G Model
NSM-5988; No. of Pages 9
ARTICLE IN PRESS

Journal of Neuroscience Methods xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Single trial classification of magnetoencephalographic recordings using Granger causality

Wojciech Kostelecki[a,b,*], Luis Garcia Dominguez[a], José Luis Pérez Velázquez[a,b,c]

[a] Neuroscience and Mental Health Program, Hospital for Sick Children, 555 University Avenue, Toronto, ON M5G1X8, Canada
[b] Institute of Medical Science, University of Toronto, 1 King's College Circle, Toronto, ON M5S1A8, Canada
[c] Department of Paediatrics, Hospital for Sick Children, 555 University Avenue, Toronto, ON M5G1X8, Canada

ARTICLE INFO

Article history:
Received 11 March 2011
Received in revised form 18 April 2011
Accepted 21 April 2011

Keywords:
Granger causality
Autoregression models
Classification
Magnetoencephalography
Short trials

ABSTRACT

The use of Granger causality (GC) for studying dependencies in neuroimaging data has recently been gaining popularity. Several frameworks exist for applying GC to neurophysiological questions but many rely heavily on specific statistical assumptions regarding autoregressive (AR) models for hypothesis testing. Since it is often difficult to satisfy these assumptions in practical settings, this study proposes an alternative statistical methodology based on the classification of individual trials of data. Instead of testing for significance using statistics based on estimated AR models or prediction errors, hypotheses were tested by determining whether or not individual magnetoencephalography (MEG) recording segments belonging to either of two experimental conditions can be successfully classified using features derived from AR and GC concepts. Using this novel approach, we show that bivariate temporal GC can be used to distinguish button presses based on whether they were experimentally *forced* or *free*. Additionally, the methodology was used to determine useful parameter settings for various steps of the analysis and this revealed surprising insight into several aspects of AR and GC analysis which, previously, could not be obtained in a comparable manner. A final mean accuracy of 79.2% was achieved for classifying *forced* and *free* button presses for 6 subjects suggesting that classification using GC features is a viable option for studying MEG signals and useful for evaluating the effectiveness of parameter variations in GC analysis.

1. Introduction

Considerable insight into cognitive processes is obtained by studying how neuroanatomical structures interact while subjects perform experimental tasks or exhibit various mental states. Magnetoencephalography (MEG) is being used increasingly to studying these types of neural processes and how they support high level brain function. A notable advantage of MEG that has made it popular is the degree of temporal precision with which important neural events can be detected. With sophisticated analysis, useful information can be extracted from MEG signals and new methodologies are constantly being developed and built upon to better exploit data-rich neuroimaging methodology and to reveal the neurophysiology underlying brain states of interest.

Various metrics exists for quantifying relationships between two or more neural signals. One such measure that is gaining popularity is Granger causality (GC) which is a proven useful tool in many neuroimaging contexts (Bressler and Seth, 2010). It provides a straightforward estimation of the interdependencies between neural signals with very little *a priori* knowledge about the specifics of those relationships. However, the statistical analysis for evaluating GC is not always straightforward because many key model assumptions are difficult to satisfy and, as a result, adaptations need to be considered. To avoid some of these problems, this study proposes a novel classification methodology for distinguishing the underlying experimental conditions of single trials of MEG data using features based solely on the concept of autoregression (AR) and temporal GC.

Classification of neural signals is an approach that is gaining popularity (Soon et al., 2008; Mitchell et al., 2004; Pereira et al., 2009) due to its ease of implementation and flexibility for hypothesis testing. However, there is no clear procedure for implementing a classification analysis based on GC. In particular trials that are short in duration are not suitable for estimating AR parameters and this often results in ineffective estimation of GC between time series. As a result, classification can be poor or the exact relationships in the data cannot be known with great certainty. In this study, a new method is developed that allows for classification of neural signals using a GC feature that effectively indicates the experimental condition to which an individual MEG trial belongs.

* Corresponding author.
E-mail address: w.kostelecki@gmail.com (W. Kostelecki).

0165-0270/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2011.04.032

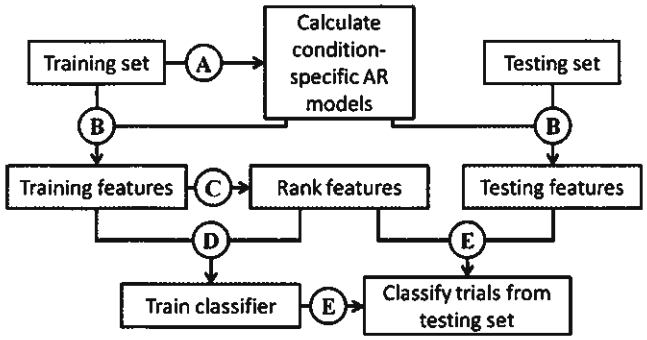

Fig. 1. Overview of the GC classification methodology. The schematic corresponds equally well to one fold of the inner cross-validation loop (Section 3.6.1) and the final cross-validation step (Section 3.6.2). The dataset was partitioned into a training and testing set. The training set was used to estimate condition-specific univariate AR models for all MEG sensors and bivariate AR models for all sensor pairs (A). Using these AR models, features were calculated for individual trials using Eq. (5) (B). All sensor pairs were ranked according to the *t*-statistics between the *forced* and *free* samples (C) to determine which possessed the greatest differences in feature distribution. The classifier was trained using the N_F highest ranking features (D). Classification of trials from the testing set was performed to obtain an estimate of the classification accuracy (E).

2. Methods

2.1. Overview

The methodology developed for classifying single trials of MEG data using GC is outlined in Fig. 1. Prior to each step, the data was partitioned into a training and testing set. Condition-specific univariate and bivariate AR models were then estimated for all MEG sensors and sensor pairs for data belonging to the training set (Fig. 1A). The estimated AR models were then used to calculate GC features (using Eq. (5)) for individual trials belonging to both the training and testing sets (Fig. 1B). GC features from the training set were ranked (Fig. 1C) according to the *t*-statistic between the samples from the two experimental conditions and a naïve Bayes classifier was then trained using the highest ranking features (Fig. 1D). Trials from the testing set were then classified using the same high-ranking features (Fig. 1E).

In this study, the classification analysis was performed using nested cross-validation to initially determine reasonable parameter settings for various aspects of the procedure. These parameters included the number of features to use for classification, AR model order, MEG signal sampling rate, and the number of principal components to remove as a form of artifact deletion. After the best parameter settings were obtained from this inner cross-validation procedure, the resulting classifier was applied to a withheld testing set in a final outer cross-validation step to estimate the resulting effectiveness of the classifier.

