Tractable Bayesian Inference of Time-Series Dependence Structure

Michael R. Siracusa
CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139

John W. Fisher III
CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We consider the problem of Bayesian inference of graphical structure describing the interactions among multiple vector time-series. A directed temporal interaction model is presented which assumes a fixed dependence structure among time-series. Using a conjugate prior over this model's structure and parameters, we focus our attention on characterizing the exact posterior uncertainty in the structure given data. The model is extended via the introduction of a dynamically evolving latent variable which indexes dependence structures over time. Performing inference using this model yields promising results when analyzing the interaction of multiple tracked moving objects.

1 INTRODUCTION

We consider the problem of inference of evolving dependence *structures* of multiple vector time-series. Specifically, we develop a Bayesian inference method over directed models as a means of examining the influence of time-series on each other. Influence between time-series is characterized by dynamically changing directed models in which the graphical structure (*i.e.* the composition of edges) is of primary interest and the underlying parameters are treated as nuisances.

The general problem of structure learning has garnered a great deal of attention in the past two decades, cf. (Heckerman, 1995). There are two primary problems within this area: learning better predictive models and structure discovery. Here we focus on the latter as a tool for data analysis. Bayesian inference over *structure* is complicated by two factors. First, priors on parameters must be chosen such that they can be efficiently marginalized. Second, Bayesian structure inference generally requires reasoning over a super-exponential number of graphical structures within a model class. These two factors lead to difficulty in evaluating the partition function, precluding exact evaluation of event probabilities over the posterior distribution on graphical structures. Thus, much of research in this area, with notable exceptions discussed herein, focus on obtaining point estimates or approximations to the exact posterior. While these methods allow one to approximate certain event probabilities, it is highly desirable to compute exact quantities to characterize posterior uncertainty when possible (Friedman and Koller, 2003; Koivisto et al., 2004).

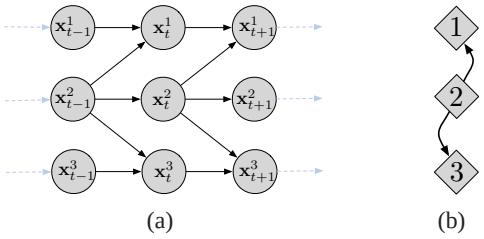


Figure 1: Example TIM(1) (a) and corresponding interaction graph (b)

Directed models are of particular interest when analyzing time-series owing to the assumption of temporal causality. That is, when performing inference over directed edges it is desirable to preclude edges which predict past observations from future observations. This provides a strict temporal ordering which we exploit to define a temporal interaction model. Figure 1(a) illustrates the structure of one such model.

A conjugate prior on the directed structure and parameters of our temporal interaction model is presented which allows the set of all directed structures to be reasoned over in exponential-time. Furthermore, by imposing simple local or global structural constraints we show that one can reduce the exponential-time complexity to polynomial-time complexity for reasoning over a still super-exponential number

Appearing in Proceedings of the 12^{th} International Confe-rence on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

of candidate structures. Specifically we focus on bounded in-degree structures with directed trees and forests being special cases with global constraints. These constraints yield tractable Bayesian inference over directed structures, allowing exact calculation of the partition function as well as additional marginal event probabilities. The method we present for reasoning about a fixed structure is closely related to that of (Friedman and Koller, 2003; Koivisto et al., 2004), but extended to the analysis of time-series for which the strict temporal ordering provides a computational advantage.

Furthermore, we augment our model via the introduction of an additional dynamically evolving latent variable. This variable indexes structure, allowing switching between a finite set of temporal interaction models over time. Using this model we present empirical results on the task of analyzing the interaction of multiple tracked moving objects. We make no assumptions as to the existence of a "correct structure." Our problem of interest is that of quantifying uncertainty in the interaction among the time-series.

Spectral methods for point estimation of graphical models over stationary time-series is considered in (Bach and Jordan, 2004). The related work of (Bilmes, 2000; Kirshner et al., 2004; Xuan and Murphy, 2007) consider alternative models in which structure is changing over time. The methodology here differs in that while we restrict ourselves to directed models, we marginalize over parameters and calculate an exact posterior over structures. These models as well as the model proposed here fall into the general class of multinets (Geiger and Heckerman, 1996) or Contingent Bayesian Networks (Milch et al., 2005).

2 TEMPORAL INTERACTION MODEL

Here, we present a temporal interaction model (TIM) for multivariate time series. This model is an rth order Markov model with additional structural constraints. We begin by introducing some notation for the purpose of explicitly denoting individual time-series, past values of individual time-series as well as sets thereof.

Consider N time-series and let $\mathbf{x}_t^v$ be a random variable representing the value of the vth time-series at time t. The r past values of time-series v is defined as $\widetilde{\mathbf{x}}_t^v$. That is, $\widetilde{\mathbf{x}}_t^v$ represents $\mathbf{x}_{t-1}^v, \ldots, \mathbf{x}_{t-r}^v$. As we will be explicit on the temporal model order, r is suppressed in the notation, $\widetilde{\mathbf{x}}_t^v$, for brevity. Furthermore, the vector $\widetilde{\mathbf{x}}_t^{\mathbf{S}}$ indexed by set $\mathbf{S}$ is $\widetilde{\mathbf{x}}_t^{S(1)}$,...,$\widetilde{\mathbf{x}}_t^{S(m)}$ stacked in a vector where $|\mathbf{S}| = m$ (e.g. $\widetilde{\mathbf{x}}_t^{v,u}$ is $\widetilde{\mathbf{x}}_t^v$ and $\widetilde{\mathbf{x}}_t^u$ stacked). The random variables denoting the present and past of all time-series at time t are defined as $\mathbf{X}_t$ and $\widetilde{\mathbf{X}}_t$ respectively. Multiple time points can be indexed by a vector $\mathbf{t} = [t_1, t_2, ..., t_T]$ such that $\mathbf{X}_\mathbf{t} = \left[\mathbf{X}_{t_1}, ..., \mathbf{X}_{t_T}\right]$. Note that $\widetilde{\mathbf{X}}_\mathbf{t}$ can be formed from $\mathbf{X}_\mathbf{t}$ and a set $\mathcal{C}$ containing past values not available in $\mathbf{X}_\mathbf{t}$ (initial conditions).

