Quantifying evidence for—and against—Granger causality with Bayes factors

Zita Oravecz[a,*] **and Joachim Vandekerckhove**[b]

Author final version. To appear in *Multivariate Behavior Research*.

Testing for Granger causality relies on estimating the capacity of dynamics in one time series to forecast dynamics in another. The canonical test for such temporal predictive causality is based on fitting multivariate time series models and is cast in the classical null hypothesis testing framework. In this framework, we are limited to rejecting the null hypothesis or failing to reject the null – we can never validly accept the null hypothesis of no Granger causality. This is poorly suited for many common purposes, including evidence integration, feature selection, and other cases where it is useful to express evidence against, rather than for, the existence of an association. Here we derive and implement the Bayes factor for Granger causality in a multilevel modeling framework. This Bayes factor summarizes information in the data in terms of a continuously scaled evidence ratio between the presence of Granger causality and its absence. We also introduce this procedure for the multilevel generalization of Granger causality testing. This facilitates inference when information is scarce or noisy or if we are interested primarily in population-level trends. We illustrate our approach with an application on exploring causal relationships in affect using a daily life study.

Granger causality | Bayes factor | multilevel vector autoregressive modeling

Technological advances are making multivariate, intensive longitudinal data increasingly prevalent. The general upsurge of such intense multivariate data affords new insights: We can now closely examine within-person dynamics, with unforeseen potential for addressing complex questions related to human behavior. When zooming in on within-person dynamics, we are often interested in the predictive capacity of our variables: can we predict current values of one time series from past values of another? For instance, does a parent's soothing behavior during a child's anger episode calm the child, aggravate the behavior, or have no effect? Or if an individual's emotional arousal increases, would they in turn also feel more or less pleasant?

Statistical inference can be performed on such predictive associations over time by testing for predictive causality, often called *Granger causality* (Granger, 1969). The idea behind this inference is that if over time, changes in some variable X "Granger cause" changes in variable Y, then past values of variable X should contain information that helps predict Y, above and beyond the information already contained in past values of Y alone. Certainly, Granger causality does not necessarily represent truly causal relations due to the possibility of omitted variables, but the inclusion of previous measurements of the variable itself enables this framework to provide more information on possible causal relations than simple correlational measures. That is to say, predictive causality analysis is limited in terms of inferring actual causality, and the causality concept here refers specifically to the forecasting of variables. All in all, Granger causality testing is a useful tool to determine whether a set of variables contains useful information for improving the predictions of another set of variables.

Granger causality testing in time series analysis is routine in, for example, the field of econometrics. For the social sciences, it has been Peter Molenaar's work that emphasized the usefulness of this approach and extended it to better fit the typical goals of social science research. Velicer and Molenaar (2013) gives a broad overview of time-series analysis methods for psychological research and highlights the utility of Granger causality testing within this framework. Molenaar and Lo (2016) summarizes generalizations of Granger causality testing by scaffolding on standard and structural vector autoregressive (VAR; see later) models, and includes approaches for handling heterogeneity and nonstationarity. Liu and Molenaar (2016) further adds to this by tackling challenges related to nonlinearities between frequency domain measures. Moreover, Molenaar (2019) describes a data-driven approach for unifying standard and structural VAR models in order to consolidate conclusions from Granger causality testing in these two VAR variants, and emphasizes how these causal relationships can be exploited for designing intervention studies.

Granger causality testing is typically done in the classical (frequentist) inference framework (for some recent exceptions using financial models, see Droumaguet, Warne, & Wozniak, 2016; Wozniak, 2016; Sen, Majumdar, & Sikaria, 2022). The classical hypothesis test for predictive causality, the Wald test (see, e.g., Lütkepohl, 2005, p. 102) can only 'reject' or 'fail to reject' the null hypothesis of *no* predictive causality. This means that we can only have binary conclusions, and cannot quantify degrees of evidence in our data in favor or against predictive causality. Even in terms of binary conclusions, we can never 'accept the null' of no Granger causality.

The Bayesian approach we introduce here allows researchers to quantify the evidence for or against Granger causality with a single number. The direct quantification of evidence is crucial for a number of purposes, including the support of incremental science – that is, allowing small amounts of hard-to-obtain evidence to stack across publications. Similarly, the ability to quantify support *against* an association can be essential in practice. As

[a]The Pennsylvania State University; [b]University of California, Irvine
All authors contributed to the final draft.
[*]Correspondence concerning this article should be addressed to Zita Oravecz (zita@psu.edu).
This work was supported by The John Templeton Foundation grant #48192. JV was additionally supported by National Science Foundation grants #1658303, #1850849, and #2051186.

an example, in studies of problematic child behavior, it is critical to know how much evidence we have that certain factors *do not* aggravate an undesirable outcome. More broadly, in the era of big data, it becomes increasingly important to be able to make informed decisions regarding which variables are worth monitoring, necessitating an informed metric for evidence in favor of or against predictive causality.

We introduce a Bayesian hypothesis test, using the Bayes factor (Jeffreys, 1961), in order to quantify evidence in favor or against Granger causality between variables changing over time. Our proposed Bayesian approach performs simultaneous inference on predictive dynamics on both the group and the individual level from multivariate time series data of multiple people. For more information on Bayesian hypothesis testing in general see, *inter alia*, Dienes (2016), Etz, Haaf, Rouder, and Vandekerckhove (2018), Mulder and Wagenmakers (2016), Rouder, Haaf, and Vandekerckhove (2018), and Vandekerckhove, Rouder, and Kruschke (2018). This Bayesian approach can help applied researchers make decisions based on substantive goals – an applied example using core affect measurements in experience sampling settings is given in the Application section. Our work represents an initial step towards developing a novel way of making inference about Granger causality, with a proof-of-concept illustration. We have made scripts and data accessible to carry out the inference featured in the Application section of this paper on the Open Science Framework (OSF).[1]

Granger causality testing in vector autoregressive models

We start by specifying the time series model on which our Bayesian hypothesis test will be based. Define the temporal evolution of a single random variable y over time t as a univariate autoregressive model, specified as $y_t = \nu + \alpha_1 y_{t-1} + \ldots + \alpha_l y_{t-l} + u_t$, with ν representing a possibly nonzero intercept; u_t some forecast error; and α_l the dependencies on past observations of the variable quantified in terms of autocorrelation coefficients, at different lags l. Lagged relationships have important value for forecasting, as some variables of interest only change gradually, so that current and past data can reliably predict future trends.

