Statistica Sinica **15**(2005), 569-595

ASYMPTOTIC PROPERTIES OF ESTAR MODELS

Thomas Jagger and Xu-Feng Niu

Florida State University

Abstract: Motivated by modeling and forecasting annual hurricane activity in the North Atlantic, we introduce a class of exponential space-time autoregressive (ESTAR) models for count processes by describing the local characteristics as members of finitely supported exponential families. We show that the joint distribution of a space-time count process conditioned on previous observations is a Gibbs field, and demonstrate that an exponential space-time model can be represented as a finite primitive aperiodic Markov chain. The space-time models are identifiable in the parameter space. Asymptotic properties of the log-likelihood function for the parameters in the processes are investigated. Under mild conditions, the maximum likelihood estimates of the parameters are proved to be consistent and asymptotically normally distributed. In order to solve the intractable-constant problem in the likelihood function, the Maximum Pseudo Likelihood Estimation (MPLE) method and the Markov Chain Monte Carlo Maximum Likelihood (MCML) method are proposed to estimate the parameters in an ESTAR model. Simulation results show that both MPLE and MCML estimates appeared relatively unbiased. The MCML method is preferred primarily due to this method providing reasonable standard error estimates of the estimated parameters.

Key words and phrases: Auto-Poisson process, conditionally specified space-time models, Fisher Information regularity conditions, Hessian matrix, reference and objective functions.

1. Introduction

Hurricanes are some of the most devastating natural catastrophes. They rival earthquakes in destructive potential and loss of life. In 1992 Hurricane Andrew caused approximately $30 billion in damage, and in 1998 Hurricane Mitch killed over 10,000 people in Central America. Unfortunately, the potential for widespread destruction from hurricanes is increasing as development continues in the areas where hurricanes tend to strike: the warm subtropical shorelines and islands of the Atlantic Ocean, Gulf of Mexico, and Caribbean Sea.

Forecasting hurricanes in the North Atlantic is very important but quite difficult. Some of the difficulty in hurricane forecasting stems from the variability in hurricane activity. For instance, during the last century in the North Atlantic, there have been years without any hurricanes and years with as many as twelve.

More difficulty in forecasting concerns the spatial and temporal nature of hurricanes: each hurricane forms at a different time during the year and a different location in the ocean.

Accurate forecasts of hurricane activity with long lead times can reduce potential damage and loss of life. Meteorologists have been using covariates to create yearly hurricane activity forecasts for the North Atlantic since 1984 (Gray (1984a, 1984b), Elsner, Lehmiller and Kimberlain (1996)). Currently, these forecasts provide total basin and sub-basin yearly activity estimates using both linear and generalized linear regression models. However, the current hurricane forecast models do not combine the spatial and temporal information from the data that are available since 1887. Thus, they cannot provide specific monthly or yearly activity forecasts for a given region.

New statistical models need to be developed that combine both the spatial and temporal structures of the data. In an effort to improve current forecast models, we propose a class of exponential space-time autoregressive (ESTAR) models for count processes defined on lattice systems. The ESTAR models developed in this paper can be used to analyze and forecast any type of space-time count processes, such as the number of tornadoes and storms in different locations over a given time period, the number of earthquakes at different times and locations, traffic flow and accident patterns in different cities, and crime patterns or disease incidences in different countries. Jagger, Niu, and Elsner (2002) applied this type of model to the analysis and forecasting of annual Atlantic hurricane activities, and found that the space-time model could provide valuable guidance for issuing seasonal hurricane forecasts. This article focuses on investigating statistical properties of the space-time models, including the existence of consistent and efficient estimates of the parameters in the models. The uniqueness of the maximum likelihood estimates of parameters is examined. Under mild conditions, we prove that the estimates are consistent and asymptotically normally distributed.

It was first noted by Brook (1964) that space-time lattice models can be either simultaneously or conditionally specified. In a simultaneously specified model, a system of equations, usually one for each lattice site, is derived. Each equation is a function of the values at the other lattice sites and a random error term. The statistical properties of the models are introduced through this error term.

Simultaneously specified spatial models where introduced by Whittle (1954), and extended by Cliff et al. (1975) to a class of space-time autoregressive moving average (STARMA) models. Niu and Tiao (1995) applied a class of STARMA models for the analysis of satellite ozone data on a fixed latitude, which used the temporal and longitudinal spatial dependence structure of the data with a Gaussian error term. Furthermore, Niu (1995) studied the consistency and efficiency of the maximum likelihood estimates of parameters in the space-time models of Niu and Tiao (1995). More recently, Niu, McKeague and Elsner (2003) proposed a class of seasonal space-time models for general lattice systems, which extended the models considered by Niu and Tiao (1995) and addressed both longitudinal and latitudinal spatial structures of environmental data.

In a conditionally specified model the conditional distribution of the values at each lattice site is a function of the values at the other lattice sites. Conditionally specified models were introduced by Bartlett ((1955, Section 2.2), 1967, 1968). When random variables on a multidimensional lattice take only two values 0 and 1 (or -1, 1), Bartlett (1971, 1972) considered the relationship between conditional "nearest-neighbor" models and space-time autoregressive models. Besag (1974) proved that under mild conditions, joint probability distributions exist for the random variables in a conditional "nearest-neighbor" system, and suggested some conditionally specified models for spatial lattice systems, such as the auto-Poisson process for modeling spatial counts data, and the autologistic and autobinomial processes for modeling spatial binary data. Huffer and Wu (1998) used an autologistic process to model the distribution of plant species, where they employed the Markov Chain Monte Carlo Maximum Likelihood (MCML) method to estimate the parameters in the spatial models.

Hurricane frequencies in different time and spatial locations form a space-time counts process. The dependence structure of this type of data can be modeled by the conditional probability approach (see Bartlett (1968), Whittle (1963), Besag (1972, 1974) and Gilks, Richardson and Spiegelhalter (1996)). The auto-Poisson model proposed by Besag (1974) imposes some restrictions on the parameter space that can only be applied to spatial data sets in which the interaction coefficients are non-positive. Furthermore, few inference results are available for this type of model. The ESTAR models proposed in this article will relax the restrictions on the parameter space and show how the seasonal hurricane frequency at a specific location is related to other atmospheric events and neighborhood observations.

The ESTAR models are defined in Section 2 in terms of conditional distributions. Two special classes of the models, one based on the truncated Poisson distribution and the other based on the binomial distribution, are discussed. We show that the joint distribution of any ESTAR model conditioned on the past is a Gibbs field and derive its potential function. We demonstrate that the space-time models can be represented by a finite state aperiodic primitive Markov Chain and are identifiable. In Section 3, we derive the log-likelihood function for the parameters in the models and discuss its properties. The conditions for the uniqueness of the maximum likelihood estimates are specified. Asymptotic properties of the maximum likelihood estimates (MLE) are investigated in Section 4. Under some mild conditions, we prove that the estimates are consistent and asymptotically normal. The proofs of the main results are presented in the Appendix. In practice, the likelihood function of an ESTAR model cannot be calculated directly due to an intractable constant of proportionality that depends not only on the parameter values but also the past data values of the process. Two methods, the Maximum Pseudo Likelihood Estimation (MPLE) method and the Markov Chain Monte Carlo Maximum Likelihood (MCML) method, are proposed to estimate the parameters in an ESTAR model. Simulation results in Section 5 show the finite sample properties of the estimators. The estimates based on both methods appeared relatively unbiased, while the MCML method is preferred in practice since it provides reasonable standard error estimates of the estimated parameters. Finally, conclusions and discussion are given in Section 6.

2. The ESTAR Models

In this section we define the Exponential Space Time Autoregressive (ESTAR) models using the conditionally specified approach and study the statistical properties of these models. We show that the joint distribution of any ESTAR model conditioned only on the past is a Gibbs field and the models are identifiable.

Let $\{X_{t,s}\}$ be a space-time count process defined on a lattice where $t \in \mathbb{Z}$, $s \in S$ and $S \subset \mathbb{Z}^d$ with $|S| < \infty$. For each $T > 0$, we can represent an ESTAR model as a collection of random vectors $\{X_1, \ldots, X_T\}$ with $X_t = \{X_{t,s}, s \in S\} \in \{0, 1, \ldots, M\}^{|S|}$. For an element $\omega \in \Omega$, $x_t = X_t(\omega) = \{X_{t,s}(\omega), s \in S\}$ is called a configuration. Using this notation, the ESTAR models are defined as follows.

Definition 2.1. The conditional distribution of an ESTAR process $\{X_t\}$ is given by

$$\Pr(X_{t,s} = x_{t,s} | \{X_v : v < t; X_{t,u} : u \neq s\}) = \frac{[\lambda_{t,s}]^{x_{t,s}} h_s(x_{t,s})}{c_s(\lambda_{t,s})}, \tag{2.1}$$

where $c_s(\lambda_{t,s}) = \sum_{k=0}^{M} (\lambda_{t,s})^k h_s(k)$, $\lambda_{t,s} = \exp\{\alpha_s + \sum_{j=0}^{p} \sum_{u \in S} \gamma_j(u,s) x_{t-j,u}\}$ with parameters $\gamma_0(u,s) = \gamma_0(s,u), \gamma_0(s,s) = 0$ and $\boldsymbol{\theta} = \{\alpha_s, \gamma_j(u,s) : u, s \in S;$ $j = 0, \ldots, p\}$.

In (2.1), the components $\lambda_{t,s}$ represent generalized rates. The parameters α_s are similar to constants in a standard time series model, and the coefficient $\gamma_j(u,s)$ is a measure of the coupling from $X_{t-j,u}$ to $X_{t,s}$.

In this study, we assume that the functions $\{h_s(\cdot), s \in S\}$ are known whereas the parameter $\boldsymbol{\theta}$ is unknown. Without loss of generality, we set $h_s(0) = 1$. Thus each collection of functions $h_s(\cdot)$ defines a unique family of ESTAR models. Notice that $c_s(\cdot)$ at any given time t is indirectly a function of the parameter $\boldsymbol{\theta}$, the previous values $\{x_{t-1}, \ldots, x_{t-p}\}$, and the "instantaneous values" $\{x_{t,u} : u \neq s\}$.

The truncated Poisson space-time autoregressive (TPSTAR) model, where $h_s(x) = 1/x!$ and the components $\lambda_{t,s}$ represent conditional Poisson rates, is an example of the ESTAR models. In the temporal direction, it is an order p autoregressive time series. In the spatial direction, conditioned on the observations in the last p time periods, it is an Auto-Truncated Poisson distribution. That is, the distribution at a single site at some point in time, conditioned on all the sites' values for the last p periods and the other site values for the current time, is a right truncated Poisson distribution, truncated at a fixed value M.

Another example of the ESTAR models is the binomial space-time autoregressive (BSTAR) model, where $h_s(y) = n_s!/(y!(n_s - y)!)$ and the components $\lambda_{t,s}$ represent the odds ratio $p/(1-p)$. This model has applications in modeling disease propagation such as AIDs incidence rates, where the sites in the model could represent counties in a state, and the time series at each site could be the number of people in that county that contracted the disease each month.

The following three restrictions exist on the ESTAR process $\{X_{t,s}, t \in \mathbb{Z}; s \in S\}$. Note that the second and third restrictions are implied by Definition 2.1.