Given a directed structure $\bar{E}$, a set of parameters Θ and $\mathcal{C}$, an rth order temporal interaction model TIM(r) is Markov:

$$p\left(\mathbf{X}_\mathbf{t}|\bar{E}, \Theta, \mathcal{C}\right) = \prod_{i=1}^{T} p\left(\mathbf{X}_{\mathbf{t}_i}|\widetilde{\mathbf{X}}_{t_i}, \bar{E}, \Theta\right). \quad (1)$$

In order to simplify notation we will drop the $\mathcal{C}$ when it is clear from the context. $\bar{E}$ is a directed structure on N nodes/vertices defining the factorization of a TIM(r) model:

$$p\left(\mathbf{X}_\mathbf{t}|\bar{E}, \Theta\right) = \prod_{i=1}^{T}\prod_{v=1}^{N} p\left(\mathbf{x}_{t_i}^v|\widetilde{\mathbf{x}}_{t_i}^{v,\mathbf{pa}\left(v,\bar{E}\right)}, \Theta_{v|\mathbf{pa}\left(v,\bar{E}\right)}\right) \quad (2)$$

where $\mathbf{pa}\left(v, \bar{E}\right)$ returns the parents of vertex v given the structure $\bar{E}$. We will drop the $\bar{E}$ and use $\mathbf{pa}\left(v\right)$ when it is clear from the context. The v-th time-series at time t is dependent on its own past $\widetilde{\mathbf{x}}_t^v$ as well as the past of its parent set $\mathbf{S} = \mathbf{pa}\left(v\right)$, $\widetilde{\mathbf{x}}_t^{\mathbf{S}}$. Note that we use $\Theta_{v|\mathbf{S}}$ rather than the more explicit notation $\Theta_{v|v,\mathbf{S}}$ to represent the parameters of this relationship for brevity.

Figure 1(a) illustrates a TIM(1) for $N = 3$ time-series with $\bar{E}$ containing two edges; one from 2 to to 1, and one from 2 to 3. Here $\mathbf{pa}\left(1\right) = \mathbf{pa}\left(3\right) = 2$ and $\mathbf{pa}\left(2\right) = \emptyset$. Figure 1(b) shows an alternative and more compact view for this model in which a single node represents a time-series over all time. We refer to this as the *interaction graph* and use diamond shaped nodes to emphasize it is not meant to be interpreted as a directed Bayesian network, though there is a one to one mapping between these graphs. A directed edge from u to v in the interaction graph implies a directed edge from $\widetilde{\mathbf{x}}_t^u$ and $\mathbf{x}_t^v$ in the TIM. Note that the TIM(1) in Figure 1(a) has an interaction graph that happens to be a directed tree (arborescence). However, in general the interaction graph need not be a tree or even acyclic. In other words, the space of interaction graphs is the space of all directed graphs. Cycles in the interaction graph do not result in cycles in the TIM(r) due to the strict temporal ordering assumptions.

3 CONJUGATE PRIOR

We adopt a prior on the structure and parameters similar to those presented in (Meila and Jaakkola, 2006; Friedman and Koller, 2003), using the factorization

$$p_0\left(\bar{E}, \Theta\right) = p_0\left(\bar{E}\right) p_0\left(\Theta|\bar{E}\right). \quad (3)$$

The parameters are assumed to be independent and *modular* given a structure $\bar{E}$ and hyper-parameters Γ. That is, they factorize according to the edges in $\bar{E}$:

$$p_0\left(\Theta|\bar{E}\right) = \prod_{v=1}^{N} p_0\left(\Theta_{v|\mathbf{pa}(v)}|\Gamma\right) \quad (4)$$

and $p_0\left(\Theta_{v|\mathbf{S}}|\Gamma\right)$ is the same for all structures $\bar{E}$ for which $\mathbf{S}$ is the parent set of v. Thus, for each time-series v one needs to specify a parameter prior for all potential parent sets. Given this finite set of priors for each v one is able to supply a full prior on parameters for any given structure.

We place a prior on directed structures which has the form:

$$p_0\left(\bar{E}\right) = \frac{1}{Z(\beta)} \prod_{v=1}^{N} \beta_{\mathbf{pa}(v),v} \tag{5}$$

where the partition function $Z(\beta)$ ensures proper normalization. Each scalar hyper-parameter $\beta_{\mathbf{S},v}$ can be interpreted as a weight on the parent set $\mathbf{S}$ for v and the prior is simply a proportional to the product of these weights. Note that if all $\beta_{\mathbf{S},v}$ are set to 1, one obtains a uniform prior and $Z(\beta)$ is the number of possible structures. If all $\beta_{\mathbf{S},v}$ are set proportional to the size of $\mathbf{S}$, $|\mathbf{S}|$, the prior will favor dense structures. Equivalently, sparse structures are favored by making $\beta_{\mathbf{S},v}$ inversely proportional to $|\mathbf{S}|$.