2.2. MEG dataset

This study used an existing MEG dataset from the experiments described by Dominguez et al. (2011) in which subjects performed *left*- and *right*-handed button presses in response to cues belonging to two conditions. The *forced* cue specifically instructed the subject which hand to use for the button press and the *free* cue instructed the subject to press a button with the hand of their choosing. In subsequent notation, the *forced* and *free* conditions are denoted by $y=-1$ and $y=1$ respectively. Subjects performed this task while undergoing MEG recordings which were segmented into short windows that were believed to possess the critical information useful for classification of these trials: 50 ms after cue presentation until 25 ms before the subject's average response time. Subjects consisted of 5 males and 1 female, all with right-hand dominance and with a mean age of 33. Subject-specific information about the datasets is given in Table 1. Classification of *forced* and *free* conditions was performed for *left* and *right* button presses separately to determine whether *forced* and *free* button presses could be distinguished without the side of the button press confounding the analysis. In addition, the *combined* analysis was performed to determine whether *forced* and *free* button presses could be classified despite the training and testing sample of trials containing button presses using both hands.

3. Theory/calculation

3.1. Basic notation

In this article, multiple simultaneous MEG recordings will be denoted by the $N_{MEG} \times T$ matrix

$$\mathbf{x} = [x_1, \ldots, x_{N_{MEG}}]^T$$

where x_i is a $T \times 1$ vector representing a MEG recording sampled at times $t = 1, \ldots, T$ and N_{MEG} is the total number of MEG sensors. Additionally, the superscript (n, y) will be used to denote data from a specific trial and experimental condition (*e.g.* $x_i^{(n,y)}(t)$ denotes the data point for trial n, MEG sensor i, time point t, and with label y).

3.2. Autoregression

AR is a statistical tool commonly used to model time-lagged dependencies in neuroimaging data. It is typically the case that MEG signals possess some of the properties of autoregressive processes in that an observation of a signal at a given time, t, can be explained using a scaled linear combination of the immediate signal history as well as a random innovations process. However, a strict correspondence between AR models and neural time series usually does not exist since the neural mechanisms underlying brain processes cannot be simplified to this type of relationship. Despite this, AR models and GC have been shown to describe relationships

Table 1
Subject-specific information about the analyzed dataset. Number of trials per fold indicates the summed number of trials in both the *forced* and *free* conditions. In all cases, folds contained equal amounts of *forced* and *free* button presses (as well as *left*- and *right-handed* button presses in the case of the *combined* analysis). The number of data points in each trial is given at the original MEG sampling rate (625 Hz) although, for the analysis, the length of each trials varies with M (as described in Section 3.9.2). The number of MEG sensors, N_{MEG}, is provided to indicate the number of bivariate comparisons available in each dataset.

Subject	Number of trials per fold			Length of each trial		Number of MEG sensors
	Left	Right	Combined	Data points	ms	
1	54	52	104	247	395	148
2	56	54	108	142	227	143
3	42	46	84	266	426	146
4	34	38	68	311	498	146
5	62	64	124	143	229	143
6	42	60	84	160	256	148
Mean	48.3	52.3	95.3	211.5	338.4	145.7

Please cite this article in press as: Kostelecki W, et al. Single trial classification of magnetoencephalographic recordings using Granger causality. J Neurosci Methods (2011), doi:10.1016/j.jneumeth.2011.04.032

in data that does not necessarily follow the exact form of an AR process (Schelter et al., 2006a,b) and it is now accepted that applying AR models to neuroimaging data often uncovers time-dependent relationships that are otherwise difficult to detect. If applied appropriately, AR and GC can be useful for studying the underlying neural pathways that generate MEG signals and inferring the neural mechanisms underlying various behaviours, cognitive states or pathologies. Problems arise, however, when formulating statistical frameworks for GC analysis with which hypotheses can be tested since many assumptions related to AR models are difficult to satisfy in a neuroimaging context. Among other problems, datasets are often incomplete and there is no guarantee the sought after neural signals and relationships can even be uncovered. At best, a method using AR and GC should utilize all available data to uncover possible relationships between neural signals that may be scattered and difficult to detect.

3.2.1. AR formulation

For the univariate case, an AR process is described by the equation

$$x_i(t) = \sum_{r=1}^{R} a_i(r)x_i(t-r) + \varepsilon_i(t)$$

where r is the time lag, R is the model order, $\{a_i(r)\}_{r=1}^{R}$ are the AR model coefficients, and $\varepsilon_i(t)$ is a random innovations process. AR models can be generalized to the multivariate case but, in practice bivariate models are commonly used for simplification purposes. A bivariate AR process is described by the equation

$$\begin{bmatrix} x_i(t) \\ x_j(t) \end{bmatrix} = \sum_{r=1}^{R} \begin{bmatrix} a_{ii}(r) & a_{ij}(r) \\ a_{ji}(r) & a_{jj}(r) \end{bmatrix} \begin{bmatrix} x_i(t-r) \\ x_j(t-r) \end{bmatrix} + \begin{bmatrix} \varepsilon_i(t) \\ \varepsilon_j(t) \end{bmatrix} \quad (1)$$

where, in this case, the immediate history of a second signal is also incorporated into the model. Describing MEG signals in terms of this framework has provided useful insight into the neural mechanisms underlying various brain states (Bressler et al., 2007, 2008; Brovelli et al., 2004; Gow et al., 2008; Supp et al., 2007) and has also been useful with neuroimaging methodologies other than MEG and EEG (Goebel et al., 2003). As described in Fig. 1A, univariate and bivariate AR models of these forms were estimated for all MEG sensors and sensor pairs.

Due to the short duration of the trials used in this study, AR parameters were estimated using ordinary least squares regression to minimize the prediction error for data after the first R datapoints of a trial. This way, AR coefficients for all lags received an identical amount of data to support their estimate. However, many alternative AR model estimating algorithms exist that scale better with the number of time series used and model order (Schlögl, 2006) and may be more suitable in other situations.

