The Annals of Applied Statistics
2023, Vol. 17, No. 1, 860–880
https://doi.org/10.1214/22-AOAS1655

MODELING A NONLINEAR BIOPHYSICAL TREND FOLLOWED BY LONG-MEMORY EQUILIBRIUM WITH UNKNOWN CHANGE POINT

BY WENYU ZHANG[1,a], MARYCLARE GRIFFIN[2,c] AND DAVID S. MATTESON[1,b]

[1]*Department of Statistics and Data Science, Cornell University,* [a]*wz258@cornell.edu,* [b]*dm484@cornell.edu*
[2]*Department of Mathematics and Statistics, University of Massachusetts Amherst,* [c]*maryclaregri@umass.edu*

Measurements of many biological processes are characterized by an initial trend period followed by an equilibrium period. Scientists may wish to quantify features of the two periods as well as the timing of the change point. Specifically, we are motivated by problems in the study of electrical cell-substrate impedance sensing (ECIS) data. ECIS is a popular new technology which measures cell behavior noninvasively. Previous studies using ECIS data have found that different cell types can be classified by their equilibrium behavior. However, it can be challenging to identify when equilibrium has been reached and to quantify the relevant features of cells' equilibrium behavior. In this paper we assume that measurements during the trend period are independent deviations from a smooth nonlinear function of time, and that measurements during the equilibrium period are characterized by a simple long memory model. We propose a method to simultaneously estimate the parameters of the trend and equilibrium processes and locate the change point between the two. We find that this method performs well in simulations and in practice. When applied to ECIS data, it produces estimates of change points and measures of cell equilibrium behavior which offer improved classification of infected and uninfected cells.

1. Introduction. We propose a model for time-series data that is characterized by two consecutive regimes which correspond to a highly nonstationary and nonlinear trend period and a stable equilibrium period. Often, researchers are interested in estimating the features of each regime as well as the timing of the transition or change point between the two.

We are motivated by the problem of detecting contamination of mammalian cell cultures by mycoplasma using electric cell-substrate impedance sensing (ECIS) data. Contamination of mammalian cell cultures is pervasive, costly, and can be challenging to detect (Gustavsson et al. (2019)). Specifically, contamination by mycoplasma is especially prevalent, occurring in up to 20% of cell cultures, while also expensive and time consuming to identify. As a result, there is a pressing need for the development of additional methods for detecting contamination by mycoplasma.

ECIS is a relatively new noninvasive method used to study cell attachment, growth, morphology, function, and motility (Keese (2019)). ECIS measurements have been used in numerous cell biology studies, from cancer biology and cytotoxicity (Hong et al. (2011), Opp et al. (2009)). Because ECIS measurements have been used to differentiate between cancerous and noncancerous cells and to classify cell lines (Gelsinger, Tupper and Matteson (2020), Lovelady et al. (2007)), it is hypothesized that they may also be used to identify cell cultures contaminated by mycoplasma.

ECIS measurements are obtained by growing cells in a well on top of small gold-film electrodes, between which alternating current is applied and electrical impedance is measured. As cells grow, they cover the electrode, and resistance, a component of impedence,

Received November 2020; revised May 2022.
Key words and phrases. Applied biophysics, change-point analysis, fractionally integrated process, long memory, time series.

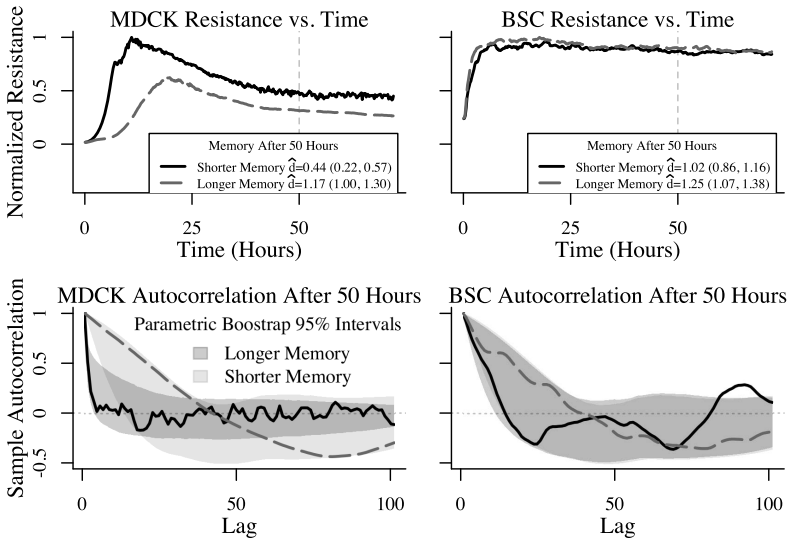


FIG. 1. *The first row shows examples of resistance measurements at* 500 *hertz for MDCK and BSC cell line samples cultivated in BSA and gel, respectively. For each cell line, one example of resistance measurements displaying shorter memory and one example of resistance measurements displaying longer memory are selected. Approximate* 95% *confidence intervals for memory parameter estimates* $\hat{d}$ *are based on* 10,000 *parametric bootstrap samples. A light gray vertical line at* 50 *hours is provided to indicate a conservative estimate of the onset of the equilibrium regime. The second row shows the corresponding sample autocorrelations for resistance measurements after* 50 *hours, accompanied by approximate* 95% *confidence intervals for sample autocorrelations, based on* 10,000 *parametric bootstrap samples.*

increases. Eventually, the cells fill the well and growth ceases. In some cases, cell death occurs due to overcrowding, causing a small drop in resistance measurements after the peak. After this point an equilibrium period begins. Resistance fluctuations during equilibrium are caused by cell micromotion. The equilibrium period is sometimes called confluence in the ECIS literature, and it continues until the cells exhaust their resources and begin to die. The first row of Figure 1 shows a subset of resistance measurements for two different cell types, Madin–Darby canine kidney (MDCK) cells and epithelial cells of African green monkey kidney origin (BSC cells). All show a nonlinear trend period, followed by an equilibrium period, with a more visually obvious change point present in the MDCK cells.

Equilibrium measurements are especially informative. They are believed to be less sensitive to initial conditions than features of the trend period and are possibly nonstationary and display long-range dependence, meaning that the correlations between successive measurements decay very slowly over time and can be characterized parsimoniously by a very simple three-parameter long memory time series model with parameters that are constant over the entire equilibrium period(Lovelady et al. (2007), Tarantola et al. (2010)). Long-range dependence has also been observed in wind speed and inflation data (Haslett and Raftery (1989), Doornik and Ooms (2004)) and can be modeled as a long memory time series, also known as a Gaussian fractionally integrated (FI) or long-memory process which has three parameters, an overall mean μ, variance ν^2, and a scalar long-memory (fractional differencing) parameter d that governs how quickly autocorrelations decay. Ideally, if these parameters could be estimated well, they could be used to quantify features of the equilibrium regime in the context of ECIS data.

The subset of resistance measurements shown in Figure 1 were chosen to illustrate the fact that equilibrium measurements can show evidence of stationary and nonstationary behavior and variable strength of long-range dependence. Let $y_1, y_2, \ldots, y_T \in \mathbb{R}$ be a sequence of time-ordered measurements at $t = 1, 2, \ldots, T$, respectively, and let τ_{50} refer to the index of

the first measurement collected after 50 hours. We quantify the strength of long-range dependence by comparing maximum likelihood estimates of the differencing parameter d under the model $\sum_{i=0}^{t-1} \binom{d}{i}(-1)^i (y_{t-i} - \mu)\mathbb{1}_{t-i \geq \tau_{50}} \sim \mathrm{N}(0, \nu^2)$ applied to each selected series. Estimates of d and corresponding 95% confidence intervals, based on the parametric bootstrap with 10,000 simulated time series, are provided in Figure 1; they range from weaker long-range dependence with $\hat{d} = 0.44$ to much stronger long-range dependence with $\hat{d} = 1.25$. Estimates of d also provide evidence of both stationarity and nonstationarity, as values of $d < 0.5$ and $d \geq 0.5$ correspond to stationary and nonstationary processes, respectively. The longer memory MDCK resistance measurements and both BSC resistance measurements show strong evidence of nonstationarity, with 95% confidence intervals for the differencing parameter exceeding 0.5. The second row of Figure 1 shows sample autocorrelation functions for the subset of resistance measurements shown in the first row after 50 hours, at which point equilibrium has been achieved for all four selected time series. Because the true autocorrelations are not defined for nonstationary processes, we include a comparison to approximate 95% intervals for sample autocorrelations obtained under the estimated model for each selected set of resistance measurements acquired from using the parametric bootstrap with 10,000 simulated time series.

Unfortunately, the long-memory parameter d is notoriously difficult to estimate in finite samples. Furthermore, the change point from trend to confluence phase, which determines the amount of data available to estimate d, is typically not precisely known in practice. Standard practice is to use a fixed time point, for example, 50 hours, as a conservative estimate of the start of the confluence regime (Tarantola et al. (2010)). This underutilizes the data, potentially resulting in poorer estimates of the parameters of interest. Furthermore, such a conservative estimate could incorrectly characterize the preceding trend phase. This suggests the need for a change-point detection method, which can identify when the trend phase gives way to confluence, specifically an unsupervised method that can detect the transition from a nonstationary model to a FI model.

To our knowledge, existing methods for change-point detection are not appropriate. Some existing methods assume short-memory autoregressive moving average (ARMA) models, long-memory FI models, or other restrictive parametric models both before and after the change point (Chen and Liu (1993), Dufrenot, Guegan and Peguin-Feissolle (2008), Killick, Fearnhead and Eckley (2012)). Others, including the popular E-Divisive algorithm, assume that measurements between change points are independent or identically distributed or assume that change points strictly correspond to level shifts or isolated outliers (Matteson and James (2014), Zhang, Gilbert and Matteson (2019)). Alternative methods in the biomedical fields tend to be too domain specific to apply to the the problem of detecting the change point between the trend and confluence phase in ECIS data (Olshen et al. (2004), Nika, Babyn and Zhu (2014)).

In this paper we develop a novel method for estimating a change point between a highly nonstationary and nonlinear trend period and a stable equilibrium period that is characterized by an FI process. The method we introduce simultaneously obtains estimates of the nonlinear trend function and the FI parameters and can accommodate heavier-than-normal tailed data which is often observed in real biological applications. We apply this method to the detection of contamination by M. hominis, a species of mycoplasma, in MDCK cells and BSC cells using ECIS measurements. Related literature on ECIS measurements of cell behavior supports the presence of a single change point in this data (Lovelady et al. (2007), Tarantola et al. (2010)). The available data consists of four experiments per cell type. Each experiment corresponds to ECIS measurements on cells on a single tray of 96 wells obtained over the course of at least 72 hours. Of the 96 wells, 16 are left empty, 32 contain uncontaminated cells, and 48 contain cells contaminated by mycoplasma. In order to mimic lab-to-lab variability in cell

culture preparation, wells were prepared using either of two different types of media. Half were prepared using bovine serum albumin (BSA), and half were prepared using gel. Within an experiment, wells containing the same media and cells with the same contamination status can be thought of as replicates.

