Support Consistency of Direct Sparse-Change Learning in Markov Networks

Song Liu
song@sg.cs.titech.ac.jp
Tokyo Institute of Technology

Taiji Suzuki
s-taiji@is.titech.ac.jp
Tokyo Institute of Technology

Masashi Sugiyama
sugi@k.u-tokyo.ac.jp
University of Tokyo

Abstract

We study the problem of learning sparse structure changes between two Markov networks P and Q. Rather than fitting two Markov networks separately to two sets of data and figuring out their differences, a recent work proposed to learn changes *directly* via estimating the ratio between two Markov network models. Such a direct approach was demonstrated to perform excellently in experiments, although its theoretical properties remained unexplored. In this paper, we give sufficient conditions for *successful change detection* with respect to the sample size n_p, n_q, the dimension of data m, and the number of changed edges d. More specifically, we prove that the true sparse changes can be consistently identified for $n_p = \Omega(d^2 \log \frac{m^2+m}{2})$ and $n_q = \Omega(n_p^2/d)$, with an exponentially decaying upper-bound on learning error. Our theoretical guarantee can be applied to a wide range of discrete/continuous Markov networks.

Introduction

Learning changes in interactions between random variables plays an important role in many real-world applications. For example, genes may regulate each other in different ways when external conditions are changed. The number of daily flu-like symptom reports in nearby hospitals may become correlated when a major epidemic disease breaks out. EEG signals from different regions of the brain may be synchronized/desynchronized when the patient is performing different activities. Identifying such changes in interactions helps us expand our knowledge on these real-world phenomena.

In this paper, we consider the problem of learning changes between two undirected graphical models. Such a model, also known as a Markov network (MN) (Koller and Friedman 2009), expresses interactions via the conditional independence between random variables. Among many types of MNs, we focus on *pairwise MNs*, whose joint distribution can be factorized over single or pairwise random variables.

The problem of learning structure of MN itself has been thoroughly investigated in the last decade. The graphical lasso method (Banerjee, El Ghaoui, and d'Aspremont 2008; Friedman, Hastie, and Tibshirani 2008) learns a sparse precision (inverse covariance) matrix from data by using the ℓ_1-norm, while the neighborhood regression methods (Lee, Ganapathi, and Koller 2007; Meinshausen and Bühlmann 2006; Ravikumar, Wainwright, and Lafferty 2010) solve a node-wise lasso program to identify the neighborhood of each single node.

One naive approach to learning changes in MNs is to apply these methods to two MNs separately and compare the learned models. However, such a two-step approach does not work well when the MNs themselves are dense (this can happen even when the change in MNs is sparse). A recent study (Zhang and Wang 2010) adopts a neighbourhood selection procedure to learn sparse changes between Gaussian MNs via a fused-lasso type regularizer (Tibshirani et al. 2005). However, no theoretical guarantee was given on identifying changes. Furthermore, extension of the above mentioned methods to general non-Gaussian MNs is hard due to the computational intractability of the normalization term.

To cope with these problems, an innovative algorithm has been proposed recently (Liu et al. 2014). Its basic idea is to model the changes between two MNs P and Q as the ratio between two MN density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, and the ratio $p(\boldsymbol{x})/q(\boldsymbol{x})$ is directly estimated in one-shot without estimating $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ themselves (Sugiyama, Suzuki, and Kanamori 2012). Since parameters in the density ratio model represent the parametric difference between P and Q, sparsity constrains can be directly imposed for sparse change learning. Thus, the density-ratio approach can work well even when each MN is dense as long as the change is sparse. Furthermore, the normalization term in the density-ratio approach can be approximately computed by the straightforward sample average and thus there is no computational bottleneck in using non-Gaussian MNs. Experimentally, the density-ratio approach was demonstrated to perform excellently. However, its theoretical properties have not been explored yet.

The ability of recovering a sparsity pattern via a sparse learning algorithm has been studied under the name of *support consistency* or *model consistency* (Wainwright 2009; Zhao and Yu 2006), that is, the support of the estimated parameter converges to the true support. Previous works for *successful structure recovery* are available for ℓ_1-regularized maximum (pseudo-)likelihood estimators (Ravikumar, Wainwright, and Lafferty 2010; Yang et al. 2012). However, Liu et al.'s density ratio estimator brought us a new question: what is the sparsistency of identifying

correct sparse changes *without* learning individual MNs? Such a concern is very practical since in many change detection applications, we only care about changes rather than recovering individual structures before or after changes.

In this paper, we theoretically investigate the success of the density-ratio approach and provide sufficient conditions for *successful change detection* with respect to the number of samples n_p, n_q, data dimension m, and the number of changed edges d. More specifically, we prove that if $n_p = \Omega(d^2 \log \frac{m^2+m}{2})$ and $n_q = \Omega(\frac{n_p^2}{d})$, changes between two MNs can be consistently learned under mild assumptions, regardless the sparsity of individual MNs. Technically, our contribution can be regarded as an extension of support consistency of lasso-type programs (Wainwright 2009) to the *ratio* of MNs.

Note that the theoretical results presented in this paper are fundamentally different from previous works on learning a "jumping MN" (Kolar and Xing 2012), where the focuses are learning the partition boundaries between jumps, and the successful recovery of graphical structure within each partition, rather than learning sparse changes between partitions.