Certainly, more still can be learned by studying the *joint dynamics among multiple phenomena*. The vector autoregressive (VAR) model extends the predictive framework of the autoregressive model by also accounting for interdependencies among time series of multiple related variables evolving over time. This way each variable's temporal evolution is not only predicted from its own past values, but also by past values of related variables. In this project we limit our attention to linear dependencies in the standard VAR model. Formally, a K-dimensional VAR(L) is specified as[2]:

$$\mathbf{y}_t = \boldsymbol{\nu} + \mathbf{A}_1 \mathbf{y}_{t-1} + \ldots + \mathbf{A}_L \mathbf{y}_{t-L} + \mathbf{e}_t, \qquad [1]$$

where $\mathbf{y}_t = (y_{1t}, \ldots, y_{Kt})'$ is a $(K \times 1)$ vector, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)'$ is a $(K \times 1)$ vector of possibly nonzero intercepts, $\mathbf{e}_t = (e_{1t}, \ldots, e_{Kt})'$ is a K-dimensional white noise or innovation process, which in its simplest form can be represented as a sequence of independent and identically distributed random K-vectors with zero mean vector and covariance matrix $\boldsymbol{\Sigma}$. Finally, each $\mathbf{A}_i$ is a $K \times K$ coefficient matrix of the lagged and cross-lagged effects at lag l. Specifically, $\boldsymbol{\Sigma}$ and $\mathbf{A}_l$ are defined as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2_{1,1,e} & \ldots & \sigma_{1,K,e} \\ \vdots & \ddots & \vdots \\ \sigma_{K,1,e} & \ldots & \sigma^2_{K,K,e} \end{bmatrix}$$

and

$$\mathbf{A}_i = \begin{bmatrix} \alpha_{1,1,l} & \ldots & \alpha_{1,K,l} \\ \vdots & \ddots & \vdots \\ \alpha_{K,1,l} & \ldots & \alpha_{K,K,l} \end{bmatrix}.$$

Let us consider a two-dimensional ($K = 2$) lag 1 ($L = 1$) VAR as an example. This is a special case of Equation 1 (i.e., the bivariate VAR(1) model) specified as:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} + \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,1} & \alpha_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix} \qquad [2]$$

where ν_1 and ν_2 represent the two elements of the intercept vector $\boldsymbol{\nu}$, $\alpha_{1,1}$ and $\alpha_{2,2}$ are the autocorrelations in the two dimensions, respectively, and parameters $\alpha_{1,2}$ and $\alpha_{2,1}$ are cross-lagged effects. The last part denotes innovation errors, $e_{1,t}$ and $e_{2,t}$, which can be assumed to be bivariate normally distributed with mean zero and covariance matrix $\boldsymbol{\Sigma}$. This 2×2 matrix $\boldsymbol{\Sigma}$ contains the residual variances for the two time series, σ_1^2 and σ_2^2, in its diagonals, while the off-diagonal element $\sigma_{1,2} = \sigma_{2,1}$ expresses contemporaneous association: the residual covariation between the two time-series. While the contemporaneous association must be symmetrical, the cross-lagged effects are not. The $\alpha_{1,2}$ and $\alpha_{2,1}$ coefficients quantify the predictive power of one time series component on the other, after controlling for past history of this latter component, hence capturing the directionality of predictive dynamics.

As an illustration, we generated two sets of time series data from the bivariate VAR(1) model described by Equation 2, with different settings for the cross-effect parameters. These simulated time series are depicted in the panels of Figure 1. For the bivariate set in the left panel, we set the cross-effect parameters by assuming no lagged dependency between the time series ($\alpha_{1,2} = \alpha_{2,1} = 0$), while in the right panel the time series were simulated with predictive association: past values of Series 2 are predictive for Series 1 ($\alpha_{1,2} \neq 0$, $\alpha_{2,1} = 0$). The contemporaneous association $\sigma_{1,2}$ is equal to 0 in both sets. As can be seen, in the left panel there does not appear a systematic dependence between changes in the two time series, while in the right panel, changes in Series 2 (upper trajectory) are followed by similar changes in Series 1 (lower trajectory).

Granger causality testing in the classical framework. Granger (1969) defined a concept of causality in the context of VAR models that built on the idea that a cause cannot come after the effect. Based on his work, predictive causality or Granger causality testing was developed into a statistical tool to infer whether one time series, $y_{1,t}$, can be used to forecast another, or more specifically if past and current values of time series $y_{1,t}$ contain additional information on future values of another time series $y_{2,t}$, above and beyond what is already contained in past and current $y_{2,t}$ alone. The idea behind predictive causality testing is that having nonzero cross-lagged effects (e.g., $\alpha_{1,2} \neq 0$ or $\alpha_{2,1} \neq 0$) decreases the error variation ($\sigma^2_{1,1,e}$ or $\sigma^2_{2,2,e}$), meaning that the prediction becomes more precise. Currently, significance testing using the Wald test is the default method for assessing predictive causality on the time domain, and the test tool is limited to a single subject design (see, e.g., Lütkepohl, 2005; Liu & Molenaar, 2016).

[1] https://osf.io/qr82d/?view_only=8bb143074543486fa231f122a62e4d4e

[2] We will only deal with the time-domain representation of the VAR model.

