UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Non-Stationarity in Time-Series Analysis: Modeling Stochastic and Deterministic Trends

Ryan, Oisin; Haslbeck, Jonas M.B.; Waldorp, Lourens J.

DOI
10.1080/00273171.2024.2436413

Publication date
2025

Document Version
Final published version

Published in
Multivariate Behavioral Research

License
CC BY

Link to publication

Citation for published version (APA):
Ryan, O., Haslbeck, J. M. B., & Waldorp, L. J. (2025). Non-Stationarity in Time-Series Analysis: Modeling Stochastic and Deterministic Trends. *Multivariate Behavioral Research*, *60*(3), 556-588. https://doi.org/10.1080/00273171.2024.2436413

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Nov 2026

MULTIVARIATE BEHAVIORAL RESEARCH
2025, VOL. 60, NO. 3, 556–588
https://doi.org/10.1080/00273171.2024.2436413

Routledge
Taylor & Francis Group

TUTORIAL

OPEN ACCESS Check for updates

Non-Stationarity in Time-Series Analysis: Modeling Stochastic and Deterministic Trends

Oisín Ryan[a]*, Jonas M. B. Haslbeck[b,c]*, and Lourens J. Waldorp[c]

[a]Department of Data Science and Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands; [b]Department of Clinical Psychological Science, Maastricht University, Maastricht, Netherlands; [c]Psychological Methods Group, University of Amsterdam, Amsterdam, Netherlands

ABSTRACT

Time series analysis is increasingly popular across scientific domains. A key concept in time series analysis is stationarity, the stability of statistical properties of a time series. Understanding stationarity is crucial to addressing frequent issues in time series analysis such as the consequences of failing to model non-stationarity, how to determine the mechanisms generating non-stationarity, and consequently how to model those mechanisms (i.e., by differencing or detrending). However, many empirical researchers have a limited understanding of stationarity, which can lead to the use of incorrect research practices and misleading substantive conclusions. In this paper, we address this problem by answering these questions in an accessible way. To this end, we study how researchers can use detrending and differencing to model trends in time series analysis. We show *via* simulation the consequences of modeling trends inappropriately, and evaluate the performance of one popular approach to distinguish different trend types in empirical data. We present these results in an accessible way, providing an extensive introduction to key concepts in time series analysis, illustrated throughout with simple examples. Finally, we discuss a number of take-home messages and extensions to standard approaches, which directly address more complex time-series analysis problems encountered by empirical researchers.

KEYWORDS

Time series; stationarity; non-stationarity; time-varying models; trends; detrending; differencing

1. Introduction

Time series analysis is becoming increasingly common across a variety of scientific disciplines, largely due to technological advances that make high-frequency data collection more feasible in many domains. For example, the environmental sciences are using an ever-growing arsenal of sensors to track variables related to the global climate (Mudelsee, 2010), pollution (Law et al., 2010), or the ocean currents (O'Brien et al., 2017); in medicine, time series of various health related variables are increasingly available and used to predict negative health outcomes (Kennedy & Turley, 2011) or the effects of interventions (Schaffer et al., 2021); and in psychological research time series of sensor-based and self-reported measures are analyzed to investigate the dynamic functioning of individuals (Kuppens et al., 2022), how this functioning is related to mental health outcomes (van de Leemput et al., 2014), and how to design interventions (Olthof et al., 2020). In behavioral research in particular time-series data can be produced by a wide variety of different sources and be used to study vastly different phenomena, including for instance: self-report measures of affect and mood (e.g., Hamaker et al., 2016; Kuppens, 2015) or physical and psychomatic sensations (e.g., Ariens et al., 2020; Stone & Broderick, 2007) collected through ecological momentary assessment surveys; heart-rate variability, movement and/or location tracking measured passively by smartphones and wearable sensors (Nakamura et al., 2016); user behavior tracking in online learning environments (e.g., Henrie et al., 2015) or google search trends about jobs opportunities or COVID-19 symptoms (e.g., Brodeur et al., 2021; Jebb et al., 2015). Regardless of the source of the data or the phenomenon under investigation, this data is often analyzed using tools and models from the field of *time-series analysis* (Hamilton, 1994).

A key concept in time series analysis is *stationarity*, both because it gives us a language to characterize

CONTACT Oisín Ryan o.ryan@umcutrecht.nl Department of Data Science and Biostatistics, University Medical Center Utrecht, Utrecht, Netherlands.
*The first two authors contributed equally to this work and are considered joint first authors.

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

different processes that we can study with time-series data, and because it plays an important role in determining when we can perform the types of statistical inference researchers are used to in cross-sectional settings, such as estimating population parameters from samples and performing hypothesis tests. Understanding the concept of stationarity is necessary to answer basic questions in applied time series modeling such as: When should I detrend my data? Which problems can be caused by ignoring non-stationarity? How can I separate effects at different time scales? What mechanisms can generate changes in a process's distribution over time? How can I determine which mechanisms are at play in a given time series? When should we difference data instead of detrending? And which ways are there to *model* non-stationarity?

In this paper, we provide a gentle introduction to the topic of stationarity and then provide answers to the above questions. We take an approach that is geared toward researchers without a strong mathematical background for whom standard treatments from the econometics literature (e.g., Brockwell & Davis, 2009; Hamilton, 1994; Pesaran, 2015) are difficult to read. While a full treatment of all methods which concern non-stationarity is beyond the scope of a single paper, we hope to provide a useful overview for researchers by adopting the following strategy: In Section 2, we begin with providing an introduction to stationarity and non-stationarity that provides a definition and builds intuition using simulated example time series. Once key concepts are clarified, we discuss how (non-) stationarity can be used to characterize systems and why the property of stationarity is so useful for statistical analyses, providing us with guarantees that the types of statistical inferences readers may be familiar with in cross-sectional settings can also be performed in the time-series setting. In Section 3, we concretely discuss how to model different types of non-stationarity for the simplest case of mean-stationarity. While researchers likely encounter more complicated forms of non-stationarity in their data, we think that understanding the rationale for modeling the simplest case is necessary for being able to navigate more complex scenarios. Finally, in Section 4 we discuss key take-aways for researchers that go beyond the simple models discussed throughout the paper.

2. Stationarity, science and statistics

In this section, we establish what stationarity is and discuss why it is such an important and commonly discussed concept in time-series analysis. The terms *stationary* and *non-stationary* are labels which we use to describe different types of processes, depending on whether their distributional characteristics remain stable or change over time. The concept of stationarity is important in time-series analysis for a number of reasons, and is frequently presented as an *assumption* of time-series models, such as the popular auto-regressive and moving average models (Hamilton, 1994, more about these models later). We will begin by introducing readers to the different types of (non-)stationarity which are typically discussed in the time-series literature (Section 2.1). Our treatment will include some technical terms and notation that the reader may not be familiar with, but throughout we will rely on illustrations based on simulated time-series to build intuition for what these terms mean. Next, we give an overview of how stationary and non-stationary time-series may fit into the bigger picture of empirical research and how we think about real-world processes (Section 2.2). Finally, we discuss what it means for stationarity to be an assumption of a time-series model, by showing how stationary processes allow us to apply many of the tools of statistical estimation and inference readers may be familiar with from cross-sectional data analysis (Section 2.3).

2.1. What is stationarity?

Stationarity is a characteristic of a stochastic process X. To unpack this statement, we first define some basic concepts about stochastic processes and relate them to distributions (or random variables), which the reader is likely more familiar with. A stochastic process is a collection or set of distributions for random variables X_t which are ordered by the index time t. For instance, we can consider X_t to represent the number of reported covid cases in a city on a particular day; the self-reported mood of a participant at a given time during the day, or the GDP of a country in a specific year. In panel (a) of Figure 1 we sampled seven different versions of the same random variables X_t for $t \in \{1, ..., 20\}$ with time series length $n_t = 20$, each one in a different color. We could use the seven observations at $t = 1$ to construct an empirical distribution of X_1. If we took an infinite amount of samples at $t = 1$, we would obtain the population distribution at X_1. This would allow us to calculate the population properties of these distributions such as the mean or the variance like for any other distribution. Instead of inspecting the distribution of X_t at a fixed t, we can also inspect individual "runs" of the

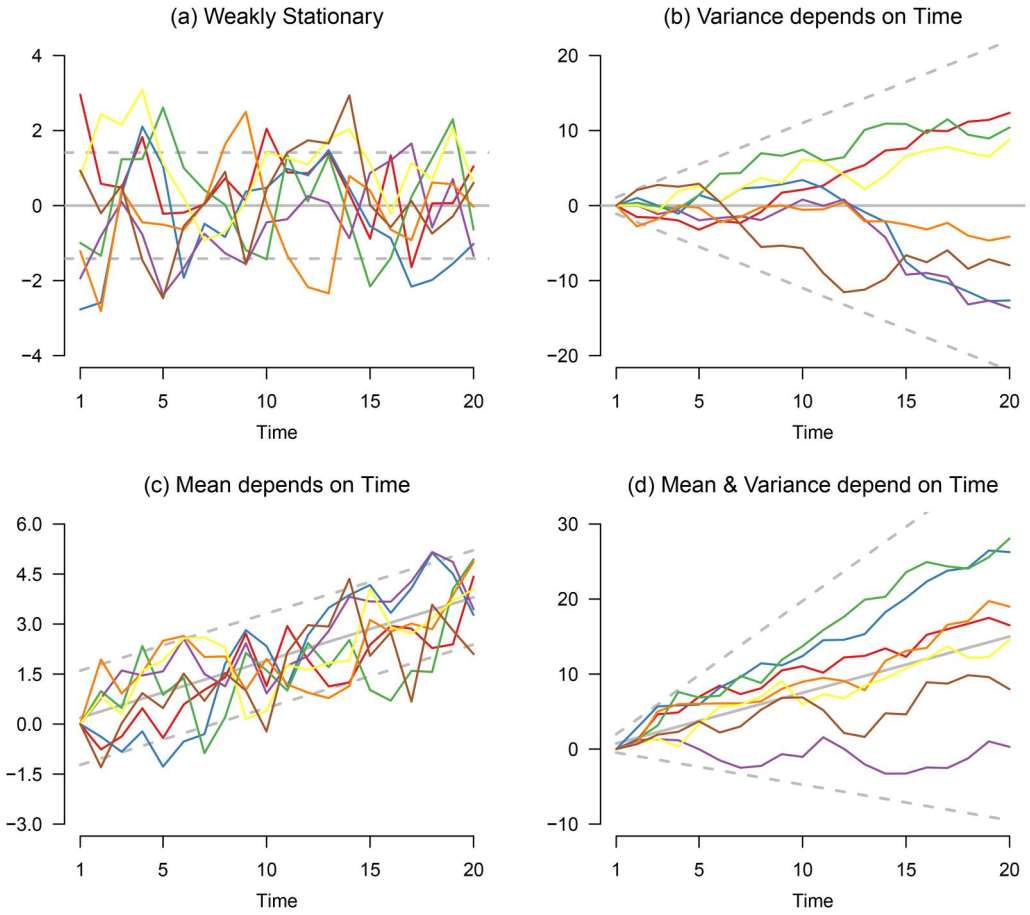

Figure 1. Examples of time series that (a) have constant mean and variance and are therefore weakly stationary, (b) have a variance that depends on time, (c) have a mean that depends on time, and (d) have both a variance and a mean that depends on time. The last three time series are therefore not weakly stationary.

stochastic process X, which are indicated by different colors. Each of these are called *realizations* of the stochastic process, in other words, an observed time-series. Infinitely many realizations, or equivalently, the set of all distributions of X_t, give rise to the *ensemble* implied by the stochastic process X.

Whether a stochastic process is stationary or not depends on whether certain properties of the distributions at X_t are the same at each time point t or not. Two of those properties are the *mean* and the *variance* of X_t. However, unique to the time-series case, we can also define how X_t co-varies with itself at a later point in time, X_{t+h}, where h is referred to as the *lag*. This property, represented by the auto-covariance function $\gamma_X(h) = \mathbb{E}(X_{t+h} - \mu)(X_t - \mu)$, is called the *auto-covariance*, where $\mathbb{E}$ is the expected value with respect to X, and μ is the mean of the stochastic process X. The term stationarity is a label we give stochastic processes X if they satisfy one or several of these properties. A time-series is *mean-* or *first-order* stationary if the mean of X_t is constant across time, that is $\mathbb{E}(X_t) = \mu$ for all t. If in addition the auto-covariance function of the time series is constant across time, we say the time series is *second-order* stationary. If in addition $\gamma_X(h)$ is finite, the time series is *weakly stationary*. *Strict stationarity* is the strongest form of stationarity, which applies when all other statistical properties of a time series are also invariant over time, for example its higher-order moments such as *skewness* and *kurtosis*.

Let's get a better feeling for different types of (non-) stationary processes by looking at simulated time series. There are many ways to generate stationary and non-stationary time series. Here we use the simple first-order Autoregressive (AR(1)) model since we suspect that it is the model the reader will be most familiar with:

$$X_t = \alpha + \phi X_{t-1} + \beta t + \varepsilon_t \tag{1}$$

where α is an intercept or drift parameter, ϕ is the autoregressive parameter, indicating how much X_t depends on X_{t-1}, β determines the slope of a possible linear trend across time, and ε_t represents innovation variance in the form of Gaussian noise $\mathcal{N}(0, \sigma_\varepsilon^2)$. The

AR(1) model is the simplest member of a large class of time-series models known as Autoregressive Integrated Moving Average (ARIMA) models (Hamilton, 1994), a model class we will return to later in the paper.

We use different choices of α, ϕ and β to generate time series that are stationary or violate weak stationarity in different ways. Since the properties defining stationarity are with respect to distributions at X_t, visualizing possible dependencies of means on time requires inspecting several realizations of the process X. We first consider a model with parameters $\alpha = 0, \phi = 0.5, \beta = 0$. This time series is shown in panel (a) of Figure 1. In this version of the AR(1) model neither the mean nor the variance are functions of time and consequently, the time series is weakly stationary. Accordingly, eyeballing the seven realizations of this process suggests that the mean and variance of the time series remain constant across time (for details and proof that this is the case, see Appendix A1). Panel (b) shows a time series generated from an AR(1) model with parameters $\alpha = 0, \phi = 1, \beta = 0$. This type of AR(1) model is also called a *random walk*, which is known to have a constant mean, but a variance that depends on time t (see e.g., Hamilton, 1994). Consequently, this process is *not* weakly stationary. The increasing variance over time can be seen from the multiple realizations or time-series drawn from that process. While the mean of the time-series remains constant at zero, we see that the variance increases with time: the realizations of the process are close together early in the time series, but become more dispersed across time. Specifically, the variance of a random walk is $t\sigma^2$ (see Appendix A1). In panel (c) we see a process generated with $\alpha = 0, \phi = 0.5, \beta = 0.1$ which has a constant variance across time, but its mean increases over time, which renders it not weakly stationary. Specifically, the mean increases with a slope determined by both β and φ (see Appendix B1 for details). Finally, in panel (d) we see a process we generated with $\alpha = 0.75, \phi = 1, \beta = 0$ in which both the variance and the mean depend on time, which is therefore again not weakly stationary. Understanding why these different parameter settings create time series with these different properties requires the concept of stochastic and deterministic trends, which we will discuss in Section 3.

2.2. Stationarity and Science

We have seen that stationarity is a label we give to processes for which all or some statistical properties remain the same for all time. However, this does not immediately make clear how or why understanding processes in these terms is useful for empirical research. To make this connection, we discuss where in the real world we could find stationary processes and then consider how stationarity can be used in empirical research.

The statistical properties that come with the label stationary have substantive implications for real-world processes. For example, if we know that a process is strictly stationary, we know that the system is not statistically (i.e., in terms of the mean, variance, etc.) changing in any structural way over time. It could be argued that no such system can exist in the real world: the universe expands, mountains erode, and humans develop, adapt and eventually die. So how can stationarity be useful to empirical research? There are at least three answers. First, stationarity may hold for *a part* of a system. For example, one's mood might be terrible in the morning and gradually improve throughout the day. This overly simplistic mood process might be modeled by a cyclic process plus some additional variation, where the additional variation itself is stationary. If true, this would give us the important information that all the systematic variation in mood is captured by the within-day cycle. Second, systems can be stationary *with respect to a given time scale*. While one's mood is clearly changing throughout the day, when measuring at a time scale of milliseconds, the changes are probably negligible. Similarly, while the number of cases of a disease may not be stationary on a timescale of years, it may be stationary on a day-to-day basis for some time window. Finally, stochastic processes are models, which are useful because they abstract away and simplify. We therefore never expect that stationarity strictly holds in the real world, but only require that stationarity is *a useful approximation* and that the conclusions derived from it are valid.

In most situations, all three qualifications are likely to be operating. For example, one could define a process for a person's body height. Clearly, there are systematic increases at least in the first 20 or so years, so we could define the time scale as adulthood. However, even when fully-grown, our body height varies by about 1% from morning until evening (Reilly et al., 1984). We could address this by either separately modeling this within-day variation such that the remaining variation is indeed stationary; or we decide that this type of model-misfit is acceptable in the given context.

One way to use (non-)stationarity in empirical research is to treat it as an empirical phenomenon (Woodward, 1989) that describes parts of systems in the real world. The non-stationarity is captured by models (e.g., for a trend), similarly to other phenomena such as mean differences, variances and correlations are captured by models. For example, if a stationary time-series approximates a time-series of mood at the time scale of minutes well, then we can consider the stationarity of mood at this time scale as an empirical phenomenon. On the other hand, a non-stationary model including a single cycle may explain one's mood at the time scale of a single day. Such empirical phenomena should be explained by theories and therefore can be used to test and develop theories (Haig, 2008; Haslbeck, Ryan, Robinaugh, et al., 2022).

The research context in which we expect that most readers encounter the topic of stationarity, however, is when performing statistical inference on time-series data. Whether a time-series model is used for prediction, description, or to aid in understanding the causal mechanisms underlying the data, as is the case with cross-sectional data, making inferences based on time-series data requires a number of assumptions to guarantee that the inference works appropriately. Statistical inference in time-series and cross-sectional settings are not fundamentally different from one another; in both cases we require certain statistical assumptions to guarantee that certain statistical inferences are valid. For example, when analyzing cross-sectional data and knowing that our data consists of independent samples from a fixed distribution, then, by the Law of Large Numbers (LLN), the sample mean will converge on the population mean. In the time-series setting, observations are typically not independent, and we therefore need different assumptions to ensure that inference works. As we will explain in the next section, the assumption of stationarity puts constraints on the dependence between variables, which guarantees that theorems the reader might be familiar with from cross-sectional data analysis, such as the Law of Large Numbers mentioned above, and the Central Limit Theorem on which we base confidence intervals and statistical tests, also hold in the time-series setting.

