University of British Columbia
Department of Statistics
Technical Report #248
June 2009

r-th order categorical Markov chains.

by

Reza Hosseini[1], Nhu D Le[2] and James V Zidek[1]
[1] University of British Columbia
[2] British Columbia Cancer Agency & University of British Columbia

---

Deaprtment of Statistics, University of British Columbia

ABSTRACT

We prove a representation theorem for $\{X_t\}$ (t denotes time), an r-th order categorical Markov chain. We prove that the conditional probability $P(X_t|X_{t-1}, \cdots, X_{t-r})$ can be written as a linear combination of the monomials of past process responses $X_{t-1}, \cdots, X_{t-r}$. Simulations show that the "partial likelihood estimation" and the representation together give us satisfactory results. We also check the performance of "BIC" criterion for selecting optimal models and find that to be quite satisfactory. An advantage of this model over pre-existing models is its capacity to admit covariates as linear terms by extension. For example, we can add some seasonal processes to get a non-stationary chain for daily precipitation values.

1 Introduction

In this report, we study r-th order categorical Markov chains and more generally, categorical discrete-time stochastic processes. By "categorical", we mean chains that have a finite number of possible states at each time point. Such chains have important applications in many areas, one of which is modeling weather processes such as precipitation over time, the genesis of the paper. In fact, we use these chains to model the binary process of precipitation as well as dichotomized temperature processes. In r-th order Markov chains, the conditional probability of the present given the past is modeled. Such a conditional probability is a function of the past r states, where each one of them only takes finite possible values.

It is useful and intuitively appealing to specify or model a discrete process over time by the conditional probabilities rather than the joint distribution. However, one must check the consistency of such a specification i.e. to prove that it corresponds to a full joint distribution. In the case of discrete-time categorical processes, we prove a theorem that shows the conditional probabilities can be used to specify the process. Also we prove a representation theorem which states that every such conditional probability after an appropriate transformation can be written as a linear summation of monomials of the past processes. In fact, we represent all categorical discrete-time stochastic processes over time, in particular r-th order Markov chains and more particularly stationary r-th order Markov chains. For the binary case the result is a consequence of an expansion theorem due to Besag [2]. However, Besag did not provide a rigorous proof and the statement of the theorem is flawed as also pointed out by Cressie et al. in [4]. We provide a rigorous statement and proof in Section 5. To generalize the result to arbitrary categorical Markov chains, we prove a new expansion theorem which generalizes the result to the case of arbitrary categorical r-th order Markov chains (rather than binary only).

The result simplifies the task of modeling categorical stochastic processes. Since we have written the conditional probability as a linear combination, we can simply add other covariates as linear terms to the model to build non-stationary chains. For example, we can add seasonal terms or geographical coordinates (longitude and latitude). The theory of partial likelihood allows us to estimate the parameters of such chain models for the binary case. By restricting the degree of those polynomials or by requiring that some of their coefficients be the same, we can find simpler models. Simulation studies show that the "BIC" criterion (Bayesian information criterion) combined with the partial likelihood works well in that they recover the correct simulation model.

2 Markov chains

Let $\{X_t\}_{t\in T}$ be a stochastic process on the index set T, where $T = \mathbb{Z}$, $T = \mathbb{N}$ (the integers or natural numbers respectively) or $T = \{0, 1, \cdots, n\}$. It is customary to

call $\{X_t\}_{t\in T}$ a chain, since T is countable and has a natural ordering. $\{X_t\}_{t\in T}$ is called an rth-order Markov chain if:

$$P(X_t|X_{t-1},\cdots) = P(X_t|X_{t-1},\cdots,X_{t-r}),\ \forall t \text{ such that } t-r \in T.$$

We call the Markov chain homogenous if

$$P(X_t = x_t|X_{t-1} = x_{t-1},\cdots,X_{t-r} = x_{t-r}) =$$
$$P(X_{t'} = x_t|X_{t'-1} = x_{t-1},\cdots,X_{t'-r} = x_{t-r}),$$

$\forall t, t' \in T$ such that $t-r$ and $t'-r$ are also in T. Note that Markovness can be defined as a local property. We call $\{X_t\}_{t\in T}$ locally r-th order Markov at t if

$$P(X_t|X_{t-1},\cdots) = P(X_t|X_{t-1},\cdots,X_{t-r}).$$

Hence, we can have chains with a different Markov order at different times.

Let X_t be the binary random variable for precipitation on day t, with 1 denoting the occurrence of precipitation and 0 non-occurrence. In particular, consider the precipitation (PN) for Calgary site from 1895 to 2006. This process can be considered in two possible ways:

1. Let $X_1, X_2, \cdots, X_{366}$ denote the binary random variable of precipitation for days of a year. Suppose we repeatedly observe this chain year–by–year from 1895 to 2006 and take these observed chains to be independent and identically distributed from one year to the next. With this assumption, techniques developed in [1] can be applied in order to infer the Markov order of the chain. However, this approach presents three issues. Firstly independence of the successive chains seems questionable. In particular, the end of any one year will be autocorrelated with the beginning of the next. Secondly this model unrealistically assumes the 0-1 precipitation stochastic process is identically distributed over all years. Thirdly and more technically, leap years have 366 days while non-leap years have 365. We can resolve this last issue by formally assuming a missing data day in the non-leap years, by dropping the last day in the non-leap year or by using other methods. However, none of these approaches seem completely satisfactory.

2. Alternatively, we could consider the observations of Calgary daily precipitation as coming from a single process that spans the entire time interval from 1895 to 2006. In this case, we will show below that we can still build models that bring in the seasonality effects within a year.

3 Consistency of the conditional probabilities

To represent a stochastic process, we only need to specify the joint probability distributions for all finite collections of states. The Kolmogorov extension theorem then guarantees the existence and uniqueness of an underlying stochastic process from which these distributions derive, provided they are consistent as described below. (See [3] for example.)

To state the version of that celebrated theorem we require, let T denote some interval (that can be thought of as "time"), and let $n \in \mathbb{N} = \{1, 2, \dots, \}$. For each $k \in \mathbb{N}$ and finite sequence of times $t_1, \cdots, t_k$, let $\nu_{t_1 \cdots t_k}$ be a probability measure on $(\mathbb{R}^n)^k$. Suppose that these measures satisfy two consistency conditions:

1. **Permutation invariance.** For all permutations π (a bijective and one-to-one map from a set to itself) of $1, \cdots, k$ and measurable sets $F_i \subset \mathbb{R}^n$,

$$\nu_{t_{\pi(1)} \cdots t_{\pi(k)}} (F_1 \times \cdots \times F_k) = \nu_{t_1 \dots t_k} \left(F_{\pi^{-1}(1)} \times \cdots \times F_{\pi^{-1}(k)} \right).$$

2. **Marginalization consistency.** For all measurable sets $F_i \subseteq \mathbb{R}^n$, $m \in \mathbb{N}$:

$$\nu_{t_1 \dots t_k} (F_1 \times \cdots \times F_k) = \nu_{t_1 \dots t_k t_{k+1}, \dots, t_{k+m}} (F_1 \times \cdots \times F_k \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n).$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a stochastic process $X : T \times \Omega \to \mathbb{R}^n$ such that: $\nu_{t_1 \dots t_k} (F_1 \times \cdots \times F_k) = \mathbb{P} (X_{t_1} \in F_1, \dots, X_{t_k} \in F_k)$ for all $t_i \in T$, $k \in \mathbb{N}$ and measurable sets $F_i \subseteq \mathbb{R}^n$, i.e. X has the $\nu_{t_1 \dots t_k}$ as its finite-dimensional distributions. (See [6] for more details.)

Remarks:

1. Note that Condition 1 is equivalent to

$$\nu_{t_{\pi(1)} \cdots t_{\pi(k)}} \left(F_{\pi(1)} \times \cdots \times F_{\pi(k)} \right) = \nu_{t_1 \dots t_k} (F_1 \times \cdots \times F_k).$$

This is seen by replacing $F_1 \times \cdots \times F_k$ by $F_{\pi(1)} \times \cdots \times F_{\pi(k)}$ in the first equality.

2. We are only concerned about the case $n = 1$. This is because we consider stochastic processes, a collection of random variables from the same sample space to $\mathbb{R}^1 = \mathbb{R}$.

