To cite this article:

Adolf, J. K., & Ceulemans, E. (in press). Improved estimation of autoregressive models through contextual impulses and robust modeling. *Psychological Methods*.

Improved estimation of autoregressive models through contextual impulses and robust modeling

Janne K. Adolf[1] & Eva Ceulemans[1]

[1] KU Leuven - University of Leuven

Postprint, March 12, 2025; accepted for publication in *Psychological Methods*.

Author Note

Janne K. Adolf & Eva Ceulemans, Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven - University of Leuven, Belgium.

This manuscript is accompanied by the following materials: the Supplementary Materials document, the *R* package *simARO* enabling *R*-level reproducibility of the raw simulation results, the *use_simARO* *R*-project supporting the usage of the *simARO* package, the raw simulation results, and the analysis code for the empirical illustrations. All of these components are accessible via the Open Science Framework (OSF) repository *Supplementary materials for manuscript: Improved estimation of autoregressive models through contextual impulses and robust modeling* (Adolf & Ceulemans, 2023); The data used for the empirical illustrations can be found in the Open Science Framework repositories *One does not simply correct for serial dependence: Supplementary materials* (Ariens, 2023).

This work was supported by a research fellowship from the German Research Foundation (DFG) awarded to Janne Adolf, and by research grants from the Fund for Scientific Research-Flanders (FWO, Project No. G.074319N) and the Research Council of KU Leuven (C14/19/054 and iBOF/21/090) awarded to Eva Ceulemans. Some resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

Excerpts (of earlier versions) of this work were presented at the 2021 Society for Ambulatory Assessment Conference in (virtual) Zurich, Switzerland, the 2022 International Meeting of the Psychometric Society in Bologna, Italy, the 2023 Society for Ambulatory Assessment Conference in Amsterdam, the Netherlands, and the 2023 10th European Congress of Methodology in Gent, Belgium.

The authors would like to thank Ruben Crevits and Christophe Croux for sharing their code, and Kristof Meers for his support with using the VSC and setting up the GitLab infrastructure accompanying this work. This document was created using R Markdown and papaja (Aust & Barth, 2022).

Correspondence concerning this article should be addressed to Janne K. Adolf, Tiensestraat 102, 3000 Leuven, Belgium. E-mail: janne.adolf@kuleuven.be

Abstract

The aim of the dynamic paradigm of affect research is to characterize daily life affective processes by means of intense longitudinal data and dynamic, typically autoregressive-type models. The contextual conditions accompanying and potentially influencing affective processes obviously form an important part of the picture. Especially distinct contextual events – ranging from salient daily events to major life events – are regularly assessed and their effects modelled. Such an explicit approach to studying context can however reach its limits, if relevant events are hard to define, measure or model. In that case one finds oneself in a situation where contextual events play out as hidden contaminators possibly obscuring the affective process of interest. Interestingly, specific forms of such contamination can also have beneficial effects in that they can trigger autoregressive dynamics and leverage their estimation. In this paper we take a closer look at this phenomenon focusing on events that exert short-lasting main effects on affective processes. We also demonstrate that robust autoregressive models have the capacity to outperfom classical models given this and other forms of contextual contamination as they not only capitalize on positive leverage effects but also mitigate the negative obscuring effects contextual events might have.

Keywords: affective processes, contextual influences, outliers, autoregressive modeling, robust estimation

Improved estimation of autoregressive models through contextual impulses and robust modeling

1 Introduction

The dynamic paradigm of affect research conceives of daily-life affective phenomena as continuously changing within-person processes. The properties of such processes, especially the (ir-)regularities they display over time, are typically referred to as *affect dynamics*. Studying those affect dynamics is supposed to reveal principles of affect (dis-)regulation, which might in turn forward the understanding of (mal-)adaptive long-term development (Bringmann et al., 2016; Hamaker & Wichers, 2017; Kuppens & Verduyn, 2015). To assess affective processes, one typically relies on the experience sampling method and thus repeatedly gathers self-reports of a person's momentary affective states (Ebner-Priemer & Trull, 2009). The result are so-called intensive longitudinal (IL) affect data. A class of statistical models then often employed to estimate affect dynamics from IL affect data is the one of autoregressive (AR) models. These models regress present affect states onto past ones, yielding estimates of AR parameters that quantify to what extent momentary - possibly extreme or out-of-balance - affective states are lasting in a person's affective experiences. AR parameters have therefore been mapped onto the concept of 'emotional inertia' (Kuppens, Allen, & Sheeber, 2010; Suls, Green, & Hillis, 1998), which is in turn believed to be a possible marker of 'regulatory weakness' (De Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2016) associated with negative developmental trajectories (Kuppens et al., 2012).

When it comes to affect and affective processes in general, it is almost trivial to state *that* context matters. Here we distinguish two stances on context and *how* it might matter. First, context might be conceived of as a "sum of (infinitely many) internal and external events, which are somehow relevant for this particular process (...) [such as] conversations (...) with colleagues, peers, or family members; (...) information the person was confronted with through books, media, and e-mails; and the appraisal, thoughts and memories experienced by the person" (Hamaker, 2012, p. 54). This sum of events forms a stable and continuous stream of input that is likely to continuously impact an affective process, in one direction or the other, to a larger or a lesser extent. Contextual influences of this sort may even be viewed as an integral part of affective processes. In fact, this is already reflected in the very definition of emotions as quick and ongoing reactions to changing contextual demands that facilitate adaptive behavior (Aldao, 2013; Cole, Martin, & Dennis, 2004; Gross, 1998). Likewise, when it comes to modeling, it is commonly assumed that the dynamic errors or innovations of AR models readily accounts for this omnipresent contextual stream (Hamaker, 2012). One might additionally include measurement errors to capture continuous contextual influences that do not affect the process of interest but only the measurements taken at each time point (e.g., situational influences

that leads to errors when filling in electronic diaries, Schuurman, Houtveen, & Hamaker, 2015).

Second, and this contrasting stance underlies the present paper, context might also come and matter in the form of discrete events that stand out in that they happen at limited times and are somewhat important. Examples warranting this perspective range from salient daily events such as arguments or gatherings with friends to major life events such as job loss or the birth of a child. Such *distinct contextual events* will likely also have *distinct effects* on affective processes, clearly exceeding the effects of the before-mentioned contextual stream. Studies therefore often collect contextual information and model it explicitly, as covariates exerting main effects on affective processes or even moderating their dynamics (Adolf, Voelkle, Brose, & Schmiedek, 2017; Aldao, 2013; Brose, Schmiedek, Lövdén, & Lindenberger, 2011; Koval et al., 2015; Zautra, Berkhof, & Nicolson, 2002).

It can however be complicated to study contextual events and their effects in such an explicit way. Specifically, relevant events can be hard to define, measure or model: What to do when it is unknown which events matter for a particular affective process (e.g., beyond the popular practice to probe for the occurrence of negative and positive events, which event categories should be considered)? What if events cannot be reliably assessed (e.g., retrospective reporting of events might be affected by momentary affective states, Brose, Lindenberger, & Schmiedek, 2013) or not assessed at all as one does not want to overburden participants (Eisele et al., 2022)? What to do when it is unknown how events take effect (e.g., immediately or delayed, similarly or differently across different events from the same category, such as negative events) and thus how to incorporate them into statistical analyses? Under these circumstances, it is not straightforward to think of and go about events as one or multiple *covariates* when conceptualizing, designing or analyzing IL studies and data. Rather, it seems more appropriate to conceive of events as contextual *contaminators*: That is, they can obscure the affective process of interest by affecting it at unknown times and in unknown ways. Speaking in statistical terms, events can cause *outliers* that contaminate the process of interest.

At first sight, this take sounds not exactly like a very optimistic perspective. A closer look however reveals that, under the dynamic paradigm, contextual contamination can even be beneficial: Certain outliers, more specifically extreme innovations called innovation outliers, can, if caused by the above described distinct contextual events, function as 'natural' impulses in AR systems. As such, their direct impact is obscuring, but their indirect impact is informative in the sense that they cascade through the system triggering its dynamics and leveraging its estimation via AR models. Innovation outliers can thus improve estimation of AR models - and we can capitalize on them further by using robust as opposed to classical modeling strategies. Robust methods - like classical ones - benefit from the leverage effects of innovation outliers, but -

unlike classical ones - they in addition mitigate the direct, disrupting effects of outliers. Also, they are supposed to handle other types of outliers that might be less beneficial, most typically measurement outliers that come about through extreme measurement errors. Measurement outliers might also be caused by distinct contextual events in case these affect the measurement of an affective process and not the process itself. Rather than triggering the system's dynamics, measurement outliers only impact momentary observations and can thereby obscure dynamics quite substantially, as we will show later on.

In this paper, we thus take a look at the above specified contextual outliers in AR systems, their potential to trigger AR dynamics, and to improve AR model estimation especially if robust versus classical modeling approaches are used (Rousseeuw, 1991). We thereby put our main focus on occasional contextual events with short-lasting effects and confine the paper to the basic case of observing and modeling one individual AR system at a time (i.e., the $N = 1$ case). That outliers can have beneficial effects in AR systems is not a new finding from the perspective of the robust statistics literature, where the phenomenon is occasionally discussed (e.g., Croux & Joossens, 2008; Maronna, Martin, & Yohai, 2006; Zoubir, Koivunen, Chakhchoukh, & Muma, 2012). Also, robust approaches to AR modeling are established in the field. We therefore think that it is all the more important that these insights and methods finally get introduced to psychological researchers who approach affective phenomena under the dynamic paradigm. With the present paper, we lay a strong emphasis on an accessible presentation of the subject matter as well as on providing open materials, including the functionality to fit the considered modeling approaches bundled in an *R* package.

The paper's structure is as follows: In Section 2, we recapitulate basic univariate first-order AR systems, their behavior and classical estimation approaches. In Section 3, we turn to the situation where such systems are contaminated with contextual outliers. We define contextual outliers reconsidering the two outlier types mentioned above, innovation and measurement outliers. We demonstrate their effects on AR systems and on classical estimation thereof. We then explain the rationale behind robust estimation and introduce as well as illustrate three different robust modeling techniques for AR systems. In Section 4, we present results from a comprehensive simulation study, in which we compare the performance of the considered robust approaches to classical ones in the absence and presence of contextual contamination. In Section 6 we close with a discussion of the conceptual as well as the simulation contributions of the paper.

2 Autoregressive systems

2.1 The basic autoregressive system

We focus on a simple AR system that is often applied in affect research: A univariate AR process of first order (AR(1), Hamilton, 1994). We can express such a system via the following equations holding for time points $t = 1, 2, ..., T$:

$$\eta_t = \alpha + \beta * \eta_{t-1} + \zeta_t \qquad \text{with} \qquad \zeta_t \overset{iid}{\sim} N(0, \psi) \tag{1}$$

$$y_t = \tau + \eta_t + \varepsilon_t \qquad \text{with} \qquad \varepsilon_t \overset{iid}{\sim} N(0, \theta) \tag{2}$$

In the first equation - the *dynamic* equation - the process of interest on the left-hand side, η_t, gets decomposed into a deterministic and a random part. The random part, ζ_t, represents an error process that constantly perturbs the process of interest, also called dynamic error or innovation process. Across time, the innovations are assumed to be normally distributed with constant variance ψ, serially independent and independent of the process of interest up to and including time $t - 1$. As mentioned in the introduction, the innovation process can be thought of as reflecting the impact of those contextual factors that together form a stable and continuous contextual stream (Hamaker, 2012). The deterministic part of the dynamic equation, $\alpha + \beta * \eta_{t-1}$, consists of a constant or dynamic intercept, α, and, the regression onto the process of interest at earlier time-points, here $t - 1$, via the AR parameter $\beta \in (0, 1)$. The AR parameter is usually the parameter of main interest, as its size determines to what extent process states and hence the innovations that had previously affected them are carried forward through time. That is, AR parameters with values closer to 0 imply more transient process states whereas AR parameters with values closer to 1 imply rather lasting states.[1] As mentioned in the introduction, the AR parameter has therefore been interpreted as reflecting 'emotional inertia' or 'regulatory weakness' in the context of affect research (De Haan-Rietdijk et al., 2016; Kuppens, Allen, et al., 2010), as it indicates the extent to which a system shows prolonged states of emotional imbalance, rather than a quick recovery and return to an emotional "home base" or equilibrium (i.e., the processes mean level, Kuppens, Oravecz, & Tuerlinckx, 2010). The dynamic intercept α finally is the parameter that uniquely relates to this equilibrium of an AR(1) process.

The second equation of the AR model constitutes a *measurement* equation. It defines an observable

[1] Note that the AR parameter is directly meaningful as an effect size measure as it is equivalent to the first-order autocorrelation in an AR(1) model.

process, y_t, that is reflective of the process of interest η_t, but possibly also affected by a measurement intercept τ and a measurement error process, ε_t. Note that unless we are concerned with multiple indicators of η_t the measurement intercept is usually assumed to be zero and thus absent from the measurement equation. We retain it here because it provides one way of introducing measurement outliers (see Section 4). The measurement errors are, much like the innovations, assumed to be normally distributed with constant variance θ, serially independent and independent of the process of interest and the innovations at all time points. Unlike innovations, however, measurement errors have no influence on the process of interest and thus also no lasting effect on the system. While the measurement error process can thus also be viewed as reflecting a stream of continuous contextual influences, these only concern the level of measurement and are only of momentary effect.

2.2 Demonstrating the system's behavior

One can demonstrate the dynamic behavior of an AR(1) system, and hence the implications of an AR parameter of a particular size, by means of data generated from the above model equations 1 & 2. Such data can be visualized in at least two ways, as Figure 1 - we focus solely on Panels A and B in this section - shows: A trajectory plot displays the sampled process states (on the y-axis) over time (on the x-axis) and a regression plot depicts present process states (i.e., y_t; on the y-axis) relative to lag-one process states (i.e., y_{t-1}; on the x-axis). A trajectory plot thus highlights the temporal properties of the generated data, while a regression plot illuminates the relationship between past and present system states aggregated over time.