1. The support of the conditional distribution for each $s \in S$ and all t is $\mathcal{X}_s = \{0, 1, 2 \ldots n(s)\}$, where $1 \leq n(s) \leq M$ and M is a finite global constant.
2. The joint distribution of $\{X_{t,s}, s \in S\}$ given $\{X_{t',s'} : t' < t, s' \in S\}$ is pairwise dependent, as defined in Cressie ((1993), Section 6.4.2). This restricts interactions in a space-time model to pairs of sites.
3. The conditional distribution of two lattice points separated by time, say $X_{t,s}$ and $X_{t',s}$, must come from the same exponential family, varying only in the natural parameter, for example, the right truncated Poisson distribution with fixed M and possibly different λ. Another example is the binomial distribution with fixed n but possibly different success probability p.

The first restriction on the support of $X_{t,s}$ implies that $h_s(y) = 0$ for all $s \in S$ whenever $y > M$ or $y < 0$. This restriction can be generalized from a finite set of positive integers to a finite set of real values. We have arbitrarily restricted $\mathcal{X}_s$ to be a subset of the positive integers including zero and one, because this set covers all of our examples and simplifies the notation used in our proofs. At each lattice point $s \in S$, $\mathcal{X}_s$ must contain at least two values so that the process is not deterministic. Note also that $\mathcal{X}_s$ is a function only of lattice position and not of time, a necessary condition for stationarity of our ESTAR models.

The second restriction defines the form of the dependence, (see, e.g., Besag (1974)). This restriction is implied by Definition 2.1 and the Factorization Theorem (Besag (1974)).

The third restriction is required so that the models are stationary, which implies that $h_s(y)$ is time invariant. However, the space-time models proposed allow us to have different exponential families at different spatial locations. (See Definition 2.3 for an exponential family). For example, if we are using members from the binomial(n, p) distribution, the restriction does not allow us to vary the total, n, temporally, but does allow us to vary n spatially.

These three restrictions may be relaxed in future models to allow for continuous support, nonstationarity, and complex interactions between sites.

For the purpose of investigating properties of the parameter estimates in the space-time models, we use the most general case, namely that each site's distribution at time t depends on every site's values at times $\{t-p, \ldots, t-1\}$, and on the other sites' values at time t.

For the conditional distribution specified in Definition 2.1, we denote $\Pr(X_{t,s} = x_{t,s} | \{X_v : v < t; X_{t,u} : u \neq s\})$ as $\Pi_{\boldsymbol{\theta}}(x_{t,s} | x_v : v < t;\ x_{t,u} : u \neq s)$. This will simplify notation in keeping with standard image analysis notation. For the most general case of the space-time models we can define the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as a real valued vector with N distinct components $\{\alpha_s, \gamma_j(u, s) : j \in 0, \ldots, p; u, s \in S\}$. In order to ensure that the conditional distributions are consistently specified, following Besag (1974) we assume that $\gamma_0(s, u) = \gamma_0(u, s)$ and $\gamma_0(s, s)$=0.

Restrictions on the parameters of the space-time models may be imposed as needed. For example, we may set $\gamma_0(s, u) = 0$. In this case, there are no instantaneous site dependencies, and the sites are conditionally independent of each other given the previous observations. Compared to the case with instantaneous dependencies, the conditionally independent case is easy to simulate, since the underlying process is a Markov chain of known distributions. This also makes parameter estimation easier, since we can use a modified generalized linear model method. Unfortunately, space-time processes in practice are rarely conditionally independent, as one expects spatial regions to interact with each other over small time periods.

We can also restrict the space-time models by imposing translation invariance on the parameters. That is, if the sites are on a regular n-dimensional lattice, we may require $\gamma_j(s, u) = \gamma_j(s-u, 0)$. Other restrictions may also be considered. For example, it is reasonable in most cases to assume that $\gamma_j(s, u) = \gamma_j(u, s)$ so that we have a well defined neighborhood structure. Other than the need to specify the order of the temporal autoregressive models, the additional neighborhood structure is not needed to prove any of the asymptotic results in this paper.

The purpose of the first investigation of the space-time models is to determine the conditional distribution of X_t given the information from the past. We need some terminology to describe these models.

Definition 2.2.(Winkler (1995, Section 3.2)) A *Gibbs field* is a probability measure on a configuration $\boldsymbol{x} = \{x_s : s \in S, x_s \in X\}$ of the form

$$\Pi(\boldsymbol{x}) = \frac{e^{-U(\boldsymbol{x})}}{\sum_{\boldsymbol{z}\in X^{|S|}} e^{-U(\boldsymbol{z})}},$$

where the state space X is countable and S is finite. Π is called the *Gibbs field* induced by the *energy function* $U(\boldsymbol{z})$ and its denominator is called the *partition function.*

Definition 2.3. An *exponential family* is a collection of Gibbs fields, parameterized by $\boldsymbol{\alpha} \in \mathbb{R}^d$, that have the form

$$\Pi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \frac{e^{\langle \boldsymbol{\alpha}, \boldsymbol{T}(\boldsymbol{x})\rangle + T_0(\boldsymbol{x})}}{c(\boldsymbol{\alpha})},$$

where $\boldsymbol{T}(\boldsymbol{x}) = (T_1(\boldsymbol{x}), \ldots, T_d(\boldsymbol{x}))', \{T_i(\boldsymbol{x}); i = 0, 1, \ldots, d\}$ are measurable functions of $\boldsymbol{x}$, and $\langle \cdot, \cdot \rangle$ represents the usual inner product in R^d.

For a space-time autoregressive process defined by (2.1), $X_t = \{X_{t,s}, s \in S\}$ for a given t is a spatial process. The finite sample space of X_t will be denoted by $\mathcal{X} = \prod_{s\in S} \mathcal{X}_s$. In order to simplify notation, we define $Z_t = \left(X_t', \ldots, X_{t-p+1}'\right)'$ and $z_t = (x_t', \ldots, x_{t-p+1}')'$, i.e., z_t is a sample configuration from Z_t. The following proposition specifies the joint distribution of X_t given $Z_{t-1} = z_{t-1}$.

Proposition 2.1. *For any time t, the distribution of a single configuration, X_t, of the ESTAR process conditioned on the past observations is a Gibbs field. The Gibbs energy function of X_t given z_{t-1} is denoted as $U_{\boldsymbol{\theta}}(z_{t-1}, x_t) + K(z_{t-1})$, where*

$$U_{\boldsymbol{\theta}}(z_{t-1}, x_t) = \sum_{s\in S} V_s(x_{t,s}) - \sum_{j=1}^{p} \sum_{(u,s)\in S^2} \gamma_j(u,s) x_{t-j,u} x_{t,s} - \sum_{\{u,s\in S\}} \gamma_0(u,s) x_{t,u} \cdot x_{t,s},$$

with $-V_s(x) = \alpha_s x + \ln(h_s(x))$, $\gamma_0(u,s) = \gamma_0(s,u)$, and $\gamma_0(s,s) = 0$.

In Proposition 2.1, $(s,u) \in S^2$ denotes the set of all possible pairs in S^2 and $\{u, s \in S\}$ denotes the set of all $|S||S-1|/2$ distinct pairs from S. Thus, $\sum_{\{u,s\in S\}} \gamma_0(u,s) x_{t,u} x_{t,s} = (1/2) \sum_{(u,s)\in S^2} \gamma_0(u,s) x_{t,u} x_{t,s}$, since we assume that $\gamma_0(u,s) = \gamma_0(s,u)$ and $\gamma_0(s,s) = 0$.

Since the conditional distribution of the configuration X_t, given the past observations, depends only on the configurations at $\{t-1, \dots, t-p\}$, that is, z_{t-1}, we can write the conditional probability $\Pr(X_t = x_t | Z_{t-1} = z_{t-1})$ as $P_{\boldsymbol{\theta}}(z_{t-1}, x_t)$ to express the dependence on both z_{t-1} and the vector parameter $\boldsymbol{\theta}$:

$$\begin{aligned} &P_{\boldsymbol{\theta}}(z_{t-1}, x_t) \\ &= \frac{e^{-U_{\boldsymbol{\theta}}(z_{t-1}, x_t) - K(z_{t-1})}}{\sum_{x \in \{0,\dots,M\}^{|S|}} e^{-U_{\boldsymbol{\theta}}(z_{t-1}, x) - K(z_{t-1})}} = \frac{e^{-U_{\boldsymbol{\theta}}(z_{t-1}, x_t)}}{\sum_{x \in \{0,\dots,M\}^{|S|}} e^{-U_{\boldsymbol{\theta}}(z_{t-1}, x)}} \\ &= \frac{e^{\sum_{s \in S}(\alpha_s x_{t,s} + \ln h_s(x_{t,s})) + \sum_{j=1}^{p} \sum_{u,s \in S^2} \gamma_j(u,s) x_{t-j,u} x_{t,s} + \sum_{\{u,s \in S\}} \gamma_0(u,s) x_{t,u} x_{t,s}}}{c(\boldsymbol{\theta}, z_{t-1})}, \qquad (2.2) \end{aligned}$$

where $c(\boldsymbol{\theta}, z) = \sum_{x \in \{0,\dots,M\}^{|S|}} e^{-U_{\boldsymbol{\theta}}(z,x)}$.

Remark 2.1. Since the distribution of X_t given the past is not affected by the choice of $K(z_{t-1})$, we can set it to 0. One can verify the joint distribution by deriving the conditional distribution at each site from $P_{\boldsymbol{\theta}}(z_{t-1}, x_t)$, or by deriving it directly from the conditional distributions using the Factorization Theorem (Cressie (1993, Equation 6.4.3)).

Remark 2.2. Even though the distribution of X_t conditioned on $\{X_{t-1} \dots X_{t-p}\}$ forms an exponential family for each $t \in \{1, \dots T\}$, the joint distribution of $\{X_1, \dots X_T\}$ does not form an exponential family. Thus, the energy function cannot be written as $U(x) = \langle \boldsymbol{\theta}, T(x) \rangle$ where $x = (x_1', \dots, x_T)'$. Also, the space-time model does not exhibit pairwise-only dependence, because the energy function cannot be defined by potential functions on sets containing only one or two points.

The following lemma shows that an ESTAR process $\{X_t : t \in \mathbb{Z}\}$ defined by (2.1) can be described as a Markov chain in the temporal direction. This is important for investigating statistical properties of this process. Consider the chain $\{Z_t, t \in \mathbb{Z}\}$, where $Z_t = (X_t', \dots, X_{t-p+1}')'$ with $p \cdot |S|$ components. This chain is a Markov chain, because the distribution of Z_t given $Z_u : u < t$ depends only on Z_{t-1}, since the model is autoregressive of order p. The members of z_t are denoted by $[z_t]_i = x_{t-i+1}$ for $i \in 1, \dots p$, so that $x_{t-i} = [z_{t-1}]_i = [z_t]_{i+1}$ for $i \in 1, \dots, p-1$.