2.3. Stationarity and statistics

In order to estimate meaningful parameters from sample data, as well as perform inference on those parameters, certain statistical theorems need to hold. Different sets of assumptions are sufficient for those theorems to hold, but weak stationarity is perhaps the most frequently used assumption. In this section, we explain how weak stationarity guarantees these theorems hold and explain how they are important for the statistical analysis of time-series data.

2.3.1. Law of large numbers

Often we are interested in a general property of some process. The LLN states that if we take more and more independent samples of the same distribution, then we get better and better estimates of the property we are interested in. Taking the mean as an example, we can estimate the population mean μ of a random variable by taking the sample average $\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$. The LLN tells us that $\bar{X}$ converges to μ with the sample size $n \to \infty$. In a cross-sectional setting n could be the number of persons sampled from the same population, but in the time-series setting the sample size is equal to the time-series length n_t. One way to ensure that the LLN holds in a time-series setting is to assume weak stationarity. We illustrate this by discussing how the LLN fails when the properties making up weak stationarity do not hold.

Recall that weak stationarity refers to the situation in which the mean, variance, and auto-covariance function of a process are constant over time and finite. The consequence of a mean μ_t that is changing across time t for the LLN is immediate: we cannot hope for the sample mean to converge on a stable mean μ when increasing the number of samples n, because no such stable mean exists. The value we obtain by taking the average of the sampled time-series therefore does not refer to a stable property of the system if mean-stationarity does not hold. The sample mean may still contain information, but because the LLN logically cannot hold in this situation, we are not estimating a stable property of the system under investigation: For example, if we collected data over a larger or smaller time interval, or over an earlier or later time-window, we would by necessity obtain a different sample mean.

A perhaps less obvious requirement for the LLN to work is that the variance of the *estimate* $\text{var}(\bar{X})$ goes to zero with increasing observations n. Intuitively, this means that if we have enough observations, the parameter estimate is arbitrarily close to the population parameter. This variance has two components, and both are guaranteed to converge to zero with increasing observations n, when stationarity holds. The first component is a term including the variance of the process divided by n. Weak stationarity implies finite variances of the process ($\text{var}(X) < \infty$), and therefore

we know that the first component converges. The second component is the covariance between different observations, which cannot be too large. This requirement becomes intuitive when considering the extreme case in which the correlation between all observations is equal to 1. Here, we effectively have only a single observation and there is no reason to believe that an estimate based on this single observation is equal to the population parameter. In the cross-sectional setting this covariance is guaranteed to be zero by assuming independent samples. Clearly, this is not the case in the time-series setting; if, for example, observations are generated by the AR(1) model we introduced above, then observations will display *autocorrelation*, that is, consecutive observations will correlate with each other. In the time-series case we require that the dependence between observations decays quickly enough as we increase the time between them. This condition is guaranteed by weak stationarity, which ensures finite variances σ^2 and auto-covariance functions that do not depend on t. For instance, in the AR(1) examples we shown above, the auto-covariance is determined by the ϕ parameter, and when it is less than one in absolute value, then observations become less and less correlated the further apart in time they are. In Appendix A2 we provide a detailed discussion of how weak stationarity can imply quickly enough decaying autocovariances. Note that the requirement of decaying autocovariance between time points is model independent, but that imposing (or assuming) a model may lead to quickly enough decaying autocorrelations.

Figure 2 displays the consequences of violating weak stationarity for estimating the mean of a time-series. We re-use the models from Figure 1 as examples. In panel (a) the time-series is stationary and the mean converges to the true value zero when increasing the time-series length. However, for the random walk in (b) we see that the mean does not converge because the variance of the process becomes infinite as $n \to \infty$. In panel (c) the mean increases linearly with time and therefore the mean cannot converge.

2.3.2. Central limit theorem

While the LLN provides us with guarantees that we can use sampled information to obtain *point estimates*, the Central Limit Theorem (CLT) is needed to guarantee that we can *quantify our uncertainty* about those estimates. The CLT states that when summing independent samples X_j from the same distribution, then the distribution of the (properly scaled) sum approaches a normal distribution $\mathcal{N}(0, \sigma^2)$, irrespective of the distribution X_j are sampled from. The CLT is important for statistical inference, because if the CLT holds then we can compute sampling distributions of parameter estimates *via* the normal distribution; this in turn allows us to compute standard errors, confidence intervals, and to perform hypothesis tests. Since the distribution of the parameters approaches a normal distribution, which has a parameter for the mean, the CLT implies the LLN.

Similar to the LLN, the CLT also requires the existence of a true population parameter that does not change across time. In addition, and also similar to the LLN, the CLT requires that the correlations between time points decreases as the number of time points between them h (i.e., the lag) increases. On a technical level, these two theorems differ in their requirements mainly in the sense that the CLT

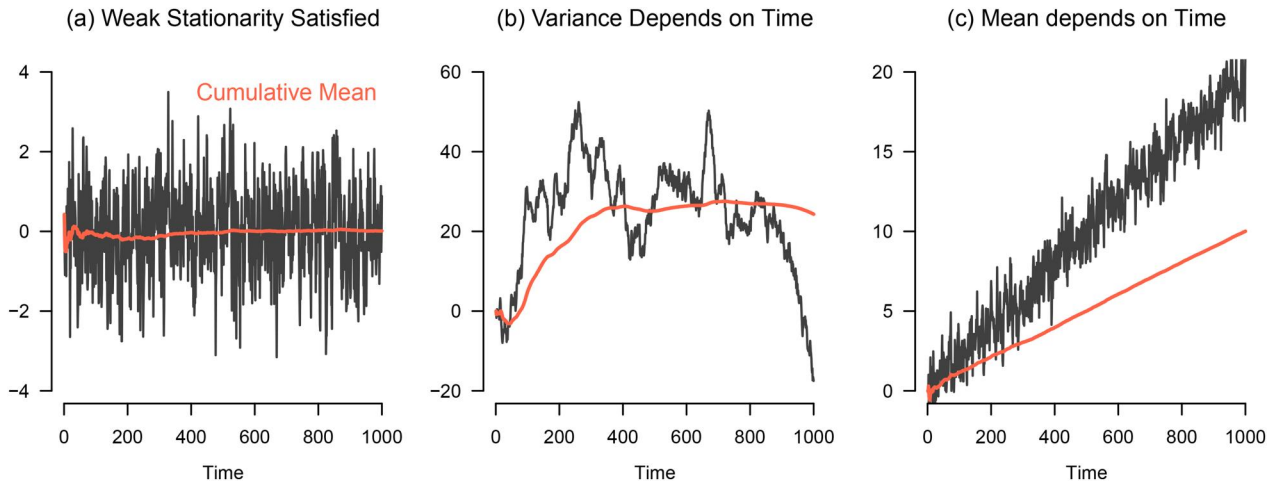

Figure 2. Realizations of time-series with (a) weak stationarity, (b) variance depending on time and (c) mean depending on time (black lines). We see that the cumulative mean (orange lines) only converges for the weakly stationary time-series.

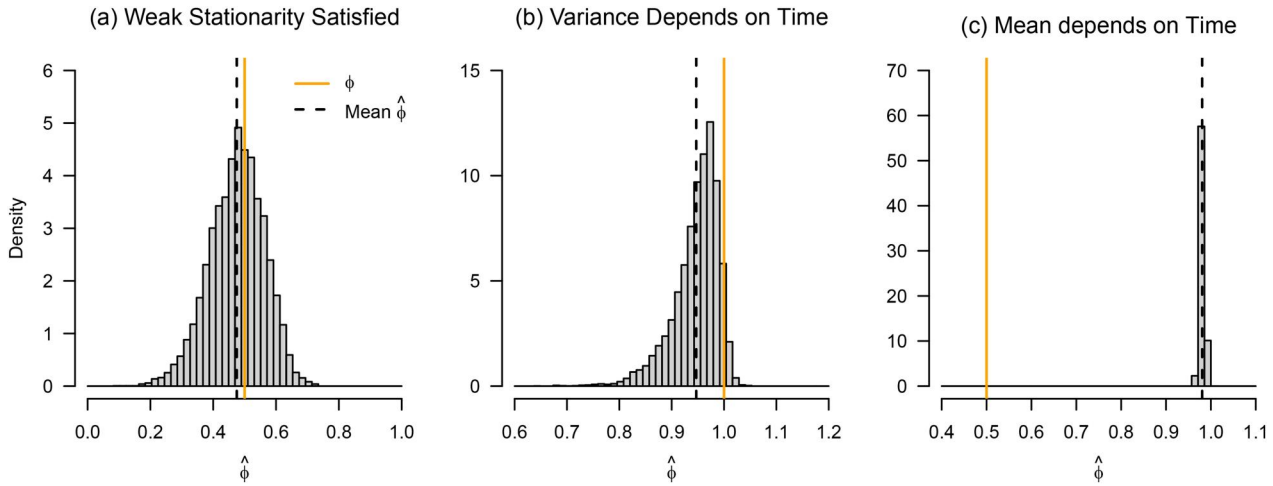

Figure 3. Sampling distribution of ϕ estimated based on time-series without trend, with deterministic trend, and with stochastic trend. The solid line indicates the ϕ in the true model; the dashed line indicates the mean of the sampling distribution.

requires a somewhat stronger restriction than weak stationarity. This restriction is essentially that the auto-covariance function approaches zero faster (that is, at shorter lags) than is required for the LLN. One way to achieve this is to assume that for lags larger than m there is independence. So, X_t and X_{t+m+1} are independent but X_t and X_{t+k}, with $k \leq m$ are dependent. If this holds for some finite m, then we obtain the CLT (White, 2001). In Appendix A3 we provide a more detailed explanation of this, but the main practical implication we consider in this paper is that, while weak stationarity alone may not be sufficient for the CLT, if weak stationarity is *violated*, then the CLT will not hold.

To illustrate how the CLT fails when the data generating process is not weakly stationary, we again consider the AR-models in panels (a), (b) and (c) in Figure 1. For each model we simulate 10,000 time-series with length $n = 100$ and estimate the autoregressive parameter ϕ. Figure 3 displays the resulting sampling distributions of $\hat{\phi}$ in the three cases. In panel (a) we see the case in which weak stationarity is satisfied and therefore the CLT holds. As predicted by the CLT, we see that the sampling distribution of $\hat{\phi}$ takes the form of a normal distribution, centered close to the true value $\phi = 0.5$.[1] Panel (b) shows the sampling distribution for the random walk in which the variance is a function of time and therefore the CLT does not hold. We see that the sampling distribution is skewed and its mean is considerably below the true parameter $\phi = 1$. This case illustrates one of the problems with simply ignoring the issue of stationarity: we would both obtained biased point estimates and our inference would be incorrect when constructing standard errors based on the normal distribution, while the true sampling distribution is clearly not normal. Finally, in panel (c) we see the sampling distribution of $\hat{\phi}$ for the model with a time-varying mean (i.e., the process depicted in Figure 1(c)). We see that the sampling distribution is far off centered around a value just below 1, while the true parameter is $\phi = 0.5$. Clearly, both point estimates and standard errors are extremely poor and would likely lead to incorrect conclusions.

2.3.3. Forecasting and model selection

While the CLT and LLN are concepts we use in almost all settings involving statistical inference, there are also a number of other reasons why stationarity is an appealing characteristic when studying systems that evolve over time. The most straightforward of these is that stationary processes are more amenable to prediction or *forecasting* than non-stationary processes.

If we are studying a stationary process, such as the stationary AR(1) process depicted in Figure 1(a) then forecasting is relatively straightforward. To make this explanation easier, let's assume that we know the generating model and its parameters, which in this case was an AR(1) model with $\phi = 0.5$. If we want to predict the value of the process at the next time point from the current value of the process, we simply use $\hat{X}_{t+1} = \hat{\phi} X_t$; that is, we use the AR(1) model equation, dropping the unknown white-noise or error term. If the process is, say, above its mean value at time t (e.g., the red line at $t = 1$), we would predict that the process is closer to its mean at the next time

[1] In fact, the sampling distribution is *not exactly* centered at the true value. The reason is Nickell's bias (Nickell, 1981), which is given by $-(1+\phi)/(T-1)$ and therefore goes to zero when increasing the time-series length T.

point $t+1$. How precise this forecast is depends on the amount of noise variance σ_ε^2 in the process. Forecasts further into the future can be obtained by iterating this process, with $\hat{X}_{t+h} = \hat{\phi}^h X_t$, with forecasts further into the future having more uncertainty. Of course, at some point, since ϕ is less than one in absolute value, the current value of X_t no longer contains any information with which we can predict X_{t+h} (since $\hat{\phi}^h \to 0$). In that case, if we want to make predictions about the process in the long-term (i.e., far into the future), we can instead use the *stationary distribution*; for a stationary process, we know that even far into the future, the mean and variance of the process will be the same. This stationary distribution is represented in Figure 1(a) with grey solid and dashed lines representing the mean and standard deviation respectively. When forecasting far into the future, our point prediction is the mean of the stationary distribution, with our uncertainty given by the standard deviation.

In contrast, consider the task of making forecasts about the non-stationary random walk process depicted in Figure 1(b). For the sake of simplicity, let's consider again that the parameters of the generating model are known, and we would like to make forecasts based on the initial value of the process depicted in the figure, $X_0 = 0$. Repeating the process we described above, we can see that our forecasts are simply equivalent to the horizontal line in the figure; since $\phi = 1$, then $\hat{X}_{t+h} = \ldots = \hat{X}_{t+1} = X_0 = 0$. Again here our uncertainty about those forecasts is depicted by the dotted line representing the standard deviation of the process at each point in time. We can see that, as we get further and further away from the last observed time point, these forecasts become worse and worse. However, unlike the stationary case above, since the variance of this process grows over time, our uncertainty becomes so large that eventually we are practically making the prediction that that the future value of the process will be somewhere between plus and minus infinity. This example shows that stationary processes are much more amenable to forecasting than non-stationary processes. It also shows why correctly identifying or modeling non-stationarity is so crucial when making forecasts: In both cases we considered here, in the long-term we make the same point forecast of $\hat{X}_{t+h} = 0$, but would have radically different ideas about the degree of confidence or uncertainty we should have about that prediction.

The concept of building forecasting models is also central to other important issues in time-series analysis, such as when Moving Average (MA) models can be related to AR models in the ARIMA modeling framework; for a detailed discussion of issues relating to forecasting with these models and so-called *invertible* MA models, we refer reads to Appendix A4. While for simplicity, we have here assumed that the true model and its parameters are known, of course in practice researchers must often perform some kind of *model selection* as well as parameter estimation. Here again, stationarity provides us with useful guarantees. For example, Wold's theorem (Hamilton, 1994; Wold, 1938) implies that, when we are studying a stationary process, some combination of AR and MA terms is capable of fully capturing the deterministic part of the system. This means that for stationary processes, ARMA models are an appropriate model class in which to perform model selection.

2.4. The importance of modeling (non)-stationarity

Above, we explained how stationarity can be seen as an *assumption* of time-series analysis: When stationarity holds, then theorems such as the Law of Large Numbers and Central Limit Theorem hold, giving us guarantees about our ability to estimate population properties from time-series data, among other useful properties. We have also seen how different configurations of even the simple AR(1) model can lead to processes which are stationary or non-stationary in different ways (such as with regards the mean, the variance and auto-covariance, or both). This shows that, even if we assume a relatively simple data-generating model, we must still contend with the possibility that we are studying a non-stationary process. In Figures 2 and 3, we have seen how violations of non-stationarity can have severe consequences for statistical inference, leading variously to biased estimates and sampling distributions which do not satisfy the requirements of standard inference, such as in hypothesis tests.

This raises a number of questions for researchers analyzing time-series data: Is there a way for researchers to detect when the time-series they are studying is non-stationary? And, if the researcher has a non-stationary time-series, is there any way to still make valid statistical inferences about the data-generating process? The remainder of this paper aims to provide answers to these questions. In the spirit of keeping the discussion as accessible as possible, in Section 3 we will approach these questions by focusing on processes which are not mean stationary, which is a relatively simple form of non-stationarity. After introducing different types of mechanisms for non-

stationary means we discuss how differencing and detrending the time-series can be used to model these different mechanisms, and how to determine the appropriate mechanism for non-stationarity in empirical data. In Section 4, where we will discuss key takeaways for researchers, we will also discuss cases that go beyond mean non-stationarity.

3. Modeling non-stationarity of means

In this section, we first introduce two mechanisms which might be responsible for producing changes in the mean of a stochastic process, also known as a *trend*, over time. These two mechansims are deterministic and stochastic trends (Section 3.1). We then discuss how these different mechanisms can be modeled with detrending and differencing, respectively (Section 3.2). Applying the right modeling approach requires knowing the type of generating mechanism at hand. We tackle this issue in Section 3.3, where we discuss and evaluate (an extension to) the popular Augmented Dickey Fuller (ADF) test to determine the type of mechanism underlying observed mean non-stationarity.

3.1. Stochastic and deterministic trends

A time-series is said to exhibit a *trend* if the mean of the time-series changes over time, $\mathbb{E}(X_t) = \mu_t$. As such, the presence of a trend necessarily implies a violation of *mean stationarity*, with all of the problems this causes for statistical inference that we have seen in the previous section. A trend (deterministic or stochastic) might be interesting in and of itself. For instance, in educational contexts, researchers may be primarily interested in understanding to what degree a students reading ability improves over time. This can be determined by inspecting whether there is an overall positive trend in reading ability test scores, and if so, determining how strong that trend is. The research may also be interesting in examining whether there is an increase in variation of reading ability; such an increase could indicate that a pupil does not seem to converge on the right strategy for a particular learning task. In other applications, however, the focus is on first modeling or removing the trend so that the remaining stationary part of the time-series can be properly analyzed.

There are many different mechanisms which might produce a trend in our time-series variable, and so many different approaches we could take to *model* a process which exhibits a trend. As we will show in this section, it is critical to determine the source of non-stationarity, because different sources require different data analysis strategies: Choosing an inappropriate way to model the trend can lead to the very same problems with biased estimates, inaccurate inferences and misleading conclusions as we saw when ignoring the presence of a trend in the previous section.