To be able to directly observe the mean-reverting behavior related to the AR parameter, we first sample from and display the AR(1) system under very controlled conditions. Specifically and as displayed in Figure 1A, we apply a single controlled impulse to the dynamic equation, while assuming that there are no innovation or measurement error effects. Following this single impulse, we can thus observe the so-called impulse response of the system (Hamilton, 1994). Looking at the trajectory plot to the left, we see that, at first, void of any input, the AR(1) process resides at its mean value - as marked by the dashed grey line - and is hence "in balance". The impulse hits after a few time points, instantly driving the process away from its mean. Afterwards we observe an exponentially decaying return to the mean, the exact rate of which is related to the exact size of the AR parameter. The decay would thus happen faster for a less autostable system and slower for a more autostable system than the one shown here (with an AR parameter of size $\beta = 0.80$). Note that the observation with the added impulse is displayed as a rhombus rather than a circle, and the impulse's decaying impact is emphasized via a gradual color scale, decaying from orange to dark blue. These visual attributes become helpful in the regression plot, where temporal order of observations is no longer visible. Here we see that the data points following on the one with the added impulse are perfectly

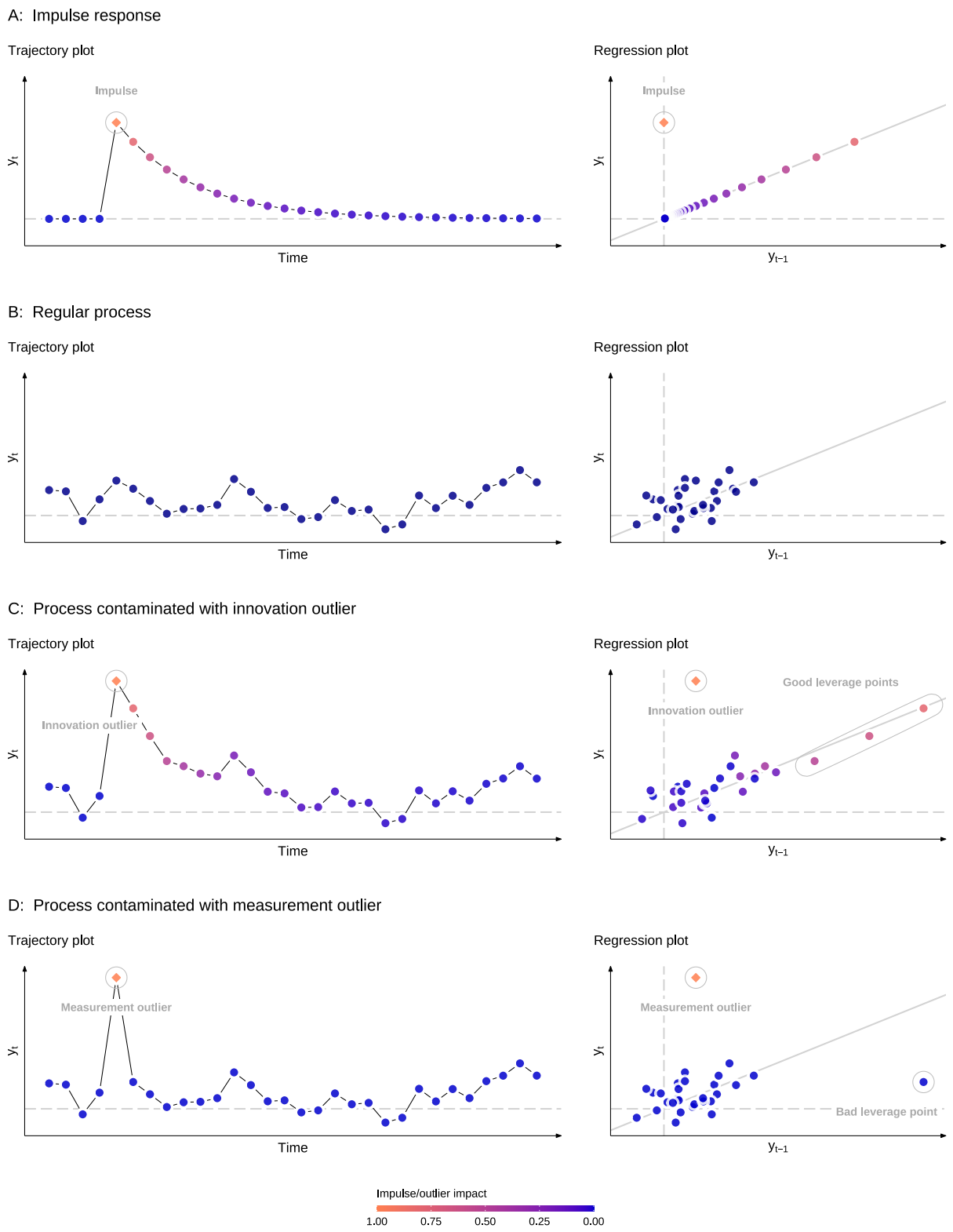

Figure 1
Behavior of AR(1) systems under different conditions. Panel A displays the system's response to a single controlled impulse and Panel B shows regular data sampled under uncontrolled conditions. Panel C and D reveal the impact of an innovation outlier and a measurement outlier, respectively.

in line with the true regression relationship, which is marked by the solid grey line.

Although quite insightful, an impulse response function of course refers to a highly idealized and artificial scenario that we will never be able to observe with psychological processes in the real world. Figure 1B therefore complements the impulse response with a visualization of the system's behavior under 'normal', uncontrolled conditions. That is, there are no impulses and the innovation and measurement error processes unfold unconstrainedly according to the above modeling assumptions. As expected from the relatively large AR effect size, we observe a somewhat smoother trajectory that shows prolonged departures from the mean (Hamilton, 1994). Also the true regression relationship is discernible to some extent in the regression plot. But there is also randomness due to dynamic and measurement error effects. This is why we usually do not try to identify the size of an AR parameter by visual inspection of the collected data, but rather rely on formal AR modeling. The following subsection hence deals with this.

2.3 Classical estimation

A modeling approach that readily accommodates the AR(1) system shown above is the one of (linear) state space modeling (SSM) - a multivariate latent variable modeling framework similar to the popular structural equation modeling (Chow, Ho, Hamaker, & Dolan, 2010; Harvey, 1989). Unlike the also popular but much more basic linear regression framework, which can also be employed for AR modeling, SSM thus allows to account for measurement error effects (Schuurman et al., 2015). In addition, it also entails measurement models with multiple indicators, and generally extended model structures involving any (linear) vector AR moving average model.

SSM affords maximum likelihood (ML) estimation of model parameters via the Kalman filter (KF). The KF is a recursive algorithm to optimally estimate the states of - among others - AR(1) processes over time. Starting from an initial process state estimate, it generates predictions of future states and updates these based on the actual observations at the respective time points and hence the committed prediction errors (i.e., discrepancy between predicted and observed states). These prediction errors in turn allow to construct a specific form of the likelihood function, the prediction error decomposition form (Harvey, 1989), which enables relatively fast estimation of AR(1) model parameters if time series from single individuals are analysed. Both problems, process state and parameter estimation, are intertwined in an iterative process as the KF relies on parameter estimates to produce state estimates, and the state estimates in turn serve to generate and update the parameter estimates. To this end this process is started by providing parameter and KF starting values and repeated until convergence is reached. For information on how we implement SSM using *R* (R Core Team, 2022) see Section 4.2.4.

Obviously, an AR(1) model without measurement error effects can also be fitted via SSM by restricting the measurement error variance to zero. Such a restricted SSM model is then comparable to an AR(1) model as formulated in linear regression modeling, which typically relies on ordinary least squares (OLS) instead of ML estimation. It has however been shown that both estimation approaches yield similar results and are equivalent asymptotically (Krone, Albers, & Timmerman, 2017). We will draw on this similarity later when comparing "classical" SSM with zero-constrained measurement error variance to robust linear regression modeling techniques. Note that the label "classical", which comes from the robust statistics literature, simply means that an estimation approach has not been robustified against possible outliers and their effects.

3 Autoregressive systems with outliers

3.1 Contextual outliers and their effects

For IL data stemming from AR systems, the robust statistics literature distinguishes two major types of outliers: Innovation and measurement (also called additive) outliers (Zoubir et al., 2012). Both gain their name from their presumed origin in an AR system: While innovation outliers stem from innovations that are more extreme than expected, measurement outliers originate from unexpectedly extreme measurement errors. A gathering with friends or an argument with one's partner might be events causing innovation outliers and according *contamination* in the innovation (and the corresponding affective) process. Being unexpectedly interrupted while filling in an electronic questionnaire might lead to an unintended extreme response and hence to a measurement outlier or contaminated measurement of affect.

Returning to the system equations (Equations 1 & 2) presented in Section 2 we might adapt them in the following way to represent contaminated innovation and measurement error processes:

$$\eta_t = \alpha + \beta * \eta_{t-1} + \zeta_t^c \tag{3}$$

$$y_t = \tau + \eta_t + \varepsilon_t^c \tag{4}$$

The added superscript c in ζ_t^c and ε_t^c indicates that the respective error process is contaminated with a few extreme realizations due to the impact of events. We have argued in the introduction that it might very well be unknown how events take effect, complicating the formulation of a model explicitly incorporating them. This was then one of the very reasons to recommend embracing the 'events as contaminators'-perspective and the use of robust modeling approaches (for model estimation), as these do not require specific assumptions about *how* contamination with possibly event-caused outliers comes about. To make the idea of contextual outliers more tangible, however, we can think of the following equations as representing *one*

possibility of how events can cause contextual outliers:

$$\zeta_t^c = \zeta_t + \gamma_1 * x_{1_t} \tag{5}$$

$$\varepsilon_t^c = \zeta_t + \gamma_2 * x_{2_t} \tag{6}$$

Here we introduce two time-varying variables x_{1_t} and x_{2_t}, which are binary and function as event-indicators. Each of these thus assumes values of zero for most time points and values of one for those few time points at which events occur. These events then impact the respective error process with relatively large effects γ_1 and γ_2. In Equations 5 and 6 we make a number of assumptions: Per error process, we allow for one event at a time, which then has a constant, deterministic effect. This is most plausible if we are concerned with one event type (e.g., an argument) re-occuring over a not too long period of time so that its impact is likely not to change. We do however not make assumptions regarding the occurence of the events. They can be indepedent or dependent over time, independent or dependent of the innovation process (e.g., an argument might be more likely if negative affect has already been pushed up due to other situational influences), and occur with constant or changing probability. If we assume events to occur independently over time and of the error processes, and with constant probability, we could for instance think of a binomial distribution as event-generating. We stick to this model (see also Croux & Joossens, 2008) in the following when outliers are *generated*, and discuss other ways in which events might affect AR systems in Section 6.

Speaking of contextual 'outliers' or 'contamination' not only implies that the focus of attention lies primarily on the affective process and not so much on the event itself; the terms also come with negative connotations. However, as we have foreshadowed in the previous sections, outliers can have beneficial effects in AR systems, since they can trigger the system's dynamics. To demonstrate this, let us again take a look at simulated data.

In Figure 1C, we observe the impact of an *innovation* outlier. The displayed data were generated by adding a value of 5 to the innovation of time point 5. The contaminated observation is again shown as a rhombus and the innovation outlier's impact via gradual coloring. Comparing Panel C to Panel A, it becomes rather obvious that an innovation outlier functions much like the earlier-discussed impulse, but in this case it is added to the regular process. Hence, innovations and measurement errors are constantly present, but the exponentially decaying impact of the outlier-based impulse remains discernible in the trajectory plot. In the regression plot, we again see the immediate effect of the outlier-based impulse, resulting in a data point outlying in the vertical direction (i.e., taking on a typical value for y_{t-1}, but a rather extreme value for y_t). In the robust statistics literature such a data point is called a *vertical outlier*

(Rousseeuw, 1991). Note that this classification is no longer about how an outlier was generated, but rather about the location of a data-point in the variable space depicted in the regression plot. Because it is caused by an innovation outlier, the vertical outlier is carried forward through time, following the dynamics of the system and hence resulting in multiple outstanding data points that reflect the true lagged relationship. In the variable space these data points are termed *good leverage points* (Rousseeuw, 1991). The term *leverage points* refers to data points that are potentially influential during estimation, because they are relatively far away from the majority of the data in the horizontal direction. How this influence plays out then depends on whether the leverage points additionally reflect the regression relationship in the majority of the data. In this case they do and are therefore called *good* leverage points.

In contrast, Figure 1D visualizes the impact of a *measurement* outlier. Here the data were generated by adding a value of 5 to the measurement error of time point 5. Unlike innovation outliers, however, measurement outliers concern the measurement equation of the model only. They are thus not carried forward through the system and can also not trigger its dynamics, as the trajectory and regression plots reveal. Instead, they even have harmful effects on estimation by introducing so-called *bad leverage points* following upon the initial vertical outlier. In the regression plot we thus observe a data point outlying in the horizontal direction, when the extreme value that used to be y_t becomes the observed value at time $t-1$ at the next time point and the new value y_t is again an uncontaminated process state. This outlier has *bad leverage* because it is outlying in the horizontal direction only and thus far from the regression line underlying the majority of the data. It will therefore tilt an estimated regression line quite a bit. But we turn to concrete estimation results in the following.

3.2 Classical estimation under contamination

We already noted that classical estimation methods such as ML estimation in the context of SSM are not designed to handle outliers and are therefore usually not robust towards their effects. However, we have also just demonstrated that outliers can have very different effects, especially when the *indirect* effects of innovation and measurement outliers are concerned. Recall that one innovation outlier causes several good leverage points as opposed to one bad leverage point caused by one measurement outlier. Good leverage points generally improve the accuracy with which a regression relationship is estimated - especially improved precision is mentioned in the robust statistics literature (Croux & Joossens, 2008; Dehon, Gassner, & Verardi, 2009; Maronna et al., 2006; Zoubir et al., 2012). This holds independent of the estimation method used. As a consequence, also classical estimation methods can show a net benefit from innovation outliers even though they are affected by the vertical outlier (Zoubir et al., 2012). Bad leverage points on the contrary distort estimation, especially increasing bias (e.g., Dehon et al., 2009; Zoubir et al., 2012).

To illustrate this, we fit a classical AR(1) model to simulated AR(1) data that are uncontaminated as well as contaminated, and compare the results in Table 1. The data-generating parameter values are similar to the ones that gave rise to the data in Figure 1 (i.e., $\beta = 0.80$, $\alpha = 0$, $\psi = 0.50$, outlier impact = 10), but we sample a larger number of time points ($T = 80$), and introduce a higher proportion of contamination (i.e., 10 % outliers, meaning that 8 time points are contaminated). For simplicity, we also assume no measurement error effects (i.e., $\theta = 0$) and thus also do not model them. Note that the differently contaminated data sets are all generated with exactly the same initial process values and dynamic error processes, which means that differences in modeling results can be attributed to differences in contamination. The first column of Table 1 contains the type of contamination, the second recapitulates the data-generating AR effect size, and the third and forth column labelled "classical" finally show the obtained results for the AR parameter. Note that we focus on this parameter due to its importance for psychological interpretation, but intercept estimation can also be affected by outliers.

The columns labelled "robust" contain the results from different robust modeling approaches, and we will turn to these in the following subsection when introducing the robust procedures. The remaining columns with label "true" result from fitting a model that mirrors the outlier-generating model presented in Equations 3 to 6. This model incorporates the perturbing events and their effects on either the dynamic or measurement process in the correct way (for the exact implementation see Section 4.2.2). This model is not included as a realistic analysis model, but as a benchmark model: By drawing upon crucial data-generating information it is expected to outperform all other models considered (that do not incorporate that information). The true model will thus provide a best-performance limit that the remaining models will approach to a larger or lesser extent. We use it in this anecdotal example as well as in the simulation study.

For now we focus on classical ML estimation in the context of SSM: The point estimate obtained for the uncontaminated data equals $\hat{\beta}_{ML} = 0.596$ with a standard error (SE) estimate of $SE(\hat{\beta}_{ML}) = 0.086$ and

Table 1
AR parameter estimates obtained via classical and robust modeling.

		classical		robust						true	
Outliers	β	$\hat{\beta}_{ML}$	(SE)	$\hat{\beta}_{cML}$	(SE)	$\hat{\beta}_{MM}$	(SE)	$\hat{\beta}_{MTL(\alpha=0.2)}$	(SE)	$\hat{\beta}_{ML(true)}$	(SE)
No	0.80	0.596	0.086	0.596	0.086	0.668	0.088	0.782	0.061	0.596	0.086
Innovation	0.80	0.801	0.067	0.801	0.067	0.782	0.016	0.804	0.013	0.783	0.015
Measurement	0.80	0.054	0.111	-0.098	0.111	0.870	0.058	0.617	0.096	0.600	0.086

Note. The following parameters values are used to generate the here-fitted data next to $\beta = 0.8$: $\alpha = 0, \sigma^2_\zeta = 0.5, \sigma^2_\varepsilon = 0$. The outlier impact is of size 10 and the data comprise 80 time points. Selected data and modeling results are also visualized in Figure 2.

resulting 95% confidence interval (CI) of $CI(\hat{\beta}_{ML}) = (0.425, 0.767)$. From the time series literature we know that this result is affected by finite sample bias, which draws AR parameter estimates towards zero, especially for larger AR effect sizes (see Krone et al., 2017).

