Momentum: a new approach to causality in functional Magnetic Resonance Imaging

Natalia Bielczyk\textsuperscript{1,2}, Alberto Arenas\textsuperscript{1,3}, Jan Buitelaar\textsuperscript{1,2}, Jeffrey Glennon\textsuperscript{1,2}, Christian Beckmann\textsuperscript{1,2,4}

\textsuperscript{1}Donders Institute for Brain, Cognition and Behavior, Nijmegen, the Netherlands
\textsuperscript{2}Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands
\textsuperscript{3}Radboud University Nijmegen, Nijmegen, the Netherlands
\textsuperscript{4}FMRIB Centre, University of Oxford, Oxford, United Kingdom

fMRI datasets have always been an enfant terrible to the causal research. This is for a few independent reasons. Firstly, human haemodynamics is delayed with respect to the neuronal activity and acts like a lowpass filter to the underlying neuronal dynamics. Secondly, background activity in neuronal networks is a confound to causal research. Thirdly, signal to noise ratio is low in fMRI in general, and varies between brain regions. In this work, we address all these. We propose Momentum: a nonparammetric, cost-efficient method for finding causal connections from fMRI data by utilizing all moments of the BOLD distributions and combining them into cumulants. Momentum is a classifier informed by Dynamic Causal Modeling used as a forward model for BOLD fMRI.

First, we evaluated performance of Momentum on synthetic benchmark, low-noise datasets. Then we moved towards DCM forward simulations that better represent natural confounds in the fMRI datasets such as varying haemodynamic lags, signal magnitudes and noise vari-
ances between the upstream and downstream regions. Momentum clearly outperforms all the existing methods on the benchmark datasets, but is also far more resilient to all these confounds. This suggests that Momentum might be further validated on the human fMRI datasets.

Furthermore, we suggest that the Momentum as a methodology for causal inference can be used across multiple disciplines beyond fMRI research, and beyond neuroscience. This inference procedure can be adapted for every research problem for which a forward model faithfully representing the data can be formulated.

1 Introduction

Limitations of fMRI data fMRI datasets have always been an enfant terrible to the causal research, for a number of reasons. Firstly, human haemodynamics is delayed with respect to the neuronal activity and acts like a lowpass filter to the underlying neuronal dynamics. It is also region-specific and subject-specific, for which it is hard to marginalize it out from the inference procedure. Furthermore, data acquisition is a few orders of magnitude slower than the intrinsic brain dynamics. In a consequence, time order of the fMRI samples yields little information about the causal dependencies between brain regions.

Secondly, the background activity in neuronal networks is a confound to causal research and its impact on the results should be reduced to a minimum by using functional parcellation. However, brain parcellation for causality problem is always a trade-off between statistical validity and interpretability. On one hand, the most popular, anatomical region definition is suboptimal for
causal discovery in fMRI from statistical point of view and will probably return lower test-retest results than the functional parcellation, but on the other hand, functional parcellation is hard to interpret in terms of biological relevance.

Thirdly, the signal to noise ratio in brain networks is hard to estimate and most probably, it varies between brain regions. Therefore, every quantity sensitive to the varying levels of the signal to noise ratio (such as skewness of the distribution) will fail to correctly estimate the directionality in connectivity between brain regions. All these obstacles are reflected by the fact that conventional methods give a poor performance on the synthetic benchmark fMRI datasets\textsuperscript{1}.

**Theoretical studies on causal inference in fMRI** The aforementioned intrinsic properties of the fMRI data contributed to the bad press of the whole concept of causal inference in fMRI. In a highly cited computational study by Smith et al.\textsuperscript{1}, synthetic data created with the Dynamic Causal Modeling forward model\textsuperscript{2} were employed to test the utility of popular methods for causal inference in fMRI. Tested methods were performing almost equally poor and close to the chance level, across multiple simulations involving a variety of connectivity patterns, haemodynamic lags, TRs and session lengths.

The methods evaluated in this study include:

1. Granger Causality (GC\textsuperscript{3,4}): infers causality between a pair of time series, assuming that both of them can be expressed as autoregressive processes, which inevitably involves a time lag.
   In order to infer if \( X \) causes \( Y \), we fit an autoregressive model to time series \( Y \) twice, first
with and then without including time series $X$ as a regressor. If the variance of the residual noise obtained from fitting the autoregressive model significantly drops after we add $X$ to the account, we can infer a causal influence from $X$ to $Y$.

2. Partial Directed Coherence (PDC): an equivalent to Granger causality in frequency domain

3. Patel’s tau (PT): based on two-stage inference procedure. In the first step, connections are spotted by means of functional connectivity. In the second step, the directionality of the connections ($X \rightarrow Y$ or $Y \rightarrow X$) is determined on the basis of the difference between conditional probabilities $P(X|Y)$ and $P(Y|X)$

Following Patel’s insight, Hyvärinen and Smith recently proposed Pairwise Likelihood Ratios approach (PW-LR): a two-stage approach to obtain a directed connectivity from a set of BOLD fMRI time series: (1) find the existing connections with partial correlation and (2) for every connection found in the previous step, apply a pairwise analysis in order to solve the two-node classification problem (i.e., distinguish between two causal models $X \rightarrow Y$ and $Y \rightarrow X$ which corresponds to the LiNGAM model for two variables). The authors compared the likelihood of these two models derived for LiNGAM and provided with a cumulant based approximation to the likelihoods ratio.

