Dangers and uses of cross-correlation in analyzing time series in perception, performance, movement, and neuroscience: The importance of constructing transfer function autoregressive models

Roger T. Dean[1] · William T. M. Dunsmuir[2]

Published online: 23 June 2015

Abstract Many articles on perception, performance, psychophysiology, and neuroscience seek to relate pairs of time series through assessments of their cross-correlations. Most such series are individually autocorrelated: they do not comprise independent values. Given this situation, an unfounded reliance is often placed on cross-correlation as an indicator of relationships (e.g., referent vs. response, leading vs. following). Such cross-correlations can indicate spurious relationships, because of autocorrelation. Given these dangers, we here simulated how and why such spurious conclusions can arise, to provide an approach to resolving them. We show that when multiple pairs of series are aggregated in several different ways for a cross-correlation analysis, problems remain. Finally, even a genuine cross-correlation function does not answer key motivating questions, such as whether there are likely causal relationships between the series. Thus, we illustrate how to obtain a transfer function describing such relationships, informed by any genuine cross-correlations. We illustrate the confounds and the meaningful transfer functions by two concrete examples, one each in perception and performance, together with key elements of the R software code needed. The approach involves autocorrelation functions, the establishment of stationarity, prewhitening, the determination of cross-correlation functions, the assessment of Granger causality, and autoregressive model development. Autocorrelation also limits the interpretability of other measures of possible relationships between pairs of time series, such as mutual information. We emphasize that further complexity may be required as the appropriate analysis is pursued fully, and that causal intervention experiments will likely also be needed.

Keywords Autocorrelation · Autoregression · Cross-correlation · Continuous responses · Time series analysis · Granger causality · Prewhitening · Transfer function modeling

The study of relationships between simultaneous time series, particularly those involving continuous human perceptions and performance, has been ongoing in many fields of psychology for several decades (e.g., Brunsdon & Skinner, 1987; Gregson, 1983; Pressing, 1999). To illustrate the ubiquity of time series data in these fields, consider the categorization provided by Pressing (1999) in his synthesis of "the referential dynamics of cognition and action". Pressing enunciates "referential behavior theory," which is a "general dynamical approach to psychological . . . systems that operate through a control or referencing process" (p. 714)—specifically, processes that operate continuously. The analytical approach to such continuous processes is most commonly discrete, in that the data are samples of the process spaced regularly in time, and this is what concerns us here. The resulting "discrete control equation" is normally an example of a vector (multivariate) autoregression time series analysis model of the paired (or several) continuous response/performance time series, which may include exogenous predictors such as the series representing the control process(es). When a time series is autocorrelated, this means that the current value of

Electronic supplementary material The online version of this article (doi:10.3758/s13428-015-0611-2) contains supplementary material, which is available to authorized users.

✉ Roger T. Dean
roger.dean@uws.edu.au

[1] MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, New South Wales 2751, Australia

[2] School of Mathematics and Statistics, University of New South Wales, Sydney, New South Wales, Australia

the series parameter is dependent on preceding values, and can be predicted (at least in part) on the basis of knowledge of those values.

Pressing distinguishes examples of such time series processes operating on different time scales: on a scale of up to about a second, one finds motor actions such as tracking, rhythmic tapping, balancing an object, and interception; on a scale between a second and about a minute occur oral presentation, musical improvisation, knowledge of results, and knowledge of performance in sport or decision-making; and at time scales of more than a minute up to decades, adaptation, learning, goal setting, and planning take place. In the development of expertise, this may include self-monitoring and self-regulation processes. Time series analysis is relevant to observations on this whole range of processes.

Time series derived from motor performance and coordination (Pressing, 1999), psychophysiological variables such as heart rate and skin conductance (Brunsdon & Skinner, 1987; Dean & Bailes, 2015), continuous perceptual responses (Dean & Bailes, 2010; Schubert, 2006), or electroencephalography (EEG; Piryatinska et al., 2009), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI) measures are almost always highly autocorrelated, so contravening the principles of independence upon which most common statistical tests rely. For a simple example of this, consider a computer mouse being moved to translate a pointer across a screen, perhaps in a visual search task. At every successive instant, the mouse position is necessarily close to, and correlated with, that of the previous instance; the case is similar for the motions of body parts during movement. Thus, any time series derived from such positions will almost inevitably be autocorrelated (Gregson, 1983). In the fields of human perception and performance, the same is generally true of experimental time series data.

For example, even in studies of tapping *apart* from an anisochronous audio pacer (i.e., tapping "off-pacer" rather than "on-pacer," and with the pacer showing irregular timings) we recently showed that amongst 34 participants, autocorrelations were almost universal (Launay, Dean, & Bailes, 2014). An autocorrelation function is a representation of the autocorrelations between values separated by any of a chosen range of lags. A partial autocorrelation function shows correlations across more than one lag, in each case with the influence of intervening lag autocorrelations "stripped out." We found that autoregressive models of order 1 (meaning that the autocorrelation at lag 1 was significant) were required to model the performance of almost everyone, and order-2 models were required for most; orders of up to 5 contributed (Launay et al., 2013). Related autoregressive time series are important in econometrics (Enders, 2004), ecology (Zuur, Ieno, & Smith, 2007), and many other fields.

More recently, the study of perceptual and performance interactions has intensified, alongside increasing interest in continuous real-time perceptual responses, as well as electrophysiological measures in EEG, MEG, and fMRI, such as interbrain coupling. In many cases, the experimental purpose is to investigate a particular version of Pressing's "referential behavior" in which one (or several) ongoing processes can control another (or several), sometimes with bidirectional influences. What is ultimately sought is usually a quantitative measure of the influences and effects at work, not simply a determination of their nature. Our purpose here is to point out that neglecting the consideration of autocorrelation, and relying instead on a simple cross-correlation analysis, may be inadequate and even dangerous in these circumstances, and then to summarize an approach providing a principled analysis.

Many of the publications in these areas have not attempted to assess, or have assessed incorrectly, the significance of cross-correlation coefficients. Consequently, without parsimonious modeling these articles may attribute relationships that are not significant. This may also be one component of the reasons for the unreliability of results in fields such as neuroimaging (Button et al., 2013; Carp, 2012), which we will not consider directly here. Specifically, Carp indicated that amongst a large set of fMRI studies analyzed, only 14.6 % "reported how or whether temporal autocorrelations were modeled, and 0.84 % described whether regressors were orthogonalized" (p. 294; see the comments on prewhitening below for more on orthogonality before cross-correlation).

We ourselves are concerned with the relationships between simultaneous acoustic, perceptual, and performative movement streams in music (Bailes & Dean, 2012; Dean & Bailes, 2010, 2011; Dean, Bailes, & Dunsmuir, 2014a, b; Launay et al., 2013) and face precisely these problems of the use and interpretation of cross-correlations. Below we show one category of solution to these problems, in the hope that this will aid participants in the fields of human perception and performance to grapple effectively with them. To explain the problem, we provide a set of simulation studies that produced series that might represent almost any perception or performance data series, and we provide analyses of data from both a continuous-perception experimental system and a movement performance system.

What is the problem with cross-correlating simultaneous autocorrelated time series?

The essential point here is that commonly, meaningless correlations exist between independent pairs of time series that are themselves autocorrelated. We seek first to explain and illustrate this and to reveal its extent by simulation. If two stationary series, $\{x_t\}$ and $\{y_t\}$, are independent of each other (i.e., if values in one series at any time do not provide information about the values in the other series at any time), then it is still possible that the cross-correlation between the series can appear to be significantly nonzero when judged against standard criteria for significance for independent pairs of measurements. Consider, for example, the situation in which both series are strongly positively autocorrelated. Such series have relatively smooth sample paths in which low-frequency behavior dominates. Now take two such series that are independent of each other and shift the sample path of one relative to the other and then correlate the aligned pairs. In this case, it is quite likely that the wave-type behavior in each series can approximately align, giving the impression (visually and numerically) that the series are correlated with each other.

We use a simple simulation to illustrate how nonsense correlations can arise between two series that are completely independent of each other when each series has strong positive serial dependence. Figure 1a shows the sample paths of segments 200 observations in length from two autoregressive series with strong positive autocorrelation that were generated to be completely independent. As can be seen, there are some moderate periods of time (e.g., over times 110 to 145) over which runs above the mean level zero are coincident. Such runs will contribute to positive correlation between the series instantaneously. In fact, for this pair of series, the correlation between x_t and y_t is small, negative, and not significant according to the usual test for a correlation—this can be seen in the cross-correlation function of Fig. 2 at lag 0, and below we explain in detail how this was constructed and why the conventional 95 % significance limits shown are incorrect. Figure 1b shows the same pair of series with y_t shifted in time by 20 time points to align it with x_t, in order to maximize the pairwise correlation between the aligned series. As is clear from Fig. 1b, there are substantially more periods of time when the "waves" in both series tend to track each other and during which both series tend to be simultaneously positive or simultaneously negative. In this situation, we would expect the correlation between the two series to be stronger than before the shift. The correlation between the series aligned in this way is .37, and this is judged to be highly significant according to the usual significance limits for independent pairs. This again can be seen in Fig. 2 by looking at the cross-correlation function at lag = –20. Figure 1 thus demonstrates how spurious correlation can arise by comparing values from one time series that has strong autocorrelation with values, possibly shifted in time, from another independent series.

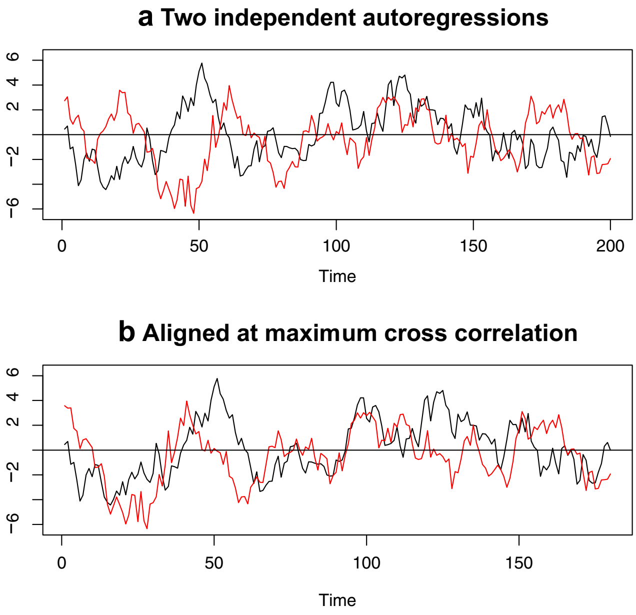


Fig. 1 Two autoregressive time series are generated (with a preset seed value, to make for reproducibility). The ar(1) coefficient (phi) is .9 for each, and there are 200 events. Panel **a** shows the series themselves, and panel **b** shows them after optimal alignment after determining the maximum cross-correlation lag, to emphasize their similarities

Some of the leading textbooks on time series analysis give good expositions of the mathematics of such "spurious correlations" between autocorrelated series (e.g., Cryer & Chan, 2008) and in-depth discussion of this originates with Yule, who referred to "nonsense-correlations" (Yule, 1926). As Yule wrote, "We must divest ourselves . . . from all our preconceptions based on sampling on fundamentally different conditions" (p. 12)—such as those of randomly distributed normal variables. The basic features of time series can be digested fully from relevant textbooks, some offering very detailed mathematical derivations (Hamilton, 1994). Amongst the relevant facts is the one that most concerns us here: that pairs of serially (that is, auto-) correlated series that are completely independent of each other can show apparent cross-correlations. Independent (non-time-series) data sets that are to be assessed by conventional parametric statistics need to comprise values that are i.i.d.—that is, *independent and identically distributed*. In contrast, as we mentioned, time series are rarely i.i.d., but rather, they are almost always autocorrelated.