Contamination with innovation outliers moves the AR parameter point estimate closer to its true value ($\hat{\beta}_{ML} = 0.801$). It also decreases the size of the estimated SE ($SE(\hat{\beta}_{ML}) = 0.067$), so that the corresponding CI becomes $CI(\hat{\beta}_{ML}) = (0.668, 0.934)$. Overall this is an improvement in the modeling solution's accuracy in this particular example. Contamination with measurement outliers instead clearly deteriorates things: The point estimate lies far from the uncontaminated estimate and the true AR effect size ($\hat{\beta}_{ML} = 0.054$) and the CI does not cover the true value ($CI(\hat{\beta}_{ML}) = (-0.167, 0.275)$) although it has also widened ($SE(\hat{\beta}_{ML}) = 0.111$). If the aim of such an application was to estimate a person's emotion dynamics, then interpretations of such a contaminated result would clearly be misleading.

3.3 Robust estimation under contamination

Although classical methods can benefit from the consequences of specific forms of contamination (i.e., good leverage points in the case of innovation outliers) they do not actually account for contamination. This is where robust modeling approaches come into play. Their rationale takes a starting point not too different from the one of classical statistics: It is assumed that we have an idea regarding what is going on in the collected data, and hence know a suited model - the AR(1) model in the present context. However, unlike in classical statistics, this model is not assumed to hold for all the data, but only the majority. We thus accept that there might be a few observations that for some reason behave differently, but it is unknown which ones and in what way they behave differently (Zoubir et al., 2012). These observations are thus called contaminated observations or outliers and can have large effects on results from classical modeling (e.g., see the above anecdotal results). To mitigate such effects, robust methods detect and weight down outliers during estimation to obtain modeling results that are close to the ones one would have obtained with uncontaminated data (Rousseeuw & Hubert, 2011). On the downside, robust methods are less efficient than classical ones if data are uncontaminated, because they do not use the full information available in the data and thus produce more variable estimates. How data points are detected and weighted - and thus also how robustness and efficiency trade off - varies between robust methods, as we will see in the following. Note again, however, that robust methods make no specific assumptions about how outliers are generated.

We employ three robust procedures in the present paper: Cleaned maximum likelihood (cML) estimation (Zoubir et al., 2012), maximum trimmed likelihood (MTL) estimation (Crevits & Croux, 2019), and MM[2] estimation (Koller & Stahel, 2011; Rousseeuw & Hubert, 2011). Whereas the two likelihood-based

[2] The full name of the estimation method is indeed 'MM'. Neither the original paper by Yohai (1987) nor more recent (overview)

approaches are implemented in a SSM framework, MM estimation is implemented in a linear regression modeling framework. Before returning to Table 1 and the results obtained with these estimation procedures, we briefly explain their principles as well as the motivation for selecting them. For information on how the three robust procedures can be implemented in *R* (R Core Team, 2022) see again Section 4.2.4.

First, MTL estimation in a SSM context has been put forward recently (Crevits & Croux, 2019) and is a natural candidate to include since we rely on SSM in the present paper. As explained in Section 2.3, SSM is an already quite powerful modeling framework for AR-type systems, but the proposed robustifications extend it even further. Specifically, Crevits and Croux suggest to combine a robust KF with one of two robust ML estimators, among them the MTL estimator. MTL estimation boils down to ML estimation with a trimmed likelihood function. That is, the estimator considers only the $1 - \alpha$ fraction of data points that are the most likely under the model, and thus excludes the α fraction of the least likely data points during estimation. In the SSM context this is achieved by excluding those data points for which the KF generates the most extreme prediction errors as measured by the Mahalanobis distance between observed and predicted states (Crevits & Croux, 2019, p. 1697). The process of trimming is thereby integrated into the iterative estimation procedure that we described in Section 2, subsection 2.3. The trimming proportion α needs to be set to a value between 0 and 1 a priori to model estimation. There are rules of thumb as to how large α should be. For example, Chan and Wei, who discuss MTL estimation of autocorrelations (which should not be too different from estimating AR(1) parameters), suggest setting $\alpha = 0.01 - 0.02$ for "general series", $\alpha = 0.03 - 0.05$ for "medium contaminated series" and $\alpha = 0.06 - 0.1$ for "heavily contaminated series" (Chan & Wei, 1992, p. 153). Unfortunately the authors leave open what "medium" or "heavily" contaminated times series are, and we will see in the following that the actual amount of contamination matters for how appropriate a chosen value of α is. Turning to the robust KF also proposed by Crevits and Croux, it is important to note that this robustification can only take effect in situations in which measurement error is modeled. To understand this we revert to the classical KF, which, if measurement error is assumed to be present and its variance is estimated, takes this estimate into account. More specifically, the larger the measurement error variance estimate, the less weight will the prediction errors and thus the actual observations get in updating the process state estimates. The observations are thus treated as less trustworthy if larger measurement error effects are expected, and the KF relies more on its own predictions. The robustification presented in Crevits and Croux further adapts this measurement error-related weighting such that the impact of *unexpectedly extreme* observations is damped additionally. This obviously concerns the estimation of process states, but also has impact on parameter estimation since both estimation processes

papers provide another full name.

are inseparably intertwined. If the measurement error variance is estimated as small, however, more weight will be put on the actual observations relative to the predicted states. In the extreme case of no measurement error effects, the KF always updates its state estimates to the observed states at a particular time point, meaning that the data are fully trusted. In these cases there is little to no space for the robustification of the KF, and in the latter case one falls back on using the MTL estimator alone. Since we constrain this paper to scenarios of small or no (estimated) measurement error effects, we do not employ the robust KF. Interested readers can however consult the supplementary materials for more information on the robust KF.

Second, cML estimation is included since the underlying principle of "cleaning" is often used to handle outliers in applied research (e.g., Bos et al., 2022; Koval et al., 2015; Malooly, Genet, & Siemer, 2013). The estimation procedure is less involved than MTL estimation and simply encompasses regular ML estimation applied after the data have been cleaned from presumably contaminated data points using the $3 - \sigma$ rule (Maronna et al., 2006). This means that observations lying outside an interval formed by the observed mean minus and plus three times the observed standard deviation (SD) are replaced by missing values. ML estimation in the context of SSM then readily handles those missing values (Chow et al., 2010; Enders, 2013). MTL and cML thus have in common that data points are weighted in a discrete fashion during estimation (i.e., observations are either included or excluded, weights take on values of 0 or 1), but they differ considerably in how the weighting is determined. In MTL estimation the weighting is realized in a model-based way, in cML estimation it is carried out in an ad-hoc and model-naïve fashion. Furthermore, the mean and SD used for cleaning are of course also estimated from the potentially contaminated sample. In consequence, using classical estimators at this stage can impair the subsequent cleaning and thus the final performance of the method (Rousseeuw & Hubert, 2011; e.g., outliers might be masked by inflated SD or mean estimates, uncontaminated data points might be discarded or 'swamped' as outliers, Zoubir et al., 2012). Here we still use classical mean and SD estimators during the preprocessing of the data, because this corresponds to common practice implementations of the $3 - \sigma$ rule (Zoubir et al., 2012).

Third and lastly, MM estimation has been proposed to robustify estimation in the context of linear regression modeling (Koller & Stahel, 2011; Rousseeuw & Hubert, 2011; Yohai, 1987). It is included here since AR(1) models without measurement error can also be - and are often - handled as simple regression models. MM estimation is a comparably sophisticated multi-stage estimation procedure that combines different robust techniques to optimize estimator performance (Zoubir et al., 2012). This then manifests in a high breakdown point (i.e., MM estimators are said to handle larger proportions of contamination) and high efficiency (Maronna et al., 2006; Zoubir et al., 2012). Like in the previous cases with α and $3 - \sigma$ there are tuning parameters to be set a priori to determine the shape of the weighting functions involved. Since their

function and setting is less obvious in this case we stick to the tuning parameter setting advised in the literature (Koller & Stahel, 2011; Maechler et al., 2021). Unlike in the previous cases, data weighting during MM estimation is continuous, resulting in individual weights that assume values between 0 and 1.

Coming back to the exemplary estimation results presented in Table 1, we see that also the robust methods benefit from the introduction of innovation outliers. This is not so much reflected in the point estimates - these are similar across all (classical and robust) methods - but more so in the SE estimates. And these change differently across robust methods: While the cML SE estimates change in the same way as the classical ML ones do (up to the number of decimal places shown), we observe a much larger decrease in SE size for MM and MTL estimation with trimming proportion $\alpha = 0.2$. In fact, these latter reduced SE sizes are comparable to the SE size obtained under the true model used as a benchmark here. The picture changes when measurement outliers are introduced. Here the cML point estimate ends up close to zero, which is again similar to the classical ML estimate. The MTL and MM point estimates are much less affected, with the former being closest to the true model's result and to the classical result under the absence of contamination. Finally, measurement outliers increase the SE estimate sizes for all methods except for MM estimation, for which the SE estimate, surprisingly, is about halved in size. [3]

Figure 2 further elucidates these results. Panel A is dedicated to contamination with innovation outliers and Panel B to contamination with measurement outliers. The top row of each panel displays the data in the same fashion as in Figure 1, that is, in terms of a trajectory and a regression plot, with contaminated observations highlighted, and the outlier impact visualized. The following rows refer to the different estimation approaches. On the one hand they show the obtained results incorporated in the regression plots in terms of the fitted regression lines. The solid black lines are thereby based on the point estimates of the intercept and slope for each method and the dashed black lines are based on the point estimates plus/minus the corresponding SE estimates. The grey solid regression lines continue to mark the true regression relationship. On the other hand, the different weighting schemes employed during estimation are visualized. The observed trajectories are therefore over-plotted with lollipop plots showing data weights (y-axis) per observation and thus time point (x-axis). Observations are crossed out if they are excluded (weight of 0, concerns cML and MTL estimation) or if they are flagged as outliers which means they get a weight close to zero (MM estimation). In the regression plots we do not visualize the continuous weighting and only keep the information on excluded or flagged observations.

[3] Note that we calculated the SE estimates for the MTL solutions analogously to how regular classical SE estimates are calculated. That is, they are based on the (numerically approximated) Hessian matrix at the modeling solution. It is unclear to what extent these SE estimates are accurate and should be used for inference. We therefore also investigate SE estimator performance as part of our simulation study. However, while this concerns the exact values of the SE estimates, the general pattern we highlight in this example agrees with what we will see later when looking at the simulated estimator variances.

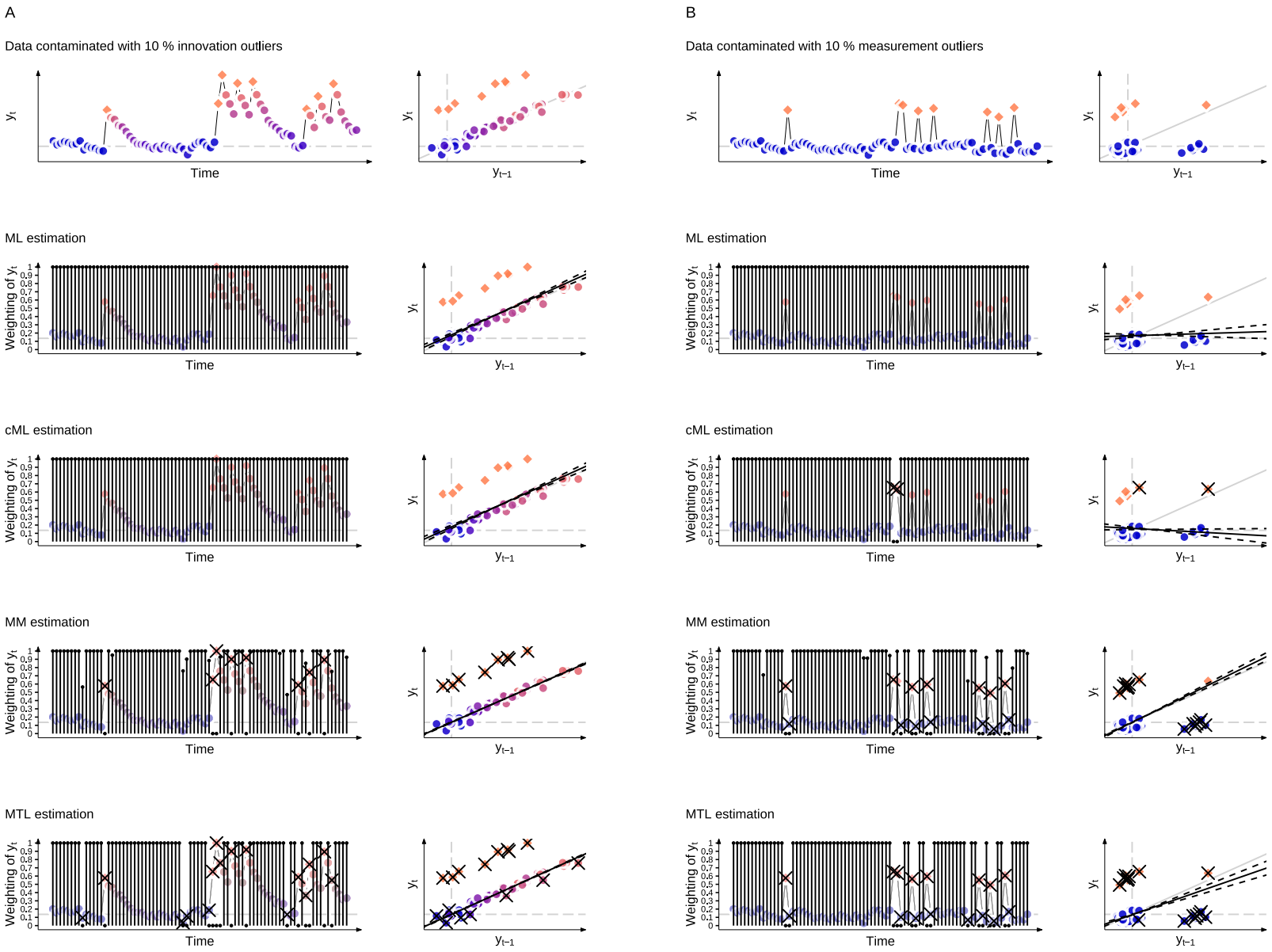

Figure 2

Outlier handling of different robust estimation methods. Panel A is devoted to innovation outliers and their effects, Panel B to measurement outliers. Data are displayed in the top row of each panel in terms of a trajectory and a regression plot. Contaminated observations, outlier impact and true regression lines are visualized as in 1. The remaining rows show the data overplotted with the results obtained from ML, cML, MM and MTL estimation. Fitted regression lines are displayed as solid black lines, dashed black lines indicate the regression lines deviating with one SE estimate per parameter from the fitted ones. Weighting schemes employed during estimation are visualized via lollipop plots (y-axis of trajectory plots). Crossed-out observations are either excluded or flagged as outliers during estimation.

Looking at Panel A of Figure 2, it becomes clear why the cML results do not differ from the classical ML results. All data points have weights of one, although the outlier effects are rather pronounced in this example. As mentioned this phenomenon is called masking and due to the data cleaning itself being distorted by the outliers. The MM estimator, in stark contrast, flags all eight vertical outliers and weights down a few more data points to a limited extent. The MTL estimator also detects all vertical outliers and excludes an additional eight data points because the trimming proportion is set to $\alpha = 0.2$, which amounts to 16 data points in this example. Interestingly these latter excluded data points are the same ones that are somewhat downweighted during MM estimation.