Let $\mathcal{D} = \{\mathbf{X}_1, ..., \mathbf{X}_T\}$ be a set of T complete observations. We use the notation $\mathcal{D}^{\mathbf{S}} = \{\mathbf{x}_1^{\mathbf{S}}, ..., \mathbf{x}_T^{\mathbf{S}}\}$ to denote observations of a set of time-series. $\mathcal{D}_t = \mathbf{X}_t$ and $\mathcal{D}_t^u = \mathbf{x}_t^u$. $\widetilde{\mathcal{D}} = \{\widetilde{\mathbf{X}}_1, ..., \widetilde{\mathbf{X}}_T\}$ can be formed using $\mathcal{D}$ and past information in $\mathcal{C}$. Given data $\mathcal{D}$, the posterior on parameters given structure has the form:

$$p\left(\Theta|\bar{E}, \mathcal{D}\right) = \prod_{v=1}^{N} p\left(\Theta_{v|\mathbf{pa}(v)}|\mathcal{D}^v, \mathcal{D}^{\mathbf{pa}(v)}, \Gamma\right) \tag{6}$$

That is, the prior on parameters given a structure is fully conjugate. If one choses a conjugate prior for each $\Theta_{v|\mathbf{S}}$ in Equation 4 the posterior has the same form and can be updated by sufficient statistic calculations from the data. In addition, the posterior on structure is:

$$p\left(\bar{E}|\mathcal{D}\right) = \frac{1}{Z(\beta \circ W)} \prod_{v=1}^{N} \beta_{\mathbf{pa}(v),v} W_{\mathbf{pa}(v),v} \tag{7}$$

where $\circ$ is an element wise / Hadamard product and

$$\begin{aligned} W_{\mathbf{S},v} &= p\left(\mathcal{D}^v|\widetilde{\mathcal{D}}^{v,\mathbf{S}}\right) \\ &= \int p\left(\mathcal{D}^v|\widetilde{\mathcal{D}}^{v,\mathbf{S}}, \Theta_{v|\mathbf{S}}\right) p_0\left(\Theta_{v|\mathbf{S}}|\Gamma\right) d\Theta_{v|\mathbf{S}} \end{aligned} \tag{8}$$

is simply the evidence for time-series v given the time-series of its parent set defined by $\mathbf{S}$. That is, the prior is updated by modifying β with a set of evidence weights W. The proof follows from the fact that the prior on structure factorizes in the same manner as Equation 2.

For continuous observations and a linear gaussian model with parameters $\Theta_{v|\mathbf{S}}$ one can choose $p_0\left(\Theta_{v|\mathbf{S}}|\Gamma\right)$ to be a matrix-normal-inverse-Wishart distribution with hyper-parameters Γ. This will yield efficient updates for Equation 6 and $W_{v|\mathbf{S}}$ will be the evaluation of a Matrix-T distribution (West and Harrison, 1997). Similarly for discrete observations one can use Dirichlet priors and have analytic forms for the evidence.

Note that there are 2^{N-1} possible parent sets for each v. This yields a super-exponential number of possible structures, 2^{N^2-N}. This suggests one may need to explicitly sum over a super exponential number of terms when calculating $Z(\beta)$. Fortunately this is not the case. When considering the set of all structures, $\mathcal{A}$, all combinations of parent sets are possible and independent of each other. Thus one can implicitly calculate this sum

$$\begin{aligned} Z(\beta) &= \sum_{\bar{E} \in \mathcal{A}} \prod_{v=1}^{N} \beta_{\mathbf{pa}(v),v} = \sum_{\mathbf{S}_1} \cdots \sum_{\mathbf{S}_N} \prod_{v=1}^{N} \beta_{\mathbf{S}_v,v} \\ &= \prod_{v=1}^{N} \sum_{\mathbf{S}_v} \beta_{\mathbf{S}_v,v} \triangleq \prod_{v=1}^{N} \gamma_v(\beta) \end{aligned} \tag{9}$$

as a product of N summations. Each $\gamma_v(\beta)$ is a summation over the 2^{N-1} possible parent sets, $\mathbf{S}_v$, for node v. That is, one can reason about all structures in exponential-time, $N2^{N-1}$, rather than super-exponential time. Furthermore, as will be discussed in the following sections, by imposing some constraints on the set of possible structures one can obtain $Z(\beta)$ in polynomial time while still considering a super-exponential number of candidate structures.

3.1 BOUNDED PARENT SET SIZE

One set of structures, $\mathcal{P}^M$, is obtained by restricting $\bar{E}$ such that each time-series/node has at most M parents in the interaction graph. This translates to each time-series at time t being dependent on at most M other time-series past in addition to its own past. This is a local constraint on the parent set of each time series. Consequently, each time-series's parent set is independent of all others and $Z(\beta)$ has the same form as Equation 9 with

$$\gamma_v(\beta) = \sum_{\mathbf{S}_v,\ s.t.\ \ |\mathbf{S}_v| \leq M} \beta_{\mathbf{S}_v,v} \tag{10}$$

which is now a sum of all possible parent sets of size less than or equal to M. One can bound the order of the summation by $\sum_{m=1}^{M} \binom{N-1}{m} \leq N^M$. Thus, only polynomial computation time, $O(N^{M+1})$, is required for calculating $Z(\beta)$ even though the total number of structures is still super exponential, $O(N^{MN})$.

3.2 DIRECTED TREES AND FORESTS

In the previous section a local constraint on parent set size was imposed. However, there may be situations in which a global structure constraint is desirable. For example, one may want to only consider structures which form a fully

connected interaction graph and/or contain no cycles. The set of directed trees, $\mathcal{T}$ is a subset of $\mathcal{P}^1$ in which the interaction graph is spanning and acyclic.