In Section 2 we propose a model, which we call Trend-to-Confluence Detector (T2CD), for data which display highly nonstationary and nonlinear trend period followed by a stable, equilibrium period with long-range dependence. In Section 3 we discuss estimation of the parameters of the model introduced in Section 2. We consider both an exact estimation procedure, which we call T2CD-step, as well as an generalized estimation procedure which has greater computational scalability for longer time series, which we call T2CD-sigmoid. We demonstrate the performance of T2CD-step and T2CD-sigmoid in simulations and find that T2CD-sigmoid provides substantial computational advantages and better estimation of the changepoint than T2CD-step in Section 4. We apply T2CD-step and T2CD-sigmoid to the ECIS data shown in Figure 1 and use the estimated change points and FI parameters to better classify cells by contamination status in Section 5.

2. Trend-to-Confluence Detector (T2CD) model.

2.1. *Overview.* We assume that the measurements y_t belong to two successive regimes, a trend regime and an equilibrium or confluent regime. Let τ denote the change point time index. We assume that

$$y_t = f(t; \boldsymbol{\beta}) + \eta_t \quad \text{for } t < \tau, \tag{1}$$

$$y_t = g(y_1, \ldots, y_{t-1}; \mu, d, \tau) + \epsilon_t \quad \text{for } t \geq \tau, \tag{2}$$

where $\tau_a \leq \tau \leq \tau_b$, $\eta_t \sim \mathrm{N}(0, \exp\{h(t; \boldsymbol{\theta})\})$ and ϵ_t are random variables with mean 0 and variance ν^2 that are either normally distributed, according to $\epsilon_t/\nu \sim \mathrm{N}(0, 1)$, or t-distributed, according to $\sqrt{\xi/(\xi-2)}\epsilon_t/\nu \sim t_\xi$ with degrees of freedom $\xi > 2$, depending on whether or not equilibrium measurements display heavier-than-normal tails. If trend regime measurements display heavier-than-normal tails, they can be captured by the time-varying trend variances $\exp\{h(t; \boldsymbol{\theta})\}$. The minimum and maximum values of the change point τ_a and τ_b are assumed to be prespecified according to application-specific domain knowledge of the change point location. In the absence of a priori information, $\tau_a = 0$ and $\tau_b = T$. The noise terms η_t and ϵ_t, which encompass measurement errors and random fluctuations due to continuous cell growth, motility, and death, are assumed to be independent within and across the two regimes.

During the first regime ($t < \tau$), the measurement at time t will be centered around a trend curve $f(t; \boldsymbol{\beta})$ which is a function of time t and fixed but unknown parameters $\boldsymbol{\beta}$. The noise terms η_t are possibly heteroscedastic with variance $\exp\{h(t; \boldsymbol{\theta})\}$ to reflect different degrees of uncertainty in the measurements when the cell culture undergoes different rates of growth and death. During the second (equilibrium) regime ($t \geq \tau$), the measurement at time t will be centered about a function $g(y_1, \ldots, y_{t-1}; \mu, d, \tau)$ of previous measurements $y_1, \ldots, y_{t-1}$ and fixed but unknown parameters μ and d. The noise terms ϵ_t are homoscedastic with fixed but unknown variance ν^2, since the cell culture is in equilibrium and not undergoing drastic changes. We describe our modeling choices of the two regimes in the following sections.

2.2. *Trend.* Resistance measurements in the first regime, or the trend phase, are characterized by a trend of initial steep increase sometimes followed by a slight drop after the peak as well as heteroscedasticity with higher variance at the stage of rapid cell growth. As mentioned above, the trend curve is denoted as $f(t; \beta)$. Depending on the trend, any appropriate parametric, semiparametric, or nonparametric model can be used to fit the first regime. The exact formulation of the trend curve can depend on the application domain and the choice

of the user. For the ECIS application that we focus on in this paper, we assume a smooth trend curve. This is in line with visual inspection of real ECIS data in Figure 1 and that cell growth, motility, death, and other functions are continuous processes. We utilize penalized splines (Ruppert, Wand and Carroll (2003)) for their flexibility to capture the ECIS trend phase, since it is highly nonstationary. We similarly use penalized splines in modeling the noise variance.

We denote the matrix of B-spline basis functions $B_{i,D}(u)$ of degree D_f, evaluated on time indices for the trend as $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_T') \in \mathbb{R}^{T \times (Q_f + D_f + 1)}$, where Q_f is the number of distinct interior knots. Similarly, we denote $\mathbf{V} = (\mathbf{v}_1', \mathbf{v}_2', \ldots, \mathbf{v}_T')$ as the $T \times (Q_h + D_h + 1)$ matrix of B-spline basis functions for the log variances of the noise terms. The model for the first regime takes the form

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \eta_t,$$

where $\eta_t \sim N(0, \exp\{\mathbf{v}_t' \boldsymbol{\theta}\})$. Let the fitted spline for the trend be $s(t) = \mathbf{x}_t' \beta$. We impose the smoothness penalty $\lambda_f \int \hat{s}''(u)^2 du = \lambda_f \boldsymbol{\beta}' \boldsymbol{M}_f \boldsymbol{\beta}$ on the spline estimate to prevent overfitting, where $\lambda_f > 0$ is a scalar that determines the smoothness of the fitted spline and $\boldsymbol{M}_f$ is a matrix with elements that are fixed given the matrix of B-spline basis functions $\mathbf{X}$. An equivalent smoothness penalty $\lambda_h \boldsymbol{\theta}' \boldsymbol{M}_h \boldsymbol{\theta}$ is imposed on the fit for the log variances of the noise terms, where λ_h is another smoothness parameter and $\boldsymbol{M}_h$ is a matrix with elements that are fixed, given the matrix of B-spline basis functions $\mathbf{V}$.

2.3. *Equilibrium.* Starting at time index τ, measurements are centered about a function $g(y_1, \ldots, y_{t-1}; \mu, d, \tau)$ of previous measurements $y_1, \ldots, y_{t-1}$ and fixed but unknown parameters μ, and d that corresponds to the conditional mean function of a fractionally integrated (FI) process,

$$g(y_1, \ldots, y_{t-1}; \mu, d, \tau) = \mu - \sum_{i=1}^{t-1} \binom{d}{i} (-1)^i (y_{t-i} - \mu) \mathbb{1}_{t-i \geq \tau}. \tag{3}$$

This captures long-range dependence of the measurements in confluence. The parameter d plays the role of the long-memory parameter in a FI model (Sowell (1992)). The FI model assumes that observed values of a time series y_t satisfy $(1-B)^d y_t = \epsilon_t$, where B is the differencing operator $B^k y_t = y_{t-k}$ and ϵ_t are random variables with mean 0 and variance ν^2 that are either normally distributed, according to $\epsilon_t / \nu \sim \mathrm{N}(0, 1)$, as FI models are commonly defined, or t-distributed according to $\sqrt{\xi/(\xi-2)}\epsilon_t / \nu \sim t_\xi$ with degrees of freedom $\xi > 2$, if equilibrium measurements show evidence of heavier-than-normal tails. Values of $d > 0$ correspond to processes that are said to have long memory with larger values of d indicating more slowly decaying autocorrelations over time. Specifically, the autocorrelation function $\mathrm{Corr}(y_t, y_{t-k})$ exhibits hyperbolic decay: as $k \to \infty$, $\mathrm{Cor}(y_t, y_{t-k}) \to (\Gamma(1-d)/\Gamma(d)) k^{2d-1}$ (Baillie (1996)). When $d < 1$, the FI process is mean reverting, and when $d < 0.5$ the FI process is stationary.

2.4. *An extension to multivariate data.* To accommodate settings where p related time series may be observed contemporaneously, we provide an extension to multivariate time series data $\mathbf{Y} \in \mathbb{R}^{T \times p}$. We assume that all p time series share a common long-memory parameter d, but have their own change point τ_j, trend parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j$, and equilibrium mean and variance μ_j and ν_j^2, and, when a t-distribution is assumed for equilibrium errors, degrees of freedom ξ_j. Specifically, we assume

$$y_{t,j} = f(t; \boldsymbol{\beta}_j) + \eta_{t,j} \quad \text{for } t < \tau_j, \tag{4}$$

$$y_{t,j} = g(y_{1,j}, \ldots, y_{t-1,j}; \mu_j, d, \tau_j) + \epsilon_{t,j} \quad \text{for } t \geq \tau_j, \tag{5}$$

where $\eta_{t,j} \sim \mathrm{N}(0, \exp\{h(t;\boldsymbol{\theta}_j)\})$ and $\epsilon_{t,j}$ are random variables with mean 0 and variance ν_j^2 that are either normally distributed, according to $\epsilon_{t,j}/\nu_j \sim \mathrm{N}(0,1)$, or t-distributed, according to $\sqrt{\xi_j/(\xi_j-2)}\epsilon_{t,j}/\nu_j \sim t_{\xi_j}$ with degrees of freedom $\xi_j > 2$, depending on whether or not equilibrium measurements display heavier-than-normal tails.

This is motivated by the ECIS measurements, described in Section 1, where the p related time series correspond to wells containing cells of the same type, contamination status, and media in the same experiment which may have varying initial conditions but common equilibrium behavior. We account for varying initial conditions, such as the number of cells deposited, by allowing each well to have its own varying change point τ_j, trend parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j$, and equilibrium means μ_j, variances ν_j^2, and, when a t-distribution is assumed for $\epsilon_{t,j}$, degrees of freedom ξ_j. A shared long-memory parameter d reflects the cells' common equilibrium or confluence behavior.

3. Estimation.

3.1. *Exact estimation for univariate data*: *T2CD-step*. First, we introduce a strategy for estimating the T2CD parameters that we call T2CD-step, because it performs a complete search over the change point location space $[\tau_a, \tau_b]$. When equilibrium errors ϵ_t are assumed to be normal, we find the change point $\tau_a \le \hat{\tau} \le \tau_b$, which maximizes

$$
\begin{aligned}
&-\sum_{t=1}^{\tau-1}\left(\frac{\boldsymbol{v}_t'\boldsymbol{\theta}}{2}+\frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\exp\{\boldsymbol{v}_t'\boldsymbol{\theta}\}}\right)\\
&-\frac{1}{2\nu^2}\sum_{t=\tau}^{T}\big(y_t-g(y_1,\ldots,y_{t-1};\mu,d,\tau)\big)^2-\left(\frac{T-\tau+1}{2}\right)\log(\nu^2)+\text{constant},
\end{aligned}
\tag{6}
$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are estimated for each candidate change point using an iterative Feasible Generalized Least Squares (FGLS) procedure (Kuan (2004)) to estimate the spline coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, with penalty parameters λ_f and λ_h for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ chosen according to leave-one-out cross-validation, as implemented in `smooth.spline` in `R` (R Foundation for Statistical Computing (2018)) which selects the smoothness penalties by golden-section search. A more detailed explanation of the FGLS procedure is provided in Web Appendix A.