Time series that are to be modeled or related to each other are normally molded first to conform within probabilistic limits to "weak stationarity." The weak stationarity (routinely simply termed "stationarity") of a time series is achieved when the mean and variance are constant and the autocorrelations between values a given number of time points (lags or leads) apart are also constant. In other words, stationarity does not require the removal of autocorrelation. Thus, a series with a trend is commonly made stationary by *differencing*: that is, by constructing a new series (which is one event shorter), comprising successive differences between adjacent values of the original series (see Sims, 1988, for discussion). Sometimes other transformations, such as taking the log or the removal of estimated temporal trends using regression-style procedures, are useful.

For a single series, the autocorrelation function (acf) comprises the correlation between the values in the series and values at a series of time lags in the same series. When an autocorrelation function of a single series is assessed, the

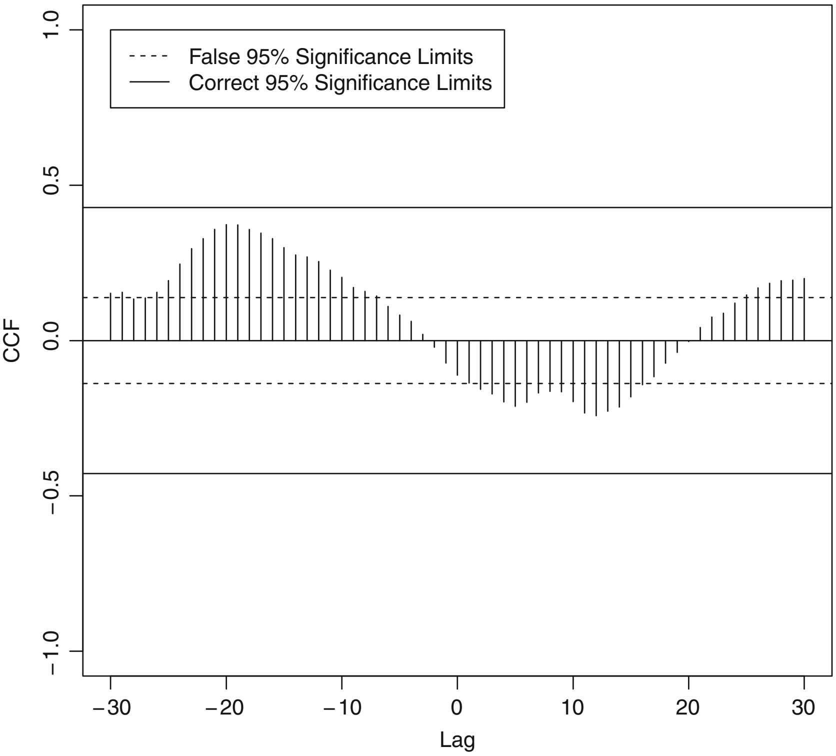


Fig. 2 Cross-correlation function (CCF) of the pair of series shown in Fig. 1

correlation coefficients are considered to be significantly different from zero at $p < .05$ when they breach a value of 1.96 standard deviations from zero. Under the assumption that the series is *not* autocorrelated, the standard deviation of an autocorrelation is approximated by $1/\sqrt{n}$, where n is the number of consecutive observations on the series. This is the approximate (large-sample) standard deviation for the usual Pearson-correlation-coefficient-based n independent observations on pairs of independent measurements, and this gives rise to the conventional autocorrelation significance limits of $\pm 1.96/\sqrt{n}$. The stationarity of both series of interest is required in order to assess cross-correlation, but in itself is not enough to avoid the risk of spurious cross-correlations. As we explain shortly, these significance limits are also used to assess the cross-correlations, although they are only correct if the two series being considered are independent *and* at least one of them is serially *un*correlated.

The autocorrelations of a single series (providing that it is stationary) are symmetrical: that is, the correlation between x_t and x_{t-l} is the same as that between x_t and x_{t+l}, where t is a time point and the time point l units of time later is $t+l$. The cross-correlations between two such series, on the other hand, are generally not symmetrical, and it is this feature that allows the determination of cross-correlation functions (CCFs: the complete set of cross-correlations across lags) to be informative about the potentially bidirectional relationships between the series—see below for further development of this point.

In order to further explain why spuriously significant cross-correlations can occur between two autocorrelated series that are independent of each other, it is necessary to define the sample cross-correlation function. To simplify the formulae while preserving the key aspects, we assume that the two series of interest, $\{x_t\}$ and $\{y_t\}$, have been adjusted to have a zero mean at all times t. On the basis of n consecutive observations of both series, the *sample* cross-covariance at lag l is

$$C_{YX}(l) = \frac{1}{n}\sum_{t=1}^{n-l} x_t y_{t+l}.$$

The *sample* cross-correlation simply normalizes this quantity by dividing by the product of the standard deviations of both series. If the cross-correlation is strong at some positive value of l, then changes in x_t are associated with changes in future values y_{t+l}, and the converse is the case if the cross-correlation is strong at a negative value of l.

When $\{x_t\}$ and $\{y_t\}$ are independent time series, that means that x_t and y_s are independent for any values of t and s; in particular, x_t and y_{t+l} are independent for any selected lag l and all values of t. Hence, on average, $x_t y_{t+l}$ is zero for all values of t, and therefore $C_{YX}(l)$ is, on average, also zero. In conclusion, the sample cross-correlation between two independent series will be centered, correctly, around zero for all lags l.

However, and this is the critical point, the variability of this estimate is not the same as would be obtained from paired data for which not only are the values in each pair independent, but all pairs are independent of each other. As we noted above, for independent repetitions of independent variables x_t and y_t, the standard deviation of a sample correlation coefficient is approximately $1/\sqrt{n}$, the classical value, and approximates the 5% significance limits $\pm L$, where $L = 1.96/\sqrt{n}$. Default graphs for cross-correlations—for example, in the widely used open-access software R—often automatically present a default criterion, usually this value or a closely similar one (see the second experimental example discussed below). This L is what we abbreviate as the *conventional cross-correlation limit* (or "ccsl," for short). But, when the pairs (x_t, y_t) are not independent temporally, this standard deviation can be higher or lower. Figure 2 shows the default limits as dashed horizontal lines, and clearly, many of the cross-correlations exceed these limits, suggesting significant associations between the two series that, as we know from how the simulation was performed, are indeed not associated at any time lag. Also clear from Fig. 2 is the tendency for the CCF to be smooth; that is, neighboring values of the CCF are positively correlated. This phenomenon has the potential to reinforce a researcher's impression that there is a relationship between the two series. Both phenomena (exceeding the ccsl and a smooth pattern in the CCF) can be mathematically explained and are completely expected for two highly autocorrelated series. The *correct* significance limits are also shown in Fig. 2, as solid horizontal lines. Relative to these, none of the CCF values are significantly different from zero. The smooth pattern remains, with all the dangers to false interpretation of significance that this could lead to.

What are these *correct* significance limits? We illustrate this in the case in which the lag 1 autoregressive autocorrelation coefficients of the two series are respectively a and b. For Fig. 1, these both took the quite large value of .9. Then the correct criterion should be $L \times F$, where the factor modifying the ccsl L is $F = \sqrt{(1+ab)/(1-ab)}$ (Cryer & Chan, 2008). Although a and b have the same sign, this value is greater than the ccsl criterion L. If, and *only if*, at least one of the two autocorrelations is zero, then $F = 1$, and this expression reduces to L, the ccsl value defined above. But when two series have positive lag 1 autoregressive autocorrelations (i.e., $a > 0$ and $b > 0$), the factor F—and hence, the criterion for significance $L \times F$—should instead become much larger. Likewise (and perhaps more challenging to intuition), when the lag autocorrelation is negative in both series ($a < 0$ and $b < 0$) then, again, $F > 1$, giving the same inflation-of-variability effect. Also, when a and b are of opposite signs, so that $a \times b < 0$, the correct significance limits are *deflated* by $F < 1$ relative to the ccsl. When the autocorrelations are more complex, the standard deviation of cross-correlations becomes again correspondingly more complex (being the sum over all lags past and future of the product of the autocorrelations for one of the series with those at the same lag for the other series). See Cryer and Chan (2008) for formulae and discussion, and see our supplemental materials for the simplest, nonparametric approach to estimating it. For a stationary pair of series, determining such a significance limit is essential to allowing an assessment of which cross-correlations are likely to be meaningful.

This formula is illustrated below for simulated data. A frequentist approach is useful for the ccsl and for interpreting the CCF, but it may be replaced in subsequent analyses by greater emphasis on effect size point estimates and confidence limits, as has been advocated since 2010 by the American Psychological Association journals. In the interpretation of CCFs, the approaches lead to the same conclusions. Note that the ccsl (L) and the correct limits $L \times F$ are not technically confidence limits, although they are often referred to as such.

The general principle also applies to pairs constructed when one series is shifted in time relative to the other, as occurs in the formula above for the sample cross-correlation. Furthermore, the pairs (x_t, y_{t+l}) are correlated through time t, and hence, when they are aggregated to form the sample covariance, the variability of $C_{XY}(l)$ is impacted. Furthermore, the covariability between $C_{XY}(l)$ at one lag and $C_{XY}(k)$ at another will not be zero. So it is quite possible to have sample cross-correlation functions (considered as a function of lag l) that show "waves" of neighboring values being higher than average, if autocorrelation in both series is strongly positive, or alternatively tending to oscillate around zero, if one series is positively autocorrelated and the other negatively so. The key point is that the variability of the individual lag cross-correlations and the covariability between these at different lags do not conform to the standard correlation paradigm for independent pairs of independent data when there is autocorrelation in both series.

Although these formulae show that even large cross-correlations may be nonsignificant, and hence misleading, they also show that converting at least one of the series into a form that does not have autocorrelation (e.g., $a = 0$ or $b = 0$, or both, in the formula above) will reduce the genuine significance cutoff so that it matches the routine ccsl. This is achieved by the process known as "prewhitening," and the result of this is a CCF that may now be realistic and informative.

Avoiding the risk of spurious cross-correlation by prewhitening

The key function of prewhitening is to remove autocorrelation from at least one of the pair of series under study. Prewhitening involves decorrelating the putative predictor (referent or control) series, and then applying the same filter (autoregressive model) required to achieve this to the response series. This is highly unlikely to decorrelate the response series. Technically, it does not matter which of the two series is used for the first step, since the objective is simply to ensure that at least one of the prewhitened series is free of autocorrelation. Consequently, prewhitening may or may not reduce or remove cross-correlation between the two, and the remaining cross-correlations would be indicative of predictive relationships.

Why is prewhitening not the end of the analysis?