**Momentum** In this work, we attempt to address all the issues with fMRI data mentioned in the previous section. We keep the same two stage inference procedure as Hyvärinen. However, we propose and advance to the second step of the inference procedure. **Momentum** is a nonparametric, cost-efficient method for finding causal connections from fMRI data by utilizing all moments
of the BOLD distributions and combining them into cumulants. Momentum is informed by Dynamic Causal Modeling (DCM) used as a forward model for BOLD fMRI.

We evaluated performance of Momentum on synthetic benchmark datasets as well as other forward simulations in which we varied the haemodynamic lags, signal magnitudes and noise variance between the upstream and the downstream regions. We also compare the performance of Momentum against all the aforementioned methods. Momentum clearly outperforms other methods on the benchmark datasets but is also far more resilient to the variability in background noise levels and haemodynamic lags. This suggests that it might be further tested on human fMRI datasets.

Furthermore, we suggest that, as a methodology for causal inference, Momentum can be used across multiple disciplines beyond fMRI research, and beyond neuroscience. This inference procedure can be adapted for every research problem for which a forward model can be formulated.

Methods

’Momentum’ for the two-node classification problem In this work, we keep the same two-stage scheme for the causal discovery as Hyvärinen and Smith. However, we improve from the previously introduced methods for solving a two-node classification problem by introducing ‘Momentum’: a novel methodology for pairwise causal analysis utilizing all moments of the BOLD fMRI distributions, and combining them into cumulants.

Fig. presents such a two-node classification problem. In this problem, one region (’up-
stream”) is sending information to another region (‘downstream’) through a connection of a weight \( w \). Both regions receive region-specific background neuronal noise \( e_i(t) \) as well as a signal \( s_i(t) \), modelled simply as trains of 'on' and 'off' states. This signal can both relate to experimental input, as well as input from other regions, projecting to this particular node.

Figure 1: Two node directed network. The upstream region sends a projection of a weight \( w \) to the downstream region. Both regions receive inputs \( s_i(t) \), related to cognition and/or projections from exterior regions. Both regions also receive a background noise of a certain magnitude \( e_i(t) \). The proportion between the amplitude of \( s_i(t) \) and the variance of the noise \( e_i(t) \) defines the SNR in the network.

We can simulate the behavior of the simple two-node networks (Fig. 1) using the he Dynamical Causal Modeling\(^2\). This golden-standard model in fMRI research expresses the beliefs upon how the BOLD response is generated from neuronal networks in the brain: as a fast neuronal dynamics convolved with a slow haemodynamic response. The details of the model and the chosen parameters are described in Supplementary Materials\(^4\). Once we simulate the behavior
of a variety of such little two node networks, we can empirically quantify statistical differences in
the moments of the BOLD distributions between the upstream and downstream regions, and build
a classifier on that basis.

Let us build a probability density function from the distribution of BOLD values. If the
moment generating function of a random variable has positive radius of convergence, then that
probability distribution is determined by its moments:

\[ M_k = \frac{1}{N\sigma^k} \sum_{i=1}^{N} \tilde{x}_i^k \]  

where \( k \geq 0 \), \( \tilde{x} \) - normalized time series, \( \sigma \) - standard deviation of the time series, \( N \) - length of
the time series.

The experimental data is a time series of a length of an order of hundreds, or thousands at
best. For such a short time series, estimation of the higher order moments becomes very inaccu-
rate. In this work, we propose a trick to fully exploit the information contained in the moments of
the distribution without falling into the high moments regime. Let us consider a normalized time
series so that \( \sigma = 1 \), and take into account not only moments of an integer order (the mean for
\( k = 1 \), the variance for \( k = 2 \), the skewness for \( k = 3 \), the kurtosis for \( k = 4 \), etc.), but also
interleaved moments of fractional orders. Since the time series is normalized, it contains negative
values and therefore, the 'fractional moments’ will become complex. Since the Eq. is continuous
\(^1\) for moments of order \( k \) in the time series \( x(t) \), the noise \( \epsilon(t) \) added to the time series also scales to the power of
\( k \), namely as \( (x + \epsilon)^k \). The 'noise’ in this case may be caused not only by the noisy background neuronal activity, but
also by experimental setup such as low time resolution (TR) or short session duration
with respect to $k$, these complex moments will form a curve in the complex plane.

