Learning Spatial-Temporal Varying Graphs with Applications to Climate Data Analysis

Xi Chen[1] and **Yan Liu**[2] and **Han Liu**[1] and **Jaime G. Carbonell**[1]
1. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
2. IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

Abstract

An important challenge in understanding climate change is to uncover the dependency relationships between various climate observations and forcing factors. Graphical lasso, a recently proposed ℓ_1 penalty based structure learning algorithm, has been proven successful for learning underlying dependency structures for the data drawn from a multivariate Gaussian distribution. However, climatological data often turn out to be non-Gaussian, e.g. cloud cover, precipitation, etc. In this paper, we examine nonparametric learning methods to address this challenge. In particular, we develop a methodology to learn dynamic graph structures from spatial-temporal data so that the graph structures at adjacent time or locations are similar. Experimental results demonstrate that our method not only recovers the underlying graph well but also captures the smooth variation properties on both synthetic data and climate data.

Introduction

Climate change poses many critical socio-technological issues in the new century (IPCC 2007). An important challenge in understanding climate change is to uncover the dependency relationships between the various climate observations and forcing factors, which can be of either natural or anthropogenic (human) origin, e.g. to assess which parameters are mostly responsible for climate change.

Graph is one of the most natural representations of dependency relationships among multiple variables. There have been extensive studies on learning graph structures that are invariant over time. In particular, ℓ_1 penalty based learning algorithms, such as graphical lasso, establish themselves as one of the most promising techniques for structure learning, especially for data with inherent sparse graph structures (Meinshausen and Bühlmann 2006; Yuan and Lin 2007) and have been successfully applied in diverse areas, such as gene regulatory network discovery (Friedman 2004), social network analysis (Goldenberg and Moore 2005) and so on. Very recently, several methods have been proposed to model time-evolving graphs with applications from gene regulatory network analysis (Song, Kolar, and Xing 2009), financial data analysis (Xuan and Murphy 2007) to oil-production monitoring system (Liu, Kalagnanam, and Johnsen 2009). Most of the existing methods assume that the data are drawn from a multivariate Gaussian distribution at each time stamp and then estimate the graphs on a chain of time.

Compared with the existing approaches for graph structure learning, there are two major challenges associated with climate data: one is that meteorological or climatological data often turn out to be non-Gaussian, e.g. precipitation, cloud cover, and relative humidity, which belong to bounded or skewed distributions (Boucharel et al. 2009); the other is the smooth variation property, i.e. the graph structures may vary over temporal and/or spatial scales, but the graphs at adjacent time or locations should be similar.

In this paper, we present a nonparametric approach with kernel weighting techniques to address these two challenges for spatial-temporal data in climate applications. Specifically, for a fixed time t and location s, we propose to adopt a two-stage procedure: (1) instead of blindly assuming that the data follow Gaussian or any other parametric distributions, we learn a set of marginal functions which can transform the original data into a space where they are normally distributed; (2) we construct the covariance matrix for t and s via a kernel weighted combination of all the data at different time and locations. Then the state-of-the-art graph structure learning algorithm, "graphical lasso" (Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008), can be applied to uncover the underlying graph structure. It is worthwhile noting that our kernel weighting techniques are very flexible, i.e. they can account for smooth variation in many types (e.g. altitude) besides time and space. To the best of our knowledge, it is the first practical method for learning non-stationary graph structures without assuming any parametric underlying distributions.

Preliminary

We concern ourselves with the problem in learning graph structures which vary in both temporal and spatial domains. At each time t and location s, we take n *i.i.d.* observations on p random variables which are denoted as $\{\mathbf{X}_i^{ts}\}_{i=1}^n$, where each $\mathbf{X}_i^{ts} := (X_{i1}^{ts}, \dots X_{ip}^{ts})^T \in \mathbb{R}^p$ is a p dimensional vector. Taking the climate data for example, we may independently measure several factors (variables), such as temperature, precipitation, carbon-dioxide (CO_2), at each location spreading at different time in a year. Our goal is to explore the dependency relationships among these variables

over time and locations.