When working on (higher order) Markov chains over the index set $\mathbb{N}$, it is natural to consider the conditional distributions of the present, time t, given the past instead of the finite joint distributions, in other words

$$P_t(x_0, \cdots, x_t) = P(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0),$$

for $\{X_t\}_{t\in\mathbb{N}\cup 0}$ plus the starting distribution

$$P(X_0 = x_0).$$

However that raises a fundamental question – does there exist a stochastic process whose conditional distributions match the specified ones and if so, is it unique? We answer this question affirmatively in this section for the case of discrete-time categorical processes, in particular higher order categorical Markov chains. We also restrict ourselves to chains for which all the joint probabilities are positive. Let $M_0, M_1, \cdots \subset \mathbb{R}$ be the state spaces for time $0, 1, \cdots$, where each one of them is of finite cardinality. A probability measure on the finite space M_0 can be represented through its density function, a positive function $P_0 : M_0 \to \mathbb{R}$ satisfying the condition

$$\sum_{m\in M_0} P_0(m) = 1.$$

The following theorem ensures the consistency of our probability model.

Theorem 3.1 *Suppose $M_0, M_1, \cdots \subset \mathbb{R}$, $|M_t| = c_t < \infty$, $t = 0, 1, \cdots$. Let b $P_0 : M_0 \to \mathbb{R}$ be the density of a probability measure on M_0 and more generally for $n = 1, \ldots,$ $P_n(x_0, x_1, \cdots, x_{n-1}, .)$ be a positive probability density on M_n $\forall (x_0, \cdots, x_{n-1}) \in M_0 \times \cdots \times M_{n-1}$. Then there exists a unique stochastic process (up to distributional equivalence) on a probability space (Ω, Σ, P) such that*

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_0 = x_0) = P_n(x_0, x_1, \cdots, x_{n-1}, x_n).$$

To prove this theorem, we first consider a related problem whose solution is used in the proof. More precisely we consider stochastic processes $\{X_i\}_{i\in\mathbb{N}\cup\{0\}}$, where the state space for X_i is M_i, $1 = 0, 1, 2, \cdots$ and finite. Suppose $p_n : M_0 \times M_1 \times \cdots \times M_n \to \mathbb{R}$ is the joint probability distribution (density) of a random vector $\{X_0, \ldots, X_n\}$, i.e.

$$p_n(x_0, \cdots, x_n) = P(X_0 = x_0, \cdots, X_n = x_n).$$

It is clear that given the $\{p_n\}_{n\in\mathbb{N}}$, other joint distributions such as
$P(X_{t_1} = x_{t_1}, \cdots, X_{t_k} = x_{t_k})$ are obtainable by summing over appropriate components. Now consider the inverse problem. Given the $\{p_n\}_{n\in\mathbb{N}}$ and some type of consistency between them, is there a (unique) stochastic process that matches these joint distributions? The following lemma gives an affirmative answer.

Lemma 3.1 *Suppose $M_t \subset \mathbb{R}$, $t = 0, 1, \cdots$ are finite, $p_0 : M_0 \to \mathbb{R}$ represents a probability density function (i.e. $\sum_{x_0\in M_0} p(x_0) = 1$) and functions $p_n : M_1 \times \cdots \times M_n \to \mathbb{R}^+ \cup \{0\}$ satisfy the following (consistency) condition:*

$$\sum_{x_n \in M_n} p_n(x_0, \cdots, x_n) = p_{n-1}(x_0, \cdots, x_{n-1}).$$

Then there exist a unique stochastic process (up to distributional equivalence) $\{X_t\}_{t \in \mathbb{N} \cup \{0\}}$ *such that*

$$P(X_0 = x_0, \cdots, X_n = x_n) = p_n(x_0, \cdots, x_n)$$

Proof

Existence: By the Kolmogorov extension theorem quoted above, we only need to show there exists a consistent family of measures (density functions)

$$\{q_{t_1,\cdots,t_k} | k \in \mathbb{N}, (t_1, \cdots, t_k) \in \mathbb{N}^k\}$$

such that $q_{1,\cdots,t} = p_t$. We define such a family of functions, prove they are measures and consistent.

For and sequence, $t_1, \cdots, t_k$, let $t = \max\{t_1, \cdots, t_k\}$ and define

$$q_{t_1,\cdots,t_k}(x_{t_1}, \cdots, x_{t_k}) = \sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,t_k\}} p_t(x_1, \cdots, x_t).$$

We need to prove three things:

a) Each $q_{t_1,\cdots,t_k}$ is a density function. It suffices to show that q_t is a measure because the $q_{t_1,\cdots,t_k}$ are sums of such measures and so are measures themselves. But p_t is non-negative by assumption. It only remain to show that p_t sums up to one. For $t = 1$ it is in the assumptions of the theorem. For $t > 1$, it can be done by induction because of the following identity

$$\sum_{x_i \in M_i, i=0,1,\cdots,t} p_t(x_0, \cdots, x_t) = \sum_{x_i \in M_i, i=0,1,\cdots,t-1} p_{t-1}(x_0, \cdots, x_t)$$

where the right hand side is obtained by the assumption $\sum_{M_n} p_n = p_{n-1}$.

b) In order to satisfy the first condition of Kolmogorov extension theorem, we need to show

$$q_{t_1,\cdots,t_k}(x_{t_1}, \cdots, x_{t_k}) = q_{t_{\pi(1)},\cdots,t_{\pi(k)}}(x_{t_{\pi(1)}}, \cdots, x_{t_{\pi(k)}}),$$

for π a permutation of $\{1, 2, \cdots, k\}$. But this is obvious since $\max\{t_1, \cdots, t_k\} = \max\{t_{\pi(1)}, \cdots, t_{\pi(k)}\}$.

c) In order to satisfy the second condition of Kolmogorov extension theorem, we need to show

$$\sum_{x_{t_i} \in M_{t_i}} q_{t_1,\cdots,t_i,\cdots,t_k}(x_{t_1}, \cdots, x_{t_i}, \cdots, x_{t_k}) = q_{t_1,\cdots,\hat{t_i},\cdots,t_k}(x_{t_1}, \cdots, \hat{x_{t_i}}, \cdots, x_{t_k}),$$

where the notation $\hat{}$ above a component means that component is omitted. To prove this, we consider two cases:

Case I: $t = \max\{t_1, \cdots, t_k\} = \max\{t_1, \cdots, \hat{t_i}, \cdots, t_k\}$, then

$$\sum_{x_{t_i} \in M_{t_i}} q_{t_1,\cdots,t_i,\cdots,t_k}(x_{t_1}, \cdots, x_{t_i}, \cdots, x_{t_k}) =$$
$$\sum_{x_{t_i} \in M_{t_i}} \sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,t_i,\cdots,t_k\}} p_t(x_0, \cdots, x_t) =$$
$$\sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} p_t(x_0, \cdots, x_t) =$$
$$p_{t_1,\cdots,\hat{t_i},\cdots,t_k}(x_{t_1}, \cdots, \hat{x_{t_i}}, \cdots, x_{t_k})$$

Case II: $\max\{t_1, \cdots, \hat{t_i}, \cdots, t_k\} = t' < t = t_i$:

$$\sum_{x_{t_i} \in M_{t_i}} q_{t_1,\cdots,t_i,\cdots,t_k}(x_{t_1}, \cdots, x_{t_i}, \cdots, x_{t_k}) =$$
$$\sum_{x_{t_i} \in M_{t_i}} \sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,t_i,\cdots,t_k\}} p_t(x_0, \cdots, x_t) =$$
$$\sum_{x_u \in M_u, u \in \{1,\cdots,t\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} p_t(x_0, \cdots, x_t) =$$
$$\sum_{x_u \in M_u, u \in \{1,\cdots,t'\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} \sum_{x_v \in M_v, v \in \{t'+1,\cdots,t\}} f_t(x_0, \cdots, x_t) =$$
$$\sum_{x_u \in M_u, u \in \{1,\cdots,t'\} - \{t_1,\cdots,\hat{t_i},\cdots,t_k\}} p_{t'}(x_0, \cdots, x_{t'}) =$$
$$q_{t_1,\cdots,\hat{t_i},\cdots,t_k}(x_{t_1}, \cdots, \hat{x_{t_i}}, \cdots, x_{t_k}).$$

Uniqueness: Suppose $\{Y_t\}_{t \in \mathbb{N} \cup \{0\}}$ is another stochastic process satisfying the conditions of the theorem with the $p'_{t_1,\cdots,t_k}$ as the joint measures.

$$p'_{1,\cdots,t} = p_t = p_{1,\cdots,t},$$

by the assumption. Taking the sums on the two sides, we get $p'_{t_1,\cdots,t_k} = p_{t_1,\cdots,t_k}$. Now the uniqueness is a straight consequence of the Kolmogorov Extension Theorem. ■

Remark:

3. Note that we did not impose the positivity of the functions for this case.

Now we are ready to prove Theorem 3.1.

Proof

Existence: In Lemma 3.1, let

$p_0 = P_0$,

$p_1 : M_0 \times M_1 \to \mathbb{R},\ p_1(x_0, x_1) = p_0(x_0)P_1(x_0, x_1)$,

$\vdots$

$p_n : M_1 \times M_2 \times \cdots \times M_n \to \mathbb{R},\ p_n(x_0, \cdots, x_n) = p_{n-1}(x_0, \cdots, x_{n-1})P_n(x_0, \cdots, x_n)$.