In the directed tree case, β is simply an $N \times N$ matrix and each hyper-parameter $\beta_{u,v\neq u}$ can be interpreted as a weight on the edge $u \rightarrow v$, and $\beta_{\emptyset,v} \triangleq \beta_{v,v}$ is a weight on a node being a root (having no parents). The edge set corresponding to the nonzero entries of β form a support graph. We assume this support graph is connected and contains at least one directed tree.

While there are N^{N-1} possible directed trees on N nodes, the Matrix Tree Theorem allows one to calculate $Z(\beta)$ in polynomial time. This theorem was used by (Meila and Jaakkola, 2006) for reasoning over undirected trees. The undirected version of theorem is a special case of the real valued directed version originally developed by (Kirchhoff, 1847). The theorem allows one to calculate the weighted sum over all directed trees rooted at r, $Z_r(\beta)$ via:

$$Z_r(\beta) = \sum_{\bar{E} \text{ rooted at } r} \prod_{u \rightarrow v} \beta_{u,v} = \mathit{Cof}_{r,r}\left(\bar{Q}(\beta)\right) \tag{11}$$

where $\bar{Q}(\beta)$ is the Kirchhoff matrix with its u, v entry defined as:

$$\bar{Q}_{uv}(\beta) = \begin{cases} -\beta_{u,v} & 1 \leq u \neq v \leq N \\ \sum_{u'=1}^{N} \beta_{u',v} - \beta_{u,u} & 1 \leq u = v \leq N \end{cases} \tag{12}$$

and $\mathit{Cof}_{i,j}(M)$ is the i,j cofactor of matrix M. $\mathit{Cof}_{i,j}\left(\bar{Q}(\beta)\right)$ is invariant to i and gives the sum over all weighted trees rooted at j. See (Tutte, 1984) for a proof.

By summing over all N possible roots one obtains

$$Z(\beta) = \sum_{r=1}^{N} \beta_{r,r} Z_r(\beta). \tag{13}$$

Thus, a straight forward implementation yields $O(N^4)$ time for calculating the partition function. However, as pointed out in (Koo et al., 2007), using the invariance of $\mathit{Cof}_{i,j}\left(\bar{Q}(\beta)\right)$ allows for $O(N^3)$ time computation of $Z(\beta)$. That is, $Z(\beta)$ can be calculated by replacing any row of the matrix $\bar{Q}(\beta)$ with $\left[\beta_{1,1}, ..., \beta_{N,N}\right]$ and taking its determinant.

The set of directed forests, $\mathcal{F}$, remove the fully connected/spanning assumption of $\mathcal{T}$ and allows $\bar{E}$ to have multiple roots. There are $(N+1)^{(N-1)}$ directed forests for N nodes, but $Z(\beta)$ can still be calculated in $O(N^3)$ time (Koo et al., 2007). Some intuition as to why this is true is that any directed forest can be turned into a directed tree by the addition of one virtual super root node which has no parents and connects to all the roots in the forest.

Note that while $\mathcal{T} \subset \mathcal{F} \subset \mathcal{P}^1$, both directed trees and forests require $O(N^3)$ computation even though the larger set $\mathcal{P}^1$ only requires $O(N^2)$. This is due to the imposed acyclic constraint which limits the parent set of one node based on the parent sets of others.

3.3 EVENT PROBABILITIES

The ability to compute the partition function and conjugacy of the prior allows one to calculate a wide variety of useful prior and/or posterior event probabilities. For example, the probability of a particular edge being present is:

$$p\left(I_{u\rightarrow v} = 1\right) = \mathbb{E}\left[I_{u\rightarrow v}\right] = 1 - \frac{Z(\beta^{-(u\rightarrow v)})}{Z(\beta)} \tag{14}$$

where $I_{u\rightarrow v}$ is an indicator variable that has value 1 when the edge $u \rightarrow v$ is present. $\beta^{-(u\rightarrow v)}$ is β with all elements involving edge from u to v set to zero. Similarly one can calculate the probability a time-series/node having no parents(root) or no children(leaf):

$$p\left(I_{v \text{ is a root}}\right) = \frac{Z(\beta^{-\mathbf{in}(v)})}{Z(\beta)}, \;\; p\left(I_{v \text{ is a leaf}}\right) = \frac{Z(\beta^{-\mathbf{out}(v)})}{Z(\beta)} \tag{15}$$

where $\mathbf{in}(v)$ and $\mathbf{out}(v)$ return the set of all possible edges in and out of time-series/node v respectively. $\beta^{-\mathbf{e}}$ indicates all elements of β which involve any edge in the set $\mathbf{e}$ that are zero.

The indicator variables used in the examples above can be expressed as a general multiplicative functions of the form $g(\bar{E}) = \prod_{v=1}^{N} g_{\mathbf{pa}(v),v}$. The expected value of a general multiplicative function can be calculated by:

$$\mathbb{E}\left[g(\bar{E})\right] = \frac{Z(\beta \circ g)}{Z(\beta)} \tag{16}$$

Note that variance or other higher order moments of multiplicative functions can also calculated in this manner (*e.g.* using $Z(\beta \circ g^2)$ in calculating posterior variance). In addition, one can calculate the expectation of additive functions of the form $f(\bar{E}) = \sum_{v=1}^{N} f_{\mathbf{pa}(v),v}$. For $\bar{E} \in \mathcal{A}$ or $\bar{E} \in \mathcal{P}^M$:

$$\mathbb{E}\left[f(\bar{E})\right] = \sum_{v=1}^{N} \frac{\gamma_v(\beta \circ f)}{\gamma_v(\beta)} \tag{17}$$

For directed trees, $\bar{E} \in \mathcal{T}$, $\mathbb{E}\left[f(\bar{E})\right] =$

$$\sum_{r=1}^{N} \frac{Z_r(\beta)}{Z(\beta)} \text{tr}\left(M_{r,r}\left(\bar{Q}(\beta \circ f)\right) M_{r,r}\left(\bar{Q}(\beta)\right)^{-1}\right) \tag{18}$$

where $M_{i,j}(Q)$ is the matrix Q with its ith row and jth column removed. A proof follows that of (Meila and Jaakkola, 2006) substituting in the directed tree partition function in place of the undirected version. A similar form is obtained for directed forests.

