THE UNIVERSITY OF BRITISH COLUMBIA
DEPARTMENT OF STATISTICS

TECHNICAL REPORT #266

Time-varying Markov models for dichotomized temperature series.

Reza Hosseini, Nhu Le, Jim Zidek

January 2012

Time-varying Markov models for dichotomized temperature series.

REZA HOSSEINI, NHU LE, JAMES V. ZIDEK

This paper uses high-order categorical non-stationary Markov chains to study the occurrence of extreme temperature events, in particular frost days. These models can be applied to estimate such things as the probability that a given period is frost-free, the probability that a given day begins a long frost-free period within a year and the distribution of the length of the frost-free period. Several stationary and non-stationary high-order (yet parsimonious) Markov models are proposed and compared using *AIC* and *BIC*. Partial likelihood theory is used to estimate the parameters of these models. We show that optimal (in terms of AIC/BIC) non-stationary Markov models that have constant "Markov coefficients" (across the year) are not adequate to estimate the aforementioned probabilities. Therefore this paper develops Markov models with a time-varying periodic structure across the year. A challenge in fitting these models (by maximizing the partial likelihood) is the large number of parameters. The paper presents a method for overcoming this challenge, one that uses parametric fits to the logit of the nonparametric estimates of the transition probability to initialize the optim function in the R package. Satisfactory results are shown to obtain from this approach. The work is applied to temperature records for the Province of Alberta, Canada.

Key Words: Time-varying Markov Coefficients; Time series; Categorical Markov chain; Partial likelihood; Frost; Minimum temperature.

1 Introduction

This paper applies an rth-order categorical non-stationary Markov chain theory developed in Hosseini et al. (2011b) to find models for extreme temperature events. It extends the models developed in Hosseini et al. (2010) for extreme temperature events to a more general framework which allows the Markov structure vary with time (season). A fundamental premise is that temperature itself, which could in some cases be handled by Gaussian space-time models, is not of specific interest. Instead its dichotomized values play the central role. For example in agroclimate risk analysis and management, the genesis of this paper, any temperature below zero destroys crops in certain periods of the crop growth. In fact, this paper only treats the case of low temperatures but the same techniques can be applied to high temperature above a critical threshold and this issue is considered in Hosseini et al.

Reza Hosseini is a postdoc fellow, Division of biostatistics, University of Southern California, Los Angeles, CA, 90089 (e-mail: reza1317@gmail.com). —— James V. Zidek is Professor Emeritus, Department of Statistics, University of British Columbia, Vancouver, British Columbia, V6T 1Z2 (e-mail: jim@stat.ubc.ca). —— Nhu Le is a Senior Scientist, BC Cancer Agency, Vancouver, British Columbia, V5Z 1L3 (email nle@bccrc.ca).

(2010). Note that by not unnecessarily specifying a full process model we gain some robustness against model misspecification say in contexts where a Gaussian model may not be appropriate.

Although the modeling strategy in this paper is used for binary temperature series, the same approach can be used for other climatological events, and in fact it was used for precipitation in Hosseini et al. (2011a). Likelihoods developed in this way can play a role in setting crop insurance premiums and managing irrigation programs, which are attaining increasing importance as the climate changes. In fact weather derivatives, which may be created as part of a risk management insurance program, can be written in terms of the attainment or non attainment of specific target-values stipulated in the contract. The models developed in this paper can be applied to estimate such things as: the probability that a given period is frost-free; the probability that a given day is the start of a long frost-free period within a year; the distribution of the length of the frost-free period and so on. By dichotomizing the minimum daily temperature process at 5 degrees the same methodology can be used to compute the probability that a given day of the year is the beginning of the growing season (the first day that the mean temperature is higher than 5 degrees for 5 consequent days) as well as the length of the growing season which are important for agricultural applications.

Calculating the probability of events, which are defined using the data over all the year, is not feasible by simply looking at their observed probability over the previous years. There are various reasons for this. Often only a few years of data are available and every year can only be considered as a single data point using such naive procedures. Sometimes only a few missing data points will prevent us from using the rest of the available data during that year. It is unrealistic to assume years are independent and identical observations of the same finite chain of length 365 (or 366) because long-term trends can be present in the temperature process (e.g. due to climate change) and also the end of one year is obviously correlated highly with the beginning of the next. Moreover, a strong seasonality effect may be present in the temperature process. For example the probability that April 1st is a frost day should be very similar to April 2nd. In other words a lot of strength can be borrowed in predicting the status of April 2nd status using the April 1st data.

Throughout this paper, temperature is measured in degrees Celsius. Let us denote the minimum temperature series by $\{mt(t)\}, t = 0,1,2,\cdots$, where t denotes time. We call day t with $mt(t) \leq 0$, a "frost day" and define a binary process $Y(t)$:

$$Y(t) = \begin{cases} 1 & mt(t) \leq 0 \text{ (deg C)} \\ 0 & mt(t) > 0 \text{ (deg C)}. \end{cases}$$

Taking 0 (deg C) to be the cut-off for low temperature seems reasonable in the absence of any other considerations, since that is the usual definition of a frost. In agriculture, where most plants contain a lot of water this can be considered as an important cut-off.

In order to study extreme events (e.g. for mt) three approaches are possible among other methods. First, fit the continuous-valued process using a Gaussian distribution. However in the tails, usually

of paramount concern, the fit does not do well as shown by the qq-plots in Hosseini et al. (2009). Second, use a specified threshold and model the values exceeding the threshold. However with this approach, we cannot answer the question about how often or in what periods of the year the extremes happen. This is because we model only the actual extreme values and ignore the non-extreme values. Moreover we need to pick the threshold high (or low) enough to make the extreme-value theory results approximately hold. This might not be an optimal threshold from a practical point of view. Third, based on practical needs, use a threshold to define a new binary process for [extreme]/[not extreme] realizations and model the binary process. This is the method we use as it does not have the issues mentioned in connection with the first two approaches because the threshold is not taken to satisfy a mere statistical requirement. In fact, we make few assumptions about the binary chain.

We use high-order non-stationary Markov models as a natural framework for modeling the minimum temperature occurrence process. But what form should the model have and what is the order of the chain? We use a representation of these chains, Hosseini et al. (2011b), which is quite general (makes no restricting assumptions about the chain), while being suitable for statistical estimation due to its linear form in the coefficients (Appendix). As it is evident from the representation, in general, the Markov structure (both the intercept and Markov coefficients) can change with time (season) and one need to accommodate that in modeling. This is shown to be an important feature to capture the frost process properties in the data used here. Moreover this modeling framework enables us to come up with parsimonious models by restricting the number of terms (or assuming some are equal) in the linear representation as shown in Hosseini et al. (2011a). Also this framework can easily accommodate other exogenous variables, if needed, by adding terms to the linear representation; a feature we did not need for this work but can be quite useful in many applications.

The problem with increasing the order of a Markov chain is the exponential increase in number of parameters in the model. For modeling precipitation occurrences as a special case, Hosseini et al. (2011a) propose models that increase with the order of the Markov chain by using only 1 extra parameter. They even propose high-order Markov models with only 2 parameters by considering the number of precipitation days in a specified period prior to the date of interest. Several stationary and non-stationary high-order (yet parsimonious) Markov models are proposed and compared using AIC/BIC. Partial likelihood theory is used to estimate the parameters of these models. To elaborate, let $\{Y_t\}, t = 1,2,3,\cdots$ be a categorical Markov chain of order r with conditional probability given by

$$P(Y_t = y_t | Y_{t-1} = y_{t-1}, \cdots, Y_{t-r} = y_{t-r}),$$

which we also denote by $p_{y_{t-r}\cdots y_{t-1}y_t}(t)$. For example $p_{01}(t) = P(Y_t = 1|Y_{t-1} = 0)$ is a 1st-order transition probability curve for a 1st-order Markov chain. The Categorical Expansion Theorem (Appendix A) provides a representation for the conditional probabilities and therefore the chain, in terms of linear combinations of monomials of past processes (e.g. $y_{t-1}, y_{t-2}, y_{t-3}, y_{t-1}y_{t-2}$). Parsimonious models are obtained by restricting the number of terms in the linear representation of the conditional

probability. For the binary precipitation occurrence chain, other parsimonious models are obtained by considering covariates such as $N^7(t-1) = \sum_{i=1}^{7} Y(t-i)$ which counts the number of precipitation days in a week prior.

Other papers that address parsimonious high-order categorical Markov models include Raftery (1985), Bühlmann et al. (1999), Jacobs et al. (1983) and Weiß (2011). Suppose $\{Y_t\}$ is a stationary rth-order Markov chain taking values in $\{1,2,\cdots,m\}$. Then the model in Raftery (1985) is given by

$$P(Y_t = j_0 | Y_{t-1} = j_1, \cdots, Y_{t-r} = j_r) = \sum_{i=1}^{r} \lambda_i q_{j_0 j_i},$$

where $\lambda_1 + \cdots + \lambda_r = 1$ and $Q = \{q_{jk}\}$ is a non-negative $m \times m$ matrix with column sums equal to 1, such that

$$0 \leq \sum_{i=1}^{r} \lambda_i q_{jk_i} \leq 1, \ \ (j, k_1, k_2, \cdots, k_r = 1, \cdots, m).$$

Estimation is done by a constrained maximum likelihood method. In Bühlmann et al. (1999) *variable length markov chain* (VLMC) models are proposed where the portion of the past that influences the next outcome depends on the history of the chain. In Jacobs et al. (1983) the authors propose DARMA and NDARMA processes which mimic the definition of ARMA processes for continuous-valued time series. Weiß (2011) extends these models to the new class of *generalized choice* (GC) models which include NDARMA models as a special case. The main limitations of the aforementioned models are: including exogenous (continuous) covariates is not possible; not much work has been done on extending these models to non-stationary chains with a complex non-stationary nature. The models proposed in this paper overcome both of these limitations.