Now we can generate the transition matrix Q of the Markov chain $\{Z_t, t \in \mathbb{Z}\}$ from $P_{\boldsymbol{\theta}}(z, x)$ since

$$\begin{aligned} Q(z_{t-1}, z_t) &\overset{\Delta}{=} \Pr\left(Z_t = z_t | Z_u = z_u : u < t\right) \\ &= \Pr(Z_t = z_t | Z_{t-1} = z_{t-1}) \\ &= P_{\boldsymbol{\theta}}(z_{t-1}, [z_t]_1) = P_{\boldsymbol{\theta}}(z_{t-1}, x_t), \qquad (2.3) \end{aligned}$$

where $P_{\boldsymbol{\theta}}(\cdot,\cdot)$ is defined in (2.2). We can show that $Q(\cdot,\cdot)$ has a strict positive power $Q^p(\cdot,\cdot)$, i.e., $Q^p(x,y) > 0$ for any $x, y \in \mathcal{X}^p$. According to the definition given by Winkler (1995, Section 4.3)), this type of Markov chains is called *primitive.*

Lemma 2.1. *The process* $\{Z_t = (X'_t, \ldots, X'_{t-p+1})' : t \in \mathbb{Z}\}$*, where* X_t *is defined in* (2.1)*, is a finite, primitive, and aperiodic Markov chain.*

Since the Markov chain $\{Z_t : t \in \mathbb{Z}\}$ has finite state space and is primitive, it is irreducible and has a unique stationary distribution μ where $\lim_{n\to\infty} ||\nu Q^n - \mu|| = 0$ for any initial distribution of the state space, ν, such as the distribution of Z_t (Winkler (1995, Theorem 4.3.1)). Let $\mu(z) = \Pr(Z_t = z)$. The Markov chain $\{Z_t : t \in \mathbb{Z}\}$ is stationary with marginal distribution μ and joint distribution $\mu(x)Q(x,y)$, which can be written as

$$\begin{aligned}\Pr(Z_{t+1} = z_{t+1}, Z_t = z_t) &= \Pr((X'_{t+1}, \ldots, X'_{t-p+1}) = (x'_{t+1}, \ldots, x'_{t-p+1}))\\ &= \mu(x_t, x_{t-1}, \ldots, x_{t-p+1}) P_{\boldsymbol{\theta}}((x'_t, \ldots, x'_{t-p+1}), x'_{t+1}).\end{aligned}$$

Results based on the above arguments are summarized in the following theorem.

Theorem 2.1. *The ESTAR process* $\{Z_t : t \in \mathbb{Z}\}$ *is a stationary stochastic process with* $\Pr(Z_t = z) = \mu(z)$ *and* $\Pr(Z_t = z, X_{t+1} = x) = \mu(z) \cdot P_{\boldsymbol{\theta}}(z,x)$*.*

From now on, let $\boldsymbol{\Theta}$ represent the parameter space of the ESTAR models. Since $|S| < \infty$, we know that $\boldsymbol{\Theta} \subset \mathbb{R}^{p|S|^2+|S|(|S|+1)/2}$, a finite dimensional vector space. The following theorem is one of our main results, its proof is given in the Appendix.

Theorem 2.2. *The ESTAR models in* (2.1) *parameterized with* $\boldsymbol{\Theta}$ *are identifiable.*

In the proof of this theorem, we use the first restriction to the ESTAR models, that is, the support of the conditional distribution at any given site is a finite set $\mathcal{X}_s$ including the value of 0, with $h_s(0) = 1$.

3. Log-Likelihood Function of the Models

This section demonstrates the properties of the log-likelihood function for the space-time models. We show that the MLE is not unique, and give conditions on the observations to guarantee a unique MLE.

Let $f(x|\boldsymbol{\theta})$ be the joint density function of $\{X'_T, \ldots, X'_{1-p}\}$ for $T > 0$. We use $\ell_T(\boldsymbol{\theta})$ to denote the likelihood function as a function of $\boldsymbol{\theta}$ given the observations $\{x_t, t \in 1-p, \ldots, T; x_{t,s} \in \{0, \ldots, M\}, \forall s \in S\}$. In Section 2, we have shown the process $\{Z_t, t \in \mathbb{Z}\}$ is a stationary process. Assume that the process has reached its steady state $Z_t \overset{d}{=} \mu$, the stationary distribution. Also, let $Z_0 = (X_0', \ldots, X_{1-p}')'$ represent the initial conditions of the model with $x = (x_{1-p}', \ldots, x_T')'$ denoting the sample path up to time T, with p initial conditions $(x_{1-p}', \ldots, x_0')$.

The likelihood that we consider treats the initial distribution of Z_0 as fixed at μ, independent of $\boldsymbol{\theta}$, but equal to the invariant distribution associated with the transition matrix $P_{\boldsymbol{\theta}_0}$ that is associated with the true parameter $\boldsymbol{\theta}_0$. However, the likelihood function for the joint distribution under $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ involves calculating the invariant distribution for each value of $\boldsymbol{\theta}$. We simplify matters by maximizing the log-likelihood function which fixes the distribution of $Z_0 \overset{d}{=} \mu$, i.e., the model where the transition matrix belongs to a family, but the initial marginal distribution is not a function of $\boldsymbol{\theta}$.

Since Z_t is a Stationary Markov Chain, then the joint distribution of $\{X_T, \ldots, Z_0\}$ or $\{X_T, \ldots, X_{1-p}\}$ can be specified by the chain rule, which gives the following result.

Proposition 3.1. *The log-likelihood function for the space-time models, with respect to the counting measure $P_{\boldsymbol{\theta}}(\cdot,\cdot)$, is*

$$\begin{aligned}\ell_T(\boldsymbol{\theta}|x) &= \sum_{t=1}^{T} \ln(P_{\boldsymbol{\theta}}(z_{t-1}, x_t)) + \ln(\mu(z_0))\\ &= \sum_{t=1}^{T} \ln(P_{\boldsymbol{\theta}}((x_{t-1}', \ldots, x_{t-p}'), x_t)) + \ln(\mu(z_0)).\end{aligned}$$

Expressing the log-likelihood function in terms of the energy function $U_{\boldsymbol{\theta}}$ (z_{t-1}, x_t) for the ESTAR models, we have

$$\ell_T(\boldsymbol{\theta}|x) = \sum_{t=1}^{T} \{-\ln(c(\boldsymbol{\theta}, z_{t-1})) - U_{\boldsymbol{\theta}}(z_{t-1}, x_t)\} + \ln(\mu(z_0)).$$

For simplicity of presentation, we denote the log-likelihood function by $\ell_T(\boldsymbol{\theta})$ instead of $\ell_T(\boldsymbol{\theta}|x)$ from now on. Moreover, for the purpose of investigating the properties of $\ell_T(\boldsymbol{\theta})$, it is convenient to increase the parameter space of the ESTAR models by additional parameters $\gamma_j(\{s,u\}, v)$ with $\{s, u, v \in S\}$ and $j \in \{1, \ldots, p\}$. These parameters are used to add additional terms $\gamma_j(\{s,u\}, v) \cdot x_{t,s}x_{t,u}x_{t-j,v}$ to the energy function. Furthermore, suppose that we order the sites as $\{s_1, \ldots, s_{|S|}\}$. Then we can define a new random vector of dimension $|S|(|S|+1)/2$

$$W_t = (X_{t,s_1}, \ldots, X_{t,s_{|S|}}, X_{t,s_1}X_{t,s_2}, \ldots, X_{t,s_1}X_{t,s_{|S|}}, X_{t,s_2}X_{t,s_3}, \ldots, X_{t,s_{|S|-1}}X_{t,s_{|S|}})'$$

with each component of W_t denoted by $W_{t,\{s,u\}} = X_{t,s}X_{t,u}$ and $W_{t,s} = X_{t,s}$. The sample values of $W_{t,\{s,u\}}$ and $W_{t,s}$ are denoted by $w_{t,\{s,u\}}$ and $w_{t,s}$, respectively. Using this notation, the additional terms are $\gamma_j(\{s,u\},v)w_{t,\{s,u\}}x_{t-j,v}$. Similarly, we can replace the terms $\gamma_0(s,u)x_{t,s}x_{t,u}$ with $\gamma_0(s,u)w_{t,\{s,u\}}\cdot 1$ and the term $\alpha_s x_{t,s}$ with $\alpha_s w_{t,s}$.

The new parameter space is called $\boldsymbol{\Theta}$. The previous parameter space, now denoted as $\boldsymbol{\Theta'}$, spans a subspace of the new space $\boldsymbol{\Theta}$, defined as $\gamma_j(\{s,u\},v) = 0$ for all elements $\{s,u,v \in S; j = 0,\dots,p\}$. For simplicity of expression, we order the parameter vector as a column vector consisting of $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\gamma}')'$, where the orderings are $\boldsymbol{\alpha} = (\alpha_{s_1},\dots,\alpha_{s_{|S|}})$ and

$$\begin{aligned}\boldsymbol{\gamma} = \Big(&\gamma_0(s_1,s_2),\dots,\gamma_0(s_1,s_{|S|}),\gamma_0(s_2,s_3),\dots,\gamma_0(s_{|S|-1},s_{|S|}),\\ &\gamma_1(s_1,s_1),\dots,\gamma_1(s_1,s_{|S|}),\dots,\gamma_1(s_{|S|},s_{|S|}),\\ &\gamma_1(\{s_1,s_2\},s_1),\dots,\gamma_1(\{s_{|S|-1},s_{|S|}\},s_1),\dots,\gamma_1(\{s_{|S|-1},s_{|S|}\},s_{|S|}),\\ &\gamma_2(s_1,s_1),\dots,\gamma_p(\{s_{|S|-1},s_{|S|}\},s_{|S|})\Big).\end{aligned}$$

Moreover, we use the Kronecker products for matrices and vectors to simplify notation. As used in this paper, the Kronecker product has a higher precedence than addition and multiplication, but a lower precedence than exponentiation. Since w_t, z_t, and $\boldsymbol{\theta}$ are column vectors, $[1, z'_{t-1}]' \otimes w_t$ is the column vector $[w'_t, x_{t-1,s_1}w'_t,\dots,x_{t-p,s_{|S|}}w'_t]'$ and

$$U_{\boldsymbol{\theta}}(z_{t-1},x_t) = -\left\langle \boldsymbol{\theta}, [1, z'_{t-1}]' \otimes w_t\right\rangle - \sum_{s\in S}\ln(h_s(x_{t,s})).$$

Using the previous notation we can express the log-likelihood function as

$$\ell_T(\boldsymbol{\theta}) = \ln\mu(z_0) + \sum_{t=1}^{T}\left(-\ln(c(\boldsymbol{\theta},z_{t-1})) + \left\langle \boldsymbol{\theta}, [1, z'_{t-1}]' \otimes w_t\right\rangle\right) + \sum_{s\in S, t=1}^{T}\ln(h_s(x_{t,s})), \tag{3.1}$$

where

$$c(\boldsymbol{\theta},z_{t-1}) = \sum_{x_t\in\{0,\dots,M\}^{|s|}}\left(\exp\left\langle \boldsymbol{\theta}, [1, z'_{t-1}]' \otimes w_t\right\rangle \prod_{s\in S} h_s(x_{t,s})\right). \tag{3.2}$$

Theorem 3.1.