3.3. Granger causality formulation

Commonly, GC is defined around the AR framework. This is accomplished by first, fitting a univariate AR model to a time series, x_i and a bivariate AR model to $[x_i, x_j]$ to obtain model coefficients, $\{\hat{a}_i(r), \hat{a}_{ii}(r), \hat{a}_{ij}(r)\}_{r=1}^{R}$. The estimated AR model coefficients minimize the variance of the prediction errors, $\varepsilon_{i|i}$, by the equation

$$\varepsilon_{i|i}(t) = x_i(t) - \sum_{r=1}^{R} \hat{a}_i(r)x_i(t-r) \quad (2)$$

Table 2
AR parameters used to generate the artificial bivariate AR time series analyzed in Fig. 2. The subscript, y, denotes one of two distinct AR processes used to generate artificial data.

Lag (r)	a_{11}	$a_{12\|y=1}$	$a_{12\|y=-1}$	a_{22}	a_{21}
1	0.4	0.2	0.1	0.4	0
2	0.2	0.1	0.3	0.2	0
3	0.1	0.3	0.1	0.1	0
4	0	0.1	0.2	0	0

and that of $\varepsilon_{i|ij}$ by equation

$$\varepsilon_{i|ij}(t) = x_i(t) - \sum_{r=1}^{R} \hat{a}_{ii}(r)x_i(t-r) - \sum_{r=1}^{R} \hat{a}_{ij}(r)x_j(t-r). \quad (3)$$

If the variance of the prediction errors for the bivariate model, $\sigma^2_{i|ij}$, is statistically lower than the variance of the prediction errors for the univariate model, $\sigma^2_{i|i}$, then the inclusion of x_j provides some added benefit for estimating x_i. GC is defined by the presence of this type of relationship and is often summarized by the equation

$$F_{j\to i} = \log \frac{\sigma^2_{i|i}}{\sigma^2_{i|ij}}. \quad (4)$$

If $F_{j\to i} > 0$, then x_j is said to Granger cause x_i and *vice versa*. Since the datasets being used in this study contain over 140 recording sensors for each subject, it is not always clear which combination of sensors should be considered simultaneously in a multivariate AR model. As a result, the analysis was restricted to all bivariate AR relationships to reduce the computational requirements. Although not discussed here, generalizations of GC to blocks of time series (Wang et al., 2007) and conditional GC (Geweke, 1984; Guo et al., 2008; Zhou et al., 2009) (as well as combinations of both approaches (Barrett et al., 2010)) exist and may be useful in some contexts.

3.4. Extension of GC to short trial classification

Short time series are of particular interest in cognitive neuroscience because brain states that support cognition and behaviour can be transient and distinct from baseline neural activity. Typically, measures of GC can be applied to short signals by treating each trial as a realization of a single AR process and estimating common AR models to entire sets of trials (Ding et al., 2000). However, due to insufficient data, AR models and GC cannot be estimated accurately for individual short trials. In this case, AR model estimation would result in coefficients with considerable uncertainty, which, although may be used to classify trials with some success, fails to reliably provide details about the underlying relationships in the data.

To illustrate this point, AR coefficients were specified (given in Table 2) and used to generate artificial datasets for two distinct processes using Eq. (1) and Gaussian innovations processes with a variance of 1. For both conditions, $\{a_{11}(r), a_{21}(r), a_{22}(r)\}_{r=1}^{R}$ were held constant, but two different sets of coefficients were used for $\{a_{12}(r)\}_{r=1}^{R}$ and were labeled $y = -1$ and $y = 1$. Artificial datasets of various lengths were generated and GC was calculated using several approaches. For every trial length, 200,000 instances of the AR process were generated and the calculated GC distribution was displayed in Fig. 2A–C as a normalized histogram and its interquartile range for each length of time series.

In Fig. 2A, GC for time series with labels $y = 1$ ($F^{(y=1)}_{2\to 1}$) is calculated using the standard procedure in which a univariate and bivariate AR model is estimated for every individual trial and GC is subsequently calculated using the prediction errors. The empirical mean of $F^{(y=1)}_{2\to 1}$ (0.26; as determined from Fig. 2B at $\log_2(T) = 11$) is indicated by

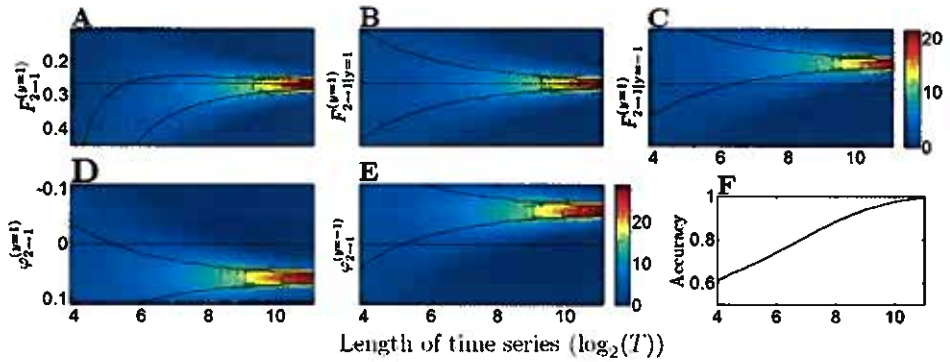

Fig. 2. Empirical probability densities of GC and GC feature estimates for simulated AR[4] time series of variable length. Empirical probability densities for three types of GC estimates: $F_{2\to1}^{(y=1)}$ (A), $F_{2\to1|y=1}^{(y=1)}$ (B), and $F_{2\to1|y=-1}^{(y=1)}$ (C) with empirical mean from B at $\log_2(T)=11$ indicated by a horizontal dotted line. Empirical probability densities for GC feature estimates: $\varphi_{2\to1}^{(y=1)}$ (D) and $\varphi_{2\to1}^{(y=-1)}$ (E) with classification boundary indicated by a horizontal dotted line. (F) Accuracy with which classification using $\varphi_{2\to1}^{(y=1)}$ and $\varphi_{2\to1}^{(y=-1)}$ was possible.

the horizontal dotted line here, as well as in Fig. 2B and C. As the time series become shorter, AR models become overparameterized, and consequently GC estimation, becomes both less precise and less accurate. This hinders the analysis of individual trials using common GC methods because reasonable estimates of GC cannot be obtained.