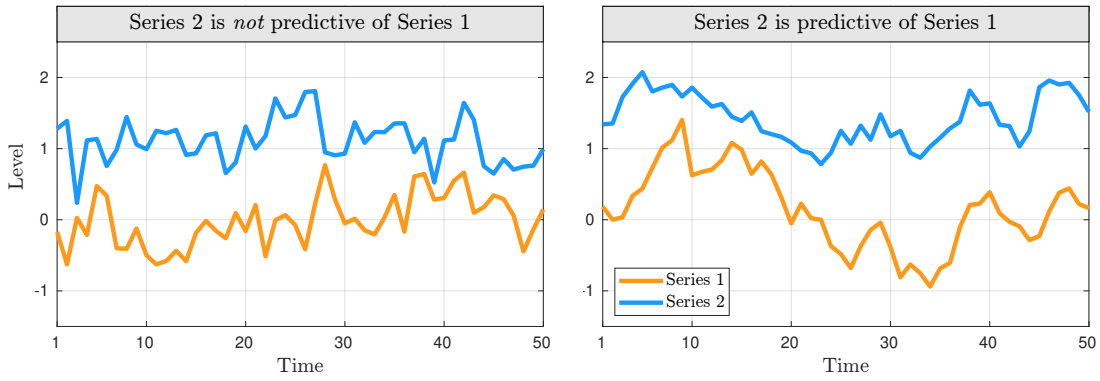

Fig. 1. Two sets of generated bivariate time series. The time series in left panel are independent, while the time series in the right panel have predictive dependence: Changes in Series 2 at time t are followed by similar changes in Series 1 at time $t + 1$.

Bayes factors for Granger causality

To recapitulate our goals, we will derive and apply Bayes factors for Granger causality testing in order to quantify relative evidence in favor or against the predictive association of one time series forecasting the other. In this section we provide a general introduction to the Bayes factor and derive its specific cases for single and multilevel VAR models.

Defining the Bayes factor. One way of comparing competing models in the Bayesian probabilistic inference framework is to use Bayes factors (Jeffreys, 1961). The Bayes factor derives immediately from Bayes' rule as follows: Let H_1 and H_2 be two competing accounts (i.e., hypotheses or models) for the data X, and let $p(X \mid H_1)$ and $p(X \mid H_2)$ be the likelihood of X under these accounts. By Bayes' rule, the posterior probability of H_1 (Hypothesis 1, which may be a "null" hypothesis but need not be) is then:

$$p(H_1 \mid X) = \frac{p(X \mid H_1)p(H_1)}{p(X)}. \quad [3]$$

As can be seen, the posterior probability of Hypothesis 1, $p(H_1 \mid X)$, is calculated by multiplying the likelihood of the data, $p(X \mid H_1)$, with the prior probability of Hypothesis 1, $p(H_1)$, and dividing this by the marginal likelihood, $p(X)$. We can replace H_1 with H_2 for the posterior probability of Hypothesis 2 (an alternative hypothesis or model).

To obtain the *relative* posterior probability of H_1 and competing account H_2, given the data, we need to formulate Equation 3 once for H_1 and again for H_2 and then divide each side of the equivalence. Conveniently, the marginal likelihood $p(X)$ then cancels out, resulting in:

$$\underbrace{\frac{p(H_1 \mid X)}{p(H_2 \mid X)}}_{\text{Posterior ratio}} = \underbrace{\frac{p(H_1)}{p(H_2)}}_{\text{Prior ratio}} \times \underbrace{\frac{p(X \mid H_1)}{p(X \mid H_2)}}_{\text{Bayes factor}}. \quad [4]$$

Here we have already re-grouped remaining factors into the prior ratio (relative prior probability of the accounts before seeing the data) and the Bayes factor (relative evidence in the data), which multiply to obtain the posterior ratio (relative probability of the accounts after seeing the data). Equation 4 can be restated again as follows:

$$\underbrace{\frac{p(X \mid H_1)}{p(X \mid H_2)}}_{\text{Bayes factor}} = \underbrace{\frac{p(H_1 \mid X)}{p(H_2 \mid X)}}_{\text{Posterior ratio}} \Big/ \underbrace{\frac{p(H_1)}{p(H_2)}}_{\text{Prior ratio}} \quad [5]$$

Table 1. Descriptive labels for certain Bayes factors.

Label	$B_{2:1}$	$p(H_2 \mid X)^*$
Data strongly support H_2	10	91%
Data weakly support H_2	3	75%
Data provide ambiguous information	1	50%
Data weakly support H_1	$1/3$	25%
Data strongly support H_1	$1/10$	9%

*: $p(H_2 \mid X)$ is the posterior probability of H_2 assuming prior equiprobability between H_1 and H_2. Adapted from Etz and Vandekerckhove (2016).

If we assume prior equiprobability between the hypotheses, $p(H_1) = p(H_2)$, the prior ratio cancels out and the Bayes factor equals the relative posterior probability (i.e., posterior probability ratio) of H_1 over H_2. More generally, the Bayes factor expresses the degree to which the data cause this probability ratio to shift. In our notation, we will indicate by subscripts which probability ratio is being shifted: $B_{1:2}$ will refer to the ratio of Hypothesis 1 over Hypothesis 2, while $B_{2:1}$ will refer to the inverse. Note that these two are just reciprocal transforms of each other, $B_{1:2} = 1/B_{2:1}$, so that we may choose either one to state our results – whichever is more convenient.

If the Bayes factor of H_1 over H_2 is large, meaning much greater than 1 (see Table 1 for indicative values), then the relative probability of H_1 over H_2 increases. If instead it is small, meaning less than 1 and closer to 0, then the relative probability of H_1 over H_2 decreases (or, equivalently, the probability of H_2 over H_1 increases). Since the Bayes factor expresses, in a single number, the degree to which a rational observer should change their belief in one hypothesis over another, we interpret the Bayes factor as the *amount of evidence provided by the data* (Evans, 2014).

