetic experiments. Distribution (over conditions) of differences in AUC (left) and Feature Importance ΔR (right) metrics between our method and the baselines. Mass under the dashed line corresponds to B2B outperforming the baseline.

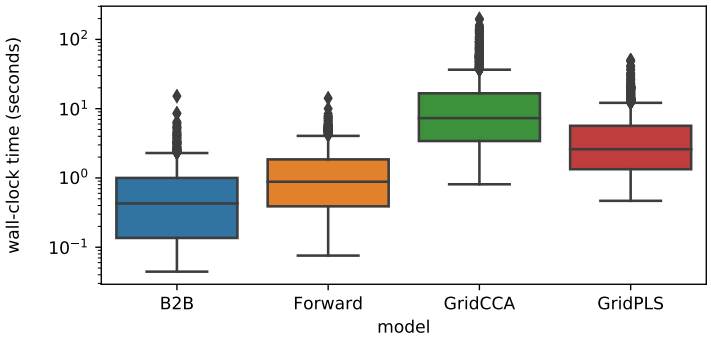

Figure 7: Wall-clock run-time for our method B2B and for the baselines. We see that B2B runs much faster than baselines.