Direct Change Learning between Markov Networks

In this section, we review a direct structural change detection method (Liu et al. 2014).

Problem Formulation

Consider two sets of independent samples drawn separately from two probability distributions P and Q on $\mathbb{R}^m$:

$$\{\boldsymbol{x}_p^{(i)}\}_{i=1}^{n_p} \overset{\text{i.i.d.}}{\sim} P \text{ and } \{\boldsymbol{x}_q^{(i)}\}_{i=1}^{n_q} \overset{\text{i.i.d.}}{\sim} Q.$$

We assume that P and Q belong to the family of *Markov networks* (MNs) consisting of univariate and bivariate factors, i.e., their respective probability densities p and q are expressed as

$$p(\boldsymbol{x};\boldsymbol{\theta}^{(p)}) = \frac{1}{Z(\boldsymbol{\theta}^{(p)})} \exp\left(\sum_{u\geq v}^{m} \boldsymbol{\theta}_{u,v}^{(p)\top} \boldsymbol{\psi}(x_u, x_v)\right), \quad (1)$$

where $\boldsymbol{x} = (x_1, \ldots, x_m)^\top$ is the m-dimensional random variable, $u \geq v$ is short for $u, v = 1, u \geq v$ (same below), $\top$ denotes the transpose, $\boldsymbol{\theta}_{u,v}^{(p)}$ is the parameter vector for the elements x_u and x_v with dimension b, and

$$\boldsymbol{\theta}^{(p)} = (\boldsymbol{\theta}_{1,1}^{(p)\top}, \ldots, \boldsymbol{\theta}_{m,1}^{(p)\top}, \boldsymbol{\theta}_{2,2}^{(p)\top}, \ldots, \boldsymbol{\theta}_{m,2}^{(p)\top}, \ldots, \boldsymbol{\theta}_{m,m}^{(p)\top})^\top$$

is the entire parameter vector. $\boldsymbol{\psi}(x_u, x_v) : \mathbb{R}^2 \to \mathbb{R}^b$ is a basis function, and $Z(\boldsymbol{\theta}^{(p)})$ is the normalization factor defined as

$$Z(\boldsymbol{\theta}^{(p)}) = \int \exp\left(\sum_{u\geq v}^{m} \boldsymbol{\theta}_{u,v}^{(p)\top} \boldsymbol{\psi}(x_u, x_v)\right) \mathrm{d}\boldsymbol{x}.$$

$q(\boldsymbol{x};\boldsymbol{\theta}^{(q)})$ is defined in the same way.

Given two parametric models $p(\boldsymbol{x};\boldsymbol{\theta}^{(p)})$ and $q(\boldsymbol{x};\boldsymbol{\theta}^{(q)})$, the goal is to discover *changes in parameters* from P to Q, i.e., $\boldsymbol{\theta}^{(p)} - \boldsymbol{\theta}^{(q)}$.

Density Ratio Formulation for Structural Change Detection

The key idea in (Liu et al. 2014) is to consider the ratio of p and q:

$$\frac{p(\boldsymbol{x};\boldsymbol{\theta}^{(p)})}{q(\boldsymbol{x};\boldsymbol{\theta}^{(q)})} \propto \exp\left(\sum_{u\geq v} (\boldsymbol{\theta}_{u,v}^{(p)} - \boldsymbol{\theta}_{u,v}^{(q)})^\top \boldsymbol{\psi}(x_u, x_v)\right),$$

where $\boldsymbol{\theta}_{u,v}^{(p)} - \boldsymbol{\theta}_{u,v}^{(q)}$ encodes the difference between P and Q for factor $\boldsymbol{\psi}(x_u, x_v)$, i.e., $\boldsymbol{\theta}_{u,v}^{(p)} - \boldsymbol{\theta}_{u,v}^{(q)}$ is zero if there is no change in the factor $\boldsymbol{\psi}(x_u, x_v)$.

Once the ratio of p and q is considered, each parameter $\boldsymbol{\theta}_{u,v}^{(p)}$ and $\boldsymbol{\theta}_{u,v}^{(q)}$ does not have to be estimated, but only their difference $\boldsymbol{\theta}_{u,v} = \boldsymbol{\theta}_{u,v}^{(p)} - \boldsymbol{\theta}_{u,v}^{(q)}$ is sufficient to be estimated for change detection. Thus, in this density-ratio formulation, p and q are no longer modeled separately, but the changes from p to q *directly* as

$$r(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{N(\boldsymbol{\theta})} \exp\left(\sum_{u\geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{\psi}(x_u, x_v)\right), \quad (2)$$

where $N(\boldsymbol{\theta})$ is the normalization term. This direct nature would be more suitable for change detection purposes according to *Vapnik's principle* that encourages avoidance of solving more general problems as an intermediate step (Vapnik 1998). This direct formulation also halves the number of parameters from both $\boldsymbol{\theta}^{(p)}$ and $\boldsymbol{\theta}^{(q)}$ to only $\boldsymbol{\theta}$.