When equilibrium errors ϵ_t are assumed to be t-distributed, we find the change point $\tau_a \le \hat{\tau} \le \tau_b$, which maximizes

$$
\begin{aligned}
&-\sum_{t=1}^{\tau-1}\left(\frac{\log(2)}{2}+\frac{\boldsymbol{v}_t'\boldsymbol{\theta}}{2}+\frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\exp\{\boldsymbol{v}_t'\boldsymbol{\theta}\}}\right)\\
&-\frac{\xi+1}{2}\sum_{t=\tau}^{T}\log\left(1+\frac{(y_t-g(y_1,\ldots,y_{t-1};\mu,d,\tau))^2}{\nu^2(\xi-2)}\right)\\
&+(T-\tau+1)\left[\log\left(\Gamma\left(\frac{\xi+1}{2}\right)/\Gamma\left(\frac{\xi}{2}\right)\right)-\log(\nu^2(\xi-2))\right]+\text{constant},
\end{aligned}
\tag{7}
$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are estimated for each candidate change point using the same iterative feasible generalized least squares (FGLS) procedure used when normal errors are assumed.

Setting a candidate range for change point τ restricts the search region to reduce computational costs and ensures the availability of points in both regimes to estimate the respective model parameters. The candidate range hence should leave sufficient number of points at both ends of the time series, while still containing the true change point. Determination of the candidate range is application specific and should be guided by domain knowledge.

Given a candidate change location, the log-likelihood can be decomposed into one component that involves the values of the time series during the trend regime $y_1, \ldots, y_{\tau-1}$, and the parameters of the trend regime, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ and another component that involves the values of the time series during the equilibrium period $y_\tau, \ldots, y_T$, and the parameters of the equilibrium period, μ, d, ν^2, and, when a t-distribution is assumed for ϵ_t, ξ. It follows that the parameters of the trend and equilibrium regime can be estimated simultaneously from the trend and equilibrium data, respectively. We estimate the long memory parameter d by maximizing equation (6) over $d \in \mathbb{R}$.

3.2. *Generalized estimation for univariate data*: *T2CD-sigmoid*. In practice, maximizing (6) can be prohibitively computationally demanding and time consuming if there are many candidate change points, as is the case when the observed time series is long. Accordingly, we introduce a generalization to the estimation procedure that we call T2CD-sigmoid. Let $w(t; \boldsymbol{\alpha})$ denote a transition function that takes on values in the interval $[0, 1]$, then we can define the mean function in the second regime $g(y_1, \ldots, y_{t-1}; \mu, d, \tau)$, as defined in (3), as a special case of

$$\mu - \sum_{i=1}^{t-1} \binom{d}{i} (-1)^i (y_{t-i} - \mu) w(t-i; \boldsymbol{\alpha}), \tag{8}$$

where $w(t; \boldsymbol{\alpha})$ has a single parameter α that corresponds to the change point τ and $w(t; \alpha) = \mathbb{1}_{t-i \geq \alpha}$ takes the form of a step function. This suggests that an alternative approach would be to replace the discrete step transition function $\mathbb{1}_{t-i \geq \tau}$ with a continuous sigmoid transition function $w(t; \boldsymbol{\alpha}) = (1 + \exp\{-\alpha_0 - \alpha_1 t\})^{-1}$ which is parameterized by a pair of real-valued parameters $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1\}$. We denote the corresponding second regime mean function as

$$\begin{aligned} &\tilde{g}(y_1, \ldots, y_{t-1}; \mu, d, \alpha_0, \alpha_1) \\ &\quad = \mu - \sum_{i=1}^{t-1} \binom{d}{i} (-1)^i (y_{t-i} - \mu) \big(1 + \exp\{-\alpha_0 - \alpha_1 (t-i)\}\big)^{-1}. \end{aligned}$$

The parameters α_0 and α_1 determine the timing of the transition from trend to equilibrium phase which corresponds to the inflection point of the transition function $(1 + \exp\{-\alpha_0 - \alpha_1 t\})^{-1}$. The change point is estimated as when the transition function is at 0.5, that is, $\hat{\tau} = -\frac{\hat{\alpha}_0}{\hat{\alpha}_1}$. The timing of the transition can be constrained to the interval $[\tau_a, \tau_b]$ by adding a penalty $C(w(\tau_b; \alpha_0, \alpha_1) - w(\tau_a; \alpha_0, \alpha_1))$ with fixed penalty parameter $C > 0$ to the objective function.

Using a smooth transition function can offer computational speed-ups because the log-likelihood can be differentiated with respect to the parameters that determine the timing of the transition, α_0 and α_1, and accordingly does not require an exhaustive search over all candidate change points.

When equilibrium errors ϵ_t are assumed to be normal, T2CD-sigmoid finds the values of α_0 and α_1, which maximize

$$\begin{aligned} &- \sum_{t=1}^{T} \big(1 - w(t; \alpha_0, \alpha_1)\big) \left(\frac{\boldsymbol{v}_t' \boldsymbol{\theta}}{2} + \frac{(y_t - \boldsymbol{x}_t' \boldsymbol{\beta})^2}{2 \exp\{\boldsymbol{v}_t' \boldsymbol{\theta}\}} \right) \\ &- \frac{1}{2\nu^2} \sum_{t=1}^{T} w(t; \alpha_0, \alpha_1) \big(y_t - \tilde{g}(y_1, \ldots, y_{t-1}; \mu, d, \alpha_0, \alpha_1)\big)^2 \\ &- \frac{1}{2} \sum_{t=1}^{T} w(t; \alpha_0, \alpha_1) \log\big(\nu^2\big) + C\big(w(\tau_b; \alpha_0, \alpha_1) - w(\tau_a; \alpha_0, \alpha_1)\big) + \text{constant}, \end{aligned} \tag{9}$$

where $C > 0$ is a constant that can be set to ensure that the inflection point of the smooth transition function occurs between τ_a and τ_b. When equilibrium errors ϵ_t are assumed to be t-distributed, T2CD-sigmoid finds the values of α_0 and α_1 which maximize

$$
\begin{aligned}
&-\sum_{t=1}^{T}(1-w(t;\alpha_0,\alpha_1))\left(\frac{\log(2)}{2}+\frac{\boldsymbol{v}_t'\boldsymbol{\theta}}{2}+\frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\exp\{\boldsymbol{v}_t'\boldsymbol{\theta}\}}\right)\\
&\quad-\frac{\xi+1}{2}\sum_{t=1}^{T}w(t;\alpha_0,\alpha_1)\log\left(1+\frac{(y_t-\tilde{g}(y_1,\ldots,y_{t-1};\mu,d,\alpha_0,\alpha_1))^2}{\nu^2(\xi-2)}\right)\\
&\quad+\sum_{t=1}^{T}w(t;\alpha_0,\alpha_1)\left[\log\left(\Gamma\left(\frac{\xi+1}{2}\right)/\Gamma\left(\frac{\xi}{2}\right)\right)-\log(\nu^2(\xi-2))\right]\\
&\quad+C\big(w(\tau_b;\alpha_0,\alpha_1)-w(\tau_a;\alpha_0,\alpha_1)\big)+\text{constant}.
\end{aligned}
\tag{10}
$$

The generalized log-likelihood resembles a weighted log-likelihood, which has been intuitively motivated as downweighting the model likelihood when model misspecification is suspected as in the literature on tempered likelihoods for Bayesian inference (Thomas and Corander (2019)).

The log-likelihood used by T2CD-sigmoid cannot be decomposed into two components, one of which involves the values of the time series during the trend period and corresponding parameters and another component that involves the values of the time series during the equilibrium period and the corresponding parameters. Fortunately, B-spline bases are flexible enough to fit local trends. Accordingly, the first step of T2CD-sigmoid is to estimate the trend regime parameters from the entire time series by maximizing

$$
-\sum_{t=1}^{T}\left(\frac{\boldsymbol{v}_t'\boldsymbol{\theta}}{2}+\frac{(y_t-\boldsymbol{x}_t'\boldsymbol{\beta})^2}{2\exp\{\boldsymbol{v}_t'\boldsymbol{\theta}\}}\right). \tag{11}
$$

Again, we use an iterative feasible generalized least squares (FGLS) procedure (Kuan (2004)) to estimate the spline coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, with tuning parameters λ_f and λ_h chosen according to leave-one-out cross validation, as implemented in `smooth.spline` in the `R` environment (R Foundation for Statistical Computing (2018)). Via simulations provided in Web Appendix A, we show that RMSEs of trend component estimated over the entire time series and over the true trend regime data alone are similar.

Having obtained estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we can set C to be on the order of the log-likelihood component in equation (9) at the estimated values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in order to place approximately equal weight on model fitting and change-point regularization. While alternative procedures, such as using cross-validation to choose C can be used, we find that this simpler strategy performs well empirically by encouraging the inflection point of the transition function to occur in the interval $[\tau_a, \tau_b]$. Having now also fixed C, we can maximize (9) with respect to $\boldsymbol{\alpha}$, d, μ, ν^2, and, when a t-distribution is assumed for ϵ_t, ξ.