Markov Random Fields (MRFs) have been widely adopted for modeling dependency relationships (Kinderman and Snell 1980). For a fixed time and location, denote each observation as a p-dimensional random vector $\mathbf{X} = (X_1, \dots X_p)$. We encode the structure of $\mathbf{X}$ with an undirected graph $G = (V, E)$, where each node u in the vertex set $V = \{v_1, \dots v_p\}$ corresponds to a component of $\mathbf{X}$. The edge set encodes conditional independencies among components of $\mathbf{X}$. More precisely, the edge between (u, v) is excluded from E if and only if X_u is conditionally independent of X_v given the rest of variables $V_{\setminus u,v} \equiv \{X_i, 1 \le i \le p, i \ne u, v\}$:

$$(u, v) \notin E \Leftrightarrow X_u \perp\!\!\!\perp X_v \,|\, V_{\setminus u,v} \tag{1}$$

A large body of literature assumes that $\mathbf{X}$ follows a multivariate Gaussian distribution, $N(\mu, \Sigma)$, with the mean vector μ and the covariance matrix Σ. Let $\Omega = \Sigma^{-1}$ be the inverse of the covariance matrix (a.k.a. the precision matrix). One good property of multivariate Gaussian distributions is that $X_u \perp\!\!\!\perp X_v \,|\, V_{\setminus u,v}$ if and only if $\Omega_{uv} = 0$ (Lauritzen 1996). Under the Gaussian assumption, we may deduce conditional independencies by estimating the inverse covariance matrix. In real world applications, many variables are conditionally independent given others. Therefore, only a few essential edges should appear in the estimated graph. In other words, the estimated inverse covariance matrix $\widehat{\Omega}$ should be sparse with many zero elements.

Inspired by the success of "lasso" for linear models, Yuan and Lin proposed "graphical lasso" to obtain a sparse $\widehat{\Omega}$ by minimizing the negative log-likelihood with ℓ_1 penalization on $\widehat{\Omega}$ (Yuan and Lin 2007). More precisely, let $\{\mathbf{X}_1, \mathbf{X}_2, \dots \mathbf{X}_n\}$ be n random samples from $N(\mu, \Sigma)$, where each $\mathbf{X}_i \in \mathbb{R}^p$ and let $\widehat{\Sigma}$ be the estimated covariance matrix using maximum likelihood:

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T,$$

where $\overline{\mathbf{X}}$ is the sample mean. The estimator $\widehat{\Omega}$ is obtained by minimizing:

$$-\ell(\mu, \Omega) + \lambda \sum_{j<k} |\Omega_{jk}|, \tag{2}$$

where

$$\ell(\mu, \Omega) = \frac{1}{2} \left(\log |\Omega| - \text{tr}(\Omega \widehat{\Sigma}) - p \log(2\pi) \right), \tag{3}$$

is the log-likelihood and λ is the tuning parameter that controls the sparsity of $\widehat{\Omega}$. The minimization can be done efficiently using the algorithm in (Friedman, Hastie, and Tibshirani 2008), which is a block coordinate descent algorithm that updates a single row and column of Ω at each iteration. It has been proven that, under certain conditions, $\widehat{\Omega}$ can recover the edge set of the underlying true graph with high probability (Ravikumar et al. 2008).

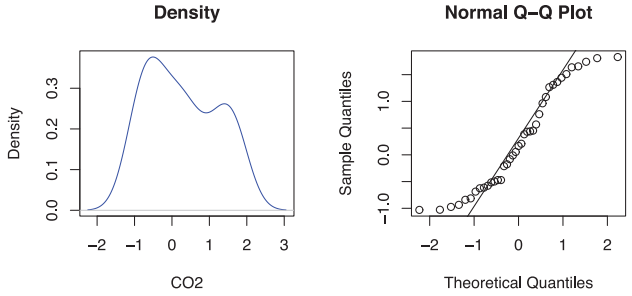

Figure 1: Density plot (left) and Q-Q plot (right) of raw CO_2 data

Nonparanormal

As discussed in the previous section, graphical lasso can estimate an inverse covariance matrix with good statistical properties as long as the data are drawn from a multivariate Gaussian distribution. However, this is not the case in many applications. Take our climate data for example, we have 39 measurements of CO_2 at a location on California coast in the first quarter over 13 years (1990 $\sim$ 2002). The density and Q-Q plot are presented in Figure 1. The p-value of Anderson-Darling normality test is $0.0133 < 0.05$, which rejects the null hypothesis and hence indicates that samples are not normally distributed.