Figure 2: Complex moments for the BOLD fMRI time series in a simulated 2-node network, in the noiseless case. Blue: upstream region. Red: downstream region. The ‘Momentum’ curve starts from $(1, 0)$ for $k = 0$, traverses the upper half-plane and arrives at $(0, 0)$ for $k = 1$. It travels back through the lower half-plane towards $(1, 0)$ for $k = 2$ since the variance is equal to 1. Every time $k$ becomes an integer, the curve crosses the real axis.

In Fig. 2 we show phase diagram for all moments in range $k \in [0.0, 5.0]$, for a long simulated BOLD fMRI time series generated in a simple two-node network in the noiseless case. The moments are computed separately for the upstream region (blue) and the downstream region (red). The curve starts at $(1, 0)$ for $k = 0$. Then it traverses the upper half-plane and arrives at $(0, 0)$ for $k = 1$. Subsequently, it goes back through the lower half-plane and comes back to $(1, 0)$ for $k = 2$ since the variance is equal to 1. Every time $k$ becomes an integer, the curve crosses the real axis. The imaginary axis describes the behavior of the left half of the BOLD distribution, since fractional moments give nonzero imaginary part for negative values of the BOLD.

Furthermore, for two time series $x(t)$, and $y(t)$, not only the sole fractional moments but
also the *asymmetry* between the moments can indicate the directionality of a connection. This asymmetry can be quantified by ‘fractional cumulants’:

\[
C_{kl} = \frac{1}{N} \sum_{i=1}^{N} (\tilde{x}_i^k \tilde{y}_i^l - \tilde{y}_i^k \tilde{x}_i^l)
\]

where \( k, l \in \{0, 0.1, 0.2, ...\} \).

In order to investigate how the cumulants differ between upstream and downstream region in our two node problem, we have run a 1,000 noiseless 2-node DCM simulations for a duration of 10 min. In order to marginalize out the haemodynamic parameters from our results, we sampled the parameters independently for the two nodes, and from the experimentally found distributions given in Friston et al. In order to marginalize out the effect of different input strengths and frequencies, we also sampled the input magnitudes and frequencies (probabilities of switch from on- to off-state and vice versa) from a Gamma distribution with mean ad variance of 1. The input signals driving the upstream and the downstream region were also sampled independently from each other, as trains on- and off-states governed by Poissonian processes. Since we aimed to investigate the ideal case, the background noise was set to 0. In order to obtain a large amount of synthetic data, we also did not subsample our synthetic BOLD every 2 – 3[s] as is typically done for the synthetic BOLD fMRI.

We performed this simulation twofold. Firstly, we fed in an empty connection in order to create a null distribution. Secondly, we added a connection with a weight of 0.9. In Fig. 3, we show the mean value over all 1,000 simulations, for all cumulants of indexes \( k, l \in [0.0, 5.0] \), both
in Cartesian and polar coordinates (centered at (0, 0) and not (1, 0), otherwise we would lose the symmetry). The patterns of cumulants are interesting especially along the imaginary dimension, where they become zeros for all the integer cumulants of $k, l = (1, 1), (2, 1), ...$, but exchange signum in the intervals in between integer cumulants in a systematic way.
Figure 3: A: mean values for all cumulants over 1000 simulations. Since cumulants are antisymmetric with respect to indexing $k, l$, the heatmaps for real and imaginary component, and for the phase, are also antisymmetric. The radius is always positive on the contrary. B: zoom into a smaller range of $[0, 0.3]$. The imaginary component becomes a zero for integer cumulants of $k, l = (1, 1), (2, 1), ...$, but exchanges signum in the intervals between in a systematic way. C: signum of the cumulants. Red: positive. Blue: negative. Green: zero. We further denote signum maps for real and imaginary components as $S_r$ and $S_i$. 

The heatmaps shown in Fig. 3A, B represent indicate for a general tendency for each cumulant to have a positive or negative signum ($S_r$ and $Si$, Fig. 3C). However, they do not represent confidence intervals, or discriminability. In order to choose cumulants which can best discriminate the connection, we created distributions of cumulant values across our 1,000 simulations in the null case and compared it against the distributions derived from simulations with a connection fed into the network. We smoothed these distributions with kernel smoothing function, and for each cumulant, we computed the percentile of samples falling beyond 95th percentile of the null distribution (in case the mean for the given cumulant is negative as in Fig. 3C, we took samples falling lower than the bottom 5th percentile of the null distribution, and higher than the 95th percentile otherwise). We will further refer to this percentile as the discriminability.

The results are shown in Fig. 4. We can observe that (1) cumulants expressed in polar coordinates are less informative than in Cartesian coordinates, for two major reasons. Firstly, the radius is uninformative for the causal inference because $\text{radius}(z) = \text{radius}(-z)$ for complex $z$. Secondly, phase of cumulants is discriminative over a very small subset of the parameter subspace, and this subspace lies within the high-moments regime - which are hard to estimate from real data. Therefore from now on, we will focus on cumulants defined in Cartesian coordinates. (2) if one of the indexes $k$, $l$ equal to zero, i.e., the cumulant reduces to simple moment, it has lower discriminative value than the full cumulants. Therefore, we will disregard simple moments from further analysis and fully concentrate on the asymmetry between moments, i.e., full cumulants ($k, l > 0$). (3) interestingly, the discriminability is different for the real and imaginary component. For the real components, all the important information seems to be contained within $k + l < 2.0$ range, whereas
there is still a lot of information preserved in the imaginary cumulants of indexes $k + l < 3.1$. This is a proof of concept that moving from integer to fractional moments of the distribution provides with additional information.