In Fig. 2B, GC is calculated for time series with labels $y=1$ by estimating an AR model for a large training set of trials having the same label ($y=1$), using the resulting AR models to calculate the residual errors for the individual trials using Eqs. (2) and (3), and using these results to calculate GC with Eq. (4). GC calculated in this manner is denoted by $F_{2\to1|y=1}^{(y=1)}$ where the subscript $y=1$ indicates the label of the trials in the large training set used to calculate the AR models. The training sets used to estimate the AR models were constructed from 130 generated time series, each 25 data points in length and are comparable in size to the MEG training sets used in this study. Using this approach, more accurate estimates of GC are obtained for shorter trial lengths although the GC distribution still becomes broader as time series for which GC is estimated become shorter.

In Fig. 2C, GC is calculated for time series with labels $y=1$ but, this time, using an AR model estimated from a training set of trials with the opposite label ($y=-1$). In this case, GC estimates are no longer centered on the expected GC measure for $y=1$, nor are they centered on the expected GC for $y=-1$. Due to the selection of the AR coefficients in Table 2, the empirical probability densities for $F_{2\to1}^{(y=-1)}$, $F_{2\to1|y=-1}^{(y=-1)}$, and $F_{2\to1|y=1}^{(y=-1)}$ are nearly identical to Fig. 2A–C, respectively, so these were omitted. From Fig. 2B and C, one can see that if the correct AR model is used to calculate GC, the result is typically an accurate estimate of the expected GC. If, however, an incorrect AR model is used to calculate GC, the result does not always approximate the desired measure of GC and can even be negative in extreme cases. As a result, one could conceivably distinguish which of the two experimental conditions a trial belongs to by comparing GC evaluated under the AR models estimated from both experimental conditions. A feature for comparing individual trials using two classes of AR models is described in the next section.

3.5. GC feature for classification

To determine the correct label of a trial from the testing set, $F_{j\to i|y=1}$ and $F_{j\to i|y=-1}$ were evaluated using the AR models estimated from the training set. The differences of these two measures of GC was then used as the classification feature which is given by the equation

$$\varphi_{j\to i} = F_{j\to i|y=1} - F_{j\to i|y=-1}. \quad (5)$$

Although this expression can be simplified further, and other variations on the feature may prove more effective for classification, this feature was introduced because it gives an intuitive metric using familiar concepts: it indicates whether GC is best described by one set of AR parameters or another. If both sets of AR models describe a trial equally well (or equally poorly), the distributions of $\varphi_{j\to i}^{(y=-1)}$ and $\varphi_{j\to i}^{(y=1)}$ typically overlap and cannot be used for classification purposes. If, on the other hand, the AR models estimated for trials labeled $y=1$ are more suitable for describing an unlabeled trial, $\varphi_{j\to i}>0$, suggesting that $y=1$ is the likely label (and *vice versa* for trials labeled $y=-1$). As a result, time series exhibiting AR relationships that differ between two experimental conditions can be classified using this feature.

In Fig. 2D and E, the distributions of $\varphi_{j\to i}^{(y=1)}$ and $\varphi_{j\to i}^{(y=-1)}$ are shown with clear separation. As a result, trials belonging to one condition can be classified, even at short time series lengths. Notice that this method allows one to distinguish trials that have nearly identical levels of GC but different underlying AR relationships explaining the data, providing added novelty and utility for different types of analysis using GC. Fig. 2F shows the classification accuracy for trials of the specified length suggesting that even trials 16 datapoints in length can be classified with over 60% accuracy. This example, however, uses only two time series that possess a unidirectional relationships. In real neuroimaging situations, datasets are frequently multivariate with many more potential relationships between signals from any combination of recording sensors. As a result, even with individually weak features, a relatively strong classifier can be constructed that takes advantage of the large dataset.

For this study, all sensor pair combinations (in both directions) were included in the initial pool of features before the highest ranking ones were selected for inclusion in the final classifier (*i.e.* Step B in Fig. 1 was performed by calculating the feature, $\varphi_{j\to i}^{(n)}$, for every trial, n, and every pair of sensors, $j\to i$).

3.6. Classification using cross-validation

Since classification using GC is a novel methodology, a nested cross-validation procedure was used to initially fine-tune several parameters of the analysis and then to subsequently verify the classification model on withheld data. For every subject, data was randomly divided into five folds of identical size and with identical amounts of trials from each experimental condition with leftover trials getting discarded. The total number of trials per fold varied across subjects and button groupings and ranged from 34 to 124 (see Table 1). An inner cross-validation loop (Section 3.6.1) with

Please cite this article in press as: Kostelecki W, et al. Single trial classification of magnetoencephalographic recordings using Granger causality. J Neurosci Methods (2011), doi:10.1016/j.jneumeth.2011.04.032

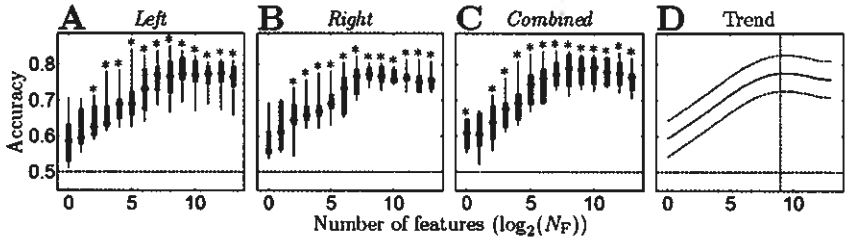

Fig. 3. Classification accuracies obtained by using a variable number of features and holding $R = 4$, $M = 14$, and $N_{PC} = 13$ constant. Plots show the spread (for the 6 subjects) of the classification accuracies for *left* (A), *right* (B), and *combined* button presses (C). Whiskers represent ranges, boxes represent interquartile ranges and circle markers represent medians. Asterisks denote significant mean classification accuracies (*t*-statistic with $p < 0.05$; adjusted for multiple comparisons using the Dunn–Šidák correction). (D) Least-squares third degree polynomial fit to all data in A–C. Dotted lines around the trend line denote 50% confidence intervals and vertical dotted line denotes $N_F = 512$.

which parameters were optimized was performed using four folds and the final estimation of classification accuracy was performed on the remaining fold (Section 3.6.2). In all cases, naïve Bayes classification was used as the classification procedure for simplicity and the classifier was trained with and applied to data from each subject separately.

3.6.1. Inner cross-validation procedure

For each loop of the inner (parameter-tuning) cross-validation procedure, the four inner cross-validation folds were divided into a training set consisting of three folds and a testing set consisting of the remaining fold. The classification was performed using a specific combination of parameter settings and was repeated identically for the other three segmentations of the inner folds. The overall classification accuracy of the inner cross-validation loop was obtained from the average classification accuracy of the four inner cross-validation iterations. These classification accuracies are displayed in Figs. 3–5 for ranges of parameter settings. The parameters that were investigated include the number of features to use for classification, (Section 3.9.1, Fig. 3), the downsampling factor for resampling the MEG signal and the AR model order (Section 3.9.2, Fig. 4), and the number of principal components to delete for artifact removal purposes (Section 3.9.3, Fig. 5).