Additive functions allow calculation of quantities such as the expected number of children or parents of a particular

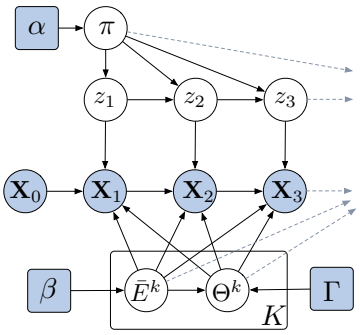


Figure 2: The structure of a STIM(1,K)

time-series/node. For example, by setting $f_{\mathbf{S},v}$ to $|\mathbf{S}|$ for a single v, $f(\bar{E})$ will count the number of parents node v has in structure $\bar{E}$.

3.4 SAMPLING STRUCTURE

Given data, we have shown how to calculate the posterior $p\left(\bar{E}|\mathcal{D}\right)$ and various event probabilities. Another important task is sampling from this posterior. When considering all structures or $\mathcal{P}^M$, sampling is done efficiently as a set of local sampling steps for each time-series/node parent set. That is the parent set $\mathbf{S}_v$ for node v is chosen with probability $\beta_{\mathbf{S},v}/\gamma v(\beta)$. Random sampling of directed spanning trees (and forests with a some modification) is a well studied problem. This paper uses Wilson's random walk based algorithm (Wilson, 1996).

4 SWITCHING INTERACTION MODELS

The TIM(r) model assumes a single $\bar{E}$ and parameters Θ over all time. Here, we introduce a switching temporal interaction model, STIM(r,K) which allows structure to change over time. Let z_t be a hidden state at time t which indexes a specific structure,E^{z_t}, and parameters, Θ^{z_t}. Figure 2 shows a first order model. Given a set of K structures $\mathbf{E} = \{\bar{E}^1, ..., \bar{E}^K\}$, parameters $\mathbf{\Theta} = \{\pi^0, \pi^1, ..\pi^k, \Theta^1, ..., \Theta^K\}$ and an observations over the time period $\mathbf{t} = [1, ..., T]$:

$$p(\mathbf{X_t}, z_\mathbf{t}|\mathbf{E}, \mathbf{\Theta}) = p(z_\mathbf{t})p(\mathbf{X_t}|z_\mathbf{t}, \mathbf{E}, \mathbf{\Theta}) \quad (19)$$
$$= \prod_{t=1}^{T} p(z_t|\pi^{z_{t-1}})p(\mathbf{X}_t|\widetilde{\mathbf{X}}_t, E^{z_t}, \Theta^{z_t})$$

where $z_0 = 0$, the transition distribution is multinomial $p\left(z_t|\pi^{z_{t-1}}\right) = \text{Mult}\left(z_t; \pi^{z_{t-1}}\right)$ and is given a Dirichlet prior $p_0\left(\pi\right) = \prod_{k=1}^{K} \text{Dir}\left(\pi^k; \alpha_1^k, ..., \alpha_K^k\right)$.

Exact inference on this model is complex due to the fact there are K^T possible state sequences. Thus, we turn to an MCMC approach in which samples are drawn from this model using a Gibbs sampler. The sampler has three main steps. Step 1 samples the state sequence given previous estimate of structures and parameters. This is done efficiently with backward message passing followed by forward sampling. This step can be modified when initializing to simply sample the state sequence from its transition prior. During Step 2 the sampled counts of state transitions are noted and transition probabilities are then sampled given these counts. In Step 3 a vector $\mathbf{t}_k$ is formed with all the time points with $z_t = k$. The structure and parameters are then sampled given $\mathcal{D}_{\mathbf{t}_k}$ for each k. It is important to note that given a state sequence, one can efficiently calculate exact event probabilities and posterior over structures.

A STIM(r,K) assumes a known number of states, K, and structures can be revisited over time. However, the TIM(r) model can also easily be embedded in a parts partition model (PPM) similar to that used by (Xuan and Murphy, 2007) which allow for an unknown number of states that are never revisited. Similarly one may adopt nonparametric Bayesian models such as the hierarchical Dirichlet process hidden Markov model (HDP-HMM)(Teh et al., 2006) to allow for an unknown number of potential revisited states. The appropriate choice of model is highly dependent on the model assumptions' match to the application of interest. The experiments presented in this paper focus on the use of the STIM(r,K) leaving the above modifications/extensions for future work.

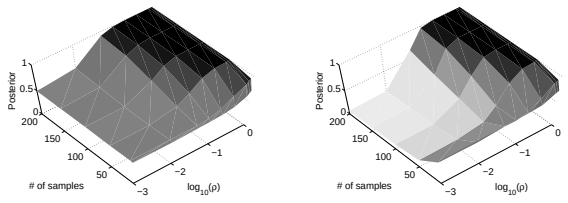
5 EXPERIMENTS

In this section we present experiments focusing on the calculation of posterior event probabilities. We begin with a simple synthetic example and move on to data involving the interaction of tracked moving objects. Specifically, we are interested in quantifying uncertainty in the dependence structure among time-series rather than obtaining point estimates. When analyzing data we do not assume there is a "true/correct structure" one would like to discover. Our goal is to fully characterizes posterior uncertainty.