The normalization term $N(\boldsymbol{\theta})$ is chosen to fulfill $\int q(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} = 1$:

$$N(\boldsymbol{\theta}) = \int q(\boldsymbol{x}) \exp\left(\sum_{u\geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{\psi}(x_u, x_v)\right) \mathrm{d}\boldsymbol{x},$$

which is the expectation over $q(\boldsymbol{x})$. This expectation form of the normalization term is another notable advantage of the density-ratio formulation because it can be easily approximated by the sample average over $\{\boldsymbol{x}_q^{(i)}\}_{i=1}^{n_q} \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x})$.

$$\hat{N}(\boldsymbol{\theta}; \boldsymbol{x}_q^{(1)}, \ldots, \boldsymbol{x}_q^{(n_q)}) := \frac{1}{n_q} \sum_{i=1}^{n_q} \exp\left(\sum_{u\geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{\psi}(x_{q,u}^{(i)}, x_{q,v}^{(i)})\right).$$

Thus, one can always use this empirical normalization term for any (non-Gaussian) models $p(\boldsymbol{x};\boldsymbol{\theta}^{(p)})$ and $q(\boldsymbol{x};\boldsymbol{\theta}^{(q)})$.

Direct Density-Ratio Estimation

Density ratio estimation has been recently introduced to the machine learning community and proven to be useful in a wide range of applications (Sugiyama, Suzuki, and Kanamori 2012). In (Liu et al. 2014), a density ratio estimator called the *Kullback-Leibler importance estimation procedure* (KLIEP) for log-linear models (Sugiyama et al. 2008; Tsuboi et al. 2009) was employed in learning structural changes.

For a density ratio model $r(\boldsymbol{x};\boldsymbol{\theta})$, the KLIEP method minimizes the Kullback-Leibler divergence from $p(\boldsymbol{x})$ to $\widehat{p}(\boldsymbol{x}) = q(\boldsymbol{x})r(\boldsymbol{x};\boldsymbol{\theta})$:

$$\begin{aligned}\mathrm{KL}[p\|\widehat{p}] &= \int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})r(\boldsymbol{x};\boldsymbol{\theta})} \mathrm{d}\boldsymbol{x} \\ &= \text{Const.} - \int p(\boldsymbol{x}) \log r(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{3}$$

Note that the density-ratio model (2) automatically satisfies the non-negativity and normalization constraints:

$$r(\boldsymbol{x};\boldsymbol{\theta}) \geq 0 \text{ and } \int q(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} = 1.$$

In practice, one minimizes the negative empirical approximation of the second term in Eq.(3)[1]:

$$\begin{aligned}\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) &= -\frac{1}{n_p} \sum_{i=1}^{n_p} \log \hat{r}(\boldsymbol{x}_p^{(i)};\boldsymbol{\theta}) \\ &= -\frac{1}{n_p} \sum_{i=1}^{n_p} \sum_{u\geq v} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{\psi}(x_{p,u}^{(i)}, x_{p,v}^{(i)}) \\ &+ \log \left(\frac{1}{n_q} \sum_{i=1}^{n_q} \exp \left(\sum_{u\geq v} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{\psi}(x_{q,u}^{(i)}, x_{q,v}^{(i)}) \right) \right).\end{aligned}$$

where

$$\hat{r}(\boldsymbol{x};\boldsymbol{\theta}) = \frac{\exp\left(\sum_{u\geq v} \boldsymbol{\theta}_{u,v}^\top \boldsymbol{\psi}(x_{q,u}, x_{q,v})\right)}{\hat{N}(\boldsymbol{\theta}; \boldsymbol{x}_q^{(1)}, \ldots, \boldsymbol{x}_q^{(n_q)})}.$$

Because $\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta})$ is convex with respect to $\boldsymbol{\theta}$, its global minimizer can be numerically found by standard optimization techniques such as gradient ascent or quasi-Newton methods. The gradient of ℓ_{KLIEP} with respect to $\boldsymbol{\theta}_{u,v}$ is given by

$$\begin{aligned}\nabla_{\boldsymbol{\theta}_{u,v}} \ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) &= -\frac{1}{n_p} \sum_{i=1}^{n_p} \boldsymbol{\psi}(x_{p,u}^{(i)}, x_{p,v}^{(i)}) \\ &+ \frac{1}{n_q} \sum_{i=1}^{n_q} \hat{r}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) \boldsymbol{\psi}(x_{q,u}^{(i)}, x_{q,v}^{(i)}),\end{aligned}$$

that can be computed in a straightforward manner for *any* feature vector $\boldsymbol{\psi}(x_u, x_v)$.

Sparsity-Inducing Norm

To find a sparse change between P and Q, one may regularize the KLIEP solution with a sparsity-inducing norm $\sum_{u\geq v} \|\boldsymbol{\theta}_{u,v}\|$, i.e., the *group-lasso* penalty (Yuan and Lin 2006). Note that the separate density estimation approaches sparsify both $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_q$ so that the difference $\boldsymbol{\theta}_p - \boldsymbol{\theta}_q$ is also sparsified. On the other hand, the density-ratio approach (Liu et al. 2014) directly sparsifies the difference $\boldsymbol{\theta}_p - \boldsymbol{\theta}_q$, and thus this method can still work well even if $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_q$ are dense as long as $\boldsymbol{\theta}_p - \boldsymbol{\theta}_q$ is sparse.

[1]Note that the ℓ_{KLIEP} is the *negative* log-likelihood.