To see that the $\{p_i\}$ satisfy the conditions of Lemma 3.1, note that

$$\sum_{x_n \in M_n} p_n(x_0, \cdots, x_n) =$$

$$\sum_{x_n \in M_n} p_{n-1}(x_0, \cdots, x_{n-1})P_n(x_0, \cdots, x_n) =$$

$$p_{n-1}(x_0, \cdots, x_{n-1}) \sum_{x_n \in M_n} P_n(x_0, \cdots, x_n) =$$

$$p_{n-1}(x_0, \cdots, x_{n-1}).$$

Lemma 3.1 shows the existence of a stochastic process with joint distributions matching the p_i. Furthermore, the positivity of the $\{P_i\}$ implies that of the $\{p_i\}$. Thus all the conditionals exist for such a process and they match the P_i by the definition of the conditional probabilities.

Uniqueness. Any stochastic process satisfying the above conditions, has a joint distribution that matches those of the $\{p_i\}$ and hence by the above theorem they are unique. ■

4 Characterizing density functions and r-th order Markov chains

The previous section saw discrete-time categorical processes represented in terms of conditional probability density functions. However such densities on finite domains satisfy certain restrictions that can make modeling them difficult. That leads to the idea of linking them to unrestricted functions on $\mathbb{R}$ in much the same spirit as a single probability can profitably be logit transformed in logistic regression.

To begin, let X be a random variable with probability density p defined on a finite set $M = \{m_1, \cdots, m_n\}$. The section finds the class of all possible such ps with $p(m_i) > 0,\ i = 1, \cdots, n$ and $g : \mathbb{R} \to \mathbb{R}^+$, a fixed bijection. The following theorem characterizes the relationship between p and g.

Theorem 4.1 *Let $g : \mathbb{R} \to \mathbb{R}^+$ a bijection. For every choice of probability density p on $M = \{m_1, \cdots, m_n\}$, $n \geq 2$, there exists a unique function $f : M - \{m_1\} \to \mathbb{R}$, such that*

$$p(m_1) = \frac{1}{1 + \sum_{x \in M - \{m_1\}} h(x)}, \tag{1}$$

$$p(x) = \frac{h(x)}{1 + \sum_{x \in M - \{m_1\}} h(x)} \; x \neq m_1, \tag{2}$$

where $h = g \circ f$. Moreover, $h(x) = p(x)/p(m_1)$. Inversely, for an arbitrary function $f : M - \{m_1\} \to \mathbb{R}$, the p defined above is a density function.

Proof
Existence: Suppose $p : M \to (0, 1)$ is given. Let $h(x) = \frac{p(x)}{p(m_1)}$, $x \neq m_1$ and $f : M - \{m_1\} \to \mathbb{R}$, $f(x) = g^{-1} \circ h(x)$. Obviously $h = g \circ f$. Moreover

$$\frac{1}{1 + \sum_{x \in M - \{m_1\}} h(x)} = \frac{1}{1 + \sum_{x \in M - \{m_1\}} p(x)/p(m_1)} =$$

$$\frac{1}{1 + (1 - p(m_1))/p(m_1)} = p(m_1)$$

and

$$\frac{h(x)}{1 + \sum_{x \in M - \{m_1\}} h(x)} = \frac{p(x)/p(m_1)}{1 + (1 - p(m_1))/p(m_1)} = p(x)$$

thereby establishing the validity of equations (1) and (2).
Uniqueness: Suppose for f_1, f_2, we get the same p. Let $h_1 = g \circ f_1$, $h_2 = g \circ f_2$, by dividing 2 by 1 for h_1 and h_2, we get $h_1(x) = p(x)/p(m_1) = h_2(x)$ hence $g \circ f_1 = g \circ f_2$. Since g is a bijection $f_1 = f_2$. ■

Corollary 4.1 *Fixing a bijection g and $m_1 \in M$, every density function corresponds to an arbitrary vector of length $n - 1$ over $\mathbb{R}$.*

Theorem 4.2 *Fix a bijection $g : \mathbb{R} \to \mathbb{R}^+$, $m_1^i \in M_i$. Let $M_i, i = 0, 1, \cdots$ be finite subsets of $\mathbb{R}$ with cardinality greater than or equal to 2 and $M_i' = M_i - \{m_1^i\}$, $\forall i$. Then every categorical stochastic process with positive joint distribution on the M_i having starting density $P_0 : M_0 \to \mathbb{R}$ and conditional probabilities P_i at stage i given the past, can be uniquely represented by means of unique functions:*

$$g_0 : M_0' \to \mathbb{R}$$
$$\vdots$$
$$g_n : M_0 \times \cdots \times M_{n-1} \times M_n' \to \mathbb{R}$$
$$\vdots$$

for $n = 1, \ldots,$ where

$$P_0(m_1^0) = \frac{1}{1 + \sum_{x \in M_0 - \{m_1^0\}} h_0(x)}, \tag{3}$$

$$P_0(x) = \frac{h_0(x)}{1 + \sum_{x \in M_0 - \{m_1^0\}} h_0(x)}, \;\; x \neq m_1^0 \in M_0, \tag{4}$$

and $h_0 = g \circ g_0$. Moreover $h_0(x) = \frac{P(X_0 = x)}{P(X_0 = m_0^1)}$.
The conditional probabilities P_i are given by

$$P_n(x_0, \cdots, x_{n-1}, m_1^n) = \frac{1}{1 + \sum_{x \in M_n - \{m_1^n\}} h_n(x)}, \tag{5}$$

$$P_n(x_0, \cdots, x_{n-1}, x) = \frac{h(x)}{1 + \sum_{x \in M_n - \{m_1^n\}} h_n(x)}, \;\; x \neq m_1^n \in M_n, \tag{6}$$

where, $h_n = g \circ g_n$. Moreover $h_n(x_0, \cdots, x) = \frac{P(X_n = x | X_{n-1} = x_{n-1}, \cdots, X_0 = x_0)}{P(X_n = m_n^1 | X_{n-1} = x_{n-1}, \cdots, X_0 = x_0)}$.
Conversely, any collection of arbitrary functions $g_0, g_1, \cdots$ gives rise to a unique stochastic process by the above relations.

Proof

The result is an immediate by Theorems 3.1 and 4.1. ■

Remarks:

4. We can view the arbitrary functions $g_0, \cdots, g_n$ on $M_0', M_0 \times M_1', \cdots, M_0 \times \cdots \times M_{n-1} \times M_n'$ as arbitrary functions g_0 on M_0', $g_1(., x_1),\ x_1 \neq m_1^1$ on M_0 and $g_n(., x_n),\ x_n \neq m_1^n$ on $M_0 \times \cdots \times M_{n-1}$. As a check we can compute the number of free parameters of such a stochastic process on $M_0, \cdots, M_n$. We can specify such a process by $c_0 c_1 \cdots c_n - 1$ parameters by specifying the joint distribution on $M_0 \times M_1 \times M_n$. If we specify the stochastic process using the above theorems and the g_i functions, we need $(m_0 - 1) + m_0(m_1 - 1) + m_0 m_1 (m_2 - 1) + \cdots + m_0 m_1 \cdots m_{n-1}(m_n - 1)$ which is the same number after expanding the terms and canceling out.

5. In the case of r-th order Markov chains, $g_n(x_0, \cdots, x_n)$ only depends on the last $r + 1$ components for $n > r$.

6. In the case of homogenous r-th order Markov chains, $M_i = M_0,\ \forall i$. Fix $m_0 \in M_0$ and suppose $|M_0| = c_0$. We only need to specify g_0 to g_r, which are completely arbitrary functions. We only need to specify g_0 on M_0' and g_r on $M_0 \times \cdots \times M_{r+1}'$. This also shows every homogenous Markov chain of order at most r is characterized by $(c_0 - 1) \sum_{i=0}^{r} c_0^r$ elements $\mathbb{R}$.

To describe processes using Markov chains, we need to find appropriate parametric forms. We investigate the generality of these forms in the following section and use the concept of partial likelihood to estimate them. We find appropriate parametric representations of g_n which are functions of $n+1$ finite variables. In the next section we study the properties of such functions. We call a variable "finite" if it only takes values in a finite subset of $\mathbb{R}$.

5 Functions of r variables on a finite domain

In this section, we study the properties of functions of r variables with finite domain. First, we present a result of Besag [2] who studied such functions in the context of Markov random fields. However the statement of the result in his paper is inaccurate and moreover it gives no rigorous proof of his result. We present a rigorous statement, proof of the result and generalization of Besag's theorem.

