Causal Time-Varying Dynamic Bayesian Networks

Victor W. Chu [1] Raymond K. Wong[1] Wei Liu[2] and Fang Chen[2]

[1] University of New South Wales, Australia
{wchu,wong}@cse.unsw.edu.au
[2] National ICT Australia
{wei.liu,fang.chen}@nicta.com.au

Technical Report
UNSW-CSE-TR-201322
September 2013

THE UNIVERSITY OF
NEW SOUTH WALES

School of Computer Science and Engineering
The University of New South Wales
Sydney 2052, Australia

Abstract

Causal network structure learning methods, e.g., IC*, FCI and MBCS*, are investigated in recent time but none of them has taken possible time-varying network structure, such as time-varying dynamic Bayesian networks (TV-DBN), into consideration. In this paper, notions of relaxed TV-DBN (RTV-DBN) and causal TV-DBN (CTV-DBN), as well as a definition of causal boundary are introduced. RTV-DBN is a generalized version of TV-DBN whilst CTV-DBN is a causal compliant version. CTV-DBN is constructed by using an asymmetric kernel, versus a symmetric kernel as in TV-DBN, to address the problem of sample scarcity and to better fit within the causal boundary; while maintaining similar level of variance and bias trade-off. Upon satisfying causal Markov assumption, causal inference can be made based on manipulation rule. We explore spatio-temporal data which is known to exhibit heterogeneous patterns, data sparseness and distribution skewness. Contrary to a naïve method to divide a space by grids, we capture the moving objects' view of space by using clustering to overcome data sparseness and skewness issues. In our experiments, we use RTV-DBN and CTV-DBN to reveal the evolution of interesting region time-varying structure from the transformed data.

1 Introduction

While there is not much previous work done in analyzing the relationships among network conditions across time, Song *et al.* [25] outline a time-varying dynamic Bayesian networks (TV-DBN) to model gene-to-gene interaction networks. Causal network structure learning methods, e.g., IC* [20], FCI [28] and MBCS* [21], are investigated in recent time but none of them has taken possible time-varying network structure into consideration. In this paper, we extend TV-DBN to model time-varying network causal relationships. For example, in the domain of spatio-temporal analysis, by replacing genes with moving objects' "traffic" conditions from different regions, we will be able to detect region-to-region interactions, which means, for example, we will be able to tell how the congestion of some regions will lead to the congestion of other regions with respect to time by applying causal inference. More recently, Dondelinger *et al.* [4] propose a non-homogeneous dynamic Bayesian networks for inferring gene regulatory networks with gradually time-varying structure. Although both proposals can be traced back to a common root of Robinson and Hartemink [23, 24], Dondelinger *et al.* work with continuous time data whilst the method proposed by Song *et al.* [25] require discretization of the data. The flexibility of the former method does come with a price. It is more complex and special attention has been put into the model to avoid over-fitting. For observations sampled periodically, a model based on discrete-time data is sufficient.

Most observations exhibit unfavorable statistical properties: heterogeneous patterns, data sparseness and distribution skewness. For example, time-varying patterns and uneven concentration of data points — data concentrated on certain regions leaving the other places rarely occupied — are common in the domain of traffic modeling [15]. We expect that similar problems also appear in other domains. By using a time-varying model, e.g. TV-DBN, the network structure learning process will allow us to observe the evolution of region connections and disconnections to address the issue of heterogeneous patterns. To address data sparseness and distribution skewness, we propose to use density-based clustering methods to reveal clusters from data directly. The clusters represent regions for a particular space-time interval from the specific type of objects' point of view, allowing for time-varying region relationship modeling using a time-varying model. Density-based clustering methods [5, 12, 22] are argued [18] to be the best solution for spatio-temporal clustering due to the following reasons: 1) they are able to handle clusters with no predefined shape, e.g. a cluster could be of any shape rather than spherical, 2) they are able to cope with noises in the data, and 3) one can base on the parameters to fine tune the methods to fit a particular problem.

In this paper, a notion of relaxed time-varying dynamic Bayesian networks (RTV-DBN) and a notion of causal time-varying dynamic Bayesian networks (CTV-DBN) are introduced. RTV-DBN is a generalized version of TV-DBN which allows a variable at time t to be regulated by all variables at time $t-1$. In CTV-DBN, causal Markov assumption [19, 26, 27, 28] is satisfied by considering causal boundary ($\mathcal{B}$). It is achieved by an asymmetric kernel [16] which limits the information sharing across time but still allowing suitable information sharing to address data scarcity while maintaining similar level of variance and bias trade-off. Asymmetric kernel provides a solution to rectify boundary problem created by real-world data where they mirror a causal function [16]. Causal

inference can be made based on manipulation rule following the assumption of faithfulness and causal sufficiency [19, 26, 27, 28].

This paper also includes an application of RTV-DBN and CTV-DBN to moving object spatio-temporal analysis under the context of moving objects' territories. The regions in a network are not only regulated by the other regions at time $t-1$, but can be regulated by the same regions (at time $t-1$). However, such a situation is forbidden in gene-to-gene regulation. In fact, the traffic conditions between adjacent time-steps of the same region is expected to be highly correlated by real-life experience.

