Nonparametric modeling of dynamical seasonality and trend with heteroscedastic and dependent errors

Yu-Chun Chen[1], Ming-Yen Cheng[2,4], and Hau-Tieng Wu[3]
[1]National Yang-Ming University, Taipei, Taiwan
[2]National Taiwan University, Taipei, Taiwan
[3]University of California, Berkeley, CA, USA.
[4]Corresponding author: Ming-Yen Cheng, email: cheng@math.ntu.edu.tw

Abstract

Seasonality (or periodicity) and trend are features describing an observed sequence, and extracting these features is an important issue in many scientific fields. However, it is not an easy task for existing methods to analyze simultaneously the trend and *dynamics* of the seasonality such as time-varying frequency and amplitude, and the *adaptivity* of the analysis to such dynamics and robustness to heteroscedastic, dependent errors is not guaranteed. These tasks become even more challenging when there exist multiple seasonal components. We propose a nonparametric model to describe the dynamics of multi-component seasonality, and investigate the recently developed Synchrosqueezing transform (SST) in extracting these features in the presence of a trend and heteroscedastic, dependent errors. The identifiability problem of the nonparametric seasonality model is studied, and the adaptivity and robustness properties of the SST are theoretically justified in both discrete- and continuous-time settings. Consequently we have a new technique for de-coupling the trend, seasonality and heteroscedastic, dependent error process in a general nonparametric setup. The incidence time series of varicella in Taiwan is analyzed.

Keywords: ARMA processes, cycles, periodic functions, synchrosqueezing transform, instantaneous frequency, time-frequency analysis

1. Introduction

Seasonality (or periodicity) is a phenomenon commonly observed in a time series. For example, incidence of varicella is known to exert seasonality with high peaks in winters, which lead to high demand for medical resource. Trend is another phenomenon commonly of interest in time series analysis; for instance, to determine if a general application of certain vaccine is effective in the society, we may like to investigate if overall trend of the disease incidence has changed. Seasonality and trend phenomena are not unique to disease incidence processes. Examples in astronomy, climatology and econometrics have been extensively discussed in the literature (Hall et al., 2000; Nott and Dunsmuir, 2002; Rosen et al., 2009; Pollock, 2009; Bickel et al., 2008).

There are abundant modern methods available to accommodate both seasonality and trend in a time series, for example, seasonal autoregressive integrated moving average (Brockwell and Davis, 2002). Although they are useful in many fields, many of the existing models have some limitations when used to analyze historical data. First, it is hard for the methods to capture *the dynamical behavior* of the seasonality such as its diminishment or changes in the period and/or strength. Indeed, the parametric model assumptions on the seasonality are often too restrictive for real world data. Another limitation is that the seasonality analysis depends on the whole time series, rendering the methods sensitive to the length of the time series. Moreover, there may exist multiple seasonal components, which cannot be handled by most of the existing methods.

To tackle the above mentioned difficulties faced by existing methods and to understand more accurately about the dynamics of a system, we introduce a phenomeno-

logical nonparametric model which captures and offers a natural decomposition of the dynamical seasonal components, the changing trend, and the heteroscedastic, dependent errors. To isolate the meaningful seasonal components based on noisy observations coming from the new model, we focus on the Synchrosqueezing transform (SST) algorithm (Daubechies and Maes, 1996; Daubechies et al., 2010), originally designed to analyze dynamical seasonality without contamination of noise or coupling with trend. It provides not only adaptive and robust estimators but also an easy visualization of the dynamical seasonal components. In addition, since the seasonality is modeled nonparametrically and the SST algorithm is local in nature, it is insensitive to the length of the observed time series, in the sense that the estimate of a seasonal component does not change much as time goes. Furthermore, since our model allows multiple seasonal components, it can extract information about both the high- and low-frequency periodic components.

After the oscillatory components are isolated from the time series, we can extract the trend and approximate accurately the heteroscedastic, dependent errors using the residuals obtained by subtracting from the time series the trend and seasonality estimates. Subsequently we can conduct further investigations on the error process, which are relevant in many directions including forecasting. A medical example is provided: incidence time series of varicella extracted from the Taiwan's National Health Insurance Research Database published by the National Health Research Institute of Taiwan.