The Savage-Dickey density ratio estimator for single-level VAR. We now derive the Savage-Dickey density ratio estimator (Dickey & Lientz, 1970; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010) for the predictive causality Bayes factor for a simple single-level VAR model. Generally speaking, the Savage-Dickey density ratio provides straightforward estimation of the Bayes factor for testing an equality constrained hypothesis against an unrestricted alternative. Let us first split the parameters of the VAR model, $\boldsymbol{\theta}$, into two subsets: $\boldsymbol{\theta} = (\delta, \epsilon)$. δ will denote the parameters of interest for the predictive causality test, while ϵ will denote other parameters of the model that are not currently of interest for testing. As a concrete example, consider the two-dimensional

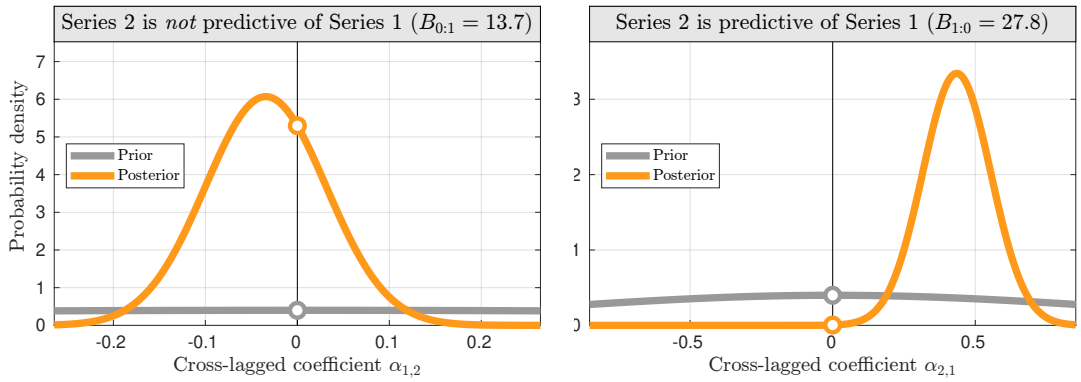

Fig. 2. Posterior and prior densities of $\alpha_{1,2}$ based on the data displayed in Figure 1. Circular markers indicate the heights of the densities at 0.

VAR with at most one lag ($L = 1$), as shown in Equation 2. When we test the predictive causality of time series 2 on time series 1, the parameter of interest is the cross-effect parameter, so we choose $\delta = \alpha_{1,2}$, while ϵ will contain the remaining parameters: $\epsilon = (\alpha_{1,1}, \alpha_{2,2}, \alpha_{2,1}, \boldsymbol{\nu}, \boldsymbol{\Sigma})$. The null hypothesis, or null model of no predictive causality, H_0, is then specified by $\alpha_{1,2} = 0$, and the alternative model/hypothesis, H_1, is $\alpha_{1,2} \neq 0$. The conditional density (denoted by $p_0(\cdot)$ to indicate conditioning on H_0 and $p_1(\cdot)$ to indicate conditioning on H_1) of $\alpha_{1,2}$ is continuous at 0, so that $\lim_{\alpha_{1,2} \to 0} p_1(\epsilon \mid \alpha_{1,2}) = p_0(\epsilon)$. Hence there is no difference between the priors of the other parameters in the null and the alternative model, that is $p_1(\epsilon \mid \alpha_{1,2} = 0) = p_0(\epsilon)$. Accordingly, the marginal likelihood under the null model of *no predictive causality* is

$$\begin{aligned} p_0(\mathbf{y}_t) &= \int p_1(\mathbf{y}_t \mid \epsilon, \alpha_{1,2} = 0) p_1(\epsilon \mid \alpha_{1,2} = 0) d\epsilon \\ &= p_1(\mathbf{y}_t \mid \alpha_{1,2} = 0). \end{aligned}$$

Now by applying Bayes' rule we get

$$p_0(\mathbf{y}_t) = \frac{p_1(\alpha_{1,2} = 0 \mid \mathbf{y}_t) p_1(\mathbf{y}_t)}{p_1(\alpha_{1,2} = 0)}. \quad [6]$$

To obtain the Bayes factor, we divide $p_0(\mathbf{y}_t)$ by $p_1(\mathbf{y}_t)$, as in Equation 4. This gives the Savage-Dickey density ratio—the ratio of the posterior and prior ordinates, evaluated at the test value—which is a simple estimator for the Bayes factor:

$$B_{(\alpha_{1,2})0:1} = \frac{p_1(\alpha_{1,2} = 0 \mid \mathbf{y}_t)}{p_1(\alpha_{1,2} = 0)}. \quad [7]$$

As can be seen, this is the ratio of the posterior for the parameter of interest under the alternative model evaluated at 0, divided by the prior of that parameter under the alternative model evaluated at 0. Inverting the right hand side of the expression in Equation 7 gives us $B_{(\alpha_{1,2})1:0}$, the Bayes factor *in favor of predictive causality*.

Example results from a single-level VAR model. Figure 2 shows a graphical representation of some of the results from fitting a VAR model for the two sets of bivariate time series depicted in Figure 1. In both panels of Figure 2, smoothed posterior densities of $\alpha_{1,2}$ are displayed in orange, and the prior on $\alpha_{1,2}$, which was a standard normal distribution in our example, is shown in grey. The circular markers indicate the heights of these densities at $\alpha_{1,2} = 0$, which we need for testing the hypotheses of $\alpha_{1,2} = 0$ or $\alpha_{1,2} \neq 0$.

Per Equation 7, to get the Bayes factor of no predictive causality we divide the height of the posterior density by the height of the prior density at 0. For the set of time series shown in the left panel of Figure 1, this gives 13.7, which indicates strong evidence for no predictive causality from Series 2 to Series 1. Evidence in favor of predictive causality can also be calculated for this pair of time series by dividing the height of the prior by the height of the posterior at 0, which gives 0.07 (the reciprocal of 13.7).

By contrast, in the right panel of Figure 2, there is strong evidence in favor of predictive causality of Series 2 for Series 1: there is approximately 27.8 times more support for predictive causality than for no predictive causality, based on the prior and posterior densities. This BF corresponds to the same time series data as in the right panel of Figure 1.