2 Relaxed Time-Varying Dynamic Bayesian Networks (RTV-DBN)

Relaxed time-varying dynamic Bayesian networks (RTV-DBN) is a generalized version of TV-DBN [25], which is built based on the model of dynamic Bayesian networks (DBN) [17]. Let $\mathbf{X}^t = (X_1^t, \ldots, X_r^t)^\top \in \mathbb{R}^r$ represents a random vector (which represents expression level in [25]) of r regions (which represent genes in [25]) at time t, a dynamic process of such time dependant condition can be modeled by a first-order Markovian transition model $p(\mathbf{X}^t|\mathbf{X}^{t-1})$ which defines the probabilistic distribution of variables at time t given those at time $t-1$. The probability of observing a scenario from these regions over a period $t \in \{1 \ldots T\}$ can be expressed by:

$$p(\mathbf{X}^1, \ldots, \mathbf{X}^T) = p(\mathbf{X}^1) \prod_{t=2}^{T} p(\mathbf{X}^t|\mathbf{X}^{t-1}). \tag{2.1}$$

Suppose that the structure of the networks is specified by a set of regulatory relations $\mathbf{X}_{\pi_i}^{t-1} = \left\{X_j^{t-1} : X_j^{t-1} \text{ regulates } X_i^t\right\}$, where $i, j \in \{1 \ldots r\}$ and $\pi_i \subseteq \{1 \ldots r\}$, we can factorize the transition model $p(\mathbf{X}^t|\mathbf{X}^{t-1})$ over individual regions. Equation (2.1) can then be rewritten to:

$$p(\mathbf{X}^1, \ldots, \mathbf{X}^T) = p(\mathbf{X}^1) \prod_{t=2}^{T} \prod_{i=1}^{r} p(X_i^t|\mathbf{X}_{\pi_i}^{t-1}). \tag{2.2}$$

Let graph $\mathcal{G}^t = (\mathcal{V}, \mathcal{E}^t)$ represents the conditional dependence between the random vectors $\mathbf{X}^{t-1}$ and $\mathbf{X}^t$, where the vectors represent feature values from different regions at time $t-1$ and at time t respectively. Each vertex in $\mathcal{V}$ corresponds to a sequence of variables $X_i^1, \ldots, X_i^T$, and the edge set $\mathcal{E}^t \subseteq \mathcal{V} \times \mathcal{V}$ contains directed edges from components of $\mathbf{X}^{t-1}$ to components of $\mathbf{X}^t$. The time dependent transition model $p^t = (\mathbf{X}^t|\mathbf{X}^{t-1})$ is expressed by an auto-regressive DBN form $\mathbf{X}^t = \mathbf{A}^t\mathbf{X}^{t-1} + \boldsymbol{\epsilon}$, where $\mathbf{A}^t \in \mathbb{R}^{r \times r}$ is a matrix of coefficients relating the variables at time $t-1$ from all regions to the variables of the regions in the next time point t, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is an error term. The region time-varying structure is represented by the non-zero entries (connections) and zero entries (disconnections) in the estimated matrices $\hat{\mathbf{A}}^t$ at time t. In this model, the estimation of the strength of dependencies is accomplished by minimizing a set of squared loss functions with regularization, one for each vertex at each time point $t^* \in \{1 \ldots T\}$.

Assuming that the underlying network structures are sparse and vary smoothly across time, the model is built with cross time information sharing to address the problem of sample scarcity. The functions representing the structures are smooth and with bounded second derivatives, to make it statistically tractable. We estimate the network by decomposing the problem along the time (t) and region feature vector ($\mathbf{x}$). We estimate the neighborhood for each region separately by using linear regression and then join these neighborhoods to form the overall network. The estimation problem is reduced to set of optimizations with one for each node $i \in \{1 \dots |\mathcal{V}|\}$ for time points $t^* \in \{1 \dots T\}$:

$$\hat{\mathbf{A}}_{i\cdot}^{t^*} = \underset{\mathbf{A}_{i\cdot}^{t^*} \in \mathbb{R}^{1\times r}}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} w^{t^*}(t)(\mathbf{A}_{i\cdot}^{t^*}\mathbf{x}^{t-1} - x_i^t)^2 + \lambda \|\mathbf{A}_{i\cdot}^{t^*}\|_1 \tag{2.3}$$

where λ is a regularization parameter which controls the sparsity of the networks, and $w^{t^*}(t)$ is the weighting of an observation from time t defined as $w^{t^*}(t) = \frac{K_h(t-t^*)}{\sum_{t=1}^{T} K_h(t-t^*)}$ in which $K_h(\cdot)$ is a symmetric and non-negative kernel function and h is the kernel bandwidth. In our experiments in Section 6, we have selected $\mathbf{x}$ to be the average velocity of trajectory fragments in each region. The end product of TV-DBN estimation is a set of $\hat{\mathbf{A}}_{i\cdot}^{t^*}$ (one per region) which can be combined to give an estimate of $\mathbf{A}^t$ where $t \in \{1 \dots T\}$. The non-zero and zero entries in the matrices $\hat{\mathbf{A}}^t$ represent the time-varying connections and disconnections between regions over the time period:

$$\hat{\mathcal{E}}^t = \{(i,j) \in \mathcal{V} \times \mathcal{V} | \hat{\mathbf{A}}_{ij}^t \neq 0\}. \tag{2.4}$$

Please note that we are no longer required to restrict that the regions between two time steps must be different as specified in [25] ($\hat{\mathcal{E}}^t = \{(i,j) \in \mathcal{V} \times \mathcal{V} | i \neq j, \hat{\mathbf{A}}_{ij}^t \neq 0\}$). In the domain of spatio-temporal analysis, a region in a network is not only regulated by the other regions at $t-1$, but is also regulated by itself at $t-1$.