Figure 4: Discriminative power for all cumulants in range $k, l \in [0.0, 5.0]$, in the ideal case of a very long BOLD time series and no background neuronal noise. Cumulants expressed in polar coordinates are less informative than in Cartesian coordinates. The phase is uninformative since the discriminative power is relatively high only in the high-moment regime - and high moments are hard to estimate on the short, experimental time series. The radius is uninformative as always positive quantity.
Which cumulants to pick for the causal inference? Which cumulants are the most informative?

In order to investigate how the performance of single cumulants extends from a two node problem to a network, we evaluated their success rate in estimating connectivity across all 28 benchmark simulations from the Smith’s study. Fig. 5 presents the grand mean success rate achieved by all the cumulants separately. The success rate for each of the simulations apart is presented in the Supplementary Materials 6. The maps of simulation-dependent success rate relate to the maps of discriminability (Fig. 4), but they are not identical and differ between simulations. This is for two reasons. Firstly, factors such as connectivity strengths, session duration, TR etc. affect the properties of the outcome BOLD fMRI data and therefore affect the causal inference. Secondly, the success rate for cumulants of high indexes \((k, l > 3)\) is lower than the discriminability scores derived in the ideal, noiseless case would suggest (Fig. 4). This is because the benchmark datasets contain subsampled and therefore short BOLD time series - and for such a short time series, estimation of high moments becomes hard.
Combining cumulants into a classifier

This variability in the success rates between cumulants suggests that only a subset of parameter space $k, l$ can be useful for causal inference. Because of the difficulties in estimation of high order moments, the discriminability of cumulants clearly depends on the joint index $k + l$, and this symmetry is also reflected in the maps of discriminability (Fig. 4). Therefore, we will be only exploring triangular combinations of cumulants $k + l \leq \text{Ind}_{max}$.

We then propose to combine information contained in multiple cumulants by building the
classifier based on a 'voting' between the cumulants. This voting aims to compare if the map of
cumulants obtained for a pair of time series \(X(t), Y(t)\) is closer to the benchmark maps shown in
Fig. 3A (which is an evidence for a connection \(X \rightarrow Y\)), or their inverse (which is an evidence
for a flipped connection \(Y \rightarrow X\)). Therefore, each of the cumulants \(C_{kl}\) votes due to signum \(S_{r_{kl}}\),
\(S_{i_{kl}}\) (Fig. 3C). If the signum of the cumulant is the same as in Fig. 3, it is one vote behind the
connection \(X \rightarrow Y\), and against this connection otherwise. Since there is the estimation problem
for cumulants for which the joint index \(k + l\) is high, we will discount their impact on the voting
by using a nonlinearity of a form \(f(x)=\log(\cosh(\max(x,0)))\). Therefore, the final classifier yields:

\[
\begin{cases} 
X \rightarrow Y & \text{if } \sum_{k,l} \left[ S_{r_{kl}} \log(\cosh(\max(zscore(\text{real}(C_{kl})),0))) \right. \\
+ \left. S_{i_{kl}} \log(\cosh(\max(zscore(\text{imag}(C_{kl})),0))) \right] \geq 0 \\
Y \rightarrow X & \text{otherwise}
\end{cases}
\]