In the time-series analysis literature, the distinction is typically made between two different types of mechanisms that can produce trends in the mean: *deterministic trends* and *stochastic trends*. In fact, we have already seen examples of both types of trend in Figure 1(c,d) respectively. An AR(1) model with a linear deterministic trend would be written as

$$X_t = \alpha + \beta t + \phi X_{t-1} + \varepsilon_t \tag{2}$$

where βt, $\beta \neq 0$, determines the linear trend in combination with the value ϕ and ε_t is white noise (i.e., a random variable with mean zero and variance σ^2 for all t). When $|\phi| < 1$, it can be seen that the mean of the time-series is a linear function of time (see Appendix B1 for more details). Notably, while the βt term effects the mean of the process over time, it does not effect the variance of the process, which is determined by ϕ and σ^2. In Figure 1(c) we show an example of such a process, where we set $\beta = 0.1$. From this figure we can see that while the mean across the ensemble increases with time, the variance around that mean remains constant. Another way of thinking about this is that each member of the ensemble, plotted as different lines in panel (c), has a mean value that changes in exactly the same way over time; they share exactly the same trend. This also helps us understand why this type of trend is called *deterministic*: this name comes from the fact that that the mechanism producing the trend is nonrandom, that is, a deterministic, function of time. The example process shown here is also sometimes referred to as a *trend-stationary process*, in that we can decompose the process into a trend part to which we add a stationary process ($\phi X_{t-1} + \varepsilon_t$) varying around that trend. While here we focus only on linear trends, they could also be non-linear or cyclic (e.g., Douc et al., 2014; Haqiqatkhah & Hamaker, 2024; Mills & Mills, 2003).

On the other hand, an example of an AR(1) model with a (linear) *stochastic trend* would be

$$X_t = \alpha + 1X_{t-1} + \varepsilon_t \tag{3}$$

a model which is known as a random walk with drift. The term random walk comes from the fact that $\phi = 1$, which implies that the value of X_t will be equal to the value it took on at the previous time-step X_{t-1}, plus some change determined by the other parameters

of the model. The intercept α is here called the drift term, since it determines the average value of the "step" the process takes from time-point to time-point. The drift term, in combination with the random walk produces a linear trend in the mean over time $\mathbb{E}(X_t) = \alpha t$ (if $\alpha = 0$, the model is simply known as a random walk, which has no trend in the mean, as depicted in Figure 1(b)). Intuitively, since $\phi = 1$ each time we take another step forward in time, say from t to $t + 1$, we add another value of α to the right hand-side of the equation, producing a change in the mean value over time. Unlike the deterministic example given above, this model also produces a process whose *variance* depends on time, $\text{var}(X_t) = t\sigma^2$. The intuition here is that, because the auto-regressive parameter is equal to one, the effect of past shocks (i.e., past ε_t) never die out, but instead stay in the system. For a derivation and more detail see Appendix B2. We have already seen an example of this model in Figure 1(d), where we can see that both the ensemble mean and variance around this mean grow over time. Another way of understanding the difference between a deterministic and stochastic trend is to compare the ensemble time-series in panels (d) with those of panel (c). In panel (d), we can see that not all individual time-series share exactly the same positive trend over time; some members of the ensemble in fact appear to have a negative trend over the time-window considered, while for others the positive trend is stronger. This has to be the case, and is in fact an implication of the property that the variance grows over time. In the stochastic case, then, we can consider the trend in the variables to be *random*, with the *average* trend in the mean, over the ensemble, given by αt.

The distinction between deterministic and stochastic trends is useful in part, because each mechanism implies a distinct way of modeling the underlying trend μ_t. In a typical time-series analysis setting, researchers who suspect that a trend might be present in their time-series first aim to identify which mechanism is generating that trend. Second, they use this information to model the trend, which we can think about as modeling variation in the system over a *long timescale*. Finally, once that has been achieved, researchers (in principle) should be left to model the remaining *short timescale* variation in the system, which, now that the trend has been removed, should consist of a stationary time-series, with all of the desirable statistical properties outlined in the previous section.

Both models shown above can be said to have a linear trend in the mean, but obviously differ in other ways. Since both mechanisms are implying different properties (in terms of trends in the variance), we may also choose to interpret the models as positing qualitatively different sources for this trend. In the deterministic case, we could interpret the t term on the right-hand-side as denoting a constant effect of time, or as standing in for a set of exogenous variables that produce a constant rate of development in the process at hand. On the other hand, in the stochastic case, we could interpret this model as representing the case where the mean trend is produced as a result of how X_t relates to itself at the next time point: The fact that X_t does not "forget" the effect of random shocks in the past, but these in turn partly determine the value of the process far into the future. These interpretations extend to the case where we may have quadratic deterministic terms t^2 or even stronger auto-regressive effects $|\phi| > 1$ which produce explosive exponential growth in the mean, rather than a linear trend.

3.2. Detrending and differencing

The appropriate way of building a time-series model in the presence of a mean trend depends on whether the trend was produced by a deterministic or stochastic mechanism. Intuitively, this is simply because deterministic and stochastic mechanisms are produced by different types of models, as we saw above. Crucially, however, making the wrong choice can lead to similar problems as not modeling the trend at all. In this section, we will first show how *detrending*, a method often discussed in the time-series literature (Brockwell & Davis, 2009; Hamilton, 1994), is appropriate for deterministic trends, while *differencing* is appropriate for stochastic trends. We will then discuss the issues that emerge when choosing the wrong strategy. Applying the right method requires determining the type of trend in empirical data, a topic we will cover in Section 3.3.

Deterministic linear trends can be modeled by fitting a linear trend term $\alpha + \beta t$ to the data, and then fitting a time-series model to the stationary residual of this model:

$$\begin{aligned} X_t &= \alpha + \beta t + R_t \\ R_t &= \phi R_{t-1} + \varepsilon_t \end{aligned} \tag{4}$$

where R_t represents the residual, and if the data-generating model is the one shown in Equation (2), is an AR(1) process with zero mean. Conceptually, this procedure is very similar to *centering* the time-series around its mean, except in this case the mean itself is modeled as changing linearly over time.[2] As such, this

procedure is sometimes referred to simply as "linear detrending", although, as we will see, this term should be used with caution.

Stochastic linear trends, however, are modeled differently. To model a stochastic trend, we first *difference* the data before fitting an appropriate time-series model. For instance, an AR(1) model on differenced data would be:

$$\begin{aligned} \Delta X_t &= X_t - X_{t-1} \\ \Delta X_t &= \phi \Delta X_{t-1} + \varepsilon_t \end{aligned} \tag{5}$$

More precisely, this is referred to as *first-order* differencing. More complex stochastic trends may require higher-order differencing, such as $\Delta X_t - \Delta X_{t-1}$ in order to fully model the trend. To see how differencing enables us to model a stochastic linear trend, consider the random walk with the drift model introduced above. Subtracting X_{t-1} from both sides of the equation, we obtain $\Delta X_t = \alpha + \varepsilon_t$ (see Appendix C1). The interpretation here is that the *change* from one time point to another is simply a white noise process (i.e., an autoregressive model with $\phi = 0$) and with intercept α, similar to that shown in Figure 1(a). While differencing could also be seen as a pre-processing step before fitting a time-series model, in the ARIMA modeling framework, the choice of whether to difference, and indeed the order of differencing, is treated as a model parameter, referred to by the "integrated" part of the ARIMA term. For instance, an AR(1) model would be an ARIMA(1,0,0) model, while an AR(1) model on first differenced data would be an ARIMA(1,1,0) model (for a brief introduction to ARIMA models, see Appendix A1). Much like the order of the auto-regressive part, the order of differencing is a model parameter to be chosen, and the differencing itself is used when making forecasts from any resulting model: Forecasts from ARIMA(1,1,0) models should have linearly increasing variance over longer forecasting horizons, since the process is non-stationary in the variance.

In practice, researchers must choose the appropriate method, detrending or differencing, in order to model the trends in their time-series. We have seen in Figure 3 that failing to model a trend when it is present leads to bias and improper inference. However, what is perhaps less well appreciated is that choosing an *inappropriate* method to model the trends can have similarly disastrous results. In Figure 4 we show the effect of detrending (first row) and differencing (second row) the above models with either a deterministic (Equation (2); first column) or stochastic (Equation (3) second column) trend. On the diagonal of this figure, we see that detrending a deterministic trend, and differencing a stochastic trend, respectively, allows us to obtain (asymptotically) unbiased estimates of the ϕ parameter, with a normal sampling distribution allowing the use of standard inferential tests.[3] On the off-diagonal (top-right and bottom-left panels) of Figure 4, however, we show that either detrending a stochastic trend or differencing a deterministic trend leads to biased parameter estimates.

Since we know the generating mechanisms, we can analyze exactly what is going wrong. If we detrend the random walk with drift, then we are left with a residual R_t which is a random walk *without* drift. We saw an example of a random walk with no drift in Figure 1(b). In that sense, by detrending the random walk, we actually succeed in modeling the trend in the mean; however, we fail to model the trend in the variance $\text{var}(R_t) = t\sigma^2$. As such, when we fit an AR(1) model to the residual, the assumption of weak stationarity is violated. We obtain biased estimates and a non-normal sampling distribution, meaning that standard inferential testing will not work appropriately (see Appendix C3 for details). In fact, we are confronted with many of the same problems as when we ignored the violation of non-stationarity earlier, in Figure 3(c).

If we apply differencing to the model with a deterministic trend, then we also succeed in removing the mean trend present, but at the cost of introducing complexity elsewhere in the model. For the AR(1) model with linear deterministic trend introduced above, differencing the time-series gives us:

$$\begin{aligned} X_t &= \alpha + \beta t + \phi X_{t-1} + \varepsilon_t \\ -X_{t-1} &= -\alpha - \beta(t-1) - \phi X_{t-2} - \varepsilon_{t-1} \\ \Delta X_t &= \beta + \phi \Delta X_{t-1} + \varepsilon_t - \varepsilon_{t-1} \end{aligned}$$

While differencing has succeeded in removing the mean trend term βt, the resulting process is not an AR(1) model on the differenced scores. Instead, we have introduced the additional term $-\varepsilon_{t-1}$. Although we have primarily focused on AR models in the current paper, time-series models where the current value of the process is determined by the stochastic noise

[2]Note that if the data is generated using the model in Equation (2), then we will obtain different α and β parameter to those shown in equation 4 if we fit the latter to that data. This is because the α and β parameters have a direct interpretation as the intercept and slope of the mean trend in the latter, whereas in (2) the intercept and slope of the mean trend are also determined by ϕ, as shown in Appendix B. This is conceptually similar to the difference between a (random) intercept- and a mean-terms often discussed in the context of multilevel time-series models (Jongerling et al., 2015).

[3]Here again, we must contend with Nickell's bias, which shrinks as a function of the time-series length.

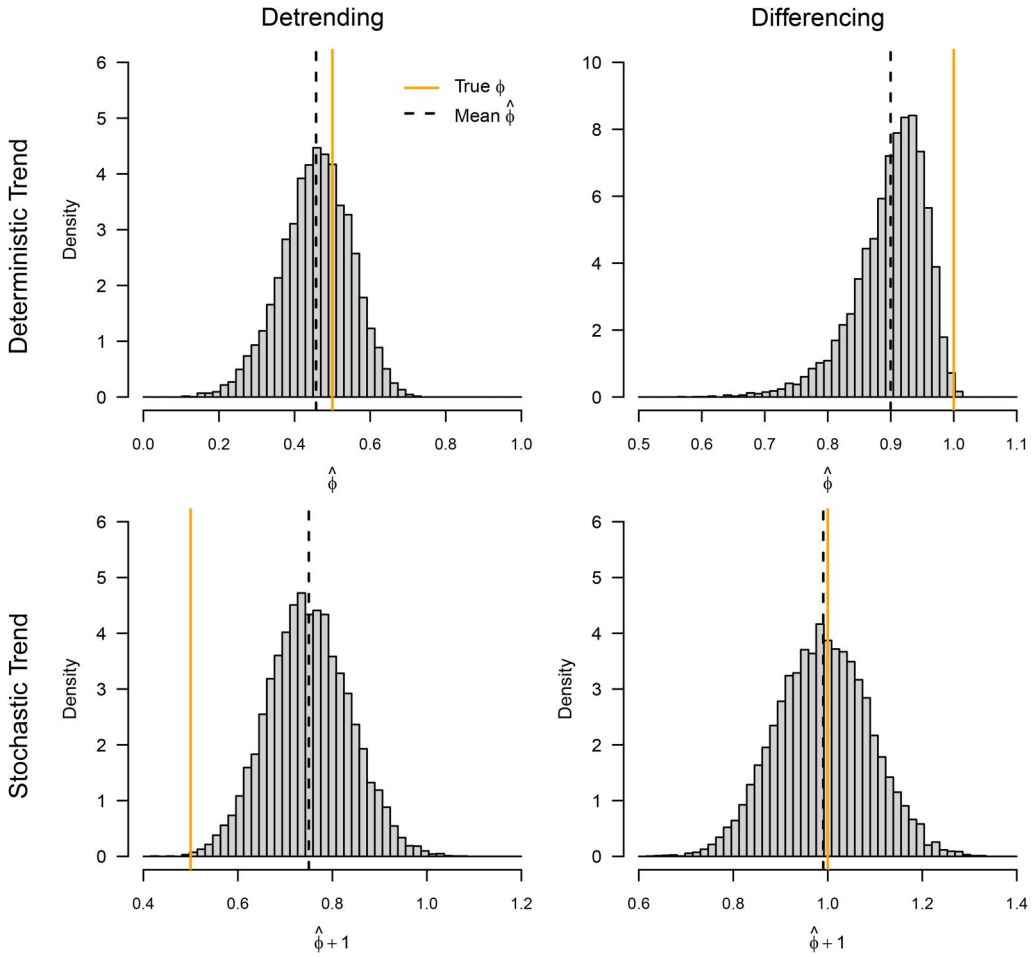

Figure 4. Sampling distribution of ϕ estimated based on time-series with either a deterministic trend (first column) or stochastic trend (second column), which are either de-trended by residualizing the time-series around a fitted linear trend (first row) or first-order differenced (second row) before analysis. The solid lines indicate the ϕ in the true model; the dashed line indicate the mean of the sampling distribution. For the differenced time-series we show the estimated sampling distribution of the parameter plus one.

term at the previous occassion are referred to as first-order *moving average* or MA(1) processes (see Appendix A1), where in the above model the MA(1) term is said to have a coefficient of -1, known as a unit root. The consequences of this are twofold. First, if we fit an AR(1) model to the differenced data, then we have fit a misspecified model; only a portion of the variance in difference scores from time-point to time-point is accounted for by an AR(1) term, with the rest explained by the unmodelled MA(1) unit root. As a consequence, we obtain highly biased estimates (Figure 4(c)). The second consequence is that fitting the "appropriate" model, including both an AR and MA term, would produce what is known as a "non-invertible model"; the right-hand side of the model can not be re-written only in terms of past values of X (see Appendix A4). This means that, to make predictions about future values of X_t, we would need to know all previous values of ε, but by definition, these are unobserved, and so, these issues cannot simply be resolved by fitting a more complex model to the differenced time-series. In the time-series analysis literature, this is known as an "over-differenced" time-series model.

This analysis shows that, when facing a non-stationary time-series, it is critically important that researchers correctly model the mechanism that produces that non-stationarity. As we have seen, there is no one-size-fits-all approach: Detrending a stochastic trend produces many of the same problems as failing to detrend at all, while differencing a deterministic trend has similar consequences but through a different mechanism. Of course, in practice, researchers may not have a priori knowledge about what mechanism produces the non-stationarity. As such, they must rely on data-driven ways of identifying whether a stochastic or deterministic trend (or a mix of both) is responsible for producing the mean trend.

3.3. Distinguishing deterministic and stochastic trends

So far, we have seen that deterministic and stochastic trends need to be modeled in different ways in order to avoid problems with biased parameter estimates and potential misleading conclusions about the processes under investigation. In practice, this requires us to be able to determine, for a given time-series, whether it is stationary, and if not, whether the trend is deterministic or stochastic. In Section 2, we saw that stationarity was a property of the ensemble, so perhaps at this point one might have the intuition that to detect non-stationarity one would need access to multiple different time-series from that ensemble. It turns out however, that we can test for the presence of stochastic or deterministic trends using a single time-series. The logic of this approach is based on the insight that (non-)stationarity is produced by certain combinations of parameter values of the underlying model. For instance, we have seen for the AR(1) model that $\beta = 0$ and $|\phi| < 1$ produces a stationary process; $\beta \neq 0$ a deterministic trend; and $|\phi| > 1$ a stochastic trend. Thus, if we have access to a single time-series, we can test for different types of non-stationarity by fitting time-series models and performing appropriate hypothesis tests on those key parameters. In what follows, we will introduce the (Augmented) Dickey Fuller (ADF) test, which is one of the primary methods used to test for the presence of a stochastic trend in time-series data using exactly the line of reasoning we have outlined above (Dickey & Fuller, 1979; Said & Dickey, 1984).

In the case of an AR(1) model, testing for the presence of a stochastic trend is the same as testing for whether the autoregressive parameter is greater than one in absolute value. In more general settings, that is, involving higher-order time-series models, this involves testing for a *unit root* in the characteristic equation of the time-series model (which we can understand as a function of the parameters of the model being greater than one in absolute value, see, e.g., Hamilton, 1994 and Appendix D). For the sake of simplicity, we will again consider the AR(1) case in this section, and therefore only discuss the original Dickey Fuller (DF) test (Dickey & Fuller, 1979). The Augmented Dickey Fuller test (Said & Dickey, 1984) generalizes what we present below to the case where the residuals have serial correlation and can be represented by higher order AR models (i.e., AR(p) models, see Appendix D1).

The Dickey Fuller test tests for the presence of a stochastic trend by fitting model of the form

$$\Delta X_t = \gamma X_{t-1} + \varepsilon_t \tag{6}$$

to the time-series data, and performing hypothesis tests on the γ parameter. If the process at hand follows a random walk, then we would expect that the γ parameter would have a value of zero. This can be seen by noting that the first-order AR(1) model can be re-written as follows

$$\begin{aligned} X_t &= \phi X_{t-1} + \varepsilon_t \\ X_t - X_{t-1} &= \phi X_{t-1} - X_{t-1} + \varepsilon_t \\ \Delta X_t &= (\phi - 1) X_{t-1} + \varepsilon_t \end{aligned}$$

and so, if the process follows a random walk, then $\phi = 1$ and so $(\phi - 1) = \gamma = 0$. Note here that, we are only differencing the *left-hand side* or outcome variable in this model, as distinct from differencing the entire time series and fitting an AR(1) model. Using this model we can in principle test for the presence of a random walk by testing the null hypothesis $\gamma = 0$. Rejecting the null hypothesis implies that no stochastic trend is present. Failing to reject the null hypothesis indicates the presence of a stochastic trend, which can be modeled by differencing. Notably, standard inferential tests may not yield proper inferences, and so this hypothesis test must be performed using the critical values supplied by Dickey & Fuller, 1981 (see also Appendix D).

While this is the basic approach of the DF test, strictly speaking, the exact approach described above will only work if the process mean does not change over time, either due to the presence of a drift term in addition to the random walk, or the presence of a deterministic trend. Violations of these assumptions can have unwanted consequences. For example, if we *exclude* a term that should be included, we may obtain biased estimates, or for instance, incorrectly difference a time-series which contains only a deterministic trend, as we showed in Figure 4(c); it has also been shown that if we *include* terms which are not present in the data-generating model (i.e., if we over-specify the model), we may have very low power to reject the null hypothesis even when it is false, meaning that we may often make the incorrect inference that the process has a unit root when no such unit root is present (Dickey & Fuller, 1981).