3 Causal Time-Varying Dynamic Bayesian Networks (CTV-DBN)

Although Bayesian networks (BN) structure may be directed, the directions of arrows do not define causal effects as the influence can flow both ways except a collider (v-structure) is hit, e.g., a directed edge from vertex α to vertex β does not require that β is causally dependent on α. All of the definitions in BN refer only to probabilistic properties, such as conditional independence [13]. Hence, the following BN are equivalent as they impose exactly the same conditional independence requirements: $\alpha \rightarrow \beta \rightarrow \gamma$ and $\alpha \leftarrow \beta \leftarrow \gamma$. Song *et al.* in their proposed TV-DBN [25] recognise the same problem – BN does not necessarily imply causality, but suggest that dynamic Bayesian networks (DBN) bears a natural causal implication in which TV-DBN is part of this family. The main reason why they favor DBN over BN is its enhanced semantic interpretability. Each edge in a DBN only points from time $t-1$ to t contributing to a natural causal implication. However, the network structure established in TV-DBN basically ignores causal relationships allowing the sharing of information across the whole time period.

As we describe in Section 2, a simple form of the transition model $p(\mathbf{X}^t|\mathbf{X}^{t-1})$ in a DBN is a linear dynamic model $\mathbf{X}^t = \mathbf{A}^t\mathbf{X}^{t-1} + \boldsymbol{\epsilon}$. The difference between causal models and probabilistic models arises when we care about interventions in the model [13]. We would like to establish causal relationships between regions to allow for causal inference by assigning manipulated probability density to a region of interest assuming that all attempted manipulations are fully successful [26]. The condition of causal Markov assumption (CMA) is invoked to make a BN isomorphic with a causal model [14], where the condition [27] is defined as: Given a causal graph $\mathcal{CG} = \langle \mathcal{V}_{\mathcal{CG}}, \mathcal{E}_{\mathcal{CG}}, \mathcal{P}_{\mathcal{CG}} \rangle$, where $\mathcal{V}_{\mathcal{CG}}$ is a set of vertices and $\mathcal{E}_{\mathcal{CG}}$ is a set of edges between vertices in $\mathcal{V}_{\mathcal{CG}}$ and $\mathcal{P}_{\mathcal{CG}}$ is a probability distribution over the vertices in $\mathcal{V}_{\mathcal{CG}}$. $\mathcal{CG}$ satisfies CMA if and only if for every $v \in \mathcal{V}_{\mathcal{CG}}$, v is independent of $\{\mathcal{V}_{\mathcal{CG}} \setminus Descendants(v) \cup Parents(v)\}$ given $Parents(v)$, where $Parents(v)$ is the set of parents of v in $\mathcal{CG}$ and $Descendants(v)$ is the set of descendants of v in $\mathcal{CG}$.

If $\mathcal{V}_{\mathcal{CG}}$ represents regional variables from all time points, i.e., $\{\cup_1^T \mathbf{X}^t\}$, the network structure estimated by optimization as defined by Equation (2.3) does not satisfy CMA. It is because the weighting function $w^{t^*}(t)$ considers the time points in $\{\mathcal{V}_{\mathcal{CG}} \setminus Descendants(v) \cup Parents(v)\}$ given $Parents(v)$, e.g., $\hat{\mathbf{A}}_{i\cdot}^{t^*}$ at time t^* is not only determined by $\mathbf{x}_i^{t^*-1}$ but also $\mathbf{x}^{t^*-z}$ where $z > 1$. As a result, it contradicts CMA. The original idea of using weighting function is based on the assumption that the structural changes of the network is smooth over time; and hence, allowing one to gather evidence across time by reweighting the observations from different time points to alleviate the problem of sample scarcity. However, in order to comply with CMA, the weighting function should only be allowed to gather evidence from $\mathcal{S} = \{Descendants(v) \cup Parents(v)\}$. We define causal boundary ($\mathcal{B}$) as a set of points in the closure of $\mathcal{S}$ but not belonging to the interior of $\mathcal{S}$. As a result, we propose to adopt a causal weighting function $w_c^{t^*}(t)$ to fulfil the requirement: $w_c^{t^*}(t) = \frac{K_h^a(t-t^*)}{\sum_{t=1}^T K_h^a(t-t^*)}$ in which $K_h^a(\cdot)$ is an asymmetric and non-negative kernel function satisfying CMA. Hence, we rewrite Equation (2.3) to:

$$\hat{\mathbf{A}}_{i\cdot}^{t^*} = \underset{\mathbf{A}_{i\cdot}^{t^*} \in \mathbb{R}^{1\times r}}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} w_c^{t^*}(t)(\mathbf{A}_{i\cdot}^{t^*}\mathbf{x}^{t-1} - x_i^t)^2 + \lambda \|\mathbf{A}_{i\cdot}^{t^*}\|_1. \tag{3.1}$$

