Structure Learning of Undirected Graphical Models with Contrastive Divergence

Jie Liu JIELIU@CS.WISC.EDU
Department of Computer Sciences, UW-Madison

David Page PAGE@BIOSTAT.WISC.EDU
Department of Biostatistics and Medical Informatics
Department of Computer Sciences, UW-Madison

Abstract

Structure learning of Markov random fields (MRFs) is generally NP-hard (Karger & Srebro, 2001). Many structure learners and theoretical results are under the correlation decay assumption in the sense that for any two nodes i and k, the information about node i captured by node k is less than that captured by node j where j is the neighbor of i on the shortest path between i and k (Netrapalli et al., 2010). In this paper, we propose to learn structure of MRFs with contrastive divergence (Hinton, 2002) and demonstrate that our structure learner can recover the structures of these correlation *non-decay* MRFs.

1. Introduction

Markov random fields (MRFs) are useful probabilistic graphical models for capturing conditional independence among variables. However, learning MRF structure from data is generally NP-hard (Karger & Srebro, 2001). So far, the structure of MRFs can be learned by combinatorial search (Bresler et al., 2008; Bromberg et al., 2009; Netrapalli et al., 2010), or by convex relaxation (Ravikumar et al., 2010; Banerjee et al., 2008; Lee et al., 2006; Lin et al., 2009), or by feature extraction (Della Pietra et al., 1997; Lowd & Davis, 2010; Davis & Domingos, 2010; Van Haaren & Davis, 2012). Some theoretical properties such as time complexity and sample complexity are also known (Karger & Srebro, 2001; Bogdanov et al., 2008; Bresler et al., 2008; Abbeel et al., 2006; Ravikumar et al., 2010; Anandkumar et al., 2012). However, many of the structure learners and theoretical results are under the correlation decay assumption in the sense that for any two connected nodes i and k in the graph, the information about node i captured by node k is less than that captured by node j where j is the neighbor of i on the shortest path between i and k (Netrapalli et al., 2010). One typical counterexample given in the work of Netrapalli et al. (2010) is as follows.

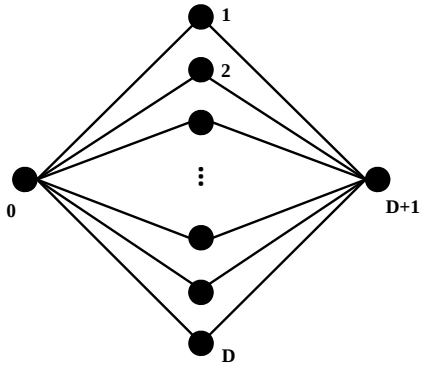

Figure 1. The structure of the correlation non-decay MRF in the work of Netrapalli et al. (2010).

Suppose that the ground truth structure of the MRF is given in Figure 1. There are in total $D+2$ nodes in the node set V, and $X_i \in \{-1, 1\}$ for $\forall i \in V$. The edge set $E = \{(0,i),(i,D+1)|1 \leq i \leq D\}$. The probability density function is $P(X;\theta) = \frac{1}{Z(\theta)} \prod_{(i,j)\in E} \exp\{\theta x_i x_j\}$, where $Z(\theta)$ is the normalizing constant. It can be shown that for a given θ, there is a D_{thres} such that if $D > D_{thres}$, the correlation between the node 0 and the node $D+1$ is the strongest among all potential pairs, and most structure learners tend to add the edge $(0, D+1)$ in the first place, even

Appearing in ICML 2013 Workshop on Structured Learning: Inferring Graphs from Structured and Unstructured Inputs.

if a sufficiently large number of training samples are provided. However, the edge $(0, D+1)$ does not exist in the ground truth structure.