However, in many cases, it is possible to find a set of marginal functions to transform the original data into another space so that they are normally distributed. More precisely, if there exists a set of univariate functions $\{f_j\}_{j=1}^p$ such that $f(\mathbf{X}) \equiv (f_1(X_1), \dots, f_p(X_p)) \sim N(\mu, \Sigma)$, we say that $\mathbf{X}$ follows a *nonparanormal* (*NPN*) distribution (Liu, Lafferty, and Wasserman 2009) and denote it as:

$$\mathbf{X} \sim NPN(\mu, \Sigma, f). \tag{4}$$

Given n p-dimensional samples, $\{\mathbf{X}_1, \mathbf{X}_2, \dots \mathbf{X}_n\}$, drawn from $NPN(\mu, \Sigma, f)$, we adopt the method in (Liu, Lafferty, and Wasserman 2009) to find a set of good estimators of $\{f_j\}_{j=1}^p$.

We constrain each f_j to be monotone and differentiable. Moreover, we demand that f_j preserves the mean and variance for the identifiability consideration:

$$\begin{aligned} \mu_j &= \mathbb{E}\left(f_j(X_j)\right) = \mathbb{E}(X_j), \\ \sigma_j^2 &= \Sigma_{jj} = \text{Var}\left(f_j(X_j)\right) = \text{Var}\left(X_j\right). \end{aligned}$$

To find a good estimator of f_j, we start by writing down the cumulative distribution function (CDF) of f_j under our basic assumption $f_j(X_j) \sim N(\mu, \Sigma)$:

$$\mathbb{P}\left(f_j(X_j) \le f_j(x)\right) = \Phi\left(\frac{f_j(x) - \mu_j}{\sigma_j}\right), \tag{5}$$

where $\Phi(\cdot)$ is the CDF of a standard Gaussian distribution. Let $F_j(x) = \mathbb{P}(X_j \le x)$ denote the CDF of X_j, the monotone property of f_j implies that

$$F_j(x) = \mathbb{P}(X_j \le x) = \mathbb{P}\left(f_j(X_j) \le f_j(x)\right). \tag{6}$$

Connecting (6) and (5), we obtain

$$F_j(x) = \Phi\left(\frac{f_j(x) - \mu_j}{\sigma_j}\right), \tag{7}$$

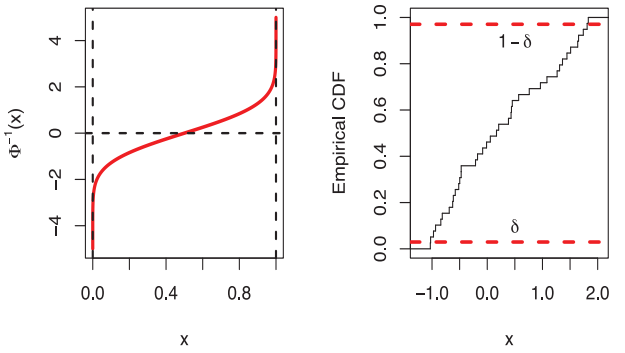

Figure 2: Inverse of the CDF of a standard Gaussian distribution (left). The truncated empirical CDF with red lines representing the truncations (right).

which implies that

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}\left(F_j(x)\right). \tag{8}$$

By substituting μ_j, σ_j and $F_j(x)$ in (8) with the sample mean $\widehat{\mu}_j$, the sample standard deviation $\widehat{\sigma}_j$ and the empirical CDF $\widehat{F}_j(x)$ defined as below, we obtain an estimator of f_j.

$$\begin{aligned} \widehat{\mu}_j &= \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad \widehat{\sigma}_j = \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}\left(X_{ij}-\widehat{\mu}_j\right)^2} \\ \widehat{F}_j(x) &\equiv \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_{ij}\le x\}} \end{aligned}$$

However, the transformation function (8) is not well defined since $\Phi^{-1}(x)$ approaches negative or positive infinity when x approaches to 0 or 1 as shown in the left panel of Figure 2. To tackle this problem, we use the approach in (Liu, Lafferty, and Wasserman 2009) to truncate the empirical CDF $\widehat{F}_j(x)$ in the following manner so that our new CDF estimator $\widetilde{F}_j(x)$ is bounded away from 0 and 1:

$$\widetilde{F}_j(x) = \begin{cases} \delta & \text{if } \widehat{F}_j(x) < \delta \\ \widehat{F}_j(x) & \text{if } \delta \le \widehat{F}_j(x) \le 1-\delta \\ (1-\delta) & \text{if } \widehat{F}_j(x) > 1-\delta, \end{cases} \tag{9}$$

where δ is a truncation parameter. The truncated empirical CDF for our motivating example is presented in the right panel of Figure 2. We set the truncation parameter δ to be $1/(4n^{1/4}\sqrt{\pi \log n})$ as in (Liu, Lafferty, and Wasserman 2009), which leads to $O_P(\log(n)/n^{1/4})$ rate of convergence of $\widehat{\Omega}$.