5.1 First representation theorem

Theorem 5.1 *Suppose,* $f : \prod_{i=1,\cdots,r} M_i \to \mathbb{R}$, M_i *being finite with* $|M_i| = c_i$ *and* $0 \in M_i,\ \forall i,\ 1 \le i \le r$. *Let* $M_i' = M_i - \{0\}$. *Then there exist a unique family of functions*

$$\{G_{i_1,\cdots,i_k} : M_{i_1}' \times M_{i_2}' \times \cdots \times M_{i_k}' \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\},$$

such that

$$f(x_1,\cdots,x_r) = f(0,\cdots,0) + \sum_{i=1}^{r} x_i G_i(x_i) +$$
$$\sum_{1\le i_1<i_2<\cdots<i_k\le r} (x_{i_1}\cdots x_{i_k}) G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) + \cdots + (x_1 x_2 \cdots x_r) G_{12\cdots r}(x_1,\cdots,x_r).$$

Proof Denote by I_A the indicator function of a set A and $N_k = \{(x_1,\cdots,x_r) : \sum_{i=1}^r I_{\{0\}}(x_i) \le k\}$.
Existence: The proof is by induction. For $i = 1,\cdots,r$, define

$$G_i : M_i' \to \mathbb{R}$$
$$G_i(x_i) = \frac{f(0,\cdots,0,x_i,0,\cdots,0) - f(0,\cdots,0)}{x_i},$$

where x_i is the i^{th} coordinate. Then let $f_1(x_1,\cdots,x_r) = f(0,\cdots,0) + \sum_{i=1}^r x_i G_i(x_i)$. Note that $f_1 = f$ on N_1.

Next define $G_{i_1,i_2} : M'_{i_1} \times M'_{i_2} \to \mathbb{R}$ by

$$G_{i_1,i_2}(x_{i_1}, x_{i_2}) = \frac{f(0,\cdots,0,x_{i_1},0\cdots,0,x_{i_2},0,\cdots,0) - f_1(0,\cdots,0,x_{i_1},0\cdots,0,x_{i_2},0,\cdots,0)}{x_{i_1}x_{i_2}},$$

where, x_{i_1}, x_{i_2} are the i_1^{th} and i_2^{th} coordinates, respectively. Using the $\{G_{i_1,i_2}\}$, we can define f_2 on N_2 by

$$f_2(x_1,\cdots,x_r) = f(0,\cdots,0) + \sum_{i=1}^{r} x_i G_i(x_i) + \sum_{1\le i_1<i_2\le r} x_{i_1}x_{i_2}G_{i_1,i_2}(x_{i_1},x_{i_2}).$$

Or equivalently,

$$f_2(x_1,\cdots,x_r) = f_1(x_1\cdots,x_r) + \sum_{1\le i_1<i_2\le r} x_{i_1}x_{i_2}G_{i_1,i_2}(x_{i_1},x_{i_2}).$$

It is easy to see that $f_2 = f$ on N_2.

In general, suppose we have defined $G_{i_1,\cdots,i_{k-1}}$ and f_{k-1}, let

$$G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) = \frac{f(0,\cdots,0,x_{i_1},0\cdots,0,x_{i_k},0,\cdots,0) - f_{k-1}(0,\cdots,0,x_{i_1},0\cdots,0,x_{i_k},0,\cdots,0)}{x_{i_1}\cdots x_{i_k}},$$

for $(x_{i_1},\cdots,x_{i_k}) \in M'_{i_1} \times \cdots \times M'_{i_k}$.
Also let

$$f_k(x_1,\cdots,x_r) = f_{k-1}(x_1,\cdots,x_r) + \sum_{1\le i_1<i_2<\cdots<i_k\le r} x_{i_1}\cdots x_{i_k}G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k})$$

We claim $f = f_k$ on N_k.

To see that, fix $x = (x_1,\cdots,x_r)$. If x has less than k nonzero elements, the second term in the above expansion will be zero and

$$f_k(x_1,\cdots,x_r) = f_{k-1}(x_1,\cdots,x_r) = f(x_1,\cdots,x_r),$$

by the induction hypothesis and we are done.

However if x has exactly k nonzero elements

$$x = (x_1,\cdots,x_r) = (0,\cdots,0,x_{j_1},0,\cdots,0,x_{j_k},0\cdots),$$

Then

$$\sum_{1\le i_1<i_2<\cdots<i_k\le r} x_{i_1}\cdots x_{i_k}G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k}) = x_{j_1}\cdots,x_{j_k}G_{j_1,\cdots,j_k}(x_{j_1},\cdots,x_{j_k}).$$

Hence

$$\begin{aligned} f_k(x_1, \cdots, x_r) &= f_{k-1}(x_1, \cdots, x_r) + (x_{j_1}, \cdots, x_{j_k}) G_{j_1, \cdots, j_k}(x_{j_1}, \cdots, x_{j_k}) \\ &= f_{k-1}(x_1, \cdots, x_r) + \\ x_{j_1} \cdots x_{j_k} & \frac{f(\cdots, 0, x_{j_1}, 0, \cdots, 0, x_{j_k}, 0, \cdots) - f_{k-1}(\cdots, 0, x_{j_1}, 0, \cdots, 0, x_{j_k}, 0, \cdots)}{x_{j_1} \cdots x_{j_k}} \\ &= f(x_1, \cdots, x_r) \end{aligned}$$

By induction, $f = f_r$ on $N_r = \prod_{i=1,\cdots,r} M_i$. Hence, the family of functions satisfies the conditions.
Uniqueness: To prove uniqueness, suppose

$$\{G_{i_1,\cdots,i_k} : M'_{i_1} \times M'_{i_2} \times \cdots \times M'_{i_k} \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\}$$

and

$$\{H_{i_1,\cdots,i_k} : M'_{i_1} \times M'_{i_2} \times \cdots \times M'_{i_k} \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\}$$

are two families of functions satisfying the equation. Also assume f_k^G and f_k^H are the summation functions as defined above corresponding to the two families. We need to show $G_{i_1,\cdots,i_k} = H_{i_1,\cdots,i_k}$ on $M'_{i_1} \times \cdots \times M'_{i_k}$. We use induction on k. It is easy to verify the result for the case $k = 1$. Now suppose, $x = (x_{i_1}, \cdots, x_{i_k}) \in M'_{i_1} \times M'_{i_2} \times \cdots \times M'_{i_k}$. Then by definition

$$G_{i_1,\cdots,i_k}(x_{i_1}, \cdots, x_{i_k}) =$$
$$\frac{f(0, \cdots, 0, x_{i_1}, 0 \cdots, 0, x_{i_k}, 0, \cdots, 0) - f_{k-1}^G(0, \cdots, 0, x_{i_1}, 0 \cdots, 0, x_{i_k}, 0, \cdots, 0)}{x_{i_1} \cdots x_{i_k}}$$

and

$$H_{i_1,\cdots,i_k}(x_{i_1}, \cdots, x_{i_k}) =$$
$$\frac{f(0, \cdots, 0, x_{i_1}, 0 \cdots, 0, x_{i_k}, 0, \cdots, 0) - f_{k-1}^H(0, \cdots, 0, x_{i_1}, 0 \cdots, 0, x_{i_k}, 0, \cdots, 0)}{x_{i_1} \cdots x_{i_k}}$$

But by induction hypothesis, $f_{k-1}^G = f_{k-1}^H$ so we are done. ■

We can think of this representation of f as an expansion around $(0, \cdots, 0)$. However, $(0, \cdots, 0)$ has no intrinsic role and we can generalize the above theorem as follows.

Theorem 5.2 *Suppose,* $f : M = \prod_{i=1,\cdots,r} M_i \to \mathbb{R}$, M_i *being finite and* $|M_i| = c_i$. *For any fixed* $(\mu_1, \cdots, \mu_r) \in M$, *let* $M_i' = M_i - \{\mu_i\}$. *Then there exist unique functions*

$$\{H_{i_1,\cdots,i_k} : M_{i_1}' \times M_{i_2}' \times \cdots \times M_{i_k}' \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\}$$

such that

$$f(x_1, \cdots, x_r) = f(\mu_1, \cdots, \mu_r) + \sum_{i=1}^{r} (x_i - \mu_i) H_i(x_i) +$$
$$\sum_{1 \le i_1 < i_2 < \cdots < i_k \le r} (x_{i_1} - \mu_{i_1}) \cdots (x_{i_k} - \mu_{i_k}) H_{i_1,\cdots,i_k}(x_{i_1}, \cdots, x_{i_k}) +$$
$$\cdots + (x_1 - \mu_1)(x_2 - \mu_2) \cdots (x_r - \mu_r) H_{12\cdots r}(x_1, \cdots, x_r).$$

