RESEARCH

Non-parametric test for connectivity detection in multivariate autoregressive networks and application to multiunit activity data

M Gilson^{1,5}, A Tauste Campo^{1,2,5}, X Chen^{3}, A Thiele^{3}, and G Deco^{1,4}

[1] Computational Neuroscience Group, Dept. de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain

[2] Epilepsy Monitoring Unit, Department of Neurology, Hospital del Mar Medical Research Institute, Barcelona, Spain

[3] Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK

[4] Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

[5] equal contribution

Keywords: Network connectivity detection; non-parametric significance method; multivariate autoregressive process; Granger causality; multiunit activity

ABSTRACT

Directed connectivity inference has become a cornerstone in neuroscience to analyze multivariate data from neuroimaging and electrophysiological techniques. Here we propose a non-parametric significance method to test the non-zero values of multivariate autoregressive model to infer interactions in recurrent networks. We use random permutations or circular shifts of the original time series to generate the null-hypothesis distributions. The underlying network model is the same as used in multivariate Granger causality, but our test relies on the autoregressive coefficients instead of error residuals. By means of numerical simulation over multiple network configurations, we show that this method achieves a good control of false positives - type 1 error - and detects existing pairwise connections more accurately than using the standard parametric test for the ratio of error residuals. In practice, our method aims to detect temporal interactions in real neuronal networks with nodes possibly exhibiting redundant activity. As a proof of concept, we apply our method to multiunit activity (MUA) recorded from Utah electrode arrays in a monkey and examine detected interactions between 25 channels. We show that during stimulus presentation our method detects a large number of interactions that cannot be solely explained by the increase in the MUA level.

INTRODUCTION

In recent years, there has been a growing interest in developing multivariate techniques to infer causal relations among time series. The initial formulation of the problem goes back to the seminal work by Granger in 1960's (Granger, 1969) motivated by the analysis of the pairwise influence between economic time series. In this work, Granger decomposes the cross-spectrum of two autoregressive time series into two directional components that account for the potential causal influences between each other. A general solution of the problem in multivariate scenarios was developed a decade later by the introduction of multivariate autoregressive (MVAR) processes, which allow the estimation of causal relationships between nodes in networks with linear feedback based on their observed activity (Amemiya, 1974; Geweke, 1982, 1984; Lütkepohl, 2005). The MVAR was further combined with spectral analysis to develop the directed transfer entropy function (Kaminski & Blinowska, 1991; Kamiński, Ding, Truccolo, & Bressler, 2001), which has been employed to analyze connectivity patterns in neurobiological systems (Babiloni et al., 2005; Wilke, Ding, & He, 2008). Granger causality analysis is nowadays often used to evaluate the influence of a group of variables onto other, which corresponds to the influence of a subgroup of nodes onto another one in networks. Nevertheless, it it also applied to detect individual connections between pairs of nodes (each subgroup being a single node), which sets the context of the present paper.

In neuroscience, this inference problem has been transposed to analyze interactions between neuronal populations from spiking activity or neuroimaging measurements such as fMRI, EEG and MEG (Lusch, Maia, & Kutz, 2016; Messé, Rudrauf, Benali, & Marrelec, 2014; Michalareas, Schoffelen, Paterson, & Gross, 2013; Rogers, Katwal, Morgan, Asplund, & Gore, 2010; Seth, Barrett, & Barnett, 2015; Storkey et al., 2007). Two types of estimation procedures may be distinguished: measures relying on an underlying interaction model such as Granger causality analysis (M. Ding, Chen, & Bressler, 2006) and dynamic causal modeling (DCM) (Friston, Harrison, & Penny, 2003) on the one hand; and model-free measures such as transfer entropy (Schreiber, 2000) and directed information (Massey, 1990), which make minimal model assumptions on the other hand. Although model-free approaches have proven useful to describe neural propagation at spike-train level (So, Koralek, Ganguly, Gastpar, & Carmena, 2012; Tauste Campo et al., 2015), certain assumptions are required when estimating interactions at the neuronal population level, in which broader spatial and temporal scales contribute to shaping the signals. Motivated by data-driven practical problems, methodological refinements of Granger causality analysis (or MVAR-based methods) have considered additive noise (Vinck et al., 2015) or measurement noise via state-space models (Barnett & Seth, 2015; Friston et al., 2014). However, in the majority of cases, the ratio behind the detection test concerns sub-model error residuals, which might become too similar when connections are placed in a highly redundant network, thus increasing the missed detection rate (Stramaglia, Cortes, & Marinazzo, 2014).

To overcome the difficulties of detecting directed connections in the general context of large networks, we propose to test the significance of the MVAR coefficients using a non-parametric procedure. As a generative model, the MVAR process is canonically related to Granger causality analysis: the linear regression in the upper right inset of Fig. 1A provides both coefficients and residuals, the latter being viewed as the remaining uncertainty in the prediction of the target time series by its source(s). By comparing the residuals of two linear regressions - one involving a supposed driver node and one without it - in a log ratio, traditional tests for Granger causality estimate the effective interaction of one node onto another (Barrett & Barnett, 2013). Since these log ratios asymptotically converge to known distributions, parametric statistical tests have been developed to assess the significance of these interactions (Barnett & Seth, 2014). Instead, our proposed method evaluates the significance of the MVAR coefficients to infer the existence of network connections. To achieve this, we propose a non-parametric significance test in the regression coefficients space. Previous literature on non-parametric testing for Granger causality has resorted to surrogate data generated by trial shuffling (Dhamala, Rangarajan, & Ding, 2008; Nedungadi, Rangarajan, Jain, & Ding, 2009), bootstrap procedures (Diks & DeGoede, 2001) or by phase randomization in frequency-domain measures (L. Ding, Worrell, Lagerlund, & He, 2007; Li et al., 2016). Here we focus on within-trial surrogate tests for time-domain coefficients and compare them across standard techniques (Faes, Marinazzo, Montalto, & Nollo, 2014; Schreiber & Schmitz, 1996; Winkler, Ridgway, Webster, Smith, & Nichols, 2014).

Our approach is motivated by the growing of multichannel recording techniques in neuroscience, which require tailored multivariate analysis. In the context of recurrent networks, which are ubiquitous in neuroscience, we provide numerical evidence that these tests can achieve a good control of the false-alarm rate and might improve the miss rate by properly adapting the null distribution to each connection. The focus of the present analysis is on the case where we observe more time samples (a few thousands per node) than the network size (about a hundred nodes), a usual ground for electrophysiological data. Within this regime, we test the robustness of the detection method for a broad range of network parameters and various non-trivial topologies inspired by neuronal networks.

METHODS: MULTIVARIATE AUTOREGRESSIVE MODEL AND CONNECTIVITY ESTIMATION

The activity in the MVAR process - a.k.a. noise-diffusion discrete-time network - is described by the following equation:

$$x^t = Ax^{t-1} + \zeta^t \, , \tag{1}$$

where the connection matrix A describes the interactions between coordinates of the vector $x^t = (x_i^t)$ with time t being an integer and node index $1 \leq i \leq N$. Here we constrain our study to the case where ζ^t is Gaussian (possibly cross-correlated noise), whose realizations are time independent for successive time steps. Without loss of generality, we assume that all variables ζ^t have zero means, giving the same property for all x_i. We only consider MVAR processes of order 1 in a first place, but will extend the work to the case of order 2 in a later section.

Granger causality analysis

Granger causality analysis is usually presented using time series and the estimation of non-zero coefficients in A from observed activity over a period $1 \leq t \leq T$ relies on the linear regression of the activity x_i^t of a given node i at time t by the past activity of a subset $\mathcal{S}$ of network nodes:

$$x_i^t = \sum_{j \in \mathcal{S}} A_{ij} x_j^{t-1} + \epsilon^t \tag{2}$$

for $2 \leq t \leq T$. When T is large, the coefficients a_{ij} converge toward A_{ij}. We define the residual ϵ as the standard deviation of the ϵ^t for the ordinary least-square (OLS) regression in Eq. (2), which is for

$$\epsilon\left(x_i^{2\leq t \leq T} | x_{\mathcal{S}}^{1 \leq t \leq T-1}\right) = \sqrt{\sum_t (\epsilon^t)^2} \, , \tag{3}$$

with a notation similar to conditional probabilities; the superscript t indicates the considered time range and the subscripts indicate the nodes involved. To detect the existence of connection $j \to i$ in a network, two types of Granger causality analysis exist: 'unconditional' and 'conditional' (Geweke, 1982, 1984). , they consider the comparisons of the following residuals:

$$\begin{aligned} \mathrm{GRu}(x_j \to x_i) &= \ln\left[\frac{\epsilon\left(x_i^{2\leq t\leq T}|x_i^{1\leq t\leq T-1}\right)}{\epsilon\left(x_i^{2\leq t\leq T}|x_{i,j}^{1\leq t\leq T-1}\right)}\right]; \\ \mathrm{GRc}(x_j \to x_i) &= \ln\left[\frac{\epsilon\left(x_i^{2\leq t\leq T}|x_{1,\cdots,j-1,j+1,\cdots,N}^{1\leq t\leq T-1}\right)}{\epsilon\left(x_i^{2\leq t\leq T}|x_{1,\cdots,N}^{1\leq t\leq T-1}\right)}\right]. \end{aligned} \tag{4}$$

For both GRu and GRc, which have a univariate target node x_i, the usual parametric test for significance relies on the F statistics, which performs better for small number of samples (Barnett & Seth, 2014). The null hypothesis of no interaction for $\mathrm{GRu}(x_j \to x_i)$ corresponds to $m = T$, $p = 1$, $n_x = 1$ and $n_y = 2$ using the notation in Barnett and Seth (2014)

$$\frac{\epsilon\left(x_i^{2\leq t\leq T}|x_i^{1\leq t\leq T-1}\right) - \epsilon\left(x_i^{2\leq t\leq T}|x_{i,j}^{1\leq t\leq T-1}\right)}{\epsilon\left(x_i^{2\leq t\leq T}|x_{i,j}^{1\leq t\leq T-1}\right)} = \left[\exp(\mathrm{GRu}_{ij}) - 1\right] > \frac{\phi(\alpha, 1, T-3)}{T-3} \tag{5}$$

with α the desired sensitivity and ϕ the inverse survival function of the F-distribution (www.scipy.org, n.d.). The equivalent for GRc corresponds to $n_y = N$, yielding

$$\frac{\epsilon\left(x_i^{2\leq t\leq T}|x_{1,\cdots,N}^{1\leq t\leq T-1}\right) - \epsilon\left(x_i^{2\leq t\leq T}|x_{1,\cdots,j-1,j+1,\cdots,N}^{1\leq t\leq T-1}\right)}{\epsilon\left(x_i^{2\leq t\leq T}|x_{1,\cdots,j-1,j+1,\cdots,N}^{1\leq t\leq T-1}\right)} > \frac{\phi(\alpha, 1, T-N-1)}{T-N-1}. \tag{6}$$

We also use non-parametric tests for GRc by performing a circular shift (see details below in Section 'Generation of surrogate time series') either on the target node for each connection (Faes et al., 2014) or independently on the time series of all nodes (in order to save time in estimating the full network's connectivity by shuffling somehow all targets simultaneously). Both cases provide a null distribution for the log ratio, with which the actual estimated log ratio can be compared.