**Choosing the optimal parameter space for the causal inference**

The question remains: which particular cumulants \(C_{kl}\) to choose for the inference? This
problem has two dimensions: (1) how does the final success rate depend on the discriminability
rate of the cumulants involved; (2) how does the maximal index \(\text{Ind}_{\text{max}}\) influence the results. Let us
first fix \(\text{Ind}_{\text{max}}\), and consider cumulants on a triangle \(k, l \geq 0.1, k + l \leq 3.1\). Let us now examine
only cumulants of discriminability exceeding a particular value. For instance, cutoff value of 0.1
means that we respect the vote from all cumulants whose discriminative value is not less than 0.1.
We then evaluate the grand mean success rate (as the mean success rate over all 28 benchmark
simulations) in the function of the threshold discriminability value. Fig. 6A, demonstrates that all
the cumulants of the nonzero discriminative value (namely, all cumulants except for $k = l$) make up for the best classifier.
Figure 6: How many cumulants to take into account? A: Grand mean success rate for real and imaginary, unweighted cumulants in range $k + l \leq 3.1$, in the function of cutoff discriminative value. The higher cutoff, the less cumulants we take into account while voting for the directionality of causal connection. The results clearly show that in order to maximize the success rate in causal inference, we should take into account all the cumulants except for the diagonal of $k = l$. B: the grand mean performance for Momentum based on cumulants of indexes $k, l$ between 0.1 and $k + l \leq \text{Ind}_{\text{max}}$, in the function of that maximal index. The optimal performance for unweighted case is achieved for $[\text{Ind}_{\text{Rmax}}, \text{Ind}_{\text{Imax}}] = (2.4, 1.7)$, and amounts to 0.835, whereas for weighted case $[\text{Ind}_{\text{Rmax}}, \text{Ind}_{\text{Imax}}] = (2.1, 3.7)$ and amounts to 0.886 - which exceeds both the grand mean performance of the 'PW-LR skew r' method (0.845) and maximal grand mean performance of any single cumulant (Fig. 5, 0.847).
Secondly, let us examine what is the optimal window $\text{Ind}_{max}$ for indexes $k, l$, and provide rationale for the weighting of cumulants with the nonlinear function. Since discriminability is generally higher for low indexes $k, l$ (Fig. 4), we will evaluate the grand mean performance for Momentum based on cumulants of indexes between 0 and a maximum $\text{Ind}_{max}$, in the function of that maximum. Since maps of discriminability differ between real and imaginary components (4), we consider the maximal indexes along real and imaginary dimension separately. The results are presented in Fig. 6 B. The optimal performance for the non-weighted classifier is achieved for $[\text{IndR}_{max}, \text{IndI}_{max}] = (2.4, 1.7)$, and amounts to 0.835. The optimal performance for the weighted classifier on the other hand is achieved for $[\text{IndR}_{max}, \text{IndI}_{max}] = (2.1, 3.7)$ and amounts to 0.886 - which slightly exceeds both the grand mean performance of the ‘PW-LR skew r’ method by Hyvärinen (0.845) and maximal grand mean performance of any single cumulant in our study (Fig. 5, 0.847). Additionally, the parameter space is smoother with respect to the grand mean performance in the weighted case, and we see more information along the imaginary axis (because the grand mean performance has a maximum further along the imaginary axis).

2 Validation with synthetic datasets

Comparison against other methods on benchmark synthetic datasets We evaluated the method on synthetic benchmark datasets and compared against the methods briefly described in the introduction:

1. simple version of Granger Causality featuring Ordinary Least Square regression with lag of 1, implemented in Multivariate Granger Causality Toolbox, obtained from
2. Partial Directed Coherence, obtained from the Extended Multivariate Autoregressive Modelling Toolbox:

http://www.science.unitn.it

3. Patel’s tau, implemented similarly as in \( \text{[1]} \) by recalculating each time series into the range \([0, 1]\), setting samples under the 10th percentile to 0, over the 90th percentile to 1, and linearly mapping the remaining samples to the range \([0, 1]\). Then, we infer the directionality of connection from the difference between \( P(X|Y) \) and \( P(Y|X) \). In addition to the previous implementation however, we also integrate the results over all the possible thresholds in order to eliminate the thresholding problem while calculating the conditional probabilities \( P(X|Y), P(Y|X) \)

4. Pairwise Likelihood Ratios methods (PW-LR), obtained from

https://www.cs.helsinki.fi/u/ahyvarin/code/pwcausal/\(^2\)

Just like Hyvärinen, we chose inverse covariance thresholded through permutation testing for the functional connectivity research in the first step of causal inference.

\(^2\)Hyvärinen utilized cumulant \( k, l = (2, 1) \) weighted with the covariance for synthetic benchmark datasets. Therefore, for this particular comparison, we use equivalent of our method weighted with covariance. The covariance between connected nodes is positive in general so that this weighting should not have the impact on the result of causal inference (Fig. \(^7\) blue dots), it can only influence z-scores of connections and therefore shape of the violins
The comparison on the simulation no 2 is presented in Fig. 7. Comparison on the rest of the benchmark simulations is presented in Supplementary Material 7. The violin plots denote the distribution of the z-scores for connections as compared to the null distribution. Blue dots denote the percentage of correct assignments for the true connections. Momentum’s performance only slightly outperforms ‘PW-LR skew’, but the benchmark dataset is based on a very low-noise simulation.

**Comparison against other methods in noisy conditions** The benchmark datasets represent a variety of experimental conditions, such as time resolution of the scanner or length of the experiment. They, however, always assume very low, temporally uncorrelated noise in the neuronal communi-
cation and equal levels of input signals between the upstream and the downstream region. In terms of the biological relevance, these are unrealistic assumptions. Therefore, we further explore the performance of the methods given the variability in the noise levels and input amplitudes.

**Impact of the background noise on Momentum’s performance**

In most computational studies using DCM forward model, the background neuronal noise \( e_i(t) \) is implemented as a white noise of the same variance across the whole network. However, this is not a representative feature since, most likely, the spectra of neurological noise in the human brain is scale-free (which we refer to as a pink noise), and the noise level might highly vary between brain regions. The noise is, in principle, every signal which we cannot explain with known regressors, therefore it may also relate to different cognitive processes and vary in one subject between experiments.