The multilevel VAR case. Bayesian vector autoregressive models are popular in economics (e.g., Litterman, 1986; Wozniak, 2016) and natural sciences (e.g., Lee, Chapman, Henderson, Chen, & Cane, 2016), where it is often possible to measure variables with high precision. Due to the different focus of these applications they do not typically incorporate multilevel designs. In social and behavioral sciences, however, we are often left to contend with fewer observations and noisy data. While this means less information about each individual of the analysis, in social and behavioral research it is common to measure multiple subjects who are jointly representative of some population. By relying on multilevel modeling techniques, we can then pool information across subjects and increase estimation accuracy for subjects with more noisy data (e.g., Baribault et al., 2018; Vandekerckhove, Verheyen, & Tuerlinckx, 2010). Moreover, this framework helps us explore group-level trends. For an overview, rationale and implementation of the two-dimensional multilevel VAR model with a social and behavioral science focus, see Li, Wood, Ji, Chow, and Oravecz (2022).

The multilevel VAR(L) model can be defined as follows:

$$\begin{aligned} \mathbf{y}_{p,t} &= \boldsymbol{\nu}_p + \mathbf{A}_{1,p}\mathbf{y}_{p,t-1} + \ldots + \mathbf{A}_{L,p}(\mathbf{y}_{p,t-L}) + \mathbf{e}_{p,t}, \quad [8] \\ \mathbf{e}_{p,t} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_p). \end{aligned}$$

The above equation represents the *within-person* model – that is, it describes the time-dynamics for each person p. Specifically, $\boldsymbol{\nu}_p$ is a $K \times 1$ vector of *person-specific* intercepts; $\mathbf{A}_{l,p}$ with $(l = 1, \ldots, L)$ is a $K \times K$ *person-specific* coefficients matrix of the lagged and cross-lagged effects at lag l; and $\mathbf{e}_{p,t}$ is a $K \times 1$ vector of random innovations following a multivariate normal distribution with a *person-specific* covariance matrix $\boldsymbol{\Sigma}_p$. Matrices $\mathbf{A}_{l,p}$ and $\boldsymbol{\Sigma}_p$ have the same structure as defined for the single-level case.

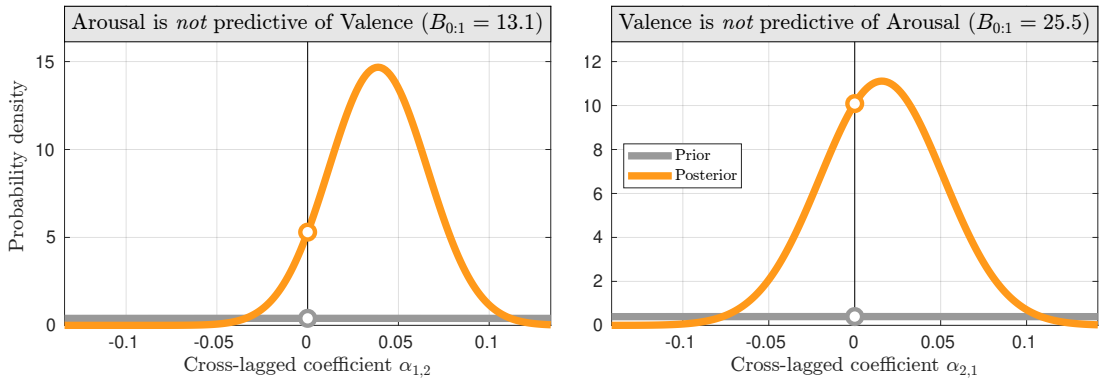

Fig. 3. Results on the group-level Bayes factor. Aggregating over participants, the Bayes factors support the null hypothesis of no Granger causality in both directions (Arousal-to-Valence nor Valence-to-Arousal).

Here we limit our focus to the stationary VAR(1) process, with stationarity defined for each person's time series as all the roots of the determinant of matrix $\mathbf{D} - \mathbf{A}_{1,p}\mathbf{y}_{p,t-1}$ having moduli greater than 1 (Lütkepohl, 2005), with $\mathbf{D}$ denoting the identity matrix.

All *within-level* parameters were person-specific and come from joint level-2 (i.e., group-level or population-level) or *between-person* distributions. Specifically, elements of the $\mathbf{A}$ matrix were also assigned normal distributions with population mean and population standard deviation estimated as $\alpha_{k_1,k_2,p} \sim N(\mu_{\alpha_{k_1,k_2}}, \sigma_{\alpha_{k_1,k_2}})$, for all subscripts k_1 and k_2, where these subscripts refer to the pair of dimensions connected by coefficient α. Furthermore, the person-specific intercepts were assumed to be normally distributed with group mean and standard deviation estimated from the data: $\nu_{k_1,p} \sim N(\mu_{\nu,k_1}, \sigma_{\nu,k_1})$.

The multilevel VAR model was cast in a Bayesian framework, where all model parameters are required to have prior probability distribution. For the person-specific parameters, the above defined level-2 (population) distribution function as priors. In case of the parameters of these level-2 distributions, the intercept parameters were assigned normal *hyperprior* distributions, with slightly wider range for the intercept's population mean $\mu_{\nu,k_1} \sim N(0, 10)$ than for the population mean of elements of the $\mathbf{A}$ matrix $\mu_{\alpha_{k_1,k_2}} \sim N(0, 1)$. The corresponding population standard deviations, σ_{ν,k_1} and $\sigma_{\alpha_{k_1,k_2}}$ had standard half-normal priors assigned to them (i.e., a standard normal distribution truncated to the positive real line to ensure that these parameters cannot take negative values). Finally, the $\mathbf{\Sigma}$ matrix was Cholesky decomposed, with priors set to default values suggested in the Stan manual (Stan Development Team, 2017). All these prior settings codify our a priori uncertainty regarding the exact values of each parameter.[3]

For the multilevel VAR case, we will test Granger causality both at the p-individual level and at the population level. For each participant in a study, we may compute the evidence for (or against) the hypotheses that $\alpha_{k_1,k_2,p} = 0$, and/or we may evaluate the population-level hypothesis that $\mu_{\alpha_{k_1,k_2}} = 0$.