To obtain reasonable parameter settings, some parameters were tested while holding the remaining parameter(s) constant to reduce the computational requirements of this analysis. The number of deleted principal components was initially held constant (at $N_{PC} = 0$) while all indicated combinations of the number of features (N_F), AR model order (R), and downsampling factor (M) were tested to obtain initial guesses for the latter three variables. The inner cross-validation procedure was then continued for a range of N_{PC} using the N_F, R, and M settings obtained from the previous step. Finally, the procedure was repeated for the same initial range of settings for N_F, R, and M with N_{PC} set to its updated value. No further steps were required at this point because nearly identical optimal settings were reached for N_F, R, and M for the updated value of N_{PC}.

Table 3
Classification accuracies obtained from outer step of the cross-validation procedure described in Section 3.6.2. Means are shown for each subject and for each set of trials that were analyzed. Non-averaged classification accuracies are all statistically significant (binomial probabilities of $p < 0.003$).

Subject	Classification accuracies (%)			
	Left	Right	Combined	Mean
1	74.1	82.7	76.9	77.9
2	82.1	81.5	81.5	81.7
3	78.6	89.1	79.8	82.5
4	79.4	81.6	79.4	80.1
5	71.0	70.3	69.4	70.2
6	88.1	81.7	78.6	82.8
Mean	78.9	81.1	77.6	79.2

3.6.2. Outer cross-validation step

After parameters settings were obtained from the inner cross-validation procedure, the four folds from that step were used as the training set for training a final classifier whose accuracy was tested on the remaining withheld fold. The classification accuracies obtained for these final folds are given in Table 3.

3.7. Feature selection

The features described by Eq. (5) were calculated for every trial and sensor pair (separately for both directions) in the training set. Features for specific sensor pairs were selected for training the classifier if the *t*-statistic between the two condition-specific samples of GC feature estimates ranked highest among all sensor pairs (Step

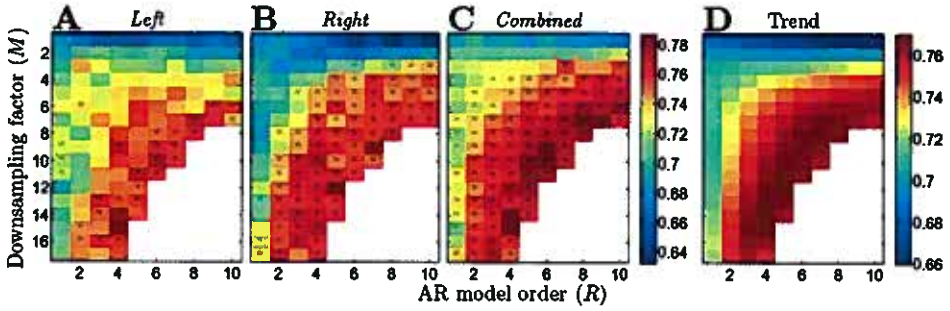

Fig. 4. Classification accuracies obtained using variably downsampled MEG data and various AR model orders while holding $N_F = 512$ and $N_{PC} = 13$ constant. Plots show mean classification accuracies for *left* (A), *right* (B), and *combined* button presses (C). Asterisks denote significant mean classification accuracies (*t*-statistic with $p < 0.05$; adjusted for multiple comparisons using the Dunn–Šidák correction). (D) Least-squares surface fit to A–C using a bivariate polynomial of the form $f(R, M) = \sum_{h,k \geq 0, h+k \leq 4} a_{hk} R^h M^k$. White space indicates $\{R, M\}$ combinations for which trials were less than $2R$ data points in length for at least one subject.

Please cite this article in press as: Kostelecki W. et al. Single trial classification of magnetoencephalographic recordings using Granger causality. J Neurosci Methods (2011), doi:10.1016/j.jneumeth.2011.04.032

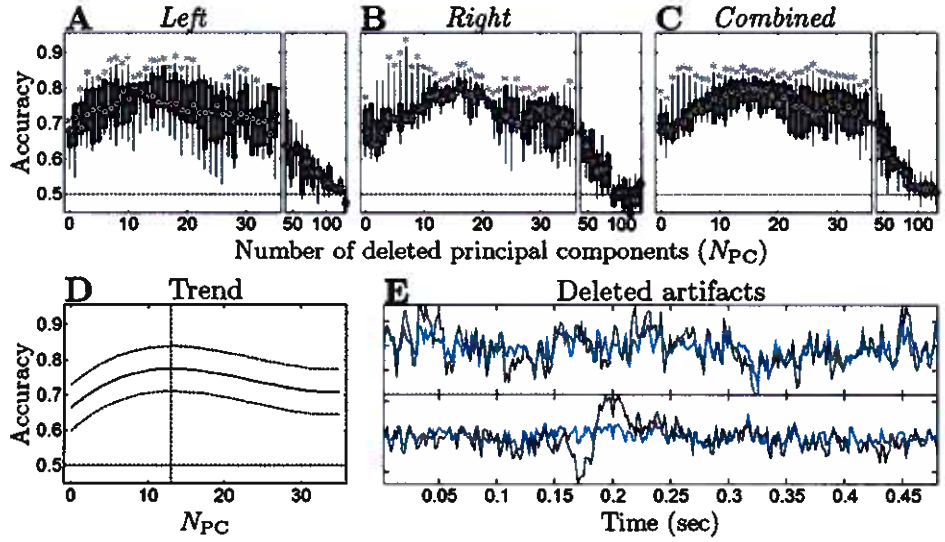

Fig. 5. Classification accuracies obtained for data with first N_{PC} principal components removed while holding $N_F = 512$, $R = 4$, and $M = 14$ constant. Plots show spread of classification accuracies for *left* (A), *right* (B), and *combined* button presses (C). Whiskers represent ranges, boxes represent interquartile ranges and circle markers represent medians. Asterisks denote significant mean classification accuracies (t-statistic with $p < 0.05$; adjusted for multiple comparisons using the Dunn–Šidák correction). (D) Least-squares third degree polynomial fit to all data in A–C. Dotted lines around the trend line denote 50% confidence intervals and vertical dotted denotes $N_{PC} = 13$. (E) Illustrative example of the effects of the artifact-removal procedure (using $N_{PC} = 13$) for a medial precentral MEG sensor with no obvious artifact (top panel) and a left frontal sensor with a blink artifact starting at 0.15 s (bottom panel). Both panels show the original signal in black and the processed signal in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

C in Fig. 1). This method for feature selection is commonly used with reasonable levels of success (*e.g.*, Pereira et al., 2009). The highest ranking directed sensor pairs were used to train the naïve Bayes classier (Step D in Fig. 1) and to subsequently classify the trials in the testing set (Step E in Fig. 1).