For this reason, there are actually three different versions of the (A)DF test which are used in practice. The first we have seen above, with the second two given as

$$\Delta X_t = \alpha + \gamma X_{t-1} + \varepsilon_t \tag{7}$$

and

$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \varepsilon_t \tag{8}$$

respectively. Both models can be understood as representing different data-generating processes, and so, different null-hypotheses of the Dickey Fuller test. For example, in the first model, if $\gamma = 0$ but $\alpha \neq 0$, then we are dealing with a random walk with drift. In the latter model, if $\gamma = 0$ but α and β are non-zero, then we are dealing with a random walk which has both a drift term, and a deterministic trend, resulting in a quadratic trend in the mean over time. As in the simpler model above, Dickey and Fuller (1981) provide nonstandard critical value tables for null hypothesis tests on the γ, α and β parameters.

3.4. Dickey Fuller tests and model selection

In practice, the correct use of the DF test relies on knowledge about which of these three DF models to use. Of course, without prior knowledge of what types of trend are present, researchers must in some way choose which version of the DF test to apply to test for a random walk, with a number of different approaches described in the literature. One popular approach described by Enders (2008) is to perform sequential or conditional hypothesis tests. In this approach, researchers begin by testing the most complex DF model (Equation (8)), including both an intercept and a deterministic trend term. If researchers fail to reject the stochastic trend hypothesis, they test if a simpler DF model can be used, thereby accounting for the fact that DF models that are too complex have low power. This procedure is outlined as a flowchart in Figure 5, and is implemented in an R-function available to readers in the online supplementary materials:[4]

Starting with the most complex DF model (Equation (8)) researchers test the γ null hypothesis; if this is rejected, then the researcher concludes that there is no stochastic trend, and the sequential testing procedure ends. If one fails to reject the γ null hypothesis, then the researcher tests if a simpler model is more appropriate, by testing if the deterministic trend term can be removed (i.e., testing the null hypothesis $\beta = 0$). If this is the case, the simpler DF model is used, the γ hypothesis test is repeated, and the process of either concluding there is a stochastic trend or potentially selecting a simpler model (in the last case by testing whether $\alpha = 0$) repeats until a decision is made.

This approach in principle avoids problems that may be caused by selecting an inappropriate model when performing a DF test, by testing for the presence of a stochastic trend even in the presence of either a drift term or a deterministic trend. As such, this sequential testing approach also provides a means of testing for combinations of stochastic and deterministic trends in the process under investigation: Once a unit root decision has been made, the user simply continues to test for model simplifications which would rule out, for instance, the presence of an additional deterministic trend term, as described above. For instance, if we find in the first step that there is no unit root, we can now test for the presence of a deterministic trend by testing $\beta = 0$ and $\alpha = 0$. The disadvantage of this procedure is that it relies on sequences of null hypothesis significance tests which have non-zero error rates; If any one hypothesis test returns the incorrect conclusion, this will produce an error in the model selection or unit root testing procedure.

We study the performance of the sequential DF testing procedure with a small simulation study. We generate from five different models, in all cases with a noise SD of 0.5: a stationary AR(1) model with $\phi = .5$; the same AR(1) model with an added deterministic trend $\beta = 0.5$; a random walk; a random walk with drift $\alpha = .25$; and a random walk with deterministic trend $\beta = 0.5$. Using 500 replications, we simulate the performance of the DF test at sample sizes, $n_t \in \{25, 50, 75 \ldots 500\}$. Throughout we use an alpha level of $\alpha = 0.01$ for all hypothesis tests. The left panel of Figure 6 shows the performance of the DF test when the correct DF model (Equations (6)–(8)) is known, while the right panel shows the performance of the sequential DF test procedure.

As we would expect, the DF test performs well when a random walk is present at all sample sizes, and performs well for stationary and trend stationary processes at sample sizes $n_t > 75$. For the sequential DF test procedure, the performance is noticeably worse for most conditions, and in particular the random walk and random walk with drift model. In particular the latter model shows somewhat unusual behavior, with performance initially worsening with increasing sample size before eventually improving to an acceptable level. The reasoning for this is as follows: With a strict alpha level, for small sample sizes, the null hypothesis is rarely rejected. In the case of the random walk with drift this yields good unit root detection rates at $n_t = 25$ but very poor unit root detection (false positives) for the stationary and trend stationary processes. As the sample size increases to $n_t = 100$ the performance is quite poor for the

[4] https://github.com/ryanoisin/NonStationaryTS.

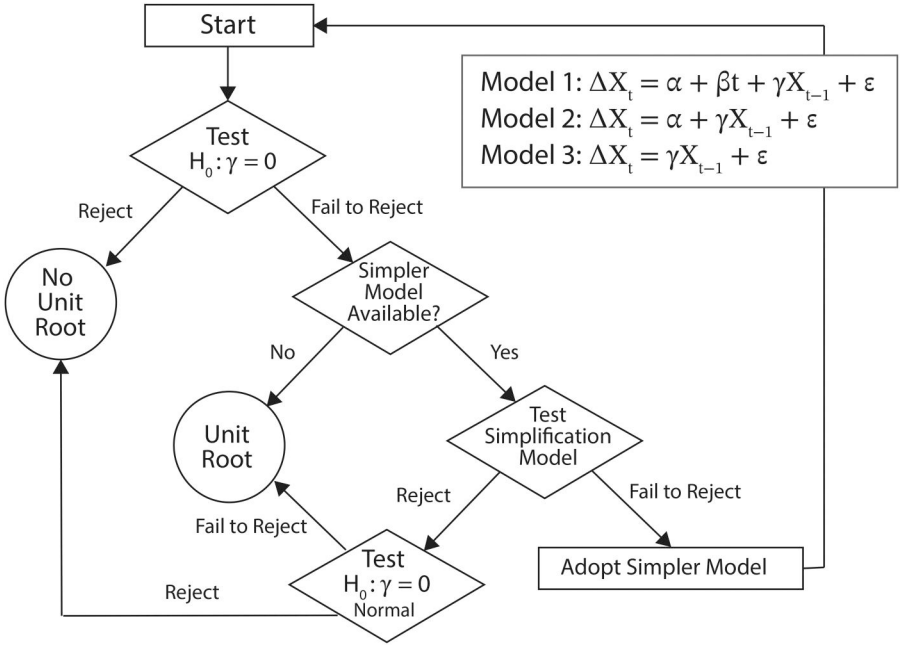

Figure 5. Flowchart representing a sequential testing approach to unit root testing with the (Augmented) Dickey-Fuller test. The procedure begins by using the most complex model (Equation (8)). If we fail to reject the null hypothesis $\gamma = 0$, this may be for two reasons: Due to model over-specification, or due to the presence of a unit root. As such, the next step of the testing procedure is to test whether the slope of the time co-efficient is equal to zero. If this is the case, then we proceed to use the simpler DF model, (Equation (7)). If $\gamma = 0$ using this model, we test for model over-specification using a hypothesis test on the intercept parameter. If we fail to reject the null hypothesis that the intercept is zero, then we proceed to use the simplest DF model (Equation (6)).

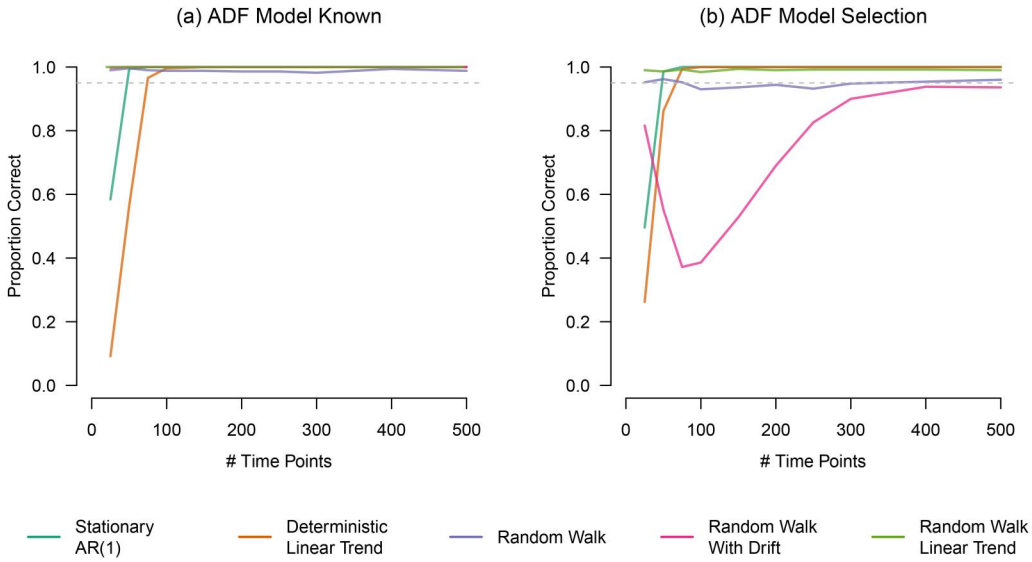

Figure 6. Left Panel: Performance of the DF test when the appropriate DF model is known. Right Panel: Performance of the sequential DF testing procedure which relies on selection of the correct DF model.

random walk with drift model, largely because the testing procedure rarely progresses past the first stage. Recall that a drift process implies that there is some non-zero trend in the mean over time. The result is that, the most complex DF model (Equation (8)) is often preferred (the time slope is non-zero) even though it is in fact over-specified relative to the correct DF model for this process (Equation (7)). However, that over-specification means that inappropriate critical values are used to assess the unit root, yielding poor performance. For larger sample sizes the sequential testing procedure works reasonably well for all processes, though necessarily performs worse than when the appropriate DF model is known a-priori. In

Appendix E we show that very similar qualitative patterns are present when varying either the size of the auto-regressive term or the standard deviation of the noise term. We also show that the performance "dip" of the model selection method for the random walk with drift is dependent on the size of the noise SD, such that it is accentuated when the SD is larger and disappears when it is sufficiently small.

The implication of this simulation study for researchers is that the sequential DF procedure outlined above provides a way, in principle, of distinguishing whether a stochastic trend, deterministic trend, or a combination of both, is present in the process which generates a given time-series. By allowing for both types of trends simultaneously, but first focusing on identifying random walk terms, we avoid making false positive conclusions about deterministic trends; for instance, if there is truly a stochastic trend present, we will always expect to find a non-zero deterministic trend term if we apply the linear detrending approach described in Equation (2), but we cannot use this as evidence that a deterministic and not stochastic trend underlies the data. In principle this approach can also be extended to higher-order AR models, as with the augmented version of the DF test. However, as our small simulation study shows, this approach comes with practical limitations; in small sample sizes ($n_t < 75$) we saw that this method performed relatively poorly for processes which had no random walk; for the case of a random walk with drift, the performance was poor with the sequential method for $n_t < 400$. The performance of such approaches is also likely to vary depending on the size of the autoregressive parameter in the true system; when ϕ is close to one it becomes more and more difficult to distinguish a system with a trend from a system with no trend using finite time points (see Appendix E for a demonstration). As such, purely data driven methods of identifying trend types should always be supplemented by theoretical knowledge on the part of the researcher applying them.

Note also that, while we have here focused on explaining the logic and demonstrating the performance of the popular ADF test, other data driven methods of testing for different types of (non-)stationarity have also been developed. These include the KPSS test (Kwiatkowski et al., 1992), which tests the null hypothesis that a process is stationary around a deterministic linear trend, and the PP (Phillips & Perron, 1988), ADF-GLS (Elliott et al., 1992) and NGP tests (Ng & Perron, 1995), which share many features with the ADF tests described here. For more detailed simulation studies comparing the relative performance of these methods across different settings, we refer readers to, amongst others, Afriyie et al. (2020) and Arltová and Fedorová (2016).

3.5. Interpretation of parameters depends on type of non-stationarity

Throughout this section we have addressed the importance of modeling trends appropriately, and how researchers might go about achieving that aim. However, it is important to note that, even when the mechanism of non-stationarity is correctly modeled, researchers still need to ensure to correctly interpret the parameters of the resulting models. This is a point which is less important when the time-series model is used only for forecasting or prediction purposes. However, in the social and behavioral sciences researchers are often interested in interpreting the parameters of these models, either to describe the system they are studying, or to aid in understanding the mechanisms which might generate the data.

Although all models described above yield estimates of a regression parameter ϕ such that when the predictor increases by 1 unit, the outcome increases by ϕ units, what can easily be missed is that the predictor and outcome variables are defined differently when fitting either a regular stationary AR(1) model (Equation (1)), an AR(1) on the residuals of a linear trend (Equation (4)) or an AR(1) model on differenced data (Equation (5)).

In the stationary setting, ϕ determines how the time series X moves back to its equilibrium across subsequent time-points. For example, a person's mood may be, on average, relatively stable over time, but might be momentarily perturbed by their environment: In this setting, ϕ determines how quickly their mood returns to its equilibrium state. However, in non-stationary settings, ϕ should be interpreted differently.

In the setting with deterministic trend, shown in Equation (4), ϕ determines how the *residuals* around a (linear) trend vary from time-point to time-point. For example, a child's cognitive ability may systematically increase from year-to-year during childhood. However, on a day-to-day basis, their cognitive ability may fluctuate around this systematic increase; a good night's rest may allow them to concentrate well today, while being placed in a distracting environment may lower their scores tomorrow. These perturbations, however, have no effect on the overall trend. In this

case, ϕ determines how quickly their cognitive ability score returns to its current *time-varying* equilibrium.

In the case of a stochastic trend, shown in Equation (5), ϕ determines how the *change* in X determines the amount of change in the following time-points. An example of such a process could be the extent to which one holds a positive or negative opinion about a person. We can imagine, for instance, that our opinion is determined by how that person acts, a random process modeled by ε_t : If a person does something positive, our opinion changes proportional to how positive that behavior was, and vice versa for negative behavior. This process could be modeled as a *random walk*: One's new opinion is equal to one's old opinion plus the effect of their behavior, leading to a stochastic trend over time. However, in the differenced representation Equation (5) we would expect to observe a value of $\phi = 0$: The amount one's opinion *changes* based on a new interaction is not influenced by the amount one's opinion changed based on the last interaction.[5] In all cases, ϕ describes a stationary *part* of the system, but the system itself, and what parts of that system are stationary, are very different.

4. Key takeaways for applied researchers

So far, we have introduced the concept of stationarity and how it is relevant in empirical research. We then used the simple case of trends to discuss why it is important to model non-stationarity appropriately and demonstrated how to identify different types of trends. The principles we have outlined already have a number of immediate and practical takeaways for applied researchers who wish to analyze time-series data, which we outline in the following. In many cases researchers may find themselves in scenarios more complex than those we have explicitly dealt with so far, which come with additional challenges, and which in some cases may require the development of a new methodology.

4.1. The issue with treating non-stationarity as "nuisance"

Researchers are often interested in fitting time series models to describe the short-term dynamics in a process and perform inference, such as hypothesis tests, on the parameters of those models. In this context stationarity is often discussed as a necessary assumption of those models. This is also how we discussed the concept of stationarity in Section 2.3, where we showed how stationary processes allow for more straightforward statistical inference in a time series context. However, this framing of stationarity could suggest that when we are studying a non-stationary process, researchers necessarily need to somehow "fix" the "problem" of non-stationarity in their data. This perspective can be problematic for two reasons.

First, it suggests that non-stationarity is a relatively trivial topic that can be dealt with with a simple correction, transformation or data analysis trick. This impression is strengthened by some common recommendations and analysis pipelines in empirical research. For example, when discussing the application of Vector Autoregressive models (a multivariate extension of AR-models) to time series in psychology, Jordan et al. (2020) only discuss detrending as a way to model non-stationarity. On the other hand, one may also be tempted to model all trends by using differencing techniques, for instance by utilizing the integrated part of the ARIMA modeling framework. Indeed, in the popular software package *fpp3* (Hyndman, 2023) the main ARIMA modeling function performs automated model selection, but does not test for deterministic trends as part of the model building procedure. Similarly, Jebb et al. (2015) discuss differencing as the main technique to obtain stationary time series. As a result, all trends in the data are treated as stochastic in nature. However, as we have shown in this paper, non-stationarity can take many forms and requires careful consideration since modeling it incorrectly can lead to incorrect conclusions. These problems are not only limited to obtaining biased parameter estimates and misleading test statistics, as we have focused on in this paper, but also extend to other uses of time-series models, such as forecasting. As we show in Figure 7, different approaches to detrending can yield similar point forecasts, but very different estimates about our uncertainty around future values of the process. Here we fit a deterministic trend (left panel) and an ARIMA model which includes differencing assuming a stochastic trend (right panel) to the same data. We see that, while the point forecasts are very similar, the uncertainty around those forecasts is much higher for the ARIMA model.

In addition, focusing on non-stationarity as a "problem to be fixed" distracts from the fact that the non-stationary aspect of the time series may often be the most interesting component. The reason is that

[5]Conversely, $\phi = 0.5$ might indicate that the effect of new interactions is partially determined by how one's opinion changed in the last interaction, etc.

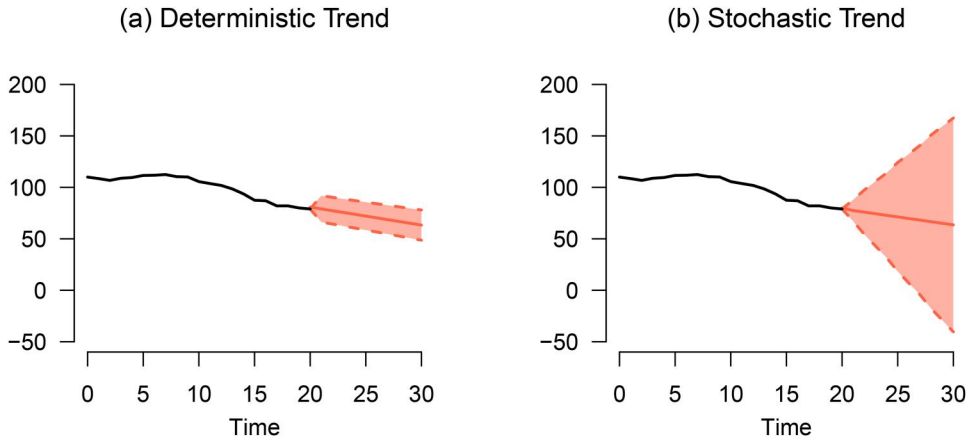

Figure 7. Figure showing the forecasts made by a linear growth-curve type model $X_t = a + b * T + e_t$ (left) modeling a deterministic trend and an ARIMA (0,1,0) model $X_t - X_{t-1} = a + e_t$ (right) modeling a stochastic trend, both fit to the same time-series data. Point forecasts are shown by the solid red line, with 95% prediction intervals denoted by the shaded region. Both models make approximately the same point forecasts, but with very different uncertainty regions.

often we are not (only) interested in the short-term behavior of a system, but also in how it behaves on the long term. For example, when studying economic development researchers may have weekly or monthly measurements, but be primarily interested in trends over the years. Similarly, in clinical psychology researchers may be interested in the etiology, maintenance and recovery from mental disorder over the course of years even if measurements are taken at a much higher frequency. Furthermore, although in this paper we have primarily discussed explicit modeling of non-stationarity in the form of stochastic and deterministic trends, there are many other approaches to modeling or using non-stationarity in more complex ways. For instance, the Invariant Causal Prediction method (Pfister et al., 2019) uses non-stationarity as a source of variance which allows, under certain conditions, researchers to use data from different time windows to identify stable causal sub-parts of a process.