Another relevant class of models are *Hidden Markov Models*. For example Hughes et al. (1999) describe a hidden Markov model which relates unobserved broad-scale atmospheric circulation patterns to local rainfall. A comparison between the performance of high-order Markov models and hidden Markov models in describing weather patterns such as frost or precipitation occurrence would be desirable in future work but is not considered further here.

Fahrmeir et al. (1987), Kaufmann (1987) and Fokianos et al. (2003) present regression models for non-stationary categorical time series. These models extend generalized linear models for independent data to the case of temporally dependent data. To be more precise, let $\{Y_t\}, t = 1,2,\cdots$ be a binary Markov chain of order r with observations denoted by $\{y_t\}$. Then Fahrmeir et al. (1987) model the conditional probabilities π_t by

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_r y_{t-r} + \gamma_1 x_{t1} + \cdots + \gamma_k x_{tk}), \ t > r,$$

where $\beta_0, \cdots, \beta_r, \gamma_1, \cdots, \gamma_k$ are unknown parameters and the *link function* h is a one-to-one mapping from a subset $D \subset \mathbb{R}$ to (0,1) e.g. the logistic distribution function. Then the observations $\{y_t\}$ form a Markov chain of order r, which is generally non-homogenous since π_t depends on exogenous vari-

ables $x_{1t}, \cdots, x_{tk}$. We call $y_{t-1}, \cdots, y_{t-r}$ *Markov Coefficients*. This paper extends the above model by allowing the Markov coefficients $\beta_1, \cdots, \beta_r$ to vary with time and more specifically with season. An empirical demonstration shows this extension to be necessary to get adequate fits. In fact we show that optimal (using *AIC*/*BIC*) non-stationary Markov models that have constant Markov coefficients (across the year) cannot estimate the aforementioned probabilities and instead propose Markov models with periodic Markov coefficients that vary across the year.

Maximum partial likelihood estimation proves challenging for our proposed models because of the large number of parameters involved. Thus, as a novel feature of the paper, we present a method that uses parametric fits to the logit of the nonparametric estimates of the transition probability to initialize the optim function in the R package. This method, which yields satisfactory results, may well have application in other multi-parameter modeling situations.

We now summarize the paper. Section 3.1 describes the theoretical foundations on which the models of this paper rest. To understand the nature of the binary frost-day stochastic process, we perform an exploratory data analysis in Section 2. That section reveals the complexity of that process and the need for a more refined method that is presented in Section 3.2. Although better than the naive approach used in the exploratory analysis, that non-parametric filtering method is still unable to adequately represent the process. However, it has a vital function in Section 3.3, that of providing initial estimates for an optimization routine for computing model estimates. That method proves successful and yields a model for the stochastic process. Section 3.5 gives ways in which our modeling approach could be extended to chains of higher order given sufficient computing power. Finally Section 4 discusses some applications and extensions of this work.

2 Exploratory analysis

This section presents results from an exploratory analysis of the binary process $Y(t)$ using data records from Medicine Hat, Alberta recorded over the period 1895 to 2006. The transition probabilities are crudely estimated from that historical data, where years are assumed to be independent observations. Similar analyses for other locations are given in Hosseini et al. (2010).

The left panels of Figure 1 and Figure 2 show respectively, the probability of a frost day over the course of a year and the first order transition probability curves for the Medicine Hat station. A regular seasonal pattern is seen, which is not surprising since the temperature changes during a year are caused by smooth cyclic changes of the Earth's location relative to the Sun.

On this basis and to see that pattern more clearly, for the next figures, we use a "cyclical" moving average (i.e. filter) of length 11 to calculate empirical probabilities and transition probabilities (See Section 3.2, Method 1 for more details). By "cyclical" here, we mean that the last days of any one year are run into the first days of the next when computing the moving average. Figures 1 and 2 also show the corresponding filtered curves in the right panels. In particular, Figure 2 shows the estimated

transition probabilities, $\hat{p}_{01}$ and $\hat{p}_{11}$ for the Medicine Hat station. If the chain were a 0th-order Markov chain then these two curves would overlap. [In fact the true transition curves would be identical.] This is not the case and a Markov chain of at least 1st-order seems needed. In Figure 2, $\hat{p_{11}}$ is missing for a period over the summer. This is because no freezing day is observed over this period in the summer and hence $\hat{p}_{11}$ could not be estimated.

Figure 3 depicts the filtered 2nd-order transition probabilities for Medicine Hat. The two curves are separated again indicating that a 2nd-order Markov chain might be appropriate. However the separation is rather small and difficult to detect by model selection procedures. Moreover this separation can only be seen by "cyclical filtering" of the annual data as the original curves are too noisy to allow detection.

Finally, Figure 4 depicts the annually-averaged daily probability of frost in Medicine Hat with the median line added. More precisely, we use the temperature data in a given year and calculate the proportions of frost days for that year. The proportion is fairly constant across the years. In particular no clear monotonic trend is seen here as might be expected under climate change scenarios. On the other hand, such trends can be seen at other locations in Alberta (Hosseini et al. (2010)). Notice some clustering across time in the proportions that tend to stay above or below the median line for a few years, especially in earlier years. In this paper our main focus is capturing the seasonality of the Markov chains and their seasonal evolution. Hence we do not consider long-term effects further.

3 Statistical models

Subsection 3.1 summarizes for completeness the theoretical foundations on which the models of this paper rest. Subsection 3.2 describes a non-parametric method in which the naive non-parametric estimates in Section 2 are improved. Although better than the naive approach used in the exploratory analysis, that non-parametric filtering method is still unable to adequately represent the process. However, it has a vital function in Sections 3.3 and 3.4, that of providing initial estimates for an optimization routine for computing model estimates. That method proves successful and yields a model for the stochastic process. Finally Subsection 3.5 gives ways in which our modeling approach could be extended to chains of higher order given sufficient computing power.

3.1 Theoretical foundations

This subsection presents appropriate models for the binary process $Y(t)$ (or Y_t) of [frost]/[not frost] days. The *Categorical Expansion Theorem* (Appendix) in Hosseini et al. (2011b) gives the form of all such rth-order stationary and non-stationary Markov chains. We give a special case of the theorem in the following example.

Example: For every (possibly non-stationary) binary (0-1) Markov chain of order $r = 3$ and $t \geq 3$

and a fixed transformation $g : \mathbb{R} \to \mathbb{R}^+$, there exists a unique collection of functions $a(\cdot)$ such that:

$$\begin{aligned} g_t &= g^{-1}\{\frac{P(Y_t = 1|Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}{P(Y_t = 0|Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}\} \\ &= a_0(t) + a_1(t)y_{t-1} + a_2(t)y_{t-2} + a_3(t)y_{t-3} \\ &+ a_{12}(t)y_{t-1}y_{t-2} + a_{23}(t)y_{t-2}y_{t-3} + a_{13}(t)y_{t-1}y_{t-3} + a_{123}(t)y_{t-1}y_{t-2}y_{t-3}. \end{aligned}$$

Conversely every collection of arbitrary functions over time $\{a(t)\}$ corresponds to a unique 3rd-order binary (0-1) Markov chain that satisfy the above equations. If we take $g : \mathbb{R} \to \mathbb{R}^+$, $g(x) = \exp(x)$ then $g^{-1}(x) = \log(x)$ in the above. If we also assume the $\{a(t)\}$ are fixed over time, we get a unique representation of stationary 3rd-order binary (0-1) Markov chains.

Most natural processes (such as temperature or precipitation) are non-stationary due to seasonal (within year cyclic variation) and *long-term non-stationarity* due to long-term climate shifts (such as those related to Global warming). This seasonal and long-term non-stationarity can be modeled by letting the parameters change over time.

In order to model the probability of precipitation, Hosseini et al. (2011a) consider models of the form

$$g_t = \alpha_0 + \alpha_1 t + \alpha_2 \cos(\omega t) + \alpha_3 \sin(\omega t) + \beta_1 y_{t-1},$$

where $\omega = 2\pi/366$, as well as extensions to higher order chains. In the above example's notation, this amounts to letting $a_0(t) = \alpha_0 + \alpha_1 t + \alpha_2 \cos(\omega t) + \alpha_3 \sin(\omega t)$. The term $\alpha_1 t$ captures any long-term trend in the probability of precipitation and $\alpha_2 \cos(\omega t) + \alpha_3 sin(\omega t)$ captures the seasonal patterns. Obviously, we can accommodate more complicated long-term non-stationarity effects by adding terms such as t^2, and more seasonal effects by adding terms such as $2\cos(\omega t), \sin(2\omega t)$. In fact Hosseini et al. (2011a) compare many such models and observe that the one that best fits, still misses some of the features observed in transition probabilities of the chains. This suggests a problem with the implicit and restrictive assumption that β_1 or other "Markov coefficients" are considered fixed over time and specially during the year. More precisely, the above model implies that the logit ($logit(x) = \log(x/(1-x))$ of the first order transition probabilities,

$$logit\{P(Y(t) = 1|Y(t-1) = 0)\}, \quad logit\{P(Y(t) = 1|Y(t-1) = 1)\},$$

are parallel curves, by which we mean one is a vertical shift of the other. But there is no reason to believe that rain yesterday has the same effect on the logit of the probability of rain the following day, in winter as in summer. This paper considers models with fixed Markov coefficients, shows their deficiencies and then extends them to Markov chains with varying Markov strength during the year.

An important remaining challenge is the computation time needed to maximize the "partial likelihood" (see below) for Markov models with a large number of parameters. For example for a 1st-order Markov model with 8 seasonal terms to represent the common seasonal factors and 8 seasonal terms

to represent the 1st-order Markov component, maximization of the partial likelihood needs to be done in a space of 16 dimensions. This takes a long time and is unreliable using standard functions such as "optim" in the R package (a free widely used object-oriented statistical package), even if the number of random initial values for the optim function are increased to 50. Instead, we propose a method that uses parametric fits to the logit of the nonparametric estimates of the transition probability to initialize the optim function in the R package and find our approach provides satisfactory results.

