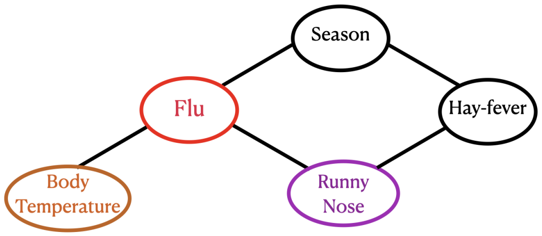
On Learning Continuous Pairwise Markov Random Fields

Abhin Shah, Devavrat Shah, Gregory Wornell

Markov Random Fields

- Diagrammatic representations of probability distributions with a Markovian structure.



Local Markov Property

- Given value of neighbors, a node is independent of the remaining nodes e.g., Body Temperature $\perp\!\!\!\perp$ Runny Nose | Flu.

Pairwise Markov Random Fields

- Consider an undirected graph $G = ([p], E)$ where the nodes correspond to a $p-$dimensional random vector $\mathbf{x}$, and E denotes the edge set.
- Any strictly positive distribution in the family of pairwise MRF represented by G factorizes as :

$$f_{\mathbf{x}}(\mathbf{x}) \propto \exp\Big(\sum_{i\in[p]} g_i(x_i) + \sum_{(i,j)\in E} g_{ij}(x_i, x_j)\Big).$$

Examples	$g_i(x_i)$	$g_{ij}(x_i, x_j)$
Ising	$\theta^{(i)} x_i$	$\theta^{(ij)} x_i x_j$
Discrete	$\boldsymbol{\theta}^{(i)}(x_i)$	$\boldsymbol{\theta}^{(i)}(x_i)$
Gaussian	$\theta_1^{(i)} x_i + \theta_2^{(i)} x_i^2$	$\theta^{(ij)} x_i x_j$

- Limited progress for continuous (non-Gaussian) MRFs.

Learning Markov Random Fields

- *Structure Recovery* : Given independent samples of $\mathbf{x}$, estimate the underlying graph structure.
- *Parameter Recovery* : Given independent samples of $\mathbf{x}$, estimate all the associated parameters.

Generalized Interaction Screening Objective

Vuffray, Misra, Lokhov - NeurIPS 2020

- GISO is a node specific convex objective function.
- GISO can recover the graph structure and the 'edge' parameters (but not the 'node' parameters) in discrete graphical models with $\Omega(\log p)$ samples.
- GISO is ingenious but unusual as it does not contain any partition function.
- If $f_{x_i}(x_i|\mathrm{x}_{-i} = x_{-i}; \boldsymbol{\theta}) \propto \exp\big(g(\boldsymbol{\theta}, \mathbf{x})\big)$, then population GISO for node i is $\mathbb{E}\Big[\exp\big(-g(\boldsymbol{\theta}, \mathbf{x})\big)\Big]$.

Problem Formulation

Continuous random variables

- Bounded Domain: Random variables are bounded.
- Parametric potentials : $g_i(\cdot) = {\boldsymbol{\theta}^{(i)}}^T \boldsymbol{\phi}(\cdot)$, and $g_{ij}(\cdot,\cdot) = {\boldsymbol{\theta}^{(ij)}}^T \boldsymbol{\psi}(\cdot,\cdot)$.

Examples : 1. Polynomial basis 2. Harmonic basis

- Bounded parameters : Parameters are bounded.
- Sparsity : Maximum degree of any node of the underlying graph is at-most d.

Algorithm

Step 1 - Learn graph structure and edge parameters

- For any $i \in [p]$ the conditional density of x_i is :

$$f_{x_i}(x_i|\mathrm{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{(i)}) \propto \exp\Big({\boldsymbol{\vartheta}^{(i)}}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big).$$

where $\boldsymbol{\vartheta}^{(i)}$ consists of node parameters and edge parameters involving node i and $\boldsymbol{\varphi}^{(i)}(\cdot)$ is a function of the node and edge basis.

$$\text{Population GISO} = \mathbb{E}\Big[\exp\Big(-\boldsymbol{\vartheta}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)\Big].$$

- GISO be adapted to recover the graph structure and the 'edge' parameters in continuous graphical models!

Algorithm

Step 2 - Learn node parameters

- For any $i \in [p]$ the conditional density of x_i is: $f_{x_i}(x_i|\mathrm{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{(i)}) \propto \exp\Big(\boldsymbol{\lambda}^T(x_{-i})\boldsymbol{\phi}(x_i)\Big)$ where $\boldsymbol{\lambda}(x_{-i})$, the canonical parameter, is linear function of node and edge parameters.
- By duality of exponential family, if we know $\boldsymbol{\mu}(x_{-i}) := \mathbb{E}[\boldsymbol{\phi}(x_i)|X_{-i} = x_{-i}]$, we can learn $\boldsymbol{\lambda}(x_{-i})$.
- Learning $\boldsymbol{\mu}(x_{-i})$ is a regression problem.
- The regression function $\boldsymbol{\mu}(\cdot)$ is Lipschitz $\longrightarrow$ approximately *linearize* $\boldsymbol{\mu}(\cdot)$ $\longrightarrow$ a sparse linear regression problem.

Main Results

For discrete or continuous random variables

1 - Population GISO is equivalent to "local" MLE

- $\arg\min$ Population GISO $= \arg\min D(\cdot \,\|\, \cdot)$ where $D(\cdot \,\|\, \cdot)$ is a node-specific KL divergence.

2 - GISO is asymptotically consistent and normal

- The traditional MLE is intractable.
- 'Local' M-estimation is tractable (but not asymptotically efficient).

For continuous random variables

3 - Structure recovery with $\Omega\big(\exp(d)\log p\big)$ samples

4 - Parameter recovery with $\Omega\big(\exp(d)\log p\big)$ samples

- All of the existing methods require some stringent conditions, for example - *incoherence, dependency, sparse eigenvalue or restricted strong convexity.*
- Our work does not require any of these conditions.