Proof Let $N_i = M_i - \mu_i$ (meaning that we subtract μ_i from all elements of M_i) so that N_i and M_i have the same cardinality. Also let $N = \prod_{i=1,\cdots,r} N_i$ and $N_i' = N_i - \{0\}$. Then define a bijective mapping

$$\phi_i : N_i \to M_i,$$

$$\phi_i(x_i) = x_i + \mu_i.$$

This will induce a bijective mapping Φ between N and M that takes $(0, \cdots, 0)$ to $(\mu_1, \cdots, \mu_r)$. Now consider $f \circ \Phi : \prod_{i=1,\cdots,r} N_i \to \mathbb{R}$. By the previous theorem, unique functions

$$\{G_{i_1,\cdots,i_k} : N_{i_1}' \times N_{i_2}' \times \cdots \times N_{i_k}' \to \mathbb{R},\ 1 \le k \le r,\ 1 \le i_1 < i_2 < \cdots < i_k \le r\}$$

exist such that

$$f \circ \Phi(x_1, \cdots, x_r) = f \circ \Phi(0, \cdots, 0) + \sum_{i=1}^{r} x_i G_i(x_i) +$$
$$\sum_{1 \le i_1 < i_2 < \cdots < i_k \le r} x_{i_1} \cdots x_{i_k} G_{i_1,\cdots,i_k}(x_{i_1}, \cdots, x_{i_k}) + \cdots + x_1 x_2 \cdots x_r G_{12\cdots r}(x_1, \cdots, x_r).$$

Hence,

$$f(\phi_1(x_1), \cdots, \phi_r(x_r)) = f(\phi_1(0), \cdots, \phi_r(0)) + \sum_{i=1}^{r} x_i G_i(x_i) +$$
$$\sum_{1 \le i_1 < i_2 < \cdots < i_k \le r} x_{i_1} \cdots x_{i_k} G_{i_1,\cdots,i_k}(x_{i_1}, \cdots, x_{i_k}) + \cdots + x_1 x_2 \cdots x_r G_{12\cdots r}(x_1, \cdots, x_r)$$

We conclude,

$$f(x_1+\mu_1,\cdots,x_r+\mu_r)=f(\mu_1,\cdots,\mu_r)+\sum_{i=1}^{r}x_iG_i(x_i)+$$
$$\sum_{1\le i_1<i_2<\cdots<i_k\le r}x_{i_1}\cdots x_{i_k}G_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k})+\cdots+x_1x_2\cdots x_rG_{12\cdots r}(x_1,\cdots,x_r)$$

This gives:

$$f(x_1,\cdots,x_r)=f(\mu_1,\cdots,\mu_r)+\sum_{i=1}^{r}(x_i-\mu_1)G_i(x_i-\mu_i)+$$
$$\sum_{1\le i_1<i_2<\cdots<i_k\le r}(x_{i_1}-\mu_{i_1})\cdots(x_{i_k}-\mu_{i_k})G_{i_1,\cdots,i_k}(x_{i_1}-\mu_{i_1},\cdots,x_{i_k}-\mu_{i_k})+$$
$$\cdots+(x_1-\mu_1)(x_2-\mu_2)\cdots(x_r-\mu_r)G_{12\cdots r}(x_1-\mu_1,\cdots,x_r-\mu_r).$$

To prove the existence, let

$$H_{i_1,\cdots,i_k}(x_{i_1},\cdots,x_{i_k})=G_{i_1,\cdots,i_k}(x_{i_1}-\mu_{i_1},\cdots,x_{i_k}-\mu_{i_k}).$$

The uniqueness can be obtained as in the previous theorem. ■
We call this expression the Besag expansion around $(\mu_1,\cdots,\mu_r)$.

Corollary 5.1 *In the case of binary $\{0,1\}$ variables, the G functions are simply real numbers, since $M'_{i_1}\times\cdots\times M'_{i_k}$ has exactly one element: $(1,\cdots,1)$. Hence, we have found a linear representation of f in terms of the $x_{i_1}\cdots x_{i_k}$.*

Corollary 5.2 *Suppose that $\{X_t\}$ is an r-th order Markov chain, X_t taking values in $M_t=\{0,1\}$ and the conditional probability*

$$P(X_t=1|X_{t-1},\cdots,X_0),$$

is well-defined and in (0,1). Let $g:\mathbb{R}\to\mathbb{R}^+$ be a given bijective transformation. Then

$$g_t(x_{t-1},\cdots,x_0)=g^{-1}\{\frac{P(X_t=1|X_{t-1}=x_{t-1},\cdots,X_0=x_0)}{P(X_t=0|X_{t-1}=x_{t-1},\cdots,X_0=x_0)}\},$$

is a function of t variables, $(x_{t-1},\cdots,x_0)$, for $t<r$ and is a function of r variables, $(x_{t-1},\cdots,x_{t-r})$, for $t>r$. Hence there exist parameters α_0^t, $\{\alpha^t_{i_1,\cdots,i_t}\}_{1\le i_1,\cdots,i_t\le t}$ for $t<r$ and $\alpha_0^t,\{\alpha^t_{i_1,\cdots,i_r}\}_{1\le i_1,\cdots,i_r\le r}$ for $t\ge r$ such that for $t<r$

$$g^{-1}\{\frac{P(X_t=1|X_{t-1},\cdots,X_0)}{P(X_t=0|X_{t-1},\cdots,X_0)}\}=$$
$$\alpha_0^t+\sum_{i=1}^{t}X_{t-i}\alpha_i^t+\cdots$$
$$\sum_{1\le i_1<i_2<\cdots<i_k\le t}\alpha^t_{i_1,\cdots,i_k}X_{t-i_1}\cdots X_{t-i_k}+\cdots+$$
$$\alpha^t_{12\cdots t}X_{t-1}X_{t-2}\cdots X_0.$$

and for $t\ge r$

$$g^{-1}\circ\frac{P(X_t=1|X_{t-1},\cdots,X_0)}{P(X_t=0|X_{t-1},\cdots,X_0)}=$$
$$\alpha_0^t+\sum_{i=1}^{r}X_{t-i}\alpha_i^t+\cdots$$
$$\sum_{1\le i_1<i_2<\cdots<i_k\le r}\alpha^t_{i_1,\cdots,i_k}X_{t-i_1}\cdots X_{t-i_k}+\cdots+$$
$$\alpha^t_{12\cdots r}X_{t-1}X_{t-2}\cdots X_0.$$

In the case of homogenous Markov chains the $\alpha_0^t,\ \alpha^t_{i_1,\cdots,i_k}$ *do not depend on* t *for* $t>r$.

The above corollary shows that the conditional probability of a Markov chain after an appropriate transformation can be uniquely represented as a linear combination of monomial products of previous states.

One might conjecture that the same result holds for all categorical-valued Markov chains (with a finite number of states) using the above theorem. This is not true in general since the $\{G_{i_1,\cdots,i_k}\}$ are functions. In the next section, we prove another representation theorem which paves the way for the categorical case. As it turns out, we need more terms in order to write down the transformed conditional probability as a linear combination of past processes.

5.2 Second representation theorem

In this section, we prove a new representation theorem for functions of r finite variables. We start with the trivial finite-valued one-variable function and then extend the result to r-variable functions. The proof for the general case is non-trivial and is done again by induction.

Lemma 5.1 *Suppose* $f:M\to\mathbb{R}$, $M\subset\mathbb{R}$ *being finite of cardinality* c. *Let* $d=c-1$. *Then* f *has a unique representation of the form*

$$f(x)=\sum_{0\le i\le d}\alpha_i x^i,\ \forall x\in M$$

Remark.

7. The lemma states that, if we consider the vector space $V = \{f : M \to \mathbb{R}\}$, then the monomial functions $\{p_i\}_{0\le i\le d}$, where $p_i : M \to \mathbb{R},\ p_i(x) = x^i$ form a basis for V.