Therefore, we test the methods against each other on a two-node simulation with fixed input strengths, and a pink background neuronal noise. The amplitude of the inputs \( s_i(t) \) was fixed to 1. We vary both the the variance of the noise in the upstream and downstream region, both in range \([0,2,5.0] \), and perform 500 realisations of 10 [min] simulation for each configuration of the noise variances. Fig. 8 presents the comparison between ’Momentum’ and the four competitive methods: GC, PDC, PT and ’PW-LR r skew’. We can observe that only one method gives a performance better than chance across the whole parameter space: Momentum. It is highly resilient to the amplitude of the noise in the upstream region, and a bit less resilient to the noise amplitude in the downstream region. PT seems to be fully resilient to the downstream background noise, but not at
all to the upstream noise. 'PW-LR r skew' relatively quickly drops down to the chance level with respect to both the noise level in the upstream and downstream region, whereas GC and PDC are performing poorly under any combination of noise variances.

![Figure 8: Resilience of the methods against the background pink neuronal noise. The variance of the noise differs between upstream and downstream region, both in range [0.2, 5.0]. For each method, the element of parameter space which gives the lowest performance was marked with a square. Only two methods give a performance better than chance across the whole parameter space: Patel’s tau and Momentum.](image)

**Impact of the varying signals amplitudes on Momentum’s performance**

In the original version of the DCM procedure\[2\], as well as in most computational studies, equal stimulus strengths $s_i(t)$ are also assumed. This is, again, an unrealistic assumption since there is no reason to believe that different brain regions receive equally strong signals. This time, we fixed noise variance to zero, and we were varying signal strengths in the upstream and down-
stream region. We perform 500 realisations of 10[\text{min}] simulation for each configuration of the input strengths, both in range [0.2, 5.0]. Fig. 9 presents the comparison between Momentum and the competitors. Again, Momentum is the only method whose performance does not fall towards the chance level within the parameter space. Patel’s tau is even more fragile than ’PW-LR r skew’ whereas GC and PDC give performance about the chance level across the whole parameter space.

Figure 9: Resilience of the methods against the variability in signal strengths. The variance of the signal differs between upstream and downstream region, both in range [0.2, 5.0]. For each method, the element of parameter space which gives the lowest performance was marked with a square. Again, momentum is the only method whose performance does not fall towards the chance level within the parameter space. Patel’s tau is equally fragile as ’PW-LR r skew’ in this case. GC and PDC give performance on the chance level across the whole parameter space.
3 Discussion

Last remarks Momentum is following recent drift in the field of causal research in fMRI towards data driven methods which do not include regression in time but focus on the properties of BOLD distributions instead, as proposed in Patel’s tau and Pairwise Likelihoods methodology. Momentum attempts to extract all the available information contained in the data by combining multiple features on the distribution (i.e., moments) together. Our evaluation on synthetic datasets in highly noisy regime demonstrates that Momentum is resilient to all the main confounding factors in fMRI data: the haemodynamic variability, the noise variability and the input variability. It can therefore be recommended for the whole brain research.

'Momentum’ is based on the same forward model as the DCM. However, using benchmark maps of cumulants derived from forward simulations, as opposed to fitting the full DCM model to the data, allows for marginalizing out all the unimportant parameters. In the DCM, we fit all the parameters of the model at once even though we might not be interested in some of them (such as region-specific haemodynamic parameters or variance of the noise in the nodes of the network). Since there is no way of marginalising out parameters different than connectivity strengths in the DCM inference procedure, we need a lot of data in order to obtain the parameters of interest (among others). In ’Momentum’, this issue does not exist since the benchmark maps are derived from multiple forward simulations with parameters samples from experimentally informed distributions. Therefore, 'Momentum' procedure can fully focus on classifying a pair of regions into upstream and downstream.
One remark to make is that this method is meant to retrieve the net connectivity. Namely, once two brain regions speak to each other, it will retrieve the stronger out of the two connections. Since we sort out connections in the first step of the causal inference, we can either interpret ambiguous output of the voting as bidirectional connection, or withdraw it from the results. One interesting direction for the method development would be also classification between excitatory and inhibitory connectivity, which is missing in the current version of 'Momentum'. Since inhibition is difficult to infer from the time series in general, and the vast majority of the connections in the macroscopic scale are excitatory, we left this issue for the future developments of 'Momentum'.

Furthermore, since the discriminative power for single cumulants changes smoothly along dimensions $k, l$ (Fig. 4), we did not increase the granularity of fractional moments to less than 0.1. Choosing the optimal resolution is a material for further investigation, although we believe that increasing index resolution to substantially less than 0.1 would not be beneficial: what it would add to the table would be highly correlated cumulants at a high computational cost.

References

1. Smith, S. et al. Network modelling methods for fmri. *NeuroIMAGE* **54**, 875–91 (2011).

2. Friston, K. J., Harrison, L. & Penny, W. Dynamic causal modeling. *NeuroIMAGE* **19**, 1273–302 (2003).

3. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrika* **37**, 424–38 (1969).
4. Seth, A. K., Barrett, A. B. & Barnett, L. Granger causality analysis in neuroscience and neuroimaging. *J Neurosci* **35**, 3293–7 (2015).