So prewhitening allows a proper assessment of the CCF relating two (previously autocorrelated) time series. Is that enough, and furthermore, does that suffice to define the control function or to answer the experimental question behind the study? It is actually quite difficult to envisage a circumstance in which simply knowing the CCF would be the real underlying purpose of the analysis. It is much more likely that the wish is to understand quantitative mechanisms, which for example requires a means of integrating the impacts of multiple cross-correlations between lags, and leads of two series *x* and *y*. Given that the cross-correlations are not necessarily symmetric, the question of the directionality of any possible influence is not necessarily resolved by the CCF per se.

One influential case (Repp, 1999, which has >70 citations; see also Repp, 2002) in which it was envisaged that the CCF in itself would be sufficient was the attempted separation of participants in rhythmic co-tapping experiments into “predictors” and “trackers,” respectively potentially providing and responding to the control function. These articles used cross-correlations within short event windows, but without assessing the significance of those cross-correlations. Furthermore, they considered only differences between lag –1 and lag +1 (lead) cross-correlation magnitudes—in a sense disallowing the influences of any other cross-correlations, which is probably dangerous for the purpose. As several authors have found, higher-order auto- and cross-correlations occur routinely in tapping performance (Launay et al., 2013; Pressing, 1999). Thus, the categorizations of participants based on Repp’s original approach may be faulty, and some of the implications of studies flowing from these may bear further investigation. For example, the question of predictor/tracker distinctions versus “mutual” relationships has been investigated further by Konvalinka, Vuust, Roepstorff, and Frith (2010), and this work implies that the processes are rather more reciprocal.

A later article considering possible mutual participant influences included a diverse set of analyses of the relationships between the movements of pairs of participants, when there was no explicit referent to which the pairs were relating, but rather they were conducting a cooperative team-task (Strang, Funke, Russell, Dukes, & Middendorf, 2014). The authors envisaged that both uni- and bidirectional influences might pertain. The CCF analyses were a small part of the work, which used a variety of additional techniques. But the CCF analyses are interesting, in that the concern of the study was synchronization and entrainment: both concepts implying that pairs of participants make some movements at the same frequency, but not necessarily that the movements are simultaneous. The CCF analyses in this particular article did not consider the possibility of lag 0 cross-correlation (simultaneity), nor discriminate between various degrees (phases) of synchrony. The lag 0 cross-correlation (Strang et al., 2014) might also have been informative, and have pointed to other aspects of the interrelations being studied. More generally, we suggest that the information in all of the multiple significant CCF lags should be considered.

In so doing, the implications even of a prewhitened CCF should be subjected to stringent tests; and as we will see below, prewhitening is not always entirely effective or informative. Given the arguments above, in the remainder of the article we assume that the analytical purpose at hand is mechanistic and goes beyond solely determining a reliable CCF, toward interpreting its integrated impact.

Testing for potentially causal correlations between variables: Granger causality and the transfer function

Granger causality assesses whether the relationship between the two series (now again in their original stationarized forms, not in their prewhitened forms) is likely to contain a causal element, considered from a statistical perspective. To be more precise, a variable *x* can be said to be Granger-causal of variable *y* if preceding values of *x* help in the prediction of the current value of *y*, *given all other relevant information*. The italicized condition is important, because it is this that “isolates” the relationship between *x* and *y*, even when there are other predictors in the system, both exogenous (not open to influence by others in the system) and endogenous (open to such influence). Granger causality involves estimating simultaneous autoregressive models of *x* and *y*, each involving both variables. Then the Granger causality of *x* on *y* is established if the coefficients on the lags of *x* in the equation for *y* are jointly significant. A complementary joint significance test can determine whether there is also causality of *y* on *x*. As with all statistical tests, some assumptions are required (and alternative versions have often been developed to overcome them). Notably, it is usually required that the error term be normally distributed and that *x* show zero covariance with the error; it is also accepted that in some circumstances both *x* and *y* may be driven by a common (but unknown, “latent”) third process. The test can be adapted to multivariate systems (using vector autoregression), as well as to the bivariate type mainly discussed here.

An assessment of Granger causality is now becoming quite common in EEG studies, for investigating the relations between different parts of the brain or between different cooperating brains (Müller, Sänger, & Lindenberger, 2013), but sometimes the issue of stationarity is neglected in such work, and autocorrelation is often neglected as well. Given Granger causality, then, a first conclusion about likely causality can be made, but this does not provide the quantitative integration of impacts that may really be sought. How much does x influence y? What are the contributions of the various lags? Given that this information (the so-called *transfer function*) is required, parsimonious modeling should follow (see below). The transfer function is the part of the model of the response that expresses the relationship between the input predictor series and the output series, alongside the autocorrelations, which remain a separate part of the model.

A example of a linear transfer function model is

$$y_t = \sum_{k=1}^{K} d_k x_{t-k} + u_t,$$

where u_t is stationary (but not necessarily a process of independent random errors), K is a maximum lag at which the x_{t-k} have an impact on y_t, and the d_k are the transfer function coefficients, some of which could be zero. In this example, the $\{x_t\}$ series "leads" y_t, and changes in the former at time t influence the latter at times $t+1, \ldots, t+K$. In general, before an assessment of significance, the limits in the summation above could also extend into the future (using negative values of k). The technique of prewhitening the predictor series $\{x_t\}$ and applying the prewhitening filter to y_t as well, to properly assess the significance of any cross-correlations, can also be used to determine the transfer function coefficients: the CCF between the prewhitened input series and the filtered output series is proportional to the set of transfer function coefficients. In the formula above, the CCF is zero for lags other than $k = 1, \ldots, K$. In general, the CCF of the prewhitened series indicates that transfer function coefficients are statistically significant as the basis for detailed modeling—see below for an example.

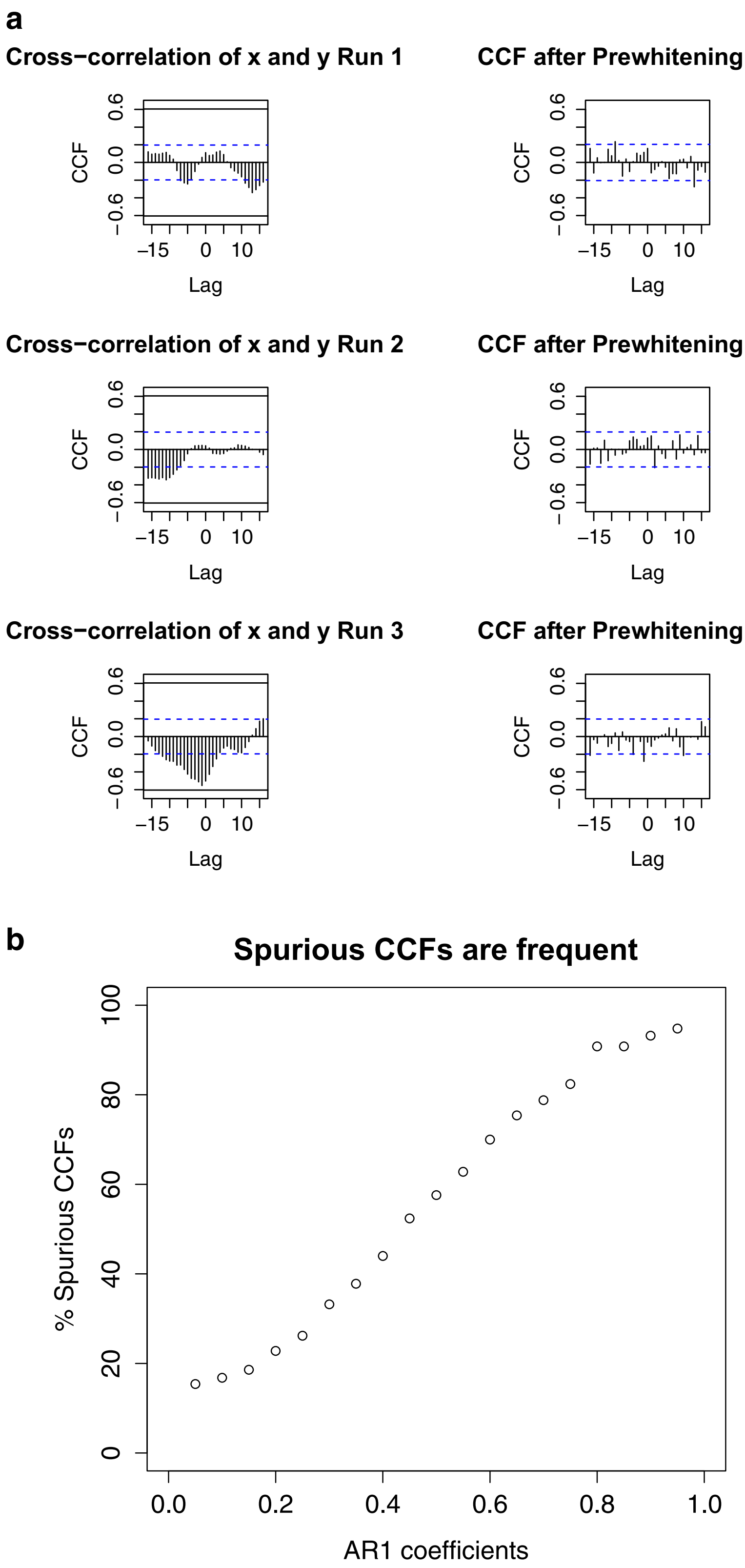
We next illustrate the common occurrence of misleading cross-correlations, using simulation, and show how they can be evaluated and dismissed by using reasonable estimates of significance limits, prewhitening, and Granger causality. We follow through with two illustrations of an appropriate method for developing an analysis of a pair of series that overcomes these problems when the series are significantly interrelated. These illustrations show that CCFs can guide the necessary subsequent modeling and provide a valuable step in data exploration.

Illustration by simulation of spurious cross-correlations, their identification, their high prevalence, and their removal

Figure 3a shows the CCFs for three successive simulations of pairs of ar1 (first-order) autoregressive time series of 100 points, with identical purposefully large positive autoregressive coefficients (see the caption). By their mode of simulation, the members of each pair of series have no mechanistic relation with each other. These series can be considered to be analogous to those obtained in most of the types of referential behavior summarized in the introduction to this article, in that the time separation between the successive data points might be of any magnitude, and one series might be the candidate control process, the other the measured response or performance. The supplemental materials summarize some key commands and packages in R that can be used to achieve the simple analyses we present here.

Figure 3a, left-hand column, shows that many of the cross-correlation coefficients for the simulated series breach the ccsl (the dotted lines), sometimes centered around lag 0, sometimes not. By simulating large numbers of series with a given ar1 coefficient, Fig. 3b shows the high frequency of potentially spurious CCFs for ar1 coefficients between .95 and .05: for example, if both ar1 coefficients are ≥.8, around 90 % of CCFs are spurious; even with coefficients of .05, the frequency is still around 20 %. As is detailed in the supplemental R code, here we take a spurious CCF as one in which three or more CCF coefficients breach the ccsl (as only ≤1.65 should do so by chance). A smaller set of parametric determinations of this prevalence (but with reference only to individual coefficients) is shown in Cryer the Chan (2008), with similar implications. Unfortunately, most assessments of the significance of cross-correlation coefficients are indeed made on the basis of the ccsl, if the assessment is undertaken at all. On the other hand, the solid horizontal lines in the left panels of Fig. 3a show the theoretically correct much larger 95 % significance limit described above, indicating that none of the apparently significant coefficients in those cases are real. These series are unconnected with each other, given their mode of simulation, and the Granger causality test confirms that none show causality at $p < .05$. When the simulated series are 1,000 points long, similar problems remain frequent (not shown).