Multivariate autoregressive (MVAR) estimation

To detect the existence of connections $A_{ij} > 0$, another possibility is to estimate the coefficients themselves, which can be done using the covariances of the observed activity variables x^t (Lütkepohl, 2005):

$$\widehat{Q}_{ij}^{\tau} = \frac{1}{T - \tau_{\max} - 1} \sum_{1\leq t\leq T-\tau_{\max}} (x_i^{t+\tau} - \bar{x}_i)(x_j^t - \bar{x}_j), \tag{7}$$

where T denotes the number of successive samples indexed by t, $\tau \in \{0, 1\}$ is the time shift (here $\tau_{\max} = 1$) and the observed mean activity for each node is $\bar{x}_i = \frac{1}{T}\sum_t x_i^t$. The Yule-Walker equation gives a consistency equation for the theoretical covariance matrices (without hat) in terms of the connectivity A

in the dynamics described by Eq.(1):

$$Q^1 = A\,Q^0 \,. \tag{8}$$

The estimation of network connections relies on evaluating A from Eq. (8) for the empirical covariance matrices defined as Eq. (7) and calculated for a given time series:

$$A = \widehat{Q}^1 (\widehat{Q}^0)^{-1} \,. \tag{9}$$

Note that this OLS estimate corresponds to the linear regression related to $\epsilon\big(x_{1,\cdots,N}^{2\leq t\leq T} | x_{1,\cdots,N}^{1\leq t\leq T-1}\big)$ and is also to the linear model with maximum likelihood under the assumption that the observed process is Gaussian.

MVAR of order 2

Eq. (1) can be extended to the case where the activity vector x^t is determined by the two previous time steps:

$$x^t = A^1 x^{t-1} + A^2 x^{t-2} + \zeta^t \,. \tag{10}$$

For the second order, we use $\tau_{\max} = 2$ in Eq. (7) and the estimation of A^1 and A^2 via the Yule -Walker equation is given by (Lütkepohl, 2005, p. 86)

$$\tilde{A} = \tilde{Q}^1\,(\tilde{Q}^0)^{-1}, \tag{11}$$

with the block matrices

$$\begin{aligned} \tilde{A} &= \begin{pmatrix} A^1 & A^2 \end{pmatrix} , \\ \tilde{Q}^0 &= \begin{pmatrix} \widehat{Q}^0 & \widehat{Q}^1 \\ (\widehat{Q}^1)^\dagger & \widehat{Q}^0 \end{pmatrix} , \\ \tilde{Q}^1 &= \begin{pmatrix} \widehat{Q}^1 & \widehat{Q}^2 \end{pmatrix} . \end{aligned} \tag{12}$$

The coefficients of A^1 and A^2 can thus be estimated using a matrix multiplication and an inversion involving the covariances, as with the first-order case in Eq. (9).

Generation of surrogate time series

In this paper, we consider circular shifts (CS), random permutations (RP) and phase randomization (PR) to shuffle the time points of the observed time series. From the original x_i^t with $1 \leq t \leq T$,

- CS draws a random integer $t_0 \in \{1, \cdots, T\}$ and returns $(x_i^{t_0}, \cdots, x_i^T, x_i^1, \cdots, x_i^{t_0-1})$;
- RP draws a random permutation σ of $\{1, \cdots, T\}$ such that each integer appears once (and only once) and returns $x_i^{\sigma(t)}$;
- PR calculates the discrete Fourier transform $\mathcal{F}(x_i^t)$ of the original x_i^t, then multiplies each of the T coefficients of $\mathcal{F}(x_i^t)$ by $\exp(2\pi\imath\phi^t)$ with ϕ^t randomly chosen in $[0, 2\pi]$, and performs the inverse transform.

Importantly, these operations are applied to each time series independently of the others.

In addition, we consider the replacement of all time series in the network by T normally distributed variables with a standard deviation equal to the mean of the standard deviations of x_i^t along the time axis, then averaged for all nodes. We refer to these surrogates as STD.

Experimental setup and processing of electrode measurements to extract MUAe activity

All procedures were carried out in accordance with the European Communities Council Directive RL 2010/63/EC, the US National Institutes of Health Guidelines for the Care and Use of Animals for Experimental Procedures, and the UK Animals Scientific Procedures Act. Two male macaque monkeys (5 - 14 years of age) were used in the experiment; only the data for the first one is used here. A surgical operation was performed under sterile conditions, in which a custom-made head post (Peek, Tecapeek) was embedded into a dental acrylic head stage. Details of surgical procedures and post-operative care have been published previously (Thiele, Delicato, Roberts, & Gieselmann, 2006). During the surgery microelectrode chronic Utah arrays (5*5 grids), attached to a CerePort base (Blackrock Microsystems) were implanted into V1. Electrodes were 1 mm in length in line with procedures described in Supèr and Roelfsema (2005). Stimulus presentation was controlled using CORTEX software (Laboratory of Neuropsychology, NIMH, http://dally.nimh.nih.gov/index.html) on a computer with an Intel Core i3-540 processor. Stimuli were displayed at a viewing distance of 0.54 m, on a 25" Sony Trinitron CRT monitor with a resolution of 1280 by 1024 pixels, yielding a resolution of 31.5 pixels / degree of visual angle (dva). The monitor refresh rate was 85 Hz for monkey 1, and 75 Hz for monkey 2. A gamma correction was used to linearize the monitor output, and the gratings had 50% contrast. Monkeys performed a passive viewing task where they fixated centrally while stationary sinusoidal grating of either horizontal or vertical orientation and 2 cycle per degree spatial frequency, were presented in a location that covered all receptive fields recorded from the 25 electrode tips. Stimuli were presented 500 ms after fixation onset for 150 ms. Raw data were acquired at a sampling frequency of 32556 Hz using a 64-channel Digital Lynx 16SX Data Acquisition System (Neuralynx, Inc.). Following each recording session, the raw data were processed offline using commercial (Neuralynx, Inc.). Signals were extracted using Cheetah 5 Data Acquisition Software, with bandpass filtering set to allow for spike extraction (600-9000 Hz) and saved at 16-bit resolution.

In the present study, we focus on the period starting 200 ms before and finishing 200 ms after the stimulus onset, for 4 conditions (vertical gradings with pre/post cue in the receptor/opposite field) that will not be compared in details. The electrode recordings is firstly down-sampled from 32556 Hz to 1000 Hz. A high-pass filter above 400 Hz is then applied - 3rd-order Butterworth filter at 0.8 of the Nyquist frequency (www.scipy.org, n.d.) - followed by a smoothing of 4 ms to extract the envelope of the resulting signal, by retaining the 250 time points of 1000-ms period surrounding the stimulus onset.

BENCHMARK OF DETECTION PERFORMANCE FOR SYNTHETIC DATA

The workflow of the benchmark for the estimation procedure is schematically represented in Fig.1A. We first consider a MVAR process defined by Eq. (1) with given connectivity matrix A and input covariances (obtained by mixing independent Gaussian processes) to generate the activity of the network. From the observed activity over a period of duration T, we estimate the coefficients matrix A using the covariances as described in Eq. (9). We also perform the linear regressions of each node activity over the past activity of given subsets of nodes corresponding to the unconditional (GRu) and conditional (GRc) Granger causality analysis, from which we calculate the ratios of residuals in Eq. (4). Actually, these estimates correspond to the same ordinary least-square (OLS) regression (top right in Fig. 1A) and the difference resides in the spaces where they lie: coefficients versus residuals.

For each method, the prediction power can be measured by the relationship between the estimated values and the original connectivity values, as illustrated in Fig. 1B (left). To discriminate between existing and absent connections, one can apply a common threshold for all connections (top thread); by sliding this threshold, we obtain the ROC curve with the rate of false alarms on the x-axis and true positives on the y-axis. The area under the curve indicates in a single value how well the ranking of estimated value performs for the detection of connections in the original connectivity. Alternatively, an

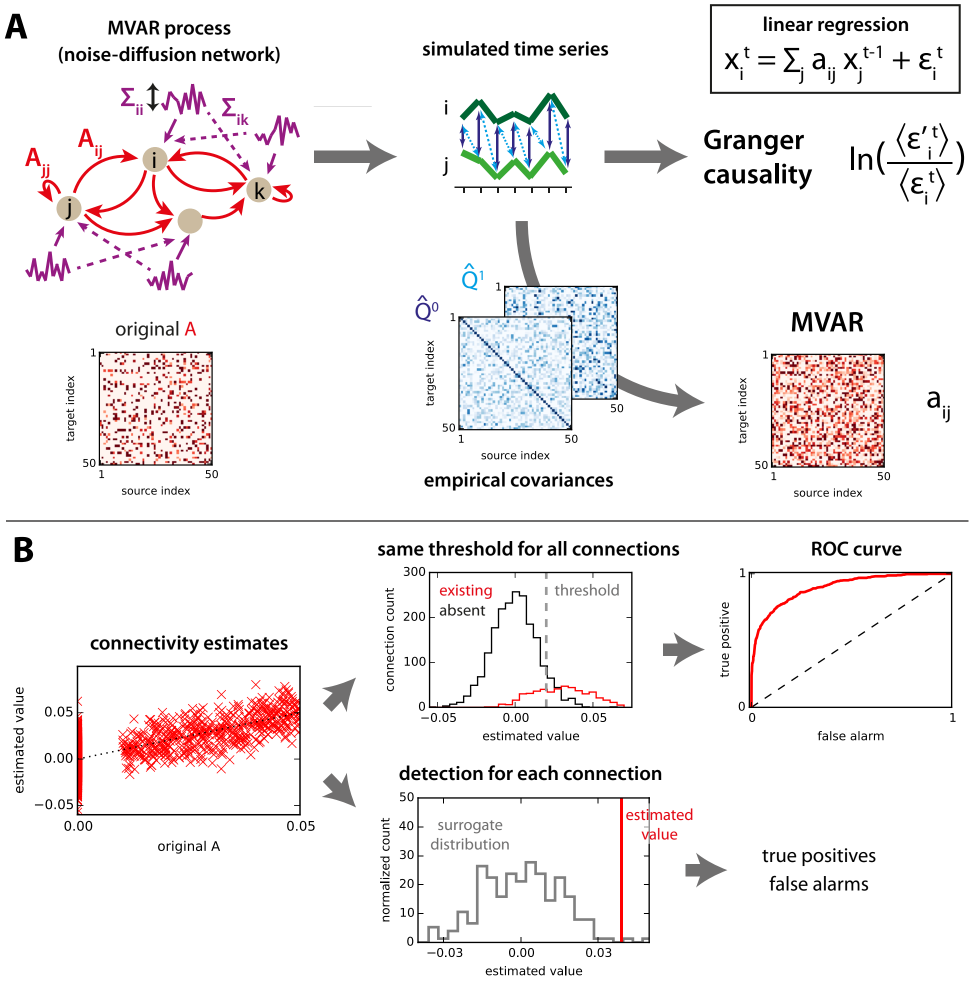


Figure 1. Network model and connectivity estimation. A: For a given directed connectivity A and input covariances Σ (left), the network activity (middle) is simulated using Eq. (1). From the observed time series, the existing interactions in the original connectivity can be estimated (right): Granger causality analysis uses the residuals of linear regressions (ϵ in the upper right equation; see Methods for details about the residuals used in the log ratio), whereas MVAR corresponds to the coefficients. Note that MVAR can be obtained using the empirical covariance matrices $\widehat{Q}^0$ and $\widehat{Q}^1$, see in Eq. (7) for $\tau = 0$ and 1. **B:** The left panel compares the estimated values to the original values for all connections in the network. The upper thread displays the distributions of estimated values for existing and non-existing connections in the original network. Using a sliding threshold (vertical dashed gray line) on the estimated values, one can calculate the ROC curve (right). The lower thread compares the estimated value for a single connection to a null distribution. From this, the choice whether the connection exists or not is made for each individual connection, yielding a single pair of true-positive and false-alarm rates.

individual test can be made for each connection in the network, for example by comparing the estimated value to a null distribution (bottom thread). Here again, we obtain two rates of false positives and negatives.