Switching to Panel B, and thus to data contaminated with measurement outliers, we see that the cML estimator now excludes two of the vertical outliers, but none of the bad leverage points, allowing them to excert their large influence. The MM estimator flags almost all vertical outliers and all bad leverage points as outlying. The one outlying data point that is retained with a high weight is special, because it actually is a combination of the vertical outlier at that time point and a bad leverage point that resulted from the vertical outlier at the previous time point. For the settings chosen here this combination accidentally produces a good leverage point. This data point thus probably explains the surprising reduction in SE size displayed in the table. The MTL estimator performs comparably in that it excludes all bad leverage points, but also all vertical outliers. Since two of these coincide, there is one regular data point additionally excluded. Note that the trimming proportion is crucial here. Had we set it to a lower value (e.g., $\alpha = 0.1$) then that could have led to a solution subtantially affected by bad leverage points.

At this point it is in order to again emphasize that the results we have presented so far are of illustrative nature. They do not support precise statements regarding the performance of the different methods more generally, for instance regarding the amounts of bias and variance expected for the considered point estimators. Achieving that is instead the purpose of the now following simulation study.

4 Simulation study

4.1 Aims

The aim of this simulation study is two-fold: First, we want to confirm the beneficial effects of innovation outliers on estimating AR(1) systems. With this we follow up on the literature on robust statistics and especially robust AR modeling, where the phenomenon is occasionally mentioned (e.g., Croux & Joossens, 2008; Maronna et al., 2006; Zoubir et al., 2012). Second, we aim to compare the performance of the above-introduced robust approaches to the classical one in the absence and presence of contamination with innovation or/and measurement outliers. For a more differentiated picture, we vary a number of

parameters used for data generation and model fitting, including the strength of the AR dynamics, the frequency with which outliers occur and the robustification of some of the estimators.

4.2 Design

4.2.1 Data generation

Figure 3 provides a schematic display of the data-generating models used. It shows that we employ two population-level model structures, an AR(1) process without measurement error and an AR(1) process with measurement error. Measurement error influences are reflected by the non-zero, but small measurement error variance (i.e., $\theta = 0.1$). We opt for a small variance, because we want to study the effects of measurement outliers - which suggest the presence of a baseline measurement error process - but we also want to compare models of which some cannot accommodate measurement error. By restricting measurement errors to vary closely around zero and thus minimizing their impact, we aim to mitigate effects of model missspecifcation on our simulation results. In both model structures the innovation variance is set to $\psi = 1$ and the size of the AR parameter is varied as $\beta = \{0.25, 0.5, 0.75\}$ to mark three different levels of process autostability, low, medium and high. We vary the AR parameter not only because it is the parameter of central interest in AR(1) models, but also because we expect that it will be the most relevant parameter to affect simulation results. The number of time points on which we sample from the different baseline processes is set to $T = 80$, which is a typical length for IL affect time series (Dejonckheere & Erbas, 2021; Vachon, Viechtbauer, Rintala, & Myin-Germeys, 2019).

The baseline models presented so far are then combined with different types of outliers. For the model without measurement error, there are either no outliers or innovation outliers. For the model with measurement error, we additionally include measurement outliers. Outliers are generated according to the model presented in Equations 3 to 6 evoking outlier-indicator variables x_{1_t} and x_{2_t}. For convenience, we rely on the fixed moderated time series analysis framework (Adolf et al., 2017) and corresponding model formulation, and let the outlier-indicator variables moderate the dynamic and measurement intercepts of the baseline models, respectively. This is mathematically equivalent to the original formulation of adding a constant effect to some innovations and/or measurement errors. The dynamic and measurement error intercepts assume their baseline values (i.e., $\alpha = \alpha^0$ and $\tau = \tau^0$) if they are un-moderated or if x_{1_t} and x_{2_t} equal zero, and (composite) values $\alpha = \alpha^0 + \alpha^x$ and $\tau = \tau^0 + \tau^x$ if they are moderated and x_{1_t} and x_{2_t} equal one. The baseline intercepts always equal $\alpha^0 = 1$ and $\tau^0 = 0$, the size of both moderation effects τ^x and α^x is set to five times the model-implied overall standard deviation of y_t. The overall standard deviation is thereby calculated as $\sigma_y = \sqrt{\frac{\psi}{1-\beta^2} + \theta}$ (Hamilton, 1994). The implemented moderations are suited to generate innovation and measurement outliers, because they can be viewed as temporary changes to the

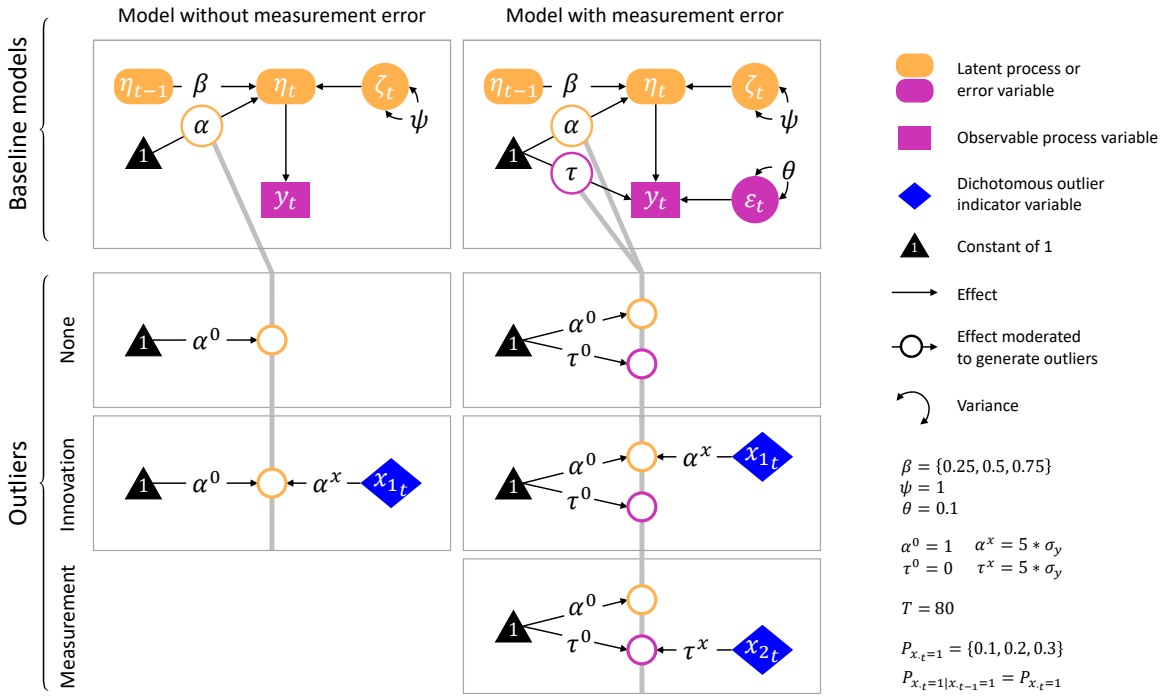

Figure 3
Data-generating models and parameter values employed in the simulation study. Two baseline model structures are combined with two and three outlier scenarios, respectively. Outliers are introduced by letting dichotomous outlier indicator variables moderate the dynamic and measurement intercepts. These parameters are therefore encircled, which is based on Curran and Bauer's suggestions on how to incorporate varying parameters into path diagrams (Curran & Bauer, 2007). Following the grey paths from the baseline models downwards, one can trace all the different moderation settings that the intercept parameters undergo. It is immediately visible when intercepts assume their baseline values reflected in the regression effect of the constant and when the outlier indicators additionally have an effect.

expected values of the innovation and measurement error processes from zero to some large non-zero value, which thereby very likely generate extreme innovations and measurement errors. In our case, the two outlier indicator variables take effect separately causing either innovation or measurement outliers.

The outlier-indicator variables themselves are generated with different probability distributions. The marginal probabilities of x_{1_t} being one and x_{2_t} being one are set to a range of $P_{x_{\cdot t}=1} = \{0.1, 0.2, 0.3\}$. These probabilities are implemented in an exact manner by sampling the respective absolute number of time points at which an outlier indicator variable is to be set to one instead of zero. In the innovation outlier setting, given $T = 80$ and $P_{x_{1_t}=1} = 0.1$ for example, we then have exactly 8 occasions contaminated with innovation outliers for all simulation replications. If, in the measurement outlier setting, $T = 80$ and $P_{x_{2_t}=1} = 0.2$, we always have exactly 16 occasions contaminated with measurement outliers. Outlier occurrence is temporally random under all outlier generation conditions.

Our simulation study encompasses a number of additional data generation settings not described

above. These concern additional sample size settings (i.e., $T = \{160, 800\}$) to check the generalizability of results to larger samples, mixed innovation and measurement outliers as well as outliers occurring with medium and high temporal dependence to mimic more complex outlier settings. More information on these settings as well as a presentation of the corresponding simulation results can be found in the supplementary materials. In total, the partly nested data-generation design of our simulation study yields 288 condition combinations or design cells.

4.2.2 Models fitted

We re-apply the models described in Sections 2.3 and 3.3 in the simulation study. Each simulated data set is thus modeled using the three robust AR(1) model variants relying on MM, MTL, and cML estimation, as well as the classical and the true AR(1) model variants relying on ML estimation. For MM estimation we use the recommended default settings (Koller & Stahel, 2011; Maechler et al., 2021). For MTL estimation, we vary the trimming proportion and set it to a broad range of values (i.e., $\alpha = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$). With the chosen values, we match all marginal outlier probabilities while also always having at least one smaller and at least one larger proportion.[4] Also, these trimming proportions can again be implemented exactly, that is, if $T = 80$ and $\alpha = 0.05$ then 4 observations are discarded during estimation. For cML estimation we retain the commonly used $3 - \sigma$ rule in combination with classical mean and SD estimators. In addition, we adapt the σ factor to match all above used MTL trimming proportions: To parallel a trimming proportion of $\alpha = 0.05$, for example, we employ a factor of ≈ 1.96. The maximum trimming proportion of $\alpha = 0.35$ is matched by a factor of ≈ 0.93. As with the $3 - \sigma$ rule these values are based on the quantiles of the standard normal distribution that exclude the respective α fractions of observations in the tails of the distribution.[5] Since such alternative σ factors are rather uncommon in outlier detection, we include a summary and discussion of the results in the manuscript, but present the visualizations in the supplementary materials. The true AR(1) model is implemented by fitting the fixed moderated AR(1) model that we already used for data generation (Adolf et al., 2017). We thus condition on the generated outlier-indicator data to recover the intercept moderations. Specifically, we constrain the intercept parameters to vary in concordance with the outlier indicators, while estimating their effect. The baseline measurement intercept thereby remains always fixed to zero for identification purposes. As detailed in Section 3, this model is included as a comparison model to mark 'optimal' performance.

[4] Note however that matching marginal outlier probabilities does not always mean that the amount of actual contamination is matched. A certain amount of measurement outliers introduces that amount of vertical outliers, but in addition also bad leverage points. We discuss possible implications of not matching the actual amount of contamination in these cases in Section 6.

[5] Unlike in the MTL estimator case these fractions are *theoretical* fractions based on the probability density function of the standard normal distribution. These need not correspond to the actually excluded fractions due to masking effects as explained in Section 3.

We fit all models without and some also with measurement error. The reason for the first is to ensure comparability across all models - while the SSM framework can handle measurement error, the regression framework and thus MM-based AR modeling cannot. The reason for the latter is that we want to get an idea of the benefit of modeling measurement error in the current simulation context (i.e., apart from outliers and robust methods, this concerns a small-ish sample size, varying AR effect sizes, limited measurement error influences). The classical, true, MTL and cML models are thus also included with the measurement error variance being freely estimated instead of fixed to zero. Next to measurement error handling, two more minor differences between regression modeling and SSM concern parameter starting values and handling of the initial observations. We provide more information on these issues in the supplementary materials.

Finally, for MTL estimation, we investigate the possibility of tuning the trimming proportion, that is, estimating α from the data. To this end, we use the out-of-sample trimmed likelihood as a model selection criterion.[6] Following a basic procedure, we split each simulated time series into a first and a second half, to be used as training and test set respectively. Each MTL estimator variant is applied to the training set and parameter estimates are obtained. These - and the corresponding trimming proportion - are then used to calculate the out-of-sample trimmed likelihood on the test set. The MTL variant with the smallest out-of-sample trimmed negative log-likelihood per test set is selected as the 'tuned MTL estimator', and final parameter estimates are obtained using the full time series. Our procedure mimics the typical out-of-sample-error-based tuning procedure in the context of regularized estimation (e.g., Orzek & Voelkle, 2023).

4.2.3 Assessing model performance

To assess model performance we primarily focus on estimation accuracy. This is an obvious criterion since it is clearly affected by outliers and many other performance measures depend on it (e.g., predictive accuracy). Estimation accuracy is cast in terms of the mean squared error of an estimator, a compound accuracy measure combining an estimator's bias and variance. Here we limit ourselves to the different AR parameter estimators employed, since estimating the AR parameter is of primary interest in the present context. For any of those estimators denoted $\hat{\beta}_\bullet$ the mean squared error (MSE) is defined as the sum of its variance and its squared bias. We obtain the observed MSE, $MSE\left(\hat{\beta}_\bullet\right)_{obs}$, by calculating the observed variance and observed bias across the N simulation replications using the observed AR estimates $\hat{\beta}_{\bullet_{obs,\,i}}$ and

[6] Using the out-of-sample mean squared error obtained via cross validation would be a more established selection criterion. It is however not a useful criterion in this case: Robust methods are designed to produce modeling results close to results under uncontaminated data (Rousseeuw & Hubert, 2011). If we are dealing with contaminated test data then (well-performing) robust methods should be - and are - associated with larger out-of-sample mean squared errors. The out-of-sample trimmed likelihood instead offers a natural way of taking outliers in the test set into account and is comparable across different trimming proportions (Crevits & Croux, 2019). It therefore seems a suitable selection criterion in this case.

their means:

$$
\begin{aligned}
MSE\left(\hat{\beta}_{\bullet}\right)_{obs} &= Var\left(\hat{\beta}_{\bullet}\right)_{obs} + \left(Bias\left(\hat{\beta}_{\bullet}\right)_{obs}\right)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\beta}_{\bullet_{obs,\,i}} - \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_{\bullet_{obs,\,i}}\right)^2 + \left(\frac{1}{N}\sum_{i=1}^{N}\hat{\beta}_{\bullet_{obs,\,i}} - \beta\right)^2 . \qquad (7)
\end{aligned}
$$

To quantify the error in our simulation due to 'sampling' random replications we bootstrap the sampling distribution of each observed MSE with 500 bootstrap replicates. In the supplementary materials we present the 16th and 84th percentiles of the sampling distribution as two thirds of the bootstrapped MSEs lie between those.

Next to estimation accuracy, predictive accuracy of models with respect to unseen data is gaining more and more attention (Yarkoni & Westfall, 2017). While it is a function of estimation accuracy, predictive accuracy additionally captures results' generalizability and the performance of each model as a whole. We thus complement the MSE with the mean squared prediction error (MSPE) as observed across replications, but keep the presentation of the latter criterion and the associated results rather brief in the current manuscript. More information, including details on how the MSPE is calculated can be found in the supplementary materials. The supplementary materials also feature results on how the different robust methods handle outliers (e.g., how many vertical outliers are weighted-down during estimation?) and all other data points.