What is needed in general in such analyses is prewhitening, so that any lags or leads (negative lags) of x that are real predictors of y can be detected. As we described above, prewhitening orthogonalizes the various lags of x, meaning the lags are no longer dependent on each other. Following this, exactly the same autoregressive lag structure is used to filter the second series. Now, as one of the resultant series (the residuals of the first) has no autoregressive components, the ccsl is the appropriate limit for assessing likely significant coefficients of cross-correlation between



the pair of residual series. This is shown in Fig. 3a (right column) for the same three pairs of series as in the left column. It can be seen that at most four coefficients narrowly breach the criterion, and even these are most likely false positives, since with a $p < .05$ cutoff, one would expect about five such positives from the 99 values shown.

◀ **Fig. 3** **a** Three immediately successive simulations of a pair of ar(1) autocorrelated series comprising 100 events were generated, and their cross-correlation functions are shown (left side). The dotted lines show the conventional cross-correlation significance limit at 95 % (what we term the ccsl), whereas the solid horizontal lines show a realistic significance limit (see the text for discussion). A Granger causality assessment confirms that none of these cross-correlations are significant, in the sense of "predictive." **b** For each indicated ar1 coefficient, 500 pairs of series, each of 100 events and each with the specified coefficient, were generated. Since the cross-correlation function (CCF) has 33 members (lag 16 to lead 16) at p = .05, one might expect by chance 1.7 coefficients to lie outside the cross-correlation significance limit. Thus, we define a spurious CCF as one in which there are three or more such values. The y-axis shows the percentage of CCFs judged spurious in this way for each condition.

Prewhitening thus may reveal any meaningful cross-correlations. Its fundamental purpose is not just their evaluation, but rather data exploration: to display the likely leading or lagging relationships, if any, between the pair of series. The guiding information from both the raw cross-correlations, considered against a realistic significance limit, and from the prewhitened cross-correlations considered against the ccsl, are used in the analysis of the transfer function model relating the two series, as we describe later. In the case of the simulated series, the analysis would proceed no further, since there are no interesting or Granger-causal relationships—as was of course known from their mode of construction.

Dangers in windowed (rolling) cross-correlation between time series

It is apparent that the dangers described above apply to any individual CCF determined on subsets (windows or "chunks") of a pair of time series. Quite often, relatively short windows are used for such analyses, and the window is "rolled" forward by one time unit for each successive CCF determination. One source of this approach (Boker et al., 2002) has >73 citations, and advocates have windowed cross-correlations using small windows (e.g., ten events for a dance analysis). The argument presented is that local short portions of series may be stationary. However, the article gives no consideration to the influence of autocorrelation and does not assess the significance of the reported cross-correlation coefficients. Unfortunately, as we discussed above, stationarity does not require nor usually achieve a removal of autocorrelation, and so does not obviate the problems under discussion. Indeed, this is exactly the issue elaborated by Yule in relation to windowed cross-correlations (Yule, 1926).

There are many examples of this dangerous approach. In substantial studies of tapping by interacting pairs of participants, investigating possible leader/follower relationships, windows of six taps were used (Konvalinka et al., 2010), finding maximal cross-correlations of about .8. Windowed cross-correlation has also been used in dance studies to discuss leading/following relationships (e.g., Himberg & Thompson, 2011). Their cross-correlations are often quite low, and significances are not reported; it is difficult to find the CCF window size from the articles. For a window of size 6, even assuming no autocorrelation in at least one of the series, the ccsl is very large: .80. Given positive autocorrelation of both series, the relevant significance limit values might be larger still for the small windows.

Yule (1926) gave a striking illustration of this problem, highly relevant to movement studies, in considering cross-correlations between short windows of pairs of oscillatory series (which may be periodic or otherwise). He showed with pairs of sine curves in various phase relationships that the frequency distribution of cross-correlation coefficients between simultaneous elements "*always remains U-shaped, and values of the correlation as far as possible removed from the true value (zero) always remain the most frequent*" (p. 10, italics in the original). The salience of this statement for time series is emphasized by the facts that they can be formulated and/or analyzed in the frequency domain (like sine functions), just as much as in the time domain that we are considering, and that oscillator models of movement production are relevant and current: see the in-depth discussion in Pressing (1999).

Returning to the temporal domain, it is also important to note that the series of CCFs of successive sliding windows are in any case autocorrelated, in part simply because they contain overlapping data. A more subtle component of this autocorrelation between successive CCFs can be grasped by again considering the pairs of time series simulated above: because the autoregressive pattern of each individual time series in a pair remains constant, there will be commonalities also between CCFs derived from them, even using windows that do not overlap. In turn, this means that any averaging or other aggregation of the measured windowed CCFs, with values that may not be significant individually, cannot safely rely on i.i.d. analyses, because the successive CCF values are not independent.

Thus, the information in any aggregation of the successive, potentially spurious, windowed CCFs is difficult to interpret, and certainly cannot be tested by conventional difference assessments. What is required, rather, is the use of realistic significance limits and prewhitening, followed by the development of a time-varying model assessed by stringent information criteria, if there are any interesting trends in the windowed cross-correlations.

We can integrate our analysis of cross-correlations between series, and of windowed cross-correlation, with the following recommendation: authors using these techniques should provide correct standard errors for their reported estimates of CCFs, and through these, a conclusion as to their statistical significance. Do they breach the realistic significance limit, and not just the ccsl? This can then lead to Granger causality and transfer function assessment.

Individual and group behaviors compared: Analyses based on aggregations of time series across groups of participants

The demonstration above that spurious CCFs are common with pairs of autocorrelated time series may elicit the initial response in a reader that it is not common to need to understand the relation between the members of single pairs of series; rather, multiple repetitions of such pairs will be aggregated before the analyses, and this will remove the spurious features. However, whether this is actually the case does not seem to have been demonstrated in simulation studies, or any others that we have found. We assess these issues here directly through further simulations.

First, it is worth pointing out that in many experimental situations in perception and performance studies, it is indeed essential to be able to analyze individual pairs of series. For example, it may be that the modes of behavior of individuals vary substantially one from another, and that it is necessary to be able to distinguish categories of behavior. This discrimination might imply either a biphasic distribution (one distribution of parameters for the followers, another for the leaders) or a single, unimodal distribution encompassing both. Given the occurrence of a distribution of behaviors in relation to some psychological issue, it is likely that analyses will need to include random effects for individuals, as in mixed-effects analyses (Baayen, 2008; Baayen, Davidson, & Bates, 2008), and/or researchers will have to separately analyze the behaviors of the different groups of participants.

For many other reasons, it can sometimes be both informative and necessary to be able to undertake analyses on individual event series. But equally, it is important to understand whether, even given a unimodal distribution of behavior in relation to some particular process, standard empirical approaches of constructing averages across multiple participants do permit secure cross-correlation and transfer function analyses. So we next provide some further analyses through simulation of an experimental system in which it is known that the distribution of behavior is unimodal, and for which we seek estimates of population parameters and transfer functions from the behavior of groups of participants.

A common group sample size in psychological experiments is 30 participants. Thus, we provide code in the supplemental materials for the simulation of 30 pairs of response time series, each series generated exactly as described for those discussed above, all with large positive order-1 autocorrelation. As before, our purpose is to analyze the predictive relationships between the pairs of series—for example, one series considered as an input (referent/independent variable) series, the other as an output (response/dependent variable) series. One common and reasonable approach to analyzing such data sets is to average the 30 CCFs generated by analyzing each pair of time series, and to determine the empirical significances of each resulting coefficient on the basis of its mean and standard error. The plausible assumption is that this will largely remove the spurious cross-correlations just demonstrated for the individual series pairs, even though the values being analyzed are mostly individually nonsignificant. Figure 4a shows the results of such a simulation: as expected, no cross-correlations remain, and the coefficients are modest (smaller than those obtained on some coefficients for each individual series pair), confirming the importance of the issues discussed above. The parallel analysis in which each individual series pair is prewhitened before the averaging shows a similar output, with no significant cross-correlations (Fig. 4b).

It might seem that the averaging process is thereby validated, even without prewhitening. But what of the pattern of the CCFs? For a pair of unrelated series, the CCF itself ought not to show autocorrelation, whereas genuine cross-correlation coefficients at different lags would be correlated with each other. However, as is clear from Fig. 2, even for two series that are not related, the CCF can show patterns of correlation between values at neighboring lags ("smooth behavior"). This also can be explained mathematically. This issue is not overcome by averaging CCFs across an ensemble of pairs of series. It can be seen in runs of significant cross-correlations above or below the zero line in prewhitened plots (see, e.g., Fig. 8 below). In contrast, if these were autocorrelation functions for white noise, there should be few such runs. Unfortunately, it seems in Fig. 4a (without prewhitening) that the individual CCF values are not randomly placed on either side of zero, but rather do form runs, implying autocorrelation. Figure 5 confirms this suspicion, showing that autocorrelation and partial autocorrelation (lags 1 and 2, in this case) is observed within the averaged raw data series CCF, but is largely removed by the prewhitening. Thus, the routine approach of obtaining multiple CCFs and averaging the resultant coefficients is still incompletely secure, even for a known unimodal distribution of participant performance parameters. The approach using prewhitening is acceptable, and may be informative.

There are circumstances in which one might consider averaging the multiple time series themselves prior to cross-correlation and transfer function analysis. This would be reasonable when a given participant performed a particular response time series several times (as in many EEG experiments), especially when the input referent was constant across repetitions. It might also be useful with a constant input referent or referents and multiple participants responding. With multiple simultaneous input/output referent series, prewhitening is no longer readily applicable, because there may be multiple distinct autoregressive features in the different inputs/outputs. Vector autoregression is often required in such a multivariate situation.