The use of asymmetric kernel for non-parametric regression can be found in economic literature [7, 8] but rarely discussed in machine learning. Mackenzie and Tieu [16] is an example which discusses an application of asymmetric kernel regression to radial-basis neural networks. The authors mention that the available real-life data reproduce a causal function; and therefore, are naturally bounded by an interval. Hence, a truncation of a symmetric kernel at the boundary makes the model suffers from material bias error, known as boundary problem. Although there are several attempts to resolve boundary problem [11, 29], most of them cannot correct bias without increasing noise level and/or variance error [16]. Mackenzie and Tieu propose to correct boundary error by replacing a symmetric kernel with an asymmetric kernel. Apart from the favorable boundary property [16] – maintaining the same level of variance and bias trade-off, asymmetric kernel, like gamma distribution function, can also provide a weighting function which is within causal boundary ($\mathcal{B}$) in $\mathcal{S}$ in our

application. We define $K_h^a(\cdot)$ by using gamma distribution function:

$$f(t;\theta;\mathcal{K}) = \frac{1}{\theta^{\mathcal{K}}\Gamma(\mathcal{K})}(t-\eta)^{\mathcal{K}-1}e^{-(t-\eta)/\theta} \tag{3.2}$$

where $t > 0$, $\mathcal{K}$ is a shape parameter which determines the basic shape of the function, θ is a scale parameter and η is a location parameter [2]. Under the scenario of using kernel regression to estimate the functional relationship between two variables y and t ($y_i = y(t_i) + \epsilon$, where $i \in 1 \ldots N$, $0 \leq t_i \leq T$ and ϵ is random noise [16]) by using symmetric Gaussian kernel ($K_{\mathcal{G}}$) with boundary ($K_{\mathcal{G}}$ is the Gaussian density function with mean μ and variance σ^2), we obtain significant bias at the boundary as the odd moments of $K_{\mathcal{G}}$ are no longer zero due to truncation [16], which is:

$$\begin{aligned} bias[\hat{y}(\eta)] = \Bigg\{ & y(\eta)\int_0^\infty K_{\mathcal{G}}(t-\eta)dt + y'(\eta)\int_0^\infty (t-\eta)K_{\mathcal{G}}(t-\eta)dt \\ & + \frac{1}{2}y''(\eta)\int_0^\infty (t-\eta)^2 K_{\mathcal{G}}(t-\eta)dt + \ldots \Bigg\} - y(\eta) \end{aligned}$$

where $\hat{y}(\eta)$ is a Priestley-Chao estimator [16] of $y(\eta)$, versus the scenario of no boundary:

$$bias[\hat{y}(\eta)] = \int_{-\infty}^{\infty} y(t_i)K_{\mathcal{G}}(\eta - t_i)dt_i - y(\eta) \cong \frac{\sigma^2}{2}y''(\eta) + \frac{\sigma^4}{8}y''''(\eta).$$

However, the boundary error term is vanished by replacing symmetric kernel $K_{\mathcal{G}}$ by asymmetric Gamma kernel (K_Γ) in the case of kernel regression with boundary:

$$bias[\hat{y}(\eta)] = \frac{\sigma^2}{2}y''(\eta) + \frac{\sigma^4}{3\eta}y'''(\eta) + \frac{\sigma^4}{8}\left\{1 + 2\left(\frac{\sigma}{\eta}\right)\right\}y''''(\eta) + \ldots$$

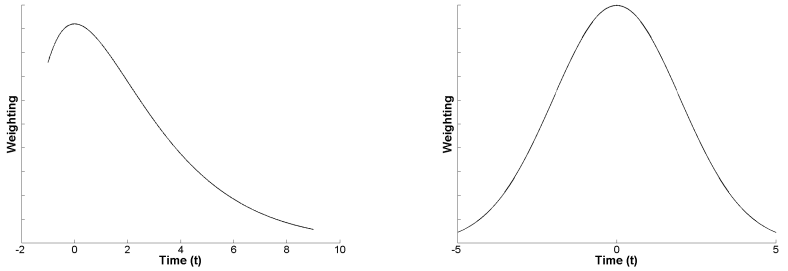
Assigning $\theta = h = 2$, $\mathcal{K} = 2$ and $\eta = 1$ gives an asymmetric weighting function:

$$K_2^a(t) = \frac{1}{4}(t-1)e^{-(t-1)/2} \tag{3.3}$$

We call it a CMA compliant Gamma kernel (K_Γ) with a shape shown in Figure 3.1(a) versus a typical symmetric Gaussian kernel ($K_{\mathcal{G}}$) displayed in Figure 3.1(b). Alternatively, assuming that the data sparseness and distribution skewness problems are *fully* addressed by density-based clustering, the necessity to gather evidence across time could be *totally* diminished. As a result, we might be able to simply drop the weighting function from Equation (2.3). Based on our experimental results on taxi trajectories in Section 6, we are in favour of using Equation (3.1), which gathers evidence only within the causal boundary ($\mathcal{B}$) in $\mathcal{S}$. As a result, we transform RTV-DBN to causal time-varying dynamic Bayesian networks (CTV-DBN) by adopting Equation (3.1) to be our new objective function for estimating time-varying region structure, while Equations (2.1) and (2.2) remain applicable to CTV-DBN.