In this paper, we propose a structure learning algorithm with contrastive divergence (Hinton, 2002) which recovers a distribution by iteratively comparing the current estimated distribution with training data. Our structure learner estimates that the edge $(0, D+1)$ exists in the beginning, as other structure learners do. However, when it gradually recovers the potential functions on edges $\{(0,i),(i,D+1)|1 \leq i \leq D\}$, it removes the edge $(0, D+1)$ eventually. Section 2 introduces contrastive divergence. Section 3 introduces our contrastive divergence structure learner. Section 4 demonstrates the performance of our structure learner and explains why our structure learner can deal with correlation non-decay situations. Finally, we conclude in Section 5.

2. Contrastive Divergence

Contrastive divergence (Hinton, 2002) is an effective parameter learner for MRFs, and we build our MRF structure learner on contrastive divergence by removing an edge during learning if its associated parameter is estimated to be close to zero. In order to present our structure learner in full detail in Section 3, we first review the details of contrastive divergence in this section.

Suppose for simplicity that we have a pairwise Markov random field on a random vector $\mathbf{X} \in \mathcal{X}^d$ described by an undirected graph $G(V, E)$ with node set V and edge set E. $\mathcal{X} = \{0, 1, ..., m-1\}$ is a discrete space, and in this paper we focus the binary situations for simplicity, namely $m = 2$. The probability of a sample $\mathbf{x}$ given a known parameter vector $\boldsymbol{\theta} = \{\theta_\alpha | \alpha \in \mathcal{I}\}$ ($\mathcal{I}$ is some index set) is

$$P(\mathbf{x}; \boldsymbol{\theta}) = \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\psi}(\mathbf{x}) - A(\boldsymbol{\theta})\right\}, \tag{1}$$

where $\boldsymbol{\psi} = \{\psi_\alpha | \alpha \in \mathcal{I}\}$ is a vector of sufficient statistics, and $A(\boldsymbol{\theta})$ is the log partition function as follows,

$$A(\boldsymbol{\theta}) = log \sum_{\mathbf{x} \in \mathcal{X}^d} \exp\left\{\boldsymbol{\theta}^T \boldsymbol{\psi}(\mathbf{x})\right\}. \tag{2}$$

Assume that we have s independent samples $\mathbb{X} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^s\}$ generated from (1), and we want to find the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ which maximizes the log-likelihood function

$$\mathcal{L}(\boldsymbol{\theta}|\mathbb{X}) \propto \frac{1}{s} \sum_{j=1}^{s} \boldsymbol{\theta}^T \boldsymbol{\psi}(\mathbf{x}^j) - A(\boldsymbol{\theta}). \tag{3}$$

It can be shown that $\mathcal{L}(\boldsymbol{\theta}|\mathbb{X})$ is concave. Therefore, we can use gradient ascent to find the global maximum of the likelihood function and find the MLE of $\boldsymbol{\theta}$. The partial derivative of $\mathcal{L}(\boldsymbol{\theta}|\mathbb{X})$ with respect to θ_α is

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}|\mathbb{X})}{\partial \theta_\alpha} = \frac{1}{s} \sum_{j=1}^{s} \psi_\alpha(\mathbf{x}^j) - E_{\boldsymbol{\theta}} \psi_\alpha = E_{\mathbb{X}} \psi_\alpha - E_{\boldsymbol{\theta}} \psi_\alpha. \tag{4}$$

Therefore, the key question is to calculate $E_{\boldsymbol{\theta}} \psi_\alpha$, the moment of statistic under a specific parameter vector $\boldsymbol{\theta}$. Exact computation of $E_{\boldsymbol{\theta}} \psi_\alpha$ takes time that is exponential in the treewidth of the graph. Contrastive divergence (CD) methods generate samples (particles) according to $\boldsymbol{\theta}$ using a Markov chain. Usually, the chain needs to reach equilibrium to generate an accurate sample, but CD's rationale is that only a rough estimate of the gradient is sufficient to determine the direction to update the parameters. Accordingly, two versions of CD methods have been proposed. One is CD-n which generates a sample by running Markov chain for n steps under parameter $\boldsymbol{\theta}^{(i)}$ (starting from a training sample) in iteration i. The other one is persistent contrastive divergence or PCD-n (Tieleman, 2008) which will advance the particles (from last iteration) for n step under the new parameters $\boldsymbol{\theta}^{(i)}$. Since n is usually chosen to be 1 in CD-n, the Markov chains for generating particles are usually far from equilibrium. Because $\boldsymbol{\theta}^{(i)}$ is close to $\boldsymbol{\theta}^{(i+1)}$ when the learning rate is small, persistent Markov chains are attractive.