Now we have reached the final objective provided in (Liu et al. 2014):

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) + \lambda_{n_p} \sum_{u\geq v} \|\boldsymbol{\theta}_{u,v}\|. \tag{4}$$

Support Consistency of Direct Sparse-Change Detection

The above density-ratio approach to change detection was demonstrated to be promising in empirical studies (Liu et al. 2014). However, its theoretical properties have not yet been investigated. In this section, we give theoretical guarantees of the convex program (4) on structural change learning. More specifically, we give *sufficient conditions* for detecting correct changes in terms of the sample size n_p and n_q, data dimensions m, and the number of changed edges d, followed by the discussion on the insights we can gain from the theoretical analysis.

Notation

Before introducing our consistency results, we define a few notations. In the previous section, a sub-vector of $\boldsymbol{\theta}$ indexed by (u, v) corresponds to a specific edge of an MN. From now on, we use new indices with respect to the "oracle" sparsity pattern of the true parameter $\boldsymbol{\theta}^*$ for notational simplicity. By defining two sets of *sub-vector indices* $S := \{t' \mid \|\boldsymbol{\theta}_{t'}^*\| \neq 0\}$ and its complement $S^c := \{t'' \mid \|\boldsymbol{\theta}_{t''}^*\| = 0\}$, we rewrite the objective (4) as

$$\begin{aligned}\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) &+ \lambda_{n_p} \sum_{t'\in S} \|\boldsymbol{\theta}_{t'}\| \\ &+ \lambda_{n_p} \sum_{t''\in S^c} \|\boldsymbol{\theta}_{t''}\|.\end{aligned} \tag{5}$$

The support of estimated parameter and its complement are denoted as $\hat{S}$ and $\hat{S}^c$. Sample Fisher information matrix $\mathcal{I} \in \mathbb{R}^{\frac{b(m^2+m)}{2} \times \frac{b(m^2+m)}{2}}$ is the Hessian of the log-likelihood: $\mathcal{I} = \nabla^2 \ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}^*)$. $\mathcal{I}_{AB}$ is a sub-matrix of $\mathcal{I}$ indexed by two sets of indices A and B on rows and columns.

Assumptions

We start our analysis with assumptions and some discussions. Similar to previous researches on sparsity recovery analysis (Wainwright 2009; Ravikumar, Wainwright, and Lafferty 2010), the first two assumptions are made on Fisher Information Matrix.

Assumption 1 (Dependency Assumption)**.** *The sample Fisher Information Matrix $\mathcal{I}_{SS}$ has bounded eigenvalues:*

$$\Lambda_{min}(\mathcal{I}_{SS}) \geq \lambda_{min}.$$

This assumption is to ensure that the model is identifiable. Although Assumption 1 only bounds the smallest eigenvalue of $\mathcal{I}_{SS}$, the largest eigenvalue of $\mathcal{I}$ is in fact, also upper-bounded, as we stated in later assumptions.

Assumption 2 (Incoherence Assumption)**.** *The unchanged edges cannot exert overly strong effects on changed edges:*

$$\max_{t''\in S^c} \|\mathcal{I}_{t''S} {\mathcal{I}_{SS}}^{-1}\|_1 \leq 1 - \alpha,$$

where $\|\boldsymbol{Y}\|_1 = \sum_{i,j} \|\boldsymbol{Y}_{i,j}\|_1$ and $\alpha \in (0, 1]$.

We also make the following assumptions as an analogy to those made in (Yang et al. 2012).

Assumption 3 (Smoothness Assumption on Log-normalization Function)**.** *We assume that the normalization term* $\log \hat{N}(\boldsymbol{\theta})$ [2] *is smooth around its optimal value and has bounded derivatives*

$$\max_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|} \left|\left|\left| \nabla^2 \log \hat{N}(\boldsymbol{\theta}^* + \boldsymbol{\delta}) \right|\right|\right| \leq \lambda_{max}, \quad (6)$$

$$\max_{t \in S \cup S^c} \max_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|} \left|\left|\left| \nabla_{\boldsymbol{\theta}_t} \nabla^2 \log \hat{N}(\boldsymbol{\theta}^* + \boldsymbol{\delta}) \right|\right|\right| \leq \lambda^{(3)}_{max},$$

where $|||\cdot|||$ is the spectral norm of a matrix or tensor. Note that (6) also implies the bounded largest eigenvalue of Fisher Information Matrix $\mathcal{I}$, because $\mathcal{I} = \nabla^2 \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*) = \nabla^2 \log \hat{N}(\boldsymbol{\theta}^*)$.

A key difference between this paper and previous proofs is that we make no explicit restrictions on the type of distribution P and Q, as KLIEP allows us to learn changes from various discrete/continuous distributions. Instead, we make the following assumptions on the density ratio:

Assumption 4 (The Correct Model Assumption)**.** *The density ratio model is correct, i.e. there exists* $\boldsymbol{\theta}^*$ *such that*

$$p(\boldsymbol{x}) = r(\boldsymbol{x}; \boldsymbol{\theta}^*) q(\boldsymbol{x}).$$

Assumptions 1, 2, and 3 are in fact related to distribution Q. However, the density ratio estimation objective is an M-estimator summed up over samples from P. Assumption 4 provides a transform between P and Q and allows us to perform analysis on such an M-estimator using an "importance sampling" fashion. See supplementary material (Liu, Suzuki, and Sugiyama 2014) for details.