Proof First note that the dimension of V is c. To show this, suppose $M = \{m_1, \cdots, m_c\}$ and consider the following isomorphism of vector spaces,

$$I : V \to \mathbb{R}^c$$
$$f \mapsto (f(m_1), \cdots, f(m_c)).$$

It only remains to show that $\{p_i\}_{0\le i\le d}$ is an independent set. To prove this suppose,

$$\sum_{0\le i\le d} \alpha_i x^i = 0, \quad \forall x \in M.$$

That would mean that the d-th degree polynomial $p(x) = \sum_{0\le i\le d} \alpha_i x^i$ has at least $c = d + 1$ disjoint roots which is greater than its degree. This contradicts the fundamental theorem of algebra. ■

Theorem 5.3 *Suppose M_i is a finite subset of $\mathbb{R}$ with $|M_i| = c_i,\ i = 1, 2, \cdots, r$. Let $d_i = c_i - 1, M = \prod_{i=1,\cdots,r} M_i$ and consider the vector space of functions over $\mathbb{R}$, $V = \{f : M \to \mathbb{R}\}$ with the function addition as the addition operation of the vector space and the scalar product of a real number to the function as the scalar product of the vector space. Then this vector space is of dimension $C = \prod_{i=1,\cdots,r} c_i$ and $\{x_1^{i_1} \cdots x_r^{i_r}\}_{0\le i_1\le d_1,\cdots,0\le i_r\le d_r}$ forms a basis for it.*

Proof To show that the dimension of the vector space is C, suppose $M = \{m_1, \cdots, m_c\}$ and consider following the isomorphism of vector spaces:

$$I : V \to \mathbb{R}^C,$$

$$f \mapsto (f(m_1), \cdots, f(m_C)).$$

To show that $\{x_1^{i_1} \cdots x_r^{i_r}\}_{0\le i_1\le d_i,\cdots,0\le i_r\le d_r}$ forms a basis, we only need to show that it is an independent collection since there are exactly C elements in it. We proceed by induction on r. The case $r = 1$ was shown in the above lemma. Suppose we have shown the result for $r - 1$ and we want to show it for r. Assume a linear

combination of the basis is equal to zero. We can arrange the terms based on powers of x_r:

$$p_0(x_1, \cdots, x_{r-1}) + x_r p_1(x_1, \cdots, x_{r-1}) + \cdots + x_r^{d_r} p_d(x_1, \cdots, x_{r-1}) = 0, \qquad (7)$$
$$\forall (x_1, \cdots, x_r) \in M_1 \times \cdots \times M_r.$$

Fix the values of $x'_1, \cdots, x'_{r-1} \in M_1 \times \cdots \times M_{r-1}$. Then Equation (7) is zero for c_r values of x_r. Hence by Lemma 5.1, all the coefficients:

$$p_0(x'_1, \cdots, x'_{r-1}), p_1(x'_1, \cdots, x'_{r-1}), \cdots, p_d(x'_1, \cdots, x'_{r-1}),$$

are zero and we conclude:

$$p_0(x_1, \cdots, x_{r-1}) = 0, p_1(x_1, \cdots, x_{r-1}) = 0, \cdots, p_d(x_1, \cdots, x_{r-1}) = 0,$$
$$\forall (x_1, \cdots, x_{r-1}) \in M_1 \times \cdots \times M_{r-1}.$$

Again by the induction assumption all the coefficients in these polynomials are zero. Hence, all the coefficients in the original linear combination in Equation (7) are zero. ■

Corollary 5.3 *Suppose X_t is a categorical stochastic process, where X_t takes values in M_t, $|M_t| = c_t = d_t + 1 < \infty$. Also assume that the conditional probability*

$$P(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0),$$

is well-defined and in (0,1). Fix $m_t^1 \in M_t$. Let $g : \mathbb{R} \to \mathbb{R}^+$ be a bijective transformation, then there are unique parameters $\{\alpha^t_{i_0,\cdots,i_t}\}_{t\in\mathbb{N}, 0\le i_0\le d_t-1, 0\le i_1\le d_{t-1}, 0\le i_2\le d_{t-2}, \cdots, 0\le i_t\le d_0}$ such that

$$P(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0) = P_t(x_0, \cdots, x_t),$$

where

$$P_t(x_0, \cdots, x_{t-1}, m_1^t) = \frac{1}{1 + \sum_{y \in M - \{m_1^t\}} h_t(y)}, \qquad (8)$$

$$P_t(x_0, \cdots, x_{t-1}, x) = \frac{h(x)}{1 + \sum_{y \in M - \{m_1\}} h_t(y)}, x \neq m_1^t \in M_t, \qquad (9)$$

for $h_t(x_0, \cdots, x_t) = g \circ g_t(x_0, \cdots, x_{t-1}, x_t)$ and

$$g_t(x_0, \cdots, x_{t-1}, x_t) = \sum_{0\le i_0\le d_t-1, 0\le i_1\le d_{t-1}, \cdots, 0\le i_t\le d_0} \alpha^t_{i_0,\cdots,i_t} x_{t-0}^{i_0} \cdots x_{t-t}^{i_t},$$
$$(x_0, \cdots, x_t) \in M_0 \times \cdots \times M_{t-1} \times M'_t.$$

On the other hand any set of parameters $\alpha^t_{i_0,\cdots,i_t}$ gives rise to a unique stochastic process with the above equations.

Corollary 5.4 *Suppose that $\{X_t\}$ is an r-th order Markov chain where X_t takes values in M_t a finite subset of real numbers, $|M_t| = c_t = d_t + 1 < \infty$, the conditional probability*

$$P(X_t = x_t | X_{t-1}, \cdots, X_0),$$

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

$$g_t(x_t, \cdots, x_0) = g^{-1}\{\frac{P(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}{P(X_t = m_t^1 | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}\},$$

is a function of $t+1$ variables for $t < r$, $(x_t, \cdots, x_0)$ and is a function of $r+1$ variables,$(x_t, \cdots, x_{t-r})$, for $t > r$. Hence 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

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

such that for $t < r$

$$g^{-1}\{\frac{P(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}{P(X_t = m_t^1 | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}\} = \sum_{0 \le i_0 \le d_t - 1, 0 \le i_1 \le d_{t-1}, \cdots, 0 \le i_t \le d_0} \alpha^t_{i_0,\cdots,i_t} x_{t-0}^{i_0} \cdots x_{t-t}^{i_t},$$
$$(x_0, \cdots, x_t) \in M_0 \times \cdots M_{t-1} \times M_t',$$

and for $t \ge r$

$$g^{-1}\{\frac{P(X_t = x_t | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}{P(X_t = m_t^1 | X_{t-1} = x_{t-1}, \cdots, X_0 = x_0)}\} = \sum_{0 \le i_0 \le d_t - 1, 0 \le i_1 \le d_{t-1}, \cdots, 0 \le i_r \le d_{t-r}} \alpha^t_{i_0,\cdots,i_r} x_{t-0}^{i_0} \cdots x_{t-r}^{i_r}$$
$$(x_0, \cdots, x_t) \in M_0 \times \cdots M_{t-1} \times M_t'.$$

In the case of homogenous Markov chains the $\alpha^t_{i_1,\cdots,i_r}$ do not depend on t for $t > r$.

One might question the usefulness of such a representation. After all we have exactly as many parameters in the model as the values of the original function. In the following, we explain the importance of linear representations of such functions.

1. A vast amount of theory has been developed to deal with linear models. Generalized linear models in the case of independent sequence of random variables is a powerful tool. As we will see in sequel, these ideas can be imported into time series using the concept of partial likelihood.

2. Although we have as many parameters in the model as the values of the original function, the representation gives us a convenient framework for modeling, in particular for making various model reductions by omitting some terms or assuming certain coefficients are equal.

3. Although this is a representation for stationary r-th order Markov chains (or representation for arbitrary locally r-th order chains at time t), this representation allows us to accommodate other explanatory variables simply by as additive linear terms and extend the model to the non-stationary cases. This cannot be done in the same way if we try to model the original values of the function.

Example 5.1 *As an example consider a categorical response variable Y and r categorical explanatory variables*

$$X_1, \cdots, X_r,$$

are given. Suppose the X_i takes values in the M_i which include 0. Our purpose is to model Y based on $X_1, \cdots, X_r$. In order to do that, we consider the conditional probability

$$P(Y = y | X_1 = x_1, \cdots, X_r = x_r).$$

Again, we assume that the conditional probability is well-defined everywhere and takes values in $(0, 1)$. The above theorem shows that after applying a transformation the conditional probability can be written as a linear combination of multiples of powers of the X_i.

Although, the theorem above shows the form of the conditional probability in general and paves the way to the estimation of the conditional probabilities by estimating the parameters, the large number of parameters makes this a challenging task which might be impractical in some cases. In the next section, we introduce some classes of r variable functions that can be useful for some applications.

5.3 Special cases of functions of r finite variables

The first class of functions we introduce are obtained by power restrictions. We simply assume that g_r can be represented only by powers less than k. Suppose X_t takes values in $0, 1, \cdots, c_t - 1$. Then for a k-restricted power model the g_t, $t > r$ is given by:

$$\sum_{0 \le i_1 \le d_1, \cdots, 0 \le i_r \le d_r, \sum_j i_j \le k} \alpha_{i_1, \cdots, i_r} X_{t-1}^{i_1} \cdots X_{t-r}^{i_r}.$$

In particular, we can let $k = 1$ and get

$$\beta_0 + \sum_i \beta_i X_{t-i}.$$

This is useful especially for binary Markov chains.