4 Causal inference by CTV-DBN

We now revisit causal graph $\mathcal{CG}$ defined in Section 3 to establish manipulation rule for CTV-DBN.



(a) Gamma kernel (K_Γ) (b) Gaussian kernel ($K_\mathcal{G}$)

Figure 3.1: Shape of kernel functions

A probability density $\mathcal{P}_{\mathcal{CG}}(\mathcal{V}_{\mathcal{CG}})$ factors according to $\mathcal{CG}$ if and only if

$$\mathcal{P}_{\mathcal{CG}}(\mathcal{V}_{\mathcal{CG}}) = \prod_{v \in \mathcal{V}_{\mathcal{CG}}} \mathcal{P}_{\mathcal{CG}}(v|Parents(v)) \tag{4.1}$$

where $Parents(v)$ is the set of parents of v in $\mathcal{CG}$ [26].

In a causal Bayesian network and under CMA, assuming $\mathbf{n} \subset \mathcal{V}_{\mathcal{CG}}$ with only non-descendants of m, a manipulation of $m \in \mathcal{V}_{\mathcal{CG}}$ to $\mathcal{P}_{\mathcal{CG}}{}'(m|\mathbf{n})$ can be achieved by replacing $\mathcal{P}_{\mathcal{CG}}(m|Parents(m))$ in Equation (4.1) by a manipulated density $\mathcal{P}_{\mathcal{CG}}{}'(m|\mathbf{n})$ to form a manipulation rule:

$$\mathcal{P}_{\mathcal{CG}}(\mathcal{V}_{\mathcal{CG}}||\mathcal{P}_{\mathcal{CG}}{}'(m|\mathbf{n})) = \mathcal{P}_{\mathcal{CG}}{}'(m|\mathbf{n}) \prod_{v \in \mathcal{V}_{\mathcal{CG}} \setminus \{m\}} \mathcal{P}_{\mathcal{CG}}(v|Parents(v)) \tag{4.2}$$

where the double bar indicates an assignment of probability and $\mathcal{P}_{\mathcal{CG}}{}'$ is a new probability density. Therefore, based on Equation (2.2), the manipulation rule for CTV-DBN at time $t = \zeta$ (where $\zeta \in \{2 \dots T\}$) and region $i = m$ can be written as:

$$p(\mathbf{X}^1, \dots, \mathbf{X}^T||p(X_m^\zeta|\mathbf{X}_{\pi_m}^{\zeta-1})) = p(X_m^\zeta|\mathbf{X}_{\pi_m}^{\zeta-1}) \left(p(\mathbf{X}^1) \prod_{\substack{t=2\dots T \\ t \neq \zeta}} \prod_{\substack{i=1\dots r \\ i \neq m}} p(X_i^t|\mathbf{X}_{\pi_i}^{t-1}) \right) \tag{4.3}$$

By using a CMA compliant CTV-DBN to model the time-varying causal relationships, predictions can be made based on the manipulation rule by assigning the conditional probability $p(X_m^\zeta|\mathbf{X}_{\pi_m}^{\zeta-1})$. For example, $p(X_m^\zeta|\mathbf{X}_{\pi_m}^{\zeta-1})$ could be equal to $p(X_5^{20} \leq 20|\mathbf{X}_{\pi_m}^{19})$ looking for the impact to $\mathbf{X}^{21} \dots \mathbf{X}^T$ if X_5^{20} has been significantly reduced from its average value, says a speed of 100km/h.

5 Related work

Dynamic Bayesian networks (DBN) [17] have been used to model sequences of variables and regarded as a method to overcome the expressive power limitation in Hidden Markov models and Kalman filter models. This is accomplished by allowing the state-space to be represented in factored form. Although the

name of DBN may give us an impression that it can model time-varying process, the reality is that DBN is in fact a time-invariant model. The structure of the network is fixed but is capable to model dynamic systems [25]. Non-stationary dynamic Bayesian networks (NS-DBN) are introduced by Robinson and Hartemink [23, 24] in recent time. They allow one to model the structure of a network which is not fixed but to evolve over time. Markov chain Monte Carlo (MCMC) sampling method is proposed to be used in its structure learning; however, Song *et al.* [25] point out that such an approach is unlikely to be scalable, and it is prone to the problem of over-fitting. In parallel, Grzegorczyk and Husmeier [9, 10] also developed an alternative approach. Their assumption of a fixed network structure is deemed to be too restrictive [4], even though the interaction parameters of the model is allowed to vary with time to cater for non-stationary systems. Song *et al.* propose TV-DBN [25] to overcome those weaknesses.

DBSCAN [5] is one of the popular density-based methods that groups related objects by using density threshold. OPTICS [1] is an alternative method similar to DBSCAN but can better handle data with varying densities. A more recent variant to DBSCAN is ST-DBSCAN [3]. ST-DBSCAN clusters spatial-temporal data by non-spatial, spatial and temporal attributes, which makes the need to consider noise objects redundant by assigning density factor to each cluster; hence, no noise object is required to be detected. The conflicts of borderline objects are overcome by comparing the average value of a cluster with new incoming value. Density-based clustering methods are argued to be the best solution for trajectory clustering [18].