Parameter estimation for the multilevel VAR. The Bayesian statistical inference framework offers flexible tools for implementing complex multilevel models, such as a multilevel extension of the VAR. Markov chain Monte Carlo methods (Robert & Casella, 2004) provide for efficient estimation of high-dimensional parameter spaces and the resulting posterior distributions of the model parameters can be used to make probabilistic statements about quantities of interest.

Li et al. (2022) implemented a one-step estimation for the two-dimensional multilevel VAR model in JAGS (Plummer, 2003), Stan (Carpenter et al., 2017) and Mplus (Muthén & Muthén, 1998-2017). Here we follow their steps and use a one-step estimation of a multi-dimensional VAR model implemented in Stan. The Stan software is a generic Bayesian inference engine and can be interfaced for example with R (R Core Team, 2016) or MATLAB (Baribault & Collins, 2021; Matzke, Boehm, & Vandekerckhove, 2018). These features together will lead to a tool that is easily adaptable for the needs of complex behavioral science applications. We have developed and tested Stan code to estimate a single-level VAR model with K dimensions and L lags with an R wrapper that includes the Bayes factor calculations. Alternatively, there has also been an R package developed (Epskamp, Deserno, & Bringmann, 2016) that can estimate multilevel VAR models; however the estimation approach implemented in this package relies on post-hoc estimation of the residual covariances and it is primarily non-Bayesian (but can call Mplus for Bayesian estimation).

Application

We demonstrate inference for Granger causality with Bayes factors using data from a 28-day long experience sampling study. Participants ($N = 52$) were asked to provide momentary self-reports on various aspects of their psychological states in their everyday life environments. All procedures were approved by the local Internal Review Board (protocol 00001017). Our analysis will focus on their reported *core affect* (CA; Russell, 2003). CA is a two-dimensional psychological construct that captures how *pleasant* and how *active/aroused* a person feels at any given moment. CA is assumed to fluctuate over short time scales time due to the influence of internal and external factors and its in the core of all our emotional experience (Barrett, 2016). These two dimensions are theorized to represent independent features of emotional experiences.

Under this theoretical framework, we would *not* expect the average valence experience one day to Granger cause the daily average arousal experience the next day, or vice versa (i.e., the theory predicts a *lack* of predictive causality). The existing classical inference framework would not allow us to quantify evidence in favor of such a theoretical position. By contrast, we will demonstrate that the Bayes factor allows us to summarize evidence favoring the

[3] Each of these is only one of many possible prior distributions for its parameter. These distributions capture the relative plausibility of different values before taking into account the data, and in our case involve some amount of theoretical commitment to the implications of these choices. Other prior distributions are possible, and might encode slightly different models and research questions (Etz et al., 2018).

absence of Granger causality (for more discussion on the differences between Bayes factors and null hypothesis testing, see, i.a., Dienes & McLatchie, 2018; Vandekerckhove et al., 2018; Wagenmakers et al., 2018). To recapitulate, we do not only test for the presence of an effect (Granger causality) here, but also directly for a null effect (absence of Granger causality). This goes beyond supporting a null hypothesis by "failing to reject" it, as is routinely done in the classical inference framework. Moreover, the magnitude of evidence in favor or against Granger causality will also be quantified.

Because our focus here is on introducing and illustrating Bayes factors for Granger causality, we simplify our analysis for ease of exposition. First, we only fit a bivariate VAR(1) model – that is, we restrict our focus to lag=1 effects. This model is reasonably for our data given that we work with daily aggregates of two dimensional core affect measures. Second, we removed data from three participants who did not have complete data on all 28 days, as well as from one participant whose data did not meet our criterion for stationarity in their time series.[4]

We fit the above specified multilevel VAR model to the data of the remaining participants ($N = 48$) in R and Stan using the package `rstan` (Stan Development Team, 2016). We called Stan from R and ran 6 chains in parallel with 1,000 warm-up and 10,000 iterations each, resulting in a final posterior sample size of 60,000 for each parameter. We found no issues with convergence (all $\hat{R}$ below 1.1; Gelman et al., 2013) and quality of sampling (effective sample size was more than 1000 for 90% of the parameters and at least 150 for each). For the Savage-Dickey approximation of the Bayes factor, we used the `polspline` package (Kooperberg, 2020) to kernel smooth the posterior distributions of relevant parameters (Wagenmakers et al., 2010).

Results. We start by looking at group-level results. We used the Bayes factor to infer whether (a) changes in arousal predict (i.e., Granger cause) changes in valence and whether (b) changes in valence predict (i.e., Granger cause) changes in valence. To obtain the Bayes factor for (a), we calculated the heights of the prior and posterior densities of $\mu_{\alpha,1,2}$ at 0. To obtain the Bayes factor for (b), we calculated the heights of the prior and posterior densities of $\mu_{\alpha,2,1}$ at 0. We found around 12 times more support for *no* Granger causality for arousal predicting valence than for Granger causing it, and 27 times more support for *no* Granger cause as opposed to Granger cause for valence predicting arousal. Results are displayed in Figure 3. They represent strong evidence on the group level for *no* predictive causality between the two core affect dimensions.

Next we look at the person-level results. We test the same propositions as above, but now for each person separately. For calculating the evidence for each person in terms of a Bayes factor for whether (a) changes in arousal Granger cause changes in valence, we calculated the heights of the prior and posterior densities of $\alpha_{1,2,p}$ at 0. For calculating the evidence for each person in terms of a Bayes factor for whether (b) changes in valence Granger cause changes in arousal, we calculated the heights of the prior and posterior densities of $\alpha_{2,1,p}$ at 0. The person-level results mirror the group level: almost all participants show evidence *against* Granger causality and none show evidence *for* Granger causality.

The person-level results are visualized in Figure 4. The markers are person-level point estimates of the cross-effect parameters

[4] We examined stationarity by first detrending each person's time series data via Hodrick-Prescott filtering using the `mFilter` package in R (Balcilar, 2019) and then running the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test from the `tseries` package (Trapletti & Hornik, 2022).