The second class of functions are useful in the case when relationships exist between the states in terms of a semi-metric d. Suppose $\{X_t\}$ is an r-th order Markov chain and X_t takes values in the same finite set $M = \{1, \cdots, m\}$. Also let

$$d : M \times M \to \mathbb{R},$$

be a semi-metric being a mapping on M that satisfies the following conditions:

$$\begin{aligned} d &\geq 0; \\ d(x,z) &\leq d(x,y) + d(y,z); \\ d(x,x) &= 0. \end{aligned}$$

Then we introduce the following model:

$$g^{-1} \circ \frac{P(X_t = j|X_{t-1}, \cdots, X_{t-r})}{P(X_t = 1|X_{t-1}, \cdots, X_{t-r})} = \alpha_{0,j} + \sum_{i=1}^{k} \alpha_{i,j} d(j, X_{t-i})$$

for $j = 2, \cdots, m$. For this model

$$P(X_t = 1|X_{t-1}, \cdots, X_{t-r}) = 1 - \sum_{j=1,\cdots,c_{t-1}} P(X_t = j|X_{t-1}, \cdots, X_{t-r}).$$

Finally, we introduce a simple class for the binary Markov chain of order r. For any bijective transformation $g : \mathbb{R} \to \mathbb{R}^+$

$$g^{-1} \circ \frac{P(X_t = 1|X_{t-1}, \cdots, X_{t-r})}{P(X_t = 0|X_{t-1}, \cdots, X_{t-r})} = \alpha_0 + \alpha_1 N_{t-1},$$

where $N_{t-1} = \sum_{j=1}^{r} X_{t-j}$. For example in the 0-1 precipitation process example seen in the Introduction, N_{t-1} counts the number of the days out of r days before today that had some precipitation.

6 Generalized linear models for time series

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. What follows in this section is a summary of the first chapter in Kedem and Fokianos [5], which we have included for completeness.

Definition 6.1 *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 given by*

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

Example 6.1 *As an example, suppose Y_t represents the 0-1 PN process in Calgary, while MT_t denotes the maximum daily temperature process. We can define $\mathcal{F}_t$ as follows:*

1. *$\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots\}$. In this case, we are assuming the information available to us is the value of the process on each of the previous days.*

2. *$\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots MT_{t-1}, MT_{t-2}, \cdots\}$. In this case, we are assuming we have all the information regarding the 0-1 process of precipitation and maximum temperature for previous days.*

3. *$\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots MT_t, MT_{t-1}, MT_{t-2}, \cdots\}$. In this case, we add to the information in 2 the knowledge of today's maximum temperature.*

The vector θ that maximizes the above equation is called the maximum partial likelihood (MPLE). Wong [7] has studied its properties. Its consistency, asymptotic normality and efficiency can be shown under certain regularity conditions.

In this report, we are mainly interested in the case: $\mathcal{F}_t = \sigma\{Y_{t-1}, Y_{t-2}, \cdots\}$. We assume that the information $\mathcal{F}_t$ is given as a vector of random variables and denote it by Z_t, which we call the covariate process:

$$Z_t = (Z_{t1}, \cdots, Z_{tp})'.$$

Z_t might also include the past values of responses $Y_t, Y_{t-1}, \cdots$.

Let $\mu_t = E[Y_t|\mathcal{F}_{t-1}]$, be the conditional expectation of the response given the information we have up to the time t.

Kedem and Fdokianos in [5] address time series following generalized linear models satisfying certain conditions about the so-called random and systematic components:

- Random components: For $t = 1, 2, \cdots, N$

$$f(y_t; \theta_t, \phi|\mathcal{F}_{t-1}) = \exp\{\frac{y_t\theta_t - b(\theta_t)}{a_t(\phi)} + c(y_t; \phi)\}.$$

- The parametric function $\alpha_t(\phi)$ is of the form ϕ/w_t, where ϕ is the dispersion parameter, and w_t is a known parameter called "weight parameter". The parameter θ_t is called the natural parameter.

- Systematic components: For $t = 1, 2, \cdots, N$,

$$g(\mu_t) = \eta_t = \sum_{j=1}^{p} \beta_j Z_{(t-1)j} = Z'_{t-1}\beta,$$

for some known monotone function g called the link function.

Example 6.2 *Binary time series: As an example consider* $\{Y_t\}$*, a binary time series. Let us denote by* π_t *the probability of success give* $\mathcal{F}_{t-1}$*. Then for* $t = 1, 2, \cdots, N$*,*

$$f(y_t; \theta_t, \phi|\mathcal{F}_{t-1}) = \exp(y_t \log(\frac{\pi_t}{1-\pi_t}) + \log(1-\pi_t))$$

with $E[Y_t|\mathcal{F}_{t-1}] = \pi_t$, $b(\theta_t) = -\log(1-\pi_t) = \log(1+\exp(\theta_t))$, $V(\pi_t) = \pi_t(1-\pi_t)$, $\phi = 1$, and $w_t = 1$.
The canonical link gives rise to the so–called "logistic model":

$$g(\pi_t) = \theta_t(\pi_t) = \log(\frac{\pi_t}{1-\pi_t}) = \eta_t = Z'_{t-1}\beta.$$

In order to study the asymptotic behavior of the maximum likelihood estimator, we consider the conditional information matrix. To establish large sample properties, the stability of the conditional information matrix and the central limit theorem for martingales are required. Proofs may be found in Kedem and Fokianos [5].

Inference for partial likelihood

The definitions of partial likelihood and exponential family of distributions imply that the log partial likelihood is given by

$$l(\beta) = \sum_{t=1}^{N} \log f(y_t; \theta_t, \phi|\mathcal{F}_{t-1}) =$$

$$\sum_{t=1}^{N}\{\frac{y_t\theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t, \phi)\} = \sum_{t=1}^{N}\{\frac{y_t u(z'_{t-1}\beta) - b(u(z'_{t-1}))}{\alpha_t(\phi)} + c(y_t, \phi)\} = \sum_{t=1}^{N} l_t,$$

where $u(.) = (g \circ \mu(.))^{-1} = \mu^{-1}(g^{-1}(.))$, so that $\theta_t = u(z_{t-1}\beta)$. We introduce the notation,

$$\nabla = (\frac{\partial}{\partial\beta_1}, \cdots, \frac{\partial}{\partial\beta_p})'$$

and call $\nabla l(\beta)$ the partial score. To compute the gradient, we can use the chain rule in the following manner

$$\frac{\partial l_t}{\partial \beta_j} = \frac{\partial l_t}{\partial \beta_j}\frac{\partial \theta_t}{\partial \mu_t}\frac{\partial \mu_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial \beta_j}.$$

Some algebra shows

$$S_N(\beta) = \nabla l(\beta) = \sum_{t=1}^{N} Z_{(t-1)}\frac{\partial \mu_t}{\partial \eta_t}\frac{Y_t - \mu_t(\beta)}{\sigma_t^2(\beta)},$$

where, $\sigma_t^2(\beta) = Var[Y_t|\mathcal{F}_{t-1}]$. The partial score process is defined from the partial sums as

$$S_t(\beta) = \nabla l(\beta) = \sum_{s=1}^{t} Z_{(s-1)}\frac{\partial \mu_s}{\partial \eta_s}\frac{Y_s - \mu_s(\beta)}{\sigma_s^2(\beta)}.$$

One can show the terms in the above sums to be orthogonal:

$$E[Z_{(t-1)}\frac{\partial \mu_t}{\partial \eta_t}\frac{Y_t - \mu_t(\beta)}{\sigma_t^2(\beta)}Z_{(s-1)}\frac{\partial \mu_s}{\partial \eta_s}\frac{Y_s - \mu_s(\beta)}{\sigma_s^2(\beta)}] = 0,\ s < t.$$

Also, $E[S_N(\beta) = 0]$.

The cumulative information matrix is defined by

$$G_N(\beta) = \sum_{t=1}^{N} Cov[Z_{(t-1)}\frac{\partial \mu_t}{\partial \eta_t}\frac{Y_t - \mu_t(\beta)}{\sigma_t^2(\beta)}|\mathcal{F}_{t-1}].$$

The unconditional information matrix is simply

$$Cov(S_N(\beta)) = F_N(\beta) = E[G_N(\beta)].$$

Next let

$$H_N(\beta) = -\nabla\nabla' l(\beta).$$

Kedem and Fokianso ([5]) show that

$$H_N(\beta) = G_N(\beta) - R_N(\beta),$$

where

$$R_N(\beta) = \frac{1}{\alpha_t(\phi)}\sum_{t=1}^{N} Z_{t-1}d_t(\beta)Z'_{t-1}(Y_t - \mu_t(\beta)),$$

and $d_t(\beta) = [\partial^2 u(\eta_t)/\partial \eta_t^2]$.
S_t satisfies the martingale property:

$$E[S_{t+1}(\beta)|\mathcal{F}_{t-1}] = S_t(\beta).$$

To prove the consistency and other properties of the estimators, we need:
Assumption A:
A1. The true parameter β belongs to an open set $B \subset \mathbb{R}$.
A2. The covariate vector Z_t almost surely lies in a non random compact set Γ of $\mathbb{R}^p$, such that $P[\sum_{t=1}^{N} Z_{t-1}Z'_{t-1} > 0] = 1$. In addition, $Z'_{t-1}\beta$ lies almost surely in the domain H of the inverse link function $h = g^{-1}$ for all $Z_{t-1} \in \Gamma$ and $\beta \in B$.
A3. The inverse link function h-defined in (A2) is twice continuously differentiable and $|\partial h(\lambda)/\partial\lambda| \neq 0$.
A4. There is a probability measure ν on $\mathbb{R}^p$ such that $\int_{\mathbb{R}^p} zz'\nu(dz)$ is positive definite, and such that for Borel sets $A \subset \mathbb{R}^p$,

$$\frac{1}{N}\sum_{t=1}^{N} I_{[Z_{t-1}\in A]} \to \nu(A).$$

Theorem 6.1 *Under assumption A the maximum likelihood estimator is almost surely unique for all sufficiently large N, and*

1. *the estimator is consistent and asymptotically normal,*

$$\hat{\beta} \xrightarrow{p} \beta$$

in probability, and

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N_p(0, G^{-1}(\beta)),$$

in distribution as $N \to \infty$, for some matrix G.