To fit our models, we use the *partial likelihood* maximization. By the way of an introduction, we would note that generalized linear models were developed to extend ordinary linear regression to the case that the response is not normal. However, that extension required the assumption of independently observed responses. The notion of partial likelihood was introduced to generalize these ideas to time series where the data are dependent. (See Kedem et al. (2002) for example.) The following definition gives a more precise description.

Definition 3.1. *Let $\mathcal{F}_t$, $t = 1,2,\cdots$ be an increasing sequence of σ-fields, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2, \cdots$ and let $Y_1, Y_2, \cdots$ be a sequence of random variables such that Y_t is $\mathcal{F}_t$-measurable. Denote the density of Y_t, given $\mathcal{F}_t$, by $f_t(y_t;\theta)$, where $\theta \in \mathbb{R}^p$ is a fixed parameter. The partial likelihood (PL) is defined by*

$$PL(\theta; y_1, \cdots, y_N) = \prod_{t=1}^{N} f_t(y_t;\theta).$$

The reader unfamiliar with σ-fields notion can think of $\mathcal{F}_t$ as the information available to us up to time t. As an example, suppose Y_t represents the 0-1 frost day process in Medicine Hat. We can define $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots\}$. In this case, we are assuming the information available to us is the value of the process on each of the previous days. If moreover we assume that Y_t is a fixed-coefficient 2nd-order stationary Markov chain, then $P(Y_t|\mathcal{F}_t) = P(Y_t|Y_{t-1}, Y_{t-2})$. We define $\mathbf{Z_{t-1}} = (1, Y_{t-1}, Y_{t-2}, Y_{t-1}Y_{t-2})$ to be the covariate process in the sense that $P(Y_t = 1|\mathbf{Z_{t-1}}) = logit^{-1}(\theta \mathbf{Z_{t-1}})$, which is a linear form. Other useful covariate processes can be considered. For example $\mathbf{Z_{t-1}} = (1, Y_{t-1}, \cos(\omega t), \sin(\omega t))$ corresponds to a non-stationary 1st-order Markov chain. For any such $\mathbf{Z_{t-1}}$, by definition the log-partial-likelihood is equal to:

$$\sum_{t=1}^{N} \log P(Y_t|\mathbf{Z_{t-1}}) =$$

$$\sum_{1 \le t \le N, Y_t = 1} \log(logit^{-1}(\theta \mathbf{Z_{t-1}})) + \sum_{1 \le t \le N, Y_t = 0} \log(1 - logit^{-1}(\theta \mathbf{Z_{t-1}})).$$

The vector θ that maximizes the above equation is called the maximum partial likelihood (MPLE). Wong (1986) has studied its properties. Its consistency, asymptotic normality and efficiency can be shown under certain regularity conditions.

3.2 Non-parametric estimation

This subsection proposes non-parametric methods to estimate the seasonal transition probabilities. First we present more details for creating the plots in Section 2 and then propose a refined method that result in transition curves with less missing points during the year.

To rigorously present these methods, let $n_{ij}(t),\ i,j \in \{0,1\},\ t = 1,2,\cdots,365$ be the number of pairs of binary event pairs (i,j) over the years that end on day t i.e.

$$n_{ij}(t) = card\{year|(Y(year,t-1),Y(year,t)) = (i,j)\},$$

and

$$n_1(t) = n_{11}(t) + n_{10}(t),\ \ n_0(t) = n_{01}(t) + n_{00}(t).$$

Method 1: In Figure 2 (left panel), the transition probability $p_{ij}(t)$ is estimated by

$$\hat{p}_{ij}(t) = \frac{n_{ij}(t)}{n_i(t)},\ \ n_i > 5.$$

Requiring $n_i(t) > 5$ avoids zero in the denominator while ensuring relative stability in the estimates. Whenever $n_i(t) \leq 5$, p_{ij} was not estimated. The resulting estimates seem noisy and hence a cyclical filter of span 11 was applied i.e. the smoothed estimate is given by

$$filt(\hat{p}_{ij}(t),5) = \sum_{t'=t-5}^{t+5} \hat{p}_{ij}(t')/11,\ \ t = 1,2,\cdots,365.$$

The filter is cyclical in the sense that we let $\hat{p}_{ij}(t') = \hat{p}_{ij}(t'+365)$ for $t' < 0$, and $\hat{p}_{ij}(t') = \hat{p}_{ij}(t'-365)$ for $t > 365$, to ensure that the end of one year is included with the beginning of the next. We thereby reflect the strong annual frost cycle. The filter's window of 11 was chosen so that the changes in the transition probabilities from one day to another would be small (less than 3 %). Here we are assuming that the temperature's annual seasonal pattern (as seen empirically during the period 1895–2006), due to the relative positions of the Sun and Earth, changes smoothly over the year. This assumption seems reasonable on substantive grounds; it is intuitively implausible that for example, from Jan 20th to Jan 21st the transition probability, p_{11} (frost given frost) would change by more than 3 percent.

Empirical application of this seemingly natural approach reveals a serious shortcoming. When $\hat{p}_{ij}(t)$ is missing, not only $filt(\hat{p}_{ij}(t),5)$ will be missing but also $filt(\hat{p}_{ij}(t),5)$ for the neighboring points will be missing as evident in Figure 2, right panel. A different a approach was called for.
Method 2: This method applies a moving *window* with an 11 day span initially and only then estimates the transition probability using all the days. Again we require $n_i > 5$ to avoid missing estimates

and stability. More precisely,

$$n_{ij}(t) = card\{year|(Y(year,t'-1),Y(year,t')) = (i,j),\ t' \in [t-5,t+5]\},$$
$$n_1(t) = n_{11}(t)+n_{10}(t),\ \ n_0(t) = n_{01}(t)+n_{00}(t),\ \ \hat{p}_{ij}(t) = \frac{n_{ij}(t)}{n_i(t)},\ \ n_i > 5. \tag{1}$$

Note that in this method the definition of $n_{ij}(t)$ is altered to include $t' \in [t-5,t+5]$ (as opposed to only t). Figure 5 shows the result, a more complete curve than that depicted in Figure 2.

Figure 6 applies the techniques explained above to get non-parametric estimates of the 2nd-order transition probability. This time, having less data than before, we take the moving window to be of length 15 and require $n_{ij} > 5$, as before, thereby obtaining estimates of $\{p_{ijk}\}$.

One could use these non-parametric estimates as an estimate of the frost day process. However the transition probabilities are not completely defined over the whole year; a gap appears mid-year for the $\{p_{11}(t)\}$. This anomaly is unimportant in practice since the chain is extremely unlikely to enter its "1" state during that period. But for completeness we can avoid it by reconstructing the mid-part of the chain. The reconstruction can either be done by assigning a small non-negative probability value or by using another filtering method, which we call the "NA filling" filtering. The latter approach works as follows. For any point with a missing estimate, consider its two neighboring days. If only one (respectively both) of them has (have) a value, enter that value (their average value) for the missing one. Applying this filter repetitively often enough will fill all the gaps. The number of such repetitions is half of the size of the biggest gap. The result is shown in Figure 7.

There we see a jump in mid-year for the $\hat{p}_{11}$ curve, an anticipated although unnatural feature. But the reconstruction does what it is supposed to do, that is ensure the chain will not terminate in mid-year. Note that the chain has a small chance of entering the top curve in the first few days where it was missing originally, but a negligible chance to be on the top curve where the jump occurs (because it is very unlikely to have a frost during that period).

In order to compare this non-parametric estimate to other estimates, we can calculate the log partial-likelihood (*LPL*) of this chain given the data. The log partial-likelihood turns out to be $LPL = 9418.9$ and hence $-2LPL = 18837.7$. Note that with 2×366 parameters and the same partial likelihood value, we get $BIC = -2LPL + (2 \times 366)\log(n) = 26601.0$. This shows that if we maximize the likelihood with a vector of 2×366 parameters (one for each $p_{i1}(t)$) then the *BIC* would be bounded by 26601.0.

If we use fixed parameters $p_{11}(t)$ and $p_{01}(t)$ for every week, we will need 2×52 parameters since every year has roughly 52 weeks. In order to get an upper bound, we need to fit a model with $2 \times 52 = 104$ parameters and calculate the likelihood. To do so, for every week we compute the average of the transition probabilities and replace the transition probability for the whole week by that average. The results are given in Figure 8 (Left panel). Then for this set of parameters $-LPL = 9430.0$ and hence $BIC = -2LPL + 2 \times 52\log(n) = 19962.9$ while *AIC*=19067.9, giving upper bounds for a

BIC/AIC of a model with 104 parameters, 52 for each of the transition probabilities. This model is computationally challenging to fit due to high dimension of the parameter space. Therefore we use the initial value obtained by the non-parametric estimate and get $BIC = 19936, AIC = 19041$, which is a slight improvement on the upper bound obtained above. The fits are given in Figure 8 (Right panel). The fit differs only slightly from the initial non-parametric values.

We showed above that non-parametric estimates can give satisfactory results in terms of fits. However the application of this method has some drawbacks. For example substantial amounts of data are needed to get reliable estimates and extensions to include other possibly continuous covariates or spatial-temporal models are not feasible.

3.3 Model with fixed coefficients

Building on the background acquired in previous sections, this subsection finds models for the extreme minimum temperature process $Y(t)$ using non-stationary high-order Markov chains. We start with fixed Markov coefficients and show their deficiencies. That leads to models with time-varying Markov coefficients and we fit those models by initializing the optim function at initial values obtained using non-parametric methods discussed in the previous subsection. We investigate the following predictors for varying integers k:

- $Y^k(t) \equiv Y(t-k)$, whether or not it was a frost day k days ago.
- N^k, the number of freezing days during the k previous days.
- sin, cos, $sin2$ and $cos2, \cdots$ which are abbreviations for $\sin(\omega t)$, $\cos(\omega t)$, $\sin(2\omega t)$ and $\cos(2\omega t), \cdots$ respectively with $\omega = \frac{2\pi}{366}$.