4.2. Trends should be modeled "within-unit"

Traditionally, time-series analysis has been concerned with the analysis of repeated measures data from a single unit. For example, we might be interested in the GDP of a single country for a given period of years. However, when the number of available time-points is relatively small, then the estimation of time-series models may be difficult or ill-advised (e.g., Bulteel et al., 2018; Dablander et al., 2020; Mansueto et al., 2023); as we have seen in the previous section, we may need a relatively high number of time points to distinguish different trend types, and even when we are willing to assume a stationary process, the parameter estimates may be too variable (i.e., our parameters to uncertain) for us to make accurate inferences. If multiple units are available that are relatively similar to each other, we can use this similarity to estimate a model that makes use of the data of all units using multilevel (also known as hierarchical) models (e.g., McElreath, 2020). These examples are often found in the social and behavioral sciences, where the units are often individual persons (Epskamp et al., 2018; Hamaker et al., 2018; McNeish & Hamaker, 2020). One advantage which motivates the use of these methods is that these models regularize individual parameter estimates to the group average relations, thereby providing more stable estimates than models fit on each individual separately. A disadvantage of this approach is that we need to assume some kind of similarity across all units in terms of which time series model is appropriate to describe each units time series.

Our treatment of non-stationarity above has one clear implication for the multilevel context: The assessment of deterministic and stochastic trends, and any differencing or detrending procedure, must be conducted for each individual separately. The reasoning behind this implication is straightforward: As we have shown, inappropriately detrending a time-series can introduce bias in the parameter estimates in a similar way that failing to detrend a time-series with a deterministic trend does. If some individuals exhibit a deterministic trend, but others do not, or if individuals differ in the nature of the deterministic trend they exhibit, then it is impossible to apply a uniform detrending procedure without inducing bias in the parameter estimates. We illustrate this in Appendix F, where we show that subtracting the average trend,

rather than the unit-specific trends, can lead to a similar level of bias in multilevel time-series models as failing to remove any trend at all. The practical caveat to testing for the presence of a trend for each unit separately is that a relatively large number of time points is needed for each unit, as we have shown above.

A challenging consideration relates to the case where some individuals exhibit a stochastic trend and others do not, meaning that in some cases individual time series should be differenced, while in others not: Again, applying the same differencing procedure to all individuals would yield bias due to over-differencing (see panel (c) in Figure 4). Pre-differencing some individuals before fitting a multilevel model would appear to be a pragmatic approach, but, this might lead to problems in the interpretation of the fixed parameters, since some of the parameters relate to first order differences, and some to undifferenced time-series. As far as we are aware, there is no default or best practice for investigating differencing or detrending per person in the most popular multilevel time series modeling software. We therefore consider this as an important issue to be addressed in future methodological work on time series modeling.

4.3. Extension to multivariate Vector Autoregressive (VAR) models

Throughout the paper we have used the example of an AR(1) model to illustrate (non)-stationarity as well as different approaches to testing for and modeling non-stationarity. Although these approaches, and tools such as the Dickey Fuller test, can readily be extended to more complex univariate time series models (i.e., the family of ARIMA models in general), modeling non-stationarity becomes less straightforward when we model *multivariate* time-series.

A natural extension of the univariate AR models we have been discussing is a Vector Autoregressive (VAR) model. In a VAR model, several variables are modeled as AR processes simultaneously (Brockwell & Davis, 2009), while allowing for cross-lagged relations between variables over time. For instance, in a bivariate VAR(1) model, each variable is predicted by itself (through the autoregressive parameter) and the other variable (through the cross-lagged parameter) at the previous measurement occassion. Often in these settings the primary targets of inference are the cross-lagged relations themselves, often interpreted by researchers as indicative of potential causal relations (Granger, 1969).

Modeling non-stationarity in VAR models can quickly become relatively complicated. Staying with our simple bivariate VAR(1) model, let's assume that both processes are non-stationary. We then face the question whether we should detrend them, difference them, or should each variable be treated differently? These are difficult decisions, and for each variable the consequence of making an incorrect decision is similar to what we have shown above in the univariate case. The consequences of non-stationarity for testing dependence between two processes can also be problematic, since processes that exhibit a trend may be correlated even when they are not directly dependent. A well known issue in the econometrics literature is that two *independent* random walk processes will often be correlated, which can lead researchers to incorrectly conclude, on the basis of non-zero cross-lagged relationships, that two processes are dependent on one another (Granger & Newbold, 1974; Phillips, 1986).

In some cases there are solutions to these issues which can help researchers avoid erroneous conclusions in multivariate time-series settings. Approaches using co-integration allow researchers to study multivariate time series in the presence of trends (Engle & Granger, 1987; Hamilton, 1994) and involves finding a linear combination of the original variables (similar to a weighted average) such that the resulting linear combination is a stationary time series. This can then be used to estimate coefficients and transform the linear combination back to the original version. In Appendix H we provide an example of such a co-integrated VAR model. For an introduction to the core principles of co-integration, we refer readers to Engle and Granger (1987) as well as an excellent discussion of the topic in Hamilton (1994). The principles of the tests involved for stationarity are similar to those discussed above. For more than two time series other methods have been developed by Johansen (1988), which rely on a multivariate version of the augmented Dickey-Fuller test. While some ideas on how to model different forms of non-stationarity in the multivariate VAR setting exist, to our knowledge no workable solutions or software exists for in this setting. We consider this an important avenue for future research.

4.4. Modeling non-stationary with time-varying parameters

When discussing trends in Section 3 we discussed two mechanisms to generate them: stochastic and deterministic trends. The deterministic trend can be seen

as an intercept parameter that is a deterministic function of time. In principle we could also consider more general forms of deterministic mechanisms to produce non-stationarity. For instance, we might consider a model class in which all of the parameters of a time-series model are direct functions of time, such as the auto-regressive parameter $\phi(t)$ of an AR(1) model, or its innovation standard deviation $\sigma_\epsilon(t)$.

This approach of modeling parameters as functions of time gives rise to highly flexible time series models. There are many implementations available to estimate AR(p) and VAR(p) models in which the parameters are functions of time. For example, Casas and Fernandez-Casal (2019) allows estimating VAR models using a kernel-smoothing approach, and Krueger (2015) discusses a method using a Bayesian framework. Haslbeck, Bringmann, et al. (2021) discuss and evaluate two such methods, which use either Generalized Additive Models (Bringmann et al., 2017) or kernel smoothing (Haslbeck & Waldorp, 2020). However, while models with time-varying parameters often provide a good fit to the data, they do not necessarily recover the data generating mechanism. In Appendix G we demonstrate this issue by simulating from the four models in Figure 1 and estimating time-varying AR models using the two methods discussed in Haslbeck, Bringmann, et al. (2021). While both methods fit the data relatively well, they do not in all cases recover the data generating model. For example, the GAM-based method models the random walk with drift with a time-varying intercept instead of an autocorrelation of 1; the kernel method, on the other hand, models the deterministic trend with an autocorrelation close to 1. This suggests that we should be cautious about drawing strong conclusions about the data generating model based on these models and interpret them descriptively instead.

There are also other approaches to more explicitly model different mechanisms which produce time-varying parameters. For instance, with *regime-switching* methods researchers can model time-series which switch between different states, with each state possibly differing in their mean, variance and auto-covariance function. Popular methods include threshold auto-regressive models (Tong & Lim, 1980) which model these changes as dependent on either the value of the process itself or some external control variable, and Hidden Markov Models (HMMs; e.g., Hamaker et al., 2010; Visser & Speekenbrink, 2010; Zucchini et al., 2017), which describe systems that switch between states as determined by a latent or hidden variable. Alternatively, Generalized Autoregressive Conditional Heteroskedastic (GARCH) models allow us to model systems where deviations from the mean are dependent on previous time-points, allowing the process to have different variance at different time points. Typically, a GARCH model has alternating periods of small fluctuations and periods of large fluctuations, and is therefore sometimes called the volatility model (Engle, 1982). An example of where this may be interesting is in psychopathology research. During unhealthy periods not only will the mean value of affective variables (likely) change, but also the variation around these means will be larger. For more details on this model see Appendix I.

4.5. Discrete valued time-series require different approaches

Throughout our treatment of stationarity so far, we have considered time-series of *unbounded* continuous variables, that is, variables that can take on any value between plus and minus infinity. In principle, high-level concepts such as different types of (non)-stationarity, as discussed in Section 2, carry over to non-continuous variables, for which we can still define expectations, variances and different moments, and for which non-stationarity still complicates statistical inference. In practice, however, some of the details of how to model non-stationarity may differ for different types of variables, and some of the intuitions we have built up in this paper may not carry over.

This is relevant for many researchers, because in practice measurements are typically discrete and bounded. For example, in public health, the number of disease transmissions, cigarette sales per capita, or total deaths are bounded below by zero. In the social sciences, self-report scales may measure constructs using ordinal Likert scales, limited to, for instance, five or seven response types, or 0–100 scales (Haqiqatkhah et al., 2024; Haslbeck et al., 2023; Mader et al., 2023). In these situations, some of the intuitions we have built up around stationarity for continuous variables would not carry over directly. For instance, a random walk as we defined it in Section 2 cannot occur in its original form in a bounded variable. Instead, if there is a random walk process for bounded variables the random walk would get to the bound and stay there. When the range of the time series is relatively short, there is also a limit to the size of a mean or variance change that we can observe in practice. This may mean that linear detrending or differencing operations might be inappropriate to use for these types of variables.

Fortunately, there are alternative methods to analyze these types of time-series which have been well studied

in a number of fields. Discrete-valued time-series can be analyzed using so-called *Markov models* (Howard, 2012; Norris, 1997), which model how a system changes between discrete *states* from one time-point to the next using a matrix of transition probabilities $P(s, r) = \mathbb{P}(D_{t+1} = r|D_t = s)$ denoting the probability of the discrete variable D changing from state r to state s at the next time point. The basic principles of (non)-stationarity also apply to markov processes: if the conditional probability $P(s, r)$ remains constant over time, then we have a stationary distribution, since eventually the probability of being in any state s will become fixed (Norris, 1997). A number of R packages exist for the analysis of time-series data using Markov models, such as markovchain (Spedicato, 2017) and msm (Jackson, 2011). In addition, a number of extensions to classic ARMA models have been developed to model systems with discrete or bounded distributions (Haqiqatkhah et al., 2024; McKenzie, 1985; Weiß, 2020; Weiß & Kim, 2013) How exactly we can deal with or model non-stationarity in these more complex cases is beyond the scope of the current paper to review (Weis, 2018, for details we refer readers to). More research is needed to investigate the degree to which standard detrending and differencing approaches are appropriate or lead to problems when applied to bounded or discrete time-series.

5. Conclusion

In this paper, we aimed to provide an accessible introduction to the topic of (non)-stationarity in the context of analyzing empirical time series. We clarified which types of stationarity there are and illustrated violations of stationarity using the well-known first-order autoregressive AR(1) model. We then discussed how stationarity can serve as a statistical assumption that allows us to make valid inferences from time series data. Taking mean trends in AR(1) models as an example, we explained the difference between stochastic and deterministic trends and illustrated that detecting and correctly modeling trends is critical for valid inference. With the (augmented) Dickey–Fuller test we introduced the most popular test to identify the presence of a stochastic trend, and showed how this approach can be used in principle to distinguish between stochastic and deterministic trends.

While researchers likely encounter more complicated forms of non-stationarity in their data, understanding the rationale for modeling the simplest case of trends helps us to reason through more complex problems in empirical research. We demonstrated this in the last section, in which we discussed key take-aways about (non-)stationarity that are relevant to empirical researchers, and connected the topic in simple settings with more complex ones involving multiple variables, multiple units, time-varying parameters or variables that are not continuous. In many cases there are directly applicable take-aways for researchers: Non-stationarity may be the most interesting part of a system to model, and not simply a nuisance, and in multiple-unit settings, such as multilevel time series modeling, trends should be modeled "within-unit". Some of these take-aways are less directly applicable, and point instead toward areas in need of further development to create solutions which are easy to understand and straightforward to apply in broad social and behavioral settings. In particular, the recent growth in popularity of multilevel multivariate time-series modeling in certain areas of psychological science (Epskamp et al., 2018; Hamaker et al., 2021) point to a need for more work on how non-stationarity can be effectively modeled in those settings. To develop solutions in those settings requires us to understand the basic principles of stationarity, and so we hope that this paper lowers the barriers to entry for more researchers to work on developing those solutions.

We hope that our introduction provides the many researchers without strong technical background working on time series analysis with a better understanding of non-stationary time series and the importance of adequately modeling non-stationarity, as well as providing a good foundation which researchers can use to reason through their time series analysis problem at hand.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant 024.004.016 from the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NOW, gravitation grant for the project 'New Sciences

of Mental Disorders' and Grant 865468 from the European Research Council (ERC, Consolidator grant).

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Tessa Blanken, Denny Borsboom, Julian Burger, Sacha Epskamp and Alberto Jover for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or funding agencies is not intended and should not be inferred.

Open Scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at https://github.com/ryanoisin/NonStationaryTS

References

Afriyie, J. K., Twumasi-Ankrah, S., Gyamfi, K. B., Arthur, D., & Pels, W. A. (2020). Evaluating the performance of unit root tests in single time series processes. *Mathematics and Statistics*, *8*(6), 656–664. https://doi.org/10.13189/ms.2020.080605

Ariens, S., Ceulemans, E., & Adolf, J. K. (2020). Time series analysis of intensive longitudinal data in psychosomatic research: A methodological overview. *Journal of Psychosomatic Research*, *137*, 110191. https://doi.org/10.1016/j.jpsychores.2020.110191

Arltová, M., & Fedorová, D. (2016). Selection of unit root test on the basis of length of the time series and value of ar (1) parameter. *Statistika*, *96*(3), 47–64.

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling*, *26*(1), 119–142. https://doi.org/10.1080/10705511.2018.1511375

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling*, *25*(3), 359–388. https://doi.org/10.1080/10705511.2017.1406803

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*, *22*(3), 409–425. https://doi.org/10.1037/met0000085

Brockwell, P. J., & Davis, R. A. (2009). *Time series: Theory and methods*. Springer Science & Business Media.

Brodeur, A., Clark, A. E., Fleche, S., & Powdthavee, N. (2021). Covid-19, lockdowns and wellbeing: Evidence from google trends. *Journal of Public Economics*, *193*, 104346. https://doi.org/10.1016/j.jpubeco.2020.104346

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). Var (1) based models do not always outpredict AR (1) models in typical psychological applications. *Psychological Methods*, *23*(4), 740–756. https://doi.org/10.1037/met0000178

Casas, I., & Fernandez-Casal, R. (2019). Tvreg: Time-varying coefficients linear regression for single and multiple equations [computer software manual]. (R package version 0.3. 0). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3363526

Dablander, F., Ryan, O., & Haslbeck, J. M. B. (2020). Choosing between ar (1) and var (1) models in typical psychological applications. *PLOS One*, *15*(10), e0240730. https://doi.org/10.1371/journal.pone.0240730

DasGupta, A. (2008). *Asymptotic theory of statistics and probability*. Springer-Verlag.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427–431. https://doi.org/10.2307/2286348

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, *49*(4), 1057–1072. https://doi.org/10.2307/1912517

Douc, R., Moulines, E., & Stoffer, D. (2014). *Nonlinear time series: Theory, methods and applications with R examples*. CRC press.

Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge University Press.

Elliott, G., Rothenberg, T. J., & Stock, J. H. (1992). Efficient tests for an autoregressive unit root. *Econometrica*, *64*(4), 813–836.

Enders, W. (2008). *Applied econometric time series*. John Wiley & Sons.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, *50*(4), 987–1007. https://doi.org/10.2307/1912773

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, *55*(2), 251–276. https://doi.org/10.2307/1913236

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791

Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111–120. https://doi.org/10.1016/0304-4076(74)90034-7

Haig, B. D. (2008). An abductive perspective on theory construction. *Journal of Theory Construction & Testing*, *12*(1), 1046–1068.

Hamaker, E. L., Asparouhov, T., & Muthén, B. (2021). Dynamic structural equation modelling as a combination of time series modeling, multilevel modeling, and

structural equation modeling. *In* The *handbook of structural equation modeling*, (vol. 31). Guildford Press.

Hamaker, E. L., Asparouhov, T., Brose, A., Schmiedek, F., & Muthén, B. (2018). At the frontiers of modeling intensive longitudinal data: Dynamic structural equation models for the affective measurements from the cogito study. *Multivariate Behavioral Research*, *53*(6), 820–841. https://doi.org/10.1080/00273171.2018.1446819

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2010). Regime-switching models to study psychological processes. In P. C. M. Molenaar & K. M. Newell (Eds.), Individual pathways of change: Statistical models for analyzing learning and development (pp. 155–168). American Psychological Association.

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2016). Modeling bas dysregulation in bipolar disorder: Illustrating the potential of time series analysis. *Assessment*, *23*(4), 436–446. https://doi.org/10.1177/1073191116632339

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

Haqiqatkhah, M. M., & Hamaker, E. L. (2024). Daily dynamics and weekly rhythms: A tutorial on seasonal ARMA models combined with day-of-week effects. PsyArXiv preprint. https://doi.org/10.31234/osf.io/duvqh

Haqiqatkhah, M. M., Ryan, O., & Hamaker, E. L. (2024). Skewness and staging: Does the floor effect induce bias in multilevel AR (1) models? *Multivariate Behavioral Research*, *59*(2), 289–319. https://doi.org/10.1080/00273171.2023.2254769

Haslbeck, J. M. B., & Waldorp, L. J. (2020). Mgm: Estimating time-varying mixed graphical models in high-dimensional data. *Journal of Statistical Software*, *93*(8), 8. https://doi.org/10.18637/jss.v093.i08

Haslbeck, J. M. B., Bringmann, L. F., & Waldorp, L. J. (2021). A tutorial on estimating time varying vector autoregressive models. *Multivariate Behavioral Research*, *56*(1), 120–149. https://doi.org/10.1080/00273171.2020.1743630

Haslbeck, J. M. B., Ryan, O., & Dablander, F. (2023). Multimodality and skewness in emotion time series. *Emotion*, *23*(8), 2117–2141. https://doi.org/10.1037/emo0001218

Haslbeck, J. M. B., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2022). Modeling psychopathology: From data models to formal theories. *Psychological Methods*, *27*(6), 930–957. https://doi.org/10.1037/met0000303

Henrie, C. R., Bodily, R., Manwaring, K. C., & Graham, C. R. (2015). Exploring intensive longitudinal measures of student engagement in blended learning. *International Review of Research in Open and Distributed Learning*, *16*(3), 131–155.