4.2.4 Implementation and Software

The simulation study is - as any analysis or visualization presented in this paper - implemented in *R* (R Core Team, 2022). We provide the *R* package *simARO* ("**Sim**ulating **AR** models with **o**utliers") to increase transparency and enable *R*-level reproducibility of the raw simulation results. The package is available via a GitLab repository and contains and documents all functions and data needed to (re-)run the simulation study. To this end, the GitLab repository *use_simARO* and the associated *R* project support the use of the *simARO* package and guides the reader through the simulation workflow. Also, the *renv* package (Ushey, 2022) is employed to make sure the *simARO* package and its dependencies are used in the correct versions. For the here-reported simulation study, we used *simARO* version 0.3.12 and the resources and services of the *VSC* (*Flemish Supercomputer Center*). Both GitLab repositories can be accessed via the OSF repository accompanying this manuscript.

For model estimation, *simARO* depends on the following *R* packages: The true, classical and cML AR(1) models are formulated as regular state space models and estimated using the corresponding

functionality in *OpenMx* (Hunter, 2018; Neale et al., 2016). MM estimation is carried out using the *robustbase* package (Maechler et al., 2021) and its *lmrob*-function. For the MTL estimator we rely on our own implementation drawing upon the accurate description of the trimmed likelihood function in Crevits & Croux (2019) and *OpenMx*'s capability to handle user-defined fit functions (see supplementary materials). We verified our implementation using the author's original code, which they kindly shared. Next to these model estimation packages, *simARO* also requires *dplyr* (Wickham, François, Henry, & Müller, 2022), *MASS* (Venables & Ripley, 2002), *purrr* (Henry & Wickham, 2020), and *tidyr* (Wickham & Girlich, 2022) to produce simulation results.

The raw simulation results are also available via the GitLab repository *simAROdata*, again accessible via the OSF repository. We further summarize these results here with the help of *dplyr* (Wickham et al., 2022) and *tidyr* (Wickham & Girlich, 2022). We also use the *boot* package (Canti & Ripley, 2021) to bootstrap the sampling distributions of model performance measures. Visualizations are created using the *ggplot2* package (Wickham, 2016) and some of its extensions.

4.3 Results

Per design cell we generated $N = 1000$ data sets. There were small numbers of non-converging modeling solutions across design cells, which are visualized in the supplementary materials and excluded from the following result presentation. For the design cells relevant for the results presented in the manuscript the mean number of non-converged solutions is 0.07, the median is 0, and the maximum is 4. Given the small absolute numbers of these problematic solutions, however, we do not expect that excluding those will significantly affect the following results.

4.3.1 Improved estimation due to innovation outliers

First, our results confirm that innovation outliers being introduced can improve estimation accuracy of the AR parameter. As mentioned this is sometimes presented as a side-finding in the robust statistics literature (e.g., Croux & Joossens, 2008; Maronna et al., 2006; Zoubir et al., 2012) and we highlighted the underlying mechanism and provided some evidence in our anecdotal example. Looking at it in more detail in the context of this simulation reveals some of the conditions determining the extent to which innovation outliers are beneficial.

Figure 4 shows the observed MSE of the AR parameter estimator (y-axis) across the different proportions of innovation outliers in the data (x-axis). The figure is confined to measurement error-free data. Each row of the figure refers to one of the considered modeling approaches, here all estimated without measurement error. Columns differentiate results according to the size of population-level AR parameter.

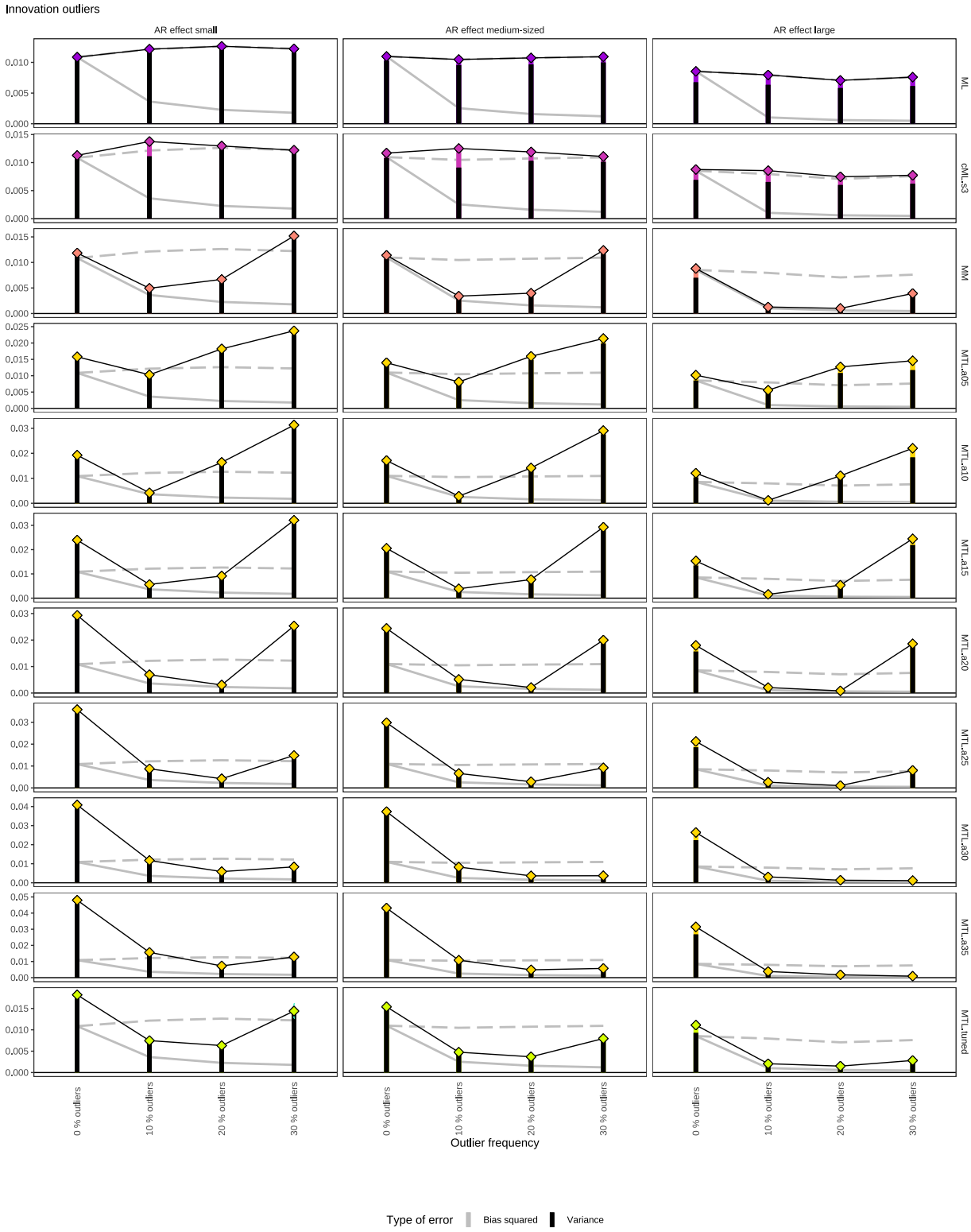

Figure 4
Estimation error AR effect for data with no measurement error (y-axis) as a function of the proportion of innovation outliers (x-axis) in measurement error-free data. Estimation methods are distinguished by rows, population-level AR effect sizes by columns. The bars marking estimation errors are decomposed into variance (black; lower bar part) and squared bias (color depending on method; upper bar part). Dashed and solid grey lines (re-)mark the performance of the classical and the true model respectively. All models displayed are fitted without measurement error, outliers occur randomly and $T = 80$. Note that the scaling of the y-axis can vary to enhance comparability within conditions.

The MSE is visually decomposed in that the part due to variance is displayed in black and the part due to squared bias in color depending on the modeling approach. For reference, we include in all tiles the MSE of the classical ML approach (dashed grey line) and the true model (solid grey line). The figure is further restricted to outliers occurring randomly and $T = 80$.

We find a reduction of estimation error due to innovation outliers under at least *some* conditions for all methods. This mainly concerns variance, but bias can also be reduced if it plays a role (i.e., when the AR effect is large and/or there is measurement error in the data, which is unaccounted for, see figure in the supplementary materials). For classical ML and $3 - \sigma$ cML estimation improvements are found in the largest AR effect size condition. Larger AR effect sizes lead to a larger number of pronounced good leverage points, since a larger AR effect size implies a slower return to the mean. Such pronounced good leverage points are probably needed here to compensate for the estimation-disrupting effects of the vertical outliers, since neither classical ML nor cML estimation handles them (either not at all or not well, see figure on outlier handling in the supplementary materials). Another factor influencing the results is the amount of contamination. For the more sophisticated robust estimators (i.e., MM and MTL) we widely observe a curvilinear relationship with estimation error. In fact, patterns are more or less convex for most of these estimator variants, suggesting that attuning robustification to the amount of contamination can yield optimal estimation accuracy results.

We will take a closer look at this in the following subsection, in which we compare the performance of the different robust methods directly. Before doing so we note that, first, the overall results look very similar if data with measurement error are considered instead of data without. Small differences mainly concern the above-mentioned clearer reductions in bias due to innovation outliers, especially if AR effects are large. Second, as expected, measurement outliers do not induce any accuracy improvements, instead there is especially increasing bias with increasing contamination, especially for large AR effect sizes. And, third, also predictive accuracy is not improved under any type of contamination. For visualizations of these results, please consult the supplementary materials.

4.3.2 Improved estimation and prediction due to robust methods

Figure 5 again shows the observed MSE of the AR parameter estimator (y-axis) - but now as a function of the different modeling approaches used (x-axis), while the outliers proportions are distinguished by row. Panel A pertains to contamination with innovation outliers, Panel B to contamination with measurement outliers. Unlike in the previous figure, data are now affected by measurement error, which is, again, not modelled. Turquoise vertical lines indicate the 'winning' model with the smallest observed MSE per tile (solid line) and those models that closely follow-up (i.e., models of which the observed MSE is smaller than the bootstrapped 84th MSE percentile of the winning model; dashed lines).

Looking at the innovation outlier case (Panel A) and the rows referring to contaminated data (i.e., ignoring the first row for now), we notice that it is always one of the MTL estimators that 'wins', more specifically in almost all cases the one of which the chosen trimming proportion *matches* the true proportion of contamination (e.g., $MTL_{\alpha=0.1}$ when there are 10 % outliers). A look at how the MTL estimators handle the data during estimation (see supplementary materials) indeed reveals that (on average) all vertical outliers are handled once the trimming proportion allows this (remember that the percentage of vertical outliers equals the percentage of innovation outliers). The observed MSEs thereby approach the ones of the true model. At the same time, it is often one of the MTL estimators that 'looses'. This happens when contamination is stronger (i.e., ≥ 20 % outliers) and trimming proportions are *too small*, where α-values smaller than the actual amount of contamination result in high, variance-driven MSEs that clearly exceed the MSEs of the classical ML estimator. When contamination is weak (i.e., 10 % outliers) and the AR effect is small (i.e., $\beta = 0.25$) the MTL variant with the highest - and thus also *too high* - trimming proportion seems to be 'loosing', but for the larger AR effect sizes the $3 - \sigma$ cML estimator performs the worst. It thus seems that unhandled vertical outliers are the primary reason for the MTL estimator's mal-performance (again, see supplementary materials showing that vertical outliers typically remain unhandled by the MTL estimator if trimming proportions undercut the actual amount of contamination). And these unhandled vertical outliers seem to have a relatively large negative impact on estimator performance and especially estimator variance. Zooming into the set of MTL estimators and inspecting performance in between the extremes, we additionally observe an increasing estimation error with increasing trimming proportions if trimming proportions are *too high*. The uncontaminated data in the first row of Panel A represent a prime example of that effect. Here we likely see the earlier-mentioned robustness-efficiency trade-off at work. If trimming proportions exceed the actual amount of contamination, uncontaminated data points are disregarded in addition (see again supplementary materials). While the estimator in principle gets more robust, it also gets less efficient and hence more variable, as it weights down more (uncontaminated) data during estimation. Eventually inspecting the tuned MTL estimator, we, unsurprisingly perhaps, observe performances that lie in between the ones of the respective best- and worst-performing MTL variant. Performance levels are however always considerably closer to the best- than the worst-performing MTL variant and, under contamination, almost always better than the classical ML estimator. The MM estimator shows less variable behavior over conditions and performs rather well generally. Although, returning to the above mentioned curvilinear relationship between contamination and estimation error, we seem to observe a generally increasing estimation error with an increasing amount of contamination. This makes sense since the robustification of the MM estimator is not varied in the present simulation. Under the strongest contamination (i.e., 30 % outliers) and small to moderate AR effect sizes the MM estimator is outperformed

by the classical ML estimator. Otherwise this only happens in the absence of contamination if the AR effect is small. For larger AR effect sizes, the MM estimator even wins although there is no contamination. In terms of trading robustness for efficiency, the MM estimator thus preserves a high level of efficiency as compared to other estimators (Maronna et al., 2006; Zoubir et al., 2012). Employing the $3 - \sigma$ cML estimator is generally not a good idea, since it performs either similarly to the classical ML estimator or worse. Performance seems poorest under 10 % of outliers with quite some bias contributing to the estimation error. As stated (and indirectly shown) in the previous subsection, the overall innovation outlier results for measurement error-free data are largely comparable to the presented ones for data with measurement error. Only bias is a slightly less important source of the MSE, especially if AR effect sizes are larger. An extended version of Figure 5 including data without measurement error can be found in the supplementary materials.

Turning to the measurement outlier case (Panel B), the MTL estimator - or one of its variants - still always wins against the other methods if there is contamination. The trimming proportion of the winning MTL variants exceed the true proportion of contamination in the majority of cases (i.e., when the AR effect is medium or large). This makes sense since a certain amount of measurement outliers introduces that amount of vertical outliers plus the same amount of bad leverage points. The tuned MTL estimator again performs worse that than the respective best- and better than the respective worst-performing MTL variant. It also seems to outperform the classical ML estimator in the majority of cases, but less so under smaller AR effects or more outliers. The performance of the MM estimator relative to the other methods is less good overall compared to its relative performance under innovation outliers. It now produces rather large biases especially under large AR effect sizes across all contamination levels. Being prone to contamination bias is the downside of keeping a high efficiency level. Note however that the MM estimator usually still outperforms the classical ML estimator when the data are contaminated with measurement outliers (or when the data are uncontaminated and the AR effect size is moderate to large). The $3 - \sigma$ cML estimator again performs comparably to the classical ML estimator or worse.

If we consider alternative σ factors for the cML estimator, we get rather mixed results: Under innovation outliers, the chosen alternative factors worsen performance in most cases, with smaller factor values (i.e. larger theoretically excluded fractions of data) leading to larger MSEs. Especially bias is increased to such large extents that most alternative cML variants exceed the MSEs of the 'loosing' MTL variants by far. Only under medium and large AR effects and 10 % of outliers do we see a subtle convex relationship between the σ factor values and the MSE. In these cases some alternative cML variants outperform the classical ML estimator, but none of the other robust estimators. Looking at how the alternative cML variants handle the data during estimation and comparing it to the behavior of the

corresponding MTL variants, we see that, generally, less vertical outliers and more other data points (potentially good leverage points) are excluded. Under measurement outliers results are quite complementary. Here we find that factor values smaller than the established value 3 often also lead to smaller MSEs. For 10 % of outliers this is the case for all alternative factor values, for stronger contamination this holds for smaller factor values and larger AR effect sizes (i.e., factors ≤ 1.44 for 20 % of outliers and medium and large AR effects, and factors ≤ 1.04 for 30 % of outliers and large AR effects). In these cases the corresponding cML variants outperform the classical ML estimator. Moreover, the MSEs are sometimes so strongly reduced that a number of cML variants also clearly outperform the previously 'winning' MTL variants. In terms of data handling, we see that under $10 - 20$ % of outliers the alternative cML estimators seem to exclude less bad leverage points than their MTL counterparts, but also less other data points. Under 30 % of outliers they seem to exclude more other data points than their MTL counterparts if factor values are smaller, but also more bad leverage points. The MSE differences due to alternative σ factors are again mainly attributable to differences in bias, this time more often bias reductions than increases. If there is no contamination we recover the pattern under innovation outliers: MSEs increase with decreasing factor values, exceeding the MSEs of the different MTL variants by far. Visualizations of these results concerning alternative σ factors can be found in the supplementary materials.