3.8. Naïve Bayes classification

Classification of individual trials is an approach that is gaining popularity for the analysis of large datasets and naïve Bayes classification—due to its simplicity—is commonly used as a basic procedure to perform this type of analysis. Although it is expected that a more sophisticated classification procedure would provide improved classification results in this analysis, this is not the primary purpose of this study, but rather, it is to illustrate the usefulness of the GC feature for classification. Naïve Bayes classification was performed by calculating the means and variances ($\mu_{j\to i}$ and $\sigma^2_{j\to i}$) of the training set feature distributions (assumed to be Gaussian) for both experimental conditions. The classification of an individual trial was then performed using the selected features which are denoted by the feature vector, $\Phi = [\varphi_{j\to i}, \ldots]$, containing the features for all selected sensor pairs. Trials were assigned to conditions reporting the larger posterior probability according to the equation

$$\hat{y} = \arg\max_{k\in\{-1,1\}} P(y = k|\Phi)$$

where the posterior probability for each label is given by

$$P(y|\Phi) = \frac{P(y)P(\Phi|y)}{P(\Phi)},$$

with the prior set to $P(y) = 0.5$ due to the construction of the experiment and datasets, the likelihood given by

$$P(\Phi|y) = \prod_{(j\to i)} P(\varphi_{j\to i}|\mu_{j\to i|y}, \sigma^2_{j\to i|y}),$$

and

$$P(\Phi) = P(y = -1)P(\Phi|y = -1) + P(y = 1)P(\Phi|y = 1).$$

3.9. Optimization of parameters

3.9.1. Number of features

Classification was performed with a variable number of features for training and classifying with the naïve Bayes classification procedure to find a concise set of sensor pairs that describe the spatial regions involved in distinguishing the two conditions under which MEG data was collected. Fig. 3A–C shows the spread of the results for this step for *left*, *right*, and *combined* sets of button presses when $R = 4$, $M = 14$, and $N_{PC} = 13$ are held constant. A setting of $N_F = 512$ features was determined from Fig. 3D to be sufficient for obtaining the best classification rates and this amount comprises approximately 2% of the total number of features available for each subject.

3.9.2. Time series downsampling and AR model order

When fitting AR models to real data, it is not always obvious what an appropriate sampling rate is for a signal of interest or how a sampling rate affects the modeling of a signal using AR (Florin et al., 2010). Furthermore, there are also uncertainties for determining an appropriate AR model order. Common guidelines suggest using a sampling rate of approximately 200 Hz and determining the AR model order using the Akaike or Bayesian information criterion (AIC and BIC) (Ding et al., 2006) which are based on statistical assumptions about the form of the prediction errors. However, very little justification exists for these guideline in most neuroimaging settings. AIC and BIC perform very well when determining the model order of AR processes but the underlying mechanisms governing the neural processes that produce detectable neural signal are not AR processes. The motivation for using AR models in neural signal contexts is to try to understand dependencies between multiple data series when very little is known about those dependencies *a priori*. For this reason the structure of the AR parameters can reveal interesting relationships in data but any biophysical relevance is

Please cite this article in press as: Kostelecki W, et al. Single trial classification of magnetoencephalographic recordings using Granger causality. J Neurosci Methods (2011), doi:10.1016/j.jneumeth.2011.04.032

speculative. As a result, AIC and BIC may not be the most useful for determining suitable AR model orders for neural signals. Instead, the method used here determines the model order that best distinguishes two classes of MEG signals that are expected to have fundamentally different underlying neural origins.

In this study, AR models of various orders are estimated for data sampled at various rates (achieved by signal decimation) to determine the optimal settings of these two parameters when classifying trials. Better classification was assumed to be a result of the improved ability of the AR parameters to capture characteristic properties of the data that allowed for *forced* and *free* MEG signals to be distinguished.

The downsampling was performed similarly to that described by Florin et al. (2010): First, an eighth-order low-pass Chebyshev Type I filter with a cutoff frequency of $250/M$ Hz was applied to the signal. The resulting smoothed time series was then resampled at a rate M times lower than the original sampling frequency (625 Hz).

Fig. 4A–C shows the mean classification accuracy at various parameter settings for *left*, *right*, and *combined* sets of button presses when $N_F = 512$ and $N_{PC} = 13$ are held constant. From the least squares trend shown in Fig. 4D, the best parameter settings were determined to be $R = 4$ and $M = 14$. However, a range of parameter values resulted in comparable classification rates which included downsampling factors ranging between 6 and 14 and AR model orders ranging from 4 to 9. In most cases, however, the combination of downsampling factor and AR model order that gave the best performance consisted of a model extending approximately 90 ms into the history of the signal.

3.9.3. Artifact deletion using principal component analysis

Artifacts are commonplace in neuroimaging data and many precautions are taken to obtain clear access to neural signals by removing known MEG sources of contamination such as blinks, eye movement, and heart rhythms. In this study, visual inspection revealed that the recordings possessed occasional artifacts, primarily in the form of short deviations from the baseline signal in frontal MEG sensors during eye blinks (*e.g.* black trace in Fig. 5E, bottom panel). Initial observations also revealed that these were not isolated to the first few principal components and varied across subjects such that no clear systematic approach existed to remove the artifacts. To determine the appropriate extent of preprocessing, a similar method to that outlined by Jung et al. (2000) was used. PCA was performed and artifacts were deleted by reconstructing signals using the last $N_{MEG} - N_{PC}$ principal components according to the equation

$$\tilde{x}_i = \sum_{i=N_{PC}+1}^{N_{MEG}} a_{ij} b_j$$

where b_j is the $T \times 1$ vector corresponding to the jth principal component (ordered from highest to lowest variance), a_{ij} is the weight of principal component j in signal i, and $\tilde{x}_i$ is the reconstruction of signal x_i. In practice, this step was always performed prior to any downsampling of the signal.