2. *The following limit holds in probability, as $N \to \infty$:*

$$\sqrt{N}(\hat{\beta} - \beta) - \frac{1}{\sqrt{N}}G^{-1}(\beta)S_N(\beta) \xrightarrow{p} 0.$$

7 Simulation studies

This section presents the results of some simulation studies about the partial likelihood applied to categorical r-th order Markov chains. We also investigate the performance of the BIC to pick the appropriate ("true") model. In particular, we generate samples from a seasonal Markov chain X_t where,

$$Z_{t-1} = (1, X_{t-1}, cos(\omega t)),\ \omega = \frac{2\pi}{366}.$$

We consider this Markov chain over 5 years between 2000 and 2005 and assume

$$logit\{P(X_t = 1|Z_{t-1})\} = \beta' Z_{t-1},$$

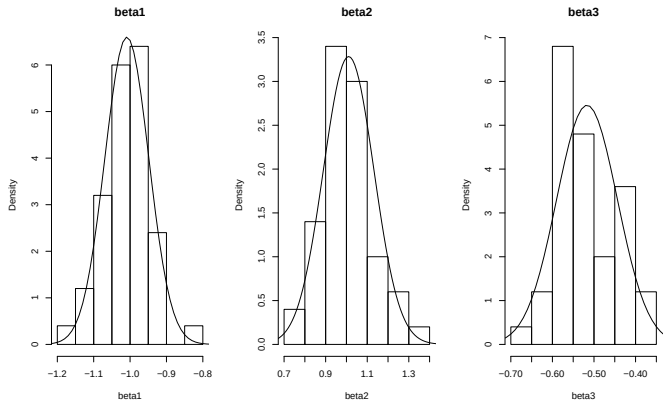

Figure 1: The distribution of parameter estimates for the model with the covariate process $Z_{t-1} = (1, X_{t-1}, cos(\omega t))$ and parameters $(\beta_1 = -1, \beta_2 = 1, \beta_3 = -0.5)$.

where $\beta = (-1, 1, -0.5)$.

To generate samples for this chain, we need an initial value of the past two states, which we take it to be $(1, 1)$. We denote the process X_{t-k} by X^k for simplicity.

To check the performance of the partial likelihood and estimates of the variance using G_N, we generate 50 chains with this initial value and then compare the parameter estimates with the true parameters. We also compare the theoretical variances with the experimental variances.

			simulated sd			theoretical sd		
$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$sd(\hat{\beta}_1)$	$sd(\hat{\beta}_2)$	$sd(\hat{\beta}_3)$	$sd(\hat{\beta}_1)$	$sd(\hat{\beta}_2)$	$sd(\hat{\beta}_3)$
-0.985	1.02	-0.415	0.067	0.104	0.0730	0.061	0.121	0.0731

Table 1: The estimated parameters for the model $Z_{t-1} = (1, X_{t-1}, cos(\omega t))$ with parameters $\beta = (-1, 1, -0.5)$. The standard deviation for the parameters is computed once using G_n and once using the generated samples.

In Kedem and Fokianos [5] other simulation studies have been done to check the validity of this method.

To check the normality of the parameter estimates, we plot the three parameter estimates histograms in Figure 1.

Next we check the performance of the BIC criterion in picking the optimal ("true") model. We use the same model as above and then compute the BIC for a few models to see if BIC picks the right one. We denote X_{t-k} by X^k and $cos(\omega t)$ by COS for simplicity. For an assessment, we simulate a few other chains.

Model: Z_{t-1}	BIC	parameter estimates
1	2380.0	-0.605
$1, X^1$	2267.12	(-1.03, 1.11)
$1, X^1, X^2$	2273.75	(-1.064,1.091,0.101)
$1, X^1, COS$	2217.73	(-1.00,0.970,-0.558)
$1, X^1, SIN$	2274.49	(-1.037,1.117,0.026)
$1, X^1, COS, SIN$	2225.087	(-1.00,0.970,-0.559,0.028)
$1, X^1, X^2, X^1X^2$	2281.142	(-1.055,1.0615,0.0647,0.077)
$1, X^1, X^2, X^1X^2, COS$	2232.424	(-0.985,0.943,-0.0870,0.0915,-0.564)
$1, X^1, X^2, X^1X^2, COS, SIN$	2239.834	(-0.981,0.957,-0.0946,0.0723,-0.575,0.0232)

Table 2: BIC values for several models competing for the role of the true model, where $Z_{t-1} = (1, X^1, COS),\ \beta = (-1, 1, -0.5)$.

As we see in Table 2, the true model has the smallest BIC so BIC performs well in this case. Also note that models which include the true model have accurate estimates for the parameters associated with $1, X^1, COS$, while giving very small magnitude for other parameters associated with other covariates in the full but not true model.

Model: Z_{t-1}	BIC	parameter estimates
1	2537.353	0.0799
$1, X^1$	2329.58	(-0.649,1.417)
$1, X^1, X^2$	2245.584	(-1.022,1.144,0.998)
$1, X^1, COS$	2265.96	(-0.553,1.236,-0.617)
$1, X^1, SIN$	2336.71	(-0.648,1.415,-0.0433)
$1, X^1, COS, SIN$	2273.01	(-0.552,1.235,-0.617,-0.0480)
$1, X^1, X^2, X^1X^2$	2251.32	(-1.08,1.287,1.140,-0.278)
$1, X^1, X^2, X^1X^2, COS$	2213.706	(-0.936,1.11,0.966,-0.175,-0.511)
$1, X^1, X^2, X^1X^2, COS, SIN$	2221.275	(-0.927,1.101,0.940,-0.160,-0.549,-0.0441)
$1, X^1, X^2, COS$	2206.865	(-0.899,1.0263,0.875,-0.515)

Table 3: BIC values for several models competing for the role of true model given by $Z_{t-1} = (1, X^1, X^2, COS),\ \beta = (-1, 1, 1, -0.5)$.

Table 3 presents the true model in the last row. Ignore that row for a moment. The smallest "BIC" corresponds to $1, X^1, X^2, X^1X^2, COS$, which has an component X^1X^2 added to the true model. However, the coefficients of this model are very close to the true model and the coefficient for X^1X^2 is relatively small in magnitude. The true model has the smallest BIC again and the parameter estimates are close to the correct values.

8 Concluding remarks

In summary, this report shows that a categorical discrete-time stochastic process can be represented using a small number of ascending joint distributions

$$p_0(x_0), p_1(x_0, x_1), p_2(x_0, x_1, x_2), \cdots .$$

We showed that a categorical discrete-time stochastic process can be represented using the conditional probabilities $P_0(X_0), P_1(X_1|X_0), P_2(X_2|X_0, X_1), \cdots$. A parametric form was found for the conditional probability distribution of categorical discrete-time stochastic processes. The parameters can be estimated for stationary binary Markov chains using partial likelihood.

References

[1] Anderson, T. W. and Goodman, L. A. (1957). Statistical inference about markov chains. *Ann. Math. Statist.*, 89–110.

[2] Besag, J. (1974). Spatial interactions and the statistical analysis of lattice systems. *JRSS–B*, 192–225.

[3] Breiman, L. (1992). *Probability*. SIAM.

[4] Cressie, N. and Subash, L. (1992). New models for markov random. *J Applied Prob.*, 877–884.

[5] Kedem, B. and Fokianos, K. (2002) *regression models for time series analysis.* Wiley Series in Probability and Statistics.

[6] Öksendal, B. (2003). *Stochastic differential equations: and introduction with applications*. Berlin: Springer.

[7] Wong, W. (1986). Theory of partial likelihood. *Ann Statist.*, 1, 88-123.