By plugging $\widetilde{F}_j(x)$, $\widehat{\mu}_j$ and $\widehat{\sigma}_j$ back into (8), we obtain our estimator of f_j:

$$\widetilde{f}_j(x) \equiv \widehat{\mu}_j + \widehat{\sigma}_j \Phi^{-1}\left(\widetilde{F}_j(x)\right), \tag{10}$$

After taking the transformations in (10), the data are mapped into $\{\widetilde{f}(\mathbf{X}_1), \widetilde{f}(\mathbf{X}_2), \ldots \widetilde{f}(\mathbf{X}_n)\}$. The maximum likelihood estimator of the mean and the covariance matrix

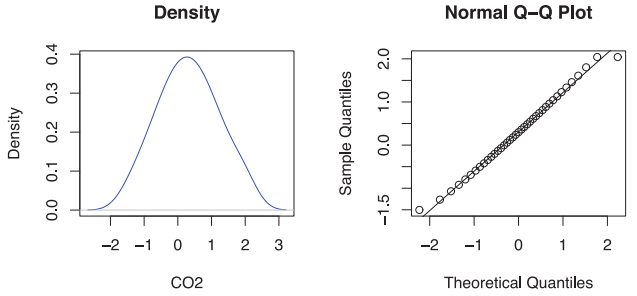

Figure 3: Density plot (left) and Q-Q plot (right) of the transformed CO_2 data

takes the following form:

$$\begin{aligned} \widetilde{\mu} &\equiv \frac{1}{n}\sum_{i=1}^{n} \widetilde{f}(\mathbf{X}_i), \\ \widetilde{\Sigma} &\equiv \frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{f}(\mathbf{X}_i)-\widetilde{\mu}\right)\left(\widetilde{f}(\mathbf{X}_i)-\widetilde{\mu}\right)^T. \end{aligned} \tag{11}$$

Back to our motivating example at the beginning of the section, we present the density and Q-Q plot of the transformed data in Figure 3. The p-value of Anderson-Darling normality test is 0.9995 which strongly indicates that transformed data are normally distributed.

Kernel Weighted Covariance Matrix

For a fixed time t and location s, we have n measurements $\{\mathbf{X}_i^{ts}\}_{i=1}^n$, where $\mathbf{X}_i^{ts} \in \mathbb{R}^p$. Assuming $\mathbf{X}^{ts} \sim NPN(\mu^{ts}, \Sigma^{ts}, f^{ts})$, we estimate f^{ts} by $\widetilde{f}^{ts}$ in (10) and obtain the covariance matrix $\widetilde{\Sigma}^{ts}$ in (11). By substituting $\widehat{\Sigma}$ in (3) with $\widetilde{\Sigma}^{ts}$ and minimizing (2), we obtain the estimated inverse covariance matrix for a single time and location.

However, this simple approach does not take into account the rich information on temporal and spatial constraints, i.e. the graph structures of two adjacent locations (e.g. New York and New Jersey) should be more similar than those of two faraway locations (e.g. New York and San Francisco). Similarly, the difference between graphs in winter and spring should be smaller than that between graphs in winter and summer. To capture this smooth variation property, we use all the data at different time and locations to construct a weighted covariance matrix $\widehat{S}^{ts}$ as an estimator for the covariance matrix at t and s:

$$\widehat{S}^{ts} = \sum_{t'}\sum_{s'} w_{tt'ss'}\widetilde{\Sigma}^{t's'}, \tag{12}$$

where $w_{tt'ss'}$ is the weighting of the difference between time location pairs (t, s) and (t', s'). The idea behind the kernel weighting technique is that all the data should contribute to the estimated covariance matrix at t and s. The smooth variation property requires that when t' is close to t and/or s' is adjacent to s, $w_{tt'ss'}$ should be large since the data from t' and s' are more important for constructing $\widehat{S}^{ts}$.