The theory of statistical causal inference developed by Perl [19] and Spirtes *et al.* [27] provides a platform allowing for causal relationship to be detected based on observations. In recent time, Pellet and Elisseeff [21] attempt to provide a causal structure-learning algorithm for causally insufficient data and show that their Markov blanket/collidet set (MBCS*) algorithm is in several orders of magnitude faster than the popular Fast Causal Inference (FCI) algorithm [28]. Another example is Inductive Causation* (IC*) [20], in which both algorithms IC* and FCI are done based on relaxing the causal sufficiency assumption. CTV-DBN is different from all of them as it takes the time-varying network structure into consideration while at the same time satisfying CMA. To the best of our knowledge, CTV-DBN is the first proposal in causal network learning which considers time-varying network structure.

6 Experiments - an application to spatio-temporal analysis

We evaluate our models on spatio-temporal data by using Beijing taxi trajectories from Complex Engineered Systems Lab, Tsinghua University, China[1]. The dataset consists of one month of trajectories of 28,000 taxis in Beijing captured in May 2009. The trajectories are firstly passed through a spatial filter with a boundary[2] of Beijing city centred at the Forbidden City (city centre) and extended to its three international airport terminals (top right hand corner). We

[1] http://sensor.ee.tsinghua.edu.cn/datasets.php

[2] a rectangle formed by latitude and longitude pairs (40.08200, 116.16054) and (39.75030, 116.62000)

then apply density-based clustering method DBSCAN on one week (Monday-Friday) of taxi trajectories at 8am to obtain a driver's view of regions in the city as shown in Figure 6.1. Please note that top right hand corner is the location of the airports (regions 2 and 6), and we can observe an expected region structure complexity from the city centre to the airports. The trajectory average speeds within clusters are calculated.

We go through the RTV-DBN and CTV-DBN structure estimation to come up with network structures as shown in Figure 6.2 for $t = 20$ (8:20am) and $t = 30$ (8:30am) by using different methods and kernels, where the cells filled with black colour represent connections and blank otherwise. Shooting algorithm [6] is used to speed up the optimization calculation in which the cost function is transformed into a standard ℓ_1-regularized least squares problem by pushing in the weighting function into the squared loss function. A nil smoothing approach of CTV-DBN with no kernel produces a rigid and discontinuous structure as observed in Figures 6.2(e) and 6.2(f). Although they might not be useful at first sight but we can clearly identify regions 3, 6, 11, 16, 22, 23, 24 and 28 (example region list) are the ones heavily depending on nearly all regions in the city. Apart from region 16, all the other regions are between the city centre and the airports.

On the other hand, because of truncation at causal boundary ($\mathcal{B}$), the method of CTV-DBN with truncated $K_{\mathcal{G}}$ suffers from bias as well as information loss (Figures 6.2(c) and 6.2(d)) and it is confirmed when comparing with the results from RTV-DBN with $K_{\mathcal{G}}$ (Figures 6.2(a) and 6.2(b)), although a subset of the structure can be recognised. Finally, the method of CTV-DBN with K_{Γ} not only comes with a theoretical strength of low bias at the causal boundary ($\mathcal{B}$) and complies with CMA, it also reveals more details of causal relationships between regions. Based on the same λ, regions with insufficient causal connections are eliminated in CTV-DBN (versus RTV-DBN) and additional connections are added based on the evidence within causal boundary ($\mathcal{B}$) in $\mathcal{S}$. The structural difference between CTV-DBN with K_{Γ} and RTV-DBN with $K_{\mathcal{G}}$ are mainly from the enforced causal relationship in the former. Out of the example region list above, only regions 3, 24 and 28 are the top 3 regions causally impacted by most of the regions and they are all located along the Beijing Airport Expressway (S12)[3] between the city and the airports, in which traffic jam is common[4]. Since the three regions are located just at or before ring roads[5] which diverge traffic to all major districts in the city, any congestion in the other regions would have ultimate impact to these three regions (major artery between the city and the airports). These findings have also been confirmed by numerous published facts such as 1) a report by China Central Television[6] and 2) a discussion (section 3.3.1) in Papers in Regional Science[7], about the relationship between ring roads and other districts in the Beijing city.