3. Contrastive Divergence Structure Learning

The reason that we can use contrastive divergence to learn MRF structure is that edge (i, j) does not exist in the structure if and only if its corresponding parameter $\theta_{(i,j)} = 0$. Therefore, when we perform contrastive divergence to learn a MRF from data, we can initially include all the possible edges and gradually remove the edges when the corresponding parameters are estimated to be close to 0. The pseudocode is provide in Algorithm 1.

Initially, we build a candidate edge set E^* (via screening every pair of two nodes by some correlation test) which includes all the potential edges, and associate each potential edge with one parameter. Then we perform standard contrastive divergence (Hinton, 2002)

Algorithm 1 Contrastive Divergence Structure Learning

Input: independent samples $\mathbb{X} = \{\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^s\}$, maximum iteration number T, MCMC step number n, threshold $\theta_{threshold}$
Output: estimated edge set $\hat{E}$
Procedure:
Create E^* which contains all potential edges and the corresponding parameter set $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_{|E^*|}\}$
Initialize $\boldsymbol{\theta}^{(1)}$ and initialize particles
Calculate $E_{\mathbb{X}}\boldsymbol{\psi}$ from $\mathbb{X}$
for $i = 1$ **to** T **do**
  Advance particles for n steps under $\boldsymbol{\theta}^{(i)}$
  Calculate $E_{\boldsymbol{\theta}^{(i)}}\boldsymbol{\psi}$ from the particles
  $\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \eta(E_{\mathbb{X}}\boldsymbol{\psi} - E_{\boldsymbol{\theta}^{(i)}}\boldsymbol{\psi})$
  for $j = 1$ **to** $|E^*|$ **do**
    If the estimated parameter for edge j is less than $\theta_{threshold}$, then fix the parameter estimate at 0 from the current iteration on
  end for
  Adjust η
end for
Set $\hat{E}$ to be empty
for $j = 1$ **to** $|E^*|$ **do**
  Add edge j to $\hat{E}$ if the corresponding estimated parameter is nonzero.
end for

or persistent contrastive divergence (Tieleman, 2008) to estimate the parameters. During learning, we keep monitoring the estimate of the parameters and fix some of the estimated parameters at 0 if they are close to 0 such as in the interval $[-\theta_{threshold}, \theta_{threshold}]$ where $\theta_{threshold}$ is a small positive real number. Eventually, the recovered edges in the MRFs are these edges whose corresponding parameters are estimated to be nonzero.

4. Experiments and Results

In this section, we demonstrate the performance of our contrastive divergence structure learner and explain why it can deal with correlation non-decay situations. We use the example in the work of Netrapalli et al. (2010), as mentioned in Section 1. We set $D = 8$ and we have an MRF of 10 variables with its ground truth structure in Figure 2. All the variables take either 1 or -1. The pairwise potential function on each edge parameterized by β $(0 < \beta < 1)$ is $\begin{pmatrix} \beta & 1-\beta \\ 1-\beta & \beta \end{pmatrix}$. Note that after we rewrite the probability density function into the canonical exponential