1. *The Fisher Information* (**FI**) *regularity conditions hold for* $\ell_T(\boldsymbol{\theta})$*:*
   - *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *and* $x \in \mathcal{X}^{T+p}$ *with* $T > 0$*, the gradient with respect to* $\boldsymbol{\theta}$*,* $\nabla\ell_T(\boldsymbol{\theta})$*, exists;*
   - *the set* $C = \{x : f(x|\boldsymbol{\theta}) > 0\}$ *does not depend on* $\boldsymbol{\theta}$*;*

- *for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $x \in \mathcal{X}^{T+p}$ with $T > 0$, the Hessian matrix $\nabla^2 \ell_T(\boldsymbol{\theta})$ is negative semidefinite.*

2. *The gradient $\nabla \ell_T(\boldsymbol{\theta})$ and Hessian $\nabla^2 \ell_T(\boldsymbol{\theta})$ are Lipschitz continuous in $\boldsymbol{\theta}$ and the modulus of continuity does not depend on x.*
3. *For all $x \in \mathcal{X}^{T+p}$, $\ell_T(\boldsymbol{\theta})$ is analytic about any point in $\boldsymbol{\Theta}$.*

The gradient $\nabla \ell_T(\boldsymbol{\theta})$ is also called the *score function.* Since the log-likelihood function satisfies the **FI** regularity condition, we have the following results based on Proposition 2.84 in Schervish (1995).

Corollary 3.1. *For all $T > 0$ and any distribution for the initial states μ_{z_0}, if the process $\{X_t, t \in \mathbb{Z}\}$ is generated by $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, then $\mathbb{E}_{\boldsymbol{\theta}_0}[\nabla \ell_T(\boldsymbol{\theta}_0)] = 0$ and*

$$I_T(\boldsymbol{\theta}_0) = \frac{E_{\boldsymbol{\theta}_0}[\nabla \ell_T(\boldsymbol{\theta}_0) \nabla \ell_T(\boldsymbol{\theta}_0)']}{T} = \frac{-\mathbb{E}_{\boldsymbol{\theta}_0}(\nabla^2 \ell_T(\boldsymbol{\theta}_0))}{T}.$$

Whenever the ESTAR process is stationary, the Fisher Information $I_T(\boldsymbol{\theta}_0)$ is the same for all $T > 0$, so we can drop the T subscript. The ESTAR process is stationary whenever the process starts at $t = -\infty$ or when $\mu_{z_0} = \mu$, the invariant distribution of Z_t.

The results in Corollary 3.1 allow us to use the negative of the expected value of the Hessian to represent the Fisher Information $I(\boldsymbol{\theta}_0)$. The negative of the Hessian is also known as the *observed information.* Unlike the exponential distribution, the Fisher information and the observed Fisher information are not the same in the space-time models defined in (2.1).

The score function $\nabla \ell_T(\boldsymbol{\theta}_0)$, expressed in equation (A.2) and being the sum of terms whose expectation is zero, can usually be described as a martingale. The result is stated in the following corollary, the proof is omitted.

Corollary 3.2. *Under $\boldsymbol{\theta}_0$, $\{\nabla \ell_T(\boldsymbol{\theta}_0), T = 1, 2, 3, \ldots\}$ is a martingale.*

The result in Corollary 3.2 is true for any parameterized transition function (see, e.g., Greenwood and Wefelmeyer (1997, p.107)). The fact that $\{\nabla \ell_T(\boldsymbol{\theta}_0), T = 1, 2, 3, \ldots\}$ is a martingale will help us prove asymptotic normality of the score function, a requirement for asymptotic normality of the MLE.

We use the following lemma on Kronecker products to derive several results concerning the Hessian $\nabla^2 \ell_T(\boldsymbol{\theta})$. The proof of this lemma is straightforward, thus omitted.

Lemma 3.1. *Let A_m be a positive semidefinite matrix, B_n be any symmetric matrix, with $m, n > 0$, and I_n an n by n identity matrix. Suppose λ_{min} is the minimum eigenvalue of B_n and λ_{max} is the maximum eigenvalue of B_n, then for* *any column vector* $\boldsymbol{a}$ *in* $\mathbb{R}^{mn}$,

$$\lambda_{min}\boldsymbol{a}'(A_m \otimes I_n)\boldsymbol{a} \le \boldsymbol{a}'(A_m \otimes B_n)\boldsymbol{a} \le \lambda_{max}\boldsymbol{a}'(A_m \otimes I_n)\boldsymbol{a} \ .$$

Based on the results in Lemma 3.1, we give the sufficient and necessary conditions under which the Hessian $\nabla^2\ell_T(\boldsymbol{\theta})$ is strictly negative definite. The proof of the following theorem is given in the Appendix.

Theorem 3.2. *The Hessian* $\nabla^2\ell_T(\boldsymbol{\theta})$ *for an ESTAR model is strictly negative definite if and only if* $\{z_t - z_0, \ t \in \{1, \dots T-1\}$ *spans* $\mathbb{R}^{p|S|}$*, if and only if* $\mathrm{Var}\,(z_t, t = 0, \dots, T-1)$ *is positive definite, where* $\mathrm{Var}\,(z_t, t = 0, \dots, T-1)$ *is the sample covariance matrix of* $\{Z_t, t = 0, \dots, T-1\}$ *based on the observed vectors* $\{z_0, z_1, \dots, z_{T-1}\}$.

For the ESTAR models, if the Hessian $\nabla^2\ell_T(\boldsymbol{\theta})$ is negative definite for a given value of $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, then it is negative definite for all values of $\boldsymbol{\theta}$. Thus if a maximum of the log-likelihood function exists then this maximum is unique, since the Hessian is a strictly concave function of $\boldsymbol{\theta}$ given x.

Now suppose that the Hessian $\nabla^2\ell_T(\boldsymbol{\theta})$ is only negative semi-definite and at least one MLE, $\hat{\boldsymbol{\theta}}$, exists. Then by the proof of Theorem 3.2, there exists some vector $\boldsymbol{a}_0$ such that $\langle \boldsymbol{a}_0, [1, z_t']'\rangle = 0$ for all $t \in \{0 \dots T-1\}$. Now for any trajectory of the form $\hat{\boldsymbol{\theta}} + t(\boldsymbol{a}_0 \otimes \boldsymbol{\beta})$ where $\boldsymbol{\beta} \in \mathbb{R}^{|S|(|S|+1)/2}$, consider the function $M(t) = \ell_T(\hat{\boldsymbol{\theta}} + t(\boldsymbol{a}_0 \otimes \beta))$. Since the log-likelihood function $\ell_T(\cdot)$ is Lipschitz continuous and the gradient $\nabla\ell_T(\cdot)$ exists everywhere, the function $M(t)$ is continuous with derivatives everywhere in $\mathbb{R}$. Notice that $\langle \boldsymbol{a}_0 \otimes \boldsymbol{\beta}, [1, z_t']' \otimes w_t\rangle = 0$. From the expression of $\ell_T(\cdot)$ given in (3.1) we can show that the derivative of $M(t)$ with respect to t is zero everywhere. Therefore $M(t)$ is a constant function on $\mathbb{R}$. Thus any point on the trajectory $\hat{\boldsymbol{\theta}}+t(\boldsymbol{a}_0\otimes\beta)$ maximizes the function $\ell_T(\hat{\boldsymbol{\theta}}+t(\boldsymbol{a}_0\otimes\beta))$. In other words, the MLE is not unique when the Hessian $\nabla^2\ell_T(\boldsymbol{\theta})$ is only negative semi-definite. The results on the uniqueness of the MLE are summarized in the following Corollary.

Corollary 3.3. *The maximum likelihood estimate of the vector parameter in an ESTAR model, if it exists, is unique if and only if the Hessian* $\nabla^2\ell_T(\boldsymbol{\theta})$ *is strictly negative definite for some value of* $\boldsymbol{\theta}$.

4. Asymptotic Properties of the MLE

In this section, we prove consistency and asymptotic normality of the maximum likelihood estimates. Let us assume that the random vectors $\{(X_t, \dots, X_{t-p}), t \in 1, \dots, T\}$ come from a stationary ESTAR model with parameter $\boldsymbol{\theta}$. Since the stationary time series admits a unique measure, we use $\mathbb{E}_{\boldsymbol{\theta}}H(\cdot)$ to denote the expected value of any measurable function $H(\cdot)$ of the random vectors

$\{X_t : t \in \mathbb{Z}\}$. We use the concept of objective and reference functions to prove consistency of the MLE. Once consistency is demonstrated, we apply a central limit theorem for martingales to prove asymptotic normality of the MLE.

4.1. Strong consistency

We prove strong consistency of the MLE instead of asymptotic consistency or convergence in probability. Winkler (1995) discussed the use of objective functions to prove asymptotic consistency. We demonstrate the strong consistency of the MLE using strengthened versions of Winkler's Theorems 13.4.1 and 13.4.2, and a modified definition of the objective function.

First, we need to give the definitions for objective functions and a reference function. These functions are described by a set of properties. The parameter space for the ESTAR models is $\boldsymbol{\Theta} \subset \mathbb{R}^d$.

Definition 4.1. A *reference function* is a function g from $\boldsymbol{\Theta}$ onto $\mathbb{R}$ such that it has a unique maximum, $\boldsymbol{\theta}_0$, and for which there exists a $\gamma > 0$ such that $g(\boldsymbol{\theta}) \leq -\gamma\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 + g(\boldsymbol{\theta}_0)$, $\forall\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r)$, where $B(\boldsymbol{\theta}_0, r)$ is a closed ball with finite radius $r > 0$.

Consider a stationary random count process $\{X_t, t = 1, 2, 3, \ldots\}$. Let $X = (X_1', \ldots, X_T')'$, and let $x = (x_1', \ldots, x_T')'$ represent a sample path of X.

Definition 4.2. We call a sequence of functions $\{G_T(\boldsymbol{\theta}, x), T = 1, 2, 3, \ldots\}$ from $\boldsymbol{\Theta} \times \mathcal{X}^T$ onto $\mathbb{R}$ *objective functions* with the reference function $g(\boldsymbol{\theta})$, if for each sample path x, each $G_T(\boldsymbol{\theta}, x)$ is a concave function of $\boldsymbol{\theta}$, and $G_T(\boldsymbol{\theta}, x) \overset{\text{a.s.}}{\to} g(\boldsymbol{\theta})$ as $T \to \infty$ uniformly $\forall\, \boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r)$, where $\boldsymbol{\theta}_0$ is the unique maximum of $g(\boldsymbol{\theta})$.

From Proposition 3.1, we notice that the log-likelihood function for the parameters in the ESTAR models is a sum of functions, $\ln(P_{\boldsymbol{\theta}}((x_{t-1}', \ldots, x_{t-p}'), x_t))$. We propose some candidate functions for a reference function and a sequence of objective functions based on the log-likelihood function. Specifically, we define $G_T(\boldsymbol{\theta}, x)$ and $g(\boldsymbol{\theta})$ as follows.

$$G_T(\boldsymbol{\theta}, x) = \frac{\ell_T(\boldsymbol{\theta}) - \mu(z_0)}{T} = \frac{\sum_{t=1}^{T} \ln(P_{\boldsymbol{\theta}}(z_{t-1}, x_t))}{T}, \tag{4.1}$$

$$\begin{aligned} g(\boldsymbol{\theta}) &= \mathbb{E}_{\mu Q}(G_1(\boldsymbol{\theta}, x)) = \mathbb{E}_{\boldsymbol{\theta}_0} \ln(P_{\boldsymbol{\theta}}(Z_0, X_1)) \\ &= \mathbb{E}_{\boldsymbol{\theta}_0}\left[\langle \boldsymbol{\theta}, [1, Z_0]' \otimes W_1 \rangle - \ln(c(\boldsymbol{\theta}, Z_0))\right] + \sum_{s \in S} \ln(h_s(x_{1,s})). \end{aligned} \tag{4.2}$$

Since $\{Z_t, t \in \mathbb{Z}\}$ is stationary, we have $\forall t > 0$,

$$\begin{aligned} g(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}_0} \ln(P_{\boldsymbol{\theta}}(Z_{t-1}, X_t)) \\ &= \mathbb{E}_{\boldsymbol{\theta}_0} \left[\langle \boldsymbol{\theta}, [1, Z'_{t-1}]' \otimes W_t \rangle - \ln(c(\boldsymbol{\theta}, Z_{t-1}))\right] + \sum_{s \in S} \ln(h_s(x_{t,s})). \end{aligned}$$

Remark 4.1. For a given $T > 0$, the candidate objective function, $G_T(\boldsymbol{\theta}, x)$, is the average contribution of each term in the log-likelihood function. For a given sample x, maximizing $\ell_T(\boldsymbol{\theta})$ is equivalent to maximizing $G_T(\boldsymbol{\theta}, x)$.