First we compare all models with an intercept, represented as the predictor that is identically "1", and any subset of the other predictors:

$$Y^1, Y^2, Y^1Y^2, N^5, N^{10}, cos, sin, cos2, sin2, cos3, sin3.$$

The number of possible models is then $2^{11} = 2048$. To cut down the computation time we initially focus on Medicine Hat's minimum temperatures during the period 1995–2000.

The result is that both the BIC and AIC criteria include Y^1, cos and sin in their best five models. That led us to include those covariates in all models subsequently considered. Those other models involve some combination of the remaining predictors, $Y^2, Y^1Y^2, N^5, N^{10}, cos2, sin2, cos3, sin3$. The analysis then used all data from the period 1895–2006 and all $2^8 = 256$ models.

Table 1 presents the top five models according to both the BIC and AIC criteria. The rankings by these criteria are consistent except that each inverts the ranking of the other's 3rd and 4th models. The best model based on both the BIC and AIC criteria involves the predictor vector $(1, Y^1, cos, sin, cos2, N^5)$,

a parsimonious 5th-order non-stationary chain with 3 seasonal terms. The second best has predictor vector $(1, Y^1, cos, sin, Y^2, Y^1Y^2, cos2, N^5)$, which is also a 5th-order chain with two more covariates Y^2 and Y^1Y^2. However the coefficient for Y^2 equals 0.04, which is rather small. *BIC*'s third best model is a 10th-order parsimonious Markov chain with three seasonal terms. The last two models by *BIC* are 2nd-order Markov models.

The gaps between the *AIC*/*BIC* values change the most between the first model and rest. Note that the predictor vector in the best model by both criteria is $(1, Y^1, cos, sin, cos2, N^5)$ whereas *BIC*'s third best model uses

$$(1, Y^1, cos, sin, cos2, N^{10}).$$

The factors lead to the concern that replacing N^5 by any one of the powers might improve *AIC* and *BIC* performance, although to cut computation times, we had not included any of $N^2, N^3, N^4, N^6, N^7, N^8, N^9$ amongst the predictors. To test this theory, we fitted those extra models and found that N^5 was indeed best amongst all these predictors.

Next we assessed the best model chosen by *BIC* and *AIC* by inspecting the conditional probability fits. Figure 9 shows the 1st-order conditional probability fit for the model $(1, Y^1, cos, sin, cos2, N^5)$. The estimated first order probabilities look satisfactory at first glance. However an issue of concern arises on closer inspection, namely the probability of frost given no frost in the summer is over-estimated during the summer. In fact if we compare the distribution of the fitted *number of frost days in summer* with the observed distribution using Kolmogorov distance or any other distance measure, a large difference is seen. The same anomaly is observed by inspecting other top models in Table 1. These fits are thus of small practical value, since the risk of frost just before, during, or just after the growing season are of great concern to farmers, managers, insurance companies and government agencies. One might think that adding more seasonality terms (i.e. more Fourier series terms) would solve this problem. However even when we considered fits that include higher frequencies $(\cos(4\omega t), \sin(4\omega t), \cdots)$, both the *AIC* and *BIC* criteria suggested even less favorable results and the fits did not improve significantly. We are thus led to seek the source of this kind of difficulty, and in particular, to see if it is due to some implicit assumption we have made.

3.4 Markov Models with time-varying coefficients

We begin with an investigation of the limitations of the models used above with fixed Markov components by fitting them to the non-parametric estimates of the transition probabilities p_{01} and p_{11}. By doing this instead of fitting the actual binary data, we dramatically reduce the computation time and can handle models with many parameters. Nevertheless, with more seasonal parameters both transition curves could not be fit simultaneously leading us to include an auxiliary binary variable that is "1" for the $logit(p_{11}(t))$ curve and is "0" for the $logit(p_{01}(t))$ curve, $t = 1, \cdots, 365$. Then we used a Fourier series of up to order 8 i.e. $\sin(\omega t), cos(\omega t), \cdots, \sin(8\omega t), \cos(8\omega t)$, which is a large enough

number of parameters to estimate the rather simple form of these curves. The fits to the logit scale are given in Figure 10 which shows the fits are not satisfactory, despite the large number of parameters . Note that even if we increase the number of the parameters, we can never fit or come close to fitting these data perfectly. That failure is because of the assumption made implicitly that the logit transition curves are parallel.

The previous analysis suggests that to make progress, we need to relax the assumption of fixed coefficients and turn to Markov chains with time-varying Markov coefficients. We first focus on the 1st-order chains and move on to higher orders later. Thus assume

$$logit\{P(Y(t)=1|Y(t-1)=y_{t-1},\cdots)=a_0(t)+a_1(t)y_{t-1},$$

where $a_0(t)$ and $a_1(t)$ are periodic functions with period 366, modeled with Fourier series. The challenge here is the large number of parameters that might potentially be needed to successfully capture the seasonal trend. Globally maximizing the partial likelihood with so many parameters is very difficult using procedures such as the optim function in R, our preferred choice. Comparing several models using *AIC*/*BIC* in this fashion is even more challenging. Initially attempts involved picking several random but reasonable initial values for the MPLEs (maximum partial likelihood estimates). But even 50 random initial values selected uniformly over a sensible neighborhood of the 0 vector, left great uncertainty about the global optimum in models with more than 16 parameters. Moreover the computations required several hours and in some cases days to fit a single model on computer with a 2.2 GH processor and 8 of GB RAM. However, that failure led us to a successful alternative.

For simplicity, we start with models for which $a_0(t), a_1(t)$ have the same number of terms involving $\sin(\omega t), \cos(\omega t), \ldots, \sin(n\omega t), \cos(n\omega t)$ and possibly different coefficients. We fit these 1st-order time-varying Markov models first by simply using 3 random initial values and optim. The results are seen in the second column of Table 2. Mathematical considerations suggest the negative likelihood (*-LPL*) should decrease with an increasing number of the parameters in the model. Instead we see that it has increased. That anomaly is due to the MPLEs for larger models being far from the true global maxima. To overcome this problem, we use the logit of the non-parametric 1st-order transition curves, fit parametric models that correspond to our target models, and then use them to initialize the optim function.

To fix ideas let

$$logit\{P(Y(t)=1|Y(t-1)=y_{t-1},\cdots)\}=a_0(t)+a_1(t)y_{t-1},$$

with

$$a_i(t)=\alpha_i+\alpha_i^1\cos(\omega t)+\beta_i^1\sin(\omega t)+\alpha_i^2\cos(2\omega t)+\beta_i^2\sin(2\omega t),\ i=0,1.$$

Then take the logit of non-parametric estimates of p_{01} and p_{11} obtained above, as a response variable

$$u_{11} = logit(p_{11} + 10^{-10}),\ \ u_{01} = logit(p_{01} + 10^{-10}),$$

where the 10^{-10} has been added to ensure the logit is well-defined. [Note that a probability of 0 and 10^{-10} are of no practical importance. Moreover $logit(10^{-10}) = -23$ is a value that neither overflows the processor's capacity nor critically changes the theoretical results in partial likelihood theory.] We define an auxiliary variable called *aux* which is equal to 0 or 1 according as we are on the curve u_{01} or u_{11}. We also include Fourier terms $\cos(\omega t), \sin(\omega t), \cos(2\omega t), \sin(2\omega t)$ and their interactions with *aux*,

$$(aux)\cos(\omega t), (aux)\sin(\omega t), (aux)\cos(2\omega t), (aux)\sin(2\omega t).$$

Then standard generalized linear models (GLM) for continuous data can be used to fit curves to the response u using the variables $1, aux, \cos(\omega t), \cdots, (aux)\sin(2\omega t)$. We use the resulting estimated parameter vector to initialize the corresponding partial likelihood model with covariates

$$1, Y^1, cos, sin, cos2, sin2, Y^1, Y^1 cos, Y^1 sin, Y^1 cos2, Y^1 sin2.$$

Table 2 shows the results of fitting 1st-order time-varying Markov models with Fourier terms up to $\cos(n\omega t), \sin(n\omega t)$ for both $a_0(t)$ and $a_1(t)$ and $n = 2,3,4,5,6,7$. The second column shows the values of the negative partial likelihood obtained from the non-parametric fits along with an upper bound for *AIC* and *BIC*. The fourth column shows the results of applying optim to that initial value. We observe a major improvement in the computed negative likelihood and consequently in *BIC* and *AIC*. The best model according to *BIC* is the one with $n = 5$ and according to *AIC* is the one with $n = 6$. Figure 11 shows the corresponding fits of these models for $n = 6$. The left-hand panels in the Figure shows the fit to the non-parametric logit curves to initialize the optim function and the righthand panels show the final fits using the initial value picked in this fashion. Tracking the non-parametric fits graphically is quite useful because we can check if the initial values themselves fit the logit scale well. It should not be expected that a poor fit on the logit scale would serve as a satisfactory initial value. Instead it would suggest the need for more Fourier terms.

To fit higher order Markov chains in a simple manner using the above techniques, we simply add 0s to the initial value to correspond to higher order terms. For example, we can add the covariate N^5 by extending the covariate process and adding one zero to the initial value vector obtained above, to correspond to the new covariate N^5. Also we can add two higher order Markov terms Y^2, Y^1Y^2 by adding two zeros to the initial value vector obtained above. The results are given in Table 3 for the models above with $n = 5,6$, which are optimal using *BIC* and *AIC* respectively. Values for the optimal models are in boldface. Comparing these results with Table 1, we observe that while *AIC* is improved dramatically, *BIC* is still beaten by overly simple models that do not capture the transition curves. Since the fits do not show any trace of over-fitting, this result seems to indicate that *AIC* is

a better measure for assessing complex Markov models. However here we did not fit all possible combinations of the models due to the high number of comparisons to be made. If we would wish to fit models that include $1, Y^1$ and Fourier terms of up to $n = 7$ for example, we would need to fit 2^{28} models, each taking about 7 or 8 minutes according to Table 2. This computation would take approximately 4000 years!