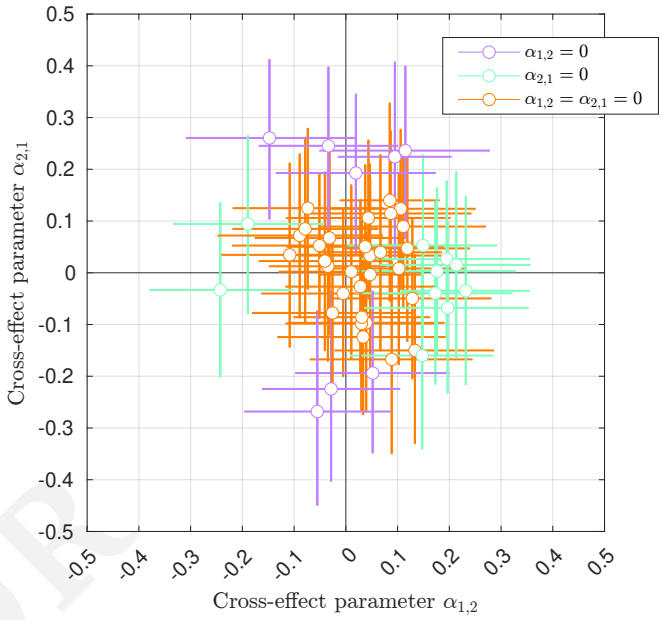

Fig. 4. Person-level estimates of the cross predictive parameters $\alpha_{1,2}$ and $\alpha_{2,1}$, color-coded based on their corresponding Bayes factor. Each marker represents a participant's α values. The lines in both dimensions indicate 3-support intervals, which are constructed such that the posterior density for values inside the interval is at least 3 times higher than the prior density. Orange marks evidence for no Granger causality for both directions, purple and green for one direction only, based on Bayes factors being larger than 3.

$\alpha_{1,2}$ and $\alpha_{2,1}$. The intervals in both dimensions are *3-support intervals* (Etz, Dablander, Gronau, & Wagenmakers, 2020): they contain those values of the cross-effect parameters whose probability density increased by at least a factor of 3 due to the data. In other words, if these intervals contain 0, then the null hypothesis for the cross-effect parameter is supported by a Bayes factor of at least 3. The intervals were color coded based on these intersections with 0, with the orange crosses indicating evidence *against* Granger causality in both directions. Out of 48 participants, 30 showed substantial evidence for the *absence* of causality in both directions (orange, $B_{(\alpha_{1,2,p})0:1}$ and $B_{(\alpha_{2,1,p})0:1}$ both larger than 3), and all showed substantial evidence for the *absence* of causality in at least one direction (purple if $B_{(\alpha_{1,2,p})0:1} \geq 3$ and green if $B_{(\alpha_{2,1,p})0:1} \geq 3$). None showed substantial evidence in support of predictive causality in either direction (all $B_{(\alpha_{1,2,p})1:0} \leq 3$ and $B_{(\alpha_{2,1,p})1:0} \leq 3$).

Discussion

We have introduced a novel inference tool, a Bayes factor for Granger causality testing, which can simultaneously evaluate evidence in favor and against Granger causality in multivariate time-series data. Moreover, the ratio of evidence for these two competing hypotheses can be quantified on a continuum. This means that for example we can state how much more evidence we have for a null hypothesis of no Granger causality in our data than for an alternative hypothesis of Granger causality (or vice versa). We have illustrated how useful such flexibility in inference can be with an example where the underlying theory suggested no predictive causality between affect qualities on a day-to-day timescale.

We derived the Bayes factor to perform inference on both the

group and the individual level. This is especially important for applications in the social and behavioral sciences, where aggregation of data over participants or groups is often critical in order to amass enough evidence. However, we note that not all the Bayes Factors calculated in this analysis are independent of one another. This makes it improper to combine Bayes factors for compound statements (e.g., $\alpha_{1,2} > 0$ **and** $\alpha_{1,2,p} > 0$). For such statements a new model comparison would need to be set up to test the compound proposition directly. We also briefly demonstrated the use of B-support intervals, which allow for a visual test of many null hypotheses at once.

Our work presents an introduction of Bayes factors to inference on Granger causality in social science. This work represents an initial step towards developing a rigorous novel way of testing Granger causality, with the aim of providing a proof-of-concept illustration. Below, we detail certain simplifications and assumptions that underlie our approach. Future research could evaluate the tenability of these simplifications and assumptions.

We made didactic simplifications in our modeling approach, including only discussing the lag=1 case, and not modeling missing data in our time series. The latter extension can easily be made by consulting Li et al. (2022). The former could be a straightforward extension in a future research project.

We would like to emphasize that this initial implementation of the Bayes Factor test for Granger causality is based on assumptions that may limit the generalizability of our results. First, all inference is conditional on the presented AR model specification and the corresponding priors on the parameters, and on our choices of measuring the modeled variables, in our case valence and arousal. For example, we used a single variable measure of valence, but the dynamics might actually be different on its positive end of the scale versus the negative end of the scale. Second, we did not examine whether the lead-lag relationships between valence and arousal change over time, but assumed that it would not fluctuate. If a lead-lag relationship oscillates over time, this could lead to estimates of zero cross-effect when inappropriately aggregated over time. Third, our conclusions are limited to the timescale we chose, in this case day-to-day, and are based on the assumption that this is a sufficient rate to discover cross-coupled dynamics.

Finally, we note that the classical significance testing framework for predictive causality provides a static testing environment: it is typically performed only after the data collection is concluded (or else multiple testing corrections need to be worked out). In the Bayesian framework sequential updating of evidence is the standard mode of operation and is straightforward (Oravecz, Huentelman, & Vandekerckhove, 2017). The Bayes factor approach we introduced presents a tool for inference involving sequential updating of evidence, for example when time series data is being acquired in real time. With continuous streaming of information becoming increasingly accessible (e.g., via passive sensing with wearables, see e.g., Brick, Mundie, Weaver, Fraleight, & Oravecz, 2020), our new Bayesian approach has great potential for prevention science.