Remark 4.2. The candidate reference function $g(\boldsymbol{\theta})$ is well defined since $\ln(P_{\boldsymbol{\theta}}$ $(Z_{t-1}, X_t))$ is a measurable function and we are sampling from the stationary distribution. Since the likelihood function is positive and bounded on the support of X_t, (although not uniformly so for all $\boldsymbol{\theta}$), the value of the reference function for a given $\boldsymbol{\theta}$ exists and is finite.

In order to prove the strong consistency of the MLE, we must show that our chosen candidate reference function is a valid reference function. Based on Definition 4.1 and Lemma C.0.3 in Winkler (1995), this is the same as proving the following lemma, the proof of which is omitted.

Lemma 4.1. *The candidate reference function g defined by* (4.2) *is twice differentiable and strictly concave with a unique maximum at $\boldsymbol{\theta}_0$. Therefore g is a valid reference function.*

Now it remains to show that $\{G_T(\boldsymbol{\theta}, x), T = 1, 2, \ldots, \}$ is a sequence of objective functions. For this purpose, we first show some relationship between the reference function, g and the candidate objective functions $\{G_T(\boldsymbol{\theta}, x), T = 1, 2, \ldots, \}$. The results on this relationship are given in the following two lemmas, where Lemma 4.3 is a stronger version of Lemma 13.4.2 of Winkler (1995). Proofs are omitted.

Lemma 4.2. *For all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:*

$$\begin{gathered} G_T(\boldsymbol{\theta}, X) \overset{\text{a.s.}}{\to} g(\boldsymbol{\theta}), \\ \nabla \frac{\ell_T(\boldsymbol{\theta})}{T} = \nabla G_T(\boldsymbol{\theta}, X) \overset{\text{a.s.}}{\to} \nabla g(\boldsymbol{\theta}) \ \textit{and}\ \nabla g(\boldsymbol{\theta}_0) = 0, \\ \nabla^2 \frac{\ell_T(\boldsymbol{\theta})}{T} = \nabla^2 G_T(\boldsymbol{\theta}, X) \overset{\text{a.s.}}{\to} \nabla^2 g(\boldsymbol{\theta}) = I(\boldsymbol{\theta}_0), \quad \textit{if}\ \boldsymbol{\theta} = \boldsymbol{\theta}_0, \end{gathered}$$

where the convergence is almost surely with respect to μQ, equivalently with respect to $\mu P_{\boldsymbol{\theta}}$, where $\nabla g(\boldsymbol{\theta}_0) = 0$ and $\nabla^2 g(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)$.

Lemma 4.3. *Let $\boldsymbol{\Theta}$ be an open subset of $\mathbb{R}^d$. Suppose that $\{G_T(\cdot, x), x \in \mathcal{X}; T \geq 1\}$ and $g(\cdot)$ are Lipschitz continuous in $\boldsymbol{\theta}$ with a common Lipschitz constant.*

Suppose further that for every $\boldsymbol{\theta} \in \Theta$, $G_T(\boldsymbol{\theta}, x) \overset{\text{a.s.}}{\to} g(\boldsymbol{\theta})$ *as* $T \to \infty$. *Then, for each* $r > 0$ *there exists a set* A *with* $\Pr_{\boldsymbol{\theta}_0}(A) = 1$ *such that for all* $x \in A$, $\lim_{T\to\infty} G_T(\boldsymbol{\theta}, x) = g(\boldsymbol{\theta})$ *uniformly for* $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r) \cap \boldsymbol{\Theta}$.

Based on the results in Lemmas 4.2 and 4.3, we have the following.

Lemma 4.4. *The functions* $\{G_T(\boldsymbol{\theta}, x), T \geq 1\}$ *form a sequence of objective functions.*

Assume that we have a family of stochastic processes $\{X_t,\ t \in \mathbb{Z}; X_t \in \mathbb{R}^{|S|}\}$ indexed by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and defined on a probability space $(\Omega, \mathcal{F}, \Pr_{\boldsymbol{\theta}})$, with sample path $\omega \in \Omega$ such that $X_t(\omega) = x_t$. Assume that we are sampling from this process with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is the true vector parameter in the interior of $\boldsymbol{\Theta}$. We can define an estimate

$$\hat{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} G_T(\boldsymbol{\theta}, x). \tag{4.3}$$

If there is more than one maximum choose one from the set. The following theorem is a stronger version of Lemma 13.4.1 in Winkler (1995).

Theorem 4.1. *Let* $\boldsymbol{\Theta} \subset \mathbb{R}^d$ *be open. If* $\{G_T(\boldsymbol{\theta}, x)\}$ *is a sequence of objective functions with reference function* g, *then* $\hat{\boldsymbol{\theta}}_T \overset{\text{a.s.}}{\to} \boldsymbol{\theta}_0$ *as* $T \to \infty$.

As we discussed in Remark 4.1, maximizing $G_T(\boldsymbol{\theta}, x)$ over $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is equivalent to maximizing the log-likelihood function $\ell(\boldsymbol{\theta})$. Therefore $\hat{\boldsymbol{\theta}}_T$ defined in (4.3) is a maximum likelihood estimate of the vector parameter $\boldsymbol{\theta}$.

Corollary 4.1. *The maximum likelihood estimate* $\hat{\boldsymbol{\theta}}_T$ *of an ESTAR model is strongly consistent, where* $\hat{\boldsymbol{\theta}}_T = \arg\max G_T(\boldsymbol{\theta}, x)$.

4.2. Asymptotic normality

When the ESTAR process defined in (2.1) is stationary, the maximum likelihood estimates of the parameters in the model are asymptotically normally distributed.

Theorem 4.2. *The MLE for the ESTAR process, when* $\boldsymbol{\theta}_0$ *is in the interior of* $\boldsymbol{\Theta}$, *satisfies*

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1}\right),$$

where $I(\boldsymbol{\theta}_0)$ *is the Fisher information matrix.*

One method of proving asymptotic normality relies on the asymptotic normality of the score function along with a smoothness property of the Fisher information. To that end we first demonstrate that the space-time models possess these properties in the following lemma, then use the results in Theorem 5.2.2 of Sen and Singer (1993, p.209) to prove the asymptotic normality of the MLE. The proof of the lemma is based on Lemma 4.2 and Corollary 3.2.

Lemma 4.5. *The ESTAR process with the true parameter* $\boldsymbol{\theta}_0$ *in the interior of* $\boldsymbol{\Theta}$ *has a positive definite and finite Fisher information Matrix* $I(\boldsymbol{\theta}_0)$*. Moreover,*

$$\mathbb{E}_{\boldsymbol{\theta}_0}\Big[\sup_{\{h:\|h\|\leq\delta\}}\|\nabla^2 G_T(\boldsymbol{\theta_0}+\boldsymbol{h},\boldsymbol{x})-\nabla^2 G_T(\boldsymbol{\theta_0},\boldsymbol{x})\|\Big]=\psi_\delta\to 0 \quad as \quad \delta\to 0,$$

and $\sqrt{T}\nabla G_T(\boldsymbol{\theta}_0,x)\xrightarrow{d}N(0,I(\boldsymbol{\theta}_0))$*.*

These properties in addition to the Fisher information conditions proved in Theorem 3.1 are sufficient for proving that the maximum likelihood estimates are asymptotically normal.

We conclude by noting that since the conditions for consistency and asymptotic normality do not require that the initial distribution of Z_0 be the limiting distribution, these results also hold for any initial distribution of the ESTAR process. For example, we may find the MLE of the parameters assuming that the initial values are zero, and the estimates will still be consistent and asymptotically normal.

5. Estimation Methods and Simulation Results

This section discusses estimation methods for the parameters in the ESTAR model and present some simulation results. If we know the likelihood of a given model exactly, then we can determine the maximum likelihood estimator. However, for the ESTAR models, the exact distribution is not known. At each time period of our model, there is an intractable constant of proportionality that depends not only on the parameter values but also the past data values.

In spatial models, where the constant of proportionality depends only on the parameter values, there are several choices for parameter estimation. For example, in models where the mean and variances of the conditional distribution at each lattice site can be calculated from the parameters, the Maximum Pseudo Likelihood Estimation (MPLE) method (see, e.g., Besag (1974) and Winkler (1995)) or the Markov Chain Monte Carlo Maximum Likelihood (MCML) method (Geyer (1994), Huffer and Wu (1998) and Jagger, Niu and Elsner (2002)) can be used to estimate the parameters.

The MPLE method is so named because the estimates are the parameter values maximizing the product of the conditional likelihood functions. The MPLE was first used by Besag (1975) for his auto-logistic models. It is easy to use with conditionally specified models, since the conditional distributions are already given. Let us assume a conditional STAR model with D parameters such that the conditional density of $\{X_{t,s}, t \in 1,\ldots,T, s \in S\}$ given $\{x_{m,s} : m < t, s \in S;\ \ x_{t,u}, u \neq s\}$, with respect to the Lebesgue measure or a counting measure is

$$f_{\boldsymbol{\theta},t,s}(x) = \frac{h_{t,s}(x) \cdot \exp \langle \boldsymbol{T}_{t,s}(x_{\partial\{(t,s)\}}, x), \boldsymbol{\theta}\rangle}{c_{t,s}(\boldsymbol{\theta}, x_{\partial\{(t,s)\}})},$$

where x_t is a time series of configurations, on a lattice S of finite size, and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ with $\boldsymbol{\Theta} \subset \mathbb{R}^D$. We use $\partial\{(t,s)\}$ to represent the space-time neighborhood of (t,s) and $x_{\partial\{(t,s)\}}$ the vector of values in this neighborhood. Also, for each (t,s), $\boldsymbol{T}_{t,s}(\cdot,\cdot)$ is a D-dimensional measurable function of the values. The MPLE is then

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \sum_{t=1}^{T} \sum_{s\in S} - \ln(c_{t,s}(\boldsymbol{\theta}, x_{\partial\{(t,s)\}})) + \langle \boldsymbol{T}_{t,s}(x_{\partial\{(t,s)\}}, x_{t,s}), \boldsymbol{\theta}\rangle.$$

If a spatial model is shift invariant with finite range, the MPLE of the parameter vector in the model is asymptotically consistent with increasing domains (Winkler (1995), Theorem 14.3.1). However, the estimator is not necessarily efficient, and does not provide standard error estimates.