Assumption 5 (Smooth Density Ratio Model Assumption)**.** *For any vector* $\boldsymbol{\delta} \in \mathbb{R}^{dim(\boldsymbol{\theta}^*)}$ *such that* $\|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|$ *and every* $t \in \mathbb{R}$*, the following inequality holds:*

$$\mathbb{E}_q \left[\exp\left(t \left(r(\boldsymbol{x}, \boldsymbol{\theta}^* + \boldsymbol{\delta}) - 1 \right) \right) \right] \leq \exp\left(\frac{10 t^2}{d} \right),$$

where d *is the number of changed edges.*

Next, we list a few consequences of Assumption 5.

Proposition 1. *For some small constants* ϵ *and any vector* $\boldsymbol{\delta} \in \mathbb{R}^{dim(\boldsymbol{\theta}^*)}$ *such that* $\|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|$*,*

$$P\left(r(\boldsymbol{x}, \boldsymbol{\theta}^* + \boldsymbol{\delta}) - 1 \geq \epsilon \right) \leq 2 \exp\left(-\frac{d\epsilon^2}{40} \right). \quad (7)$$

This proposition can be immediately proved by applying the Markov inequality and the Chernoff bounding technique.

Proposition 2. *For any vector* $\boldsymbol{\delta} \in \mathbb{R}^{dim(\boldsymbol{\theta}^*)}$ *such that* $\|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|$*,*

$$d \cdot Var_q \left[r(\boldsymbol{x}; \boldsymbol{\theta}^* + \boldsymbol{\delta}) \right] \leq 20.$$

Noting $\mathbb{E}_q \left[r(\boldsymbol{x}; \boldsymbol{\theta} + \boldsymbol{\delta}) - 1 \right] = 0$, this inequality is the consequence of sub-Gaussianity.

Using Assumption 5, we get Proposition 1 which provides a tail probability bound of the density ratio model on Q. Obtaining such an upper bound by the Hoeffding inequality (Hoeffding 1963) usually requires a bounded random variable. However, for continuous distributions, the ratio between two densities could be unbounded, and thus the boundedness cannot be assumed explicitly. Assumption 5 assumes the sub-Gaussianity of $r(\boldsymbol{x}; \boldsymbol{\theta}^* + \boldsymbol{\delta})$ on Q and guarantees an exponentially decaying upper-bound of approximation error of the normalization term. Please see supplementary material (Liu, Suzuki, and Sugiyama 2014) for details.

Proposition 2 demonstrates the limitation of density ratio estimation based algorithm: In order to guarantee the boundedness, the product between $\text{Var}_q \left[r(\boldsymbol{x}; \boldsymbol{\theta}^* + \boldsymbol{\delta}) \right]$ and d needs to be small. Since the density ratio model indicates the magnitude of change between two densities, such an assumption excludes the KLIEP algorithm from detecting significant change in parameters on many edges. We discuss a milder assumption later on.

We are now ready to state the main theorem.

Sufficient Conditions for Successful Change Detection

The following theorem establishes sufficient conditions of change detection in terms of parameter sparsity. Its proof is provided in supplementary material (Liu, Suzuki, and Sugiyama 2014). First, let's define $g(m) = \frac{\log(m^2+m)}{(\log \frac{m^2+m}{2})^2}$ which is smaller than 1 when m is reasonably large.

Theorem 1. *Suppose that Assumptions 1, 2, 3, 4, and 5 as well as* $\min_{t' \in S} \|\boldsymbol{\theta}^*_{t'}\| \geq \frac{10}{\lambda_{min}} \sqrt{d} \lambda_{n_p}$ *are satisfied, where* d *is the number of changed edges. Suppose also that the regularization parameter is chosen so that*

$$\frac{8(2-\alpha)}{\alpha} \sqrt{\frac{M_1 \log \frac{m^2+m}{2}}{n_p}} \leq \lambda_{n_p},$$

$$\frac{4(2-\alpha) M_1}{\alpha} \min\left(\frac{\|\boldsymbol{\theta}^*\|}{\sqrt{b}}, 1 \right) \geq \lambda_{n_p},$$

where $M_1 = \lambda_{max} b + 2$*, and* $n_q \geq \frac{M_2 n_p^2 g(m)}{d}$*, where* M_2 *is some positive constant. Then there exist some constants* L_1*,* K_1*, and* K_2 *such that if* $n_p \geq L_1 d^2 \log \frac{m^2+m}{2}$*, with the probability at least* $1 - \exp\left(-K_1 \lambda^2_{n_p} n_p \right) - 4 \exp\left(-K_2 d n_q \lambda^4_{n_p} \right)$*, the following properties hold:*

- *Unique Solution: The solution of* (5) *is unique*
- *Successful Change Detection:* $\hat{S} = S$ *and* $\hat{S}^c = S^c$*.*

Note that the probability of success converges to 1 as $\lambda^2_{n_p} n_p \to \infty$ and $d n_q \lambda^4_{n_p} \to \infty$. The proof roughly follows the steps of previous support consistency proofs using *primal-dual witness* method (Wainwright 2009). Here is a short sketch of the proof:

- Solve (5) with extra constrains on zero parameters.

$$\hat{\boldsymbol{\theta}}_S = \operatorname*{argmin}_{\boldsymbol{\theta}_S} \ell_{\text{KLIEP}} \left(\begin{bmatrix} \boldsymbol{\theta}_S \\ \mathbf{0} \end{bmatrix} \right) + \lambda_{n_p} \sum_{t' \in S} \|\boldsymbol{\theta}_{t'}\|;$$

[2] From now on, we simplify $\hat{N}(\boldsymbol{\eta}; \boldsymbol{x}^{(1)}_q, \ldots, \boldsymbol{x}^{(n_q)}_q)$ as $\hat{N}(\boldsymbol{\eta})$.