In summary, estimation accuracy can be improved by using robust over classical estimators if the data are contaminated with either innovation or measurement outliers. The population-level AR effect size and the amount of contamination in relation to the robustification of the estimator thereby matter for estimator performance. Among the robust estimators that (can) outperform the classical ML estimator under both measurement *and* innovation outliers and in most contamination and effect size settings, we always find (a variant of) the MTL estimator to be best performing. At the same time we find that the MTL estimator is very sensitive to a robustification-contamination mismatch. This also extends to the performance of the SE estimators of the AR parameter obtained from the trimmed likelihood function: Results presented in the supplementary materials show an underestimation of the MTL estimator variance by the SE estimator under inappropriate trimming proportions. Considering that we implemented a rather wide range of trimming proportions and thus (degrees of) robustifications, these MTL estimator results might not come as a surprise and might mainly be a function of our simulation design. What is remarkable, however, is that the performance of the MTL estimator deteriorates substantially and can be worse than the one of the classical ML estimator if trimming proportions are chosen too small and vertical outliers or bad leverage points remain unhandled. Our results thus underline that it is of crucial importance to adjust the α value to the true amount of contamination, which is, however, usually unknown. We provide initial evidence that

tuning the MTL estimator via the trimmed likelihood function seems a viable approach to deriving an appropriate level of robustification from the data. The MM estimator is usually better than the classical ML estimator, but can suffer from substantial contamination bias. The $3-\sigma$ cML estimator as implemented here cannot be recommended as a robust method. Alternative (i.e., smaller) σ factor values can, under measurement outliers, greatly improve performance in that certain cML variants even outrank the respective MTL variants as best performing estimators. Under innovation outliers, however, the pattern is reversed and none of the alternative cML variants reaches the performance any of the other robust estimators that outperform the classical ML estimator. Instead, alternative cML variants often perform (much) worse than any other robust estimator. More results for the special cases of mixed and temporally dependent outliers, and for longer time series can be found in the supplementary materials.

Finally, inspecting the predictive performance of the different methods, we note that the overall MSPE pattern resembles the pattern for the MSE of the AR parameter estimator (see supplementary materials). Unlike in the estimation error case, however, the MTL estimator almost always outperforms the classical ML and the similarly mal-performing $3-\sigma$ cML estimator except for the combination of the largest AR effect size, the highest amount of contamination with innovation outliers, and the smallest trimming proportion. Also the cML estimator variants with alternative sigma factors are performing better in the innovation outlier case: When the AR effect is small or contamination is weak(er), they often outperform the classical ML and $3-\sigma$ cML estimator, but do not reach performance of the (tuned) MTL estimator.

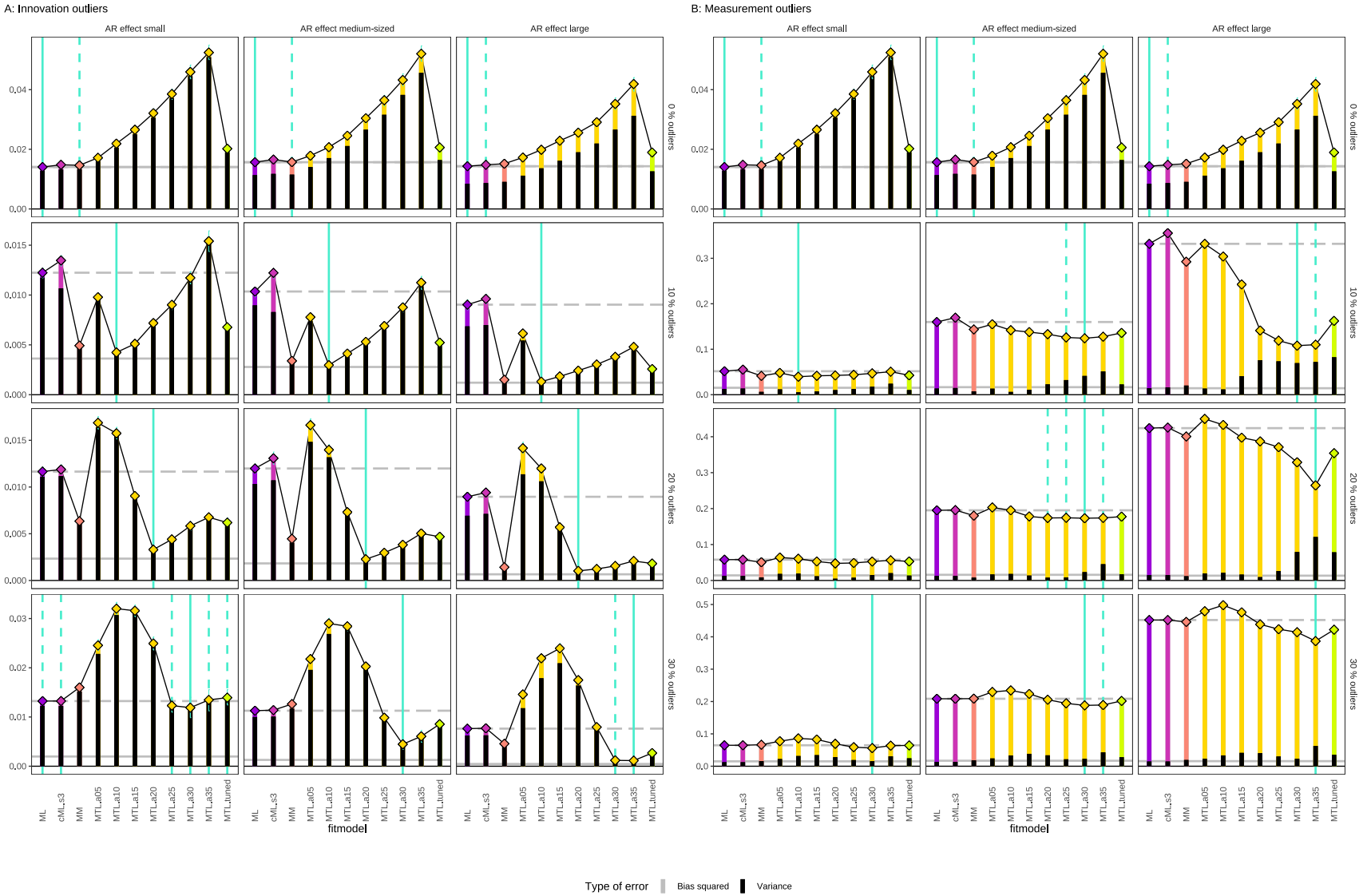

Figure 5

Estimation error AR effect for data with little measurement error (y-axis) as a function of estimation method (x-axis). Panel A displays results under innovation outliers, Panel B under measurement outliers. In both cases, data are affected by measurement error. Outlier proportions are distinguished by rows, population-level AR effect sizes by columns. The bars marking estimation errors are decomposed into variance (black; lower bar part) and squared bias (color depending on method; upper bar part). Dashed and solid grey lines (re-)mark the performance of the classical and the true model respectively. Turquoise vertical lines indicate the winning model (smallest MSE; solid lines) and the models close to it (dashed lines). All models displayed are fitted without measurement error, outliers occur randomly and $T = 80$. Note that the scaling of the y-axis can vary to enhance comparability within conditions.

4.3.3 Modeling measurement error

So far we have only considered results under the restriction of measurement error not being modeled. However, some of the included methods can in principle handle measurement error and this might have beneficial effects on estimation accuracy. Figure 6 thus shows how estimation accuracy changes when measurement error is present *and* modeled. As in Figure 5, the MSE of the AR parameter estimator is shown on the y-axis, but now only across the classical ML, the $3-\sigma$ cML, and the MTL estimator(s), each implemented in a model without (diamonds) and a model with measurement error (triangles). Bars displaying the same estimator are grouped together, with gray-scale bars referring to the models with measurement error and colored bars to the models without, as in previous figures. If the model with measurement error outperforms the model without by means of a smaller estimation error the corresponding triangle is marked in turquoise.

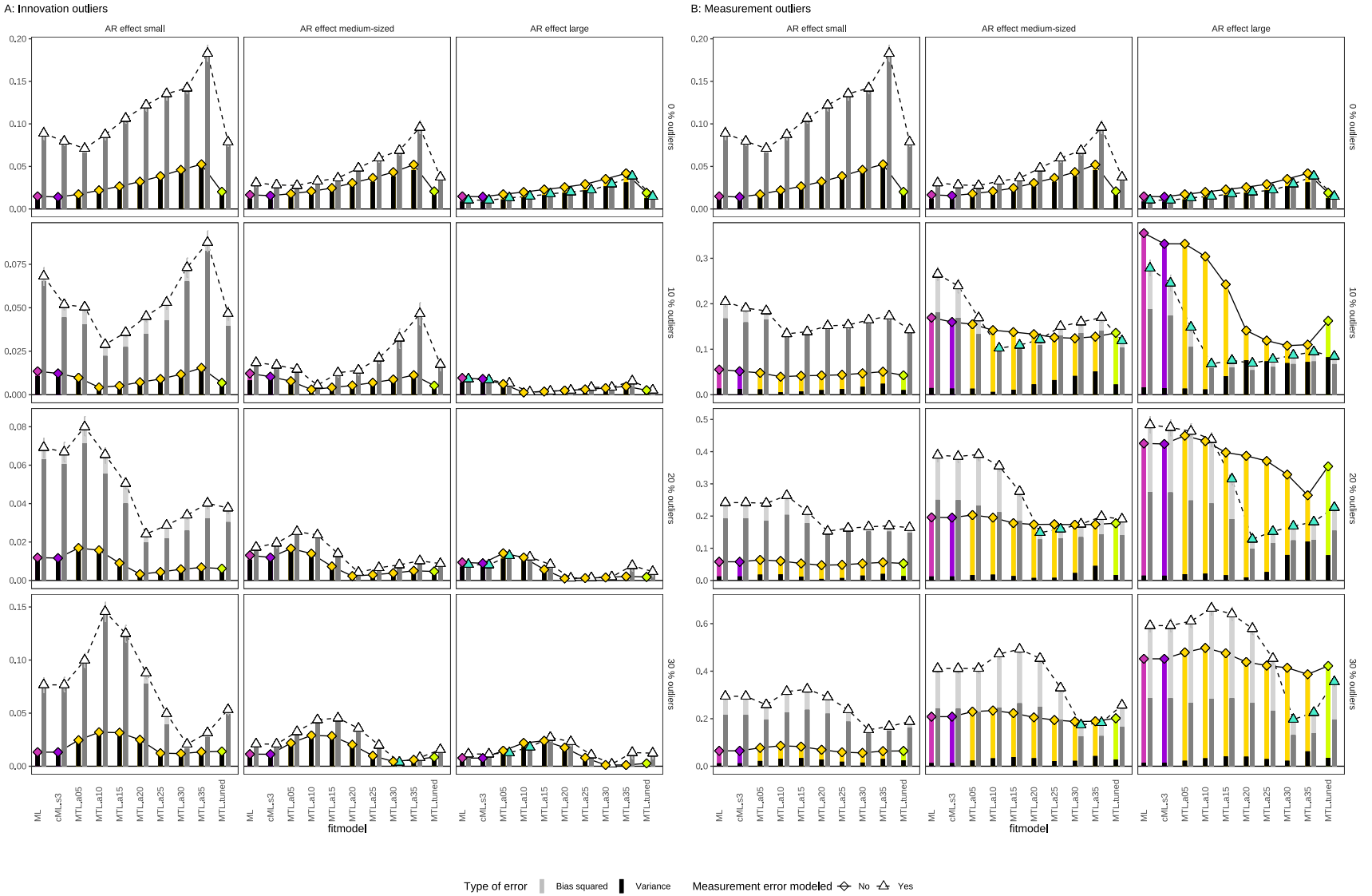

Figure 6

Estimation error AR effect for data with little measurement error (y-axis) as a function of estimation method (x-axis). Results pertain to measurement error-affected data with innovation outliers (Panel A) or measurement outliers (Panel B). Outlier proportions are distinguished by rows, population-level AR effect sizes by columns. Estimation methods are reduced to classical ML, $3-\sigma$ cML and MTL but each method is implemented in a model without (diamonds) and with measurement error (triangles). Bars displaying the same estimator are grouped together, with colored bars referring to the models without measurement error and gray-scale bars to the models with measurement error. Bars are again decomposed into variance (either black or dark grey; lower bar part) and squared bias (color depending on method or light grey; upper bar part). If the model with measurement error shows a smaller estimation error than the model without measurement error it is marked in turquoise. Outliers occur randomly and $T = 80$. Note that the scaling of the y-axis can vary to enhance comparability within conditions.

One clear result across all panels and thus outlier-types is that modeling measurement error never pays off when the AR effect size is small - at least given the small amount of measurement error induced here, no or randomly occurring outliers and $T = 80$. Estimating measurement error results in MSE increases, which are often large and mainly attributable to increases in variance. For medium and large AR effect sizes, the outlier type makes a difference.

In the case of innovation outliers (Panel A), it is only for the large AR effect size, that we sometimes observe a beneficial effect of modeling measurement error. This effect however remains relatively small and also only occurs for the classical and $3 - \sigma$ cML estimator, or for those MTL estimator variants for which the trimming proportion is chosen too small and which are thus outperformed by other MTL variants and the classical ML estimator (without measurement error).

Under measurement outliers (Panel B), modeling measurement error can be beneficial under both medium and large AR effect sizes, and more often so the smaller the amount of contamination. For these beneficial cases, we always observe reductions in bias, which are sometimes quite drastic, but there are also very often increases in variance, albeit less extreme. Interestingly, there also seems to be an effect on what the optimal trimming proportion of the MTL estimator is in these cases: If the measurement error variance is freely estimated, then the MTL variant with the trimming proportion that matches the proportion of measurement outliers seems to win. This is surprising, because the actual amount of contamination (i.e., the amount of vertical outliers and bad leverage points) will in principle be twice as large as the amount of measurement outliers and the winning MTL variants should either neglect vertical outliers or bad leverage points. But it is possible that these neglected outliers can to some extend be handled by the measurement error variance. We also return to this in the discussion.

Shifting focus to predictive performance of the different methods, we more often observe an advantage of the model variants that include modeling measurement error. Performance gains appear to be larger for non- or less robust estimators (i.e., classical ML, $3 - \sigma$ cML, and MTL estimators with smaller trimming proportions). At the same time, the differences between model variants with and without measurement error seem rather negligible as compared to differences between estimation methods.

5 Empirical illustration

5.1 Data

This empirical illustration relies on IL data collected from $N = 89$ participants using an experimental gambling paradigm involving $T = 152$ trials (Ariens, Adolf, & Ceulemans, 2023). The data can be retrieved from an OSF repository (Ariens, 2023). In each trial participants were asked to chose one out of

four doors, behind which monetary wins or losses where hidden. Participants were told that two wins and two losses had been randomly assigned to the doors whereas their distribution was in fact under experimental control and fixed across participants. After choosing a door, participants were presented with their current total (i.e., cumulative) wins or losses, and had to rate their momentary affective state. For the latter a grid with dimensions "positive" and "negative" was presented, and participants indicated one out of the $10 * 10$ possible grid positions. Here we use the positions along the "negative"-axis as measures of negative affect (Ariens et al., 2023).