Coefficients from linear regression potentially predict better existing connections than residuals

We start with the comparison between the predictability of coefficients and residuals for MV, GRu and GRc for all connections in a given network. To do so, we simulate 500 randomly connected networks, which are simulated with different sizes ($N = 50$ to 150 nodes), density and connectivity weights (uniformly drawn in a randomly chosen range $[w_{\min}, w_{\max}]$); here inputs are *not* correlated: the ζ_i are independent across node indices in Eq. (1). For each network configuration, we evaluate the accuracy for connection detection via the area under the ROC curve (see the upper thread in Fig. 1B). Fig. 2A displays this ROC-based accuracy as a function of the number of observed time samples (x-axis) represented by violin plots for 500 randomly connected networks. When considering many samples (10^4), all methods perform well. However, for smaller sample sets, the MVAR method exhibits superior performance than GRu: as measured by the Mann-Whittney test, $p < 10^{-45}$, $p < 10^{-19}$ and $p < 10^{-5}$ for the three values of observed samples, respectively.

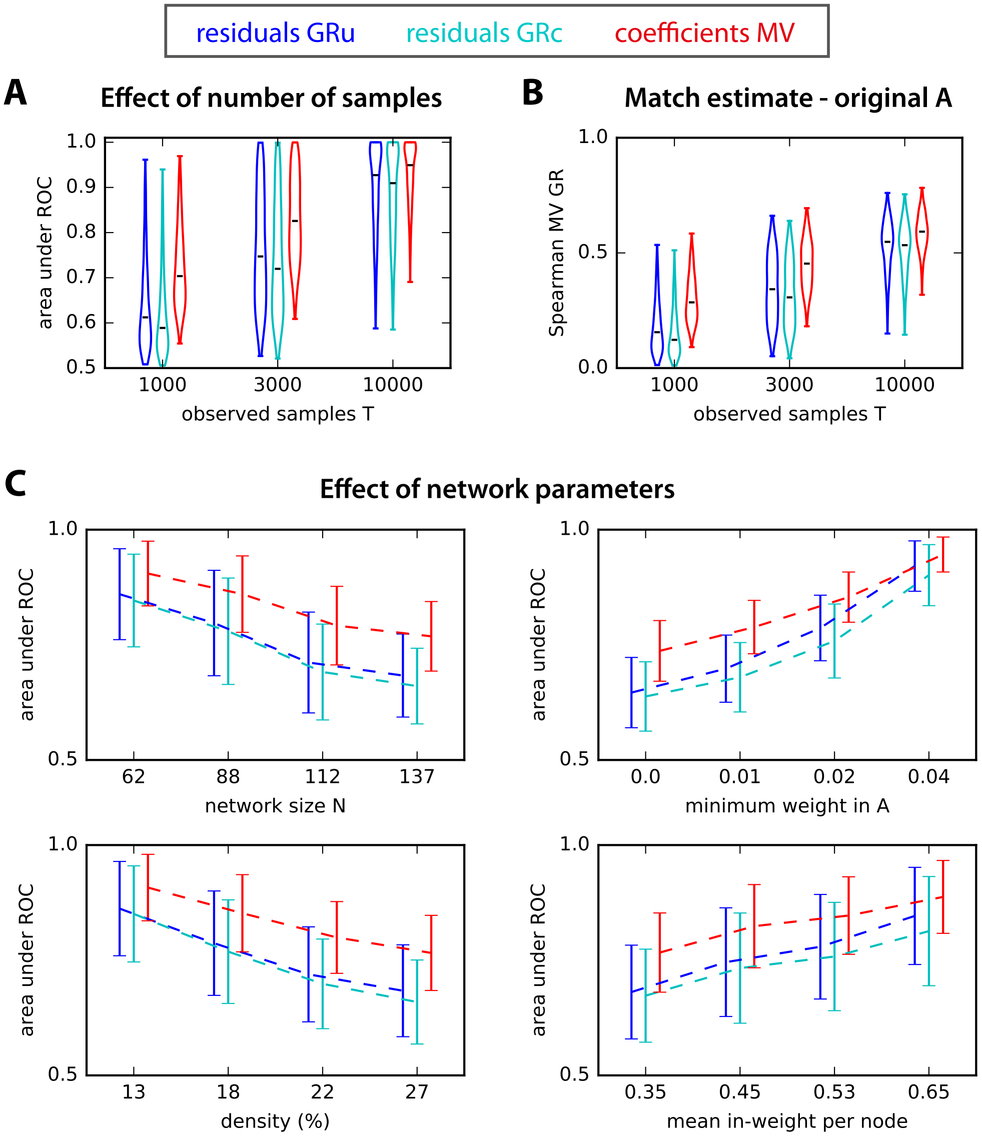
Although error residuals ratios are in a different space from the true weights in A, one expects some degree of correlation between them, such that Granger causality analysis effectively detects connections. In Fig. 2B, both GRu and GRc estimates have a ranking similar to the original A weights (as measured by the Spearman correlation) for $T = 10000$ observed time samples, but this weakens dramatically for $T \leq 3000$. In contrast, the ranking for estimated MVAR coefficients reflects much better the original A for $T \leq 3000$. In the studied networks, GRu performs slightly better than GRc. As analyzed in previous studies, this can be consequence of the balance between redundant and synergistic activity exhibited by the simulated network nodes (Stramaglia et al., 2014). To shed light into the effect of the network structure, we next examine how the ROC-based performance in Fig. 2A depends on the controlled network parameters.The four panels in Fig. 2C display the trends of the values for the 500 networks as a function of the network size N, the network density, the minimum weight in the original network ($w_{\min}$ mentioned above) and the mean sum of incoming weights per node. For illustration purpose, the 500 networks are grouped in quartiles for each parameter. Not surprisingly, the estimation accuracy of all methods decreases as a function of the network size N and density, and increases as a function of the minimum connectivity weight and the mean incoming weight per node. More interestingly, in challenging configurations with small weights, MVAR consistently shows a superior performance by a larger gap compared to Granger causality analyses. These findings support the use of coefficients to robustly detect connections in recurrently connected networks. Note that GRu performs on average slightly better than GRc here: the discrepancy decreases as a function of the network density, which may follow from lower redundancy in the recurrent network (Stramaglia et al., 2014).

A robust non-parametric significance test for MVAR

We have so far examined the performance of different estimation methods based on the area under the ROC curve, which corresponds to a single threshold for all connections in a network and combines the information about false alarms and true detection over the whole range of estimated values. However, in the context of real data, the decision for the existence of a connection typically relies on comparing the value of the connection estimate with a given statistical threshold. For GRu and GRc, such parametric tests have been developed, for example, based on the F statistics (Barnett & Seth, 2014). Equivalently, it is sufficient to know how the values of the estimates for absent connections are distributed, in order to select a desired rate of false alarms (type-1 error). In this section, we develop a significance test for the estimated MVAR coefficients by providing a null-hypothesis distribution for absent connections.

Our approach relies on the fact that covariances reflect the underlying connectivity: we thus construct the null distribution for estimates by performing a random permutation for each of the observed time samples, which “destroys” the covariance structure apart from the variances on the diagonal of $\widehat{Q}^0$ as illustrated in Fig. 3A; other methods will be tested in a later section. From the resulting covariance matrices, we evaluate a surrogate connectivity matrix. The core result underlying our surrogate approach is shown in Fig. 3B: the distribution of surrogate estimates (thick black line) is compared against the distribution of existing (red) and absent (blue) connections in a simulated random network model: the surrogate distribution in black provides a good approximation for the distribution of estimates for non-existing connections in blue.

We consider two options - corresponding to the two threads in Fig. 1B - to test the existence of a connection from an MVAR estimate while keeping the false alarm rate under control.



209 **Figure 2. ROC-based prediction power. A:** Area under ROC for estimated A obtain from log-ratios of residuals obtained from Granger causality analysis (unconditional for GRu and conditional for GRc) and MVAR. The x-axis indicates three sample size T for the observed network activity. The the violin plots correspond to 500 simulated networks of various sizes and connectivity strengths (the horizontal black bar indicates the median). **B:** Match of the ranking between GRu, GRc and MVAR estimates and the original connectivity weights A, as measured by the Spearman correlation coefficient. The plotted values correspond to the 500 networks in A and the x-axis indicates the sample size T. **C:** Effect of network parameters on ROC-based performance. Influence of network size N, connectivity density, sum of recurrent connectivity strengths, minimum weight $w_{\min}$ in A, mean noise on the diagonal of Σ and mean off-diagonal noise in Σ on the ROC-based accuracy in Fig. 2A. In each plot, the network configurations have been grouped in quartiles according to the parameter plotted on the x-axis, and the corresponding group mean and standard deviations are indicated; the curves are displaced horizontally to improve legibility.

- The *global* test relies on the null distribution corresponding to the black histogram in Fig. 3C, which is obtained by grouping together all SN^2 matrix elements of all matrices for $S = 200$ surrogates. From that surrogate distribution, we perform a detection test by setting a threshold corresponding to a percentage of the right tail equal to the desired false-alarm rate (here 2%), as illustrated by the vertical gray dashed line.
- Instead, the *local* test uses for each connection the surrogate distribution of S values, corresponding to the same matrix element in each of the S surrogates. From that distribution in red in Fig. 3C, the detection threshold is defined similarly (vertical dark red dashed line).

The rationale behind these two choices lies in the trade-off between taking into account spatial heterogeneity in the network and gaining larger sample size, as illustrated by the distinct thresholds in Fig. 3C. Note also that the F statistical test for Granger causality analysis corresponds to a global threshold on the log ratio values. When varying the desired false-alarm rate, the two tests perform well, as illustrated in Fig. 3D by their location close to the ROC curve (circles and triangles for local and global, respectively).

To assess the effect of the small variability observed in Fig. 3D over the randomness of network configurations, we simulate 500 randomly connected networks with the same parameters as in Fig. 2, except for the size $50 \leq N \leq 90$ and the presence of input cross-correlations. Note that, from Fig. 2, the chosen size N corresponds to a situation where Granger causality analysis performs relatively well as compared to MVAR. The control of the false-alarm rate is displayed in Fig. 3E for both local and global tests with various numbers S of surrogates. The control of false-alarm rates is close to perfect across various values for all S and both tests (local and global), demonstrating the robustness of the proposed method for randomly connected networks. Following, we fix the desired false-alarm rate to 2% and evaluate the miss rate (true negatives) of both methods depending on the actual weight strength: in Fig. 3F, connections are grouped in terciles for each network configuration. Interestingly, the local test improves with the number S of surrogates (right panel), whereas the global test exhibits a constant performance for all S (left panel). Note that the advantage of the local test over the global test particularly concerns connections with small weights, which are difficult to detect, in line with Fig. 2C (see influence of the minimum original weight).

