Technical Report of the 3rd Workshop on Latent Dynamics (Sep 24, 2012, Tokyo, Japan)

スパースかつ低ランク制約に基づく時変ネットワーク構造推定

平山 淳一郎* アーポ ヒバリネン† 石井 信‡
Jun-ichiro Hirayama Aapo Hyvärinen Shin Ishii

Abstract: Several authors have recently proposed sparse estimation techniques for *time-varying* Markov networks, in which both graph structures and model parameters may change with time. In this study, we extend a previous approach with a low-rank assumption on the matrix of parameter sequence, using a recent technique of nuclear norm regularization. This can potentially improve the estimation performance by reducing the effective degree of freedom of the estimation which tends to be very high in large-scale time-varying networks. We derive a simple algorithm based on the alternating direction method of multipliers (ADMM) which can effectively utilize the separable structure of our convex minimization problem. A breif summary of a simulation result is presented, which shows the nuclear norm regularization is potentially effective for improving the performance of recovering time-varying network structures.

1 Introduction

Markov networks (MNs), or Markov random fields (MRFs), are basic statistical models for representing dependency networks of multiple random variables, and have many applications in various fields related to machine learning. An MN describes a structure of conditional (in)dependences by an undirected graph, and defines a probability distribution with parametric potentials associated with their nodes and edges. Two fundamental examples of MNs are the Gaussian Graphical Model and the Ising model, the latter of which we focus on in this study.

Recently, several authors have proposed sparse estimation techniques, typically using the ℓ_1-norm regularization to prune irrelevant edges, for *time-varying* MNs [7, 6, 8] which allows the graph structure and model parameters to change with time. They showed that time-varying MNs may be estimated by incorporating certain mechanisms of temporal smoothing into the sparse estimation framework.

Here, extending an approach in [6] for the Ising model, we propose a new and effective approach to estimating time-varying Markov network based on an additional assumption that the parameter matrix, whose column is the vector of all the model parameters at a single time step, have a relatively low rank. This assumption is expected to be effective for reducing the degree of freedom of the parameter matrix, which tends to be very high in large-scale time-varying networks.

2 Proposed method

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_D)^\top \in \{-1, 1\}^D$ be a binary observed vector. Then, the Ising model is given by

$$p(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{i<j} \theta_{ij} y_i y_j + \sum_i \theta_{ii} y_i \right), \quad (1)$$

where $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}} \exp\left(\sum_{i<j} \theta_{ij} y_i y_j + \sum_i \theta_{ii} y_i \right)$ is the partition function. The first summation in the exponent is over all pairs (i, j) that satisfy $i < j$, and we put all the $C = D(D+1)/2$ parameters in a vector $\boldsymbol{\theta} \in \mathbb{R}^C$. The corresponding undirected graph to this model has nodes $i = 1, 2, \ldots, D$, and any pair of nodes (i, j) is connected if and only if θ_{ij} is non-zero.

*国際電気通信基礎技術研究所 (ATR), 619-0288 京都府相楽郡精華町光台二丁目２番地２, e-mail hirayama@atr.jp
Advanced Telecommunications Research (ATR) Institute International, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
†Department of Mathematics and Statistics/ Department of Computer Science and HIIT, University of Helsinki, Helsinki, Finland, e-mail aapo.hyvarinen@helsinki.fi
‡京都大学 大学院情報学研究科, 611-0011 京都府宇治市五ヶ庄, e-mail ishii@i.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University, Uji, Kyoto 611-0011, Japan

Now suppose the parameter vector $\boldsymbol{\theta}$ is time-dependent, indexed by a superscript $n = 1, 2, \ldots, N$, and define a parameter matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N) \in \mathbb{R}^{C\times N}$. In order to effectively estimate $\boldsymbol{\Theta}$ from a given observed time series $\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^N$, we introduce a convex minimization problem:

$$\underset{\boldsymbol{\Theta}}{\text{minimize}} \quad f(\boldsymbol{\Theta}) + \|\boldsymbol{\Lambda} \circ \boldsymbol{\Theta}\|_1 + \eta\|\boldsymbol{\Theta}\|_*, \qquad (2)$$

where the first term is a kernel-smoothed loss function as used in the previous studies on time-varying MNs [8, 6], given by

$$f(\boldsymbol{\Theta}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{N} \varphi\left(|m-n|\right) l(\boldsymbol{y}^m, \boldsymbol{\theta}^m). \qquad (3)$$

Here, we use the negative logarithm of the *pseudolikelihood* [2], i.e. $l(\boldsymbol{y}, \boldsymbol{\theta}) := -\sum_{i=1}^{D} \log p(y_i^m \mid \boldsymbol{y}_{\setminus i}^m; \boldsymbol{\theta}^n)$, for the loss measure; $\|\boldsymbol{\Theta}\|_1$ denotes the ℓ_1-norm for a long vector that concatenates all the columns in $\boldsymbol{\Theta}$, where $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}, \boldsymbol{\lambda}, \ldots, \boldsymbol{\lambda})$ contains the vector of regularization coefficients, $\boldsymbol{\lambda} \in [0, \infty)^C$, which is assumed to be common for all the time steps; $\|\boldsymbol{\Theta}\|_*$ denotes the nuclear norm (or trace norm) [3], which is defined as the summation of all the singular values of $\boldsymbol{\Theta}$, where $\eta \geq 0$ is the regularization coefficient. Since all the three terms are convex, the problem (2) itself is also convex.