5.2 Aim and Analysis

Unlike in the anecdotal example or the simulation study we do not know what the participants' true affect dynamics and hence AR effect sizes are. We will thus not be able to evaluate how well the different methods recover these effects.[7] Instead, we capitalize on the experimental design underlying the present IL data to showcase some evidence for the events-as-contaminators perspective: The main contextual factors (i.e., the task participants had to perform and the wins they did or did not book) are under experimental control and therefore known. Moreover, the experiment was designed to alter affective states. Here we speculate that, although wins and losses were experienced at every measurement occasion, more extreme total wins or losses might have functioned as contextual impulses, evoking more distinct affective responses and hence outliers.

We thus analyse the present data using the discussed robust AR(1) model variants. For the MTL estimator, we limit this illustration to the three smallest trimming proportions (i.e., $\alpha = \{0.05, 0.10, 0.15\}$)[8] and include tuning. We then investigate whether the robust methods detect outliers across participants, and/or to what extent the detected outliers might be the result of contextual impulses. All analysis code is available in a GitLab repository via the OSF repository.

5.3 Results

The numbers of outliers detected across participants strongly vary by method: Under $3 - \sigma$ cML estimation an average of 0.32 outliers are detected, under MM estimation we get an average of 1.88 outliers, and under the different MTL estimator variants the (expected) averages of 8, 15, 23, and 13.39 outliers.

[7] We do include such an evaluation in the supplementary materials, which involves adding innovation or measurement outliers to the observed data, and then comparing the performance of the various methods to the performance of classical AR(1) modeling of the original, uncontaminated data. The results are in line with the simulation study results, yet obtained under somewhat more realistic conditions (i.e., individual AR parameter estimates are used to generate innovation outliers and consequently positive leverage points; outliers had a 50 % chance per contaminated measurement occasion of being negative instead of positive and where thus symmetrical).

[8] If trimming proportions cannot be implemented exactly because they imply non-integer numbers of likelihoods to be trimmed, as in this case with $T = 152$, the proportions are slightly modified such that the nearest integer number of likelihoods gets trimmed.

While it is not surprising that the $3-\sigma$ cML estimator detects relatively few outliers (e.g., due to masking), the MM estimator also seems to detect a rather low average number. Keep in mind however, that during MM estimation data are weighted in a continuous manner and only flagged as outliers if their weights are very close to zero. In consequence, even if they are not flagged, data points can still be considerably down-weighted during estimation. If we correlate the MM estimation weights with the ones of the different MTL estimators and average those across participants, we indeed get relatively high values ranging between 0.71 and 0.78 (for the $3-\sigma$ cML and all other estimators average correlations range between 0.30 and 0.43). Also, modeling results (i.e., AR parameter point and SE estimates) obtained via the $3-\sigma$ cML estimator basically do not differ from the classical AR(1) results, whereas MM and MTL results do differ: For these we find relatively consistently across participants that AR parameter point estimates are larger and SE estimates - at least for the MTL estimator variants - are smaller under robust estimation (but see supplementary materials to the simulation study for an evaluation of SE estimator performance). Looking at whether outliers coincide with extreme wins or losses, we find that for the majority of participants wins are larger in average absolute value (i.e., further away from the average win over trials) on data points classified as outliers versus on regular data points. This holds across all robust estimation methods and is visualized in Figure 7. It shows the distribution of differences in average absolute wins, calculated for each participant by substracting the average value of absolute wins on trials classified as regular from the average value of absolute wins on trials classified as outlying. For a given participant, a positive difference score means larger absolute winnings (i.e., positive or negative) for outlying trails on average, whereas a negative difference score means larger absolute winnings for regular trails on average. We mainly observe positive difference scores across participants, supporting the notion of wins as contextual impulses. Relatively few cases lie left of the dashed line marking the point of no difference. This distribution is also evidenced by counting the number of participants for which difference scores are negative versus positive. The corresponding numbers are displayed in the top left and right corners of each tile.

6 Discussion

6.1 Summary

The potential impact of contextual events on affective processes is widely acknowledged and researchers attempt to study it by including events as covariates in IL data collection and modeling. Clearly, this requires that relevant events can be defined, measured, and modeled. But this can be complicated and one should be prepared to face situations in which events and their impact cannot be investigated explicitly. One then deals with IL affect data that are potentially contaminated with and thus obscured by event-induced outliers. In the first part of this paper, we have shown that for dynamical systems contextual

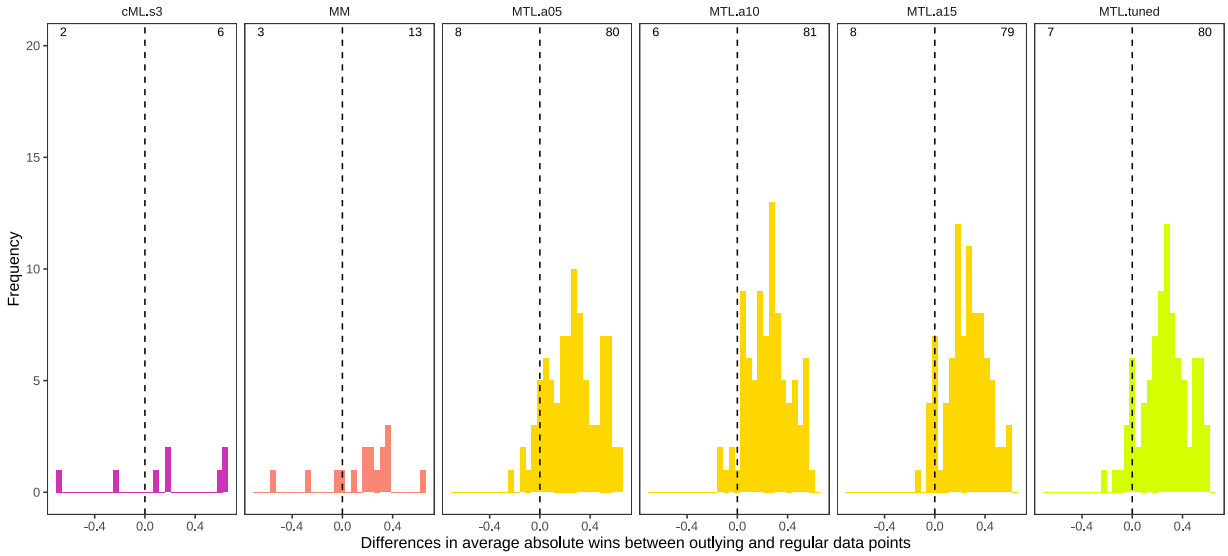

Figure 7
Results from the empirical illustration. Shown are the histograms of the frequency distribution of differences in average absolute wins between outlying and regular data points across participants. The different tiles refer to the different robust methods. A positive difference score implies that, for the respective participant, absolute winnings are on average larger for outlying than for regular trials. The number of individuals for which difference scores are positive versus negative are displayed in the left versus right corner of each tile.

contaminators are not necessarily as problematic as their name suggests, quite the contrary in certain cases: Innovation outliers caused by distinct contextual events} can trigger AR dynamics and improve estimation accuracy of AR parameters, especially when robust AR modeling is used. In addition, robust methods are designed to also handle other types of outliers such as measurement outliers.

The second part of the paper presents a simulation study and an empirical illustration. In the simulation study we compared a number of robust and classical AR models in the absence and presence of contextual contamination due to distinct contextual events with short-lasting effects. Across a range of conditions, we showed how estimation accuracy of the AR parameter is increased when contextual impulses (i.e., innovation outliers) are present rather than not, and when robust modeling techniques are used rather than classical ones if the data are contaminated with either innovation or measurement outliers. The latter performance advantage of robust models also transfers to their out-of-sample predictive accuracy. However, they did not perform better than classical methods unconditionally as tuning robustification to contamination matters. Also, their absolute performance under measurement outliers was rather poor overall. Modeling measurement error did lead to performance gains of the MTL estimator under some simulation conditions, though. Finally, a brief empirical illustration complements the simulation study by shifting focus from estimator performance to whether outliers detected by robust methods are related to contextual

changes in real-world affective time series data. To this end we analyzed a dataset collected via a gambling experiment, in which a contextual manipulation was implemented. Specifically, administered monetary wins and losses were supposed to influence participants' affective states. We applied robust AR modeling and showed that detected outliers are associated with more extreme wins or losses.

In the following we would first like to discuss some specificities and potential limitations of the simulation study and, briefly, of the empirical illustration. We then move on to some broader conceptual points.

6.2 Specificities and limitations of the simulation study

A first set of limitations of the simulation study is given by a number of design decisions regarding outlier characteristics, which might have been atypical or extreme compared to simulation studies reported in the robust statistics literature: This concerns, first, rather high outlier proportions, which likely made outlier handling quite difficult and possibly pushed some robust estimators towards or even beyond their breakdown points. The breakdown point of an estimator is the amount of contamination beyond which it cannot distinguish between uncontaminated and contaminated data anymore and its performance breaks down (i.e., it ends up maximally biased, Zoubir et al., 2012). Breakdown points are hard to determine theoretically in IL and other correlated data, but the point at which estimator bias stabilizes even though the amount of contamination is increasing can be an indication in a simulation study (Zoubir et al., 2012). This is what we seem to see for some of the robust estimators (e.g., MTL) when comparing higher proportions (i.e., 20% vs. 30%) of measurement outliers, and could thus explain their poor-ish performance under the measurement outlier setting. Remember that a certain percentage of measurement outliers in our design can lead to a contamination with vertical outliers and bad leverage points that is at maximum twice as high. For pure measurement outliers such high proportions might also not be realistic from a substantive perspective. It seems plausible that if the amount of measurement-disrupting events becomes larger (or their impact more extreme) this will also start influencing affective processes to some extent. But keeping the design symmetric between innovation and measurement outliers of course allowed us to draw direct comparisons (- and renders the performance under also high proportions of innovation outliers even more impressive). Eventually, our extreme design decision highlights that even when having robust methods available, efforts to prevent, minimize or record measurement outliers can be worthwhile. Further zooming into model performance under smaller amounts of (measurement) outliers might be an avenue for future research.

Another somewhat untypical outlier design decision concerns the one of outliers occuring randomly over time. When it comes to simulating daily event occurrence this seemed to be a more realistic choice than the often-used isolated outliers occurring in an equidistant pattern (Zoubir et al., 2012). It however also

means that outliers can follow each other closely or occur at subsequent occasions. Mapped into the variable space of the AR regression problem, this results in data points that are both bad leverage points and vertical outliers or good leverage points and vertical outlier (or 'doubly' good leverage points). Such 'multiply outlying' data points will affect estimation accuracy differently than simple outliers, sometimes in unexpected ways. In the illustrative example in Section 3, we saw how a bad leverage point can turn into a good leverage point when combined with a vertical outlier, at least if the population-level parameters of the AR model and the outlier impact have the right size. In the simulation study the frequency of multiply outlying data points varies across replications within conditions, but also varies systematically between conditions. This is rather obvious for outliers occurring with increasing temporal dependence (see results in the supplementary materials), but already increasing proportions of randomly occurring outliers make multiply outlying data points more likely. Under isolated outliers multiply outlying data points would either be ruled out completely (combination involving bad leverage points) or at least minimized (combinations involving good leverage points).

A third and last notable outlier design decision concerns the simplistic way in which event-based outliers had an impact. One characteristic of the here-employed outlier-generating Equations 3 to 6 was that event-indicators had a constant, deterministic effect. More complex outlier effects are of course possible. We already took a first step in this direction in the supplementary illustration, where an outlier effect had a 50 % chance per contaminated measurement occasion of being negative instead of positive (the results do however not diverge from the simulation results). Another strategy to implement probabilistic outlier effects involves letting the outlier indicator variables moderate the innovation and/or measurement error variances such that they become large at some time points (see Crevits & Croux, 2019). Since the expectations of the error processes would not change, however, less extreme innovations or measurement errors could still occur with relatively large probability, which seems somewhat at odds with the definition of outliers as being 'outlying'. Alternatively, a combination of moderated intercepts and error variances, which is equivalent to adding interaction effects between outlier indicators and normal random variables with non-zero means to both system equations, could be a useful extension towards more realistic outlier effects. One might also want to extend the outlier-generating equations to contain multiple event-indicators, so that different outlier-evoking events with different effects can occur over time or even simultaneously. Multiple outlier-evoking events might moreover interact so that it makes a difference for their impact if they occur simulatenously versus sequentially. Also, it might make a difference for their impact if outlier-evoking events follow-up on each or occur in a more isolated fashion (e.g., Schilling & Diehl, 2014). Finally, events in daily life are likely to be to some extent predictable, and people anticipating such events might already lead to changing emotions. There

are thus many directions along which the generation of outliers via event-indicators could be adapted or extended. However, our aim here was to introduce some rather basic insights from the robust statistics literature warranting a rather basic simulation design.

We move from outlier design decisions to what our simulation study does (not) allow to deduce regarding robustification. One thing we clearly saw was that model-naive robustification in the form of cML estimation with the typically used $3 - \sigma$ rule did perform as bad as or worse than classical, non-robust estimation. Alternative σ factor values could improve performance considerably under measurement outliers, but also worsen performance under innovation outliers. Employing the cML estimator generally as a robust estimator would therefore be a very risky endeavour. Model-based robustification in the form of MM or MTL estimation instead had the potential to perform better than classical ML estimation under both types of outliers. However, robustification still mattered here: For the MTL estimator and innovation outliers, we could impressively demonstrate that tuning the trimming proportion to contamination made a large difference in estimator performance - between worse than classical estimator performance and near-optimal performance, close to the one of true model. This does not necessarily mean that the MTL estimator is genuinely more sensitive to a contamination-robustification mismatch than the MM estimator. We just varied the trimming proportion quite drastically as compared to recommendations in the literature (Chan & Wei, 1992). The simulation results also revealed that tuning the MTL estimator is feasible. As a proof of principle we relied on the trimmed likelihood function and a simple half-half split into training and test set to calibrate the estimator's trimming proportion. Future research could map out alternative strategies of splitting the data during tuning (cf. Bulteel, Mestdagh, Tuerlinckx, & Ceulemans, 2018). It remains interesting to note that tuning does not seem to be a standard ingredient to tutorial or textbook texts of (applied) robust statistics. Another aspect that complicates general conclusions is that it is not so clear how the different estimators compare quantitatively in terms of robustification given the obvious qualitative differences in how these estimators work. Also, acknowledged measures of robustness such as breakdown points or maximum bias curves (Maronna et al., 2006; Rousseeuw & Hubert, 2011) are often not generally known and mapping them out would need extensive simulation, as we have already alluded to.

Another (quite literal) estimator-related limitation is that the studied set of robust procedures is of course far from exhaustive. For instance, for the data setting with measurement error, we could have combined the MTL estimator with the robustified KF also suggested by Crevits and Croux (2019). In addition, other forms of outlier handling can be implemented in a SSM context (You, Hunter, Chen, & Chow, 2020). Our selection here was guided by *R* packages on CRAN specifically dedicated to robust (AR) modeling and the recently proposed robustification of SSM.