We illustrate the consequences of averaging multiple time series themselves in Fig. 6, using exactly the same set of 30

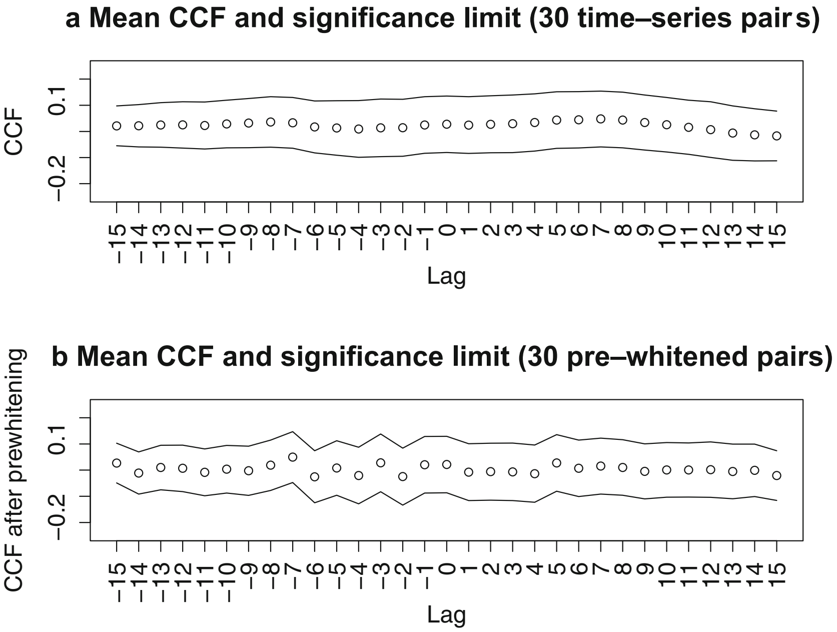


Fig. 4 In all, 30 simulations of a pair of ar(1) autocorrelated series, each comprising 100 events, were generated. The conditions were identical to those of Fig. 1, with the same autoregressive correlation coefficient for each series (.9). CCFs for each pair were obtained, and the statistics of each coefficient are shown, with the solid line representing the 95 % significance limit for each obtained from its standard error. The top graph (**a**) shows the results when the raw series are analyzed; the bottom (**b**), when each pair is prewhitened first

pairs of series used in Figs. 4 and 5. Thus, the 30 individual time series from a group are averaged, giving two grand average series, and these are studied further. Figure 6a shows that the CCF for the raw pair of grand average series has no significant coefficients when the proper significance limit (solid horizontal, ±0.6) is considered. Thus, the averaging

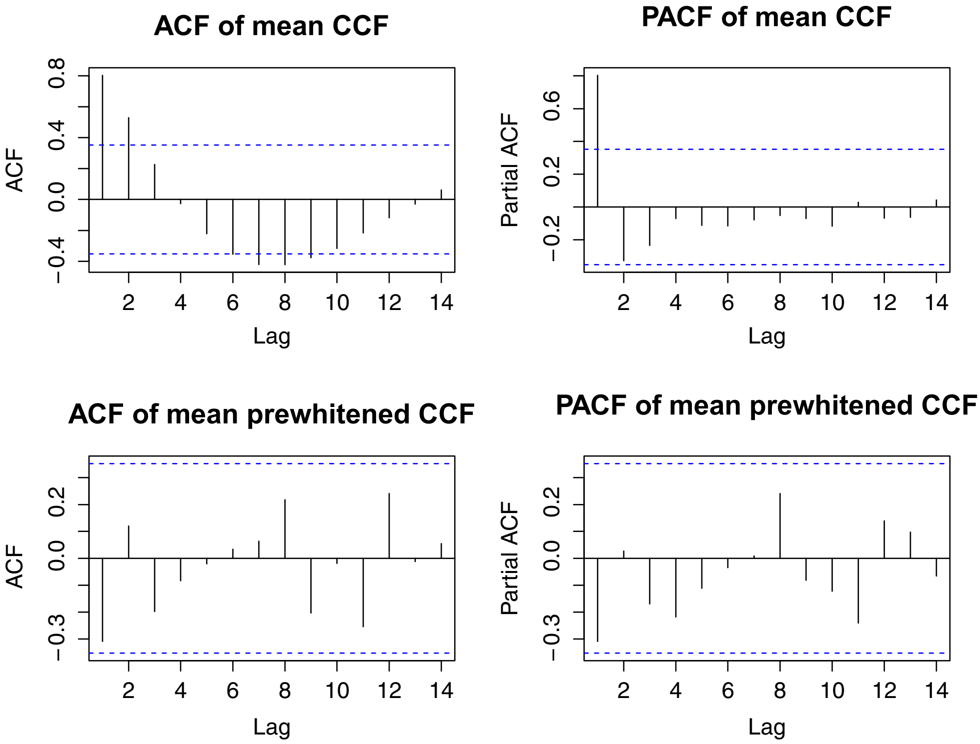


Fig. 5 This figure is based on the same data set as Fig. 2. The top pair of graphs are the autocorrelation and partial autocorrelation functions for the CCF in Fig. 2a (from the raw data series). The bottom pair of graphs are the corresponding functions for the CCF in Fig. 2b (from prewhitened data series)

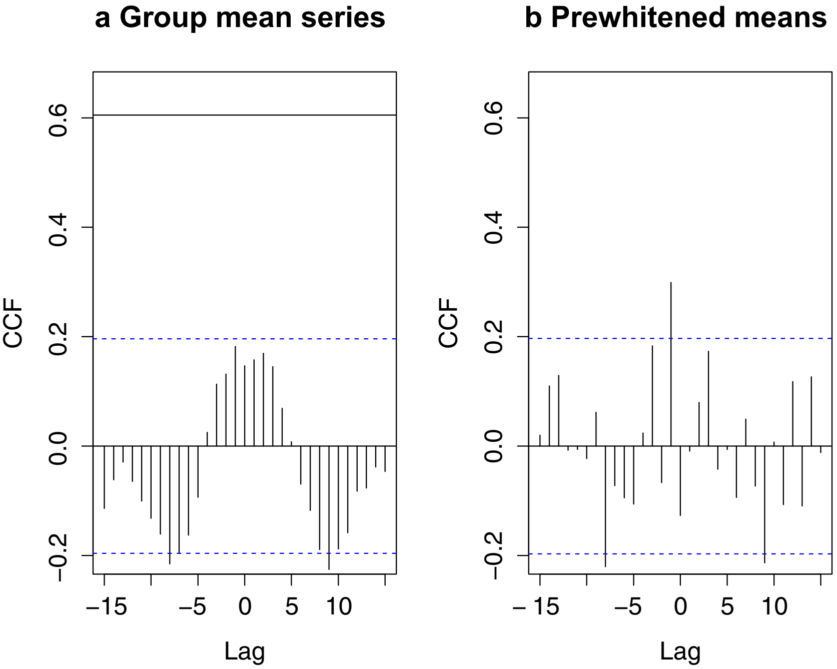


Fig. 6 The figure is again based on the same data set as Fig. 2. In this case, the two groups of 30 series were each averaged before further processing. The left panel shows the CCF for the resultant pair of grand average series, together with the realistic significance limit calculated as before (solid line), whereas the right panel shows the resultant pair of series after prewhitening, together with the appropriate cross-correlation significance limit (dotted line)

approach seems a plausible one, but again, the CCF itself indicates autocorrelation. When the pair of grand average series are prewhitened (Fig. 6b), the CCF autocorrelation is largely removed, though the lag 1 cross-correlation shows a spurious significance, compatible with the 95 % significance limit chosen, given that there are 30 coefficients. This same lag cross-correlation shows spurious significance even when each pair of time series is individually prewhitened before the averaging and CCF determination (Fig. 7); though again, this CCF appropriately shows no autocorrelation. But as we note in the figure caption, this particular approach is not strongly justified, since in constructing the grand averages of prewhitened series, autocorrelation may not only remain in the *y* set (as already noted), but also be reintroduced in the *x* set, thus nullifying the objective of prewhitening. The occurrence of an apparently significant cross-correlation again emphasizes the need for caution in both generating and interpreting CCFs, and points to using them as a data exploration tool prior to transfer function modeling.

A general approach for the autoregressive modeling of a pair of related time series

It may be useful at this point to summarize a simple approach that results from the considerations above. We will follow this with two worked examples using our own data.

The objective is to understand whether an autocorrelated time series *x* is predictive of an autocorrelated time series *y*, and to produce a model of *y* comprising its own autoregressive function, the transfer function representing the impact of *x* on *y*, and a white-noise error term (i.e., one that no longer contains any autoregression or any other structured information). We first stationarize the series (if necessary) and apply the same

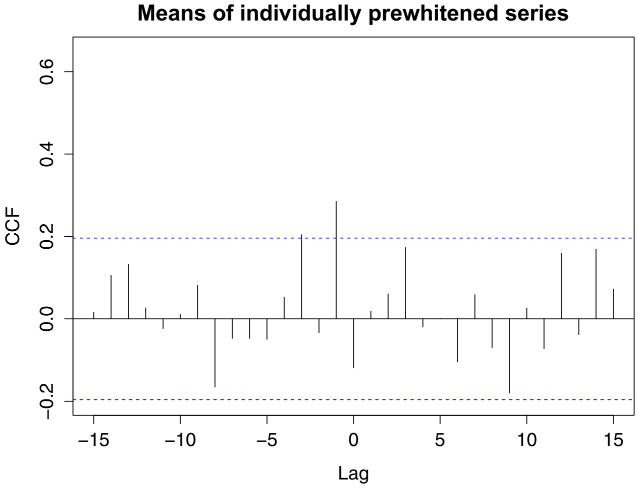


Fig. 7 Again based on the same data set as for Fig. 2, this figure shows the results of prewhitening individual series pairs and then taking the pair of grand averages of the two sets as the basis for a CCF. Note that the resultant averaged series may still be autocorrelated, but there is no strong rationale for a further prewhitening step, given that this process has already been undertaken

transformation, usually a single differencing, to both series. Cross-correlation after prewhitening is then our opening form of data exploration, and it gives cues as to which lags of *x* may be influential on *y*; it may also forewarn us whether there are signs of bidirectionality that need to be investigated separately.

Possible predictive influences are then tested by means of Granger causality, which allows us to solidify our impression of which influences are suitable for inclusion in a transfer function. Then an estimate of the required autoregressive order for the stationarized *y* series is made (using the autocorrelation and partial autocorrelation functions). This is followed by an exhaustive search through models containing the lags of *x* indicated by the CCF and the Granger causality analyses. We use optimization by means of the Bayesian information criterion (BIC), so as to aim for parsimony, rather than simply to obtain a best fit (lowest error) by overfitting. Alongside the BIC, one considers the role of individual coefficients that are in themselves not significant, since sometimes these are necessary for an optimally parsimonious model. The final model then displays transfer function coefficients whose confidence limits can be assessed, and whose overall impact can be estimated.

We illustrate the process, in outline, in the following two examples. The first case is from averaging across multiple performances of a particular response time series in relation to multiple predictor series, as we have just discussed; in the second case, we studied pairs of individual realizations of a tapping response series. The principles that we illustrate can be directly applied to most repeated measures or multiparticipant cases, and the mixed-effects modeling technique of cross-sectional time series analysis (sometimes called *panel* or *longitudinal* data analysis) can allow such transfer functions instead to be determined from the complete set of individual series, without preaveraging. We have discussed this approach in depth for time series of the perception of affect in a recent pair of studies (Dean et al., 2014a, b).

Illustrating the analysis of the transfer function model relating two series that display Granger causality

Example 1: Relationships between continuous perceptions of change and expressed affect while listening to Webern's Piano Variations Our first target is from recent work on pairs of continuous perception variables obtained from participants listening to ~3-min excerpts of four pieces of music (Bailes & Dean, 2012; Dean & Bailes, 2010, 2011; Dean et al., 2014a, b). The participants responded successively in two tasks—a continuous perception-of-change task and a perception-of-affect task (counterbalanced)—using a computer interface (sampling rate 2 Hz). In brief, participants had to move a mouse (scrub) quickly when they perceived that the music was changing fast, more slowly as they perceived slower change, and keep the mouse still for as long as they felt the music to be unchanging. In the affect task, the Russell 2-D circumplex model of affect was used, represented in a computer interface with "arousal" and "valence" at right angles to each other. Participants were trained to move the mouse in order to represent simultaneously their continuous impressions of arousal (essentially, the level of activity expressed by the music) and valence (essentially, the pleasantness expressed by the music). Although the two tasks (perceived change and perceived arousal) were successive rather than simultaneous, they were each aligned synchronously with a single musical piece (a fixed referent), and hence, it is meaningful to consider the possible interrelation of the two perceptual responses. Full details of the pieces, participants, and procedures are given in the references quoted.