A natural way to define $w_{tt'ss'}$ is to utilize the product kernel:

$$w_{tt'ss'} = \frac{K_{h_t}(|t-t'|)K_{h_s}(\|s-s'\|_2)}{\sum_{t''}\sum_{s''} K_{h_t}(|t-t''|)K_{h_s}(\|s-s''\|_2)}, \tag{13}$$

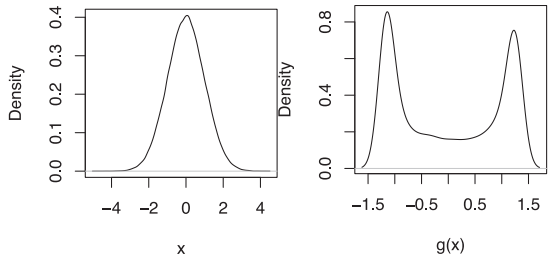

Figure 4: Density plot for the normally distributed data (left) and transformed data (right) by the Gaussian CDF transformation

where $K_h(\cdot) = \frac{1}{h}K\left(\frac{\cdot}{h}\right)$ is a symmetric nonnegative kernel function and h_t, h_s are the kernel bandwidths for time and space. For example, one of the most widely adopted kernel functions is Gaussian RBF kernel where $K_h(t) = \frac{1}{\sqrt{2\pi}h}\exp(-\frac{t^2}{2h^2})$.

In this paper, each time stamp is represented by a single discrete number. The absolute value of the difference between two time stamps is adopted to measure their distance. Each location is represented as a two dimensional vector composed of its longitude and latitude. And the distance between two locations is defined by their Euclidean distance, i.e. their vector 2 norm.

Note that this kernel weighting technique is very flexible. For example, if we have more continuity constraints besides time and space, we can easily extend the product kernel to include all these conditions to enforce the smooth variation effect.

With the weighting technique above, the estimated covariance matrices, $\widehat{S}^{ts}$, are "smooth" in time and space, i.e. estimated covariance matrices of two adjacent time stamps or places should not differ too much. Then we plug $\widehat{S}^{ts}$ into (3) to replace $\widehat{\Sigma}$ and obtain the estimated sparse inverse covariance matrix $\widehat{\Omega}^{ts}$.

Experiments

In our experiment, we compare four different methods on both synthetic data and our motivating climate dataset:

1. Kernel Weighted Nonparanormal: taking the transformation in (10) and using the kernel weighted estimator of the covariance matrix in (12).
2. Kernel Weighted Normal: using the kernel weighted estimator of the covariance matrix based on the raw data.
3. Nonparanormal: taking the transformation in (10) but without the kernel weighting step.
4. Normal: directly computing the sample covariance at each time and location.

Synthetic Data

For the synthetic data experiment, we only consider time-varying graphs for ease of illustration. In fact, we can adapt our method to time-varying graphs simply by replacing the product kernel in (13) with a kernel only involving time: $K_{h_t}(|t-t'|)$. We set the number of nodes $p = 20$, the number of edges $e = 15$, the number of time stamps $T = 20$, the sample size for each time stamp $n = 50$ and the maximum node degree to be 4. The observation sequence for synthetic Markov Random Fields are generated as follows:

1. Generate an Erdös-Rényi random graph $G^1 = (V^1, E^1)$. Then from $t = 2$ to T, we construct the graph $G^t = (V^t, E^t)$ by randomly adding one edge and removing one edge from G^{t-1} and taking care that the maximum node degree is still 4.
2. For each graph G^t, generate the inverse covariance matrix Ω^t as in (Meinshausen and Bühlmann 2006):
$$\Omega^t(i,j) = \begin{cases} 1 & \text{if } i = j, \\ 0.245 & \text{if } (i,j) \in E^t, \\ 0 & \text{otherwise,} \end{cases}$$
where 0.245 guarantees the positive definiteness of Ω^t when the maximum node degree is 4.
3. For each t, we sample n data points from a multivariate Gaussian distribution with mean $\mu = (1.5, \ldots 1.5)$ and covariance matrix $\Sigma^t = (\Omega^t)^{-1}$:
$$\mathbf{Y}_1^t, \ldots, \mathbf{Y}_n^t \sim N(\mu, \Sigma^t),$$
where each $\mathbf{Y}_i^t \in \mathbb{R}^p$.
4. For each $\mathbf{Y}_i^t$, we take the Gaussian CDF transformation $\{g_j(\cdot)\}_{j=1}^p$ on each dimension and generate the corresponding $\mathbf{X}_i^t$:
$$\mathbf{X}_i^t = \left(X_{i1}^t, \ldots, X_{ip}^t\right) = \left(g_1(Y_{i1}^t), \ldots, g_p(Y_{ip}^t)\right).$$