[3] http://en.wikipedia.org/wiki/Airport_Expressway_(Beijing)

[4] http://wikitravel.org/en/Beijing, http://www.bjjtgl.gov.cn/publish/portal1/

[5] http://en.wikipedia.org/wiki/Ring_roads_of_Beijing

[6] http://www.cctv.com/lm/124/41/90128.html

[7] The impact of urban growth on commuting patterns in a restructuring city: Evidence from Beijing, 2011

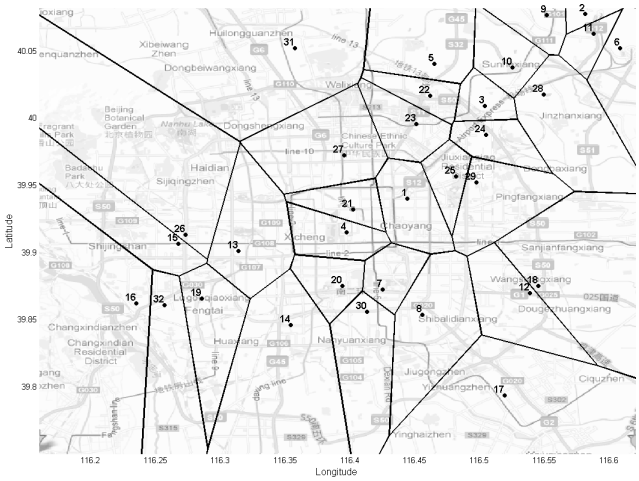


Figure 6.1: Regions drawn by Voronoi diagrams at peak hour (8am) based on one week (Monday-Friday) of Beijing taxi trajectories

7 Conclusion

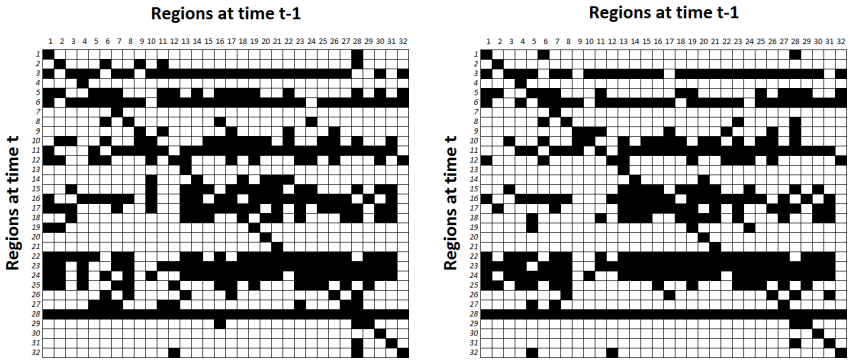
This paper presents notions of relaxed time-varying dynamic Bayesian networks (RTV-DBN) and causal time-varying dynamic Bayesian networks (CTV-DBN), as well as defines causal boundary ($\mathcal{B}$). In CTV-DBN, causal Markov assumption (CMA) is satisfied by replacing a symmetric Gaussian kernel ($K_{\mathcal{G}}$) with an asymmetric Gamma kernel (K_{Γ}). K_{Γ} does not only fit well within the causal boundary ($\mathcal{B}$) in $\mathcal{S}$ as defined in Section 3, but also provides low bias in dealing with boundary problem, that would be significant if a symmetric kernel is used due to truncation. Upon satisfying causal Markov assumption, causal inference can be made based on the manipulation rule derived in this paper. We apply RTV-DBN and CTV-DBN on spatio-temporal data by combining the model with the techniques of density-based clustering, Voronoi diagram, etc. By learning the time-varying region structures using moving objects' view of territories, causal relationships among regions are captured and available for causal inference. The findings learnt from real-life Beijing taxi data using the proposed method are consistent with the known facts as discussed in Section 6.

Bibliography

[1] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In *SIGMOD Conference*, pages 49–60, 1999.

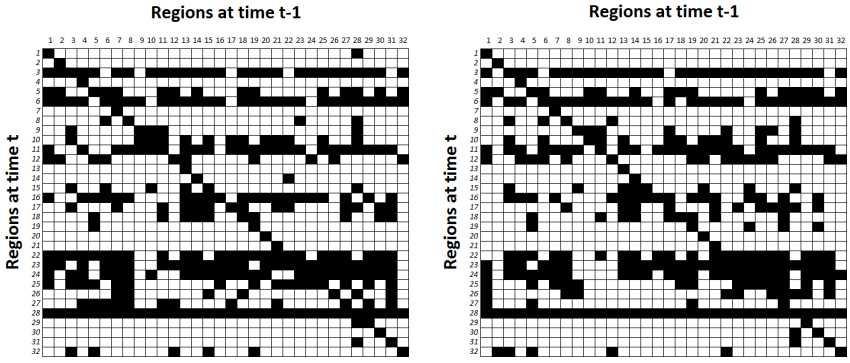
[2] L. J. Bain and M. Engelhardt. *Introduction to probability and mathematical statistics*, volume 4. Duxbury Press Belmont, CA, 1992.

[3] D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*, 60(1):208–221, 2007.