To compare many combinations then, we again propose a method based on the fits to the non-parametric estimates. We compare the fits of all the combinations of covariates

$$sin, cos, \cdots, sin6, cos6, aux, (aux)cos, (aux)sin, \cdots, (aux)cos6, (aux)sin6,$$

with the logit of the non-parametric transition curves as the response. Note that fitting such models is very fast in R (as opposed to maximizing the partial likelihood). R also provides the *AIC* for each of these fits and we use that approach to pick the model with smallest *AIC*. That model turns out to have covariates:

$$\begin{aligned} sin, cos, sin2, cos2, sin3, cos3, sin4, cos4, sin5, sin6, cos6, \\ aux, (aux)cos, (aux)sin2, (aux)cos2, (aux)sin6. \end{aligned}$$

Fitting the corresponding partial likelihood model to the original data, we get

$$(-LPL, BIC, AIC) = (9449, 19079, 18933),$$

which shows an improvement in terms of *AIC* and *BIC* compared to 1st-order Markov models in Table 2. Once again we can add higher order Markov terms N^5, Y^2 and (Y^2, Y^1Y^2) to get

$$\begin{aligned} \text{add } N^5 \quad &\rightarrow \quad (-LPL, BIC, AIC) = (9407, \mathbf{19006}, 18851), \\ \text{add } Y^2 \quad &\rightarrow \quad (-LPL, BIC, AIC) = (9422, 19035, 18880), \\ \text{add } (Y^2, Y^1Y^2) \quad &\rightarrow \quad (-LPL, BIC, AIC) = (9412, 19025, 18862). \end{aligned}$$

Therefore the model with covariates

$$1, cos, sin, \cdots, cos6, sin6, Y^1, Y^1cos, Y^1sin2, Y^1cos2, Y^1sin6,$$

has the smallest *BIC* among all compared models and is close to *AIC* for the model in the last row of Table 3. We repeated model selection with *BIC* and we got $(-LPL, BIC, AIC) = (9478, 19093, 18982)$ for the model with covariates

$$aux, sin, cos, sin2, cos2, sin3, cos3, sin4, cos4, sin5, sin6, (aux)cos2,$$

which is inferior to the non-parametric fit that *AIC* picked above with

$$(-LPL, BIC, AIC) = (9449, 19079, 18933).$$

The fits to the 1st-order and 2nd-order transition probability for the overall optimal model picked by *AIC* in Table 3 with seasonality terms up to $sin6$ and $cos6$ for both fixed and Markov components and 2nd-order covariates

$$Y^2, Y^1Y^2,$$

are given in Figures 12 and 13. The fits to the 1st-order transition probabilities are quite satisfactory. The fits to the 2nd-order transition probabilities do not fully match the estimated non-parametric companions and even though they do show separation of the pair of transition curves, the separation is not as pronounced as the non-parametric estimates. However note that here we are not fitting curves to observations and rather compare them to non-parametric estimates which are noisy due to less data in comparison to the 1st-order chains. To investigate the problem more we extended this model by adding the covariates $Y^2cos, Y^2sin, Y^1Y^2cos, Y^1Y^1sin$, thus creating a time-varying 2nd-order chain. We obtained *AIC*=18860 , *BIC*=19136 and the fits did not show much improvement. The fits to the 1st-order and 2nd-order transition probabilities for the overall optimal model picked by *BIC* with covariates

$$1, cos, sin, \cdots, cos6, sin6, Y^1, Y^1cos, Y^1sin2, Y^1cos2, Y^1sin6, N^5$$

were similar to the optimal model picked by *AIC* but slightly inferior in tracking the non-parametric curves in both 1st and 2nd-order transition probabilities and we do not show them here for brevity.

In summary in this section we have used AIC/BIC to guide us to appropriate models and we then check their fits to pick the final model. Below is a summary of the search we propose in finding appropriate models.

1. Fit the non-parametric estimates of 1st-order transition probabilities with different n for the Fourier series expansion and identify ns for which the non-parametric fits look satisfactory. Use the initial values of the satisfactory fits (chosen n) to fit the chain using partial likelihood. Also extend the models to higher orders by adding N^5, Y^2, Y^1Y^2 to the covariate and inspect AIC/BIC as well as their fits.

2. Use AIC/BIC to perform model selection relative to the non-parametric logit 1st-order transition probabilities and choose the optimal models. Try these optimal fits (by calculating the partial likelihood, AIC/BIC) as well as their extensions by adding N^5, Y^2, Y^1Y^2 to the covariate and inspect AIC/BIC along with their fits

3. Compare the best fits from 1,2 and choose the final model.

As discussed above, in Step 2, *AIC* seems to be more appropriate for the model selection part (in this

data). Also the final model with most satisfactory fit to the non-parametric estimates in this data came from Step 1, in Figure 13. However the smallest overall *BIC* was detected in Step 2.

3.5 Extensions to higher orders

Note that the techniques discussed for 1st-order time-varying Markov chain can be extended to higher orders. For simplicity, we only discuss how these techniques can be extended to 2nd-order chains. By the *Categorical Expansion Theorem for Markov Chains* (Appendix), every 2nd-order binary Markov chain (with strictly positive joint distributions for any finite collection of times) can be represented by an initial joint probability $P(Y_0, Y_1)$ and the logit conditionals

$$logit\{P(Y(t) = 1|Y(t-1), Y(t-2))\} =$$

$$a_0(t) + a_1(t)Y(t-1) + a_2(t)Y(t-2) + a_{1,2}(t)Y(t-1)Y(t-2),$$

for unique functions $a_0, a_1, a_2, a_{1,2}$. Conversely any such collection of functions corresponds to a unique chain (up to distributional equivalence). In the previous section, we discussed several Markov models with time-varying 1st-order coefficients. We can extend such models to higher orders by assuming

$$a_j(t) = \alpha_0^j + \sum_{i=1}^{n} [\alpha_i^j \cos(i\omega t) + \beta_i^j \sin(i\omega t)].$$

The computational challenge is even more acute for these higher order models. However non-parametric methods like those in above can be applied. For example, suppose we are interested in fitting the following 2nd-order Markov chain:

$$logit\{P(Y(t) = 1|Y(t-1), Y(t-2))\} = \alpha_0 + \alpha_0^1 \cos(\omega t) + \beta_0^1 \sin(\omega t) +$$

$$\alpha_1 Y(t-1) + \alpha_1^1 Y(t-1)\cos(\omega t) + \alpha_2 Y(t-2) + \beta_2^1 Y(t-2)\sin(\omega t) +$$

$$\alpha_{1,2} Y(t-1)Y(t-2) + \alpha_{1,2}^1 Y(t-1)Y(t-2)\cos(\omega t).$$

Also suppose we have non-parametric estimates available for $p_{ij1},\ i, j \in \{0, 1\}$. Then we introduce two auxiliary variables $aux1, aux2$ corresponding to $Y(t-1)$ and $Y(t-2)$ and also consider the interaction term $(aux1)(aux2)$ to correspond to $Y(t-1)Y(t-2)$. For points on the curve p_{001}, we let $aux1 = aux2 = 0$; for p_{011}, $aux1 = 1, aux2 = 0$; for p_{101}, $aux1 = 0, aux2 = 1$; and finally for p_{111}, we let $aux1 = aux2 = 1$. Then we fit a normal linear model using all the data pooled and covariates

$$1, \cos(\omega t), \sin(\omega t), aux1, (aux1)\cos(\omega t),$$

$$aux2, (aux2)\sin(\omega t), (aux1)(aux2), (aux1)(aux2)\cos(\omega t)$$

and use the obtained vector to initialize optim. It should be clear now how these can be extended to even higher Markov chains.

4 Discussion and concluding remarks

In order to model frost day occurrences, several non-stationary, high-order Markov chains are considered and compared. These models can capture the strong seasonality of these processes well and also accommodate long-term trends and dependence in the chain as discussed in Section 3.1. It turns out that to capture the evolution of the process, one needs to consider time-varying Markov coefficients. To overcome the high computational costs, we suggested the use of non-parametric fits to the conditional probabilities to guide our parametric estimation. This idea was also used in Model selection. The final model we propose based on this analysis includes covariates

$$\begin{aligned} &1, \sin(\omega t), \cos(\omega t), \cdots, \sin(6\omega t), \cos(6\omega t), \\ &Y_{t-1}, Y_{t-1}\sin(\omega t), Y_{t-1}\cos(\omega t), \cdots, Y_{t-1}\sin(6\omega t), Y_{t-1}\cos(6\omega t), Y_{t-2}, Y_{t-1}Y_{t-2}. \end{aligned}$$

It had the smallest *AIC* among all models we compared and also revealed satisfactory fits to the transition probabilities (Figures 12 and 13).

Another possible method to overcome the computational burden is to use blocks of data (say every 5 years) to fit a model several times and initialize optim at all the obtained values. We used this idea to reduce the model selection time at the beginning of Section 3.3. However, the reduction in computations in this case is far less significant than using the non-parametric methods.

Hosseini et al. (2010) use similar models to provide confidence intervals for events such as: π : *The probability of having at least 5 days without frost in the first week of October.* The confidence intervals were obtained once using the *partial information matrix* of the parameters and once using a bootstrap method. The results were quite similar. The confidence interval was (0.75, 0.85) using the partial information matrix and (0.74, 0.85) using the bootstrap method.