The Gaussian CDF transformation function $g(x)$ takes the basic form of $\Phi(\frac{x-\mu_g}{\sigma_g})$ and is scaled to preserve mean and variance. In general, it transforms a standard Gaussian data into a bi-modal distribution as shown in Figure 4. Here we omit the rigorous definition due to space limitations; interested readers may refer to (Liu, Lafferty, and Wasserman 2009).

We run four methods with the bandwidth $h_t = T \cdot \frac{5.848}{N^{1/3}} = 11.7$, where $N = n \cdot T = 1000$ is the total number of data points and $\frac{5.848}{N^{1/3}}$ is a widely adopted plug-in bandwidth for nonparametric learning. We independently simulate the above procedure for 50 times and evaluate different methods based on F1-Score which is the harmonic mean of precision and recall in retrieving the true graph edges. The result is presented in the left panel of Figure 5. As we can see, at all 20 time stamps, Kernel Weighted Nonparanormal achieves significantly higher F1-Score as compared to other methods. Moreover, we plot the ROC curve at $t = 1$ for randomly selected simulation on the right of Figure 5. Kernel Weighted Nonparanormal is still superior to other methods. For other time, the ROC curves exhibit similar patterns.

Climate Data

We run our proposed method on a climate dataset (Lozano et al. 2009), which contains monthly data of 18 different climatological factors from 1999 to 2002. The observations span 125 locations in the U.S. on an equally spaced grid with the range of latitude from 30.475 to 47.975 and the range of longitude from -119.75 to -82.25. Each location s is denoted by (s_1, s_2) where s_1 is the latitude and s_2 is

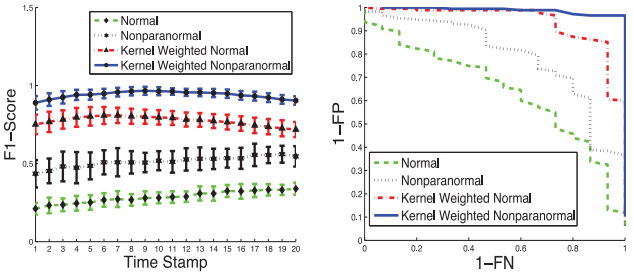

Figure 5: F1-Score (left) and ROC Curve (right)

the longitude. The 18 climatological factors measured for each month include CO_2, CH_4, H_2, CO, average temperature (TMP), diurnal temperature range (DTR), minimum temperate (TMN), maximum temperature (TMX), precipitation (PRE), vapor (VAP), cloud cover (CLD), wet days (WET), frost days (FRS), global solar radiation (GLO), direct solar radiation (DIR), extraterrestrial radiation (ETR), extraterrestrial normal radiation (ETRN) and UV aerosol index (UV). (for more details, see (Lozano et al. 2009)).

At a specific location, we divide a year into 4 quarters and treat the data in the same quarter of all years as separate *i.i.d.* observations from a nonparanormal distribution. We set the tuning parameter $\lambda = 0.15$ to enforce a moderate sparsity of graphs. We use Gaussian RBF kernel with the bandwidth $h_t = 4 \cdot \frac{5.848}{n^{1/3}} = 0.87$ where 4 is the number of quarters in a year and $n = 19500$ is the total number of observations. Similarly, the bandwidth h_s is set to be $\max_{s,s'}(\|s - s'\|_2) \cdot \frac{5.848}{n^{1/3}} = 9$.

As a typical example, we show in Figure 6 the estimated graphs for 4 quarters at the location (30.475 , -114.75) which is a place in CA south of San Diego. We see that the graph structures are quite "smooth" between every two adjacent quarters except for quarter 3 and 4. It indicates that the climate may change more significantly between 3rd and 4th quarter. Another interesting observation is that some factors are clustered on the graph, such as $\{CO_2, CH_4, H_2\}$, {DIR, UV, GLO}, etc. It indicates that these factors are highly correlated and should be studied together by meteorologists. In fact, it is quite possible that, due to the greenhouse effect, $\{CO_2, CH_4, H_2\}$ are highly correlated.