2. Model and Methodology

Consider the following classes of periodic functions The identifiability theory of functions in these classes is given in Chen et al. (2013).

DEFINITION 2.1. *For fixed* $0 < \epsilon \ll 1$ *and* $\epsilon \ll c_1 < c_2 < \infty$, *the space* $\mathcal{A}_\epsilon^{c_1,c_2}$ *of Intrinsic Mode Functions consists of functions* $f \in C^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ *having the form*

$$f(t) = A(t)\cos(2\pi\phi(t)),$$

where $A : \mathbb{R} \to \mathbb{R}$ *and* $\phi : \mathbb{R} \to \mathbb{R}$ *satisfy the following conditions for all* $t \in \mathbb{R}$:

$$A \in C^1(\mathbb{R}) \cap L^\infty(\mathbb{R}), \ \inf_{t\in\mathbb{R}} A(t) > c_1, \ \sup_{t\in\mathbb{R}} A(t) < c_2,$$

$$\phi \in C^2(\mathbb{R}), \ \inf_{t\in\mathbb{R}} \phi'(t) > c_1, \ \sup_{t\in\mathbb{R}} \phi'(t) < c_2, |A'(t)| \le \epsilon\phi'(t), \quad |\phi''(t)| \le \epsilon\phi'(t).$$

DEFINITION 2.2. *Fix* $0 < d < 1$. *The space* $\mathcal{A}_{\epsilon,d}^{c_1,c_2}$ *of superpositions of IMFs consists of functions* f *having the form*

$$f(t) = \sum_{k=1}^{K} f_k(t)$$

for some finite $K > 0$ *and for each* $k = 1, \ldots, K$, $f_k(t) = A_k(t)\cos(2\pi\phi_k(t)) \in \mathcal{A}_\epsilon^{c_1,c_2}$ *such that* ϕ_k *satisfies*

$$\phi_k'(t) > \phi_{k-1}'(t) \quad \textit{and} \quad \phi_k'(t) - \phi_{k-1}'(t) \ge d[\phi_k'(t) + \phi_{k-1}'(t)]. \tag{1}$$

We call $A_k(t)$ the *amplitude modulation*, $\phi_k(t)$ the *phase function* and $\phi_k'(t)$ the *instantaneous frequency* of the k-th component in $f \in \mathcal{A}_{\epsilon,d}^{c_1,c_2}$. We then model a random process $Y(t)$ as below:

$$Y(t) = f(t) + T(t) + \sigma(t)\Phi(t), \tag{2}$$

where the seasonality $f(t) = \sum_{k=1}^{K} A_k(t)\cos(2\pi\phi_k(t))$ is in $\mathcal{A}_{\epsilon}^{c_1,c_2}$ when $K = 1$ and in $\mathcal{A}_{\epsilon,d}^{c_1,c_2}$ when $K > 1$, the trend $T(t)$ is modeled as a C^1 real-valued function, $\Phi(t)$ is some stationary generalized random process (GRP), and $\sigma(t) > 0$ so that $\sigma \in C^\infty \cap L^\infty$ is a real-valued smooth function used to model the heteroscedasticity of the error term. For example, $\Phi(t)$ can be taken as a continuous-time autoregressive moving average (CARMA) random process of order (p, q), where $p, q \geq 0$. In practice, we can only access the continuous-time process $Y(t)$ given in model (2) on discrete sampling time-points $n\tau$, where $n \in \mathbb{Z}$ and $\tau > 0$ is the sampling interval. So, we consider the following discrete-time model

$$Y_n = f(n\tau) + T(n\tau) + \sigma(n\tau)\,\Phi_n,\ n \in \mathbb{Z}, \tag{3}$$

where f, T and σ are as in model (2), and Φ_n, $n \in \mathbb{Z}$, is a zero-mean stationary time series which can be taken as, for example, an ARMA time series.