For this section of the *Piano Variations*, we showed previously that both acoustic features and continuous perceptions of change can together predict perceptions of arousal. Most of the work has been done with grand average series, obtained by simple averaging of the responses of a group of participants, though in other work we have undertaken detailed studies of individual and group variation in such continuous responses (Dean et al., 2014a, b; Launay et al., 2013). We use the grand average series for our illustration.

Here we solely focus on the possible relationship of the two continuous perceptual variables, change and arousal. The arousal series, in particular, required differencing for stationarity, so both series were first differenced: when a time series is called "name," we refer here for simplicity to its differenced form as "dname." Figure 8 (top) shows the raw cross-correlations between the two differenced series, dchange and darousal, both of which were autocorrelated. The dotted lines automatically provided by R are again essentially the ccsl, but we reiterate that the ccsl is not to be relied on in these circumstances. Given that $n = 382$ points are in each differenced series, the $p < .05$ ccsl is now a smaller value than in Fig. 1. The solid lines, in contrast, are the nonparametric estimates of the confidence limit, based on the empirical acfs (see the supplemental materials). The results suggest that only a few early lags of dchange are likely predictors. Figure 8 (bottom) shows the CCF after prewhitening on the basis of the autocorrelation structure of the differenced change series, which we are testing as a predictor of differenced arousal (darousal). It suggests that the first three lags of dchange are likely to be influential, but a few later lags could also be predictors.

Granger causality between dchange and darousal was then tested. Consistent with the implications of the prewhitened CCF—using one to three lags of dchange, for example—Granger causality was highly significant: $p < .001$. We went on to model darousal using ARX—that is, autoregression with dchange treated as an eXogenous predictor (the source of the transfer function). Table 1 and Fig. 9 show that the optimal solution was a good model, involving three lags of dchange and an autoregressive (orders 1 and 10) process in darousal. In

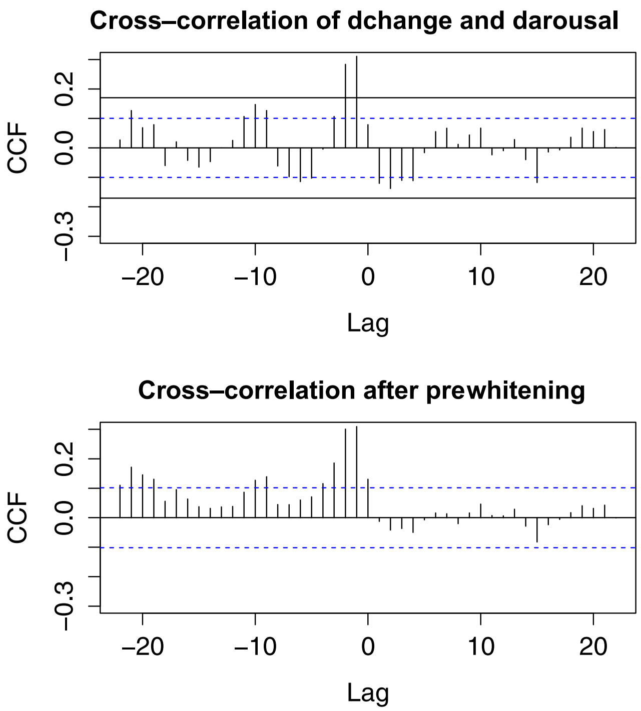


Fig. 8 The top graph shows the (misleading) cross-correlations obtained with our raw dchange and darousal time series of responses to a segment of Webern's *Piano Variations*. The bottom shows the prewhitened relationships, suggesting that lags 0–3, in particular, might be predictive. This was confirmed by Granger causality assessment, and in the subsequent modeling (see the text)

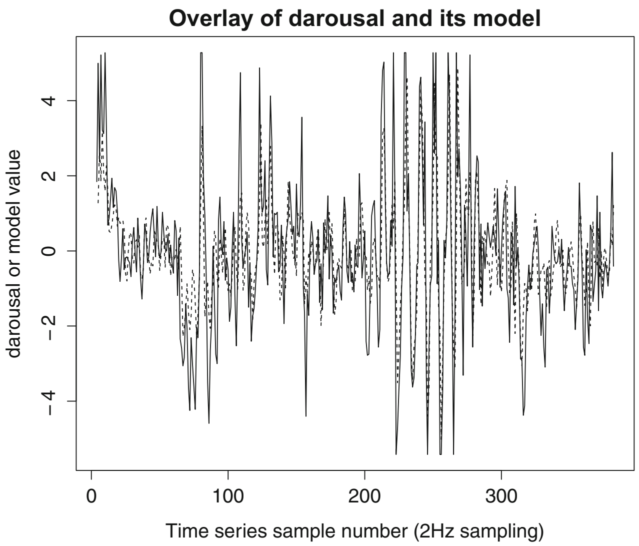


Fig. 9 Data and model fit for the Webern darousal series. Although the two series are very similar, there are numerous disparities; this is not surprising, given that this is only a partial model of the system (see the text), and a fuller model has been provided elsewhere

Fig. 9, close inspection shows numerous minor disparities between the data and model, but this is not surprising, given that in earlier studies with additional predictors, such as the acoustic variables, models of darousal can be enhanced substantially.

The result gives strong support to the view that dchange is indeed a significant predictor of darousal in this particular set of responses to the Webern piece, and provides quantitation. Previously, we used vector autoregression, with acoustic variables included (Bailes & Dean, 2012; Dean & Bailes, 2010, 2011). These additional approaches allow for assessing all the putative interacting acoustic and perceptual variables together, and allow for the possibility that influences operate in either or both directions between variables, when appropriate. As we have shown here, perceptions of change can influence

Table 1 Optimal model of darousal based on a dchange transfer function and autoregression only

The R call (as described in the main text; and see the Supplemental Material for more detail and explanation):

```
arima(x = x.darous, order = c(10, 0, 0), xreg = x[ , 3:6] , fixed = c(NA, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, NA, NA, NA, NA))
```

Coefficients:

	ar1	ar10	(lO)dchange	(l1)dchange	(l2)dchange	(l3)dchange
coefficient	0.425	0.113	11.106	25.072	23.846	10.632
SE	0.047	0.048	2.129	2.527	2.541	2.125

sigma^2 is estimated as 2.495, log likelihood = –709.36, BIC = 1,460.27

Confidence intervals of the coefficients, which can suggest testing the effect of eliminating any that might be zero (as was done with earlier models in this process):

	2.5 %	97.5 %
ar1	0.333	0.517
ar10	0.019	0.208
(l0)dchange	6.935	15.278
(l1)dchange	20.118	30.0256
(l2)dchange	18.865	28.826
(l3)dchange	6.468	14.797

Selected elements of the R results for the BIC-optimized model are shown. Ar1 and 10 are the autoregressive components for darousal, lags 1 and 10; (l0)dchange is the lag 0 of dchange, and similarly for lags 1–3 (these constitute the transfer function components). The intercept term was not significant and so was not included (as is normal with models of differenced series). The "fixed" vector in the R code is used to set the ar coefficients for lags 2–9 and for the intercept to 0, whereas "NA" leaves the corresponding parameter, such as the coefficients on the lags of dchange, free to be optimized (see the supplemental materials for more explanation of the syntax). The supplemental materials also show some of the important assessments of the quality of the model (such as a lack of autocorrelation in the residuals), which should follow before a model such as this is accepted.

perceptions of arousal. Clearly, acoustic features cannot be influenced by perceptual ones, but the perception of loudness, for example, might be influenced by the perception of expressed arousal. The procedural illustration here shows how to avoid spurious cross-correlations and proceed to a meaningful transfer function model for any particular pair of continuous perception time series variables. Subsequent analyses can then probe further.

Example 2: The relationship between tapping and an irregular pacer tone during a motor synchrony task This example provides a more complex situation, from within a large field common to music, performance, gait, and motor studies: tapping (or otherwise moving) with an audible pacer stimulus. Somewhat like studies of mutual adaptation between pairs of tappers, the pacer series was substantially varied with time, though not adaptive (not responding to the tapper). The data (previously unpublished) are taken from the set obtained during the work described in Launay et al. (2013), where full details of the experimental procedure are presented. The data used here are provided with the R script in the supplemental materials. That article solely described off-pacer tapping (i.e., trying to avoid tapping with the pacer sounds, but rather to tap between them), whereas here we discuss one individual participant's performance of our other task, in which the instruction was to tap simultaneously with the pacer. Figure 10 illustrates the interonset interval (IOI) series for the pacer (64 events), which was constructed as an irregular Kolakoski sequence, varying in IOI around a standard value of 600 ms. In the first segment of this, the deviations were 45 ms; in the second, 22.5 ms. The standard stationarity test is not powerful enough to detect the changing variance built into the construction of the series (stationarity tests are often focused on stationarity of the mean). Like tapping studies with oscillating IOIs, the Kolakoski sequence shows some higher-order features. It involves a conditional choice of whether an IOI is shorter (S) or longer (L) than the mean, such that no more than two of either type are adjacent to each other. It takes several events at the minimum to cover all of the triplet patterns this involves, even when viewing the sequence with a triplet moving window of hop one (e.g., viewing LSSLLSLS as consisting of the successive windows LSS, SSL, etc.). The lower panel shows the sequence of asynchronies for an individual response to this pacer stimulus, in which a negative asynchrony means that the tap occurred slightly before the audio tone it was matching (as is common under similar conditions), and a positive asynchrony means that the tap occurred slightly after; this suggests a clear shift at the time the pacer pattern changes. The analytical question that we address here is, is the pacer IOI (referent) series a statistical predictor of the response asynchrony series? It seems that assessing this or closely related questions is an underlying purpose common among many of the published motor cross-correlation studies that have led us to write this article.