5. Baccala, L. A. & Sameshima, K. Partial directed coherence: a new concept in neural structure determination. *Biol Cybern* **84**, 463–74 (2001).

6. Patel, R., Bowman, F. D. & Rilling, J. A bayesian approach to determining connectivity of the human brain. *Hum Brain Mapp* **27**, 267–76 (2006).

7. Hyvärinen, A. & Smith, S. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research* **14**, 111–52 (2013).

8. Marrelec, G. *et al.* Partial correlation for functional brain interactivity investigation in functional mri. *NeuroImage* **32**, 228–37 (2006).

9. Shimizu, S., Hoyer, P. O., Hyvärinen, A. & Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* **7**, 2003–30 (2006).

10. Hyvärinen, A., Zhang, K., Shimizu, S. & Hoyer, P. O. Estimation of a structural vector autoregression model using non-gaussianity. *J. of Machine Learning Research* **11**, 1709–31 (2010).

11. Billingsley, P. *Probability and Measure* (JOHN WILEY & SONS, 1995).

12. Barnett, L. & Seth, A. K. The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference. *Journal of Neuroscience Methods* **223**, 50–68 (2014).

13. He, B. Y. Scale-free brain activity: past, present, and future. *Trends Cogn Neurosci* **18**, 480–87 (2014).
14. Bédard, C., Kröger, H. & Destexhe, A. Does the 1/f frequency scaling of brain signals reflect self-organized critical states? *Physical Review Lett* **97**, 118102 (2006).

15. Dehghani, N., Bédard, C., Cash, S. S., Halgren, E. & Destexhe, A. Comparative power spectral analysis of simultaneous electroencephalographic and magnetoencephalographic recordings in humans suggests non-resistive extracellular media. *J Comput Neurosci* **29**, 405–21 (2010).
Supplemental Materials

4 DCM forward model

Neuronal level The classic DCM model\(^2\) describes generation of BOLD response from the neuronal networks across two levels: neuronal level, latent in the experiment, and the observable haemodynamic level. The neuronal level reads as follows:

\[
\frac{d\vec{x}(t)}{dt} = A\vec{x}(t) + C\vec{u}(t) + \sigma(t) \tag{3}
\]

where \(\vec{x}\) denotes the temporary activity across all nodes, \(\vec{u}(t)\) denotes binary (on-off) inputs, \(A\) denotes the desired adjacency matrix, \(C\) denotes connections from inputs to nodes, and \(\sigma(t)\) represents the noise. We leaved out the term in the neuronal interaction that relates to the modulation of connections with the inputs (\(B\) matrix in the original DCM model\(^2\)).

Inputs to the network \(u(t)\) were simulated as independent trains of up- and down-states simulated with time resolution of 5ms\(^1\). The probability of state switches was governed by a Poissonian process of an expected high-state duration of 2.5s, and mean low-state duration of 7.5s.

Observation level The golden-standard Balloon-Windkessel model of haemodynamic response function\(^2\) and reads as follows
\[
\begin{aligned}
\dot{s}_i &= \dot{x}_i - \kappa_i s_i - \gamma_i (f_i - 1), \\
\dot{f}_i &= s_i, \\
\tau_i \dot{v}_i &= f_i - v_i^{1/\alpha}, \\
\tau_i \dot{q}_i &= f_i E(f_i, \rho_i) / \rho_i - v_i^{1/\alpha} q_i / v_i 
\end{aligned}
\]

(4)

with the following expression for BOLD response:

\[
y = V_0(7 \rho_i (1 - q_i) + 2(1 - q_i / v_i) + (2 \rho_i - 0.2)(1 - v_i))
\]

(5)

where \(s_i\) - vasodilatory signal, \(f_i\) - inflow, \(v_i\) - blood volume, \(q_i\) - deoxyhemoglobin content, \(V_0 = 0.02\) - resting blood volume fraction, \(E(f, \rho) = 1 - (1 - \rho)^{1/f}\).

In order to reproduce the natural variability in the haemodynamic parameters, we sampled the parameters from Gaussian distributions centered around values given in \cite{2}, with variance such that the standard deviation of the peak delay of BOLD responses equals 0.5.

We used a setup as in Smith et al.\cite{3}: trains of binary inputs switching on- and -off due to a Poisson process. To make the results more generalizable, the frequency of inputs was varying between the two nodes and between the trials: the probability of switching on- and off- was originally set to give an average of 2.5s of on-state and 7.5s of off-state as in \cite{1} but then multiplied with square root of a number of Gamma distribution of a shape and scale of 1. The haemodynamic parameters were also sampled from distributions described in \cite{3}.
5 Benchmark synthetic datasets

Benchmark datasets by Smith et al. are build using the DCM forward model described in Supplementary Materials. The adjacency matrices fed into the model are shown in Fig. 10. All the variations of the connectivity used in this study are acyclic and sparse, and only the size varies between 5 and 50 nodes.