We show some examples to illustrate the spatial smoothness. For an adjacent location (32.975, -117.250) (in CA between San Diego and Los Angeles) and a faraway location (42.975, -84.75) (in Michigan) from the one in Figure 6, the estimated graphs for 4 quarters are shown in Figure 7 and Figure 8 respectively. As we can see, there are at most 2 different edges for each quarter between Figure 6 and 7. But Figure 6 and 8 are very different.

Moreover, by varying the tuning parameter λ in (2) from a large value to a small one, we obtain the full regularization path which could be useful to identify the influence of other factors on a specific factor of interest, e.g. CO_2. More precisely, the order in which the edges appear on the regularization path indicates their degree of influence on a particular factor. As an illustration, Figure 9 shows the changing of the edges connecting CO_2 in the first quarter at location

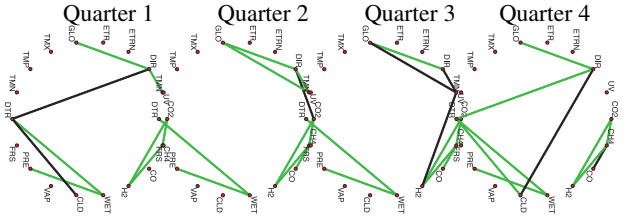

Figure 6: Estimated graphs at location (30.475 , -114.75). The common ones between time (t mod 4) and ($t + 1$ mod 4) are colored as green.

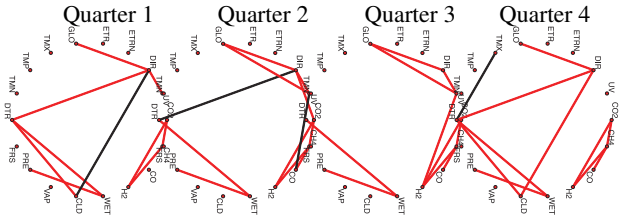

Figure 7: Estimated graphs at location (32.975, -117.250). The edges in common with the corresponding quarter at (30.475 , -114.75) (Figure 6) are colored as red

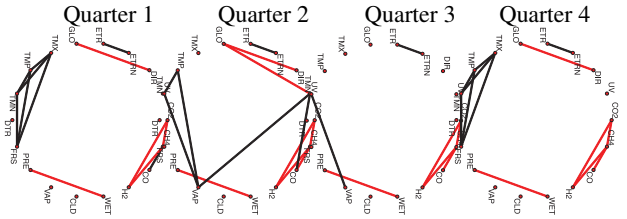

Figure 8: Estimated graphs at location (42.975, -84.75). The edges in common with the corresponding quarter at (30.475 , -114.75) (Figure 6) are colored as red

(30.475 , -114.75). From the plots, we see that the edge between CO_2 and CH_4 appears first, followed by H_2 and then DIR. It indicates that the amount of CH_4 is the most crucial factor to estimate CO_2, and the second is H_2, the third is DIR, etc. The result is quite interpretable in meteorology. In fact, CO_2 is mainly produced by burning fossil fuels which primarily consists of CH_4. In addition, the generating capacity of fossil fuel, formed by organic matters mixed with mud, is directly determined by solar radiation (Chapter 7 in IPCC 2007). These domain facts seem to suggest that the graph structures we learned are quite reasonable. Furthermore, they might be able to provide additional insights to help meteorologists better understand the dependency relationships among these factors.

Finally, we run the full regularization path of Kernel Weighted Normal method and find the graph having the smallest symmetric difference compared to the first quarter of Figure 6. The symmetric difference is plotted in the left of Figure 10, where we see that several factors, such as UV, involve several edges in the symmetric difference graph. It indicates that our marginal transformations on these factors change them substantially. To see this, we select two representative factors, UV and CO and plot their marginal

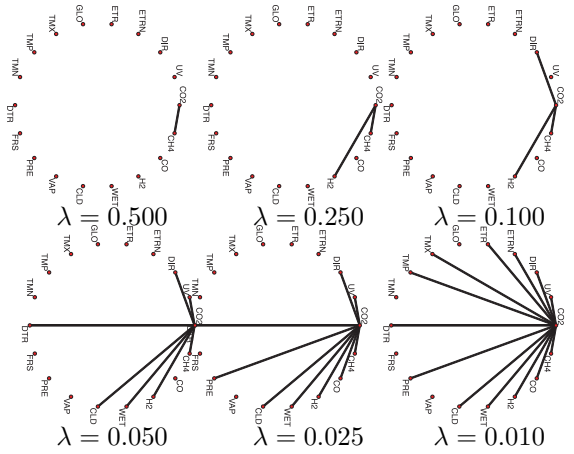

Figure 9: Estimated graphs using different λs in the first quarter at (30.475 , -114.75)