- For all $t' \in S$, $\hat{\boldsymbol{z}}_{t'} = \nabla \|\hat{\boldsymbol{\theta}}_{t'}\|$, and let $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\theta}}_S, \mathbf{0}]$;
- Obtain dual feasible sub-vectors $\hat{\boldsymbol{z}}_{t''}$ for all $t'' \in S^c$ by using the following equality:

$$\nabla \ell_{\text{KLIEP}}(\hat{\boldsymbol{\theta}}) + \lambda_{n_p} \hat{\boldsymbol{z}} = \mathbf{0}.$$

- Check the dual feasibility by showing $\max_{t'' \in S^c} \|\boldsymbol{z}_{t''}\| < 1$ with high probability under certain conditions.

There are some fundamental differences between this work and previous proofs. First, ℓ_{KLIEP} analyzed in this proof is a *likelihood ratio* between two densities which means that two sets of samples are involved in this proof and we have to consider the sparsity recovery conditions not only on one dataset, but with respect to two different MNs. Second, we did not explicitly limit the types of distribution for P and Q, and the parameter of each factor $\boldsymbol{\theta}_t$ is a vector rather than a scalar, which gives enough freedom of modelling highly complicated distributions. To the best of our knowledge, this is the first sparsity recovery analysis on learning changes from two MNs.

It is interesting to analyze the sample complexity of n_q, which is a novel element in this research. Intuitively, one should obtain sufficient number of samples from Q to accurately approximate the normalization term. Theorem 1 states n_q should grow at least quadratically with respect to n_p. Moreover, we show that as long as the density ratio model is smooth with respect to d (Assumption 5 and Proposition 2), such sample complexity can be relaxed by order $\mathcal{O}(d^{-1})$(see supplementary material (Liu, Suzuki, and Sugiyama 2014) for proof).

However, Assumption 5 together with Proposition 2 also shows that the variation allowed for the density ratio model decays as the number of changed edges d grows. This implies that, if d is large, we are only able to detect weak changes that do not cause huge fluctuations in the density ratio model, which is rather restrictive. Below, we consider another more relaxed scenario, where the assumption on the smoothness of the density ratio model is irrelevant to d.

Assumption 6. *For any vector $\boldsymbol{\delta} \in \mathbb{R}^{dim(\boldsymbol{\theta}^*)}$ such that $\|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|$ and every $t \in \mathbb{R}$, the following inequality holds:*

$$\mathbb{E}_q \left[\exp\left(t\left(r(\boldsymbol{x}, \boldsymbol{\theta}^* + \boldsymbol{\delta}) - 1\right)\right)\right] \leq \exp\left(10t^2\right).$$

Corollary 1. *Suppose that Assumptions 1, 2, 3, 4, and 6 are satisfied, $\min_{t \in S} \|\boldsymbol{\theta}_t^*\|$ satisfies the condition in Theorem 1, and the regularization parameter is chosen so that*

$$\frac{2-\alpha}{\alpha} \sqrt{\frac{M_1 \log \frac{m^2+m}{2}}{n_p^{\frac{3}{4}}}} \leq \lambda_{n_p},$$
$$\frac{4(2-\alpha)M_1}{\alpha} \min\left(\frac{\|\boldsymbol{\theta}^*\|}{\sqrt{b}}, \frac{1}{n_p^{1/8}}\right) \geq \lambda_{n_p},$$

where $M_1 = \lambda_{max} b + 2$, and $n_q \geq M_2 n_p g(m)$ where M_2 is some positive constant. Then there exist some constants L_1 such that if $n_p \geq L_1 d^{\frac{8}{3}} \left(\log \frac{m^2+m}{2}\right)^{\frac{4}{3}}$, KLIEP has the same properties as those stated in Theorem 1.

Corollary 1 states that it is possible to consider a relaxed version of Assumption 5 with the cost that the growth of n_p with respect to d has now increased from 2 to $\frac{8}{3}$, while the growth rate of n_q on n_p has decreased from 2 to 1. This is an encouraging result, since with mild changes on sample complexities, we are able to consider a weaker assumption that is irrelevant to d.

So far, we have only considered the scaling quadruple (n_p, n_q, d, m). However, it is also interesting to consider that the scalability of our theorem relative to b. This is a realistic scenario: It may be difficult to know the true underlying model of MN in practice, and thus we may adopt a model that contains many features to be "flexible enough" to describe the interactions among data. In the following corollary, we restate Theorem 1 with b and a new scalar s, which is the maximum number of non-zero elements in a pairwise feature vector.