5.1 DISTINGUISHING STRUCTURE

A simple experiment on two one dimensional data streams ($N = 2$) is carried out in order to explore how well one can distinguish whether a an edge is present as a function of the number of samples and strength of dependence. T samples are drawn from a TIM(1) model with a static structure of $\mathbf{x}_\mathbf{t}^1$ influencing $\mathbf{x}_\mathbf{t}^2$: $\bar{E}$ contains a single edge, $1 \rightarrow 2$. $\mathbf{x}_\mathbf{t}^1$ is a random walk ($\mathbf{x}_t^1 = \mathbf{x}_{t-1}^1 + n_t^1$) with unit variance Gaussian noise. The amount of influence $\mathbf{x}_\mathbf{t}^1$ has on $\mathbf{x}_\mathbf{t}^2$ is controlled via a variable ρ such that $\mathbf{x}_t^2 = \rho\mathbf{x}_{t-1}^1 + (1-\rho)\mathbf{x}_{t-1}^2 + n_t^2$ where n_t^2 is unit variance. Note that if $\rho = 1$, $\mathbf{x}_t^2$ moves to $\mathbf{x}_{t-1}^1$ plus noise, and if $\rho = 0$ it simply follows its own random walk independent of the other time-series.

Given the T samples the posterior probability of edge $1 \rightarrow 2$ is calculated using a TIM(1). A weak matrix-normal-inverse-Wishart prior is placed on the parameters



(a) $\bar{E} \in \mathcal{T}$ (b) $\bar{E} \in \mathcal{A}$

Figure 3: Edge posterior for different ρ and # of samples

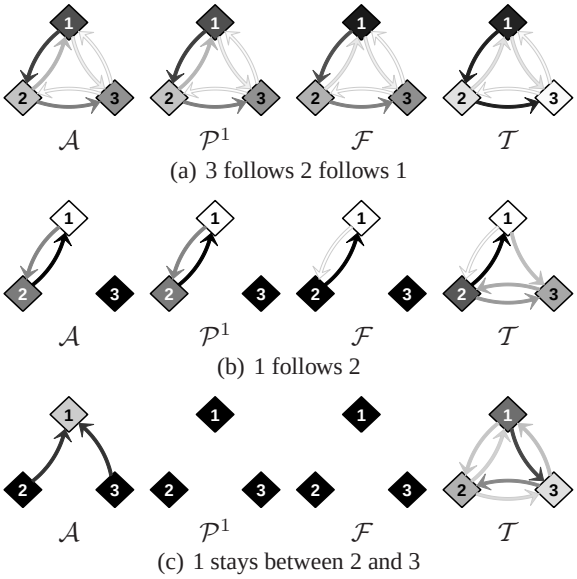
(zero mean, large variance, small degrees of freedom). Various settings of ρ and T are explored. For each setting, 100 trials are performed and the average posterior is recorded.

Note that for $N = 2$ the structure set $\mathcal{P}^1 = \mathcal{A}$ and that $|\mathcal{T}| = 2$, $|\mathcal{F}| = 3$ and $|\mathcal{A}| = 4$. Using a uniform prior over structures we display results in Figure 3(a) and Figure 3(b) for $\bar{E}$ in the set $\mathcal{T}$ and $\mathcal{A}$ respectively. The results follow intuition: few samples and/or $\rho < 0.1$ results in a posterior close to chance (1/2 for $\mathcal{T}$, 1/4 for $\mathcal{A}$) while once $\rho > 0.1$ there is a sharp increase in the posterior of the correct structure. Note that, here, we are quantifying uncertainty in terms of posterior edge appearance probabilities. One can also calculate higher order information such as the posterior variance. For example, running a single trial with $T = 50, \rho = 0.1$ one obtains a posterior edge probability of .6327 and posterior variance of .2324 when $\bar{E} \in \mathcal{T}$.

5.2 INTERACTIONS OF MOVING OBJECTS

Next we explore a dataset comprised of recordings of three individuals playing a simple interactive computer game. Each player's computer mouse controls a specific dot/marker on their screen. In addition, each player's screen shows the dots representing the other players as well. The players are instructed to perform a particular interactive behavior. The position of each player is obtained at 8 samples per second. Each behavior is recorded for approximately 30 seconds.

The players are first instructed to follow each other in order. That is, player 1 moves around the screen while player 2 is instructed to follow him. Player 3 is instructed to follow player 2. Using a uniform prior on structure and weak matrix-normal-inverse-Wishart prior on parameters a posterior on structure is obtained given this data using a TIM(1). For increasingly restrictive sets of structures, results are shown in Figure 4(a) depicted as weighted interaction graphs. In these graphs edge color represents the posterior probability of that edge. Node color represents the probability a time-series has no parents/is a root (white = 0, black = 1). Note that this behavior is described well by all structure classes/sets and the most certainty is obtained when considering the most restrictive set $\mathcal{T}$.



(a) 3 follows 2 follows 1

(b) 1 follows 2

(c) 1 stays between 2 and 3

Figure 4: Posterior interaction graphs for specific behaviors. Columns corresponds to the posterior over increasingly restrictive sets of structures: From $\bar{E} \in \mathcal{A}$ to $\bar{E} \in \mathcal{T}$.