Finally, it remains to be noted that our simulation design incorporates measurement error only to a limited extent. This decision was taken because the presence of measurement error in itself was not of primary interest in the present paper. In addition, modeling measurement error is generally complicated with single indicator AR(1) models if time series are short (Schuurman et al., 2015). To generate reliable results for other measurement error scenarios one would require longer time series and/or AR modeling drawing upon multiple indicators per process to better disentangle measurement and indicator-specific error from dynamic process error.

6.3 Limitations of the empirical illustration

Turning to the empirical illustration, a few limitations are to be mentioned. While the gambling experiment was designed to alter participants' affective states, the experimental manipulation did not involve distinct contextual events (i.e., wins or losses that are extreme in comparison to and thus distinct from the majority of wins and losses). Such a more tailored design might have produced even clearer results.

6.4 Conceptual specificities, limitations and implications

Expanding as well as complementing the specificities and limitations of the simulation design and the empirical illustration, we would like to make a number of conceptual points. First, as noted in Section 3, shifting the perspective from events as covariates to events as contaminators also shifts the focus of attention, away from the event and its impact towards the affective process. This should however not be misinterpreted as an evaluation of events as being unimportant or uninteresting and thus unworthy of research. The shift is a pragmatic one simply owing to a possible lack of information, preventing the explicit study of contextual events and their effects. In analogue, the term 'outlier' is neutral in the robust statistics literature - any data point that does not behave like the majority of data points is outlying and it is acknowledged that reasons for this can vary (Maronna et al., 2006; Rousseeuw, 1991). On the side, we would like to highlight here that in the simulation study we found that the true model was always performing best. So, if it is possible to model contextual events and their effects explicitly, then this would indeed be the preferred option. However, and as we have argued initially, the information necessary for this might not be available, because events can be hard to theorize about, measure, or model.

Second, we confined ourselves to the situation where distinct (i.e., somewhat important occasional) events exert (immediate or delayed) short-lasting main effects on an affective process or its measurement causing innovation or measurement outliers (or both). In Subsection 6.3, we already discussed additional ways in which events might perturb an affective process and cause outliers. But events might not only have short-lasting effects and evoke 'a few' outliers. More extreme events such as major life events (e.g., death of a close person, job loss) may rather lead to enduring changes in the equilibrium and/or dynamics of an

affective process Lucas (2007). If events cause enduring changes in affective processes it becomes more difficult to conceptualize them as contamintors causing occassional outliers obscuring an underlying affective process. Instead one would probably want to account for the change in process characteristics that has happened, also in terms of modeling. Consequently, robust modeling seems less of a suited approach, while efforts to explicitly measure and model the event and its effect seem more appropriate.

A third point concerns the general status of good leverage points. In our case, we know the good leverage points are caused by impulses and are therefore truly good in the sense that they reveal the actual dynamics of interest. But in other - often cross-sectional - contexts good leverage points can occur due to other reasons and do thus not automatically carry such valuable information. Even in the present context we saw such alternative good leverage points arise, if only rarely, when bad leverage points met vertical outliers. In those cases estimation accuracy can be impaired by good leverage points in that they can deflate SE estimates of the slope parameter and lead to overly liberal inferences (Dehon et al., 2009). Note however that good leverage points are only problematic if they are directly due to contamination and not if they are a consequence of certain outliers as shown in the present paper.

Finally, the presented principles also apply to intervention research. Here, a shift in perspective similar to the one described above can be taken: From doing interventions to learn about their effects on a system to doing interventions to learn about the system that one intervenes on (e.g., Driver & Voelkle, 2018). The latter perspective is common in engineering (e.g., Keesman, 2011) but not established in psychology yet. A difficulty will probably lie in designing interventions that are powerful enough to elicit strong reactions but remain specific and controlled at the same time (for a general discussion of difficulties with designing psychological interventions see Eronen, 2020). For instance, if one intervenes to learn about the size of the AR parameter in an AR(1) system, implementing strong impulse-type interventions causing extreme innovation outliers will likely be a good strategy. But these interventions should really only temporarily affect the state of the system and not the system and its dynamics itself. In addition, having control over an intervention and its effect also means that one possesses the information necessary to explicitly model the two. As we have mentioned above this was always the most accurate option at least in our simulation study.

7 References

Adolf, J. K., & Ceulemans, E. (2023). *Supplementary materials for manuscript: Improved estimation of autoregressive models through contextual impulses and robust modeling.* Retrieved from https://osf.io/u5yda/

Adolf, J. K., Voelkle, M. C., Brose, A., & Schmiedek, F. (2017). Capturing Context-Related Change in Emotional Dynamics via Fixed Moderated Time Series Analysis. *Multivariate Behavioral Research*, *52*(4), 499–531. https://doi.org/10.1080/00273171.2017.1321978

Aldao, A. (2013). The Future of Emotion Regulation Research: Capturing Context. *Perspectives on Psychological Science*, *8*(2), 155–172. https://doi.org/10.1177/1745691612459518

Ariens, S. (2023). *One does not simply correct for serial dependence: Supplementary materials.* Retrieved from https://osf.io/esmwa/

Ariens, S., Adolf, J. K., & Ceulemans, E. (2023). One Does Not Simply Correct for Serial Dependence. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–15. https://doi.org/10.1080/10705511.2023.2173203

Aust, F., & Barth, M. (2022). *Papaja: Prepare reproducible APA journal articles with R Markdown.* Retrieved from https://github.com/crsh/papaja

Boker, S. M. (2015). Adaptive equilibria: A balancing act of regulation in development. *Journal for Person-Oriented Research*, *1*(1-2), 99–109. https://doi.org/10.17505/jpor.2015.10

Bos, F. M., Schreuder, M. J., George, S. V., Doornbos, B., Bruggeman, R., Krieke, L. van der, … Snippe, E. (2022). Anticipating manic and depressive transitions in patients with bipolar disorder using early warning signals. *International Journal of Bipolar Disorders*, *10*(1), 12. https://doi.org/10.1186/s40345-022-00258-4

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., … Kuppens, P. (2016). Assessing Temporal Emotion Dynamics Using Networks. *Assessment*, *23*(4), 425–435. https://doi.org/10.1177/1073191116645909

Brose, A., Lindenberger, U., & Schmiedek, F. (2013). Affective states contribute to trait reports of affective well-being. *Emotion*, *13*(5), 940–948. https://doi.org/10.1037/a0032401

Brose, A., Schmiedek, F., Lövdén, M., & Lindenberger, U. (2011). Normal aging dampens the link between intrusive thoughts and negative affect in reaction to daily stressors. *Psychology and Aging*, *26*(2), 488–502. https://doi.org/10.1037/a0022287

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

Canti, A., & Ripley, B. (2021). *Boot: Bootstrap R (S-Plus) Functions*. Retrieved from https://cran.r-project.org/package=boot

Chan, W., & Wei, W. W. S. (1992). A comparison of some estimators of time series autocorrelations. *Computational Statistics & Data Analysis*, *14*(2), 149–163. https://doi.org/10.1016/0167-9473(92)90170-K

Chow, S.-M., Ho, M., Hamaker, E. L., & Dolan, C. V. (2010). Equivalence and Differences Between Structural Equation Modeling and State-Space Modeling Techniques. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(2), 303–332. https://doi.org/10.1080/10705511003661553

Cole, P. M., Martin, S. E., & Dennis, T. A. (2004). Emotion Regulation as a Scientific Construct: Methodological Challenges and Directions for Child Development Research. *Child Development*, *75*(2), 317–333. https://doi.org/10.1111/j.1467-8624.2004.00673.x

Crevits, R., & Croux, C. (2019). Robust estimation of linear state space models. *Communications in Statistics - Simulation and Computation*, *48*(6), 1694–1705. https://doi.org/10.1080/03610918.2017.1422752

Croux, C., & Joossens, K. (2008). Robust Estimation of the Vector Autoregressive Model by a Least Trimmed Squares Procedure. In P. Brito (Ed.), *COMPSTAT 2008* (pp. 489–501). Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2084-3_40

Curran, P. J., & Bauer, D. J. (2007). Building path diagrams for multilevel models. *Psychological Methods*, *12*(3), 283–297. https://doi.org/10.1037/1082-989X.12.3.283

De Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2016). Get Over It! A Multilevel Threshold Autoregressive Model for State-Dependent Affect Regulation. *Psychometrika*, *81*(1), 217–241. https://doi.org/10.1007/s11336-014-9417-x

Dehon, C., Gassner, M., & Verardi, V. (2009). Beware of "Good" Outliers and Overoptimistic Conclusions. *Oxford Bulletin of Economics and Statistics*, *71*(3), 437–452. https://doi.org/10.1111/j.1468-0084.2009.00543.x

Dejonckheere, E., & Erbas, Y. (2021). Designing an experience sampling study. In I. Myin-Germeys & P. Kuppens (Eds.), *The open handbook of experience sampling methodology: A step-by-step guide to designing, conducting, and analyzing ESM studies* (pp. 33–70). Center for Research on Experience Sampling; Ambulatory Methods Leuven.

Driver, C. C., & Voelkle, M. C. (2018). Understanding the Time Course of Interventions with Continuous Time Dynamic Models. In K. van Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous Time Modeling in the Behavioral and Related Sciences* (pp. 79–109). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-77219-6_4

1132 Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood
1133 dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075
1134 Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022).
1135 The Effects of Sampling Frequency and Questionnaire Length on Perceived Burden, Compliance, and
1136 Careless Responding in Experience Sampling Data in a Student Population. *Assessment*, *29*(2), 136–151.
1137 https://doi.org/10.1177/1073191120957102
1138 Enders, C. K. (2013). Dealing With Missing Data in Developmental Research. *Child Development*
1139 *Perspectives*, *7*(1), 27–31. https://doi.org/10.1111/cdep.12008
1140 Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in*
1141 *Psychology*, *59*, 100785. https://doi.org/10.1016/j.newideapsych.2020.100785
1142 Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General*
1143 *Psychology*, *2*(3), 271–299. https://doi.org/10.1037/1089-2680.2.3.271
1144 Hamaker, E. L. (2012). Why Researchers Should Think "Within-Person": A Paradicmatic Rationale. In M.
1145 R. Mehl & T. S. Conner (Eds.), *Handbook of Research Methods for Studying Daily Life* (pp. 43–61). New
1146 York, NY: Guilford Publications.
1147 Hamaker, E. L., & Wichers, M. (2017). No Time Like the Present: Discovering the Hidden Dynamics in
1148 Intensive Longitudinal Data. *Current Directions in Psychological Science*, *26*(1), 10–15.
1149 https://doi.org/10.1177/0963721416666518
1150 Hamilton, J. D. (1994). *Time series analysis* (2nd ed.). Princeton: Princeton university press.
1151 Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, UK:
1152 Cambridge University Press.
1153 Henry, L., & Wickham, H. (2020). *Purrr: Functional Programming Tools*. Retrieved from
1154 https://CRAN.R-project.org/package=purrr
1155 Hunter, M. D. (2018). State Space Modeling in an Open Source, Modular, Structural Equation Modeling
1156 Environment. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 307–324.
1157 https://doi.org/10.1080/10705511.2017.1369354
1158 Keesman, K. J. (2011). *System Identification*. London: Springer London.
1159 https://doi.org/10.1007/978-0-85729-522-4
1160 Koller, M., & Stahel, W. A. (2011). Sharpening Wald-type inference in robust regression for small samples.
1161 *Computational Statistics & Data Analysis*, *55*(8), 2504–2515. https://doi.org/10.1016/j.csda.2011.02.014
1162 Koval, P., Brose, A., Pe, M. L., Houben, M., Erbas, Y., Champagne, D., & Kuppens, P. (2015). Emotional
1163 inertia and external events: The roles of exposure, reactivity, and recovery. *Emotion*, *15*(5), 625–636.
1164 https://doi.org/10.1037/emo0000059

Krone, T., Albers, C. J., & Timmerman, M. E. (2017). A comparative simulation study of AR(1) estimators in short time series. *Quality & Quantity*, *51*(1), 1–21. https://doi.org/10.1007/s11135-015-0290-1

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991. https://doi.org/10.1177/0956797610372634

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060. https://doi.org/10.1037/a0020962

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J. G., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depressive disorder in adolescence. *Emotion*, *12*(2), 283–289. https://doi.org/10.1037/a0025046

Kuppens, P., & Verduyn, P. (2015). Looking at Emotion Regulation Through the Window of Emotion Dynamics. *Psychological Inquiry*, *26*(1), 72–79. https://doi.org/10.1080/1047840X.2015.960505

Lucas, R. E. (2007). Adaptation and the Set-Point Model of Subjective Well-Being: Does Happiness Change After Major Life Events? *Current Directions in Psychological Science*, *16*(2), 75–79. https://doi.org/10.1111/j.1467-8721.2007.00479.x

Maechler, M., Rousseeuw, P. J., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., ... Palma, M. A. di. (2021). *Robustbase: Basic Robust Statistics R package*. Retrieved from https://cran.r-project.org/web/packages/robustbase/index.html

Malooly, A. M., Genet, J. J., & Siemer, M. (2013). Individual differences in reappraisal effectiveness: The role of affective flexibility. *Emotion*, *13*(2), 302–313. https://doi.org/10.1037/a0029980

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods* (1st ed.). Wiley. https://doi.org/10.1002/0470010940

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535–549. https://doi.org/10.1007/s11336-014-9435-8

Orzek, J. H., & Voelkle, M. C. (2023). Regularized continuous time structural equation models: A network perspective. *Psychological Methods*, *28*(6), 1286–1320. https://doi.org/10.1037/met0000550

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rousseeuw, P. J. (1991). Tutorial to robust statistics. *Journal of Chemometrics*, *5*(1), 1–20. https://doi.org/10.1002/cem.1180050103

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.2

Schilling, O. K., & Diehl, M. (2014). Reactivity to stressor pile-up in adulthood: Effects on daily negative and positive affect. *Psychology and Aging*, *29*(1), 72–83. https://doi.org/10.1037/a0035500

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n = 1 psychological autoregressive modeling. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01038

Suls, J., Green, P., & Hillis, S. (1998). Emotional Reactivity to Everyday Problems, Affective Inertia, and Neuroticism. *Personality and Social Psychology Bulletin*, *24*(2), 127–136. https://doi.org/10.1177/0146167298242002

Ushey, K. (2022). *Renv: Project Environments*. Retrieved from https://CRAN.R-project.org/package=renv

Vachon, H., Viechtbauer, W., Rintala, A., & Myin-Germeys, I. (2019). Compliance and Retention With the Experience Sampling Method Over the Continuum of Severe Mental Disorders: Meta-Analysis and Recommendations. *Journal of Medical Internet Research*, *21*(12), e14475. https://doi.org/10.2196/14475

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). *Dplyr: A Grammar of Data Manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Girlich, M. (2022). *Tidyr: Tidy Messy Data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Yohai, V. J. (1987). High Breakdown Point and High Efficiency Robust Estimates for Regression. *Annals of Statistics*, *15*, 642–656.

You, D., Hunter, M., Chen, M., & Chow, S.-M. (2020). A Diagnostic Procedure for Detecting Outliers in Linear State–Space Models. *Multivariate Behavioral Research*, *55*(2), 231–255. https://doi.org/10.1080/00273171.2019.1627659

Zautra, A. J., Berkhof, J., & Nicolson, N. A. (2002). Changes in affect interrelations as a function of stressful events. *Cognition & Emotion*, *16*(2), 309–318. https://doi.org/10.1080/02699930143000257

Zoubir, A. M., Koivunen, V., Chakhchoukh, Y., & Muma, M. (2012). Robust Estimation in Signal Processing: A Tutorial-Style Treatment of Fundamental Concepts. *IEEE Signal Processing Magazine*, *29*(4), 61–80. https://doi.org/10.1109/MSP.2012.2183773