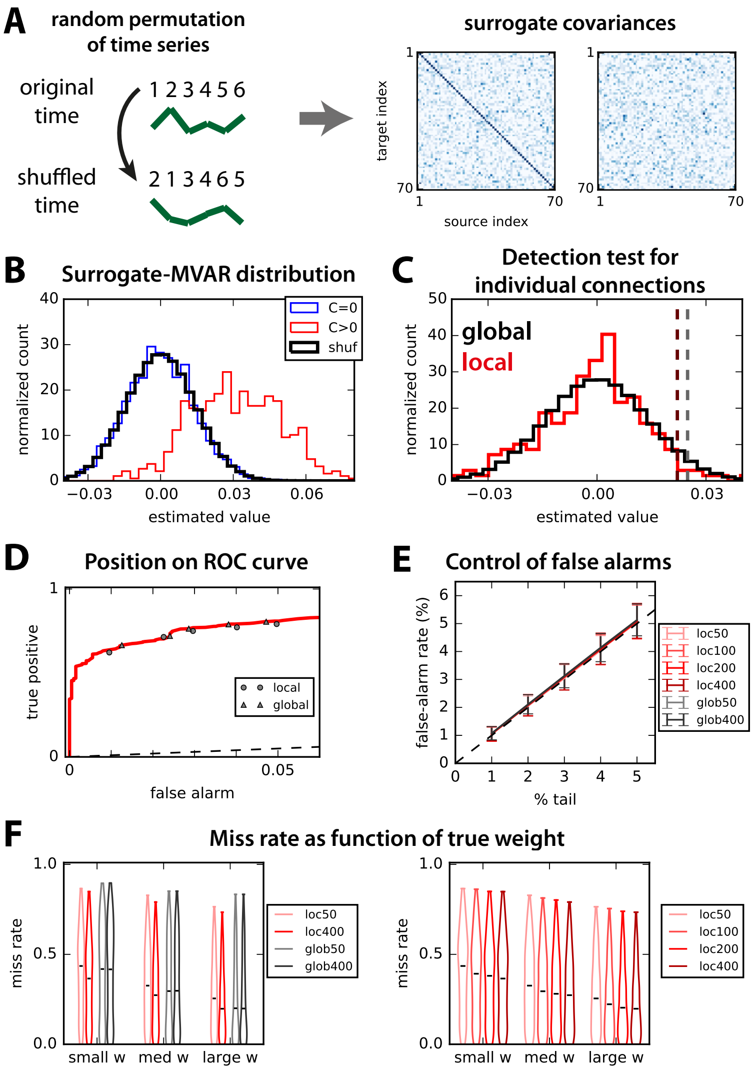


Figure 3. Non-parametric test to assess significance for MVAR coefficients based on random permutations. **A:** Schematic illustration of random permutation (numbers indicate time) applied independently to all observed time series (green curves) to generate the surrogate covariances (right panels). **B:** Pooled distributions of estimated weights for the existing (in red) and absent (blue) connections. The thick black curve indicates the distribution of connections over 100 surrogates, which closely matches the blue distribution. **C:** For a given connection, we compare two methods: for 'local' in red, the null distribution corresponds to the matrix elements for the same connection in 200 surrogates; for 'global' in black, the null distribution is the pooled distribution for all N^2 elements of the 200 surrogate matrices (same as in B). The dark red and gray dashed lines indicate the detection thresholds corresponding to the 4% tail for those two options. **D:** The performance of the two non-parametric methods for the thresholds described in B and C is displayed on the ROC curve for a desired false-alarm rate ranging from 1 to 5%. Triangles indicate the local test and circles the global test. **E:** Comparison of the desired (% of the tail of null distribution) and actual rate of false alarms for the local and global tests when varying the the number S of surrogates (see figure in legend). Error bars indicate one standard deviation for 500 random networks; importantly, inputs for these networks have cross-correlation, unlike Fig. 2. **F:** Influence of the strength of original weight on the detection performance for the 500 random networks and a desired false-alarm rate set to 2% in E. In both panels, lighter colors indicate smaller numbers of surrogates S, in red for the local test and gray for the global test (see legends).

Because GRu does not take all network nodes into account, the presence of spatially correlated noise (indicated by the purple dashed arrows in Fig. 1A) dramatically affects the false-alarm rate when using the parametric significance F-test (Barnett & Seth, 2014), as shown in Fig. 4A by the dark blue dashed curve. This is solved by the "complete" linear regression in GRc, achieving a quasi perfect control irrespective of the input correlation level for both the parametric and two non-parametric tests (cyan, green and blue-green dashed curves, respectively), as our non-parametric tests do (red and gray; recall also Fig. 3E). We consider two non-parametric tests for GRc: 'T' stands for target, where the null distribution of a connection is obtained by shuffling only the target, and 'F' stands for full, where we shuffle all time series simultaneously as in our coefficient-based tests. Both perform equally in terms of false alarms.

The main result of the paper is described in Fig. 4B, where the dashed line corresponds to the miss rate for parametric GRc: our non-parametric method exhibits a better than miss rate - i.e., decrease - for both local and global tests (in red and gray, respectively) on average over the same 500 random networks as in Fig. 4A. For $S \geq 200$, the local test even becomes better in all cases. Note that the small miss-rate improvements of about 7% actually corresponds to more than 50 existing connections per network here. In contrast, both non-parametric tests for GRc perform worse than the parametric test here, with the target-shuffling surrogate converging faster close to the non-parametric GRc. Fig. 4C displays the trends of the performance of all 5 tests in Fig. 4B as a function of two network properties: the mean incoming weight per node (left) and the density (right). The main result here is that the local test performs better especially in difficult configurations with small weights and dense connectivity.

From Figs. 3F and 4B-C, we conclude that the local test is preferable to the global test provided $S \geq 200$ surrogates are generated. However, the computational cost increases linearly with S, as illustrated in Fig. 4D by the red curves. Note that the parametric GRc (in cyan) takes the same time to calculate as $S = 50$ surrogates. However, our non-parametric method scales better than parametric GRc when the network size increases. As a comparison, the full-network non-parametric test for GRc takes longer time to compute, but further optimization of the calculations could be made that were not incorporated here.

315 ***Comparison of generation methods for surrogates for non-parametric MVAR***

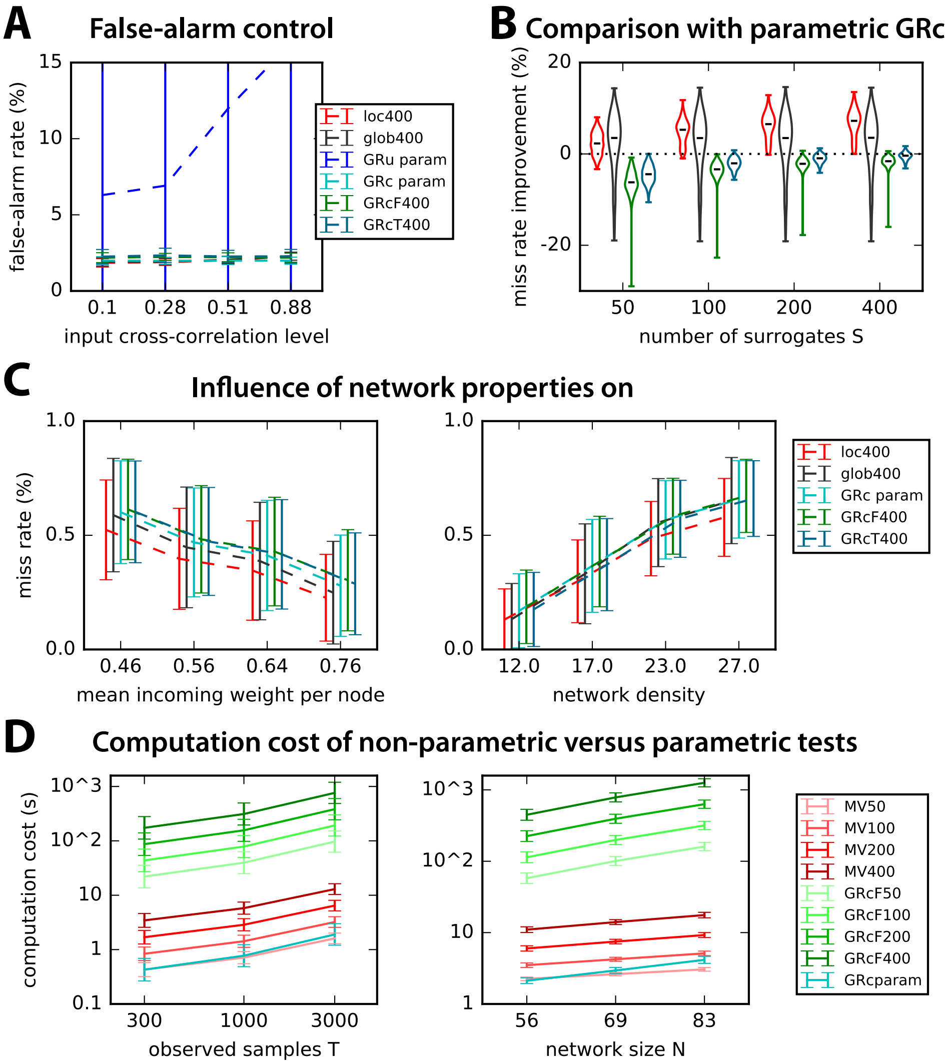


Figure 4. Comparison of our coefficient-based method with Granger causality analysis. A: Comparison of the parametric tests for GRu (blue curve) and GRc (cyan) with the non-parametric methods for GRc (green for GRcF and blue-green for GRcT, see the text for details) and MVAR (red for local test and gray for global). The x-axis indicates the strength of input correlations (i.e., pink noise) in the simulated network. The desired false-alarm rate is set to 2% as in Fig. 3F and the number of observed time samples is $T = 3000$. Error bars indicate one standard deviation over the 500 random networks as in Fig. 3E. **B:** Comparison of the miss rate improvement (decrease) with respect to parametric GRc for the 500 networks in A as a function of the number S of surrogates (x-axis). Red indicates the local test, gray the global test, green the full-network non-parametric GRc and blue-green the target-only non-parametric GRc. **C:** Details of the performance of the 5 methods in B as a function of the mean incoming weight per node (left) and the network density (right). The plots for the miss rate are similar to those for the ROC-based prediction power in Fig. 2C. **D:** Comparison of the computational cost for the surrogate-based method and parametric tests as a function of the number T of observed samples (left) and network size (right). Only GRcF is shown, as GRcT takes much longer time in the unoptimized version that we use.

The fact that the OLS MVAR estimates can be obtained via the two covariance matrices (with and without time shift, see Fig. 1A) hints at possible methods to generate surrogate by destroying the information in these covariances. Methods to generate surrogate time series have been widely used in the past: circular shifts of the time series (Faes et al., 2014), random permutation (Winkler et al., 2014) and phase randomization (Schreiber & Schmitz, 1996) to generate a null distribution for the ratios in Eq. (4); they are referred to here as CS, RP and PR, respectively. We thus consider these three methods (cf. box in Fig. 5), as well as surrogate time series that only preserve the mean standard deviation averaged over the network (STD), so as test to which extent it is important to preserve the spatial heterogeneity of the nodes' activity. See Methods for details about the calculations.