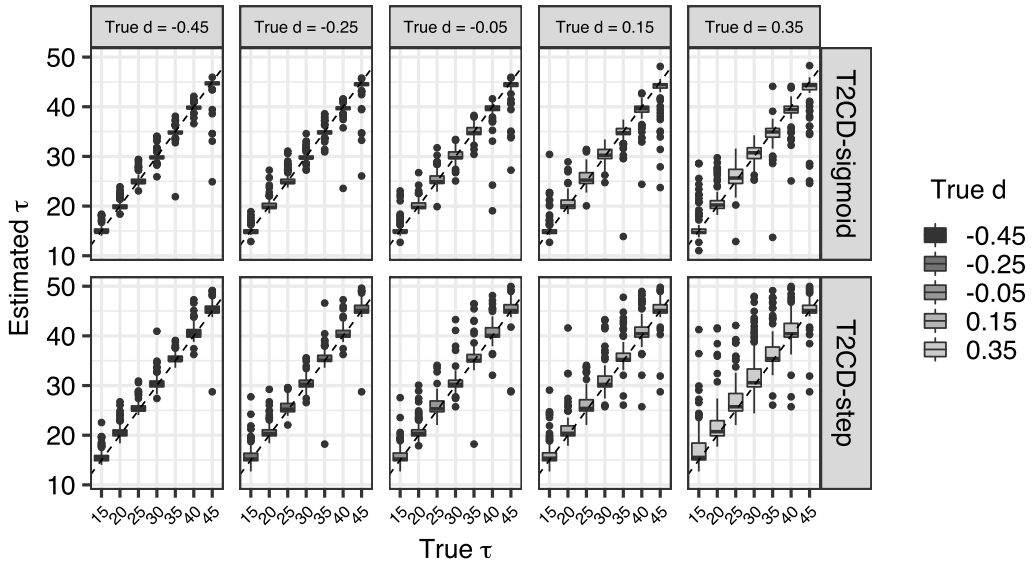
3.3. *Exact and generalized estimation for multivariate data.* When there are p replicates of the sequences, the log-likelihood function is a sum of the log-likelihoods of the individual sequences. The only constraint is that the long-memory parameter d is shared across dimensions, as described in Section 2.4. Recall that the change locations are allowed to differ across replicates, the number of possible combinations for change locations is m^p, where m is the number of time indices in $[\tau_a, \tau_b]$. An exhaustive search for the best combination is often computationally prohibitive. For this reason we use the following iterative procedure. First, we run either T2CD-step or T2CD-sigmoid on each univariate sequence to obtain estimates

of $\boldsymbol{\beta}_j, \boldsymbol{\theta}_j, d_j, \mu_j, \nu_j^2, \xi_j$ (when a t-distribution is assumed for ϵ_t) and τ_j for T2CD-step or $\boldsymbol{\alpha}_j$ for T2CD-sigmoid. Second, fixing the estimates of $\boldsymbol{\beta}_j$, $\boldsymbol{\theta}_j$, and τ_j for T2CD-step or $\boldsymbol{\alpha}_j$ for T2CD-sigmoid, we then optimize over d, μ_j, ν_j^2, and, when a t-distribution is assumed for ϵ_t, ξ_j initializing d at the mean univariate estimate across all of the time series $p^{-1}\sum_{j=1}^{p}\hat{d}_j$ and μ_j, ν_j^2, and, when a t-distribution is assumed for ϵ_t, ξ_j at the univariate estimates. We then iterate the process starting with rerunning step 1 with all parameters initialized at their current estimated values. The process terminates after a specified maximum number of iterations or when the difference between the two most recent objective values is below a tolerance threshold, whichever is earlier. In our experiments we set the maximum number of iterations to be 10 and the tolerance threshold to be 10^{-6}.

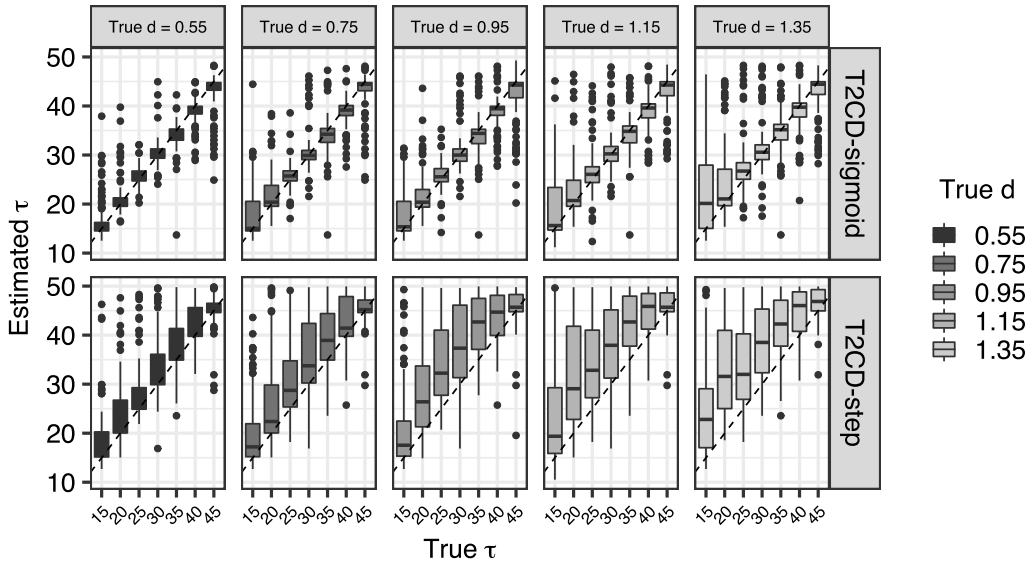
4. Simulation study. We evaluate the performance of T2CD-step and T2CD-sigmoid for estimating τ and d under several different scenarios, using both univariate and multivariate time series data. Because assuming t-distributed equilibrium errors slows computation substantially, we strictly consider normally distributed equilibrium errors in simulations. We set up the simulations to be similar to the ECIS data described in Section 1. First, we consider one simple scenario and compare estimates of the change point τ, obtained by T2CD-step and T2CD-sigmoid, in order to examine how generalizing the discrete transition using a smooth transition function affects change-point estimation. We then consider a broader set of scenarios and compare T2CD-step and T2CD-sigmoid not only to each other but also to several alternative methods.

We simulate univariate time series for comparing T2CD-step and T2CD-sigmoid as follows. Given a fixed change point τ, we simulate trend curves $\boldsymbol{f} = (f_1, \ldots, f_\tau)$ from a mean zero Gaussian process with squared exponential kernel $\mathrm{Cov}[f_t, f_s] = 10\exp(-0.5(s-t)^2)$. We simulate trend regime measurements $y_t = f_t + \eta_t$, where η_t are mean zero heteroscedastic measurement errors with standard deviation $\sigma_t = \frac{2-0.1}{\max\{f_s\}_{s=1}^{\tau} - \min\{f_s\}_{s=1}^{\tau}}[f_t - \min\{f_s\}_{s=1}^{\tau}] + 0.1$. We simulate equilibrium measurements $y_{\tau+1}, \ldots, y_T$, according to the FI model $(1-B)^d y_t = \epsilon_t$, where $\epsilon_t \sim \mathrm{N}(0, 0.25)$. For comparison with the observed ECIS data, we simulate univariate time series of length $T = 420$, which we can think of as 70 hours of data. For each combination of true change points τ set to values in the interval $[90, 270]$ chosen to correspond to change points at $\{15, 20, \ldots, 45\}$ hours and long memory parameter $d \in \{-0.25, -0.05, \ldots, 1.45\}$, we simulate 100 univariate time series. When applying T2CD-step and T2CD-sigmoid to each simulated univariate time series, we set the candidate range of τ to $[\tau_a = 10, \tau_b = 50]$. This mimics the structure of the ECIS data, where domain knowledge suggests that the candidate range $[\tau_a = 10, \tau_b = 50]$ for τ is appropriate. We also use degree three B-spline basis functions 3 with knots at every integer value of t when fitting β, and knots at every integer multiple of 5 when fitting θ. For T2CD-sigmoid we fix $C = 1000$ throughout and increase C by increments of 1000 when estimated change point does not fall within the candidate range. We check the choice of these hyperparameters in Figure 10 through residual analysis. Extensive studies on hyperparameter tuning is beyond the scope of this work.

The performance of estimates of τ are shown in Figure 2. Estimated change points for T2CD-sigmoid are set to the time index when the smooth transition function is equal to 0.5. Performance of estimates of τ, based on T2CD-step and T2CD-sigmoid, varies little with the true change point. Both T2CD-step and T2CD-sigmoid estimate the change point τ well when d is much smaller than 0.5. A more detailed investigation of how estimation of τ varies with the true value of d is available in Web Appendix B. In general, T2CD-step tends to overestimate τ and produces poorer estimates of τ as the true value of d increases. This is especially evident when d is close to or greater than 0.5; T2CD-step tends to overestimate



(a) Estimates of τ when the second regime is stationary at $d < 0.5$.

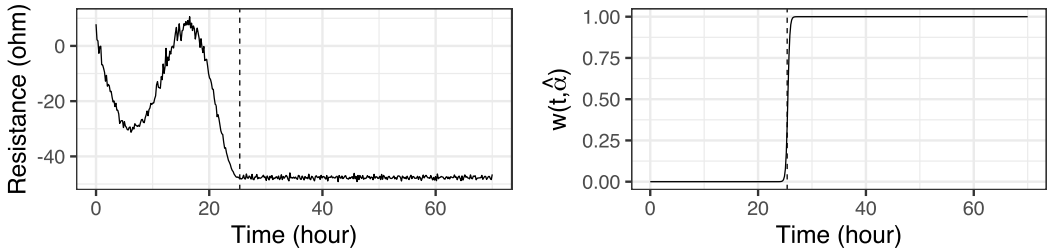


(b) Estimates of τ when the second regime is nonstationary at $d > 0.5$. T2CD-step tends to overestimate τ because nonstationarity can be mistaken for the first regime. The overestimation issue is less severe for T2CD-sigmoid because the smooth transition function accommodates uncertainty about the change point.

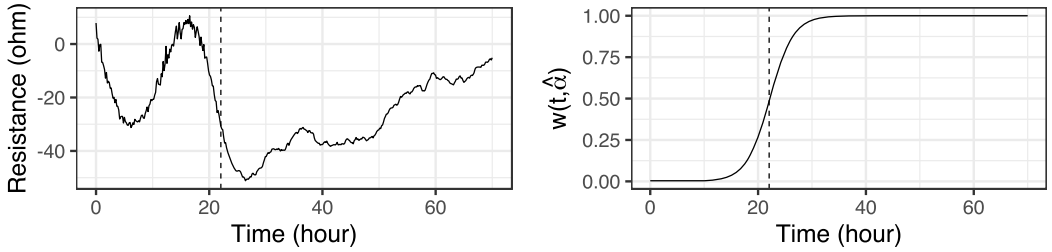
FIG. 2. *T2CD estimates of τ for simulation setup where the first regime is generated via Gaussian process with squared exponential kernel and the second regime generated via FI(d).*

τ while T2CD-sigmoid continues to estimate τ well, on average. We hypothesize that, as d approaches and exceeds 0.5, the change point is more difficult to recover because the long-range autocorrelations between equilibrium measurements can yield smoothly varying time trends during the equilibrium period which can be mistaken for a continuation of the trend period.

In order to understand why T2CD-sigmoid provides better change-point estimates than T2CD-step, when d is close to or greater than 0.5, we zoom in on a pair of estimated smooth transition functions from simulations with $d = -0.45$ and $d = 1.35$ in Figure 3, with true



(a) Estimates by T2CD-sigmoid for simulation with $d = -0.45$.