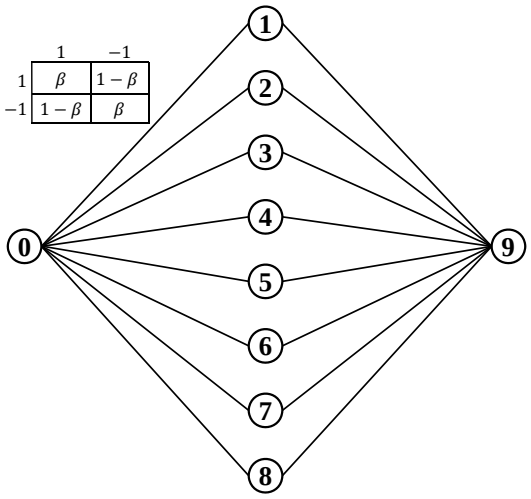

Figure 2. The structure of the MRF in the experiment.

family form, $\theta = \log \frac{\beta}{1-\beta}$ and $\psi(X_i, X_j) = I(X_i = X_j)$ where $I(X_i = X_j)$ is the indicator variable that indicates whether X_i and X_j take the same value. Note that $\theta = 0$ if and only if $\beta = 0.5$. We generate s training samples and monitor how the contrastive divergence structure learner recovers the MRF structure from the data. In our simulations, we set $\beta = 0.6$ and $s = 5{,}000$. We start our structure learner with a candidate edge set E^* which includes all the 45 possible edges. There are essentially three types of edges. The first type of edges are $\{(0, i), (i, 9) | 1 \leq i \leq 8\}$, namely the 16 edges in the ground truth. The second type only consists of the edge $(0, 9)$ which has the strongest correlation in the data, but the edge is not in the ground truth structure. The last type includes 28 edges $\{(i, j) | 1 \leq i < j \leq 8\}$, which are not included in the ground truth. A consistent structure learner should be able to recover only the first type of edges when given enough data.

We monitor the estimated parameters that are associated with the three types of edges. Specifically, we monitor the estimate of $\beta_{(0,1)}$, $\beta_{(0,9)}$ and $\beta_{(1,2)}$ during contrastive divergence learning, and the results are presented in Figure 3. The parameter estimate for edge $(1, 2)$ quickly became 0.5 (equivalently $\hat{\theta}_{(1,2)} = 0$) at the second iteration and the edge $(1, 2)$ was removed. In the first 30 iterations, the parameter for edge $(0, 9)$ was estimated to be higher than that for edge $(0, 1)$ and the reason is that node 0 shows a higher level of sample correlation with node 9 than with node 1. However, as the parameter estimate for edge $(0, 1)$ continued to increase, the parameter estimate for edge $(0, 9)$ continued to decrease. At the 388-

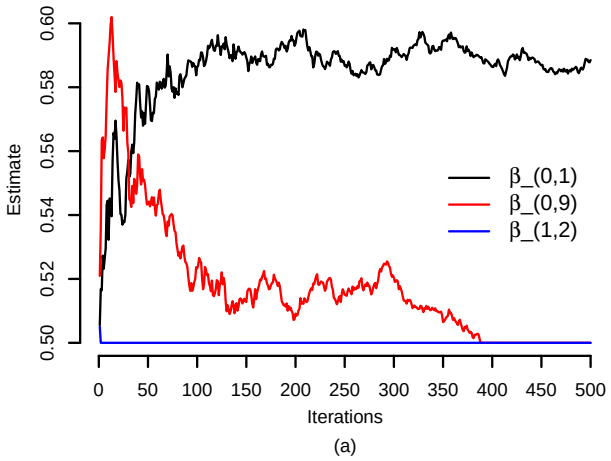

Figure 3. The parameter estimation from contrastive divergence.

th iteration, the parameter estimate for edge $(0, 9)$ was below the threshold 0.501, and the edge $(0, 9)$ was removed. Therefore in the end, we successfully recovered all the edges correctly.