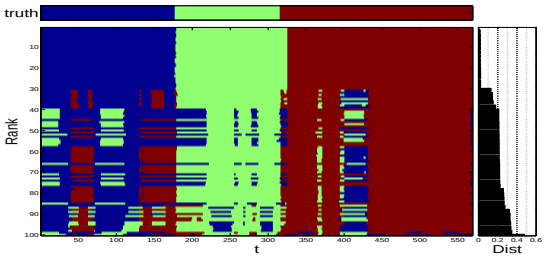
Next the players 2 and 3 are instructed to move freely, while player 1 is told to follow player 2. The resulting posteriors are shown in Figure 4(b). Note this behavior is not described well by a tree since player 3 is independent of all others. Lastly, players 2 and 3 are told to move freely while player 1 does his best to stay between both of them. The results in Figure 4(c) show that this behavior only seems represented well in set $\mathcal{A}$. It is the only set that allows more than a single parent for a time-series.

5.3 FOLLOW THE LEADER

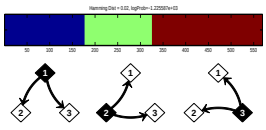
Using the same interactive game setup we record three individuals playing a game of follow the leader. One player is designated the leader. The leader moves his or her dot randomly around the screen while the other players are instructed to follow. Here, the designated leader changes throughout the game. That is, a fourth person observing the game tells the players when to switch leaders. In this case, the latent variable indicating the change of leader is observable, and consequently, nominal ground truth is available by which to evaluate performance.

We begin by using STIM(1,3) with $\bar{E} \in \mathcal{T}$ to analyze this sequence. That is, we will use the fact there are only three players and knowledge that directed trees may sufficiently describe the interaction among players. A uniform prior on structures and equivalently weak prior on parameters is used. A weak self biased prior on the state transition distribution is imposed with a bias towards self transition.

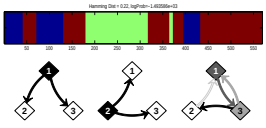
Given the data and the prior model, 100 samples of the structure, parameters and the hidden state sequence are ob-



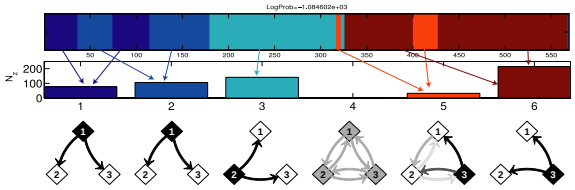
(a) Samples of z using STIM(1,3)



(b) Rank 1 sample



(c) Rank 34 sample



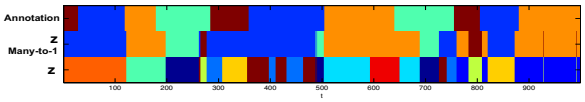
(d) Sample using STIM(1,6)

Figure 5: Follow the leader results

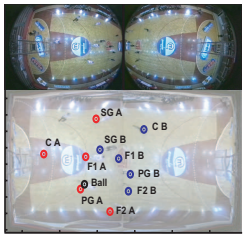
tained with a Gibbs sampler. Burn in required approximately 60 iterations. A detailed view of the results can be found in Figure 5(a). The ground truth state sequence is shown on top with players taking turns being the leader in order. This figure shows each of the Gibbs sampled state sequences. The labels were permuted to give a consistent coloring with the ground truth segmentation shown on the top. The results are ranked by the log likelihood of the data given the sampled parameters and structures, with the top being the most likely. The right side of the figure shows the normalized Hamming distance of the best mapping to the ground truth. Note that it is highly correlated with the log likelihood of the data. Each sample falls within two general categories. The top third of the samples match the ground truth closely, the bottom two thirds suggests a consistent alternative explanation.

The state labels alone simply provide a segmentation. Given this segmentation we look at the posterior on structure to analyze the interaction among the players. Figures 5(b) and 5(c) show a more detailed breakdown of two sampled models. The first row of each figure shows the most likely state sequence given the sample model. The second row shows interaction graphs representing the posterior probability of the structure for each state. Recall that while the state sequence is MCMC sampled, we obtain an exact posterior conditioned on this sequence.

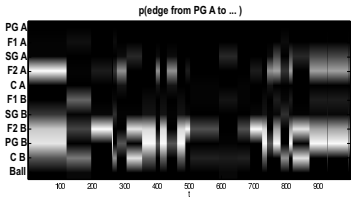
Figure 5(b) is a sample with low Hamming distance and



(a) Sample z from STIM(2,10) compared w/ annotation



(b) Sample frame



(c) Influence of PG A

Figure 6: Basketball results

high log likelihood. Notice that the posterior on structure for each state is basically a delta function on three distinct structures. These structures agree with our intuition in that each root is consistent with who was designated as the leader and the followers are conditionally independent given the root. Figure 5(c) is a sample with a mid-range Hamming distance (ranked 34 out the 100 samples). It has errors consistent with the majority of sequences shown in 5(a). The confusion between the first and third state is most noticeably reflected in the structure posterior for state 3.

While the above analysis assumed three states, consistent with our knowledge of the ground truth, Figure 5(c) gives evidence for additional modes/states. That is, for each phase of the game a better model may be a mixture of processes each with similar structure but different parameters. We repeat the experiment using $K = 6$ states. Figure 5(d)is a sample from this model.

The second row shows the occurrence of the learned states. Interestingly, state 4 indicates uncertainty in the structure. However, this state is never used and thus its posterior remains uniform. The remaining structures are consistent with the ground truth indicating who is the leader with little uncertainty.