(b) Estimates by T2CD-sigmoid for simulation with $d = 1.35$.

FIG. 3. *T2CD-sigmoid estimates for simulation setup where the first regime is generated via Gaussian process with squared exponential kernel and the second regime generated via FI(d). The vertical dashed line marks the time index when the regime transition function is estimated to cross* 0.5. *The estimated transition is less abrupt for large d.*

$\tau = 25$. We observe that the estimated transition function is much steeper and more similar to the discrete transition function assumed by T2CD-step when $d = -0.45$. By allowing a smooth transition function, T2CD-sigmoid can accommodate greater uncertainty about the change point when d is close to or greater than 0.5.

The bias in T2CD-step's estimation of τ is also accompanied by uncertainty of the estimate. To quantify the uncertainty in the parameter estimates, confidence intervals (CI) can be constructed by assuming that the sampling distribution of the parameter estimates is approximately normal and approximating the standard deviation of the parameter estimates using a parametric bootstrap procedure. The parametric bootstrap procedure uses the standard deviation of parameter estimates, obtained by fitting T2CD-step to data simulated from the data model, with unknown parameters set to the T2CD-step parameter estimates computed from the original data. For each of the two samples in Figure 3, we draw 500 bootstrap samples with T2CD-step parameter estimates to construct approximate 95% CI. When the second regime is stationary with true $d = -0.45$, T2CD-step estimates $\hat{d} = -0.532$ with CI $(-0.636, -0.447)$ and $\hat{\tau} = 24.9$ with CI $(24.4, 27.9)$. When the second regime is nonstationary with true $d = 1.35$, T2CD-step estimates $\hat{d} = 1.275$ with CI $(1.167, 1.358)$ and $\hat{\tau} = 31.4$ with CI $(24.5, 48.1)$.

Having shown that using T2CD-sigmoid and generalizing the discrete transition function assumed in T2CD-step with a smooth transition function can actually result in improved estimation of the true change point, we now evaluate the performance of T2CD-step and T2CD-sigmoid in estimating d and τ across several different scenarios, using both univariate and multivariate time series data. Throughout, we assess performance in terms of absolute bias of estimates of d and τ. We focus on relative performance, as compared to alternative methods, because the magnitude of the absolute bias of estimates of d and τ can depend on the frequency and magnitude of observed data.

We examine the relative performance of T2CD-step and T2CD-sigmoid with respect to estimating the long memory parameter d in Figure 4. For context, we also consider estimation

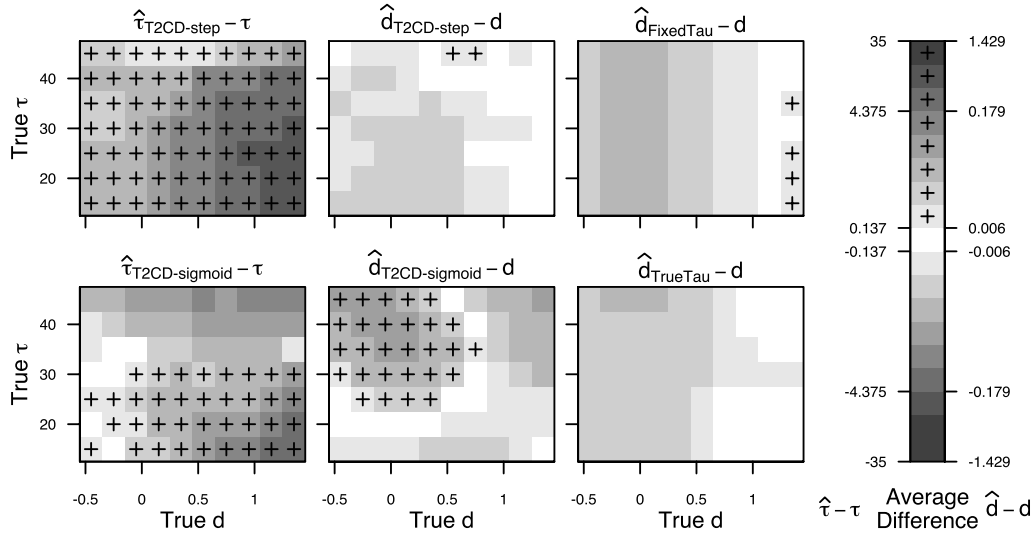


FIG. 4. *Performance of estimates for change location* τ *and long-memory parameter* d *obtained using T2CD-step, T2CD-sigmoid, FixedTau, and TrueTau across* 100 *simulated series per each combination of* τ *and* d *for each simulation configuration. The first regime is generated via Gaussian process with squared exponential kernel; the second regime is generated via FI*(d).

of d using a procedure that fixes the change point $\tau = 50$ (FixedTau) and a procedure that fixes the change point τ at its true value (TrueTau). FixedTau sets the bar for estimating d with conservative data usage common in literature (Lovelady et al. (2007), Tarantola et al. (2010)), whereas TrueTau gives the best d estimate that can be attained if the true change point were known.

We see that estimation of the change point τ and estimation of the long-memory parameter d are closely related. When the estimated change point occurs too early, we tend to overestimate the long-memory parameter. When the estimated change point occurs too late, we tend to underestimate the long-memory parameter d. This pattern is most apparent when T2CD-step is used. Both T2CD-step and T2CD-sigmoid provide better estimates of d than FixedTau, as long as the true change point occurs before 40 hours. We also observe that T2CD-step provides only slightly poorer estimation of d than TrueTau. We further investigate the relative performance of T2CD-step and T2CD-sigmoid in Figure 5.

We see that T2CD-step and T2CD-sigmoid provide comparably accurate estimates of the differencing parameter d. Recalling that T2CD-sigmoid provides vastly superior estimation of the change point τ, compared to T2CD-step when equilibrium process is nonstationary with true long memory parameter $d > 0.5$, these results lead us to prefer T2CD-sigmoid to T2CD-step in most settings.

Now, we compare the performance of T2CD-step and T2CD-sigmoid for estimating the change point τ and long-memory parameter d to the performance of two alternative procedures. Because we are not aware of alternative change-point detection methods in use in the ECIS literature (Lovelady et al. (2007), Tarantola et al. (2010)), we compare to the performance of two alternative procedures based on the popular E-Divisive algorithm introduced in Matteson and James (2014) and James and Matteson (2015). The E-Divisive algorithm is a general-use nonparametric procedure which uses the energy statistics as a distance metric for binary segmentation to find multiple change points. Because it allows for multiple change points and, as a nonparametric procedure, does not incorporate the hypothesized long-memory behavior of equilibrium measurements, the E-Divisive algorithm does not provide an ideal approach to the ECIS change-point detection problem. Nonetheless, it

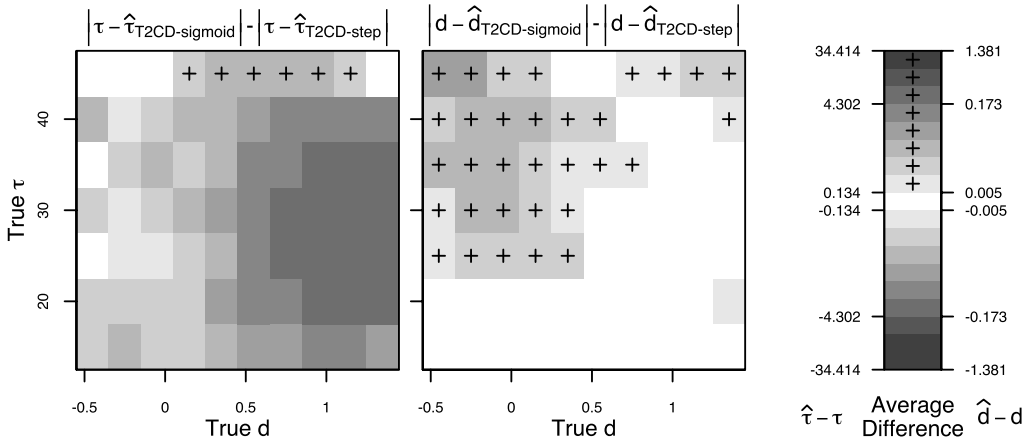


FIG. 5. *Relative performance of estimates for change location τ and long-memory parameter d obtained using T2CD-step and T2CD-sigmoid across* 100 *simulated series per each combination of τ and d for each simulation configuration. The first regime is generated via Gaussian process with squared exponential kernel; the second regime is generated via FI(d). All plots describing performance of estimates share the same x- and y-axes.*

can provide a baseline to evaluate T2CD-step and T2CD-sigmoid against. In our comparison we use E-Divisive to find a maximum of *three* change points and use the most significant change point within the candidate range $[\tau_a, \tau_b]$. We consider two different procedures based on E-Divisive: ECP applies the E-Divisive algorithm to the observed time series $\boldsymbol{y}$, whereas ECP.diff applies the E-Divisive algorithm to the first difference of the observed time series data. Once an estimated change point τ has been obtained, both ECP and ECP.diff procedures estimate the parameters of the FI model for the equlibrium period using maximum likelihood. The relative performance of T2CD-step and T2CD-sigmoid compared to ECP and ECP.diff is shown in Figure 6.

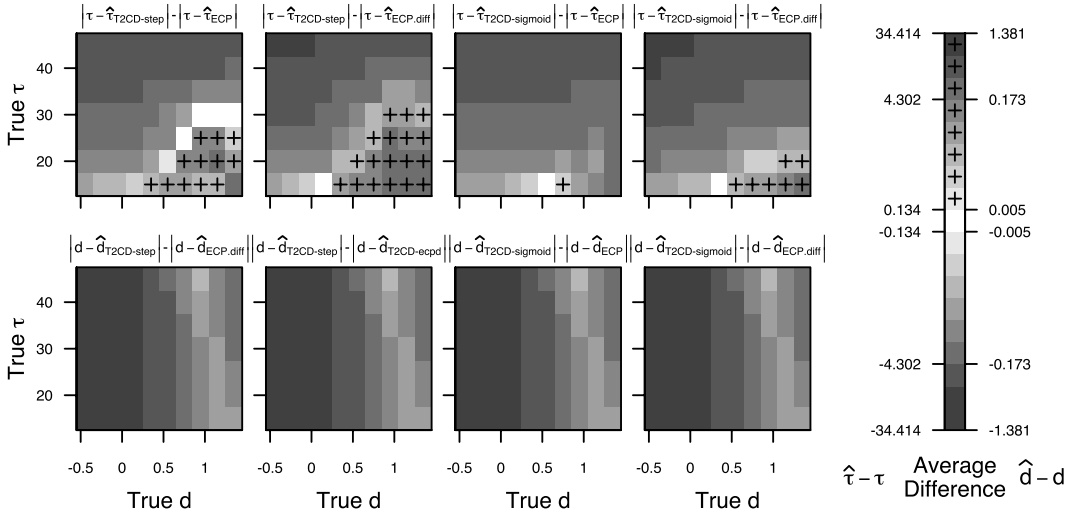


FIG. 6. *Relative performance of estimates for change location τ and long-memory parameter d obtained using T2CD-step, T2CD-sigmoid, ECP, and ECP.diff across* 100 *simulated series per each combination of τ and d for each simulation configuration. The first regime is generated via Gaussian process with squared exponential kernel; the second regime is generated via FI(d). All plots describing performance of estimates share the same x- and y-axes.*

When compared to alternative methods ECP and ECP.diff, both T2CD-step and T2CD-sigmoid estimate the change point τ better for all true change points when the equilibrium process is stationary with $d < 0.5$, and for late true change points $\tau > 35$ when the equilibrium process is nonstationary with $d \geq 0.5$. Careful examination of the change-point estimates indicates that ECP and ECP.diff tend to underestimate the change point, which is likely, due to the fact that both assume that observations between change points are independently and identically distributed. This does not hold for data that we simulated, nor do we expect it to hold for the ECIS data described in Section 1.