5. Conclusion

In this paper, we propose a new structure learning algorithm for MRFs. The new structure learner is based on contrastive divergence which makes the structure learner different from previous structure learners via combinatorial search (Bresler et al., 2008; Bromberg et al., 2009; Netrapalli et al., 2010), or via convex relaxation (Ravikumar et al., 2010; Banerjee et al., 2008; Lee et al., 2006; Lin et al., 2009), or via feature extraction (Della Pietra et al., 1997; Lowd & Davis, 2010; Davis & Domingos, 2010; Van Haaren & Davis, 2012). The contrastive divergence structure learner can handle correlation non-decay situations which cannot be handled by previous structure learners. Contrastive divergence, which uses the concavity of MRF's log likelihood function and uses MCMC based gradient ascent to bypass the intractable normalizing constant, is currently one of the most effective parameter estimation methods for undirected graphical models. We hope this technique can be further employed for MRF structure learning.

Acknowledgements

The authors gratefully acknowledge the support of NIGMS grant R01GM097618-01 and NLM grant R01LM011028-01.

References

Abbeel, Pieter, Koller, Daphne, and Ng, Andrew Y. Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.*, 7:1743–1788, December 2006. ISSN 1532-4435.

Anandkumar, A., Tan, V. Y. F., and Willsky, A.S. High-Dimensional Structure Learning of Ising Models : Local Separation Criterion. *Annals of Statistics*, 40:1346–1375, 2012.

Banerjee, Onureena, El Ghaoui, Laurent, and d'Aspremont, Alexandre. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res.*, 9:485–516, June 2008. ISSN 1532-4435.

Bogdanov, Andrej, Mossel, Elchanan, and Vadhan, Salil. The complexity of distinguishing markov random fields. In *Proceedings of the 11th international workshop, APPROX 2008, and 12th international workshop, RANDOM 2008 on Approximation, Randomization and Combinatorial Optimization: Algorithms and Techniques*, APPROX '08 / RANDOM '08, pp. 331–342, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85362-6. doi: 10.1007/978-3-540-85363-3_27.

Bresler, G., Mossel, E., and Sly, A. Reconstruction of markov random fields from samples: Some observations and algorithms. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pp. 343–356, 2008.

Bromberg, Facundo, Margaritis, Dimitris, and Honavar, Vasant. Efficient markov network structure discovery using independence tests. *J. Artif. Int. Res.*, 35(1):449–484, July 2009. ISSN 1076-9757.

Davis, J. and Domingos, P. Bottom-up learning of markov network structure. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 271–280, 2010.

Della Pietra, Stephen, Della Pietra, Vincent, and Lafferty, John. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(4): 380–393, April 1997. ISSN 0162-8828. doi: 10.1109/34.588021.

Hinton, Geoffrey. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

Karger, David and Srebro, Nathan. Learning markov networks: maximum bounded tree-width graphs. In

Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms, SODA '01, pp. 392–401, Philadelphia, PA, USA, 2001. Society for Industrial and Applied Mathematics. ISBN 0-89871-490-7.

Lee, S.I., Ganapathi, V., and Koller, D. Efficient structure learning of markov networks using l_1-regularization. In *NIPS*, 2006.

Lin, Yuanqing, Zhu, Shenghuo, Lee, Daniel D., and Taskar, Ben. Learning sparse markov network structure via Ensemble-of-Trees models. In *International Conference on Artificial Intelligence and Statistics*, 2009.

Lowd, D. and Davis, J. Learning markov network structure with decision trees. In *2010 IEEE International Conference on Data Mining*, pp. 334–343. IEEE, 2010.

Netrapalli, P., Banerjee, S., Sanghavi, S., and Shakkottai, S. Greedy learning of markov network structure. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 1295–1302. IEEE, 2010.

Ravikumar, P., Wainwright, M.J., and Lafferty, J.D. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Tieleman, Tijmen. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML*, 2008.

Van Haaren, J. and Davis, J. Markov network structure learning: A randomized feature generation approach. In *AAAI*, 2012.