The control of false alarms for local tests in Fig. 5A and B is better for CS and RP, whereas the detection of true connections is similar for the four methods over 500 random networks of size $N = 70$. However, CS fails to detect self connections (Fig. 5C). The reason is that, because CS surrogates preserve the autocovariances in the time-shifted covariance, they fail to build a proper null distribution for self-connections. The influence of the number of samples used in the estimation is similar for all methods, as illustrated in Fig. 5D. The comparison with STD (purple), which averages the covariance statistics over the whole network, suggests that the local test makes a good use of the heterogeneous information across nodes. As a conclusion, we retain RP as the best option.

Influence of network topology

In this part, we test and compare the robustness of global and local surrogate-based detection tests to specific connections and topological configurations. Here, $T = 3000$ observed samples and we compare the local and global tests with $S = 400$ surrogates for 500 networks of each type. In all cases, the simulated networks have the same size $N = 70$, but vary in connectivity density, distribution of recurrent weights and level of input cross-correlation. We compare the miss rate for unidirectional, reciprocal and self connections in the random networks examined until now (and a desired 2% of false alarms). Fig. 6A shows that the miss rate is similar in unidirectional and reciprocal connections with the local test, which performs slightly better than the global test (as in Fig. 3F).

Now we consider more elaborate network topologies than the random connectivity (Erdös-Rényi) considered so far, namely modular and hierarchical networks. In Fig. 6B, we simulate 500 modular

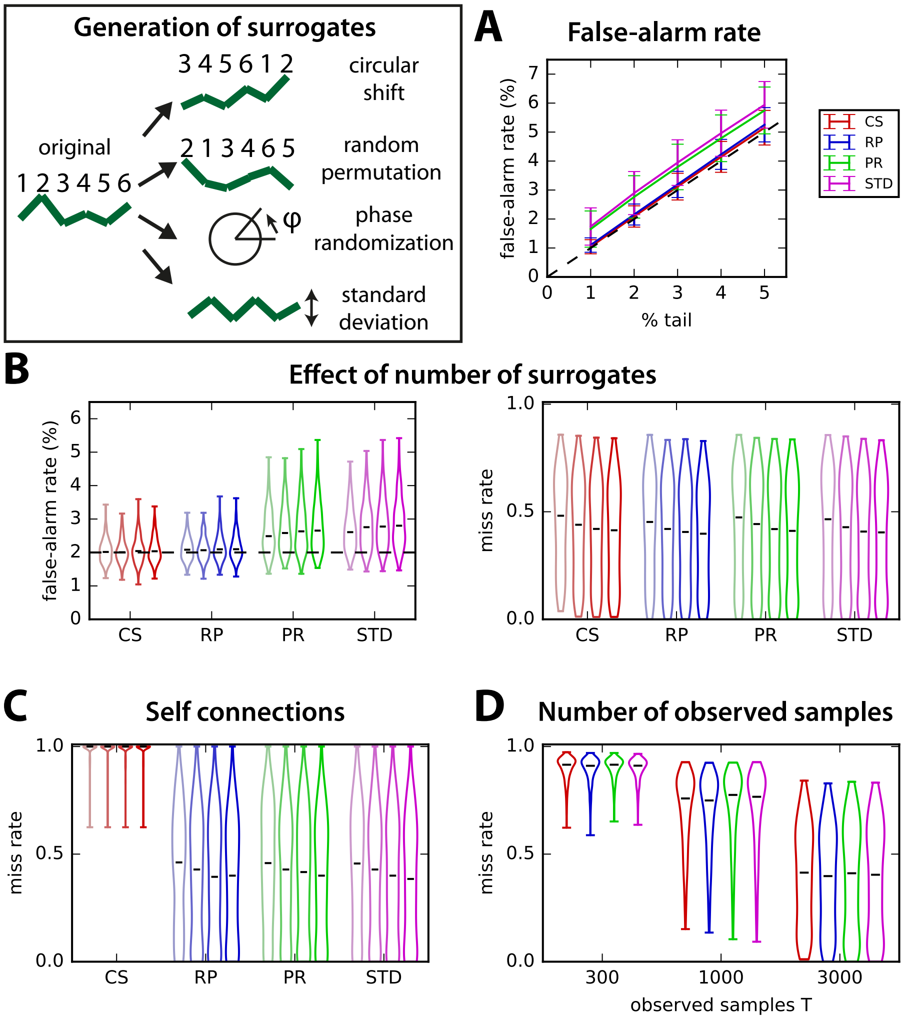


Figure 5. Comparison of the local test for 4 surrogate generation methods. Surrogates are generated by independently performing for each original time series 1) random permutations, 2) circular shifts, 3) phase randomization and 4) replacing the original by a new random time series with the mean standard deviation averaged over all of the original time series in the network. **A:** Comparison of the control of the false-alarm rate for various thresholds on the tail of the distributions of 400 surrogates (% indicated on the x-axis). The error bars correspond to the variability over 500 random networks similar to Fig. 3 with $T = 3000$ observed time samples and the local test. **B:** Influence of the number S of surrogates on the detection performance for the 4 surrogate methods (x-axis). The violin plots indicate the distribution of false-alarm rates (left panel) and miss rates (right) for the 500 networks in A with a desired false-alarm rate set to 2% (dashed line in the left panel). Lighter to darker colors correspond to 50, 100, 200 and 400 surrogates, respectively. **C:** Same as the miss rate in B, but only for self connections. **D:** Influence of the number T of observed time samples on the miss rate for $S = 400$ surrogates.

networks with two groups (green and blue) linked by hubs (red, about 5 to 15% of the nodes). Interestingly, intra-group and hub-group connections have a similar miss rate with regard to using local and global surrogates. In Fig. 6C, we simulate hierarchical networks of three layers, for which connections either link the center and an intermediate node, or link an intermediate node and a leaf, or are self connections. This network type is much sparser than the two types in A and B, yielding a quasi perfect detection performance for all types of connections (miss rate < 0.1 in Fig. 6C). In all cases, the local test performs better than the global test. However, the control of false-alarm rate is similar for both tests with all topologies, as can be seen in Fig. 6D.

Finally, we consider a network with both excitatory and inhibitory connections (with a inhibitory ratio equal to 5 to 50% of all) and perform the test by defining a threshold on both tails of the null distributions. As can be seen in Fig. 6E, the positive/negative nature of the connection weights affect neither the false-alarm nor the miss rate. However, the performance is poorer than with excitatory connections only.

We conclude that, in those networks with spatial heterogeneity as with randomly connected networks, the local test with an individual null distribution per connection performs better than the global test. Recall that an improvement of the miss rate by 1% in a network with a density of 20% actually corresponds to $N^2 0.2/100 \simeq 10$ existing connections here, so the plotted improvements concern about 50 connections.

Applicability to second-order MVAR process

As explained in Methods, an MVAR process whose state depends on the two previous time steps can be estimated with the covariances with time shifts $\tau = 0$, 1 and 2; see Eqs. (11) and (12) for details. Here we simply focus on random connectivity for the two corresponding matrices A^1 and A^2, with size N that is randomly drawn between 30 and 80; we construct A^1 and A^2 such that a connection $j \to i$ cannot be in both matrices, but at most in one. The existing connections are detected with the non-parametric local test relying on RP surrogates for each matrix separately, as a proof of concept. The control of false alarms in Fig. 7A and the overall detection performance in Fig. 7B suggest that our surrogate method can be extended satisfactorily to higher-order MVAR processes. Note that the improvement by generating more surrogates is rather weak here. Importantly, there is no difference between the detection in A^1 and

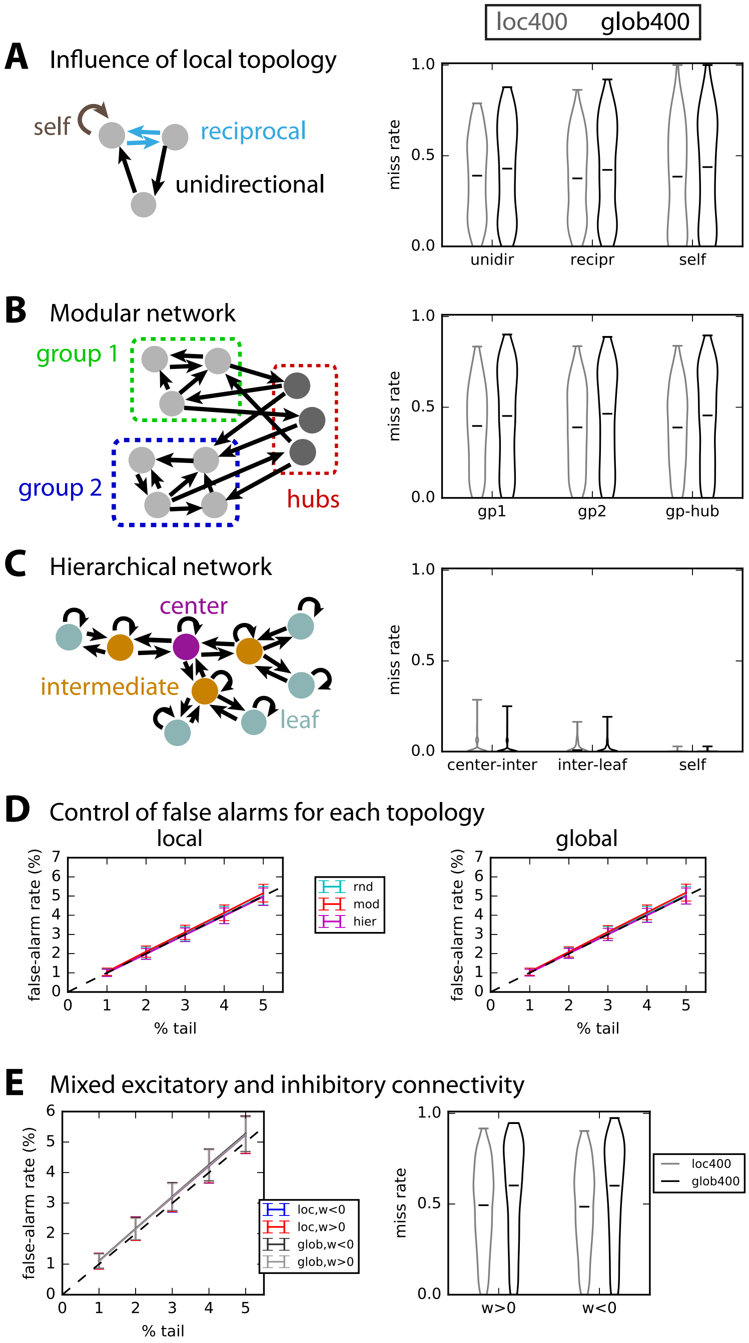


Figure 6. Robustness to non-trivial network topology. A: Detection performance for unidirectional, reciprocal and self connections in 500 the randomly connected networks used so far in Fig. 3. **B:** Detection performance for modular topology schematically represented in the left: each of the 500 networks comprises of two groups connected by hubs. The connections are separated depending on whether the connect group nodes or hubs, as indicated by the diagram on the left. **C:** Similar to B with a hierarchical topology, where connections are grouped in 3 subsets: from center to intermediate nodes; from intermediate nodes to leaves; self connections. The results concern 500 such networks, which all have very low density. **D:** Control of false-alarm rate for the local (left) and global (right) significance tests and the three network topologies; the plot is similar to Fig. 3E. **E:** Control of false-alarm rate and miss rate for networks with both excitatory and inhibitory connections.