References

Balcilar, M. (2019). mFilter: Miscellaneous time series filters [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=mFilter (R package version 0.1-5)

Baribault, B., & Collins, A. (2021, Dec). *Troubleshooting Bayesian cognitive models: A tutorial with matstanlib.* PsyArXiv. Retrieved from psyarxiv.com/rtgew doi: 10.31234/osf.io/rtgew

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., ... Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*, 2607-2612. doi: 10.1073/pnas.1708285114

Barrett, L. F. (2016). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1-23. doi: 10.1093/scan/nsw154

Brick, T. R., Mundie, J., Weaver, J., Fraleight, R., & Oravecz, Z. (2020). Low-burden mobile monitoring, intervention, and real-time analysis using the wear-IT framework: Example and usability study. *JMIR Formative Research*, *4*(6), e16072. doi: 10.2196/16072

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1-32.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214-226. Retrieved from http://www.jstor.org/stable/2239734

Dienes, Z. (2016). How bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*, 207-218.

Droumaguet, M., Warne, A., & Wozniak, T. (2016). Granger causality and regime inference in Markov switching VAR models with Bayesian methods. *Journal of Applied Econometrics*, *32*, 802-818. doi: 10.1002/jae.2531

Epskamp, S., Deserno, M. K., & Bringmann, L. F. (2016). mlvar: Multi-level vector autoregression. *R package version 0.3.3.*. Retrieved from https://rdrr.io/cran/mlVAR/

Etz, A., Dablander, F., Gronau, Q. F., & Wagenmakers, E. (2020). The support interval. *Erkenntnis*, *87*(2), 589–601. doi: 10.1007/s10670-019-00209-z

Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, *1*(2), 281–295.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS ONE*, *11*, e0149794. doi: 10.1371/journal.pone.0149794

Evans, M. (2014). Discussion of "on the Birnbaum argument for the strong likelihood principle". *Statistical Science*, *29*(2), 242 – 246. doi: 10.1214/14-STS471

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, Third edition*. Boca Raton (FL): Chapman & Hall/CRC.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424-438. Retrieved from http://www.jstor.org/stable/1912791

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Kooperberg, C. (2020). polspline: Polynomial spline routines [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=polspline (R package version 1.1.19)

Lee, D. E., Chapman, D., Henderson, N., Chen, C., & Cane, M. A. (2016). Multilevel vector autoregressive prediction of sea surface temperature in the north tropical atlantic ocean and the Caribbean sea. *Climate Dynamics*, *47*(1), 95–106. doi: 10.1007/s00382-015-2825-5

Li, Y., Wood, J., Ji, L., Chow, S.-M., & Oravecz, Z. (2022). Fitting multilevel vector autoregressive models in Stan, JAGS, and Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 452–475.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics*, *4*(1), 25-38. Retrieved from http://www.jstor.org/stable/1391384

Liu, S., & Molenaar, P. C. M. (2016). Testing for Granger causality in the frequency domain: A phase resampling method. *Multivariate Behavioral Research*, *51*(1), 53-66. (PMID: 26881957) doi: 10.1080/00273171.2015.1100528

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin: Springer-Verlag Berlin Heidelberg.

Matzke, D., Boehm, U., & Vandekerckhove, J. (2018). Bayesian inference for psychology, Part III: Bayesian parameter estimation in nonstandard models. *Psychonomic Bulletin & Review*, *25*, 77–101. doi: 10.3758/s13423-017-1394-5

Molenaar, P. C. M. (2019). Granger causality testing with intensive longitudinal data. *Prevention science: the official journal of the Society for Prevention Research*, *20*(3), 442-451. doi: 10.1007/s11121-018-0919-0

Molenaar, P. C. M., & Lo, L. (2016). Alternative forms of Granger causality, heterogeneity, and nonstationarity: Methods for applied empirical research. In W. Wiedermann & A. von Eye (Eds.), *Statistics and causality: Methods for applied empirical research* (p. 203-229). Wiley.

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1–5.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA.

Oravecz, Z., Huentelman, M., & Vandekerckhove, J. (2017). Sequential Bayesian updating for big data. In M. N. Jones (Ed.), *Big data in cognitive science* (p. 13-33). Sussex, UK: Psychology Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 1-10).

Sen, R., Majumdar, A., & Sikaria, S. (2022). Bayesian testing of granger causality in functional time series. *Journal of Quantitative Economics*, *66*, XX–XX. doi: https://doi.org/10.1007/s40953-022-00306-x

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods*. New York: Springer.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, Part IV: Parameter estimation and Bayes factors. *Psychonomic bulletin & review*, *25*(1), 102–113.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145–172.

Stan Development Team. (2016). *Stan: the R interface to Stan*. Retrieved from http://mc-stan.org/ (R package version 2.14.1)

Stan Development Team. (2017). *Stan modeling language users guide and reference manual, version 2.16.0.* Retrieved from http://mc-stan.org/

Trapletti, A., & Hornik, K. (2022). *tseries: Time series analysis and computational finance.* Retrieved from https://CRAN.R-project.org/package=tseries (R package version 0.10-51.)

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). *Bayesian methods for advancing psychological science* (Vol. 25) (No. 1). Springer.

Vandekerckhove, J., Verheyen, S., & Tuerlinckx, F. (2010). A crossed random effects diffusion model for speeded semantic categorization data. *Acta Psychologica*, *133*, 269–282. doi: 10.1016/j.actpsy.2009.10.009

Velicer, W. F., & Molenaar, P. C. M. (2013). Time series analysis for psychological research. In *Handbook of psychology. 2 : Research methods in psychology* (2nd ed., p. 628-660).

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158 - 189. doi: http://dx.doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology, Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35-57. doi: 10.3758/s13423-017-1343-3

Wozniak, T. (2016). Bayesian vector autoregressions. *Australian Economic Review*, *49*(3), 365-380. doi: 10.1111/1467-8462.12179