Note that the formulation in this paper includes Markov chains with changing orders over time as a special case. For example here we introduce a Markov chain that changes its order from 0 to 1 at time t_c:

$$logit\{P(Y(t)|Y(t-1))\} = a_0(t) + a_1(t)Y(t-1),$$

where $a_1(t) = 0$, $t < t_c$ and $a_1(t) = k \neq 0$, $t \geq t_c$. The parameters of this chain also can obviously be estimated using partial likelihood and this can be extended to higher orders. There is little to believe such jumps are useful in weather processes and so we did not use them here.

Finally a comparison with continuous-valued (Gaussian) models of minimum temperature would be desirable in future work and we believe a complex continuous-valued model is needed. The techniques developed here can also be used in the continuous setting for example by considering models

such as $Y(t) = a_0(t) + a_1(t)Y(t-1) + a_2(t)Y(t-2) + \varepsilon(t)$, where $\varepsilon(t) \sim N(0, \sigma(t)^2)$, are independent errors. Seasonal/long-term structures can be considered for functions $a_1(t), a_2(t), \sigma(t)$. This is an extension of the well-studied *Autoregressive models* (AR). Similar extensions can be considered for ARIMA processes.

Acknowledgement. We are indebted to L. A. Vincent from Environment Canada for providing the data on which the application in this paper is based.

5 Appendix

Here we state the *Categorical Expansion Theorem* for categorical Markov chains, which characterizes all discrete-time categorical Markov chains of a given order. A similar result also holds for arbitrary discrete-time processes and the proof of both can be found in Hosseini et al. (2011b).

Theorem 5.1. *(Categorical Expansion Theorem for Markov Chains) Suppose that* $\{Y_t\}, t = 0, 1, 2, \cdots$ *is an rth-order Markov chain where* Y_t *takes values in* M_t*, a finite subset of real numbers,* $|M_t| = c_t = d_t + 1 < \infty$*, the conditional probabilities*

$$P(Y_t = y_t | Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0),\ t = 1, 2, \cdots$$

are well-defined and belong to $(0, 1)$*. Fix* $m_t^1 \in M_t$*, let* $M_t' = M_t - \{m_t^1\}$ *and suppose* $g : \mathbb{R} \to \mathbb{R}^+$ *is a given bijective transformation. Then*

$$g_t(y_t, \cdots, y_0) = g^{-1}\{\frac{P(Y_t = y_t | Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}{P(Y_t = m_t^1 | Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}\},$$

is a function of $t+1$ *variables for* $t < r$*,* $(y_t, \cdots, y_0)$ *and is a function of* $r+1$ *variables,*$(y_t, \cdots, y_{t-r})$*, for* $t > r$*. Moreover there exist parameters*

$$\{\alpha^t_{i_0, \cdots, i_t}\}_{0 \le i_0 \le d_t - 1, 0 \le i_1 \le d_{t-1}, \cdots, 0 \le i_t \le d_0},\ \text{for } t < r,$$

and

$$\{\alpha^t_{i_0, \cdots, i_r}\}_{0 \le i_0 \le d_t - 1, 0 \le i_1 \le d_{t-1}, \cdots, 0 \le i_r \le d_{t-r}},\ \text{for } t \ge r,$$

such that for $t < r$*:*

$$g^{-1}\{\frac{P(Y_t = y_t | Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}{P(Y_t = m_t^1 | Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}\} = \sum_{0 \le i_0 \le d_t - 1, 0 \le i_1 \le d_{t-1}, \cdots, 0 \le i_t \le d_0} \alpha^t_{i_0, \cdots, i_t} y_{t-0}^{i_0} \cdots y_{t-t}^{i_t},$$

$$(y_0, \cdots, y_t) \in M_0 \times \cdots M_{t-1} \times M_t',$$

and for $t \geq r$:

$$g^{-1}\{\frac{P(Y_t = y_t|Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}{P(Y_t = m_t^1|Y_{t-1} = y_{t-1}, \cdots, Y_0 = y_0)}\} = \sum_{0 \leq i_0 \leq d_t - 1, 0 \leq i_1 \leq d_{t-1}, \cdots, 0 \leq i_r \leq d_{t-r}} \alpha^t_{i_0, \cdots, i_r} y_{t-0}^{i_0} \cdots y_{t-r}^{i_r}$$

$$(y_0, \cdots, y_t) \in M_0 \times \cdots M_{t-1} \times M_t'.$$

Moreover any collection of arbitrary parameters

$$\{\alpha^t_{i_0, \cdots, i_t}\}_{0 \leq i_0 \leq d_t - 1, 0 \leq i_1 \leq d_{t-1}, \cdots, 0 \leq i_t \leq d_0}, \text{ for } t < r,$$

and

$$\{\alpha^t_{i_0, \cdots, i_r}\}_{0 \leq i_0 \leq d_t - 1, 0 \leq i_1 \leq d_{t-1}, \cdots, 0 \leq i_r \leq d_{t-r}}, \text{ for } t \geq r,$$

specify a unique rth-order Markov chain (upto distribution) by the above relations.
In the case of homogenous Markov chains, the $\alpha^t_{i_1, \cdots, i_r}$ do not depend on t for $t > r$.

Remark. The binary case is a special case for which the powers of $y_{t-1}, \cdots, y_{t-r}$ in the representation are at most 1. This means while interaction terms such as $y_{t-1}y_{t-2}y_{t-3}$ appear in the representation, terms such as $y_{t-2}^2 y_{t-3}$ do not appear.

References

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19:716–723, 1974.

Biswas, A. and Song, P. X.-K. Discrete-valued ARMA processes. *Statistics & Probability Letters*, 79(17):1884–1889, 2009.

Bühlmann, P., Wyner, A.J. Variable length Markov chains. *Annals of Statistics*, 27: 480–513, 1999.

Fahrmeir, Ludwig and Kaufmann, Heinz Regression models for non-stationary categorical time series. *Journal of Time Series Analysis*, 8(2):147–160, 1987.

Fokianos, K. and Kedem, B. Regression theory for categorical time series. *Statistical Science*, 18(3):357–376, 2003.

Hosseini, R., Le, N. and Zidek, J. An Analysis of Alberta's climate. Part II: Homogenized data. TR #246, Department of Statistics, UBC, 2009

Hosseini, R., Le, N. and Zidek, J. Model selection for the binary dichotomized temperature processes. TR #257, Department of Statistics, UBC, 2010

Hosseini, R., Le, N. and Zidek, J. Selecting a binary Markov model for a precipitation process. *Environmental and Ecological Statistics*, 18(4):795–820, 2011a.

Hosseini, R., Le, N. and Zidek, J. A characterization of categorical Markov chains. *Journal of Statistical Theory and Practice*, 5(2):261–284, 2011b.

Hughes, J. P. and Guttorp, P and Charles, S. P. A non-homogeneous hidden Markov model for precipitation occurrence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(1):15–30, 1999.

Jacobs, P.A., Lewis, P.A.W. Stationary discrete autoregressive-moving average time series generated by mixtures. *Journal of Time Series Analysis*, 4(1):19–36, 1983.

Heinz Kaufmann Regression Models for Nonstationary Categorical Time Series: Asymptotic Estimation Theory. *The Annals of Statistics*, 15(1):79–98, 1987.

Kedem, B. and Fokianos, K. *Regression Models for Time Series Analysis*, Wiley Series in Probability and Statistics, 2002.

Raftery, A.E. A model for high-order Markov chains. *Journal Royal Statistical Society Series B*, 47(3):528–539, 1985.

Schwartz, G. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

Weiß, C.H. Generalized choice models for categorical time series. *Journal of Statistical Planning and Inference*, 141(8):2849–2862, 2011.

Wong, W. Theory of partial likelihood. *Annals of Statistics*, 14(1):88–123, 1986.

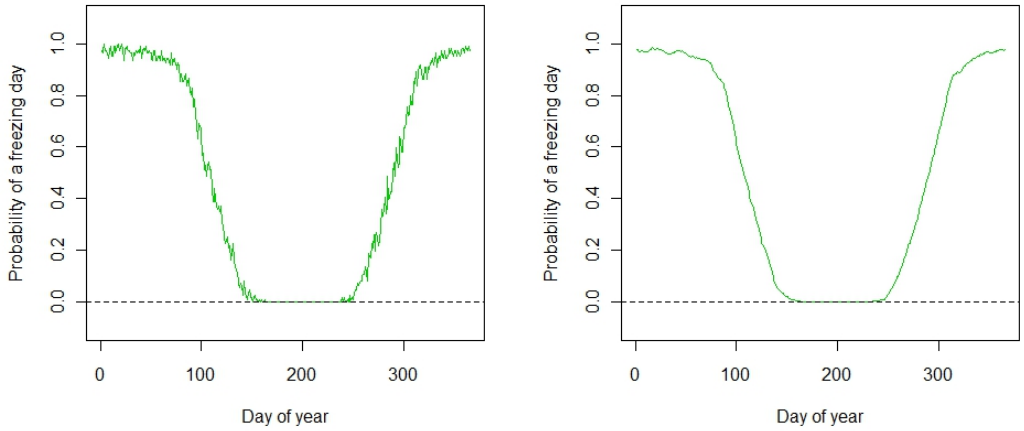

Figure 1: **(Left)** The estimated probability of a freezing day for the Medicine Hat site for different days of a year computed using the historical data. **(Right)** is a smoothed version of the right curve using a moving average filter of length 11.