The MCML method, Geyer (1994), estimates the maximum likelihood when the formula for a model's distribution contains an intractable or unknown normalizing constant. The MCML method is computationally intensive, and requires an initial parameter estimate that is reasonably close to the MLE, lest the method fail to produce any estimate. In the spatial lattice case, Wu (1994) applied this methodology to determining the parameters in an auto-logistic regression model. This method is derived by noting that the ratio of likelihood functions can be estimated using Markov Chain Monte Carlo methods. In this paper, the MCML method for spatial processes is extended and applied to parameter estimation for ESTAR models.

For illustrating the MPLE and MCML methods and assessing the final sample properties of the estimates, we used a lag one (p=1) nearest neighbor TPSTAR model for simulation and parameter estimation. The site space is $S = \{1,\ldots,m\} \times \{1,\ldots,n\}$, and the time dimension is $\{1,\ldots,T\}$ with $T > 0$. We denote the random variable at each site as $X_{t,i,j}$ taking on values $x_{t,i,j}$, with $t \in 1,\ldots,T$ and $(i,j) \in S$. We set the boundary conditions to zero. That is $X_{0,i,j} = X_{t,m+1,j} = X_{t,i,n+1} = X_{t,0,j} = X_{t,i,0} = 0$.

The model is expressed as

$$\begin{aligned} &X_{t,i,j}|X_{s,k,l} \sim \text{tpois}(\lambda_{t,i,j}, M),\\ &\quad \ln(\lambda_{t,i,j}) = \gamma_v(X_{t,i-1,j}+X_{t,i+1,j})+\gamma_h(X_{t,i,j-1}+X_{t,i,j+1})+\gamma_c X_{t-1,i,j}+\alpha+\beta z_{i,j},\\ &\quad \text{w}ith\ \ z_{i,j} = \sin(\omega(i+j)), \end{aligned} \tag{5.1}$$

where $\mathrm{tpois}(\lambda, M)$ is the notation for the right truncated Poisson distribution with parameter λ and maximum value M. Note that the neighborhood of each point consists of five lattice points, they are the two nearest horizontal, the two nearest vertical, and the previous point.

In order to verify that the MCML algorithm works, we performed simulations on five parameter vectors using the TPSTAR model defined by (5.1). Each parameter vector consists of $\{\gamma_h, \gamma_v, \gamma_c, \alpha, \beta\}$. The five parameter vectors are presented in the first column of Table 1. We set the frequency $\omega = 0.2$ for the simulation, and $\beta = 0.5$. The simulation was done on a 40 by 40 array with $T = 5$ and boundary conditions set to be zero outside the array, and we initialized $x_0 = \{x_{0,i,j}\} = \mathbf{0}$. Moreover, we set the maximum, M, or the truncation value to 10 for the simulations.

For each given parameter vector, 100 simulations were performed. The Gibbs sampler was used to produce the samples from Model (5.1). Specifically, given the initial values $x_0 = \{x_{0,i,j}\} = \mathbf{0}$ and the boundary conditions, we can generate sample x_1, a vector of length 40×40 from the distribution $\{X_1 | x_t : t < 1\}$ using the Gibbs sampler. The Gibbs sampler can be used because the ESTAR model is defined so that conditioned on the past, the distribution at each stage t is a Gibbs field, determined by its local characteristics. Next, we repeat this for each value of $\{t \in 1, \ldots, T\}$, and generate a single sample of $\{X_t | X_s : s < t\}$. After T steps the process generates one sample from the joint distribution of $\{X_1, \ldots, X_T\}$ given $\{X_s = x_s : s < 1\}$. This method provides perfect samples of the process, that is the distribution of samples under this sampling method follows the TPSTAR process in Model (5.1), under the given parameters and boundary conditions (see, e.g., Propp and Wilson (1996)).

For each simulation, the five parameters were estimated using both MPLE and MCML methods. The simulation results are shown in Table 1. The second and third columns of Table 1 show the MPLE and MCML estimates of the actual parameters, while the next two columns are the mean, standard deviation of the parameter estimates based on the 100 simulations. The last column of Table 1 presents the mean of the modified MCML standard error estimates derived from the estimated Fisher information matrix.

From the results in Table 1, we cannot conclude that the MCML is a better parameter estimator than the MPLE, as reported in Wu (1994) for the autologistic spatial model. Both the MCML and the MPLE parameter estimates appeared relatively unbiased as compared to their standard errors. However the primary reason for using MCML is that, unlike MPLE, it provides reasonable standard error estimates, Wu (1994). In our simulation, the mean MCML standard error estimates are within 10% of the actual standard errors, as estimated by simulation.

Table 1. MPLE and MCML estimates of the parameters in Model (5.1) based on 100 simulations.

		Mean of Estimates		Standard Deviations		Mean Standard Errors
Parameter		MPLE	MCML	MPLE	MCML	MCML
γ_v	0.2	0.2038	0.2040	0.0150	0.0156	0.0136
γ_h	0.2	0.1979	0.1979	0.0152	0.0155	0.0135
γ_c	0.4	0.4320	0.4189	0.0997	0.0736	0.0652
α	-2.0	-2.0113	-2.0075	0.0560	0.0546	0.0588
β	0.5	0.5034	0.4934	0.0648	0.0617	0.0623
γ_v	0.0	-0.0035	-0.0049	0.0726	0.0719	0.0670
γ_h	0.0	-0.0030	-0.0021	0.0677	0.0667	0.0670
γ_c	0.4	0.3920	0.3922	0.0549	0.0553	0.0592
α	-2.0	-1.9983	-1.9984	0.0470	0.0465	0.0433
β	0.5	0.5031	0.5034	0.0417	0.0425	0.0453
γ_v	-0.5	-0.5033	-0.5032	0.0921	0.0925	0.1106
γ_h	-0.5	-0.5066	-0.5049	0.0993	0.0967	0.1108
γ_c	0.4	0.3830	0.3834	0.0725	0.0699	0.0740
α	-2.0	-1.9993	-1.9997	0.0463	0.0456	0.0440
β	0.5	0.5033	0.5034	0.0511	0.0518	0.0517
γ_v	0.0	0.0006	0.0006	0.0025	0.0023	0.0024
γ_h	0.0	0.0002	0.0003	0.0027	0.0024	0.0024
γ_c	0.4	0.3984	0.3983	0.0050	0.0049	0.0058
α	1.0	0.9942	0.9941	0.0197	0.0196	0.0222
β	0.5	0.4977	0.4978	0.0144	0.0144	0.0137
γ_v	-0.5	-0.4989	-0.4958	0.0365	0.0342	0.0334
γ_h	-0.5	-0.5097	-0.5065	0.0363	0.0343	0.0343
γ_c	0.4	0.3998	0.4000	0.0078	0.0074	0.0065
α	1.0	1.0016	1.0006	0.0181	0.0158	0.0167
β	0.5	0.4989	0.4996	0.0198	0.0189	0.0191

6. Conclusions and Discussion

Testing spatial and temporal correlations is an important topic in understanding and modeling the structure of a space-time process. One possible approach for achieving this purpose is model selection. For example, consider two TPSTAR models with parameters

$$\ln(\lambda_{t,i,j}) = \gamma_v(X_{t,i-1,j} + X_{t,i+1,j}) + \gamma_h(X_{t,i,j-1} + X_{t,i,j+1}) + \gamma_c X_{t-1,i,j}, \qquad (6.1)$$

$$\ln(\lambda_{t,i,j}) = \gamma_v(X_{t,i-1,j} + X_{t,i+1,j}) + \gamma_h(X_{t,i,j-1} + X_{t,i,j+1}). \qquad (6.2)$$

Various criteria can be used to compare the two models. If Model (6.2) is chosen instead of Model (6.1), we conclude that the temporal dependence of the space-time process $\{X_{t,i,j}\}$ is not significant. A similar approach can be applied to test the spatial dependence of the process. Jagger, Niu and Elsner (2002) used the Bayesian Information Criterion (BIC) to select models for annual North Atlantic hurricane activity. The final model they chose showed that both spatial and temporal dependences were significant in describing the hurricane activity process. Other procedures of testing spatial and temporal independence may also be developed. We will pursue this topic in our future research.

The ESTAR models with time varying covariates are non-stationary. For simultaneously specified STAR models with nonstationary covariates, for example Niu and Tiao (1995), it is clear how to separate the model into a deterministic time varying component and a stationary component. This separation may not be possible for the general ESTAR models with nonstationary covariates, the theory in this paper only applies to stationary ESTAR processes. Statistical properties of non-stationary ESTAR models will be investigated in future studies.

Acknowledgement

The authors would like to thank the Editor, an associate editor, and a referee for their insightful comments and suggestions that let to an improved version of this paper. The authors are also thankful to Professors James B. Elsner, Fred W. Huffer and Ian McKeague, Florida State University, for their helpful comments and discussions.

Appendix. Proofs of the Main Results

Proof of Proposition 2.1. Fix t and assume that $\{X_v, v < t\}$ is known. The conditional distribution of $X_{t,s}$ is

$$\begin{aligned}
&\Pi(x_{t,s}|x_v : v < t;\ x_{t,u} : u \neq s) \cdot c_s(\boldsymbol{\theta}, z_{t-1}, x_{t,u} : u \neq s)\\
&= \exp\Big\{\alpha_s x_{t,s} + \sum_{j=1}^{p}\sum_{u\in S}\gamma_j(u,s)x_{t-j,u}x_{t,s} + \sum_{u\in S}\gamma_0(u,s)x_{t,u}x_{t,s} + \ln h_s(x_{t,s})\Big\}\\
&= \exp\Big\{\Big(\alpha_s + \sum_{j=1}^{p}\sum_{u\in S}\gamma_j(u,s)x_{t-j,u}\Big)x_{t,s} + \sum_{u\in S}\gamma_0(u,s)x_{t,u}x_{t,s} + \ln h_s(x_{t,s})\Big\}\\
&= \exp\Big\{\Big(\alpha'_s + \sum_{u\in S}\gamma_0(u,s)x_{t,u}\Big)x_{t,s} + \ln h_s(x_{t,s})\Big\}, \qquad (A.1)
\end{aligned}$$

where $\alpha'_s = \alpha_s + \sum_{j=1}^{p}\sum_{u\in S}\gamma_j(u,s)x_{t-j,u}$. The conditional distributions at each site given the values at the other sites form a family of exponential distributions with one natural parameter, $\alpha'_s + \sum_{u\in S}\gamma_0(u,s)\cdot x_{t,u}$, $T_{1,s}(x_{t,s}) = x_{t,s}$ and $T_{0,s}(x_{t,s}) = \ln h_s(x_{t,s})$.

Since $\gamma_0(u,s)=\gamma_0(s,u)$ and $\gamma_0(s,s)=0$ by the definition in (2.1), based on the arguments in Section 6.4.2 of Cressie (1993, pp.419-422) the conditional distributions in (A.1) uniquely specify a Gibbs energy function on the configuration at time t, up to a constant $K(z_{t-1})$. We denote the Gibbs energy function of X_t given z_{t-1} as $U_{\boldsymbol{\theta}}(z_{t-1},x_t)+K(z_{t-1})$.