Relative performance of the long-memory parameter d mirrors the relative performance of the change point τ. T2CD-step and T2CD-sigmoid tend to perform comparably, with better estimates of the long memory parameter d from T2CD-step for most true values of the differencing parameter, especially when the change point τ occurs later with $\tau > 20$ and then when the equilibrium process is stationary with $d < 0.5$. ECP and ECP.diff produce much poorer estimates of the long-memory parameter d than both T2CD-step and T2CD-sigmoid for all true change point and long-memory parameter values, which is unsurprising, given we observed poorer estimates of the change point τ from ECP and ECP.diff.

However, the performance advantages of T2CD-step and T2CD-sigmoid do come at a computational price. For the first univariate experiment, where the trend regime is generated via Gaussian processes, on average, on a 2.7 GHz CPU, ECP and ECP.diff both take 1.20 seconds, T2CD-step takes 196 seconds, and T2CD-sigmoid takes 19.2 seconds. While both of the T2CD methods are slower than the alternatives, T2CD-sigmoid is roughly 10 times faster than T2CD-step on average. This makes T2CD-sigmoid a competitive option in providing balance between the quality of estimation and computational speed.

Next, we consider multivariate time-series data made up of p individual time series with unique change points $\tau_1, \ldots, \tau_p$ and common long-memory parameter d. We simulate 100 multivariate time series of length $T = 420$ with $p = 3$ for each value of the long-memory parameter $d \in \{-0.25, -0.05, \ldots, 1.45\}$. For each value of d, a single simulated multivariate time series is comprised of three individual time series with different change points $\tau_1 = 15$, $\tau_2 = 25$, and $\tau_3 = 45$. As in the univariate simulations, trend curves $\boldsymbol{f}_j = (f_{j1}, \ldots, f_{j\tau_j})$ are simulated from a mean zero Gaussian process with squared exponential kernel $\mathrm{Cov}[f_t, f_s] = 10\exp(-0.5(s-t)^2)$. We simulate trend regime measurements $y_{jt} = f_{jt} + \eta_{jt}$, where η_{jt} are mean zero heteroscedastic measurement errors with standard deviation $\sigma_{jt} = \frac{2-0.1}{\max\{f_{js}\}_{s=1}^{\tau_j} - \min\{f_{js}\}_{s=1}^{\tau_j}}[f_{jt} - \min\{f_{js}\}_{s=1}^{\tau_j}] + 0.1$. We simulate equilibrium measurements $y_{\tau_j+1}, \ldots, y_T$ according to the FI model $(1-B)^d y_{jt} = \epsilon_{jt}$, where $\epsilon_{jt} \sim \mathrm{N}(0, 0.25)$. Again, we set the candidate range of τ to $[\tau_a = 10, \tau_b = 50]$, use degree three B-spline basis functions with knots at every integer value of t, when fitting β, and knots at every integer multiple of 5 when fitting θ. For T2CD-sigmoid we fix $C = 1000$ throughout. Estimates of the change point τ and long-memory parameter d are summarized in Figure 7.

The multivariate results shown in Figure 7 mirror the univariate results shown in Figure 4. Both T2CD-step and T2CD-sigmoid tend to overestimate earlier change points and underestimate the latest change point. Also, both T2CD-step and T2CD-sigmoid slightly overestimate the differencing parameter d. The performance of both T2CD-step and T2CD-sigmoid is on par with the conservative and oracle methods FixedTau and TrueTau. T2CD-step and T2CD-sigmoid provide better estimates of the long-memory parameter d than FixedTau, as long as the true long-memory parameter is not close to $d = 0.5$ and only slightly worse estimates of the long-memory parameter d than TrueTau.

Figure 8 zooms in on the relative performance of T2CD-step and T2CD-sigmoid. T2CD-sigmoid tends to provide better estimation of the change points τ_1, τ_2, and τ_3, and the two methods again provide comparable estimates of the differencing parameter d. As in the univariate case, these results lead us to prefer T2CD-sigmoid to T2CD-step in most settings.

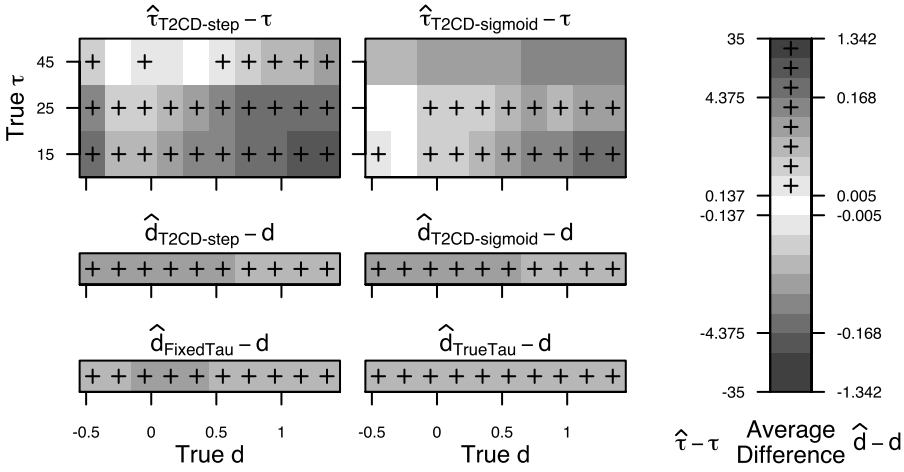


FIG. 7. *Performance of estimates for change location* τ *and long-memory parameter* d *obtained using T2CD-step, T2CD-sigmoid, FixedTau, and TrueTau across* 100 *simulated series per each combination of* τ *and* d. *Multivariate simulations with* $p = 3$, *with change points at* 15, 25, *and* 45. *For each series the first regime is generated via Gaussian process with squared exponential kernel, and the second regime is generated via FI*(d). *All plots describing performance of estimates share the same x- and y-axes.*

Figure 9 examines the relative performance of T2CD-step and T2CD-sigmoid compared to ECP and ECP.diff. T2CD-step and T2CD-sigmoid provide better estimates of the change points τ_1, τ_2, and τ_3, compared to ECP and ECP.diff, as long as the equilibrium process is stationary or the change point occurs late. Regarding estimation of the long-memory parameter d, we observe consistently better performance of T2CD-step and T2CD-sigmoid estimates relative to ECP and ECP.diff estimates.

5. Application to ECIS data. Now, we apply T2CD-step and T2CD-sigmoid to the MDCK and BSC cell data described in Section 1. ECIS resistance measurements were obtained at several frequencies; however, we focus on resistance measured at the frequency of 500 hertz. We also exclude wells that are mechanically disrupted to create a "wound-healing" assay and a single well containing MDCK cells that displayed evidence of instrument failure. In order to assess whether or not cell culture preparation affects our ability to identify cells contaminated with mycoplasma, we analyze data from BSA and gel wells separately.

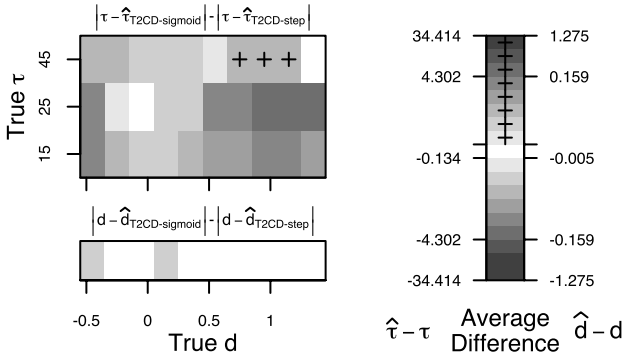


FIG. 8. *Relative performance of estimates for change location* τ *and long-memory parameter* d *obtained using T2CD-step and T2CD-sigmoid across* 100 *simulated series per each combination of* τ *and* d. *Multivariate simulations with* $p = 3$, *with change points at* 15, 25, *and* 45. *For each series the first regime is generated via Gaussian process with squared exponential kernel, and the second regime is generated via FI*(d). *All plots describing performance of estimates share the same x- and y-axes.*

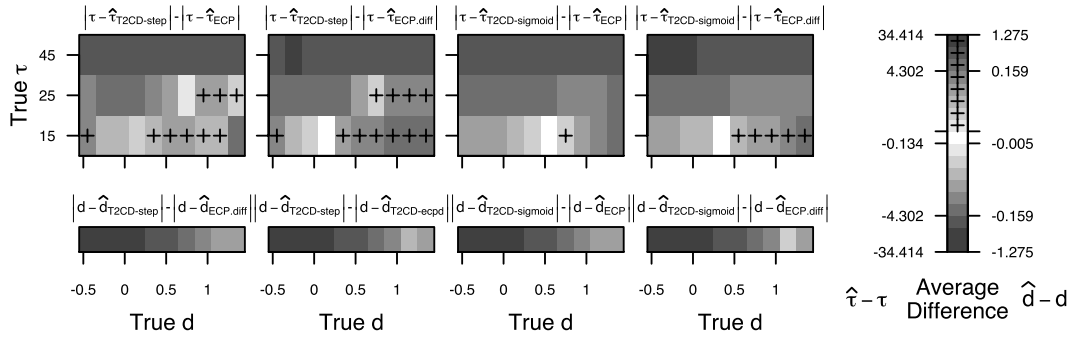


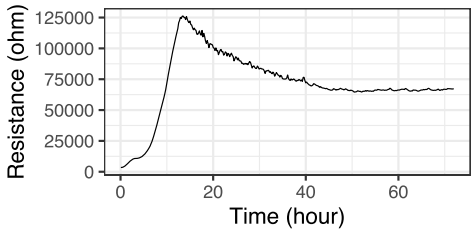
FIG. 9. *Relative performance of estimates for change location τ and long-memory parameter d obtained using T2CD-step, T2CD-sigmoid, ECP, and ECP.diff across* 100 *simulated series per each combination of τ and d. Multivariate simulations with $p = 3$, with change points at* 15, 25, *and* 45. *For each series the first regime is generated via Gaussian process with squared exponential kernel, and the second regime is generated via FI(d). All plots describing performance of estimates share the same x- and y-axes.*

For model fitting we use degree three B-spline basis functions with knots at every integer value of t, when fitting β, and knots at every integer multiple of 5 when fitting θ. As in Section 4, we set $C = 1000$ when implementing T2CD-sigmoid. Based on relevant domain knowledge and visual inspection of the MDCK and BSC data, we set the candidate range of τ to $[\tau_a = 10, \tau_b = 50]$ for MDCK cells and $[\tau_a = 5, \tau_b = 45]$ for BSC cells.