5.4 SPORTS INTERACTION ANALYSIS

Next we consider analysis of player interactions in sports data. A basketball game recording from the CVBASE 06 dataset was used (Pers et al., 2006). Players were tracked in two cameras and their positions were mapped to a common coordinate system and recorded at each frame. A total of 11 tracks were obtained; five players on each team plus the ball. A coarse annotation of the current phase of the game was created. Four phases were noted: team A on offense, team B on offense, team A transitioning to offense, and team B transitioning to offense. A sample frame with player positions is shown in Figure 6(b).

A STIM(2,10) model with $\bar{E} \in \mathcal{T}$ is used for analysis; a

second order model incorporates velocity information. We use a weak prior on parameters and uniform prior on structures. A state sequence obtained from our Gibbs sampler is shown in the bottom row of Figure 6(a). The middle row shows the best many-to-one mapping of the sampled state sequence to the coarse annotation. Note that while the sampled sequence is somewhat predictive of the annotation, a direct comparison is misleading since within each phase of the game multiple structures may be active.

Given a sampled state sequence, the posterior on $\bar{E}$ is obtained for each point in time. With 11 time-series and 6 states, displaying all posterior interaction graphs is impractical. As an alternative, we focus on calculating posterior event probabilities over time intervals. Table 1(a) shows the probability of the root being on either team or the ball given each phase in the coarse annotation. Over all phases of the game the ball has the highest probability of being the root. When on offense or transitioning to offense a player on team A has a higher probability of being the root than one on B and vice versa. Similarly, Table 1(b) shows the probability of being a leaf averaged over each team and the ball. Here the ordering is reversed and has a connection to which team is on defense. We see how analysis of these posterior event probabilities can give one an intuitive prediction of the state of the game.

Lastly, we calculate the expected number of children for each player and the ball over all time. The top four time-series were the ball, point guard A (PG A), forward 1 A (F1 A), and forward 1 B (F1B) with expectations of 1.76, 1.73, 1.08, and 0.95. Again, the results are intuitive. The ball has the largest influence on the dynamic of the game. The point guard controls the flow of the game and is usually the best ball handler. Figure 6(c) shows the posterior probability of an edge from the PG A to every other player over all time. Note that PG A tends to influence his own forward as well as the point guard and forward on the other team.

6 CONCLUSION

We have presented a framework for Bayesian inference of statistical dependence structure among multiple time series where parameters are treated as a nuisance. In the static structure setting, sets of directed structures were introduced that yield tractable inference and exact calculation of useful expectations and event probabilities for time-series interactions. In the dynamic structure setting, the inclusion of a latent index over structures allowed for inference of changing interactions. Additionally, we illustrated the utility of this framework via experiments characterizing the posterior uncertainty on the interaction of multiple moving objects.

M. Siracusa was partially supported by HSN (Heterogeneous Sensor Networks), which receives support from Army Research Office (ARO) Multidisciplinary Research Initiative (MURI) program (Award number W911NF-06-1-0076). J. Fisher was partially supported by the Air Force Office of Scientific Research under Award No. FA9550-06-1-0324. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Air Force.

Table 1: (a) Probability of root and (b) average probability of leaf given annotation. **bold** and underlined values are the maximum and second highest in each column respectively

(a)	A Offense	B Offense	B Trans to O	A Trans to O
Team A	.2475	.0832	.0420	.4769
Team B	.0468	.2902	.4234	.0459
Ball	**.7057**	**.6266**	**.5346**	**.4771**

(b)	A Offense	B Offense	B Trans to O	A Trans to O
Team A	.3942	**.4097**	**.4880**	.3254
Team B	**.5515**	.0493	.4234	**.6512**
Ball	.1685	.0671	.1133	.2986

References

F. R. Bach and M. I. Jordan. Learning graphical models for stationary time series. In *Trans. on signal processing*, 2004.

J. A. Bilmes. Dynamic bayesian multinets. In *UAI*, 2000.

N. Friedman and D. Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 50:95–125, 2003.

D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and bayesian multinets. In *Artificial Integlligence*, volume 82, pages 45–74, 1996.

D. Heckerman. A tutorial on learning with bayesian networks. *Microsoft Tech Report: MSR-TR-95-06*, 1995.

G. Kirchhoff. Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. In *Annalen der Physik und Chemie*, volume 72, pages 497–508, 1847.

S. Kirshner, P. Smyth, and A. Robertson. Conditional chow-liu tree structures for modeling discrete-valued vector time series. In *Technical Report UCI-ICS*, 2004.

M. Koivisto, K. Sood, and M. Chickering. Exact bayesian structure discovery in bayesian networks. *JMLR*, 5, 2004.

T. Koo, A. Globerson, X. Carreras, and M. Collins. Structured prediction models via the matrix-tree theorem. In *EMNLP*, 2007.

M. Meila and T. Jaakkola. Tractable bayesian learning of tree belief networks. In *Statistical Computing*, pages 77–92, 2006.

B. Milch, B. Marthi, D. Sontag, S. Russell, D.L. Ong, and A. Kolobov. Approximate inference for infinite contingent bayesian networks. In *AIStats*, 2005.

J. Pers, M. Bon, and G. Vuckovic. CVBASE dataset. In *http://vision.fe.uni-lj.si/cvbase06/dataset.html*, 2006.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical dirichlet processes. In *JASA*, volume 101(476), 2006.

W.T. Tutte. *Graph Theory*. Addison-Wesley Pub. Co., 1984.

M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 1997.

D.B. Wilson. Generating random spanning trees more quickly than the cover time. In *Theory of Computing*, 1996.

X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *ICML*, 2007.