Howard, R. A. (2012). *Dynamic probabilistic systems: Markov models*. (Vol. 1) Courier Corporation.

Hyndman, R. (2023). *Fpp3: Data for "forecasting: Principles and practice*. (3rd ed.). OTexts. https: //github.com/robjhyndman/fpp3package

Jackson, C. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, *38*(8), 1–28. https://doi.org/10.18637/jss.v038.i08

Jacod, J., & Protter, P. (2004). *Probability essentials*. Springer-Verlag.

Jebb, A. T., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: Examining and forecasting change. *Frontiers in Psychology*, *6*, 727. https://doi.org/10.3389/fpsyg.2015.00727

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control*, *12*(2-3), 231–254. https://doi.org/10.1016/0165-1889(88)90041-3

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel ar (1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, *50*(3), 334–349. https://doi.org/10.1080/00273171.2014.1003772

Jordan, D. G., Winer, E. S., & Salem, T. (2020). The current status of temporal network analysis for clinical science: Considerations as the paradigm shifts? *Journal of Clinical Psychology*, *76*(9), 1591–1612. https://doi.org/10.1002/jclp.22957

Kennedy, C. E., & Turley, J. P. (2011). Time series analysis as input for clinical predictive modeling: Modeling cardiac arrest in a pediatric ICU. *Theoretical Biology & Medical Modelling*, *8*(1), 40. https://doi.org/10.1186/1742-4682-8-40

Krueger, F. (2015). Bvarsv: Bayesian analysis of a vector autoregressive model with stochastic volatility and time-varying parameters. R Package Version, 1. https://cran.r-project.org/web/packages/bvarsv/index.html

Kuppens, P. (2015). It's about time: A special section on affect dynamics. *Emotion Review*, *7*(4), 297–300. https://doi.org/10.1177/1754073915590947

Kuppens, P., Dejonckheere, E., Kalokerinos, E. K., & Koval, P. (2022). Some recommendations on the use of daily life methods in affective science. *Affective Science*, *3*(2), 505–515. https://doi.org/10.1007/s42761-022-00101-0

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1-3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Law, K. L., Morét-Ferguson, S., Maximenko, N. A., Proskurowski, G., Peacock, E. E., Hafner, J., & Reddy, C. M. (2010). Plastic accumulation in the north Atlantic subtropical gyre. *Science*, *329*(5996), 1185–1188. https://doi.org/10.1126/science.1192321

Lehmann, E. (1999). *Elements of large sample theory*. Springer.

Mader, N., Arslan, R. C., Schmukle, S. C., & Rohrer, J. M. (2023). Emotional (in) stability: Neuroticism is associated with increased variability in negative emotion after all. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(23), e2212154120. https://doi.org/10.1073/pnas.2212154120

Mansueto, A. C., Wiers, R. W., van Weert, J. C. M., Schouten, B. C., & Epskamp, S. (2023). Investigating the feasibility of idiographic network models. *Psychological Methods*, *28*(5), 1052–1068. https://doi.org/10.1037/met0000466

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

McKenzie, E. (1985). Some simple models for discrete variate time series 1. *JAWRA Journal of the American Water Resources Association*, *21*(4), 645–650. https://doi.org/10.1111/j.1752-1688.1985.tb05379.x

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

Mills, T. C., & Mills, T. C. (2003). *Modelling trends and cycles in economic time series* (Vol. 10) Springer.

Mudelsee, M. (2010). Climate time series Classical Statistical and Bootstrap Methods (1st edition). Springer, Dordrecht.
Nakamura, T., Kiyono, K., Wendt, H., Abry, P., & Yamamoto, Y. (2016). Multiscale analysis of intensive longitudinal biomedical signals and its clinical applications. *Proceedings of the IEEE*, *104*(2), 242–261. https://doi.org/10.1109/JPROC.2015.2491979
Ng, S., & Perron, P. (1995). Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association*, *90*(429), 268–281. https://doi.org/10.1080/01621459.1995.10476510
Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, *49*(6), 1417–1426. https://doi.org/10.2307/1911408
Norris, J. (1997). *Markov chains*. Cambridge University Press.
O'Brien, T. D., Lorenzoni, L., Isensee, K., & Valdés, L. (2017). What are marine ecological time series telling us about the ocean. A status report. *IOC Technical Series*, *129*, 1–297.
Olthof, M., Hasselman, F., Strunk, G., van Rooij, M., Aas, B., Helmich, M. A., Schiepek, G., & Lichtwarck-Aschoff, A. (2020). Critical fluctuations as an early-warning signal for sudden gains and losses in patients receiving psychotherapy for mood disorders. *Clinical Psychological Science*, *8*(1), 25–35. https://doi.org/10.1177/2167702619 865969
Pesaran, M. H. (2015). *Time series and panel data econometrics*. Oxford University Press.
Pfister, N., Bühlmann, P., & Peters, J. (2019). Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, *114*(527), 1264–1276. https://doi.org/10.1080/01621459.2018.1491403
Phillips, P. C. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics*, *33*(3), 311–340. https://doi.org/10.1016/0304-4076(86)90001-1
Phillips, P. C., & Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, *75*(2), 335–346. https://doi.org/10.1093/biomet/75.2.335
Reichenbach, H. (1956). *The direction of time* (Vol. 65) Univ of California Press.
Reilly, T., Tyrrell, A., & Troup, J. (1984). Orcadian variation in human stature. *Chronobiology International*, *1*(2), 121–126. https://doi.org/10.3109/07420528409059129
Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607. https://doi.org/10.1093/biomet/71.3.599
Schaffer, A. L., Dobbins, T. A., & Pearson, S.-A. (2021). Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: A guide for evaluating largescale health interventions. *BMC Medical Research Methodology*, *21*(1), 58. https://doi.org/10.1186/s12874-021-01235-8
Spedicato, G. A. (2017). Discrete time Markov chains with R. *The R Journal*, *9*(2), 84–104. https: //journal.r-project.org/archive/2017/RJ-2017-036/index.html https://doi.org/10.32614/RJ-2017-036
Stone, A. A., & Broderick, J. E. (2007). Real-time data collection for pain: Appraisal and current status. *Pain Medicine*, *8*(suppl 3), S85–S93. https://doi.org/10.1111/j.1526-4637.2007.00372.x
Tong, H., & Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society B*, *42*(3), 245–268. https://doi.org/10.1111/j.2517-6161.1980.tb01126.x
van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., van Nes, E. H., Viechtbauer, W., Giltay, E. J., Aggen, S. H., Derom, C., Jacobs, N., Kendler, K. S., van der Maas, H. L. J., Neale, M. C., Peeters, F., Thiery, E., Zachar, P., & Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(1), 87–92. https://doi.org/10.1073/pnas.1312114110
Visser, I., & Speekenbrink, M. (2010). Depmixs4: An R package for hidden Markov models. *Journal of Statistical Software*, *36*(7), 1–21. https://doi.org/10.18637/jss.v036.i07
Weiß, C. H. (2018). *An introduction to discrete-valued time series*. John Wiley & Sons.
Weiß, C. H. (2020). Regime-switching discrete ARMA models for categorical time series. *Entropy*, *22*(4), 458. https://doi.org/10.3390/e22040458
Weiß, C. H., & Kim, H.-Y. (2013). Binomial AR (1) processes: Moments, cumulants, and estimation. *Statistics*, *47*(3), 494–510. https://doi.org/10.1080/02331888.2011.605893
White, H. (2001). *Asymptotic theory for econometricians*. Academic Press.
Wold, H. (1938). *A study in the analysis of stationary time series* [Doctoral dissertation]. Almqvist & Wiksell.
Woodward, J. (1989). Data and phenomena. *Synthese*, *79*(3), 393–472. https://doi.org/10.1007/BF00869282
Zucchini, W., MacDonald, I. L., & Langrock, R. (2017). *Hidden Markov models for time series: An introduction using R*. CRC press.

Appendix A. Stationarity and statistical properties

A1. The ARMA and ARIMA model

In this Appendix we introduce briefly the concepts of autoregressive (AR) and moving average (MA) processes, including the integrated version (ARIMA). We mostly rely on material from Brockwell and Davis (2009) and from Hamilton (1994).

We consider random variables X_t with values on the real line $\mathbb{R}$ as a function of time t. The time domain is discrete, that means we only consider integer values $t = 0, \pm 1, \pm 2, \ldots$. We sometimes use the notation $t \in \mathbb{Z}$ to indicate this. Each X_t can then be considered a random variable, obtained from the distribution at time t. An example of a very simple time series is where at each time point the draw from an identical distribution is independent for each t, i.e., $X_t = Z_t$, where Z_t is a random variable with mean 0 and variance 1 independent of Z_{t+h} for any $t, h \in \mathbb{Z}$. Then we have the familiar situation from classical statistics where observations are independent and identically distributed (i.i.d.).

A more interesting case in time series is if values of time points previous to the current one determine to some extent the value of the current time point. For instance, consider the process

$$X_t = \phi X_{t-1} + Z_t$$

where Z_t is an i.i.d. sequence with mean 0 and variance 1 (often called white noise) and ϕ is a regression parameter that relates the previous time point to the current time point. In the main text we refer to Z_t as ε_t. In time-series Z_t is often used, but ε_t is consistent with the standard statistical practice of referring to error variables with Greek letters. Here, Z_t is used to denote that this can also be seen as any other white-noise time series variable. This model is called an autoregressive model of order one, denoted by AR(1). Another important example of a time series is the model

$$X_t = \theta Z_{t-1} + Z_t$$

where, again, the Z_t are white noise. This is called a moving average of order one, denoted by MA(1). These two simple looking models are at the core of many time series analyses. The two models combined give

$$X_t - \phi X_{t-1} = \theta Z_{t-1} + Z_t$$

This is called an ARMA(1,1) model, since it contains both an AR(1) and an MA(1) model. This can be generalized to what are referred to as ARMA(p,q) models. The class of ARMA models is extensively studied. The reason is that they are simple enough to allow for detailed analytical study while still being rich enough to emulate observed data.

An important observation is that an AR model can be rewritten as an MA model. Consider the AR(1) model above. If we recursively substitute for X_t, then we obtain

$$\begin{aligned} X_t &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t = \cdots \\ &= \phi^t Z_0 + \phi^{t-1} Z_1 + \cdots + Z_t \end{aligned}$$

where $X_0 = 0$. This is clearly an MA model of order t. We can also rewrite an MA model into an AR model given certain assumptions (see Appendix A4). To do this we require the so-called *back-shift operator* B, such that $BX_t = X_{t-1}$, and so an application of the operator obtains a shift back in time. We also have that $B^2 X_t = X_{t-2}$. We can now write the MA(1) model above as

$$X_t = (1 - \theta B) Z_t$$

where we recoded θ to include the minus sign. If we are allowed to take the inverse of $1 - \theta B$ (see Appendix A4) then we obtain

$$(1 - \theta B)^{-1} X_t = (1 - \theta B - \theta^2 B^2 + \cdots) X_t = Z_t$$

and this is clearly an AR process.

The recursive substitution of the AR(1) model above indicates that there might be certain restrictions where the AR model could be extrapolated to long times (infinite time). This is important theoretically, because then we can determine stability and convergence for any time window. In such cases we can determine the statistical properties (mean and variance and the like) of such series. For instance, considering the AR(1) model, written as an MA process, gives mean and variance, respectively

$$\begin{aligned} \mathbb{E}(X_t) &= \phi^t \mathbb{E}(Z_0) + \phi^{t-1} \mathbb{E}(Z_1) + \cdots + \mathbb{E}(Z_t) = 0 \\ \mathbb{E}(X_t^2) &= \phi^{2t} \mathbb{E}(Z_0^2) + \phi^{2(t-1)} \mathbb{E}(Z_1^2) + \cdots + \mathbb{E}(Z_t^2) \end{aligned}$$

Suppose that $\mathbb{E}(Z_t^2) = \sigma^2$ is finite for any t. Now, if the time series window is very long and $\phi = 1$, we would obtain a variance of $t\sigma^2$. This means that the variance increases with the time point we are considering. But, if ϕ is smaller than 1 (and larger than -1), then the series would converge, and the variance is finite again. Intuitively, the series converges because the influence of each of the variances becomes extremely small since ϕ^{2t} becomes small for large t.

An ARIMA model is an extension of the ARMA model in that it allows for increasing (or decreasing) behavior. An ARIMA(1,1,1) process is defined as one where $(1 - B)X_t$ is an ARMA(1,1) process. Consider the following AR(1) process

$$X_t = X_{t-1} + Z_t$$

This is the AR(1) model with the parameter ϕ set to 1. We saw above already that this leads to increasing variance, and so is problematic. One way to handle this is to take differences (see main text and Appendix C1). Taking differences means we consider

$$X_t - X_{t-1} = (1 - B)X_t = Z_t$$

Since Z_t is white noise, we know we are able to determine statistical properties of this time series. The I in ARIMA comes from integrated (as in integral). If the difference is considered as differentiation, then the derivative is stationary but the series itself is integrated.

A2. Detailed discussion of the law of large numbers

We start with the classical situation where we are able to sample independent random variables X_j which are identically distributed. Suppose we can draw samples X_j from a common distribution with mean μ and variance σ^2, both finite. Using n independent and identically distributed (*iid*) samples we can estimate the mean μ by the average $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$. This works because all of the X_j are from the same distribution (identical) and were obtained independently. The law of large numbers now tells us that $\bar{X}$ converges to μ as $n \to \infty$. Intuitively, if we obtain enough observations we come arbitrarily close to the true mean (we only have to assume finite mean; Durrett, 2010). We immediately see that if the mean were to change from j to $j + h$ for some $h \neq 0$, then we have a problem because each X_j may have a different mean. Hence, the law of large numbers requires identical distributions. Now suppose that the means of X_j are equal to μ but the X_j are dependent, so that $\gamma_X(h) \neq 0$ for $h \geq 0$. If we use Chebyshev's inequality to show the weak law of large numbers, then we obtain

$$\mathbb{P}(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\mathrm{var}(\bar{X})}{\varepsilon^2}.$$

We see that in order for the law of large numbers to converge to 0, we require that $\mathrm{var}(\bar{X})$ converges to 0. This makes intuitive sense, because if the variance decreases to 0

then the mean will converge to a single point, here $\bar{X}$ converges to μ. Now (DasGupta, 2008)

$$\text{var}(\bar{X}) = \frac{1}{n}\sigma^2 + \frac{1}{n^2}\sum_{i\neq j}\text{cov}(X_i, X_j)$$

If the X_j are all independent, then $\text{cov}(X_i, X_j) = 0$ and $\text{var}(\bar{X})$ will converge to 0. However, if the X_j are not independent, then $\sum_{i\neq j}\text{cov}(X_i, X_j)$ need not converge and so the law of large numbers may not work. In general, we only need to assume that the dependence between any X_j and X_{j+h} is not too strong (i.e., is $o(n^{-2})$) so that the second term above still converges to 0 and we can apply Chebyshev's inequality.

We can make several different assumptions in order for $\sum_{i\neq j}\text{cov}(X_i, X_j)$ to converge. Stationarity is one way to do this. For instance, in an autoregressive model $X_j = \theta X_{j-1} + Z_j$, where Z_j is white noise, we obtain $\gamma_X(h) = \text{cov}(X_{j+h}, X_j) = \sigma^2\theta^{|h|}/(1-\theta^2)$. We then see that convergence is guaranteed if $|\theta| < 1$, since $\theta^{|h|}$ converges to 0 as $|h|$ increases. Another option is to assume that the time series $(X_j, j \in \mathbb{N})$ is asymptotically uncorrelated. This assumption says that $\text{cov}(X_{j+h}, X_j)$ converges to 0 at a sufficiently fast rate (i.e., $\leq \rho_h \text{sd}(X_{j+h})\text{sd}(X_j)$ with $\sum_h \rho_h$ finite and positive ρ_h and sd is the standard deviation, White, 2001). Intuitively, this means that the correlation between X_{j+h} and X_j converges to 0 as $|h|$ increases. So, the idea is similar to that of the autoregressive model. Another option is to assume that the time series $(X_j, j \in \mathbb{N})$ is mixing (Lehmann, 1999). A time series is mixing if

$$\begin{aligned}&\mathbb{P}((X_0, X_1, \ldots) \in A, (X_h, X_{h+1}, \ldots) \in B)\\ &\quad - \mathbb{P}((X_0, X_1, \ldots) \in B)\mathbb{P}((X_h, X_{h+1}, \ldots) \in A)\\ &\quad \to 0\end{aligned}$$

Intuitively, a time series is mixing if eventually (for sufficiently large lag h) the random variables become independent. If the time series is mixing then it is also ergodic (White, 2001). Take $A = B$ in the above, and we obtain $\mathbb{P}((X_j, j \in \mathbb{N}) \in A) = \mathbb{P}((X_j, j \in \mathbb{N}) \in A)^2$, which is true only if the probability equal 0 or 1. It is then also clear, that an ergodic (or mixing) time series need only be stationary eventually (after some time $n > n_0$).