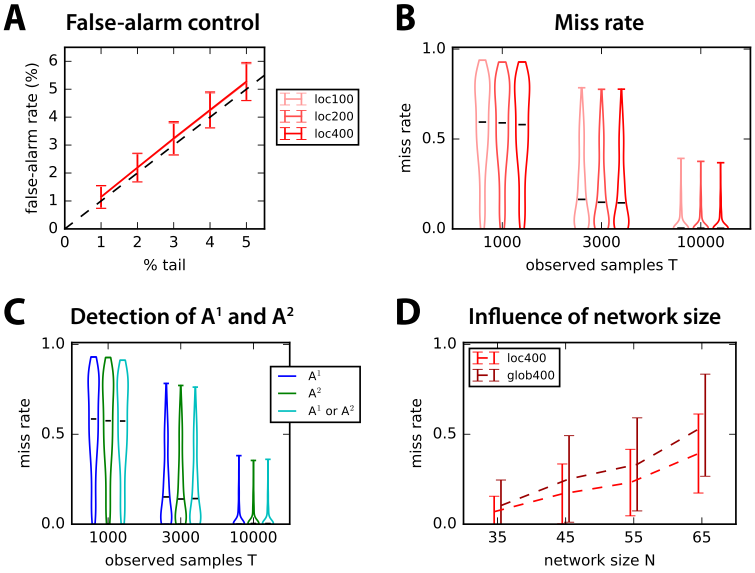


Figure 7. Connectivity detection for second-order MVAR process. A: Control of false-alarm rate for the local test with $S = 100$, 200 and 400 in both connectivity matrices $A^1$and A^2, corresponding to each time step. Error bars correspond to the variability over 500 network configurations with random connectivity and $T = 3000$ observed samples. **B:** Influence of the number T of observed samples (x-axis) on the miss rate for the 500 networks in A. Lighter to darker red indicates the number of surrogates S. **C:** Details of the detection performance for A^1 and A^2 separately, as well as connections in either A^1 or A^2. The number of observed samples is indicated on the x-axis as in B. **D:** Influence of the network size N on the detection performance in C for the local and global tests with $S = 400$ surrogates.

A^2, as demonstrated in Fig. 7C. Last, the network size worsens the miss rate in Fig. 7D, which affects more dramatically the global test as compared to the local test.

APPLICATION TO EXPERIMENTAL DATA

Multiunit activity data obtained from Utah electrode array in monkey

Now we consider data recorded from a monkey performing a visual task, where the stimulus corresponds to vertical gratings covering all recorded V1 receptive fields from the Utah arrays (see Methods for details). We aim to provide a proof of concept for the connectivity analysis for this type of data, so as to complement the more classical analysis based on the activity of individual channels; therefore we do not focus on comparing the 4 stimulus conditions with each other.

The multiunit activity envelope (MUAe) is obtained as described in Methods. In Fig. 8A, the resulting MUAe is represented for two out of the 26 channels (red and purple) for two trials in the top and middle panels, 400 ms before and 600 ms after the stimulus onset. The typical analysis of MUAe activity consists of averaging over 200 trials, which exhibits a peak immediately after the stimulus for the two channels in the bottom panel. Among the 26 channels, about a third show a large increase in activity after the stimulus onset as compared to before (namely, a post-stimulus mean activity larger by more than three standard deviations compared to the pre-stimulus activity); almost all channels show a moderate increase of one standard deviation. One channel is discarded for a much larger activity (by 5 times) than all others.

To further investigate the temporal information conveyed by MUAe jointly for pairs of channels, we calculate the pairwise covariances between them, after centering the MUAe activity individually for each trial. Fig. 8B shows the stabilization of the cross-covariance between the two channels in Fig. 8A from a single trial to averages over 20 and 200 trials. Note the asymmetry with respect to time difference: this information is extracted by the network model to estimate the interactions between the neuronal populations recorded by the channels. Then we verify that the model can be applied to these data, by examining the MUAe autocovariances in Fig. 8C, which exhibit a profile corresponding to an exponential decay up to two time shifts (i.e., 8 ms for the downsampling every 4 ms), that is, a straight line in the log plot. This suits an autoregressive model with large positive values on the diagonal of the connectivity matrix A.

Both connectivity matrices for the 25 channels estimated using the MVAR method before and after the stimulus are illustrated in Fig. 8D for condition 1: we find larger off-diagonal values for the period after the stimulus than before. This is actually true for all conditions, as indicated by the more spread distributions in red as compared to gray in Fig. 8E. The channels appear to be coordinated at the considered time scale of 4 ms and their collective interaction scheme is affected by the stimulus presentation.

Significance test for real data: interactions related to stimulus presentation

We then use the local and global tests based on 1000 surrogates (with random permutation) to retain only significant interactions from the estimates in Fig. 8D: this leaves a few interactions for the pre-stimulus period in Fig. 9A (left panel), 8 out of 650, which is of the order of the desired false-alarm rate set to 1%

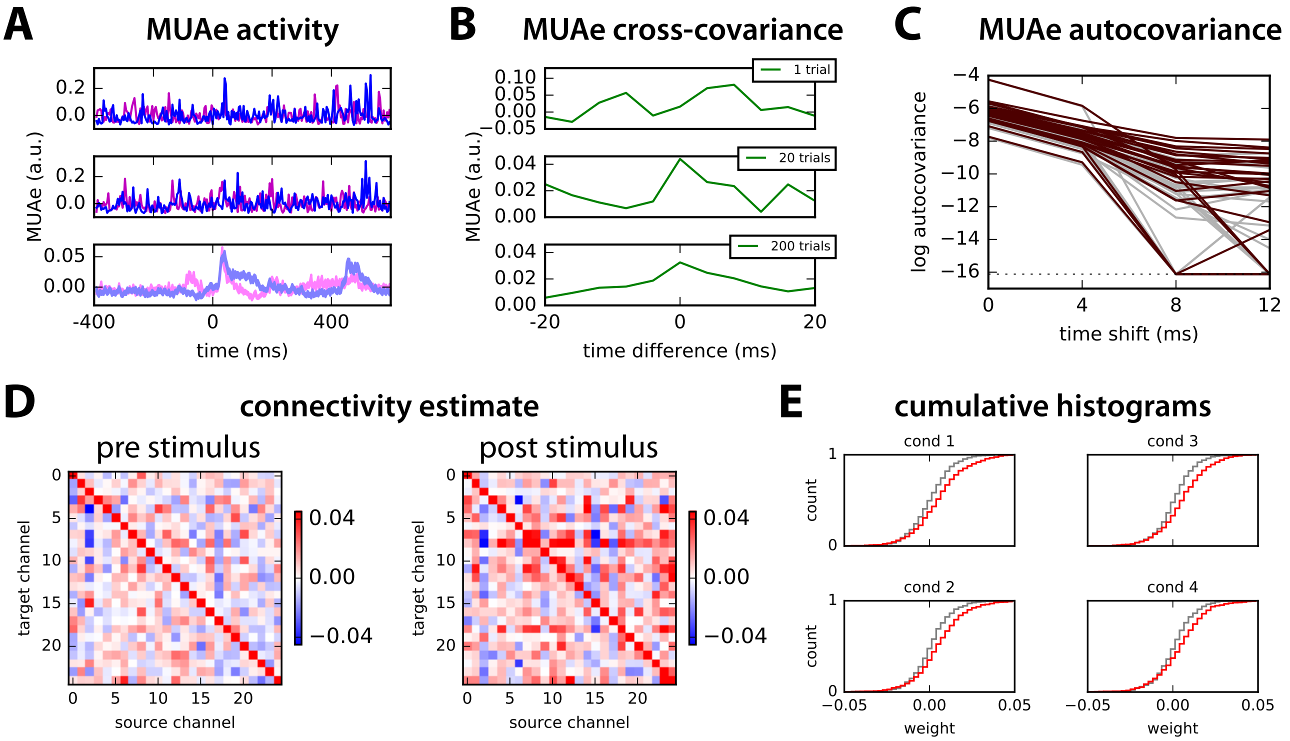


Figure 8. Application to multiunit activity (MUAe) data. A: Example of two trials (top and middle panels) of multiunit activity envelope (MUAe) for two channels of recordings using Utah electrode array in the primary visual cortex of a monkey (in arbitrary units; see text for further details). The bottom panel represents the average over 200 trials (with standard-error mean for the thickness of the curve). The stimulus is presented at time point 100 (actually 400 ms, since the smoothing corresponds to a smoothing window of 4 ms). **B:** Example of cross-covariance between the two channels in A averaged over 1, 20 and 200 trials. **C:** Autocovariance profiles of MUAe signals for all 25 channels and time shifts up to 12 ms averaged over 200 trials plotted with a log y-axis: comparison of signals before (gray) and after (red) the stimulus presentation. **D:** MVAR estimates of the connectivity between the 25 channels for the MUAe activity 200 ms before and after stimulus presentation (i.e., 50 time points each), averaged over 200 trials. The scaling has been optimized to enhance the legibility off-diagonal elements. **E:** Comparison of the cumulative distribution of connectivity weights (off-diagonal elements in D) for the 4 conditions. Gray and red indicate before and after the stimulus, respectively.

(namely, the extreme 0.5% of each tail). In contrast, many more post-stimulus interactions survive the significance tests in the right panel: almost all these interactions are unidirectional. The counterpart for circular shift for Fig. 9B involve 24 interactions in common with Fig. 9A. On average over the 4 conditions, 22 post-stimulus interactions are common between the two shuffling methods, to be compared with 7 for the pre-stimulus period (both with a standard deviation of 4); this corresponds to 3.5% of all possible interactions. Almost all detected interactions are unidirectional, as illustrated in Fig. 9C for both local and global tests for the post-stimulus period. Varying the threshold on the tail of the null distributions, we see that the number of detected interactions is close to the desired false-alarm rate for the pre-stimulus period in Fig. 9D (dark red and black curves, respectively). In contrast, post-stimulus interactions are many more for both local and global tests (light red and gray). The global test detects fewer interactions than the local test, indicating the necessity to take into account the disparities across channels. Around 57% of post-stimulus interactions detected by the global test (largest values in absolute value) are found by the local test.

Finally, we check the relationship between the strengths of significant interactions - in absolute value - and the increase of average MUAe observed in Fig. 8A (lower panel). In Fig. 9E, the plotted dots correspond to the pre-post change in the sum of incoming (left panel) and outgoing (right panel) significant interactions for each node. The summed interaction values positively correlate with the MUAe difference (post minus pre) only for the incoming interactions: $p = 0.03$ with a coefficient of 0.21; nevertheless, the plotted values exhibit a large variability, which moderates this significance. In contrast, outgoing interactions exhibit a non-significant negative correlation ($p \gg 0.1$). This suggests a stimulus-driven gating of the effective gain for incoming anatomical connections to recorded cell populations. The application of our method thus unravels stimulus-driven directed interactions and cannot be merely explained by an increase of single-channel MUAe activity.