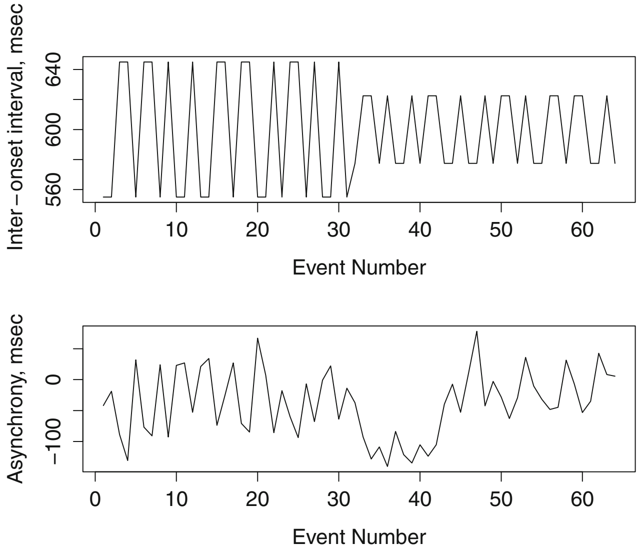


Fig. 10 Tapping synchronously with an irregular pacing audio stimulus (the referent process). The top panel shows the series of interonset intervals of the pacer, based on a mean of 600 ms and deviations derived from the application of a Kolakoski series. In the first segment, the deviations are 45 ms, and in the second, 22.5 ms

Figure 11 shows the CCFs of the raw (upper) and prewhitened (lower) series. The raw CCF, assessed against the realistic significance limit (solid line), suggests that lag 0 might be a powerful influence, though the $p < .05$ condition demands

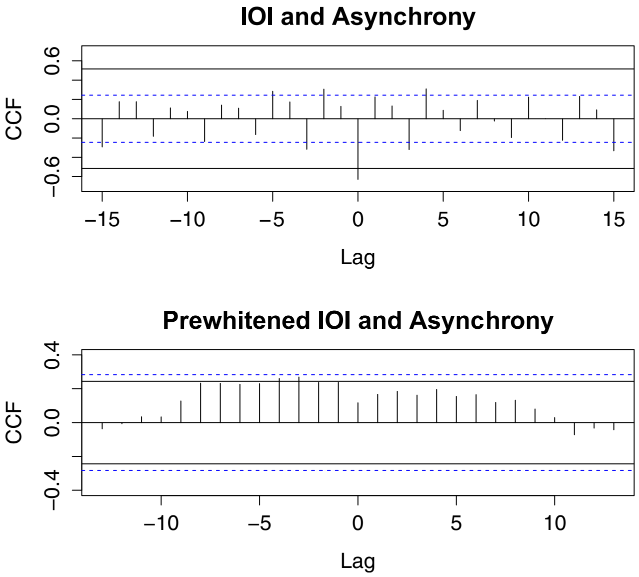


Fig. 11 Cross-correlation functions determined on the interonset interval (IOI) and asynchrony series, before (top panel) and after (lower panel) prewhitening. In each case, the horizontal dotted lines show the default significance limit presented by R. The solid horizontal lines in the top panel are a reasonable estimate of the significance limit, calculated as described in the text, taking account of the autocorrelation of both series. The solid horizontal lines in the bottom panel mark the conventional cross-correlation significance level (see the text), which is a reasonable estimate of the significance limit only given prewhitening

caution about this. In the prewhitened (lower) panel, the dotted lines mark the default estimate of the CCF significance limit provided by R (which is based on uncorrelated series), which here is actually slightly more stringent than the ccsl as we have defined it (solid lines). Only lags 3 and 4 show any sign of being significant. This is a case in which the prewhitening does not provide useful information, probably because the automatic ar order selection that is part of the prewhitening chooses a high order, reflecting the complexity of the stimulus. Similar issues can occur when prewhitening series in which successive IOI values follow, for example, a cyclic pattern spread over many IOIs. Taking the figure as a whole, no convincing suggestions as to significant cross-correlations can be made, though one should continue to consider possible influences of lag 0, and more skeptically, lags 3 and 4.

A Granger causality assessment confirmed that for lags up to 3 ($p < .003$) or 4 ($p < .006$), there is a predictive relationship. Proceeding to modeling, we observed that the IOIs form an ar(1–3) series, and the asynchronies were ar(1,3). Consequently, on the basis of the BIC, the best model of the relationship was the one summarized in Table 2. Note that because of the four lags of IOI used as predictors during model development, the final modeled series contains only 60 of the original 64 events. The simple and clear-cut result, by no means obvious from the preceding cross-correlations, is that lag 0 of the IOI is a strong predictor with a large negative coefficient, consistent with the interpretation that this participant was predicting the completion time of each IOI reasonably successfully, in spite of their irregularity. This is consistent with earlier studies of synchronization with more regular (or entirely isochronic) IOI series. When the IOI is shortened, the tap is expected to be late, and vice versa, creating the negative coefficient, and consistent with many observations (Repp, 2005).

As we noted, the audio pacer series comprised two segments, with deviations from IOI = 600 ms of 45 and 22.5 ms, respectively. However, the addition of a segment dummy variable (set at 0 and 1, respectively) worsened the model. The mean asynchronies were actually larger overall during the second (–44.4 ms) than during the first (–31.4 ms) segment, although not statistically different, and mechanistic differences between the two segments might be found on further analysis, limited of course by the short lengths of the segments. Figure 12 shows the reasonable fit given by the model. The model accounted for 78 % of the sum of squares of the data, whereas a solely autoregressive model accounted for only 52 %.

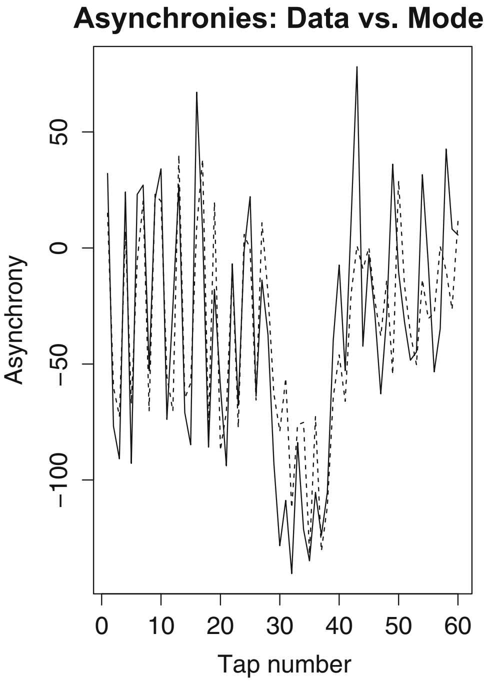


Fig. 12 Observed asynchronies (for the slightly curtailed series resulting from the modeling process: solid line), superimposed with those predicted by the model (dotted line)

Discussion

Summary and implications We have shown clearly that cross-correlations between pairs of time series, or even pairs of series derived as averages of sets of series, can be misleading. The key means of avoiding such spurious cross-correlations is to prewhiten the series being cross-correlated. But even then, some spurious correlations may remain, and the results need to be treated with caution. Not only is critical interpretation necessary, but also an awareness that certain kinds of time series may not be appropriate for the prewhitening approach—for example, when data are binomial or the series show only sparse change. Such results can still be susceptible to an informative transfer function analysis, as we

Table 2 Optimal model of the motor asynchrony series based on referent (pacer) IOI and autoregression only

Call:

```
arima(x = x.asynch, order = c(3, 0, 0), xreg = c(x[ , 3:5]), fixed = c(NA, 0, NA, NA, NA, 0, 0))
```

Coefficients:

	ar1	ar3	intercept	(l0)ioi
Coefficients	0.489	0.290	517.985	-0.918
SE	0.108	0.111	62.609	0.101

sigma^2 estimated as 924.5, log likelihood = –290.42, Akaike information criterion = 588.85, BIC = 601.32

have illustrated. Thus, in the later sections, we illustrated how Granger causality and ARX modeling can achieve the definition of an appropriate model. We emphasize that this remains a predictive model, and causal intervention experiments are commonly necessary to determine whether the model genuinely captures influences at work in the system.

We have tried to make apparent how the autocorrelated nature of most psychological and neurophysiological time series dictates that the successive data points are not i.i.d., and thus not susceptible to many conventional statistical analyses. Such autocorrelation also creates the need for the careful determination of the significance limit against which to assess the CCF relating a pair of time series, as we have detailed, and as in the code that we have provided. Autocorrelation also has critical dangers when one takes short chunks (windows) from time series for analytical purposes, as we have briefly mentioned in discussing rolling windowed cross-correlations. A few additional implications of autocorrelation bear discussion.

In evoked response potential studies, multiple repetitions of a precisely given stimulus allow the realistic averaging of the chunks of EEG signal corresponding to the periods before, during, and immediately after the stimulus. This permits the comparative analysis of such chunks. But even in EEG, if one deals with responses whose timing is not reproducibly related to the stimulus; or situations in which the stimulus is not precisely defined, yet an output signal alteration can be detected and requires analysis; or even with trials that are simply successive and repetitious (Dyson & Quinlan, 2010), then autocorrelation again requires critical assessment. We have faced the situation of undefined timing relationships in interpreting skin conductance changes and potentially triggered skin conductance responses during pianists' improvisatory performances (Dean & Bailes, 2015), and dealt with it by the use of ARX, together with change-point and segmentation analyses.

A common situation in EEG that presents analogous problems is when the overall energy in a given frequency band of the signals needs to be compared between different time periods, and/or different electrodes, or different interacting participants, as in studies of interbrain "coupling" (Dumas, Nadel, Soussignan, Martinerie, & Garnero, 2010). Detailed methodologies for the nonparametric assessment of such differences and relationships include using permutation tests (Maris & Oostenveld, 2007), which involve permuting in every possible way the chunks of the data series. This, of course, raises the dangers of chunking that we have described, and these authors are very clear and specific that autocorrelation must be removed first, otherwise the analyses are invalid. This is particularly difficult when the analyses involve multiple simultaneous time series (from different electrodes). However, Maris and Oostenveld do not present methods for removing autocorrelation, nor did they assess whether this was achieved in the data sets that they analyzed. Several other authors have addressed this point further (Langheim, Murphy, Riedner, & Tononi, 2011; Ma, Wang, Raubertas & Svetnik, 2010), but as we noted in the introduction, only a very small minority of authors have explicitly and properly addressed such considerations in relation to chunking in general, or EEG in particular.

It is also interesting that some analytical processes may themselves create serial correlation in data: this occurred in the study of Schulz and Huston (2002) on rat behavior. Their key step was to rank a set of data and then order them on that basis (Schulz & Huston, 2002); this created serial correlation in the reordered set. Consequently, the subsequent windowed cross-correlation of pairs of data sets involving such preranked data attracted the problems of precision and significance defined already. The authors demonstrated this themselves by showing cross-correlations even when they started with randomly generated data, but as with their real data, the modest cross-correlations were sufficiently small that their significance was in question (Schulz & Huston, 2002).

Within the movement literature, Livingstone, Palmer, and Schubert (2012) quoted Schulz and Huston (2002) as they used continuously expanding CCF windows (which again created serial correlations between the successive window parameters) to successfully characterize transition points in perceptual responses and musical structures. Livingstone et al. did not place statistical significances on the results, and used a somewhat arbitrary mechanism for relating the slightly different results from windowing forward from the start versus backward from the end.

The potentially misleading aspects of CCFs point to the need for formulating hypotheses and analyses sufficiently clearly that all of the relevant lags of a CCF are ultimately considered. In doing this, some of the disparities introduced by restricted CCF analyses may be understood and removed. If there are multiple significant lagged relationships, their effects need to be integrated. Thus, let us briefly consider the case of a pair of tappers adapting to each other in trying to obtain simultaneity, or a fixed phase relationship, rather like pairs of musicians commonly do, and as has been studied by several groups (Pecenka & Keller, 2011). Granger causality and ARX analyses permit an estimation of the integrated significance of the influences of tapping stream x on tapping stream y, and vice versa. VAR or VARX (i.e., vector autoregression, with or without a fixed referent eXogenous predictor) analyses allow the consideration of the influence of any defined referent elements, such as duration notations or a pacer stimulus. VAR (allowing for a reciprocal relationship between the performed stream and referent) may even be useful in that the pacer may be adaptive, or the notational implication may be transmuted by the immediate actions of the participants. Thus, if the question assessed is formed as "does x influence y from a statistical perspective," then a clear answer can be obtained—and similarly for y influencing x and for any interactions with referents. If, on the other hand, some of the significant CCF lags are neglected—or if, as we illustrated above, the CCF is misleading or uninformative—then without a transfer function or VAR(X), inferences may not be secure.