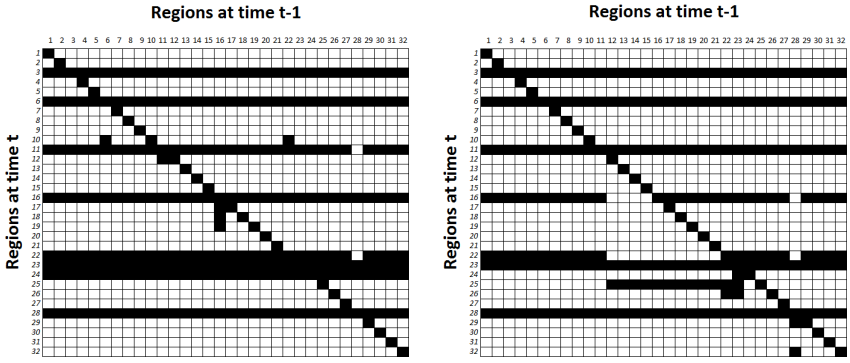
(a) RTV-DBN ($K_{\mathcal{G}}$) at t=20

(b) RTV-DBN ($K_{\mathcal{G}}$) at t=30



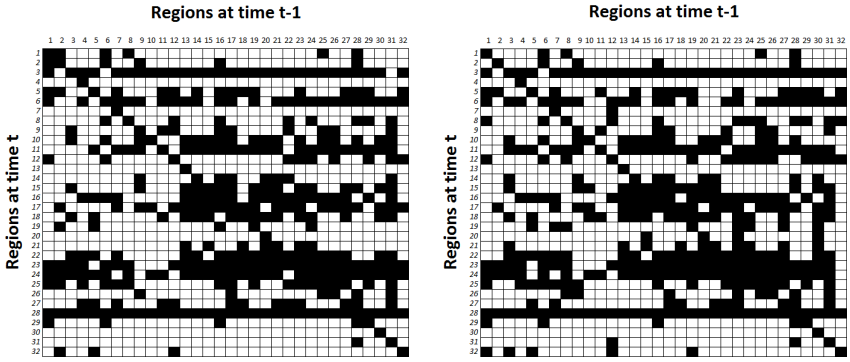
(c) CTV-DBN (truncated $K_{\mathcal{G}}$) at t=20

(d) CTV-DBN (truncated $K_{\mathcal{G}}$) at t=30



(e) CTV-DBN (No kernel) at t=20

(f) CTV-DBN (No kernel) at t=30



(g) CTV-DBN (K_{Γ}) at t=20

(h) CTV-DBN (K_{Γ}) at t=30

Figure 6.2: $\hat{\mathcal{E}}^t$ in adjacency matrix form representing the directed edges from components of $\mathbf{X}^{t-1}$ (columns) to components of $\mathbf{X}^t$ (rows) based on $\hat{\mathbf{A}}^t$ with $\lambda = 0.4$

[4] F. Dondelinger, S. Lèbre, and D. Husmeier. Non-homogeneous dynamic bayesian networks with bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Machine Learning*, 90(2):191–230, 2013.

[5] M. Ester, H. peter Kriegel, J. S, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[6] W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[7] N. Gospodinov and M. Hirukawa. Time series nonparametric regression using asymmetric kernels with an application to estimation of scalar diffusion processes. 2008.

[8] N. Gospodinov and M. Hirukawa. Nonparametric estimation of scalar diffusion models of interest rates using asymmetric kernels. *Journal of Empirical Finance*, 2012.

[9] M. Grzegorczyk and D. Husmeier. Non-stationary continuous dynamic bayesian networks. In *NIPS*, pages 682–690, 2009.

[10] M. Grzegorczyk and D. Husmeier. Non-homogeneous dynamic bayesian networks for continuous data. *Machine Learning*, 83(3):355–419, 2011.

[11] P. Hall and T. E. Wehrly. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, 86(415):665–672, 1991.

[12] S. Kisilevich, F. Mansmann, M. Nanni, and S. Rinzivillo. Spatio-temporal clustering. In *Data Mining and Knowledge Discovery Handbook*, pages 855–874. 2010.

[13] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[14] J. F. Lemmer. The causal markov condition, fact or artifact?, sigart 7(3. *SIGART Bulletin*, pages 7–3, 1996.

[15] W. Liu, Y. Zheng, S. Chawla, J. Yuan, and X. Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *KDD*, pages 1010–1018, 2011.

[16] M. Mackenzie and A. K. Tieu. Asymmetric kernel regression. *IEEE Transactions on Neural Networks*, 15(2):276–282, 2004.

[17] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, 2002.

[18] M. Nanni and D. Pedreschi. Time-focused clustering of trajectories of moving objects. *J. Intell. Inf. Syst.*, 27(3):267–289, 2006.

[19] J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[20] J. Pearl and T. Verma. A theory of inferred causation. In *KR*, pages 441–452, 1991.

[21] J.-P. Pellet and A. Elisseeff. Finding latent causes in causal networks: an efficient approach based on markov blankets. In *NIPS*, pages 1249–1256, 2008.

[22] A. Rinaldo, A. Singh, R. Nugent, and L. Wasserman. Stability of density-based clustering. *Journal of Machine Learning*, 13:905–948, 2012.

[23] J. W. Robinson and A. J. Hartemink. Non-stationary dynamic bayesian networks. In *NIPS*, pages 1369–1376, 2008.

[24] J. W. Robinson and A. J. Hartemink. Learning non-stationary dynamic bayesian networks. *Journal of Machine Learning Research*, 11:3647–3680, 2010.

[25] L. Song, M. Kolar, and E. P. Xing. Time-varying dynamic bayesian networks. In *NIPS*, pages 1732–1740, 2009.

[26] P. Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11:1643–1662, 2010.

[27] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[28] P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 499–506. Morgan Kaufmann Publishers Inc., 1995.

[29] S. Zhang, R. Karunamuni, and M. Jones. An improved estimator of the density function at the boundary. *Journal of the American Statistical Association*, 94(448):1231–1240, 1999.