In comparison, similar detection with parametric GRc testing gives more interactions for the post-stimulus period in Fig. 9F (bright green), more than twice the number for a p-value of 0.01 (corresponding to 1% in Fig. 9D). Moreover, the distributions of MVAR coefficients in Fig. 9E and the corresponding ones for GRc estimated values have similar KS distance when comparing - for each condition - the pre- and post-periods (mean of 0.17 with a std of 0.01 over the 4 conditions). This means that GRc values collectively discriminate between the two periods as well as the estimated MVAR coefficients. However, the pre-post changes in incoming and outgoing GRc values strongly correlate with the change in mean MUAe ($p < 10^{-20}$ for both incoming and outgoing connections); note that the estimated connectivity is not symmetric, though. This contrasts with the two plots in Fig. 9E and suggests that the two methods may capture distinct effects at work in the network of neuronal populations. Note that the non-parametric method for GRc did not work for the experimental data here, failing to detect more interactions than the expected false-alarm rate.

DISCUSSION

Non-parametric MVAR-based detection of linear feedback in recurrent networks

This paper proposed a non-parametric method to detect pairwise feedback connections in biological networks with possibly strong and/or dense recurrent feedback. We examined the benefit of detecting directional connections in MVAR-like models estimated using the OLS autoregressive coefficients instead of the error residuals ratios (Fig. 1). To our understanding, the good performance of the presented method has three reasons. First, the ROC-based prediction power in Fig. 2, which relies on the estimated ranking of connections in the network (i.e., from small to strong weights), is more robust for the regression coefficients than residual log ratios for recurrent networks with relatively large density ($0.1 - 0.3\%$); these networks overall imply many redundancy and convergence patterns of connections (M. Ding et al., 2006; Stramaglia et al., 2014). Second, practical coefficient-based connectivity detection performed using non-parametric significance tests based on time-series randomization (red distribution in Fig. 4B) yields better results than conditional Granger causality ratio test in the time domain, either with the standard parametric F test (Barnett and Seth (2014), dashed line) or with non-parametric testing (green and blue-green distributions). Finally, the use of connection-specific significance testing achieves higher accuracy than a network-pooled alternative, especially when asymptotic assumptions do not hold (e.g., small number of time samples), as illustrated by the local non-parametric test in Figs. 3F and 4B (red versus dark gray). Together, our results highlight the need for testing strategies that capture the heterogeneity of sufficiently large networks to detect individual connections. Further note that assessing the connectivity via the regression coefficients space brings an additional advantage for network studies: the estimated connection weights can be interpreted and compared across the whole network, for example using graph theory (Sporns, 2013).

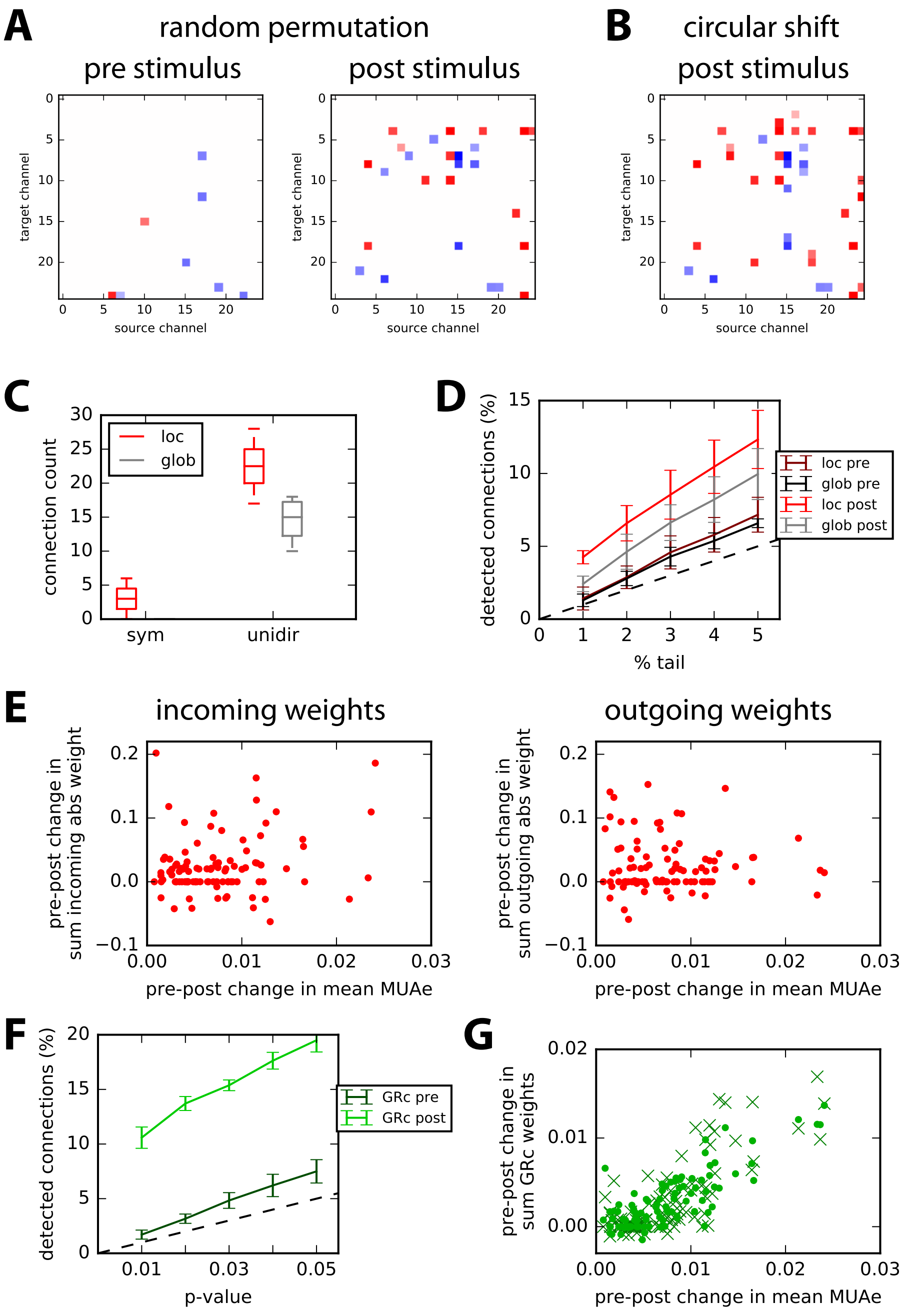


Figure 9. Detection of significant interactions. **A:** Examples of significant interactions with the time-rolling surrogate method; left and right panels correspond to pre- and post-stimulus periods, respectively. The p-value corresponds to the upper and lower 0.5% tails of 1000 surrogates (local test). Many more interactions are found for post- than pre-stimulus period. **B:** Matrix of post-stimulus interactions where green pixels indicate symmetric and purple asymmetric significant interactions; other are left blank. **C:** Comparison of number of asymmetric interactions versus symmetric interactions for the local (red) and global (gray) tests over the 4 conditions. **D:** Ratios of detected interactions for pre- and post-stimulus periods with the desired false-alarm rate equal to 1 to 5% corresponding to both local and global tests. **E:** Change between pre- and post-stimulus periods of the sum of incoming (left) and outgoing (right) significant weights - in absolute value - plotted against the change in the change in mean MUAe. Each dot represents a channel and all 4 conditions are grouped here. **F:** Similar plot to D for parametric GRc. **G:** Similar plot to E with sums of GRc values for each node over its incoming (dots) and outgoing (crosses) connections. Only significant interactions passing the parametric test with p-value of 0.01 in F are retained here.

Our approach for generating surrogate distributions can be encompassed in the family of constrained randomization methods (Schreiber, 1998; Schreiber & Schmitz, 1996). Here we have shown that random permutation provides a good estimation of all types of connections (Fig. 5) in false-alarm and miss rates. Comparatively, the circular-shift method performs as well except for self connections that are not detected at all; this also holds for non-random topologies (results not shown). Hence, preserving the autocovariance structure in the generation of surrogates does not provide a substantial advantage here (Fig. 5B-C). Both methods show a good control for the false-alarm rate in comparison to phase randomization (Schreiber & Schmitz, 1996) and a control Gaussian approximation over the whole network (STD), which lead to an excess of about 1% of false alarms (i.e., ~ 50 connections for a network of 70 nodes). These results show the importance of choosing a surrogate method adapted to the detection problem. For distinct dynamics governing the nodal activity such as nonlinearities, conclusions may differ and further research along these lines is necessary.

As mentioned earlier, the use of an individual null distribution for each connection (local test) gives better results for the miss rate (by a few %) than lumping together all matrix elements of all surrogates (global test), provided sufficiently many surrogates are generated. For the size of networks considered here, computation time is not an issue (Fig. 4D) and our results support the choice of the local test over the global test to attain between accuracy in the true-positive detection. This may be especially true for specific topologies or networks with both excitatory and inhibitory connections, see Fig. 6. In other words, the local test incorporates to a better extent the network heterogeneities in order to build the null distribution for each connection. The present study was limited to ordinary least-square (OLS) estimates for MVAR, but there exist alternative estimators such as the locally weighted least-square regression (Ruppert & Wand, 1994) that may perform better for particular network topologies. The extension of the presented surrogate techniques to the case where observations are sparser than connections - implying that the covariance matrix is not invertible - is another interesting direction to explore (Castelo & Roverato, 2006).

The problem of multiple comparison is intrinsic to brain connectivity detection as the number of testable connections across brain regions is massive (Rubinov & Sporns, 2010). In this context, different approaches have been developed to control the family-wise error rate in the weak sense. For instance, many studies on neuroimaging data (Genovese, Lazar, & Nichols, 2002; Nichols & Hayasaka, 2003) or electrophysiology (Lage-Castellanos, Martínez-Montes, Hernández-Cabrera, & Galán, 2010) have resorted to procedures that control the false discovery rate (FDR) (Benjamini & Hochberg, 1995), namely, the expected number of falsely declared connections among the total number of detections. These methods make decisions on single connections relying on the entire sequence of p-values computed for each connection and yield substantial statistical power gains over more conservative methods such as Šidák-Bonferroni (Abdi, 2007). With the ever growing application of graph theory to brain connectivity, new methods have been proposed that exploit the clustered structure of the the declared connections (Han, Yoo, Seo, Na, & Seong, 2013; Zalesky, Fornito, & Bullmore, 2010) to propose cluster-based statistical tests (Maris & Oostenveld, 2007) that attain similar performance to FDR methods. The present work can therefore be understood as a primary step before performing any or several multiple-correction procedures. By defining an accurate null model of inexistent connections, p-value estimates per connection are improved and cluster-based surrogate distributions can be better approximated, which is expected to empower the overall control of false positive rates in network connectivity analysis.

Applications to real electrophysiological and neuroimaging data

A motivation for our method is the detection of neuronal interactions between electrode channels, which is often performed using spectral Granger causality analysis on local-field potential (low-passed signal of electrode measurements) or ECOG measurements. As an alternative, we have applied our connectivity detection method to MUAe recorded from macaque area V1 in order to provide a proof of concept. Multi-channel recording devices have been developed in the past years to obtain this type of data (Fan et al., 2011; Roy & Wang, 2012) and a recent cognitive study has highlighted group properties of MUAe activity for a similar experimental setup to the one used here Engel et al. (2016). Looking at the variability for individual trials in Fig. 8A (upper and middle panels), it is rather surprising that the MUAe conveys temporal information about joint activity for pairs of channels (Fig. 8B), which can be be related to causal directed interactions. This means that the high trial-to-trial variability is not an absolute limitation to temporal coordination, even though the latter only becomes apparent over multiple trial repetitions, just as the post-stimulus increase of MUAe for single channels (lower panel in Fig. 8A). The procedure detects a number of significant directional interactions well above the false positive rate that were not merely explained by the MUAe increase (Fig. 9D and E). In comparison, the control for the pre-stimulus period in the 4 different conditions detects just above the expected false-alarm rate. We find that the parametric Granger causality test also detects many more interactions after the stimulus than before; however, these interactions happen to link channels exhibiting the strongest increases in MUAe activity. This presents the caveat of providing little information in addition to the changes observed at the single-node level, when interpreting the obtained results. The research of optimal preprocessing - in particular the filtering to obtain MUAe - to obtain a robust detection of interactions is left for a later study. Likewise, such electrode recordings are often analyzed with respect to specific frequency bands (e.g., alpha and gamma), but the adaptation of our framework along the lines of previous works (Dhamala et al., 2008; L. Ding et al., 2007) and comparison of detected interactions with established methods will be done in future research.