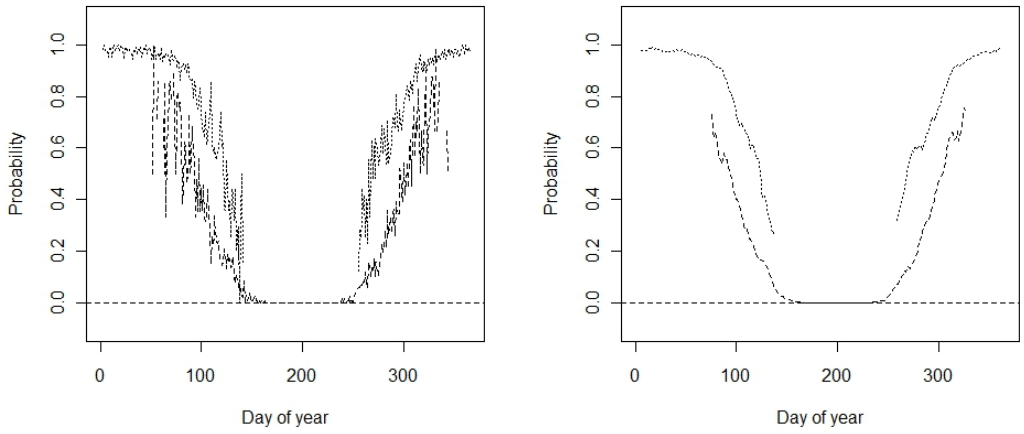

Figure 2: **(Left)** The estimated 1st-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site. The dotted line represents the estimated probability of "$Y(t) = 1$ if $Y(t-1) = 1$" ($\hat{p_{11}}$) and the dashed, "$Y(t) = 1$ if $Y(t-1) = 0$" ($\hat{p_{01}}$). **(Right)** is a smoothed version of the right curve using a moving average filter of length 11 (Section 3.2, Method 1).

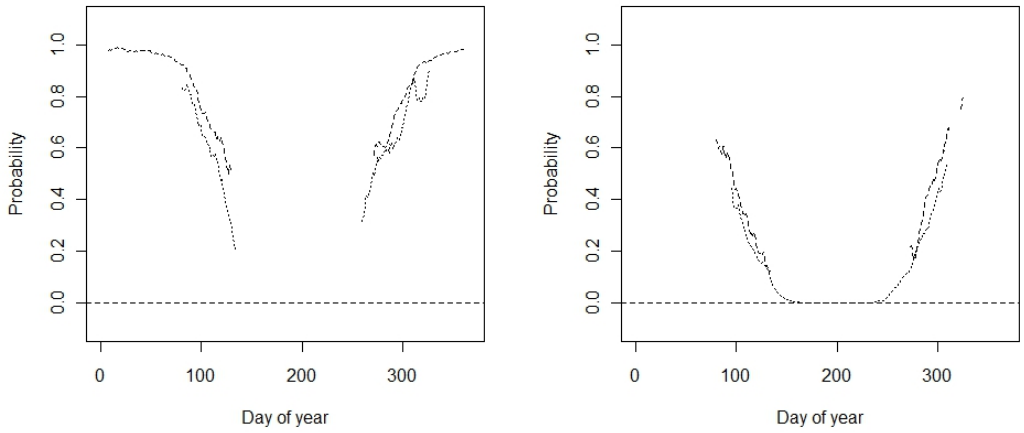

Figure 3: **(Left)** The smoothed estimated 2nd-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site with $\hat{p}_{111}$ (dashed) compared with $\hat{p}_{011}$ (dotted) calculated from the historical data. **(Right)** $\hat{p}_{001}$ (dashed) compared with $\hat{p}_{101}$ (dotted) calculated from the historical data.

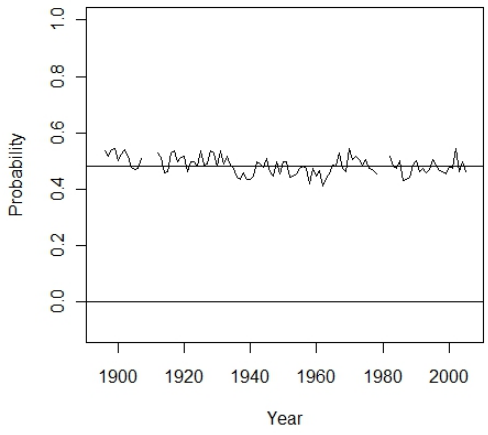

Figure 4: Medicine Hat's estimated annual proportion of frost days calculated from the historical data with the median line added.

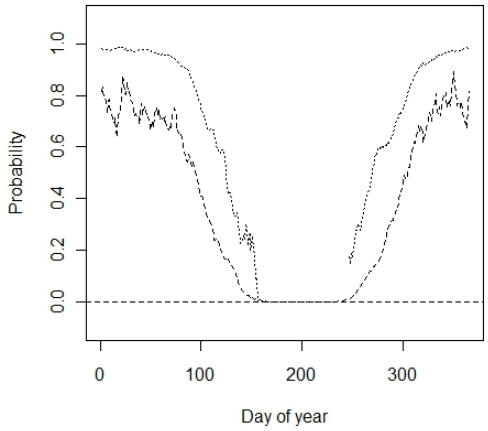

Figure 5: The non-parametric estimate of the 1st-order transition probability (dashed line p_{01} and dotted line p_{11}) of a freezing day for the Medicine Hat site for different days of a year computed using Method 2 given in Equation 1.

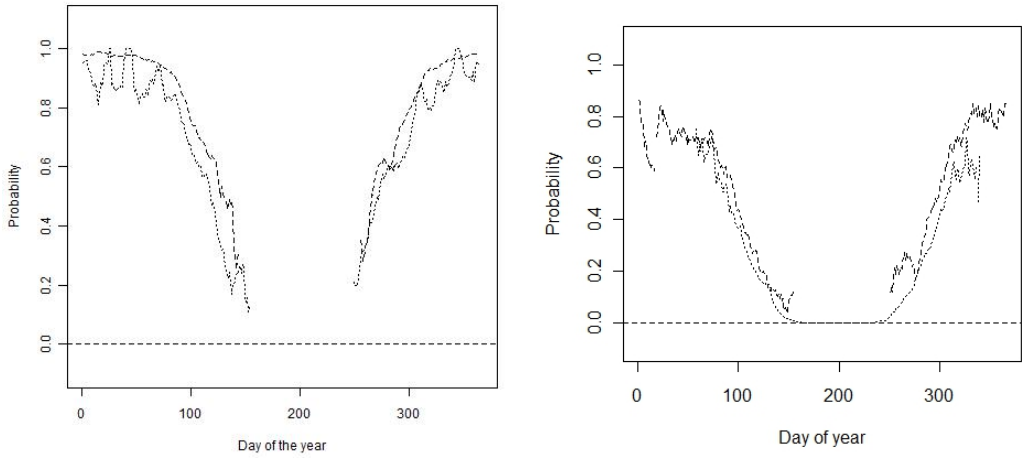

Figure 6: **(Left)** The smoothed estimates of 2nd-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site with $\hat{p}_{111}$ (dashed) compared with $\hat{p}_{011}$ (dotted) calculated from the historical data. **(Right)** The smoothed estimated $\hat{p}_{001}$ (solid) compared with $\hat{p}_{101}$ (dotted) using an extension of Method 2 (Equation 1) to higher orders.

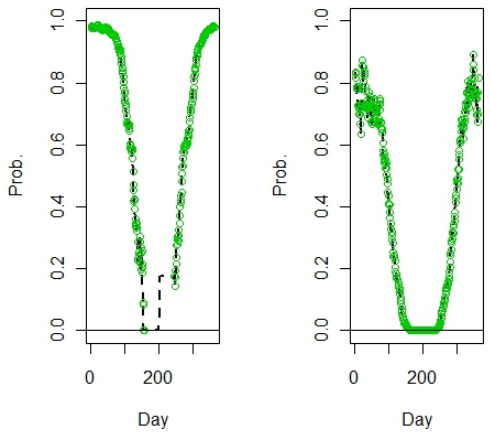

Figure 7: The estimated conditional probabilities (left panel p_{11} and right panel p_{01}) of a freezing day for the Medicine Hat site for different days of a year computed using the historical data (circles) with missing values filled-in (dashed line).

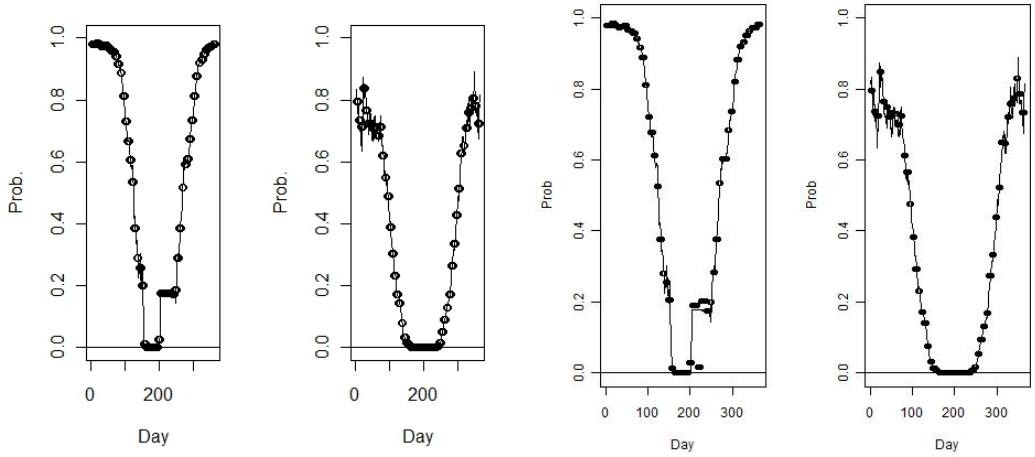

Figure 8: **(Left)** The estimated probability of a freezing day for the Medicine Hat site for different days of a year computed using the historical data average for 1 week. **(Right)** The estimated transition probability curves computed by maximizing the partial likelihood with 2×52 weekly transition parameters.