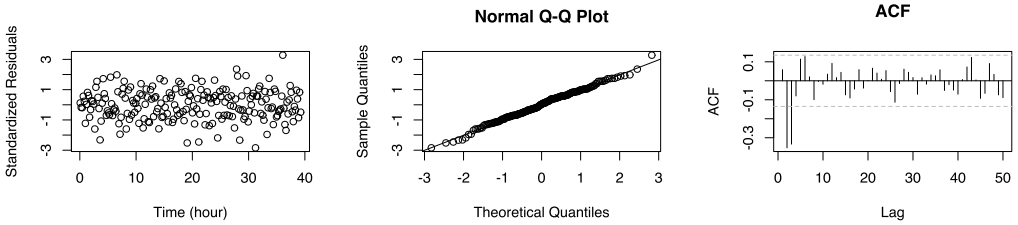
To check the choice of hyperparameters, we plot the time series for a MDCK sample in Figure 10 as well as standardized residuals from model fitting with T2CD-step. The residuals from the first regime are scaled by their estimated variances. Figure 10(b) shows that our choice of model parameters reasonably fit to the trend phase. The residuals in Figure 10(c) demonstrate heavier-than-normal tails.

To systematically assess whether or not equilibrium measurements display heavier-than-normal tails, we perform Shapiro–Wilk type tests of the null hypothesis that the equilibrium errors are normal based on Shapiro and Wilk (1965): we fit the equilibrium model assuming normal errors to the measurements after 50 hours for each well, compute the Shapiro–Wilk test statistic for the residuals, and compare the computed test statistic to the 0.05 quantile of a simulated null distribution based on 500 equilibrium time series simulated according to the fitted normal model. For the MDCK and BSC data, we reject the null hypothesis of normally distributed equilibrium measurements in 60.4% and 5.6% of wells. Accordingly, we assume t-distributed equilibrium errors for MDCK data and normally distributed equilibrium errors for BSC data. To confirm that assuming t-distributed equilibrium errors for the MDCK data, we perform additional Shapiro–Wilk type tests of the null hypothesis that the equilibrium errors are distributed according to a t-distribution: we fit the equilibrium model, assuming t-distributed errors to the measurements after 50 hours for each well, compute the Shapiro-Wilk test statistic for the residuals, and compare the computed test statistic to the 0.05 quantile of a simulated null distribution based on 500 equilibrium time series simulated according to the fitted t-distribution based model. We reject the null hypothesis of t-distributed equilibrium measurements for 1.9% of wells.

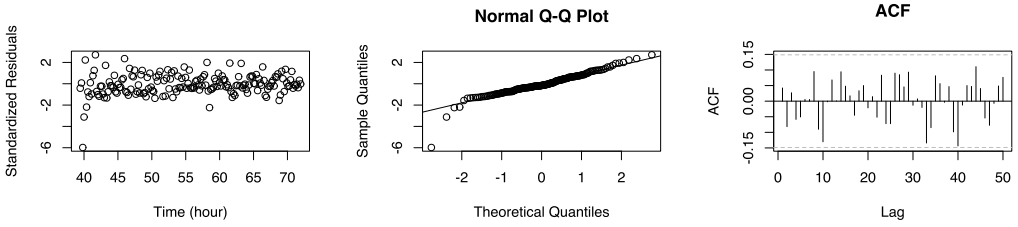
5.1. *MDCK cell line.* From Figure 1 we see that the resistance measurements for MDCK cells tend to peak before slightly decreasing and stablizing. The start of confluence or equilibrium is hypothesized to be at or slightly after the peak and, as a result, is visually distinct. Figure 11(a) plots estimates of the change points τ and long-memory parameters d in combination with approximate 95% confidence intervals for d obtained by applying T2CD-sigmoid to each well as a univariate time series and by applying T2CD-sigmoid to all replicate wells



(a) Time series of resistance measurements recorded at 500 hertz.



(b) Standardized residuals for the trend phase.

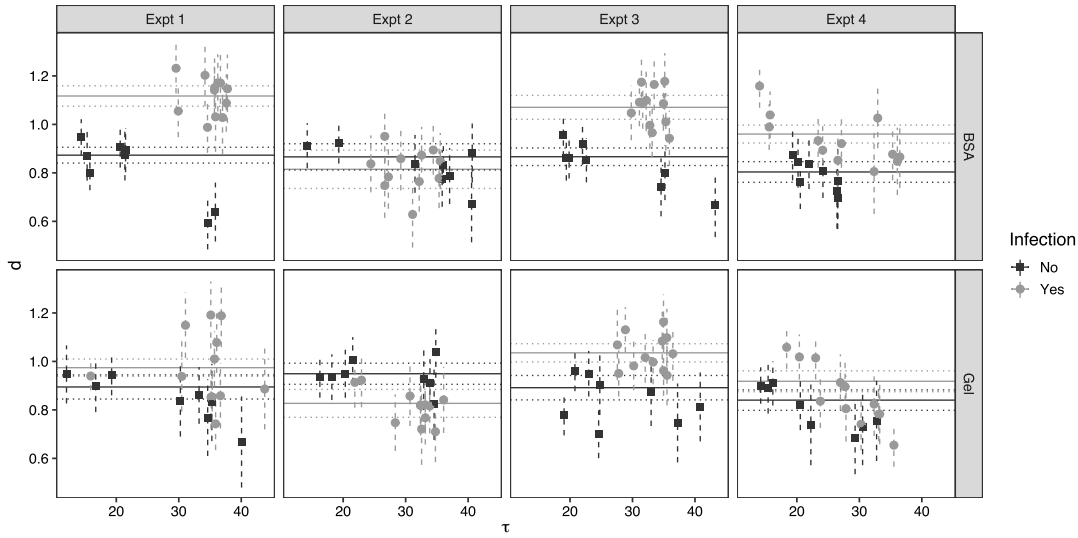


(c) Standardized residuals for the confluence phase.

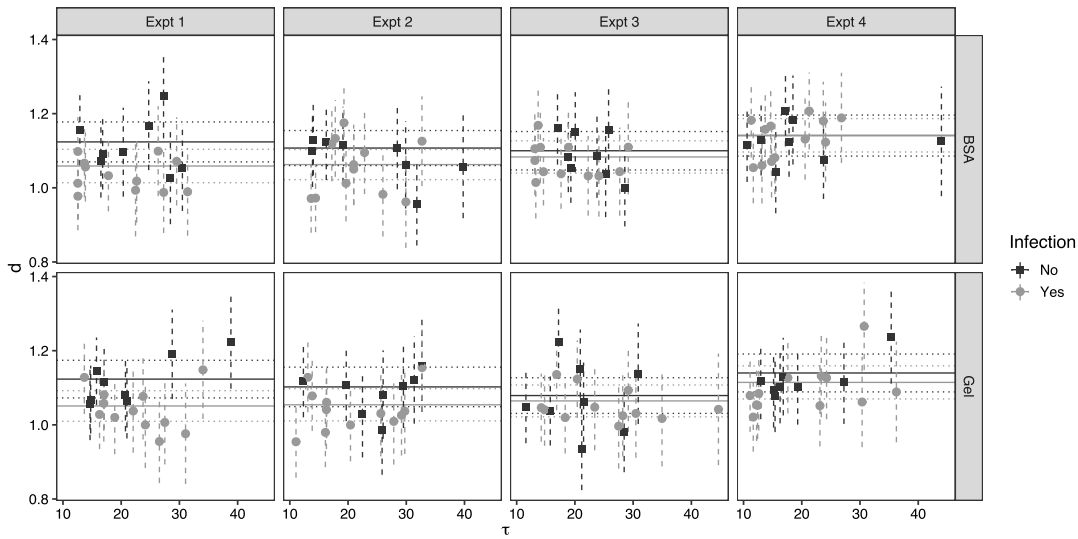
FIG. 10. *MDCK cell, infected and cultivated in gel.*

within the same experiment as a multivariate time series. Approximate 95% confidence intervals for the differencing parameter are based on assuming an approximately normal sampling distribution and obtaining an estimate of the standard deviation from 100 parametric boostrap replicates. An analogous figure for estimates, obtained using T2CD-step, is provided in Appendix C. The estimated change points are scattered within the candidate range of [10, 50], signifying varied initial conditions even in the same batch. We observe clear evidence of long-range dependence at confluence, with all estimates of the long-memory parameter above 0.5. Experiments 1, 3, and 4 suggest that MDCK cells that are contaminated by mycoplasma tend to show longer memory than MDCK cells that are uncontaminated. For wells prepared with BSA, approximate 95% confidence intervals for multivariate estimates of d for contaminated cells exceed approximate 95% confidence intervals for multivariate estimates of d for uncontaminated cells in Experiments 1, 3, and 4. Experiment 2 provides weak evidence of the opposite, with overlapping approximate 95% intervals for multivariate estimates for contaminated cells and uncontaminated cells. This may reflect the presence of batch effects. Web Appendix C contains more a detailed review of estimates of the change point τ and the long-memory parameter d, estimated by T2CD-step and T2CD-sigmoid across experiments, serum types, and infection status.