Denote by $\mathcal{S}$ the Schwartz space and let $\mathcal{S}'$ be its dual (the tempered distribution space). When $g \in \mathcal{S}'$ and $h \in \mathcal{S}$, $g(h)$ means g acting on h. Here, sometimes we use the notation $g(h) := \int gh\mathrm{d}t$ which is consistent with the case when g is an integrable function. Given a function $h \in \mathcal{S}$, its Fourier transform is defined as $\widehat{h}(\xi) := \int_{-\infty}^{\infty} h(t)e^{-i2\pi\xi t}\mathrm{d}t$. Take $\psi \in \mathcal{S}$. For $k \in \mathbb{N} \cup \{0\}$, define the following abbreviations:

$$\psi_{a,b}^{(k)}(x) := \frac{1}{a^{k+1/2}}\psi^{(k)}\left(\frac{x-b}{a}\right),\ \psi_{a,b}(x) := \psi_{a,b}^{(0)}(x) \ \text{ and } \ \psi_a^{(k)}(x) := \psi_{a,0}^{(k)}(x),$$

where $\psi^{(k)}$ is the k-th derivative of ψ, $a > 0$ and $b \in \mathbb{R}$. The CWT of a given $f(t) \in \mathcal{S}'$ is defined by

$$W_f(a,b) = \int_{-\infty}^{\infty} f(t)\overline{\psi_{a,b}(t)}\mathrm{d}t, \tag{4}$$

where $a > 0$ and $b \in \mathbb{R}$. Here we follow the convention in the wavelet literature that ψ is called the mother wavelet, a means scale and b means time.

The SynchroSqueezing Transform (SST) algorithm, tailored to analyze a clean function $f(t) \in \mathcal{A}_{\epsilon,d}^{c_1,c_2}$, is composed of three steps. First, choose the mother wavelet $\psi \in \mathcal{S}$ so that supp $\widehat{\psi} \subset [1-\Delta, 1+\Delta]$, where $\Delta \ll 1$, and calculate $W_f(a,b)$, the CWT of $f(t)$ as given in (4). Second, calculate the function $\omega_f(a,b)$ defined on $\mathbb{R}^+ \times \mathbb{R}$, which plays the role of the reassignment rule:

$$\omega_f(a,b) := \begin{cases} \frac{-i\partial_b W_f(a,b)}{2\pi W_f(a,b)} & \text{when} \quad |W_f(a,b)| \neq 0; \\ \infty & \text{when} \quad |W_f(a,b)| = 0. \end{cases} \tag{5}$$

By its definition, $\omega_f(a,b)$ contains abundant information about the instantaneous frequency functions in f. Third, the SST of $f(t)$ is defined by re-assigning the TF representation $W_f(a,b)$ according to the reassignment rule $\omega_f(a,b)$:

$$S_f^{\Gamma}(t,\xi) := \lim_{\alpha\to 0} \int\limits_{\{(a,t):\ |W_f(a,t)|\geq\Gamma\}} h_\alpha(|\omega_f(a,t)-\xi|)W_f(a,t)a^{-3/2}\mathrm{d}a \tag{6}$$

where $(t,\xi) \in \mathbb{R}\times\mathbb{R}^+$, $\alpha, \Gamma > 0$, $h_\alpha(t) := \frac{1}{\alpha}h(\frac{t}{\alpha})$, $h \in L^1(\mathbb{R})$, and $h_\alpha \to \delta$ weakly when $\alpha \to 0$ with δ denoting the Dirac delta function. According to the reassignment rule (5), $S_f^{\Gamma}(t,\xi)$ will only have dominant values around $\phi_k'(t)$ which allows us an accurate estimate of $\phi_k'(t)$. To reconstruct the k-th component $f_k(t) = A_k(t)\cos(2\pi\phi_k(t))$ in f, its amplitude modulation $A_k(t)$ and phase $\phi_k(t)$, we resort to the reconstruction formula of CWT and consider the following estimators:

$$\widetilde{f}_k^{\Gamma,\mathbb{C}}(t) := \mathcal{R}_\psi^{-1}\int_{\frac{1-\Delta}{\phi_k'(t)}}^{\frac{1+\Delta}{\phi_k'(t)}} W_f(a,t)\boldsymbol{\chi}_{|W_f(a,t)|>\Gamma}(a)a^{-3/2}\mathrm{d}a,\ \ \widetilde{f}_k^{\Gamma}(t) := \mathfrak{Re}\,\widetilde{f}_k^{\Gamma,\mathbb{C}}(t), \tag{7}$$