We employ the alternating direction method of multipliers (ADMM) [1] to solve the convex minimization problem introduced here. See [4, 5] for the detail of our algorithm.

3 Simulation result

Here, we briefly summarize an experimental result with a toy problem (see [4] for more details); other results will be found in [5].

In this experiment, the dataset was sampled from the Ising model (1) with a time-varying parameters. The dimensionality of observations was $D = 7$ and the length of time-series $N = 200$. The parameter space was then $\mathbb{R}^{28}$, but every $\boldsymbol{\theta}^n$ was constrained to be in a three-dimensional subspace, according to

$$\boldsymbol{\theta}^n = s_1^n \boldsymbol{a}^1 + s_2^n \boldsymbol{a}^2 + s_3^n \boldsymbol{a}^3, \qquad (4)$$

and thus rank$(\boldsymbol{\Theta}) = 3$. Here, we chose the three basis elements, $\boldsymbol{a}^1$, $\boldsymbol{a}^2$ and $\boldsymbol{a}^3$, so that they correspond respectively to the three graphs in the left column of Fig. 1, and their non-zero elements (*i.e.*, edge weights) were uniformly set at 0.5. The right column of Fig. 1 also shows time-series of their coefficients, s_1^n, s_2^n and s_3^n, which only took 0 or 1 for simplicity.

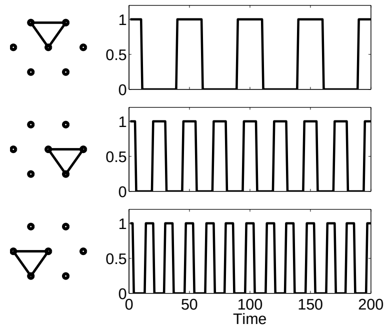

図 1: Three graphs corresponding to the basis elements (left) and the time-series of their coefficients (right) used for generating a sparse parameter time-series that are embedded in a three-dimensional subspace.

We used a rectangular window function for temporal smoothing:

$$\varphi(|m-n|) = \begin{cases} 1/w & |m-n| \leq \tau \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

where $w = 2\tau + 1$. We examined several values for the time-window width at $w = 5, 9, 13$ and 17 ($\tau = 2, 4, 6$ and 8). The regularization coefficient for the ℓ_1-norm was set by $\lambda_{ii} = 0$ and $\lambda_{ij} = \lambda$ $(i \neq j)$ with various values of λ.

We evaluated the performance of structure recovery with the Area Under the ROC Curve (AUC) by regarding it as a binary classification problem. In other words, from the final estimate of $\boldsymbol{\Theta}$ obtained as above, we have a binary classifier which says whether a single weight θ_{ij}^n belongs to the class of non-zero weights or to that of weights equal to zero for each $i \neq j$ and n. The performance of this detection can be quantified by an ROC curve, and the area under the curve is quantified by the trapezoidal rule.

Figure 2 plots the AUC versus $\log_{10}\eta$. This shows that the performance of structure recovery was improved in all the window widths by introducing low-rank regularization within an appropriate range of η.

4 Summary

We have proposed a new "sparse and low-rank" estimation framework of time-varying MNs, particularly using an Ising model as a concrete example of MNs. An

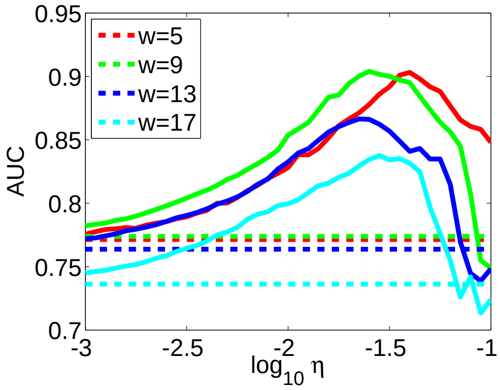

図 2: Area Under the ROC Curve versus strength of low-rank regularization ($\log \eta$). The horizontal dashed lines indicate the AUC values when $\eta = 0$ for each w.

experiment with artificially-generated dataset showed that the low-rank regularization can potentially improve the estimation performance over those only using sparsity and local smoothness. A full-length report including a real-data experiment will be found in [5].

参考文献

[1] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and distributed computation: Numerical methods.* Prentice-Hall, Inc., 1989.

[2] J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

[3] M. Fazel, H. Hindi, and S. Boyd. Rank minimization and applications in system theory. In *Proceedings American Control Conference*, pages 3273–3278, 2004.

[4] J. Hirayama, A. Hyvärinen, and S. Ishii. Sparse and low-rank estimation of time-varying markov networks with alternating direction method of multipliers. In *International Conference on Neural Information Processing (ICONIP'10), Lecture Notes in Computer Science*, volume 6443, pages 371–379, 2010.

[5] J. Hirayama, A. Hyvärinen, and S. Ishii. Sparse and low-rank matrix regularization for learning time-varying markov networks, in revision.

[6] M. Kolar, L. Song, A. Ahmed, and E. P. Xing. Estimating time-varying networks. *Annals of Applied Statistics*, 4(1):94–123, 2010.

[7] X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th International Conference on Machine Learning (ICML'07)*, pages 1055–1062, 2007.

[8] S. Zhou, J. Lafferty, and L. Wasserman. Time varying undirected graphs. *Machine Learning*, 80(2–3):295–319, 2010.