Corollary 2. *Suppose that Assumptions 1, 2, 3, 4, and 5 are satisfied, $\min_{t \in S} \|\boldsymbol{\theta}_t^*\|$ satisfies the condition in Theorem 1, and the regularization parameter is chosen so that*

$$\frac{8(2-\alpha)}{\alpha} \sqrt{\frac{M_1 s \log \frac{m^2+m}{2}}{n_p}} \leq \lambda_{n_p},$$
$$\frac{4(2-\alpha)M_1}{\alpha} \min\left(\frac{\|\boldsymbol{\theta}^*\|}{\sqrt{b}}, 1\right) \geq \lambda_{n_p},$$

where $M_1 = \lambda_{max} b + 2$ and $n_q \geq \frac{M_2 s n_p^2 g'(m)}{d}$ where M_2 is some positive constant and $g'(m) = \frac{\log\left((m^2+m)\binom{b}{s}\right)}{(\log \frac{m^2+m}{2})^2}$. Then there exist some constants L_1 such that if $n_p \geq L_1 s d^2 \log \frac{m^2+m}{2}$, KLIEP has the same properties as those stated in Theorem 1.

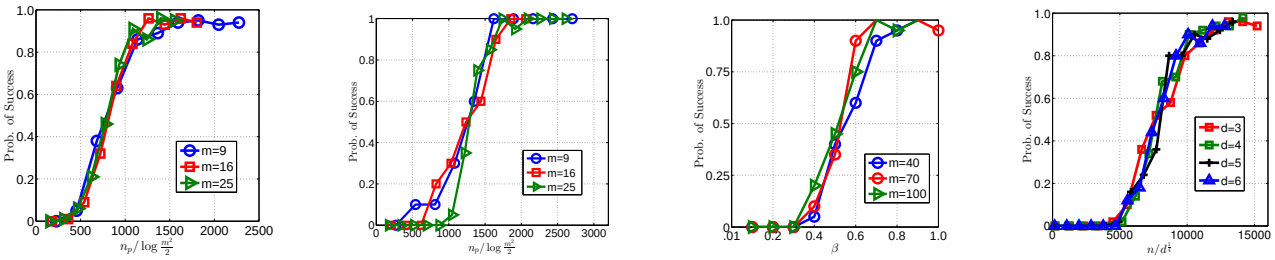
From Corollary 2, we can see that required n_p and n_q for change detection grows only linearly with s, and n_q grows mildly with $\binom{b}{s}$. Therefore, it is possible for one to consider a highly flexible model in practice.

Discussions

From the above theorem, one may gather some interesting insights into change detection based on density ratio estimation.

First, the required number of samples depends solely on d and m and is irrelevant to the number of edges of each MN. In contrast, separate graphical structural learning methods require more samples when each MN gets denser in terms of number of edges or neighborhood (Meinshausen and Bühlmann 2006; Ravikumar, Wainwright, and Lafferty 2010; Raskutti et al. 2009). This establishes the superiority of the density-ratio approach in sparse change detection between dense MNs. In other words, in order to detect sparse changes, the density-ratio approach does not require the individual MN to be sparse.

Second, the growth of n_q is also lower-bounded and grows quadratically with respect to n_p. This result illustrates the consequence of introducing a sample approximated normalization term. An insufficient number of samples from Q



(a) Lattice grid, Gaussian (b) Lattice grid, Diamond (c) Random Graph, Gaussian (d) Lattice grid, Gaussian

Figure 1: The rate of successful change detection versus the number of samples n_p normalized by $\log \frac{m^2+m}{2}$ (a-c) and $d^{\frac{1}{4}}$ (d).

would lead to poor approximation of the normalization term, and makes change detection more difficult. Fortunately, such growth rate can be further relaxed, and with slightly increased sample complexity of n_p.

Finally, our theorem also points out the limits of the density-ratio approach. Since the density-ratio approach is a conjunction of density ratio estimation and a (group) lasso program, it also inherits the drawbacks from both algorithms. Our analysis shows that the density ratio model may not deviate too much from 1 near the mean of distribution Q. A previous study on another density ratio estimator also has a similar observation (Yamada et al. 2013). Furthermore, the amount of variation allowed to diverge from 1 decreases at speed $\mathcal{O}(d^{-1})$. Since the density ratio indicates how much the change between P and Q is, this analysis generally says that the density-ratio approach is not good at detecting dramatic changes on a large number of edges.

Experiments

One important consequence of Theorem 1 is that, for fixed d, the number of samples n_p required for detecting the sparse changes grows with $\log \frac{m^2+m}{2}$. We now illustrate this effect via experiments.

The first set of experiments are performed on four-neighbor lattice-structured MNs. We draw samples from a Gaussian lattice-structured MN P. Then we remove 4 edges randomly, to construct another Gaussian MN Q. We consider the scaling of $m = 9, 16, 25$, $n_p \in [3000, 10000]$, and $n_p = n_q$. As suggested by Theorem 1, λ_{n_p} is set to a constant factor of $\sqrt{\frac{\log \frac{m^2+m}{2}}{n_p}}$. The rate of successful change detection versus the number of samples n_p normalized by $\log \frac{m^2+m}{2}$ is plotted in Figure 1(a). Each point corresponds to the probability of success over 20 runs. It can be seen that KLIEP with different input dimensions m tend to recover the correct sparse change patterns immediately beyond a certain critical threshold. All curves are well aligned around such a threshold, as Theorem 1 has predicted.