where $\mathcal{R}_\psi = \int \frac{\widehat{\psi}(\zeta)}{\zeta}\mathrm{d}\zeta$, $\boldsymbol{\chi}$ is the indicator function, and $\mathfrak{Re}$ means the real part,

$$\widetilde{A}_k(t) := |\widetilde{f}_k^{\Gamma,\mathbb{C}}(t)|$$

and an estimator for $\phi_k(t)$ can then be obtained by unwrapping the phase of the complex-valued signal $\widetilde{f}_k^{\Gamma,\mathbb{C}}(t)/\widetilde{A}_k(t)$.
Now, consider that we have discrete-time observations of $f \in \mathcal{A}_{\epsilon,d}^{c_1,c_2}$, that is, $\boldsymbol{f} := \{f(n\tau)\}_{n\in\mathbb{Z}}$ and $\boldsymbol{f_n} = f(n\tau)$, where $\tau > 0$ is the sampling interval. In this case we model the discrete-time observations as a delta chain, $f_\tau = \tau\sum_{i\in\mathbb{Z}} f(t)\delta_{n\tau}$, where $\delta_{n\tau}$ is the delta measure at $n\tau$ and plug it into (4). Since $\psi \in \mathcal{S}$, $W_{f_\tau}(a,b)$ is well-defined and equals $\tau\sum_{m\in\mathbb{Z}} f(m\tau)\frac{1}{a^{1/2}}\psi\big(\frac{m\tau-b}{a}\big)$. This is simply the discretization of (4), so for $a > 0$ and $n \in \mathbb{Z}$ denote

$$\begin{aligned} W_{\boldsymbol{f}}(a,n\tau) &:= \tau\sum_{m\in\mathbb{Z}} \boldsymbol{f_m}\frac{1}{a^{1/2}}\psi\big(\frac{m\tau-n\tau}{a}\big), \\ \partial_b W_{\boldsymbol{f}}(a,n\tau) &:= \tau\sum_{m\in\mathbb{Z}} \boldsymbol{f_m}\frac{1}{a^{3/2}}\psi'\big(\frac{m\tau-n\tau}{a}\big). \end{aligned} \tag{8}$$

Similarly, we have the discretization of (5) and (6), which are denoted as

$$\begin{aligned} \omega_{\boldsymbol{f}}(a,n\tau) &:= \begin{cases} \frac{-i\partial_b W_{\boldsymbol{f}}(a,n\tau)}{2\pi W_{\boldsymbol{f}}(a,n\tau)} & \text{when} \quad |W_{\boldsymbol{f}}(a,n\tau)| \neq 0; \\ \infty & \text{when} \quad |W_{\boldsymbol{f}}(a,n\tau)| = 0. \end{cases} \\ S_{\boldsymbol{f}}^{\Gamma}(n\tau,\xi) &:= \lim_{\alpha\to 0} \int\limits_{\{(a,n\tau):\ |W_{\boldsymbol{f}}(a,n\tau)|\geq\Gamma\}} h_\alpha(|\omega_{\boldsymbol{f}}(a,n\tau)-\xi|)W_{\boldsymbol{f}}(a,n\tau)a^{-3/2}\mathrm{d}a, \end{aligned} \tag{9}$$

where $\Gamma > 0$ and $n \in \mathbb{Z}$. Then the estimation of f_k, A_k and ϕ_k, $k = 1,\ldots,K$, follows immediately, for example, for $n \in \mathbb{Z}$ we have

$$\widetilde{f}_{k,n}^{\Gamma,\mathbb{C}} := \mathcal{R}_\psi^{-1}\int_{\frac{1-\Delta}{\phi_k'(n\tau)}}^{\frac{1+\Delta}{\phi_k'(n\tau)}} W_{\boldsymbol{f}}(a,n\tau)\boldsymbol{\chi}_{|W_{\boldsymbol{f}}(a,n\tau)|>\Gamma}a^{-3/2}\mathrm{d}a, \quad \widetilde{f}_{k,n}^{\Gamma} := \mathfrak{Re}\widetilde{f}_{k,n}^{\Gamma,\mathbb{C}}. \tag{10}$$