A3. Detailed discussion of Central limit theorem

The central limit theorem is a stronger result than the law of large numbers but it also requires stronger assumptions. In fact, the central limit theorem implies the law of large numbers. In the central limit theorem we are interested in the convergence in distribution of the random variable $\sqrt{n}(\bar{X} - \mu)$. In the classical case of iid sampling, we obtain the central limit theorem as follows (Jacod & Protter, 2004). Let $(X_j, j \in \mathbb{N})$ be a sequence of independent and identically distributed random variables with mean μ and variance σ^2 (both finite). Then the random variable

$$\sqrt{n}(\bar{X}) \to N(0, v)$$

where v is the variance $\text{var}\left(\sqrt{n}\bar{X}\right)$. For an infinite time series $(X_j, j \in \mathbb{N})$ we therefore require that this variance converges (stabalises). If we go back to the variance above we now obtain

$$\text{var}\left(\sqrt{n}\bar{X}\right) = \sigma^2 + \frac{1}{n}\sum_{i\neq j}\text{cov}(X_i, X_j)$$

We see that we need even stronger conditions in order for the second term to converge (i.e., of order $o(n^{-1})$). If we assume stationarity then we obtain for the variance (Brockwell & Davis, 2009)

$$\text{var}\left(\sqrt{n}\bar{X}\right) = \sigma^2 + \sum_{h\neq 0}\left(\frac{n-|h|}{n}\right)\gamma_X(h)$$

Since the term $(n-|h|)/n$ is upper bounded by 1, we only need to worry about $\sum_h|\gamma_X(h)|$; if this sum is finite then we obtain convergence of $\sum_h \gamma_X(h)$. We have already invoked stationarity to make the result amenable to analysis, but we require more (in comparison to the law of large numbers) to get finite variance; stationarity is necessary but not sufficient to obtain a central limit theorem. One such assumption is finite independence or m-dependence (a form of mixing). A time series is *m-dependent* if the random vectors $(\ldots, X_{j-1}, X_j)$ and $(X_{j+m+1}, X_{j+m+2}, \ldots)$ are independent for all $j \in \mathbb{N}$ (DasGupta, 2008; Lehmann, 1999). The intuition is that as soon as the lag is m or larger, then the covariance between X_j and X_{j+m+1} is 0, and so the sum $\sum_h \gamma_X(h)$ will be finite. It can be shown that if we assume m-dependence, then $\sqrt{n}(\bar{X} - \mu)$ converges in distribution to a normal random variable with mean 0 and variance $\sum_h \gamma_X(h)$. Another way to allow limited dependence is by viewing the sum $\sum_j X_j$ as a martingale difference sequence (as a function of n). See Lehmann, 1999 for a good introduction to asymptotic analysis and more details.

A4. Prediction and uniqueness

In the main text we have described, using the example of a stationary AR(1) process and a random walk, why intuitively stationary processes are more amenable to forecasting than non-stationary process. The difficulty of forecasting a random walk process has to do with the effect that random shocks Z_t play in the non-stationary AR(1) process considered here. In the stationary case, at some point the effect of a random shock "washes out"; it no longer matters what size the shock received at time t is, at some time $t + h$ the shock no longer plays any part in determining the value of the process, and instead we can predict its value using the stationary distribution. In the random walk case, however, the effects of random shocks stay in the system forever. This means that, in order to make better forecasts of X_{t+h} using X_t than we discussed above, we would need to know the value of the shocks the system has received at intermediate time points $t + 1, \ldots, t + h$. However, since the values of *future shocks* are typically not known, this doesn't provide us with a useful way to build a forecasting model for the non-stationary process.

On a technical level, we can re-state this by saying that we would like to fit an ARIMA model to our observed time-series such that the model can be re-written so that future values of the process X_{t+h} are a function only of a finite number of past observed values of the process X_t. This is only possible for stationary AR processes, and this is

typically referred to as obtaining a "unique" or sometimes "causal" AR model (Brockwell & Davis, 2009). There are closely related restrictions for MA models that can be used for forecasting, known as obtaining an *invertible* MA model, which we explain below.

To interpret and predict using time series, only past observed values are useful. That is, if we want to explain measurements, it is undesirable to use future values of a time series. Also, it is impossible to predict future values of a time series using observations from the future or unobserved variables. Restricting a time series so that each future observations can be written in (predicted by) past observations is referred to as causality in AR processes and invertibility in MA processes (Hamilton, 1994).

We start with an AR process

$$X_t = \phi X_{t-1} + Z_t$$

Suppose that we have that $\phi > 1$. Then, we run into the problem that when recursively substituting for X_{t-1} the series we obtain eventually will not converge, i.e.,

$$X_t = \phi^t X_0 + \phi^{t-1} Z_{t-1} + \cdots + Z_t$$

does not converge because the effects of the Z_t are too large. If $|\phi| < 1$ (and $\sup_t \mathbb{E}|Z_t| < \infty$) the series will converge. But, as it turns out, we can rewrite the the series as

$$X_{t-1} = -\phi^{-1} X_t + \phi^{-1} Z_t$$

But now we have that $0 < \phi^{-1} < 1$ because $\phi > 1$, and so, the series

$$X_{t-1} = \phi^{-1} Z_t + \phi^{-2} Z_{t+1} + \cdots + \phi^{-t} X_0$$

does converge. Mathematically this seems reasonable. However, the interpretation of the series is that we are explaining the past X_{t-1} by future values of Z_t. This is neither practical not theoretically plausible. Theoretically, it is unnatural to use the future to predict the past (assuming that time has one direction, Reichenbach, 1956). In terms of prediction, this is also impossible, because we will never have future observations to use for prediction; that is the point of prediction. A model where we do not allow the future to predict the past (i.e., when $|\phi| > 1$) is called noncausal (Brockwell & Davis, 2009, Chapter 3). We, therefore, see that when rewriting the AR(1) process as an (infinte) MA process, we require that $|\phi| < 1$ to obtain meaningful and practical solutions.

We also have the problem the other way around. Suppose we have the MA process

$$X_t = Z_t - \phi Z_{t-1} = (1 - \phi B) Z_t$$

For prediction we want to use the observed values X_{t-1}, X_{t-2} etc., so that we can actually predict future values of X_t. The MA process above, however, has only past values of the unobserved process Z_t. But we can simply invert the difference operator $(1 - B)$ to remedy the situation. We obtain

$$(1 - \phi B)^{-1} X_t = Z_t$$

It turns out that we can rewrite the inverse difference operator as

$$(1 - \phi B)^{-1} = 1 - \phi B - \phi^2 B^2 - \phi^3 B^3 - \cdots$$

This implies that

$$(1 - \phi B)^{-1} X_t = X_t - \phi X_{t-1} - \phi^2 X_{t-2} \cdots = Z_t$$

When this inversion can be done, then the model is called invertible. So, only when it is possible to invert the difference operator, is it possible to give the model so that past observed values can be used to predict future values. The inversion $(1 - \phi B)^{-1}$ can only be obtained if $|\phi| < 1$. This coincides exactly with the stationarity assumption for models with $|\phi| < 1$.

In both AR and MA cases, we have seen that we require that the process is either causal or invertible. When an AR process is causal and when an MA process is invertible, we can interpret such models as explaining the future from past observations, which are useful for prediction.

Appendix B. Deterministic and stochastic mean trends

B1. Deriving the mean (determinstic)

As usual we use an AR(1) process to demonstrate. We have white noise Z_t, which has mean 0 and variance $\sigma^2 < \infty$. For the AR(1) process with determinisic trend βt we have the expectation

$$\begin{aligned}
&\mathbb{E}(X_t) = \mathbb{E}(\phi X_{t-1} + \beta t + Z_t) \\
&= \mathbb{E}(Z_t) + \phi \mathbb{E}(Z_{t-1}) + \cdots + \phi^t \mathbb{E}(Z_0) + \beta(t + \phi(t-1) \\
&\quad + \phi^2(t-2) \cdots + \phi^t 0) \\
&= \sum_{j=0}^{t} \phi^j \mathbb{E} Z_{t-j} + \beta \sum_{j=0}^{t} \phi^j (t-j) \to \frac{\beta}{1-\phi} t \quad (\text{as } t \to \infty)
\end{aligned}$$

where we recursively applied the AR process. When ϕ is less than one in absolute value, the infinite sum is a geometric series and converges on $\frac{\beta}{1-\phi}$. This shows that the system has a linear trend over time.

B2. Deriving the mean (stochastic)

That a random walk with drift produces a trend in the mean becomes clear if we represent the random walk model using recursive substitution for the X_{t-1} term:

$$X_t = \alpha + (\alpha + X_{t-2} + Z_{t-1}) + Z_t \tag{9}$$

$$X_t = \alpha t + Z_t + Z_{t-1} + \ldots Z_0 \tag{10}$$

where $t = 0$ represents some (hypothetical) initial time point. Since $\phi = 1$, the effects of shocks in the past Z_{t-s} remain in the system indefinitely. Since the Z_t terms represent mean-zero noise terms, we can see that $\mathbb{E}(X_t) = \alpha t$, that is, there is a linear trend in the mean produced by the random walk in combination with the drift parameter.

Unlike the deterministic example given above, this model also produces a process whose *variance* depends on time. The variance of X_t can be computed by taking the sum of the independent noise term variances on the right hand side, giving us $\text{var}(X_t) = t\sigma^2$.

Appendix C. Detrending and differencing

C1. Differencing a stochastic trend

Let Z_t be a white noise process, i.e., for all t, Z_t has mean 0 and $\text{cov}(Z_{t+h}, Z_t) = \sigma^2$ for $h = 0$ and 0 for $h \neq 0$. For a random walk (unit root process) we have

$$X_t = Z_0 + Z_1 + \cdots + Z_t$$

A random walk is non-stationary since the variance of a random walk is $\text{var}(X_t) = t\sigma^2$. Let $\nabla = (1 - B)$ be the differencing operator, where $BX_t = X_{t-1}$. Then, the difference

$$\begin{aligned}(1 - B)X_t &= X_t - X_{t-1} \\ &= (Z_0 + Z_1 + \cdots + Z_t) - (Z_0 + Z_1 + \cdots + Z_{t-1}) \\ &= Z_t\end{aligned}$$

is certainly stationary since Z_t is stationary.

A similar procedure can be applied to a random walk with drift. Consider what happens when we take the difference $X_t - X_{t-1}$ of such a process. Substituting the recursive expression for X_t and X_{t-1} we get

$$\begin{aligned}X_t &= \alpha t + Z_0 + Z_1 + \cdots + Z_t \\ -X_{t-1} &= -\alpha(t-1) - Z_0 - Z_1 - \cdots - Z_{t-1} \\ \Delta X_t &= X_t - X_{t-1} = \alpha + Z_t\end{aligned}$$

which shows that the differenced time-series can be modeled as a stationary white-noise process (i.e., an autoregressive model with $\phi = 0$) and with intercept α.

C2. Differencing a (trend-)stationary process

Although differencing is the correct method of dealing with a stochastic trend, differencing should not be applied to stationary time series, or those which exhibit a deterministic trend. Consider the case of a stationary AR(1) model, so that

$$X_t = \phi X_{t-1} + Z_t.$$

Then, applying the differencing operator, we obtain

$$(1 - B)X_t = \phi(1 - B)X_{t-1} + (1 - B)Z_t.$$

Note that a mean α in the AR(1) model would also be removed by differencing, since $(1 - B)\alpha = 0$. And so

$$X_t - X_{t-1} = \phi(X_{t-1} - X_{t-2}) + Z_t - Z_{t-1}$$

Or, written using $\nabla = (1 - B)$, we obtain

$$\nabla X_t = \phi \nabla X_{t-1} + Z_t - Z_{t-1}$$

For $|\phi| < 1$ we then obtain an ARMA(1,1) process with respect to the difference process ∇X_t. However, the MA(1) process introduced by differencing has a unit root (Hamilton, 1994, Section 15.3), which means that it is not invertible (see Appendix A4). On an intuitive level, the differencing in this case introduces unneccesary complexity into the model.

Similar problems occur if differencing is applied to a process which exhibits a deterministic trend. If a linear deterministic trend βt were present in the AR(1) model, we get $(1 - B)\beta t = \beta t - \beta(t - 1) = \beta$. This implies that if we were to apply differncing on an AR(1) model including a linear trend, we would obtain an ARMA(1,1) model with mean β and a unit root MA part. Clearly, for an AR(1) process including a linear trend the correct way to remove the trend is simply subtract it.

C3. Detrending a stochastic trend

If we had a unit root process $X_t = Z_0 + Z_1 + \cdots + Z_t$, with Z_t white noise, then, as we saw above, differencing works. However, if we were to subtract a linear trend (detrending, as if we thought there were a linear trend), then we do not fully succeed in obtaining a stationary process. To see this, consider the unit root process $X_t = \alpha + X_{t-1} + Z_t$. By iteration we see that

$$\begin{aligned}X_t &= \alpha + (\alpha + X_{t-2} + Z_{t-1}) + Z_t = \cdots \\ &= t\alpha + Z_0 + Z_1 + \cdots + Z_t\end{aligned}$$

Clearly, subtracting αt, that is detrending, will solve the issue with the mean, because $\mathbb{E}(X_t - t\alpha) = 0$. However, it is still a unit root process and so the variance of the process is $t\sigma^2$ (Hamilton, 1994, Section 15.3). As seen above, differencing will remove the unit root process and render the process stationary.

Summarizing, if we have a (linear) trend model and difference, we introduce a unit root. If we have a unit root process and detrend, then we will not have removed the unit root process, which increases the variance of the process with time.

Appendix D. Hypothesis testing with unit roots

Many hypothesis tests rely on the central limit theorem. So do tests on regression parameters, like in the AR(1) model. Suppose we have

$$X_t = \phi X_{t-1} + Z_t$$

where Z_t is white noise. Then standard regression techniques (see, e.g., Brockwell & Davis, 2009, Chapter 8, or Hamilton, 1994, Chapter 8) can be applied to obtain estimates (ordinary least squares or method of moment estimators). Then it can be shown (given regularity conditions, see Appendix A3) that we obtain convergence (as $t \to \infty$) in distribution to a random normal variable when scaled by $\sqrt{t}$

$$\sqrt{t}(\hat{\phi} - \phi) \xrightarrow{d} N(0, (1 - \phi)^2)$$

where $N(a, b)$ is a Gaussian random variable with mean a and variance b. Recall that a unit root process (e.g., random walk) is obtained when $\phi = 1$ in the AR(1) representation above, because

$$X_t = \phi X_{t-1} + Z_t = \phi Z_{t-1} + \phi^2 Z_{t-2} + \cdots + \phi^t Z_0$$

And if $\phi = 1$ we obtain a random walk, where the influence of each new Z_t remains (is not diminished by ϕ). The asymptotic result of convergence to a Gaussian random variable shows that when $\phi = 1$ we get into trouble because convergence will be to a fixed value 0, since the variance will be 0 when $\phi = 1$. This suggests that the rate of convergence is faster than $\sqrt{t}$ near $\phi = 1$.

Dickey and Fuller (1981) show that with scaling t there is still convergence to a random variable, but a rather odd-looking one

$$t(\hat{\phi}-\phi)\xrightarrow{d}\frac{1}{2}\frac{W(1)^2-1}{\left(\int_0^1 W(r)^2dr\right)^{1/2}}$$

where $W(r)$ is Brownian motion ($W(0)=0$, $W(j)-W(i)$ are $N(0,j-i)$ for $j>i$, and $W(r)$ is continuous in r with probability 1).

Noteworthy of this random variable is that $W(1)=W(1)-W(0)$ and so is a normal random variable $N(0,1)$, and hence $W(1)^2$ a $\chi^2(1)$ random variable. This random variable has most mass, approximately 2/3, to the left of 1. And so, the estimate $\hat{\phi}$ in the AR(1) model will often be less than the true value $\phi=1$.

With this result Dickey and Fuller (1981) developed a test and tabulated critical values for significance testing. A standard "t"-test is obtained with

$$t=\frac{(\hat{\phi}-1)\left(\sum_{j=1}^{t}y_{j-1}^2\right)^{1/2}}{\hat{\sigma}_\phi}$$

where $\hat{\sigma}_\phi$ is the estimate of the standard error for the parameter estimate $\hat{\phi}$

$$\hat{\sigma}_\phi^2=\frac{1}{t-1}\sum_{j=1}^{t}(y_j-\hat{\phi}y_{j-1})^2$$

This test for the unity of ϕ is the standard unit root test by Dickey and Fuller (DF) which does not allow for serial correlation in the residuals.

D1. The augmented Dickey–Fuller test

The augmented Dickey–Fuller test is about taking into account the possibility that the residuals U_t could be serially correlated. That is, the residuals U_t and U_{t+h}, for some $h>0$ are correlated. This is induced for instance by an AR process. Suppose that we have an AR(2) process

$$X_t-\phi_1X_{t-1}-\phi_2X_{t-2}=Z_t$$

where Z_t is white noise. We can rewrite this using the back shift operator as

$$(1-\phi_1B-\phi_2B^2)X_t=Z_t$$

and so, if the inverse exists,

$$X_t=(1-\phi_1B-\phi_2B^2)^{-1}Z_t=U_t$$

It turns out that this U_t is serially correlated and is similar to a random walk (see, for instance, Hamilton, 1994, Chapter 17).

Dickey and Fuller used this to devise a general test for unit roots, where serial correlation is allowed to exist. This is called the augmented Dickey-Fuller (ADF) test. We start from the last equation and rewrite this as

$$[(1-\rho B)-\zeta_1B(1-B)]X_t=Z_t$$

with $\rho=\phi_1+\phi_2$ and $\zeta_1=-\phi_2$. Then we obtain the equation

$$X_t=\rho X_{t-1}+\zeta_1\nabla X_{t-1}+Z_t$$

where $\nabla=1-B$. Here we have a representation with ∇X_{t-1} stationary and X_{t-1} is integrated (similar to a random walk, see Appendix A1). And so, with this representation, only the parameter ρ needs to be tested for a unit root, i.e., $H_0:\rho=\phi_1+\phi_2=1$. This reparametrisation can be generalized to higher order models (e.g., AR p). See Hamilton (1994, Chapter 17) for general AR(p) models and other details of the ADF test.

Appendix E. Performance of ADF-based tests

In the main text Section 3.4 we present the results of a simulation study investigating the performance of the model-known and model-selection based DF test procedures with a fixed set of generating parameters, amongst those, $\phi=.5$ and noise SD $\sigma=0.5$. In this appendix we present supplementary results obtained by replicating the simulation study but varying either a) the noise standard deviation $\sigma\in[0.25,0.75]$, or b) the autoregressive parameter $\phi\in[0.25,0.75]$.

Figure 8 shows the performance of both testing procedures when the noise SD is varied. Overall we see a very similar pattern of results as shown in the main text. The main difference of note here is the performance of the sequential testing procedure (right hand-panels) for the random walk with drift. In panel (d), we see that increasing the noise SD appears to accentuate the performance "dip" that we noted in the corresponding panel (b) of Figure 6, with the performance here continuing to degrade up to sample sizes of $n_t=200$, and not yet reaching acceptable levels at sample sizes of $n_t=500$. Recall that this dip was explained in the main text as driven by our failure to reject the over-specified DF model; with more noise variation in the system, it appears as though the sequential procedure needs an even larger sample size to reliably reject this model, resulting in poorer performance. On the other hand, as we see in panel (b) of Figure 8, lower innovation variance leads to relatively better performance: Although it takes a relatively larger sample size (around $n_t=150$) for the sequential method to work well, relative to the stationary AR(1) and deterministic linear trend models, there is no performance "dip" with increasing sample size evident in this condition. Varying the sample size grid to be more fine tuned in the $25<n_t<150$ range also failed to produce a noticeable performance dip.