We next perform experiments on the non-Gaussian distribution with a diamond shape used in (Liu et al. 2014). The MNs are constructed in the same way as the previous experiment, while the samples are generated via slice sampling (Neal 2003). Figure 1(b) shows, for the lattice grids with dimensions $m = 9$, $m = 16$ and $m = 25$, the curves of success rates are well aligned.

We then validate our theorem on a larger scale Gaussian MNs with randomly generated structures. In this set of experiments, the structure of P is generated with 20% overall sparsity. The structure of Q is also set by removing 10 edges randomly. We consider $m = 40, 70, 100$, and $n_p = n_q$ scales as $1500\beta \log \frac{m^2+m}{2}$ where $\beta \in [0.1, 1]$. Again, curves of successful detection rate are aligned well on this graph, as Theorem 1 has predicted.

Finally, we evaluate the dependency between number of samples $n_p = n_q$ and number of changed edges d. Our theory predicts n_p required for successful change detection grows with d. We again construct a Gaussian lattice-structured MN P. Then we remove d edges randomly, to construct another Gaussian MN Q. We plot the success rate for $d = 3, 4, 5, 6$ versus $n_p/d^{\frac{1}{4}}$. Results are shown on Figure 1(d). As we can see, curves are well aligned, which suggests that n_p scales linearly with $d^{\frac{1}{4}}$. The sufficient condition from Theorem 1, $n_p = \Omega\left(d^2 \log \frac{m/2}{2}\right)$, seems to be overly conservative and might be tightened under certain regimes.

Conclusion

The KLIEP algorithm was experimentally demonstrated to be a promising method in sparse structure-change learning between two MNs (Liu et al. 2014). In this paper, we theoretically established sufficient conditions for its *successful change detection*. Our notable finding is that the number of samples needed for successful change detection is not dependent on the number of edges in each MN, but only on the number of *changed* edges between two MNs. We also provide numerical illustrations for our theories.

Ackownledgements

SL is supported by JSPS Fellowship and JSPS Kakenhi 00253189. MS is supported by JST CREST program. TS is partially supported by JST PRESTO, JST CREST, and MEXT Kakenhi 25730013.

References

Banerjee, O.; El Ghaoui, L.; and d'Aspremont, A. 2008. Model selection through sparse maximum likelihood esti-

mation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* 9:485–516.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441.

Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* 58(301):13–30.

Kolar, M., and Xing, E. P. 2012. Estimating networks with jumps. *Electronic Journal of Statistics* 6:2069–2106.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press.

Lee, S.-I.; Ganapathi, V.; and Koller, D. 2007. Efficient structure learning of Markov networks using l_1-regularization. In Schölkopf, B.; Platt, J.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems 19*, 817–824. Cambridge, MA, USA: MIT Press.

Liu, S.; Quinn, J. A.; Gutmann, M. U.; Suzuki, T.; and Sugiyama, M. 2014. Direct learning of sparse changes in Markov networks by density ratio estimation. *Neural Computation* 26(6):1169–1197.

Liu, S.; Suzuki, T.; and Sugiyama, M. 2014. Support consistency of direct sparse-change learning in Markov networks. *ArXiv e-prints*.

Meinshausen, N., and Bühlmann, P. 2006. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34(3):1436–1462.

Neal, R. M. 2003. Slice sampling. *The Annals of Statistics* 31(3):705–741.

Raskutti, G.; Yu, B.; Wainwright, M. J.; and Ravikumar, P. 2009. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc. 1329–1336.

Ravikumar, P.; Wainwright, M. J.; and Lafferty, J. D. 2010. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics* 38(3):1287–1319.

Sugiyama, M.; Nakajima, S.; Kashima, H.; von Bünau, P.; and Kawanabe, M. 2008. Direct importance estimation with model selection and its application to covariate shift adaptation. In Platt, J. C.; Koller, D.; Singer, Y.; and Roweis, S. T., eds., *Advances in Neural Information Processing Systems 20*. Curran Associates, Inc.

Sugiyama, M.; Suzuki, T.; and Kanamori, T. 2012. *Density Ratio Estimation in Machine Learning*. Cambridge, UK: Cambridge University Press.

Tibshirani, R.; Saunders, M.; Rosset, S.; Zhu, J.; and Knight, K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1):91–108.

Tsuboi, Y.; Kashima, H.; Hido, S.; Bickel, S.; and Sugiyama, M. 2009. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing* 17:138–155.

Vapnik, V. N. 1998. *Statistical Learning Theory*. New York, NY, USA: Wiley.

Wainwright, M. J. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Trans. Inf. Theor.* 55(5):2183–2202.

Yamada, M.; Suzuki, T.; Kanamori, T.; Hachiya, H.; and Sugiyama, M. 2013. Relative density-ratio estimation for robust distribution comparison. *Neural Computation* 25(5):1324–1370.

Yang, E.; Genevera, A.; Liu, Z.; and Ravikumar, P. 2012. Graphical models via generalized linear models. In Pereira, F.; Burges, C.; Bottou, L.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc. 1358–1366.

Yuan, M., and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Zhang, B., and Wang, Y. 2010. Learning structural changes of Gaussian graphical models in controlled experiments. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI2010)*, 701–708.

Zhao, P., and Yu, B. 2006. On model selection consistency of lasso. *The Journal of Machine Learning Research* 7:2541–2563.