Support Consistency of Direct Sparse-Change Learning in Markov Networks

Song Liu, Taiji Suzuki
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro, Tokyo, Japan
{song@sg.cs,s-taiji@is}.titech.ac.jp

Masashi Sugiyama
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
sugi@k.u-tokyo.ac.jp

Abstract

The interactions between covariates may change with learning domains. Discovering such transitions may offer key information helping us transfer our knowledge from one domain to another. 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. 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.

1 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. 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.

We consider the problem of learning changes between two undirected graphical models. Such a model, also known as a Markov network (MN) [2], expresses interactions via the conditional independence between random variables. Naively, one may utilize existing MN learning methods (e.g. Graphical Lasso [1]) to approximate two separated MNs and compare their differences.

One most recent effort based on *density ratio estimation*, proposes to learn the changes *directly* between MNs without modelling each individual MN [3]. In this paper, we theoretically investigate the success of such 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.

2 Direct Change Learning between Markov Networks

2.1 Problem Formulation

Consider two sets of independent samples drawn separately from two probability distributions P and Q on $\mathbb{R}^m$:$\{\boldsymbol{x}_p^{(i)}\}_{i=1}^{n_p} \overset{\text{i.i.d.}}{\sim} P$ 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

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

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, 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. $\boldsymbol{\psi}(x_u, x_v) : \mathbb{R}^2 \to \mathbb{R}^b$, 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., $\boldsymbol{\theta}^{(p)} - \boldsymbol{\theta}^{(q)}$.

2.2 Density Ratio Formulation for Structural Change Detection

The key idea in [3] 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. We *directly* model the ratio between p and q as

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

where $N(\boldsymbol{\theta})$ is the normalization term. The normalization term $N(\boldsymbol{\theta})$ is chosen to fulfill $\int q(\boldsymbol{x}) r(\boldsymbol{x};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{x} = 1$, and is defined as $N(\boldsymbol{\theta}) = \int q(\boldsymbol{x}) \exp\left(\sum_{u\geq v} \boldsymbol{\theta}_{u,v}^{\top} \boldsymbol{\psi}(x_u, x_v)\right) \mathrm{d}\boldsymbol{x}$, which is the expectation over $q(\boldsymbol{x})$. This expectation can be easily approximated by the sample average over $\{\boldsymbol{x}_q^{(i)}\}_{i=1}^{n_q} \overset{\text{i.i.d.}}{\sim} q(\boldsymbol{x})$.

2.3 Direct Density-Ratio Estimation

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

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

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

$$\ell_{\mathrm{KLIEP}}(\boldsymbol{\theta}) = -\frac{1}{n_p} \sum_{i=1}^{n_p} \log r(\boldsymbol{x}_p^{(i)};\boldsymbol{\theta})$$

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. 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 [8].

Now we have reached the final objective provided in [3]:

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

3 Support Consistency of Direct Sparse-Change Detection

3.1 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

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

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: $\mathcal{I} = \nabla^2 \ell_{\text{KLIEP}}(\boldsymbol{\theta}^*)$. $\mathcal{I}_{AB}$ is a sub-matrix of $\mathcal{I}$ indexed by two sets of indices A and B on rows and columns.

3.2 Assumptions

Similar to previous researches on sparsity recovery analysis [6, 5], 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: $\Lambda_{min}(\mathcal{I}_{SS}) \geq \lambda_{min} > 0$.*

This assumption is to ensure that the model is identifiable. Although Assumption 1 only bounds the smallest eigenvalue, 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:* $\max_{t'' \in S^c} \|\mathcal{I}_{t''S}\mathcal{I}_{SS}^{-1}\|_1 \leq 1 - \alpha, \alpha \in (0, 1]$, *where* $\|\boldsymbol{Y}\|_1 = \sum_{i,j} \|\boldsymbol{Y}_{i,j}\|_1$.

We also make the following assumptions as an analogy to those made in [7].

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

$$\begin{aligned} \max_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|} \left|\left|\left| \nabla^2 \log \hat{N}(\boldsymbol{\theta}^* + \boldsymbol{\delta}) \right|\right|\right| &\leq \lambda_{max}, \\ \max_{t \in S \cup S^c} \max_{\boldsymbol{\delta}, \|\boldsymbol{\delta}\| \leq \|\boldsymbol{\theta}^*\|} \left|\left|\left| \nabla_{\boldsymbol{\theta}_t} \nabla^2 \log \hat{N}(\boldsymbol{\theta}^* + \boldsymbol{\delta}) \right|\right|\right| &\leq \lambda^{(3)}_{max}, \end{aligned} \tag{6}$$

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:

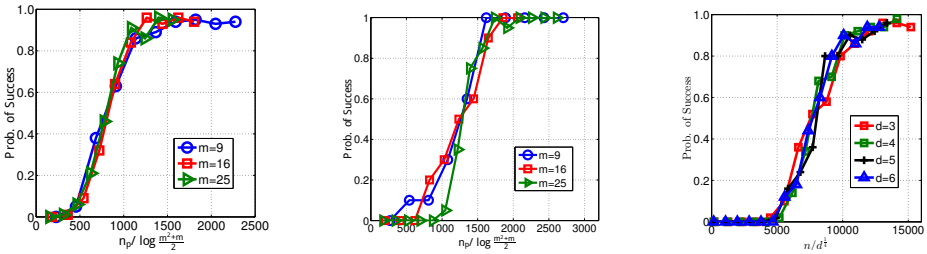
Assumption 4 (The Correct Model Assumption)**.** *The density ratio model is correct, i.e. there exists $\boldsymbol{\theta}^*$ such that $p(\boldsymbol{x}) = r(\boldsymbol{x}; \boldsymbol{\theta}^*)q(\boldsymbol{x})$.*

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:*

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

where d is the number of changed edges.

The following main theorem establishes sufficient conditions of change detection in terms of parameter sparsity. 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.



(a) Lattice grid, Gaussian (b) Lattice grid, Diamond (c) Lattice grid, Gaussian

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

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*

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

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* $1 - \exp\left(-K_1 \lambda_{n_p}^2 n_p\right) - 4\exp\left(-K_2 d n_q \lambda_{n_p}^4\right)$, *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 $\lambda_{n_p}^2 n_p \to \infty$ and $d n_q \lambda_{n_p}^4 \to \infty$. The proof follows the steps of previous support consistency proofs using *primal-dual witness* method [6] and is provided in the supplementary material [4].

4 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 $\log \frac{m^2+m}{2}$. 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 scale dimension m and n_p and let $n_p = n_q$. 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). 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 repeat the same experiment on non-Gaussian Diamond dataset [3] and results are shown in Figure 1(b).

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 Gaussian lattice-structured MNs. As we can see from Fig. 1(c), curves are well aligned, which suggests that n_p scales linearly with $d^{\frac{1}{4}}$.

Acknowledgements

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

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

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

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

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

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

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

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

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