The above discussions concern the cases when the observations are not contaminated with noise and do not contain trend. If we observe Y satisfying model (2), we simply replace f in (5), (6) and (7) by Y, and we consider the following the trend estimator:

$$\widetilde{T} := Y - \mathfrak{Re}\int_{\frac{1-\Delta}{c_2}}^{\frac{1+\Delta}{c_1}} W_Y(a,b)a^{-3/2}\mathrm{d}a,$$

which is a GRP in general. Suppose we have discrete-time observations $\boldsymbol{Y} = \{\boldsymbol{Y_n}\}_{n\in\mathbb{Z}}$ from model (3). Then we replace $\boldsymbol{f}$ in (8), (9) and (10) by $\boldsymbol{Y}$, and then reconstruct the trend by the following:

$$\widetilde{T}_n := \boldsymbol{Y_n} - \mathfrak{Re}\int_{\frac{1-\Delta}{c_2}}^{\frac{1+\Delta}{c_1}} W_{\boldsymbol{Y}}(a,n\tau)a^{-3/2}\mathrm{d}a, n \in \mathbb{Z}\,.$$

We refer to Chen et al. (2013) for the robustness theory and implementation of the proposed methods.

3. Seasonal Dynamics of Varicella

Varicella is caused by the varicella-zoster virus. It occurs primarily in children and adolescents and features a seasonal pattern with the peak incidence happening in

the winter. Beyond the existence of seasonality, the dynamics of the seasonality, in particular, the effect of the public vaccination program on the seasonal dynamics has been less studied. As varicella is a highly contagious disease but can be effectively prevented, by 70%–80%, using varicella vaccines, free varicella vaccination was made available to certain areas in Taiwan starting from 2003. A nationwide vaccination program was then launched in Taiwan in 2004. The weekly cumulative incidence rate of varicella from 1 January 2000 to 31 December 2009 was calculated using out-patient visit records of an one-million representative cohort dataset of Taiwan's National Health Insurance Research Database.

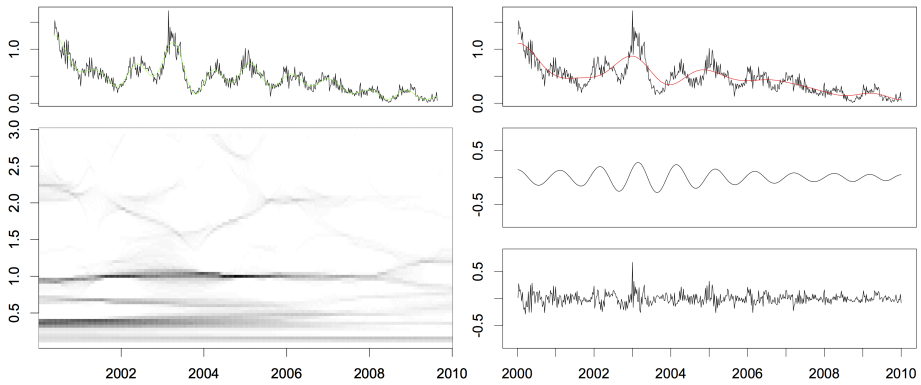

Fig. 1. *Varicella data.* Upper left: The varicella incidence time series Y (black) and $\widetilde{f}+\widetilde{T}$ (green). Lower left: The SST result, where the y-axis is the frequency and the intensity of the graph is the absolute value of the SST given in (6). Right: From top to bottom are $\widetilde{T}$ (red) and Y (black), the estimated seasonality $\widetilde{f}$, and the residuals.