![Figure 10: Adjacency matrices used for the forward modeling in Smith et al., reprint from 1](image)

The DCM forward model allows for emulating a variety of experimental conditions, such as the number of nodes in the network (N), the session duration (SD), the time resolution of the data (TR), the amount of thermal noise added to the BOLD response or the variability in delay of the haemodynamic response. In some simulations, additional features were introduced, e.g., shared inputs or backward connections. All the parameters are summarized in Table 1.
| no | N  | SD (min) | TR (s) | noise (%) | std of the HRF | Other features                       |
|----|----|----------|--------|-----------|---------------|--------------------------------------|
| 1  | 5  | 10       | 3.0    | 1.0       | 0.5           |                                      |
| 2  | 10 | 10       | 3.0    | 1.0       | 0.5           |                                      |
| 3  | 15 | 10       | 3.0    | 1.0       | 0.5           |                                      |
| 4  | 50 | 10       | 3.0    | 1.0       | 0.5           |                                      |
| 5  | 5  | 60       | 3.0    | 1.0       | 0.5           |                                      |
| 6  | 10 | 60       | 3.0    | 1.0       | 0.5           |                                      |
| 7  | 5  | 250      | 3.0    | 1.0       | 0.5           |                                      |
| 8  | 5  | 10       | 3.0    | 1.0       | 0.5           | shared inputs                        |
| 9  | 5  | 250      | 3.0    | 1.0       | 0.5           | shared inputs                        |
| 10 | 5  | 10       | 3.0    | 1.0       | 0.5           | global mean confound                 |
| 11 | 10 | 10       | 3.0    | 1.0       | 0.5           | bad ROIs (time series mixed with each other) |
| 12 | 10 | 10       | 3.0    | 1.0       | 0.5           | bad ROIs (new random time series mixed in) |
| 13 | 5  | 10       | 3.0    | 1.0       | 0.5           | backwards connections                |
| 14 | 5  | 10       | 3.0    | 1.0       | 0.5           | cyclic connections                   |
| 15 | 5  | 10       | 3.0    | 0.1       | 0.5           | stronger connections                 |
| 16 | 5  | 10       | 3.0    | 1.0       | 0.5           | more connections                     |
| 17 | 10 | 10       | 3.0    | 0.1       | 0.5           |                                      |
| 18 | 5  | 10       | 3.0    | 1.0       | 0.0           |                                      |
| 19 | 5  | 10       | 0.25   | 0.1       | 0.5           | neural lag=100 ms                     |
| 20 | 5  | 10       | 0.25   | 0.1       | 0.0           | neural lag=100 ms                     |
| no | N | SD (min) | TR (s) | noise (%) | std of the HRF | Other features                              |
|----|---|---------|--------|-----------|----------------|----------------------------------------------|
| 21 | 5 | 10      | 3.0    | 1.0       | 0.5            | 2-group test                                |
| 22 | 5 | 10      | 3.0    | 0.1       | 0.5            | nonstationary connection strengths           |
| 23 | 5 | 10      | 3.0    | 0.1       | 0.5            | stationary connection strengths              |
| 24 | 5 | 10      | 3.0    | 0.1       | 0.5            | only one strong external input               |
| 25 | 5 | 5       | 3.0    | 1.0       | 0.5            |                                              |
| 26 | 5 | 2.5     | 3.0    | 1.0       | 0.5            |                                              |
| 27 | 5 | 2.5     | 3.0    | 0.1       | 0.5            |                                              |
| 28 | 5 | 5       | 3.0    | 0.1       | 0.5            |                                              |

Table 1: All the parameters included in the benchmark simulations
6  Success rate for single cumulants on the synthetic benchmark datasets
Figure 11: Success rate for all the individual cumulants, across all 28 simulations at benchmark datasets (Supplementary Material 5). In every case, there are cumulants performing better than the cumulant (2,1). Simulation no 7 is particularly easy for the causal inference whereas simulation no 13 is particularly misfortunate. This particular simulation contains loops of excitatory projections \( A \rightarrow B \) and inhibitory backward projections \( B \rightarrow A \).
Comparison between Momentum and other methods on all synthetic datasets
Figure 12: Overall success rate of Momentum against 'PW-LR' methods. In most of the simulations, we achieved a slight improvement with respect to the main competitor, 'PW-LR r skew'. In simulations 7, 8, 9, 14, 17, 20 and 23 the performance is roughly the same.

- Acknowledgements We would like to acknowledge Rembrandt Bakker, Koen Haak, Vinod Kumar, Andre Marquand, Tim van Mourik and Thomas Wolfers for their ideas, remarks and encouragement to carry out this new methodology.

- Competing Interests The authors declare that they have no competing financial interests.
• Correspondence and requests for materials should be addressed to Natalia Bielczyk (email: natalia.bielczyk@gmail.com).