Based on the form of the conditional distribution we have

$$\begin{aligned}
&U_{\boldsymbol{\theta}}(z_{t-1},x_t)\\
&=-\sum_{s\in S}(\alpha_s' x_{t,s}+\ln h_s(x_{t,s}))-\sum_{\{u,s\in S\}}\gamma_0(u,s)x_{t,u}x_{t,s}\\
&=-\sum_{s\in S}(\alpha_s x_{t,s}+\ln h_s(x_{t,s}))-\sum_{j=1}^{p}\sum_{(u,s)\in S^2}\gamma_j(u,s)x_{t-j,u}x_{t,s}-\sum_{\{u,s\in S\}}\gamma_0(u,s)x_{t,u}x_{t,s}.
\end{aligned}$$

This expression closely follows the form given by Cressie (1993, p.421). Now the joint distribution of the vector X_t given the past exists if the partition function is finite. Since X_t has $(M+1)^{|S|}$ configurations, a finite number, the partition function is finite and always exists. Thus the distribution of X_t given the past, $\{X_{t-1},\ldots X_{t-p}\}$, is a Gibbs field.

Proof of Theorem 2.2. Suppose a space-time model defined in (2.1) is not identifiable. Then there exist two vector parameters $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2\in\boldsymbol{\Theta}$ such that the distributions of all finite combinations of $\{X_t, t\in\mathbb{Z}\}$ are the same. This implies that the conditional distributions must be the same as long as the joint distributions are positive, i.e., the model is not conditionally identifiable.

Since the process $\{X_t, t\in\mathbb{Z}\}$ is a Markov chain of order p, for any initial distribution of $\{X_0,\ldots,X_{p-1}\}$ the joint distribution of $\{X_p,\ldots,X_{2p-1}\}$ is positive by Theorem 2.1. Thus we can look at the conditional distribution of X_{2p} given $\{X_p,\ldots,X_{2p-1}\}$. Since the conditional distribution must be the same for any value of X_{2p}, let $X_{2p,s}=0\ \forall s\in S$, then for each value of $X_p,\ldots,X_{2p-1}$ we must have $\Pr(X_{2p}=0|X_p,\ldots,X_{2p-1})=c(\boldsymbol{\theta}_1,X_p,\ldots,X_{2p-1})^{-1}=c(\boldsymbol{\theta}_2,X_p,\ldots,X_{2p-1})^{-1}$. Now this implies that the conditional potential functions given in Proposition 2.1 must be the same for the two sets of parameters. Thus, for all possible values of $(X_p,\ldots,X_{2p})$,

$$U_{\boldsymbol{\theta_1}}((X_{2p-1}',\ldots,X_p'),X_{2p}')=U_{\boldsymbol{\theta_2}}((X_{2p-1}',\ldots,X_p'),X_{2p}').$$

Now $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$ differ in at least one component, either α_s or $\gamma_j(s,u)$. First assume that they differ for α_s, say $\alpha_{s1}\neq\alpha_{s2}$. In this case, we choose $\{X_{p,s}=0,\ldots,X_{2p-1,s}=0;\forall s\in S\}$, $\{X_{2p,u}=0; u\neq s\}$, and $X_{2p,s}=1$. Then $U_{\boldsymbol{\theta_1}}((X_{2p-1}',\ldots,X_p'),X_{2p}')=-\alpha_{s1}\cdot 1+\log h_s(1)$ and $U_{\boldsymbol{\theta_2}}((X_{2p-1}',\ldots,X_p'),X_{2p}')$ $=-\alpha_{s2}\cdot 1+\log h_s(1)$. These must be the same for both α_{s1} and α_{s2}, but this is impossible, since $h_s(1)>0$.

Now suppose $U_{\boldsymbol{\theta_1}}((X'_{2p-1},\dots,X'_p),X'_{2p})$ and $U_{\boldsymbol{\theta_2}}((X'_{2p-1},\dots,X'_p),X'_{2p})$ differ for some component $\gamma_j(s,u)$, say $\gamma_{j,1}(s,u)\neq\gamma_{j,2}(s,u)$. In this case, we choose $X_{2p-j,u}=X_{2p,s}=1$ and $X_{t,v}=0$ for any other vector components in $\{X_{2p-1},\dots,X_p\}$. Then for $j>0$, $U_{\boldsymbol{\theta_1}}((X'_{2p-1},\dots,X'_p),X'_{2p})=-\gamma_{j,1}(s,u)$ and $U_{\boldsymbol{\theta_2}}((X'_{2p-1},\dots,X'_p),X'_{2p})=-\gamma_{j,2}(s,u)$ and, for $j=0$,

$$\begin{aligned}
U_{\boldsymbol{\theta_1}}((X'_{2p-1},\dots,X'_p),X'_{2p}) &= -\alpha_s-\alpha_u-\log(h_s(1)h_u(1))-\gamma_{0,1}(s,u),\\
U_{\boldsymbol{\theta_2}}((X'_{2p-1},\dots,X'_p),X'_{2p}) &= -\alpha_s-\alpha_u-\log(h_s(1)h_u(1))-\gamma_{0,2}(s,u).
\end{aligned}$$

In either case, these must be the same for both $\gamma_{j,1}(s,u)$ and $\gamma_{j,2}(s,u)$, but this is impossible, since both $h_s(1)$ and $h_u(1)$ are greater than zero. Thus, our model is conditionally and unconditionally identifiable.

Proof of Theorem 3.1. From the previous lemma, we know that the log-likelihood function is the sum of terms $P_{\boldsymbol{\theta}}(z,y)$. Each of these terms is a conditionally generalized auto-distribution. Since this is an exponential distribution given any state z, each term satisfies the FI regularity conditions. Thus the log-likelihood function, being a finite sum of these distributions, satisfies the FI regularity conditions and, for each z_0, the set C is the set of all sample paths $\{z_0,x_1,\dots,x_T\}$. It should be noted that if M is not a fixed parameter these conditions cannot be met, since the support for the distribution is not fixed.

The second condition can be proved by showing that the first three derivatives of the log-likelihood function are bounded globally, for fixed T, M and S. First we find the derivatives of the conditional constant $c(\boldsymbol{\theta},z_{t-1})$. By (3.2), we have

$$\begin{aligned}
&\frac{\partial\ln(c(\boldsymbol{\theta},z_{t-1}))}{\partial\boldsymbol{\theta}}\\
&=\frac{\sum_{x_t\in\{0,\dots,M\}^{|s|}}([1,z'_{t-1}]'\otimes w_t)\exp\left(\langle\boldsymbol{\theta},[1,z'_{t-1}]'\otimes w_t\rangle\prod_{s\in S}h_s(x_{t,s})\right)}{c(\boldsymbol{\theta},z_{t-1})}\\
&=\mathbb{E}_{\boldsymbol{\theta}}([1,z'_{t-1}]'\otimes W_t|z_{t-1})=[1,z'_{t-1}]\otimes\mathbb{E}_{\boldsymbol{\theta}}(W_t|z_{t-1}).
\end{aligned}$$

Now in general we have

$$\frac{\partial^n\ln(c(\boldsymbol{\theta},z_{t-1}))}{\partial\boldsymbol{\theta}^n}=[1,z'_{t-1}]^n\otimes\kappa_n(\boldsymbol{\theta},z_{t-1}),$$

where $\kappa_n(\theta,z_{t-1})$ is the cumulant of order n for $W_t|z_{t-1}$ under $\boldsymbol{\theta}$. For $n=2$ or $n=3$ this is the same as the second and third central moment. It should be noted that the cumulant of order n is a symmetric covariate tensor of order n.

Then, by the expression of $\ell_T(\boldsymbol{\theta})$ in (3.1), the first three derivatives of the log-likelihood function with respect to $\boldsymbol{\theta}$ are

$$\nabla\ell_T(\boldsymbol{\theta})=\sum_{t=1}^{T}\left([1,z'_{t-1}]'\otimes w_t-\frac{\partial(\ln(c(\boldsymbol{\theta},z_{t-1})))}{\partial\boldsymbol{\theta}}\right)$$

$$= \sum_{t=1}^{T} \left([1, z'_{t-1}]' \otimes w_t - \mathbb{E}_{\boldsymbol{\theta}}([1, z'_{t-1}]' \otimes W_t | z_{t-1})\right)$$

$$= \sum_{t=1}^{T} \left([1, z'_{t-1}]' \otimes \{w_t - \mathbb{E}_{\boldsymbol{\theta}}(W_t | z_{t-1})\}\right), \tag{A.2}$$

$$\nabla^2 \ell_T(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \left\{([1, z'_{t-1}]'[1, z'_{t-1}]) \otimes \mathrm{Var}_{\boldsymbol{\theta}}(W_t | z_{t-1})\right\}, \tag{A.3}$$

$$\nabla^3 \ell_T(\boldsymbol{\theta}) = -\sum_{t=1}^{T} \Big([1, z'_{t-1}]^3 \otimes \mathbb{E}_{\boldsymbol{\theta}}([W_t - \mathbb{E}_{\boldsymbol{\theta}}(W_t | z_{t-1})]^3 | z_{t-1})\Big). \tag{A.4}$$

The bounds on the derivatives are determined by finding the maximum value of the derivative on the line for each path x:

$$\|\nabla^n \ell_T(\cdot)\|$$
$$= \sup_{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 : \|\boldsymbol{\theta}_1\|_2 = 1} \frac{\mathbf{d}^n \ell_T(\boldsymbol{\theta}_0 + t\boldsymbol{\theta}_1)}{\mathbf{d}t^n} \leq T \left(\sqrt{\frac{(p|S| + 1|)(|S| + 1)|S|}{2}} \cdot M^3\right)^n.$$

Thus the log-likelihood function and its first two derivatives are Lipschitz continuous in $\boldsymbol{\theta}$, with the Lipschitz constant for each derivative being a global constant depending only on T and M. Furthermore, if we are interested in $\ell_T(\boldsymbol{\theta})/T$ the Lipschitz constant depends only on M and $|S|$.