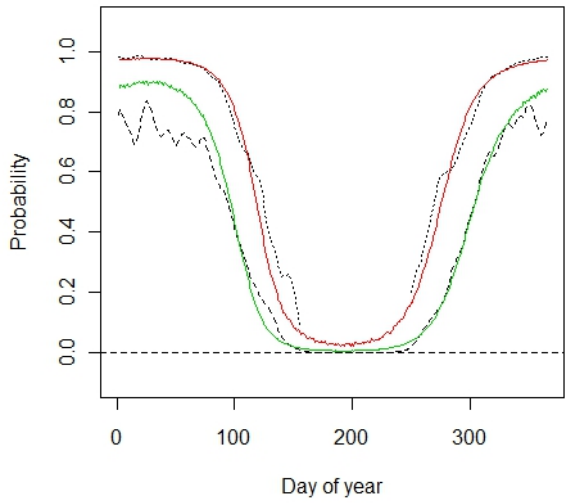

Figure 9: The estimated 1st-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site for the model with covariate process $(1, Y^1, N^5, sin, cos, cos2)$. The dotted line represents $\hat{p_{11}}$ and the dashed $\hat{p_{01}}$.

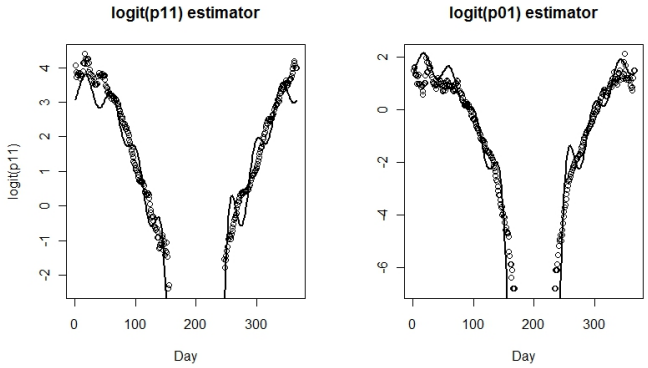

Figure 10: The fits to logit transition probability using fixed 1st-order Markov component models with seasonality covariates going up to $cos8$ and $sin8$. The fits are not satisfactory. For example we notice that at the beginning of the year $logit(p_{11})$ is under estimated and $logit(p_{01})$ is over estimated. The explanation seems to be that the fitted logit transition curves are forced to be vertical shifts of each other.

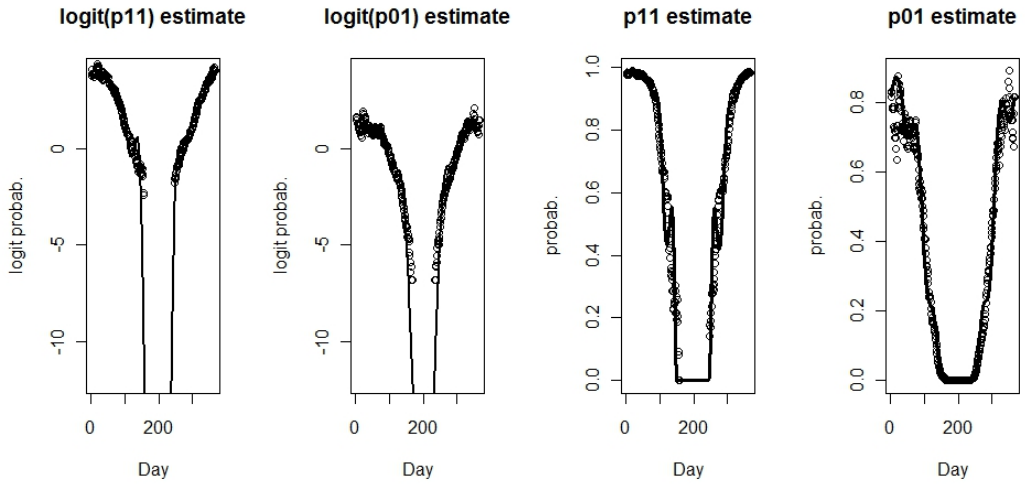

Figure 11: *(Lefthand two panels.)* The estimated logit transition probabilities computed by fitting $(1, cos, sin, \cdots, cos6, sin6)$ to the logit of nonparametric 1st-order transition probability estimates. *(Righthand two panels)* The estimated transition probabilities for $(1, cos, \cdots, sin6, Y^1, Y^1 cos, \cdots, Y^1 sin6)$, computed as MPLEs initialized at the estimates obtained from the estimated parameters in the left panel.

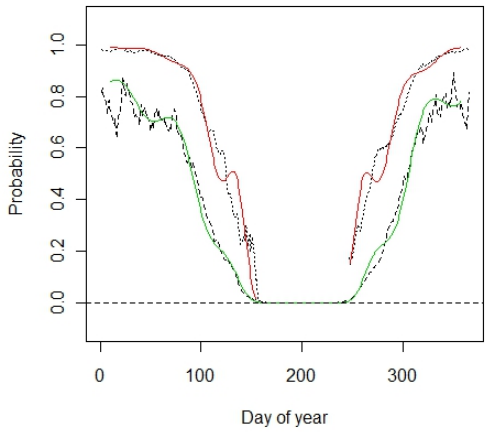

Figure 12: The smoothed estimated 1st-order transition probabilities for the 0-1 process of extreme minimum temperatures for the Medicine Hat site. The dotted line represents $\hat{p_{11}}$ and the dashed, $\hat{p_{01}}$. The fits are from the optimal model picked by *AIC* given in Table 3.

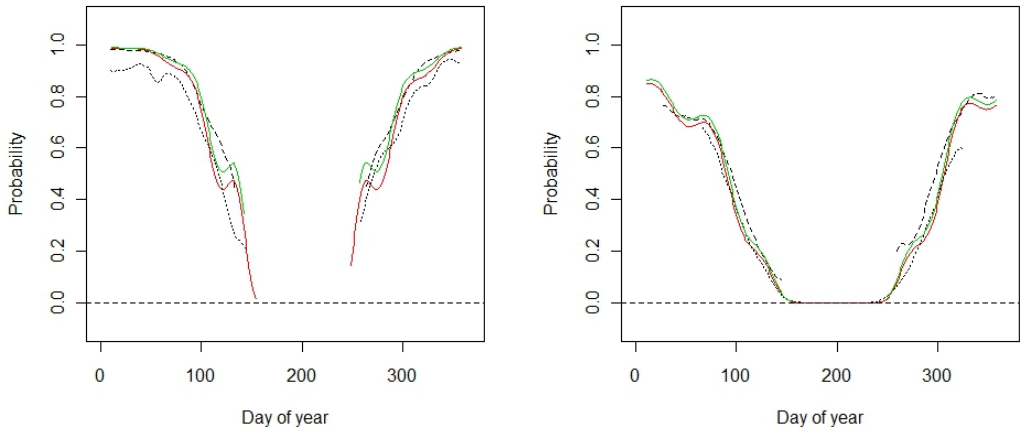

Figure 13: **(Left)** The smoothed $\hat{p}_{111}$ curve (dashed) compared with $\hat{p}_{011}$ (dotted). **(Right)** The smoothed $\hat{p}_{001}$ (dashed) compared with $\hat{p}_{101}$ (dotted). The fits from the optimal model picked by *AIC* in Table 3.

Table 1: The best five models picked by *BIC* and *AIC* for Medicine Hat, 1895–2006 for the binary process of days with frost.

Model covariates	*BIC*, rank	*AIC*, rank	parameter estimates
$(1, Y^1, cos, sin, cos2, N^5)$	**19039**, 1	18987, 1	(-0.9,1.4,-3.2,-0.8,0.4,0.2)
$(1, Y^1, cos, sin, Y^2, Y^1Y^2, cos2, N^5)$	19078, 2	19009, 2	(-0.9,1.4,-3,-0.8,0.04,0.1,0.5,0.2)
$(1, Y^1, cos, sin, cos2, N^{10})$	19088, 3	19036, 4	(-1.3,1.6,-2.7,-0.6,0.4,0.1)
$(1, Y^1, cos, sin, Y^2, cos2, sin2)$	19089, 4	19029, 3	(-0.7,1.6,-3.3,-0.9,0.6,0.5,0.1)
$(1, Y^1, cos, sin, Y^2, cos2)$	19093, 5	19042, 5	(-0.7,1.6,-3.4,-0.8,0.4,0.5)

Table 2: Comparing partial likelihood fits with random initial values (right panel) and fits using initial values obtained from non-parametric estimates (left panel).

	nonparametric initial		random initial	
n	(-*LPL*,*BIC*,*AIC*)	time	(-*LPL*,*BIC*,*AIC*)	time
2	(9506, 19118, 19032)	7 min	(9723, 19551, 19465)	21 min
3	(9893, 19934, 19813)	7 min	(9840, 19829, 19708)	21 min
4	(9494, 19179, 19024)	8 min	(10392, 20974, 20819)	22 min
5	(9432, **19096**, 18907)	8 min	(10835, 21903, 21714)	23 min
6	(9422, 19119, **18896**)	9 min	(13739, 27755, 27531)	22 min
7	(9440, 19199, 18941)	8 min	(16587, 33492, 33234)	22 min

Table 3: We add higher order covariates (N^5, Y^2, Y^1Y^2) to best 1st-order fits found by initializing partial likelihood maximization using non-parametric fits. n denotes the upper bound for the Fourier terms. The initial values for these models were obtained by extending the corresponding initial vector for the 1st-order case by simply adding zeros for the high-order covariates.

Model Complexity	(-*LPL*,*BIC*,*AIC*)
$n = 5, N^5$	(9408,19059,18862)
$n = 6, N^5$	(9402,19091,18859)
$n = 5, Y^2$	(9402,**19048**,18851)
$n = 6, Y^2$	(9396,19078,18846)
$n = 5, Y^2, Y^1Y^2$	(9402,19059,18852)
$n = 6, Y^2, Y^1Y^2$	(9392,19082,**18841**)