5.2. *BSC cells*. From Figure 1 we see that the resistance measurements for the BSC cell line tend to increase sharply before plateauing. As compared to the MDCK cell line, the end



(a) MDCK



(b) BSC

FIG. 11. *T2CD-sigmoid estimates for τ and d based on assuming t-distributed errors during equilibrium for MDCK data and normally distributed errors during equilibrium for BSC data. Points are estimates from the univariate version of the method, and horizontal lines mark estimates from the multivariate version. Approximate* 95% *intervals for d based on an approximate normal sampling distribution and standard deviation approximated by* 100 *parametric bootstrap simulations are provided by dashed vertical lines for univariate estimates and horizontal lines for multivariate estimates.*

of the BSC trend phase is less visually distinct. This makes change-point detection and subsequent estimation of the long-memory parameter more difficult. Figure 11(b) plots estimates of the change points τ and long-memory parameters d in combination with approximate 95% confidence intervals for d obtained by applying T2CD-sigmoid to each well as a univariate time series and by applying T2CD-sigmoid to all replicate wells within the same experiment as a multivariate time series. Again, approximate 95% confidence intervals for the differencing parameter are based on assuming an approximately normal sampling distribution and obtaining an estimate of the standard deviation from 100 parametric boostrap replicates. An

analogous figure for estimates obtained using T2CD-step is provided in Appendix C. We observe evidence of long memory, regardless of contamination status, with most univariate and multivariate estimates of the long memory parameter d exceeding one. We do not observe distinct separation between the contaminated and uncontaminated cells. We observe some weak evidence that contaminated BSC cells tend to have slightly shorter memory, corresponding to lower estimates of d, than uncontaminated cells in almost all experiments. However, approximate 95% intervals for multivariate estimates of d for contaminated cells and uncontaminated cells overlap substantially in all settings; see Web Appendix C for a more detailed review of estimates of the change point τ and the long-memory parameter d estimated by T2CD-step and T2CD-sigmoid across experiments, serum types, and infection status.

5.3. *Mycoplasma contamination classification*. To demonstrate the quality and utility of our change point τ and long-memory parameter d estimates, we incorporate the estimates as features in a downstream task of classifying cells by their mycoplasma contamination status. We build on the linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) classifiers built to to classify cell lines using ECIS measurements in Gelsinger, Tupper and Matteson (2020).

Let c indicate the possible classes of observations, which, in this application, corresponds to whether or not a well contains cells contaminated by mycoplasma. Letting $\boldsymbol{z}$ be a vector of features and $\bar{\boldsymbol{z}}_c$ be the average feature vector across all observations in class c, LDA, and QDA class discriminant scores can both be written as special cases of

$$\begin{aligned} \delta_c(\boldsymbol{z}) &= (\boldsymbol{z} - \overline{\boldsymbol{z}}_c)^T \widehat{\Sigma}_c^{-1}(\rho)(\boldsymbol{z} - \overline{\boldsymbol{z}}_c) + \log\left|\widehat{\Sigma}_c(\rho)\right|, \\ \widehat{\Sigma}_c(\rho) &= (1-\rho)\widehat{\Sigma}_c + \rho\widehat{\Sigma}. \end{aligned} \tag{12}$$

LDA is obtained by setting $\rho = 1$, and QDA is obtained by setting $\rho = 0$.

For each cell line we train four LDA and QDA classifiers, training each classifier on data from three experiments and computing classification accuracy on data from the remaining experiment. The average classification accuracy across all four classifiers is provided in Table 1, along with the correspond standard deviations. We compare classifiers trained using the original features described in Gelsinger, Tupper and Matteson (2020) to classifiers trained using the best feature from among the original features described in that paper as well as estimates of the change point τ and long memory parameter d, obtained by applying either T2CD-step or T2CD-sigmoid to data from each well as a univariate time series. A more detailed description of how we constructed the original features for our ECIS measurements is given in the Web Appendix D.

From Table 1 it is evident that the τ and d estimates from T2CD-step and T2CD-sigmoid are useful features that increase classification accuracy for both cell lines. For MDCK cells, LDA using the original features has a mean classification accuracy of 0.743. Including T2CD-step or T2CD-sigmoid features improved the mean classification accuracy by 11.8% and 18.1%, respectively. For BSC cells, LDA using the original features has a mean classification accuracy of 0.563. Including T2CD-step and T2CD-sigmoid features improved the mean classification accuracy by 16.5% and 23.3%, respectively.

6. Conclusion. In this paper we propose a model called T2CD for estimating a change point between a smooth, nonlinear trend period and a long-memory equilibrium period and for quantifying features of the trend and equilibrium periods. We provide exact and generalized estimation strategies, T2CD-step and T2CD-sigmoid. Via simulations, we show that T2CD outperforms a two-step comparison method, based on the popular E-Divisive algorithm for change point detection, when the equilibrium period can be characterized by a

TABLE 1

Classification accuracy for infection status. Average is taken by taking each of the four experiments as the test set and the other three as training set. Parameters τ and d estimated by T2CD, based on assuming t-distributed errors during equilibrium for MDCK data and normally distributed errors during equilibrium for BSC data, increased classification accuracy for both MDCK and BSC cell line

		LDA		QDA	
Cell line	Features	Mean	SD	Mean	SD
MDCK	Original	0.743	0.272	0.580	0.237
	T2CD-step	0.861	0.096	0.849	0.021
	T2CD-sigmoid	0.924	0.079	0.925	0.054
BSC	Original	0.563	0.060	0.588	0.072
	T2CD-step	0.656	0.085	0.575	0.068
	T2CD-sigmoid	0.681	0.072	0.650	0.061

long memory time-series model. Compared to E-Divisive, T2CD tends to produce better estimates of the change points and long-memory parameters. Between T2CD-step and T2CD-sigmoid, we show that T2CD-sigmoid offers computational efficiency gains over T2CD-step and shows more robust performance for estimating the change point. The results also suggest several directions for future research. For instance, the simulation results for T2CD-step indicate that it would be valuable to define an adjustment factor for estimation of the change point. Additionally, it would be valuable to define a data-driven approach to refining the candidate range for the change point in situations where domain knowledge is unavailable or where the candidate range implied by domain knowledge is so large that computation for T2CD-step becomes infeasible.

Practical usage on the MDCK and BSC cell lines shows that T2CD recovers meaningful estimates of change points and long-memory parameters during the confluence phase. Importantly, using T2CD reduces the amount of human supervision needed to manually identify change points, ensures that all change points are identified using the same logic, and makes full use of the available data. Furthermore, we show that estimates of the change points and long-memory parameters improve classification performance downstream.

Funding. The authors gratefully acknowledge financial support from the Cornell University Institute of Biotechnology, the New York State Foundation of Science, Technology and Innovation (NYSTAR), a Xerox PARC Faculty Research Award, National Science Foundation Awards 1455172, 1934985, 1940124, 1940276, and 2114143, USAID, and Cornell Atkinson Center for Sustainability.

SUPPLEMENTARY MATERIAL

Web appendix (DOI: 10.1214/22-AOAS1655SUPP; .pdf). Additional details on implementation of the proposed T2CD method, simulation and experiment results (Zhang, Griffin and Matteson (2023))

REFERENCES

BAILLIE, R. T. (1996). Long memory processes and fractional integration in econometrics. *J. Econometrics* **73** 5–59. https://doi.org/10.1016/0304-4076(95)01732-1

CHEN, C. and LIU, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *J. Amer. Statist. Assoc.* **88** 284–297.

DOORNIK, J. A. and OOMS, M. (2004). Inference and forecasting for ARFIMA models with an application to US and UK inflation. *Stud. Nonlinear Dyn. Econom.* **8**.

DUFRENOT, G., GUEGAN, D. and PEGUIN-FEISSOLLE, A. (2008). Changing-regime volatility: A fractionally integrated SETAR model. *Appl. Financ. Econ.* **18** 519–526.

GELSINGER, M., TUPPER, L. and MATTESON, D. (2020). Cell line classification using Electric Cell-substrate Impedance Sensing (ECIS). *Int. J. Biostat.* **16**.

GUSTAVSSON, R., MANDENIUS, C. F., L'OFGREN, S., SCHEPER, T. and LINDLER, P. (2019). In situ microscopy as online tool for detecting microbial contaminations in cell culture. *J. Biotechnol.* **296** 53–60.

HASLETT, J. and RAFTERY, A. E. (1989). Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **38** 1–50.

HONG, J., KANDASAMY, K., MARIMUTHU, M., CHOI, C. S. and KIM, S. (2011). Electrical cell-substrate impedance sensing as a non-invasive tool for cancer cell study. *Analyst* **136** 237–245.

JAMES, N. A. and MATTESON, D. S. (2015). ecp: An R package for nonparametric multiple change point analysis of multivariate data. *J. Stat. Softw.* **62** 1–25.

KEESE, C. (2019). ECIS Application Webinar Series. http://www.biophysics.com/webinar.php. Accessed: 2019-04-13.

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598.

KUAN, C.-M. (2004). Generalized least squares theory.

LOVELADY, D. C., RICHMOND, T. C., MAGGI, A. N., LO, C.-M. and RABSON, D. A. (2007). Distinguishing cancerous from noncancerous cells through analysis of electrical noise. *Phys. Rev. E, Stat. Nonlin. Soft Matter Phys.* **76** 041908. https://doi.org/10.1103/PhysRevE.76.041908

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345.

NIKA, V., BABYN, P. and ZHU, H. (2014). Change detection of medical images using dictionary learning techniques and PCA. In *Medical Imaging*.

OLSHEN, A. B., VENKATRAMAN, E. S., LUCITO, R. and WIGLER, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5** 557–572. https://doi.org/10.1093/biostatistics/kxh008

OPP, D., WAFULA, B., LIM, J., HUANG, E., LO, J.-C. and LO, C.-M. (2009). Use of electric cell–substrate impedance sensing to assess in vitro cytotoxicity. *Biosens. Bioelectron.* **24** 2625–2629.

R Foundation for Statistical Computing (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, Cambridge. https://doi.org/10.1017/CBO9780511755453

SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52** 591–611.

SOWELL, F. (1992). Maximum likelihood estimation of stationary univariate fractionally integrated time series models. *J. Econometrics* **53** 165–188.

TARANTOLA, M., MAREL, A.-K., SUNNICK, E., ADAM, H., WEGENER, J. and JANSHOFF, A. (2010). Dynamics of human cancer cell lines monitored by electrical and acoustic fluctuation analysis. *Integr. Biol.* **2** 139–50. https://doi.org/10.1039/b920815a

THOMAS, O. and CORANDER, J. (2019). Diagnosing model misspecification and performing generalized Bayes' updates via probabilistic classifiers. *Methodol.* ArXiv.

ZHANG, W., GILBERT, D. and MATTESON, D. S. (2019). ABACUS: Unsupervised multivariate change detection via Bayesian source separation. In *Proceedings of the* 2019 *SIAM International Conference on Data Mining* (*SDM*) 603–611. SIAM, Philadelphia.

ZHANG, W., GRIFFIN, M. and MATTESON, D. S. (2023). Supplement to "Modeling a nonlinear biophysical trend followed by long-memory equilibrium with unknown change point." https://doi.org/10.1214/22-AOAS1655SUPP