Finally, the log-likelihood function is analytic, that is, $\ell_T(\boldsymbol{\theta})$ has a Taylor expansion in an open ball about each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Since $c(\boldsymbol{\theta}, z_{t-1})$ is a finite sum of exponential functions it is analytic. Now for any value of z_{t-1}, or $\boldsymbol{\theta}$, this sum is strictly positive. Since the $\log(x)$ is an analytic function for $x > 0$, each term in the log-likelihood expression is analytic, using a power expansion. Thus the log-likelihood is analytic about every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Proof of Theorem 3.2. Since the conditional distribution is exponential and identifiable for every z_{t-1}, we have by Lemma 13.2.1 (c) of Winkler (1995) that the conditional covariance matrix $\mathrm{Var}_{\boldsymbol{\theta}}(W_t | z_{t-1})$ is positive definite for all z_{t-1} and $\boldsymbol{\theta}$. Now let

$$\lambda_{\min}(\boldsymbol{\theta}) = \min_{z_{t-1} \in \{0, \ldots, M\}^{p|S|}} \text{eigenvalue}\left(\mathrm{Var}_{\boldsymbol{\theta}}(W_t | z_{t-1})\right), \quad \text{so that } \lambda_{\min}(\boldsymbol{\theta}) > 0.$$

$$\lambda_{\max}(\boldsymbol{\theta}) = \max_{z_{t-1} \in \{0, \ldots, M\}^{p|S|}} \text{eigenvalue}\left(\mathrm{Var}_{\boldsymbol{\theta}}(W_t | z_{t-1})\right),$$

$$\lambda^*_{\min} = \min\ \text{eigenvalue}\ \Big(\sum_{t=1}^{T} [1, z'_{t-1}]'[1, z'_{t-1}]\Big) \geq 0.$$

Since each term $[1, z'_{t-1}]'[1, z_{t-1}]$ is positive semidefinite, we have by Lemma 3.1 that for every $\boldsymbol{a} \in \mathbb{R}^{(p|S|+1)(|S|(|S|+1)/2)}$,

$$\begin{aligned}
\boldsymbol{a}'\nabla^2\ell_T(\boldsymbol{\theta})\boldsymbol{a} &\leq -\lambda_{\min}(\boldsymbol{\theta})\boldsymbol{a}'\Big\{\sum_{t=1}^{T}([1, z'_{t-1}]'[1, z'_{t-1}]) \otimes I\Big\}\boldsymbol{a} \\
&= -\lambda_{\min}(\boldsymbol{\theta})\boldsymbol{a}'\Big\{\Big(\sum_{t=1}^{T}[1, z'_{t-1}]'[1, z'_{t-1}]\Big) \otimes I\Big\}\boldsymbol{a} \qquad \text{(A.5)} \\
&\leq -\lambda^*_{\min}\lambda_{\min}(\boldsymbol{\theta})\|\boldsymbol{a}\|_2^2,
\end{aligned}$$

where I is the identity matrix with dimension $|S|(|S|+1)/2$.

Now let $\boldsymbol{b}$ be the normalized eigenvector associated with $\lambda^*_{\min}$. For any nonzero vector $\boldsymbol{c} \in \mathbb{R}^{(p|S|+1)(\frac{|S|(|S|+1)}{2})}$, let $\boldsymbol{a}_0 = \boldsymbol{b} \otimes \boldsymbol{c}$ so $\|\boldsymbol{a}_0\|_2 = \|\boldsymbol{c}\|_2$. By Lemma 3.1

$$\begin{aligned}
0 \geq \boldsymbol{a}'_0\nabla^2\ell_T(\boldsymbol{\theta})\boldsymbol{a}_0 &\geq -\lambda_{\max}(\boldsymbol{\theta})\boldsymbol{a}'_0\Big\{\Big(\sum_{t=1}^{T}[1, z'_{t-1}]'[1, z'_{t-1}]\Big) \otimes I\Big\}\boldsymbol{a}_0 \\
&= -\lambda_{\max}(\boldsymbol{\theta})\boldsymbol{b}'\Big(\sum_{t=1}^{T}[1, z'_{t-1}]'[1, z'_{t-1}]\Big)\boldsymbol{b} \cdot \boldsymbol{c}'I\boldsymbol{c} \\
&= -\lambda_{\max}(\boldsymbol{\theta})\lambda^*_{\min}\|\boldsymbol{c}\|_2^2 \ .
\end{aligned}$$

If $\lambda^*_{\min} = 0$ there exists a nonzero vector, $\boldsymbol{a}_0$, such that $\boldsymbol{a}'_0\nabla^2\ell_T(\boldsymbol{\theta})\boldsymbol{a}_0 = 0$.

Thus, $\boldsymbol{a}'\nabla^2\ell_T(\boldsymbol{\theta})\boldsymbol{a} < 0$ for all $\boldsymbol{a} \neq \boldsymbol{0}$ if and only if $\lambda^*_{\min} > 0$, which is true if and only if the determinant of $\sum_{t=1}^{T}[1, z'_{t-1}]'[1, z'_{t-1}]$ is greater than 0 or, equivalently, the sample covariance matrix of $\{Z_{t-1} : t \in \{1, \ldots T\}\}$ is positive definite, since

$$\det\left(\mathrm{Var}\ (\{z_{t-1} : t \in \{1, \ldots T\}\})\right) = \det\Big(\frac{\sum_{t=1}^{T}[1, z'_{t-1}]'[1, z'_{t-1}]}{T}\Big).$$

Moreover, $\sum_{t=1}^{T}[1, z'_{t-1}]'[1, z'_{t-1}]$ is positive definite if and only if $[1, z_{t-1}] : t \in \{1, \ldots, T\}$ spans the space $\mathbb{R}^{p|S|+1}$, which is true if and only if $z_t - z_0 : t \in \{1, \ldots, T-1\}$ spans $\mathbb{R}^{p|S|}$.

Proof of Theorem 4.1. We need a set A with $\Pr_{\boldsymbol{\theta}_0}(A) = 1$ such that, for any $x \in A$ and every $\epsilon > 0$, there exists a $T(x, \epsilon)$ where $\hat{\boldsymbol{\theta}}_T \in B(\boldsymbol{\theta}_0, \epsilon)$ whenever $T > T(x, \epsilon)$. We choose A to be the set of $x \in \mathcal{X}$ where $G_T(\boldsymbol{\theta}, x) \overset{\text{a.s.}}{\to} g(\boldsymbol{\theta})$ uniformly for $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r) \cap \boldsymbol{\Theta}$. Since $G_T(\boldsymbol{\theta}, x)$ is a sequence of objective functions, $\Pr_{\boldsymbol{\theta}_0}(A) = 1$. Since $\boldsymbol{\theta}_0$ is in the interior of the open set $\boldsymbol{\Theta}$, we can choose $\epsilon > 0$, with $\epsilon < r$, so that $B(\boldsymbol{\theta}_0, \epsilon) \subset B(\boldsymbol{\theta}_0, r) \cap \boldsymbol{\Theta}$.

As $g(\boldsymbol{\theta})$ is a reference function, there exist both $\gamma > 0$ and $r > 0$ such that, for any $\boldsymbol{\theta}$ on the boundary $\partial B(\boldsymbol{\theta}_0, \epsilon)$ of the ball $B(\boldsymbol{\theta}_0, \epsilon)$, we have $g(\boldsymbol{\theta}) \leq g(\boldsymbol{\theta}_0) - \gamma\epsilon^2$.

Now since $G_T(\boldsymbol{\theta}, x)$ is a sequence of objective functions, $G_T(\boldsymbol{\theta}_0, x)$ converges uniformly to $g(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r)$. Thus, for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, r)$ there exists a $T(\epsilon, x)$ such that $G_T(\boldsymbol{\theta}, x) < g(\boldsymbol{\theta}_0) - \gamma\epsilon^2/2$ and $g(\boldsymbol{\theta}_0) - \gamma\epsilon^2/2 < G_T(\boldsymbol{\theta}_0, x)$ whenever $T > T(\epsilon, x)$.

In particular, the above two inequalities are true for all $\boldsymbol{\theta} \in \partial B(\boldsymbol{\theta}_0, \epsilon)$. This implies that $G_T(\boldsymbol{\theta}, x) < G_T(\boldsymbol{\theta}_0, x)$ whenever $T > T(\epsilon, x)$. This fact can be extended to all elements in $\boldsymbol{\Theta} \setminus B(\boldsymbol{\theta}_0, \epsilon)$ by the concavity of $G_T(\cdot, x)$. Thus we have $\hat{\boldsymbol{\theta}}_T \in B(\boldsymbol{\theta}_0, \epsilon)$ whenever $T > T(\epsilon, x)$.

References

Bartlett, M. S. (1955). *An Introduction to Stochastic Processes.* Cambridge University Press, Cambridge.

Bartlett, M. S. (1967). Inference and stochastic processes. *J. Roy. Statist. Soc. Ser. A* **130**, 457-477.

Bartlett, M. S. (1968). A further note on nearest-neighbor models. *J. Roy. Statist. Soc. Ser. A* **131**, 579-580.

Bartlett, M. S. (1971). Physical nearest-neighbor models and non-linear time series. *J. Appl. Probab.* **8**, 222-232.

Bartlett, M. S. (1972). Physical nearest-neighbor models and non-linear time series II: Further discussion of approximate solutions and exact equations. *J. Appl. Probab.* **9**, 76-86.

Besag, J. E. (1972). Nearest-neighbor systems and the Auto-logistic model for binary data (with Discussion). *J. Roy. Statist. Soc. Ser. B* **36**, 192-236.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36**, 192-225.

Besag, J. E. (1975). Statistical analysis of of non-lattice data. *The Statistican* **24**, 179-195.

Billingsley, P. (1995). *Probability and Measure.*, Third edition. John Wiley & Sons, New York.

Brook, D. (1964). On the distinction between the conditional probability and joint probability approaches in the specification of nearest neighbor systems. *Biometrika* **51**, 481-483.

Cliff, A. D., Hagget, P., Ord, J. K., Bassett, K. A. and Davies, R. B. (1975). *Elements of Spatial Structure: A Quantitative Approach.* Cambridge University Press, New York.

Cressie, N. (1993). *Statistics for Spatial Data.* Revised edition. John Wiley & Sons, New York.

Elsner, J. B., Lehmiller, G. S. and Kimberlain, T. B. (1996). Early August forecasts of Atlantic tropical storm activity for the balance of the 1996 season, using Poisson models. *Experimental Long-lead Forecast Bull.* **5**, 26-28.

Geyer, C. M. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* **56**, 261-274.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman Hall, London.

Gray, W. M. (1984a). Atlantic seasonal hurricane frequency: Part I. El Niño and 30 mb quasi-biennial oscillation influences. *Monthly Weather Rev.* **112**, 1649-1668.

Gray, W. M. (1984b). Atlantic seasonal hurricane frequency: Part II. Forecasting its variability. *Monthly Weather Rev.* **112**, 1669-1683.

Greenwood P. E. and Wefelmeyer W. (1997). Maximum likelihood estimator and Kullback-Leibler information in misspecified Markov Chain models. *Theory Probab. Appl.* **42**, 103-111.

Huffer F. and Wu, H. (1998). Markov Chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics* **54**, 509-524.

Jagger, T. H., Niu, X-F. and Elsner J. B. (2002). A space-time model for seasonal hurricane prediction. *Internat. J. Climatology* **22**, 451-465.

Niu, X-F. (1995). Asymptotic properties of maximum likelihood estimates in a class of space-time models. *J. Multivariate Anal.* **55**, 82-104.

Niu, X-F., McKeague, I. W. and Elsner, J. B. (2003). Improving climate prediction using seasonal space−time models. *Statist. Inference for Stochastic Processes* **6**, 111-133.

Niu, X-F. and Tiao, G. C. (1995). Modeling satellite ozone data. *J. Amer. Statist. Assoc.* **90**, 969-983.

Propp, J. and Wilson, D. (1996). Exact sampling with coupled Markov Chains and application to statistical mechanics. *Random Structures Algorithms* **9**, 223-252.

Schervisch, M. J. (1995). *Theory of Statistics.* Spring-Verlag, New York.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications.* Chapman and Hall, London.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434-449.

Whittle, P. (1963). Stochastic processes in several dimensions. *Bull. Inst. Internat. Statist.* **40**, 974-994.

Winkler G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods.* Springer-Verlag, Berlin.

Wu, H. (1994). Regression models for spatial binary data with application to the distribution of plant species. Ph.D. Thesis, Department of Statistics, Florida State University.

Department of Geography, Florida State University, Tallahassee, FL 32306, U.S.A.

Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, U.S.A.

E-mail: niu@stat.fsu.edu

(Received September 2003; accepted October 2004)