Figure 9 shows the performance of the DF testing procedures when varying the auto-regressive parameter to higher ($\phi=.75$) or lower ($\phi=.25$) stationary values than used in the simulation depicted in the main text ($\phi=.5$). Note that here only the conditions of a stationary AR(1) and deterministic linear trend are replicated, as a random walk has by definition $\phi=1$.

Here too we see a very similar qualitative pattern of results across all conditions, with performance typically poor at very small sample sizes of $n_t=25$ and increasingly relatively quickly with increasing sample size. As discussed in section 3.4, we can see that both procedures perform relatively worse for values of ϕ closer to 1 and relatively better for values of ϕ closer to 0.

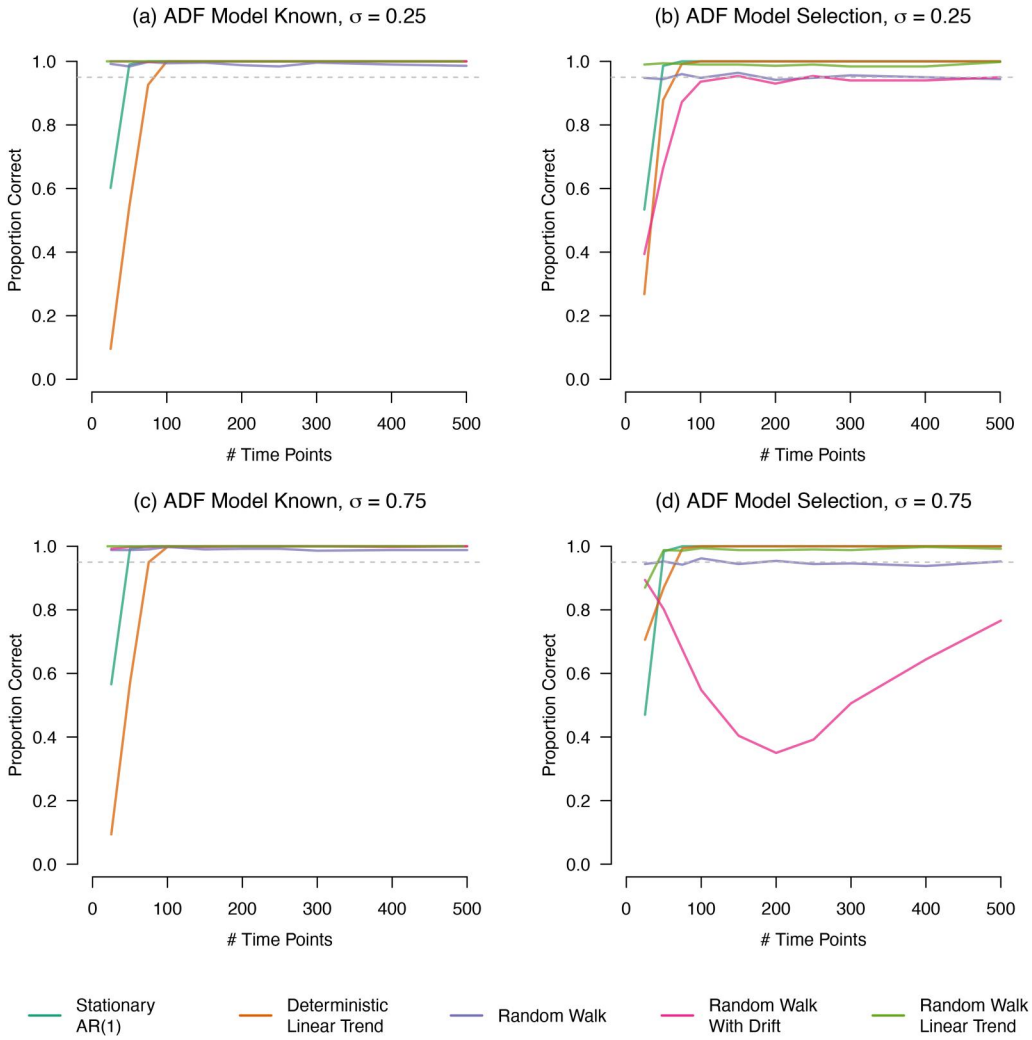

Figure 8. Left Panel: Performance of the DF test when the appropriate DF model is known. Right Panel: Performance of the sequential DF testing procedure which relies on selection of the correct DF model. Top Panels: Results obtained with a noise SD of 0.25. Bottom Panels: Results obtained with a noise SD of 0.75.

Appendix F. Detrending in a multilevel context

We illustrate the effects of detrending in a multilevel time-series context using a small simulation study. In this study we simulated time-series of 100 different units with 50 observations each, over 1000 iterations. In each iteration, each unit's time-series is governed by an AR(1) process with a deterministic trend. For each unit we draw an auto-regressive parameter from a normal distribution with mean $\mu_\phi = 0.3$ and standard deviation $\sigma_\phi = 0.1$, i.e., $\phi \sim N(0.3, 0.1)$, and a slope parameter β which determines their deterministic trend, again from a normal distribution $\beta \sim N(0.2, 1)$. On average units exhibit a positive trend, but by choosing a relatively large variance for the trend term, we ensure that some participants exhibit more strongly positive trends, while others exhibit weaker negative trends.

In each iteration, we examine estimates of the auto-regressive fixed effect $\hat{\mu}_\phi$ obtained from the simulated data under four different pre-processing regimes: (1) the raw data (2) data which has been centered around each participants within-person empirical mean $\overline{Y_i}$; (3) data which has been detrended using the sample-average trend (i.e., between-person detrending) and (4) data which has been detrended within-person, around each individuals estimated trend. In each case, estimates are obtained using the R package *lme4* (Bates et al., 2015), while trend estimates are obtained by fitting linear growth curve models to either the full dataset or individual-specific data, for the between-person and within-person detrending conditions respectively.

The results of this simulation are shown in Figure 10. As we would expect, the raw data analysis yields highly biased estimates of the auto-regressive fixed effect, driven by the strong average positive trend in the data, and these estimates show very little variability across conditions. Similarly biased estimates are obtained when the data is centered around the within-person mean, and even when the data is detrended based on the average trend. In fact, this latter mode of detrending yields more highly biased parameter estimates than the raw data analysis in this condition. Detrending the time-series per unit before analysis, however, yields the most promising result. Notably, while far outperforming the alternatives, the parameter estimates in this condition do still exhibit a small bias toward zero (with an average parameter estimate of $\hat{\phi} = 0.272$). This is likely

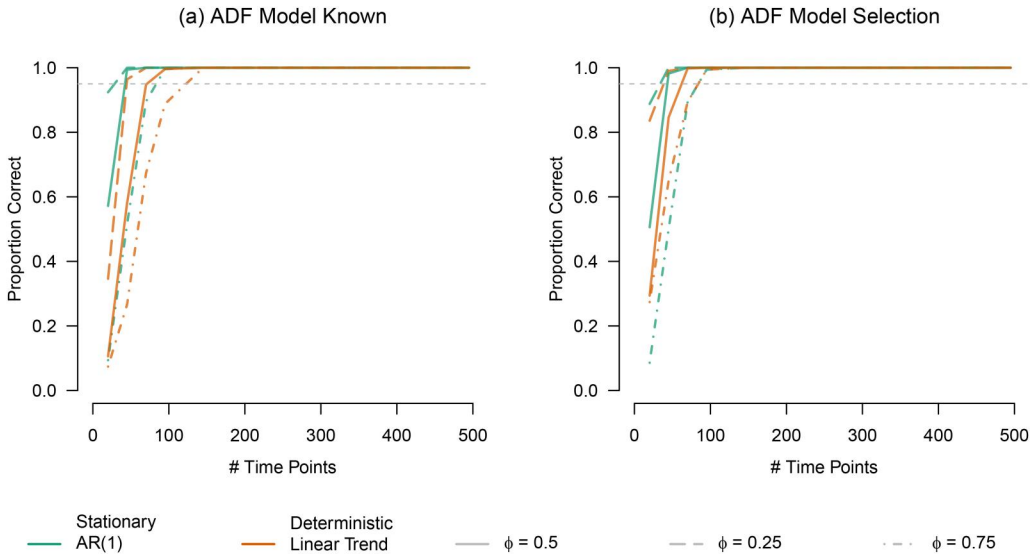

Figure 9. Left Panel: Performance of the DF test when the appropriate DF model is known. Right Panel: Performance of the sequential DF testing procedure which relies on selection of the correct DF model. Here results are only shown for conditions which depend on the autoregressive parameter φ.

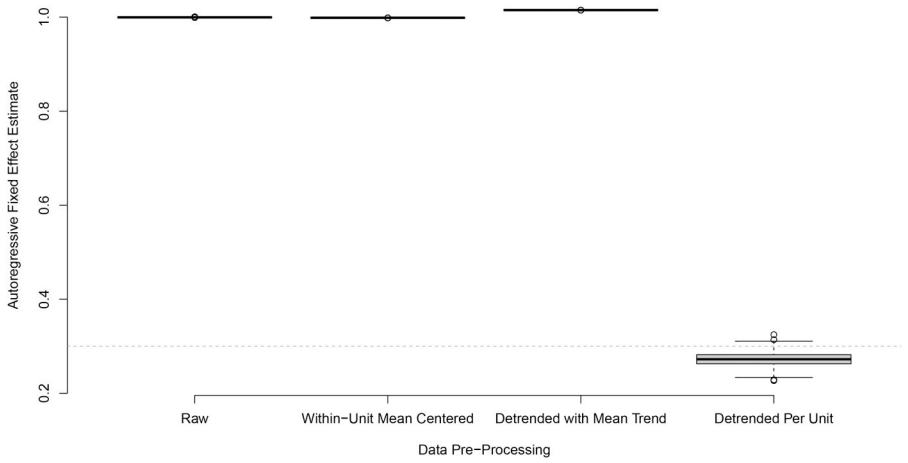

Figure 10. Boxplot showing for each method the distribution of autoregressive fixed effect estimates $\hat{\mu_\phi}$ based on the simulation described in the main text. The true parameter $\mu_\phi = 0.3$ is shown by the dashed gray line.

to be caused by the fact that the time-series is centered around the sample-estimated trend, rather than the true trend, resulting in a phenomena analagous to Nickell's bias, which has been discussed in detail in the multilevel time series modeling literature with regards within-unit centering of stationary time-series (Asparouhov et al., 2018; Asparouhov & Muthén, 2019; Nickell, 1981).

Appendix G. Fitting time-varying AR(1) model to AR(1) models discussed in the paper

Here we generated time series from the four models shown in Figure 1 and estimated time-varying AR models using the two methods using a kernel-smoothing or a Generalized Additive Model (GAM) approach discussed in Haslbeck, Bringmann, et al. (2021). These time-varying VAR models contain each of the four models as special cases and should therefore in principle be able to recover the data generating models.

The first row of Figure 11 shows the results for the stationary AR(1) model. The first column shows a time series with $N = 1000$ time points, on which we estimated the time-varying AR model. We see that both methods estimate the ϕ parameter to be a function of time. This is incorrect, since this parameter is not in fact a function of time. However, some error is expected in finite samples, and the estimates are close to the true value of $\phi = 0.5$.

The second row shows the results for the random walk. We see that while both methods estimate ϕ to be close to 1, they fail to recover the true constant parameter $\phi = 1$. The intercepts/drift parameter which is zero in the data generating model is strongly time-varying, and is capturing the random walk in the data.

In the third row, we see the results for the linear deterministic trend. Here, the kernel-smoothing method

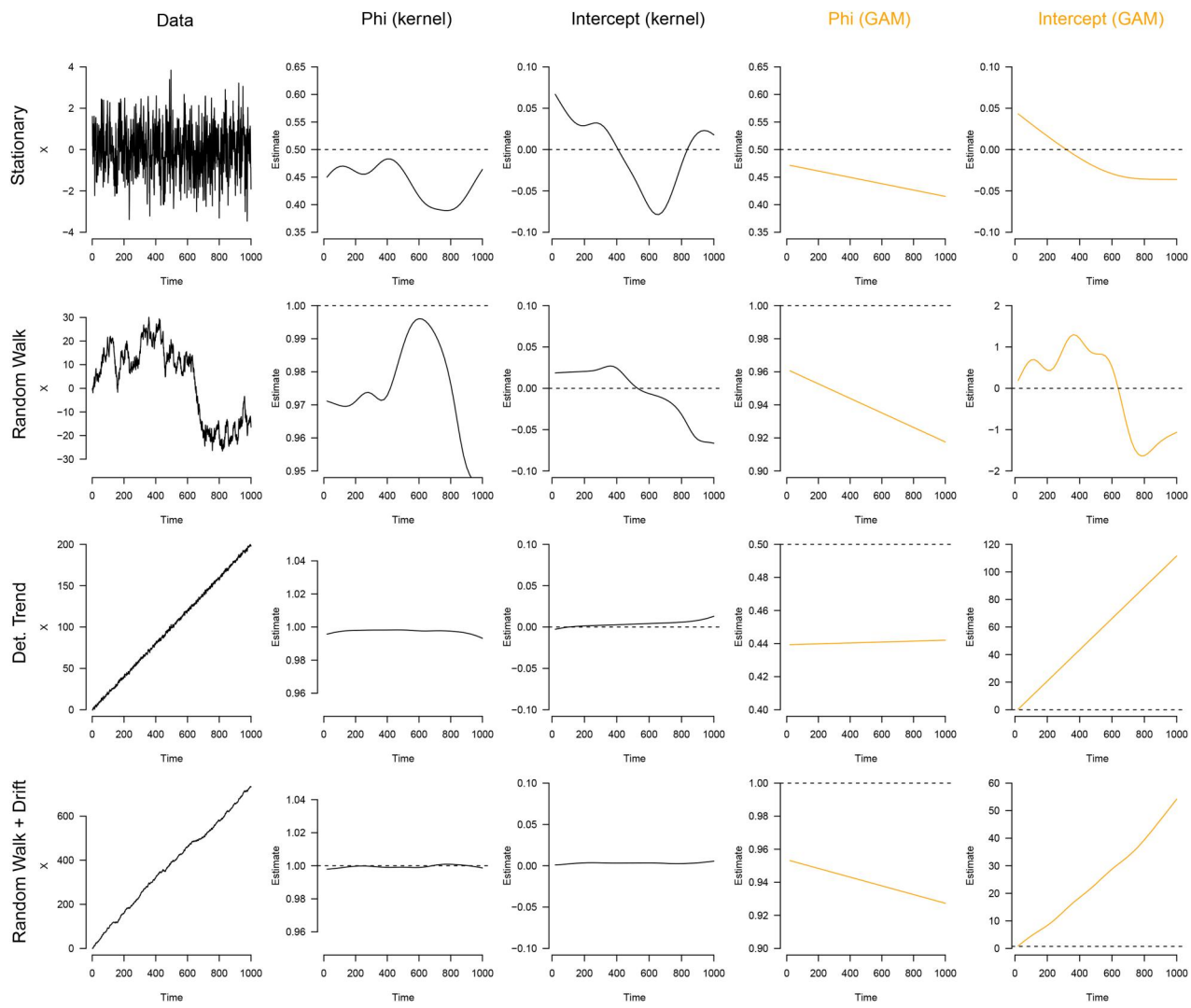

Figure 11. The rows show the four AR(1) models shown in Figure 1: a stationary model, a random walk, a deterministic trend, and a random walk with drift; The columns show the generated data, and the estimated time-varying intercept/drift and ϕ-parameters of the kernel-smoothing and the GAM-method.

estimates a ϕ close to 1, while the true parameter is $\phi = 0.5$. The intercept is essentially zero. In contrast, the GAM method slightly underestimates ϕ and models the trend using a time-varying intercept. Since the deterministic trend is the same as a time-varying intercept, the GAM method therefore better recovers the data generating model in this case.

Finally, the last row shows the results for a random walk with drift. Here, we see that the kernel method seems to correctly pick up a constant $\phi = 1$, however, it does not pick up the intercept/drift term. The GAM-method, however, does not pick up the constant $\phi = 1$ but instead models the observed trend with a time-varying intercept.

These results show that wile time-varying methods will provide good summaries of the data, they are not generally able to recover the data generating mechanism, at least not for moderate sample sizes like the one used in this illustration.

Appendix H. Vector autoregressive (VAR) model

In relatively simple settings, each variable in the VAR is treated similarly as we have described above. However, more interesting cases are the ones in which variables are interacting across time lags. For instance in this model $X_{1,t}$ does not only depend on itself at the previous time point $X_{1,t-1}$, but also on $X_{2,t-1}$:

$$X_{1,t} = \phi_1 X_{1,t-1} + \phi_2 X_{2,t-1} + Z_{1,t}$$
$$X_{2,t} = \rho X_{2,t-1} + Z_{2,t}$$

This VAR model represents the situation in which X_1 is determined by its past and X_2, and X_2 is determined only by its past. These kinds of models are extremely useful to capture interaction effects between time series (Dablander et al., 2020b). They are also often used in network analyses where connections between nodes in a network are represented by the coefficients ϕ_2, in the above example (e.g., Epskamp et al., 2018; Haslbeck & Waldorp, 2020).

But not all VAR models are a simple extension of the univariate case. Suppose we have two variables and each is an AR model

$$X_{1,t} = \phi X_{2,t-1} + Z_{1,t}$$
$$X_{2,t} = X_{2,t-1} + Z_{2,t}$$

Since $\rho = 1$ in the second time series, there is a unit root process while the first is determined by the second time series and white noise $Z_{1,t}$. However, simply differencing both variables simultaneously will fail to properly account for the stochastic trend, because we then obtain for the first time series a unit root. The remedy instead is to take $X_{1,t} - \phi X_{2,t-1} = Z_{1,t}$ as this is a stationary sequence. A process like this, where there is a linear combination of the variables such that the combination is stationary, is called a cointegrated time series. This example makes it clear that a simple generalization to vector autoregressive time series is not always possible, but many cases can be handled with more complex approaches. See for more details and examples (Hamilton, 1994, Chapter 19).

Appendix I. Generalized autoregressive conditional heteroskedastic

In the Generalized Autoregressive Conditional Heteroskedastic (GARCH) model we make innovation variances dependent on earlier states of the time series. A GARCH model has mean 0 but the deviations are dependent on previous time points. This makes the model have different variance at different time points. Typically, a GARCH model has alternating periods of small fluctuations and periods of large fluctuations. It is therefore sometimes called the volatility model (Engle, 1982). A GARCH(1,1) model is defined by

$$X_t = \sigma_t Z_t$$
$$\sigma_t^2 = \alpha + \phi\sigma_{t-1}^2 + \theta X_{t-1}^2.$$

Given certain assumptions the mean of this GARCH(1,1) model is 0 and the variance changes with time. Because of the autoregressive model for the variance of the model, the regimes of large and small fluctuations alternate. An example of where this may be interesting is in psychopathology. During episodes not only will the mean value of affective variables (likely) change, but also the variation around these means will be larger.