Due to uncertainty about how many principal components represent artifacts, increasing numbers of principal components (up to the first 35 principal components) were removed and the classification was performed to see how accuracy changed with the varying degree of preprocessing. Fig. 5A–C shows the resulting spread of classification accuracies for *left*, *right*, and *combined* sets of button presses when $N_F = 512$, $R = 4$, and $M = 14$ are held constant. The least squares fit to the data (Fig. 5D) shows that the classification accuracy is highest at $N_{PC} = 13$.

An illustrative example of effects of the artifact-removal procedure on MEG signals are shown in Fig. 5E. Removal of the first 13 principal components appears to effectively remove artifacts (Fig. 5E; bottom panel) while only moderately altering signals that do not possess obvious artifacts (Fig. 5E; top panel).

Classification was subsequently performed for $N_{PC} = 40, 48, \ldots, 128$ to confirm that the classification accuracy drops to 50% when enough principal components are discarded. Results for this are shown in panels to the right of Fig. 5A–C.

4. Results

4.1. Final parameter settings

Manipulation of the described parameter settings all had affects on the resulting classification rates, suggesting that there is a need for a statistical approach that allows for this type exploration of the parameter space and subsequent evaluation of the suitability of the determined settings. The optimal settings achieved were not always obvious and previously, no well-established method existed with which to measure the effectiveness changes in certain parameters improve or worsen the description of time series using GC. As determined by the inner cross-validation procedure, classification of short MEG trials using GC features and naïve Bayes classification accuracy was highest when 512 features were used, data was downsampled by a factor of 14 (resulting in a sampling rate of approximately 45 Hz), an AR model order of 4 was used, and the first 13 principal components were discarded.

4.2. Final cross-validation results

The final step of the cross-validation procedure was performed using the estimated parameters of $R = 4$, $M = 14$, $N_{PC} = 13$ and $N_F = 512$ (from Section 3.9). This, again, was repeated separately for every subject and for *left*, *right*, and *combined* sets of button presses. The resulting classification accuracies, as well as the means across subjects and sets of button presses, are shown in Table 3 and are comparable to the best classification accuracies obtained during the inner cross-validation procedure suggesting little or no overfitting. A final mean classification accuracy of 79.2% was obtained for all subjects and all sets of button presses.

5. Discussion

5.1. MEG sampling rate

This study found that, surprisingly, MEG signals can be resampled at 45 Hz and still retain a significant amount of the information that allows them to be classified using GC features. This suggests—at least for the dataset used in this study—that either AR models are best suited for extracting GC relationships from MEG data sampled at low frequencies or that MEG data primarily contains characteristic relationships at these low frequencies during the button pressing experiment under investigation. Note that MEG signals that are resampled at 45 Hz were first low-pass filtered with a cutoff frequency of 18 Hz such that signals containing this particular part of the frequency spectrum are the ones that provide the information being used to successfully classify trials. However, adapting this procedure for one of the many frequency domain representations of GC (Baccalá and Sameshima, 2001; Kamiński and Blinowska, 1991; Sameshima and Baccalá, 1999) may be better suited for investigating these types of relationships. Additionally, it is also possible that other datasets may reveal that other sampling rates result in the best classification and this probably depends on the experiment being conducted, underlying neurophysiology, and other factors that may not hold for all studies or subjects.

Please cite this article in press as: Kostelecki W. et al. Single trial classification of magnetoencephalographic recordings using Granger causality. J Neurosci Methods (2011), doi:10.1016/j.jneumeth.2011.04.032

5.2. AR model order

Previous methods for determining appropriate AR model orders often involved the use of AIC or BIC (Ding et al., 2006). These methods rely on certain assumptions about the behaviour of the time series and distribution of the error terms but these assumptions cannot always be satisfied in bivariate settings so the exact form prediction errors take is unknown and cannot be used for significance testing. Instead, the procedure proposed in this study bypasses the requirements on the prediction errors by evaluating significance based on classification accuracy. If a higher classification accuracy is obtained with a different AR model order, that AR model order is more suitable for describing the GC relationships in the data, presumably because that AR model order spans the time scales at which important physiological processes that distinguish *forced* from *free* button presses are occurring. AR model orders that resulted in high classification accuracies ranged from 4 to 9 but this depended considerably on the sampling rate of the MEG signal under investigation (Fig. 4). Although not determined by a strict criterion, a low model order of $R = 4$ for MEG signals sampled at approximately 45 Hz was used for the final classification model evaluation to minimize the computational requirement for AR model estimation.

5.3. Principal component analysis for artifact deletion

For the removal of artifacts, it was surprising that classification rates improved up until approximately 13 principal components were deleted and slowly dropped off with the deletion of subsequent principal components (Fig. 5). It is unlikely that the first 13 principal components are entirely composed of contaminating signals and the trade off in this procedure is evident. The results suggest a balance in artifact removal between deletion of high amplitude contaminating artifacts which are present in the first few principal components and the slow degradation of the neural components of a signal by the further deletion of subsequent principal components. Although the exact artifact is likely not removed and all neural sources are not retained, this procedure does provide a basic method for enhancing the MEG signal for further analysis using GC concepts. Prior to this study, however, there was no clear method for testing alternative methods for signal preprocessing and evaluating whether those methods provide a practical improvement for the analysis.

In this study, a single setting for N_{PC} did not maximize the classification accuracy for all subjects simultaneously and the exact number of discardable principal components can conceivably be affected by many factors. As a result, a single setting was chosen to make a basic generalization about the proposed method for artifact deletion and its overall relationship to the analysis.

5.4. Classification with AR and GC

Previous studies have successfully classified neural signals using AR model coefficients (Anderson et al., 1998) suggesting that this has been a method of interest in the past. However, the method that was previously provided does not translate well to large MEG datasets of short time series due to the number of available recording sensors and the unreliable estimation of AR models at short time series lengths. Instead, this study offers an alternative feature for use in classification which is based on GC and provides a very compact metric summarizing the effectiveness with which different AR models explain the observed data of a short time series.