More generally, our methodology requires adequate preprocessing of multivariate time series - activity aggregation over hundreds of voxels for fMRI and 4-ms smoothing of MUAe for electrode recordings - such that the autocovariance profiles match the exponential decay of the dynamic network model with linear feedback, which underlies the connectivity analysis. Although filtered and smoothed signals fall into the class of autoregressive moving-average (ARMA) models, our approach is applicable provided the autocovariances exhibit a profile resembling Fig. 8C. We further expect non-parametric testing methods for be extendable to ARMA processes, complementing approaches developed by Barnett and Seth (2015); Friston et al. (2014). In theory, stationarity of the time series remains a critical issue as we need sufficiently many observed samples to obtain precise covariances from which we estimate the connectivity, but this may not be a strong limitation for MUAe in practice in view of the post-stimulus average response shown in Fig. 8A.

ACKNOWLEDGMENTS

MG acknowledges funding from the Marie Sklodowska-Curie Action (grant H2020-MSCA-656547). MG and GD were supported by the Human Brain Project (grant FP7-FET-ICT-604102 and H2020-720270 HBP SGA1). GD and ATC were supported by the European Research Council Advanced Grant DYSTRUCTURE (Grant 295129). AT was supported by the UK Medical Research Council (grant MRC G0700976). The authors are grateful to Robert Castelo and Inma Tur for constructive discussions.

AUTHOR CONTRIBUTIONS

Project was formulated by MG and ATC. Simulation code was developed by MG. Experimental data were provided by AT and XC. Manuscript was written by MG, ATC, AT and GD.

REFERENCES

Abdi, H. (2007). The bonferonni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, *3*, 103-107.

Amemiya, T. (1974). Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica: Journal of the Econometric Society*, 999-1012.

Babiloni, F., Cincotti, F., Babiloni, C., Carducci, F., Mattia, D., Astolfi, L., . . . He, B. (2005). Estimation of the cortical functional connectivity with the multimodal integration of high-resolution eeg and fmri data by directed transfer function. *Neuroimage*, *24*(1), 118-131.

Barnett, L., & Seth, A. (2014). The MVGC multivariate granger causality toolbox: A new approach to granger-causal inference. *J Neurosci Methods*, *223*, 50-68.

Barnett, L., & Seth, A. (2015). Granger causality for state-space models. *Phys Rev E*, *91*, 040101. doi: 10.1103/PhysRevE.91.040101

Barrett, A., & Barnett, L. (2013). Granger causality is designed to measure effect, not mechanism. *Front Neuroinform*, *7*, 6. doi: 10.3389/fninf.2013.00006

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Castelo, R., & Roverato, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *The Journal of Machine Learning Research*, *7*, 2621-2650. Retrieved from http://dl.acm.org/citation.cfm?id=1248641

Dhamala, M., Rangarajan, G., & Ding, M. (2008). Analyzing information flow in brain networks with nonparametric granger causality. *Neuroimage*, *41*(2), 354-362.

Diks, C., & DeGoede, J. (2001). A general nonparametric bootstrap test for granger causality. In (p. 391-403). Bristol, UK: Institute of Physics Publishing.

Ding, L., Worrell, G., Lagerlund, T., & He, B. (2007). Ictal source analysis: localization and imaging of causal interactions in humans. *Neuroimage*, *34*(2), 575-586.

Ding, M., Chen, Y., & Bressler, S. (2006). Granger causality: basic theory and application to neuroscience. In B. Schelter, M. Winterhalder, & J. Timmer (Eds.), *Handbook of time series analysis: Recent theoretical developments and applications.* arXiv preprint q-bio/0608035: Wiley-VCH Verlage.

Engel, T. A., Steinmetz, N. A., Gieselmann, M. A., Thiele, A., Moore, T., & Boahen, K. (2016). Selective modulation of cortical state during spatial attention. *Science*, *354*, 1140-1144.

Faes, L., Marinazzo, D., Montalto, A., & Nollo, G. (2014). Lag-specific transfer entropy as a tool to assess cardiovascular and cardiorespiratory information transfer. *IEEE Trans Biomed Eng*, *61*, 2556-68. doi: 10.1109/TBME.2014.2323131

Fan, D., Rich, D., Holtzman, T., Ruther, P., Dalley, J. W., Lopez, A., … Yin, H. H. (2011). A wireless multi-channel recording system for freely behaving mice and rats. *PLoS One*, *6*, e22033.

Friston, K., Bastos, A., Oswal, A., van Wijk, B., Richter, C., & Litvak, V. (2014). Granger causality revisited. *Neuroimage*, *101*, 796-808. doi: 10.1016/j.neuroimage.2014.06.062

Friston, K., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*(4), 1273-1302.

Genovese, C., Lazar, N., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, *15*(4), 870-878.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *J Am Stat Assoc*, *77*, 304-313.

Geweke, J. (1984). Measures of conditional linear dependence and feedback between time series. *J Am Stat Assoc*, *79*, 907-915.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424-438.

Han, C., Yoo, S., Seo, S., Na, D., & Seong, J.-K. (2013). Cluster-based statistics for brain connectivity in correlation with behavioral measures. *PLoS one*, *8*(8), e72332.

Kaminski, M., & Blinowska, K. (1991). A new method of the description of the information flow in the brain structures. *Biological cybernetics*, *65*(3), 203-210.

Kamiński, M., Ding, M., Truccolo, W., & Bressler, S. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological cybernetics*, *85*(2), 145-157.

Lage-Castellanos, A., Martínez-Montes, E., Hernández-Cabrera, J., & Galán, L. (2010). False discovery rate and permutation test: an evaluation in erp data analysis. *Statistics in medicine*, *29*(1), 63-74.

Li, Y., Ye, X., Liu, Q., Mao, J., Liang, P., Xu, J., & Zhang, P. (2016). Localization of epileptogenic zone based on graph analysis of stereo-eeg. *Epilepsy Res*, *128*, 149-157.

Lusch, B., Maia, P., & Kutz, J. (2016). Physical review e covering statistical, nonlinear, biological, and soft matter physics highlights recent accepted authors referees search press about inferring connectivity in networked dynamical systems: Challenges using granger causality. *Phys Rev E*, *94*, 032220.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177-190.

Massey, J. (1990). Causality, feedback and directed information. In *Proc. int. symp. inf. theory applic.(isita-90)* (p. 303-305).

Messé, A., Rudrauf, D., Benali, H., & Marrelec, G. (2014). Relating structure and function in the human brain: relative contributions of anatomy, stationary dynamics, and non-stationarities. *PLoS Comput Biol*, *10*, e1003530.

Michalareas, G., Schoffelen, J., Paterson, G., & Gross, J. (2013). Investigating causality between interacting brain areas with multivariate autoregressive models of meg sensor data. *Human Brain Mapping*, *34*(4), 890-913. Retrieved from `http://dx.doi.org/10.1002/hbm.21482` doi: 10.1002/hbm.21482

Nedungadi, A., Rangarajan, G., Jain, N., & Ding, M. (2009). Analyzing multiple spike trains with nonparametric granger causality. *J Comp Neu*, *27*(1), 55-64.

Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, *12*(5), 419-446.

Rogers, B., Katwal, S., Morgan, V., Asplund, C., & Gore, J. (2010). Functional mri and multivariate autoregressive models. *Magnetic Resonance Imaging*, *28*(8), 1058-1065.

Roy, S., & Wang, X. (2012). Wireless multi-channel single unit recording in freely moving and vocalizing primates. *J Neurosci Methods*, *203*, 28-40.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, *52*(3), 1059-1069.

Ruppert, D., & Wand, M. (1994). Multivariate locally weighted least squares regression. *Ann Statist*, *22*, 1346-1370.

Schreiber, T. (1998). Constrained randomization of time series data. *Phys Rev Lett*, *80*, 2105.

Schreiber, T. (2000). Measuring information transfer. *Phys Rev Let*, *85*, 461.

Schreiber, T., & Schmitz, A. (1996). Improved surrogate data for nonlinearity tests. *Phys Rev Lett*, *77*, 635.

Seth, A., Barrett, A., & Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *J Neurosci*, *35*, 3293-3297. doi: 10.1523/JNEUROSCI.4399-14.2015

So, K., Koralek, A., Ganguly, K., Gastpar, M., & Carmena, J. (2012). Assessing functional connectivity of neural ensembles using directed information. *J Neu Eng*, *9*(2), 026004.

Sporns, O. (2013). Making sense of brain network data. *Nat Methods*, *10*, 491-493.

Storkey, A., Simonotto, E., Whalley, H., Lawrie, S., Murray, L., & Mcgonigle, D. (2007). Learning structural equation models for fmri. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (p. 1329-1336). MIT Press.

Stramaglia, S., Cortes, J., & Marinazzo, D. (2014). Synergy and redundancy in the granger causal analysis of dynamical networks. *New Journal of Physics*, *16*, 105003.

Supèr, H., & Roelfsema, P. (2005). Chronic multiunit recordings in behaving animals: advantages and limitations. *Prog Brain Res*, *147*, 263-282.

Tauste Campo, A., Martinez-Garcia, M., Nácher, V., Luna, R., Romo, R., & Deco, G. (2015). Task-driven intra-and interarea communications in primate cerebral cortex. *Proc Natl Acad Sci U.S.A.*, *112*(15), 4761-4766.

Thiele, A., Delicato, L., Roberts, M., & Gieselmann, M. (2006). A novel electrode-pipette design for simultaneous recording of extracellular spikes and iontophoretic drug application in awake behaving monkeys. *J Neurosci Methods*, *158*, 207-211.

Vinck, M., Huurdeman, L., Bosman, C., Fries, P., Battaglia, F., Pennartz, C., & Tiesinga, P. (2015). How to detect the granger-causal flow direction in the presence of additive noise? *Neuroimage*, *108*, 301-18. doi: 10.1016/j.neuroimage.2014.12.017

Wilke, C., Ding, L., & He, B. (2008). Estimation of time-varying connectivity patterns through the use of an adaptive directed transfer function. *IEEE Trans Biomed Eng*, *55*(11), 2557-2564.

Winkler, A., Ridgway, G., Webster, M., Smith, S., & Nichols, T. (2014). Permutation inference for the general linear model. *Neuroimage*, *92*, 381-97. doi: 10.1016/j.neuroimage.2014.01.060

www.scipy.org. (n.d.). the scientific python library.

Zalesky, A., Fornito, A., & Bullmore, E. (2010). Network-based statistic: identifying differences in brain networks. *Neuroimage*, *53*(4), 1197-1207.