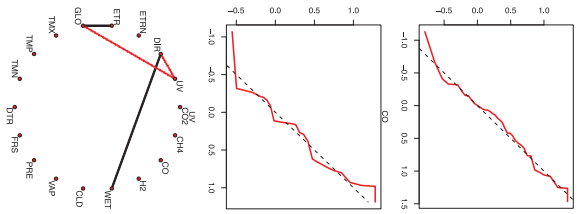

Figure 10: Symmetric difference graph (left) for the first quarter at (30.475 , -114.75). The red edges are those appear in the graph with transformations and the black edges are those appear in the graph without transformations. Estimated transformation for UV (middle) and CO (right). The black dashed lines plot identity map and the red lines indicate the transformations on the data.

transformations in the middle and right of Figure 10. As we can see, for UV, the transformation does change the original data. In contrast, for CO, there is no associated edge in the symmetric difference graph which suggests that the transformation might have no effect. This could be verified by the right panel of Figure 10 where the transformation and identity map on the data nearly coincide.

Conclusion

Motivated by the task of analyzing climate data, we develop a two-stage procedure to learn dynamic graph structures when the underlying distributions are non-Gaussian. In the first stage, we learn a set of marginal functions that transform the data to be normally distributed. In the second stage we use the kernel weighting technique to construct an estimated covariance matrix and then adopt graphical lasso to uncover the underlying graph structure. Empirical results show that our method not only better recovers each single graph structure when the distributions are highly skewed, but also captures the smooth variation property in both spatial and temporal domains.

This paper is a preliminary work aiming at modeling the dependency relationships among different climate factors with powerful structure learning tools. The next step is to collaborate with meteorologists and incorporate domain knowledge constraints to find more interesting structures. Our hope is that the structures we learned could help meteorologists better understand the climatological phenomena and lead to new discoveries in the field of climatological analysis.

Acknowledgement

We thank Aurelie Lozano, Hongfei Li, Alexandru Niculescu-mizil, Claudia Perlich and Naoki Abe for helpful discussion.

References

Boucharel, J.; Dewitte, B.; Garel, B.; and du Penhoat, Y. 2009. Enso's non-stationary and non-gaussian character: the role of climate shifts. *Nonlinear Processes in Geophysics* 16:453–473.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9:432–441.

Friedman, N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303:799–805.

Goldenberg, A., and Moore, A. W. 2005. Bayes net graphs to understand co-authorship networks. In *LinkKDD*.

Climate Change 2007 - the physical science basis *IPCC Fourth Assessment Report*.

Kinderman, R., and Snell, J. L. 1980. *Markov Random Fields and Their Applications*. American Math Society.

Lauritzen, S. L. 1996. *Graphical Models*. Oxford: Clarendon Press.

Liu, Y.; Kalagnanam, J. R.; and Johnsen, O. 2009. Learning dynamic temporal graphs for oil-production equipment monitoring system. In *ACM SIGKDD*.

Liu, H.; Lafferty, J.; and Wasserman, L. 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* 10:2295–2328.

Lozano, A. C.; Li, H.; Niculescu-Mizil, A.; Liu, Y.; Perlich, C.; Hosking, J.; and Abe, N. 2009. Spatial-temporal causal modeling for climate change attribution. In *ACM SIGKDD*.

Meinshausen, N., and Bühlmann, P. 2006. High dimensional graphs and variable selection with the lasso. *The Annals of Stat.* 34:1436–1462.

Ravikumar, P.; Raskutti, G.; Wainwright, M.; and Yu, B. 2008. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. In *Neural Information Processing Systems (NIPS)*.

Song, L.; Kolar, M.; and Xing, E. P. 2009. Time-varying dynamic bayesian networks. In *Neural Information Processing Systems (NIPS)*.

Xuan, X., and Murphy, K. 2007. Modeling changing dependency structure in multivariate time series. In *Intl. Conf. on Machine Learning (ICML)*.

Yuan, M., and Lin, Y. 2007. Model selection and estimation in the gaussian graphical model. *Biometrika* 94:19–35.