In this study, classification rates were not as high as what can typically be expected for neural signals in this type of setting. A previous analysis of the same dataset with the goal of obtaining the best classification rates obtained a mean accuracy of 82% (Dominguez et al., 2011) suggesting that superior methods exists if the goal of the study is to obtain a high classification accuracy but even these do not perform much better. Among possible reasons for the relatively low classification rates are that signals were analyzed in the original sensor space and are of a very short length. Improvements may be obtained by augmenting the analysis with source localization or blind source separation to access direct relationships between the neural sources underlying the observed MEG signals. Additionally, the behavioural experiment used in this study is short in duration and its underlying neural basis is expected to be transient which may have made classification more difficult. Consequently, better classification can be expected for datasets consisting of longer time series describing neural processes that are more stationary in time. Despite these short-comings, the study succeeded in its primary purpose of using AR process-like relationships to classify MEG data and may potentially be used to develop a method with which these types of dependencies between signals could be used to better understand neural mechanisms from MEG data.

6. Conclusion

The approach developed in this study achieved a single trial classification accuracy of 79.2% using features derived solely from the concept of GC. It offers a useful method for determining whether distinct directed relationships exist between multiple MEG recordings during the execution of a button pressing task by testing whether information about these relationships can be used to accurately classify individual MEG trials. Additionally, it provides a means for evaluating the efficacy of various adaptations to GC analysis. As a result, the approach provides a unique tool for studying time-dependent relationships in MEG signals and how MEG signals relate to AR models and GC.

Acknowledgments

This study was funded by the Bial foundation and the Natural Sciences and Engineering Research Council (awarded to WK). Funding agencies played no role in developing the study design, collecting, analyzing and interpreting the data, writing the report, and deciding to submit the paper for publication.

References

Anderson C, Stolz E, Shamsunder S. Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks. IEEE Trans Biomed Eng 1998;45(3):277–86.
Baccalá L, Sameshima K. Partial directed coherence: a new concept in neural structure determination. Biol Cybern 2001;84(6):463–74.
Barrett A, Barnett L, Seth A. Multivariate Granger causality and generalized variance. Phys Rev E 2010;81(4):041907.
Bressler S, Richter C, Chen Y, Ding M. Cortical functional network organization from autoregressive modeling of local field potential oscillations. Stat Med 2007;26(21):3875–85.
Bressler S, Seth A. Wiener–Granger causality: a well established methodology. NeuroImage 2010., doi:10.1016/j.neuroimage.2010.02.059.
Bressler S, Tang W, Sylvester C, Shulman G, Corbetta M. Topdown control of human visual cortex by frontal and parietal cortex in anticipatory visual spatial attention. J Neurosci 2008;28(40):10056–61.
Brovelli A, Ding M, Ledberg A, Chen Y, Nakamura R, Bressler S. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. Proc Nat Acad Sci U S A 2004;101(26):9849–54.
Ding M, Bressler S, Yang W, Liang H. Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. Biol Cybern 2000;83(1):35–45.
Ding M, Chen Y, Bressler S. Granger causality: basic theory and application to neuroscience. In: Schelter B, Winterhalder M, Timmer J, editors. Handbook of time series analysis: recent theoretical developments and applications. Berlin: Wiley-VCH; 2006. p. 437–60.
Dominguez L, Kostelecki W, Wennberg R, Pérez Velázquez J. Distinct dynamical patterns that distinguish willed and forced actions. Cogn Neurodyn 2011;5(1):67–76.

Please cite this article in press as: Kostelecki W, et al. Single trial classification of magnetoencephalographic recordings using Granger causality. J Neurosci Methods (2011), doi:10.1016/j.jneumeth.2011.04.032

Florin E, Gross J, Pfeifer J, Fink G, Timmerman L. The effect of filtering on Granger causality based multivariate causality measures. NeuroImage 2010;50(2):577–88.
Geweke J. Measures of conditional linear dependence and feedback between time series. J Am Stat Assoc 1984;79(388):907–15.
Goebel R, Roebroeck A, Kim D, Formisano E. Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. Magn Reson Image 2003;21(10):1251–61.
Gow D, Segawa J, Ahlfors S, Lin F. Lexical influences on speech perception: a Granger causality analysis of MEG and EEG source estimates. NeuroImage 2008;43(3):614–23.
Guo S, Seth A, Kendrick K, Zhou C, Feng J. Partial Granger causality—eliminating exogenous inputs and latent variables. J Neurosci Meth 2008;172(1):79–93.
Jung T, Makeig S, Humphries C, Lee T, Mckeown M, Iragui V, Sejnowski T. Removing electroencephalographic artifacts by blind source separation. Psychophysiology 2000;37(2):163–78.
Kamiński M, Blinowska K. A new method of the description of the information flow in the brain structures. Biol Cybern 1991;65(3):203–10.
Mitchell T, Hutchinson R, Niculescu R, Pereira F, Wang X, Just M, Newman S. Learning to decode cognitive states from brain images. Mach Learn 2004;57(1–2):145–75.
Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 2009;45(1):S199–209.
Sameshima K, Baccalá L. Using partial directed coherence to describe neuronal ensemble interactions. J Neurosci Meth 1999;94(1):93–103.
Schelter B, Winterhalder M, Eichler M, Peifer M, Hellwig B, Guschlbauer B, Lücking C, Dahlhaus R, Timmer J. Testing for directed influences among neural signals using partial directed coherence. J Neurosci Meth 2006a;152(1–2):210–9.
Schelter B, Winterhalder M, Hellwig B, Guschlbauer B, Lücking C, Timmer J. Direct or indirect? Graphical models for neural oscillators. J Physiol Paris 2006b;99(1):37–46.
Schlögl A. A comparison of multivariate autoregressive estimators. Signal Process 2006;86(9):2426–9.
Soon C, Brass M, Heinze H, Haynes J. Unconscious determinants of free decisions in the human brain. Nat Neurosci 2008;11(5):543–5.
Supp G, Schlögl A, Trujillo-Barreto N, Müller M, Gruber T. Directed cortical information flow during human object recognition: analyzing induced EEG gamma-band responses in brain's source space. PLoS One 2007;2(8):e684.
Wang X, Chen Y, Bressler S, Ding M. Granger causality between multiple interdependent neurobiological time series: blockwise versus pairwise methods. Int J Neural Syst 2007;17(2):71–8.
Zhou Z, Chen Y, Ding M, Wright P, Lu Z, Liu Y. Analyzing brain networks with PCA and conditional Granger causality. Hum Brain Map 2009;30(7):2197–206.

Please cite this article in press as: Kostelecki W, et al. Single trial classification of magnetoencephalographic recordings using Granger causality. J Neurosci Methods (2011), doi:10.1016/j.jneumeth.2011.04.032