Further solutions: ARX transfer functions, VAR, and cross-sectional time series analysis Our supplemental materials provide R code to illustrate the appropriate steps for CCF, Granger analysis, and ARX; the materials do not go deeply into assessment of the quality of models, in terms of autocorrelation-free white-noise residuals, lack of outliers amongst the residuals, and other criteria that always need to be considered carefully.

In some situations, prewhitening is not useful or even appropriate. For example, if the potential predictor variable is binary, or discrete, rather than continuous, this form of analysis will not be suitable. A related issue is that there may be individual continuous data series in which a large mass of the differenced data are zero (i.e., no alteration between successive values of the original series); this can potentially undermine standard time series approaches, because the data distribution is so skew, and it may be resolved by a binomial approach in conjunction with generalized linear ARX, developed by author W.T.M.D. (Dean et al., 2014a, b), or by aggregating multiple response series as is often done, so that zero change points are largely removed.

The first example we analyzed above for demonstrative purposes has further complexity to be considered, as we have shown in earlier publications and as has been indicated previously (Pressing, 1999). For example, there could be reciprocal relations between the perception of change and that of arousal. In the present analysis, when a Granger causality test of the influence of the perception of arousal (as darousal) on that of change (as dchange) was run (the converse of the analysis discussed above), it showed no lagged relationships. This was also confirmed with the CCF of the prewhitened series. So there was unlikely to be an influence of the perception of arousal on that of change, whereas the converse influence was demonstrable. In our second example, the pacer stimulus was not dependent on the tapping, and hence, a consideration of such reverse-direction Granger causality did not arise, whereas in mutual dyadic movement studies, it does.

Dynamic relationships may apply in time series influences, where the nature of the influence of series *x* on series *y* changes with time. As we mentioned above in relation to windowed cross-correlations, these cases require special modeling treatment; they may, for example, be revealed by changing autoregressive components and their coefficients, or by changing the exogenous predictor lag and coefficient patterns. Time-varying spectral analysis can also be useful here. Once one has multiple predictors to consider, rather than just two, VAR is often appropriate (Pressing, 1999). This occurs in part because VAR allows for treating variables that might be simultaneously mutually influential in such a way as to reveal this mutual influence. Indeed, VAR is used as part of the Granger causality analysis.

We have also alluded above to the situation in which substantial differences are of interest between the behaviors of different participants. Using the mixed-effects approach of cross-sectional time series analysis, and retaining the integrity of all of the individual data series, the influence of interindividual variability can be appropriately assigned, such that the core transfer functions are more reliably revealed (Dean et al., 2014a, b). In the process, when needed, the interindividual variation can itself be further analyzed. In many movement or perception studies, such as of tapping or the electrophysiology of affect in music (Chapados & Levitin, 2008), it is routine for a large proportion of trials or events in a continuous series to be discarded as inappropriate: mixed-effects analysis often permits retaining all or the vast majority of the data (Launay et al., 2013), hence providing a more conservative, potentially stronger, analysis.

The interpretations discussed here are those of statistical inference: predictive correlative, or what is often called *predictive causal*, relationships (Granger causality), rather than necessarily causal ones, in the philosophical or mechanistic sense. To further assess mechanistic causalities, intervention experiments are called for, such as those that we earlier undertook successfully in relation to the influence of acoustic intensity change profiles upon the continuous perception of affect (Dean, Bailes, & Schubert, 2011). The time series analysis approach is one way to reach a level of understanding of the complex temporal relationships that can guide such intervention studies. Similar thoughts apply to newer, high-potential techniques for the functional data analysis of time series (Hyndman & Shang, 2012; Ramsay & Silverman, 2002; Ramsay & Silverman, 1997).

Concluding comments The motivation for this article was to encourage the awareness of both the limitations and uses of cross-correlation analysis in the community of researchers dealing with psychological, and particularly perceptual and motor–performative, time series. As we have shown, cross-correlations between pairs of autocorrelated series may be spurious, and hence misleading. We noted here that closely related problems flowing from the occurrence of autocorrelation also make conventional measures of mutual information or spectral coherence inappropriate or difficult to interpret. For example, mutual information has been derived as a parameter of random variables, and autocorrelated variables require a different formulation.

Quite often, the data under study comprise multiple replications of particular time series conditions—for example, by different participants—and the purpose may be to get an integrated view of relationships, and/or to get a simple dependent variable representing each participant that can be used in further analyses. In our first example, four transfer function coefficients could be measured for each individual participant and then so used; in the second example, one. In either case, other, simple single parameters could be obtained for use as dependent variables or covariates, such as reduction in the summed squares of the residuals due to the transfer function (or, equivalently, in the mean squared errors), as is summarized in the R script for Example 2.

Author note We acknowledge support from the Australian Research Council.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Bailes, F., & Dean, R. T. (2012). Comparative time series analysis of perceptual responses to electroacoustic music. *Music Perception, 29,* 359–375.

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods, 7,* 338–355. doi:10.1037/1082-989X.7.3.338

Brunsdon, T. M., & Skinner, C. J. (1987). The analysis of dependencies between series in psychological experiments. *British Journal of Mathematics and Statistical Psychology, 40,* 125–139.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. doi:10.1038/nrn3475

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 63,* 289–300.

Chapados, C., & Levitin, D. J. (2008). Cross-modal interactions in the experience of musical performances: Physiological correlates. *Cognition, 108,* 639–651. doi:10.1016/j.cognition.2008.05.008

Cryer, J. D., & Chan, K. (2008). *Time series analysis with applications in R*. New York, NY: Springer.

Dean, R. T., & Bailes, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception of music. *Empirical Musicology Review, 5,* 152–175.

Dean, R. T., & Bailes, F. (2011). Modelling perception of structure and affect in music: Spectral centroid and Wishart's red bird. *Empirical Musicology Review, 6*(2), 1–7.

Dean, R. T., & Bailes, F. (2015). Using time series analysis to evaluate skin conductance during movement in piano improvisation. *Psychology of Music, 43,* 3–23. doi:10.1177/0305735613489917

Dean, R. T., Bailes, F., & Dunsmuir, W. T. M. (2014a). Shared and distinct mechanisms of individual and expertise-group perception of expressed arousal in four works. *Journal of Mathematics and Music, 8,* 207–223. doi:10.1080/17459737.2014.928753

Dean, R. T., Bailes, F., & Dunsmuir, W. T. M. (2014b). Time series analysis of real-time music perception: Approaches to the assessment of individual and expertise differences in perception of expressed affect. *Journal of Mathematics and Music, 8,* 183–205. doi:10.1080/17459737.2014.928752

Dean, R. T., Bailes, F., & Schubert, E. (2011). Acoustic intensity causes perceived changes in arousal levels in music: An experimental investigation. *PLoS ONE, 6*, e18591. doi:10.1371/journal.pone.0018591

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS ONE, 5*, e12166. doi:10.1371/journal.pone.0012166

Dyson, B. J., & Quinlan, P. T. (2010). Decomposing the Garner interference paradigm: Evidence for dissociations between macrolevel and microlevel performance. *Attention, Perception, & Psychophysics, 72,* 1676–1691. doi:10.3758/APP.72.6.1676

Enders, W. (2004). *Applied econometric time series* (2nd ed.). Hoboken, NJ: Wiley.

Gregson, R. A. M. (1983). *Time series in psychology.* Hillsdale, NJ.: Erlbaum.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

Himberg, T., & Thompson, M. R. (2011). Learning and synchronising dance movements in South African songs: Cross-cultural motion-capture study. *Dance Research, 29,* 303–326.

Hyndman, R., & Shang, H. L. (2012). ftsa: Functional time series analysis (R package version). Retrieved from http://robjhyndman.com/software/ftsa/

Konvalinka, I., Vuust, P., Roepstorff, A., & Frith, C. D. (2010). Follow you, follow me: Continuous mutual prediction and adaptation in joint tapping. *Quarterly Journal of Experimental Psychology, 63,* 2220–2230.

Langheim, F. J. P., Murphy, M., Riedner, B. A., & Tononi, G. (2011). Functional connectivity in slow-wave sleep: Identification of synchronous cortical activity during wakefulness and sleep using time series analysis of electroencephalographic data. *Journal of Sleep Research, 20,* 496–505.

Launay, J., Dean, R. T., & Bailes, F. (2014). Evidence for multiple strategies in off-beat tapping with anisochronous stimuli. *Psychological Research, 78,* 721–735. doi:10.1007/s00426-013-0513-9

Livingstone, S. R., Palmer, C., & Schubert, E. (2012). Emotional response to musical repetition. *Emotion, 12,* 552–567.

Ma, J., Wang, S., Raubertas, R., & Svetnik, V. (2010). Statistical methods to estimate treatment effects from multichannel electroencephalography (EEG) data in clinical trials. *Journal of Neuroscience Methods, 190,* 248–257. doi:10.1016/j.jneumeth.2010.05.013

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164,* 177–190. doi:10.1016/j.jneumeth.2007.03.024

Müller, V., Sänger, J., & Lindenberger, U. (2013). Intra- and inter-brain synchronization during musical improvisation on the guitar. *PLoS ONE, 9*, e73852. doi:10.1371/journal.pone.0073852

Pecenka, N., & Keller, P. E. (2011). The role of temporal prediction abilities in interpersonal sensorimotor synchronization. *Experimental Brain Research, 211,* 505–515.

Piryatinska, A., Terdik, G., Woyczynski, W. A., Loparo, K. A., Scher, M. S., & Zlotnik, A. (2009). Automated detection of neonate EEG sleep stages. *Computer Methods and Programs in Biomedicine, 95,* 31–46.

Pressing, J. (1999). The referential dynamics of cognition and action. *Psychological Review, 106,* 714–747. doi:10.1037/0033-295X.106.4.714

Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis.* New York, NY: Springer.

Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies*. New York, NY: Springer.

Repp, B. H. (1999). Control of expressive and metronomic timing in pianists. *Journal of Motor Behavior, 31,* 145–164.

Repp, B. H. (2002). The embodiment of musical structure: Effects of musical context on sensorimotor synchronization with complex timing patterns. In W. Prinz & B. Hommel (Eds.), *Common mechanisms in perception and action: Attention and performance XIX* (pp. 245–265). New York, NY: Oxford University Press.

Repp, B. H. (2005). Sensorimotor synchronization: A review of the tapping literature. *Psychonomic Bulletin & Review, 12,* 969–992. doi:10.3758/BF03206433

Schubert, E. (2006). Analysis of emotional dimensions in music using time series techniques. *Context: Journal of Music Research, 31,* 65–80.

Schulz, D., & Huston, J. P. (2002). The sliding window correlation procedure for detecting hidden correlations: Existence of behavioral subgroups illustrated with aged rats. *Journal of Neuroscience Methods, 121,* 129–137.

Sims, C. A. (1988). Bayesian skepticism on unit root econometrics. *Journal of Economic Dynamics and Control, 12,* 463–474.

Strang, A. J., Funke, G. J., Russell, S. M., Dukes, A. W., & Middendorf, M. S. (2014). Physio-behavioral coupling in a cooperative team task: Contributors and relations. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 145–158. doi:10.1037/a0033125

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?—A study in sampling and the nature of time-series. *Journal of the Royal Statistical Society, 89,* 1–63.

Zuur, A. F., Ieno, E. N., & Smith, G. M. (Eds.). (2007). *Analysing ecological data*. New York, NY: Springer.