Notice from Figure 1 that the seasonality, the dominant curve on the time-frequency plane, is graphically visible based on the SST analysis. Also, we can tell from $\widetilde{f}$ the dynamics of the seasonality. Before the nationwide public vaccination program was launched in 2004, the seasonal behavior of varicella was stable and evident: it climbed gradually after September or October, reached the peak level in December and the next January, and then declined down to the base during June and July. This finding is compatible with that of previous studies in Hong Kong and Denmark without public vaccination program (Chan et al., 2011; Metcalf et al., 2009). After 2004, accordant with the increase in the vaccination rate, the winter peak shifted slightly toward spring between 2004 and 2008 while the period remained the same. This finding is consistent with the result of vaccination program in the United States (Seward et al., 2002). More importantly, less oscillatory seasonality is observed after the launch of public immunization in 2004, which is important and less reported before. The estimated trend of varicella incidence of is compatible with the finding in Chang et al. (2011). The obvious drop in the trend starting from 2003 may be explained by the free varicella vaccination program in 2003, which was subsequently accelerated by the nationwide public vaccination program commenced in 2004. Both the sharp decline during 2000-2001 and the increase during 2002 in the trend are less conclusive by this analysis; instead they may have been simply artifacts caused by the transition of the coding system. In Taiwan, the whole medical claim system had undergone a transition from a localized coding scheme (A-code) to the international standardized coding scheme (International Classification of Disease, ICD-9), which was not completed until 2002. The gradual decrease in the trend starting from 2005 and the fact that the trend seems to level off starting from 2008 may be interpreted as the expected impact of the vaccination program. Clearly, the SST analysis

showed its robustness to the coding bias problem in 2000-2002, when recovering the trend in 2003-2010.

References

Bickel, P., B. Kleijn, and J. Rice (2008). Event weighted tests for detecting periodicity in photon arrival times. *The Astrophysical Journal 685*, 384–389.

Brockwell, P. J. and R. A. Davis (2002). *Introduction to Time Series and Forecasting*. Springer.

Chan, J., L. Tian, Y. Kwan, W. Chan, and C. Leung (2011). Hospitalizations for varicella in children and adolescents in a referral hospital in Hong Kong, 2004 to 2008: A time series study. *BMC Public health 11*(1), 366.

Chang, L.-Y., L.-M. Huang, I.-S. Chang, and F.-Y. Tsai (2011). Epidemiological characteristics of varicella from 2000 to 2008 and the impact of nationwide immunization in Taiwan. *BMC Infectious disease 11*(1), 352.

Chen, Y.-C., M.-Y. Cheng, and H.-T. Wu (2013). Nonparametric and adaptive modeling of dynamic seasonality and trend with heteroscedastic and dependent errors. *Manuscript*, http://arxiv.org/abs/1210.4672.

Daubechies, I., J. Lu, and H.-T. Wu (2010). Synchrosqueezed wavelet transforms: An empirical mode decomposition-like tool. *Appl. Comput. Harmon. Anal. 30*, 243–261.

Daubechies, I. and S. Maes (1996). A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models. In *Wavelets in Medicine and Biology*, pp. 527–546. CRC-Press.

Hall, P., J. Reimann, and J. Rice (2000). Nonparametric estimation of a periodic function. *Biometrika 87*(3), 545–557.

Metcalf, C., O. Bjornstad, B. Grenfell, and V. Andreasen (2009). Seasonality and comparative dynamics of six childhood infections in pre-vaccination copenhagen. *Proceedings of the royal society B 276*(1), 4111–4118.

Nott, D. J. and W. T. Dunsmuir (2002). Estimation of nonstationary spatial covariance structure. *Biometrika 89*, 819–829.

Pollock, D. (2009). Investigating economic trends and cycles. In *Vol. 2 Applied Econometrics, T.C. Mills and K. Patterson (eds.)*, Palgrave Handbook of Econometrics. Palgrave Macmillan.

Rosen, O., D. S. Stoffer, and S. Wood (2009). Local spectral analysis via a Bayesian mixture of smoothing splines. *J. Am. Stat. Assoc. 104*(485), 249–262.

Seward, J., B. Watson, C. Peterson, L. Mascola, J. Pelosi, J. Zhang, T. Maupin, G. Goldman, L. Tabony, K. Brodovicz, A. Jumaan, and M. Wharton (2002). Varicella disease after introduction of varicella vaccine in the united states, 1995-2000. *JAMA 287*(5), 606–611.