On Learning Continuous Pairwise Markov Random Fields

Abhin Shah, Devavrat Shah, Gregory W. Wornell

Massachusetts Institute of Technology

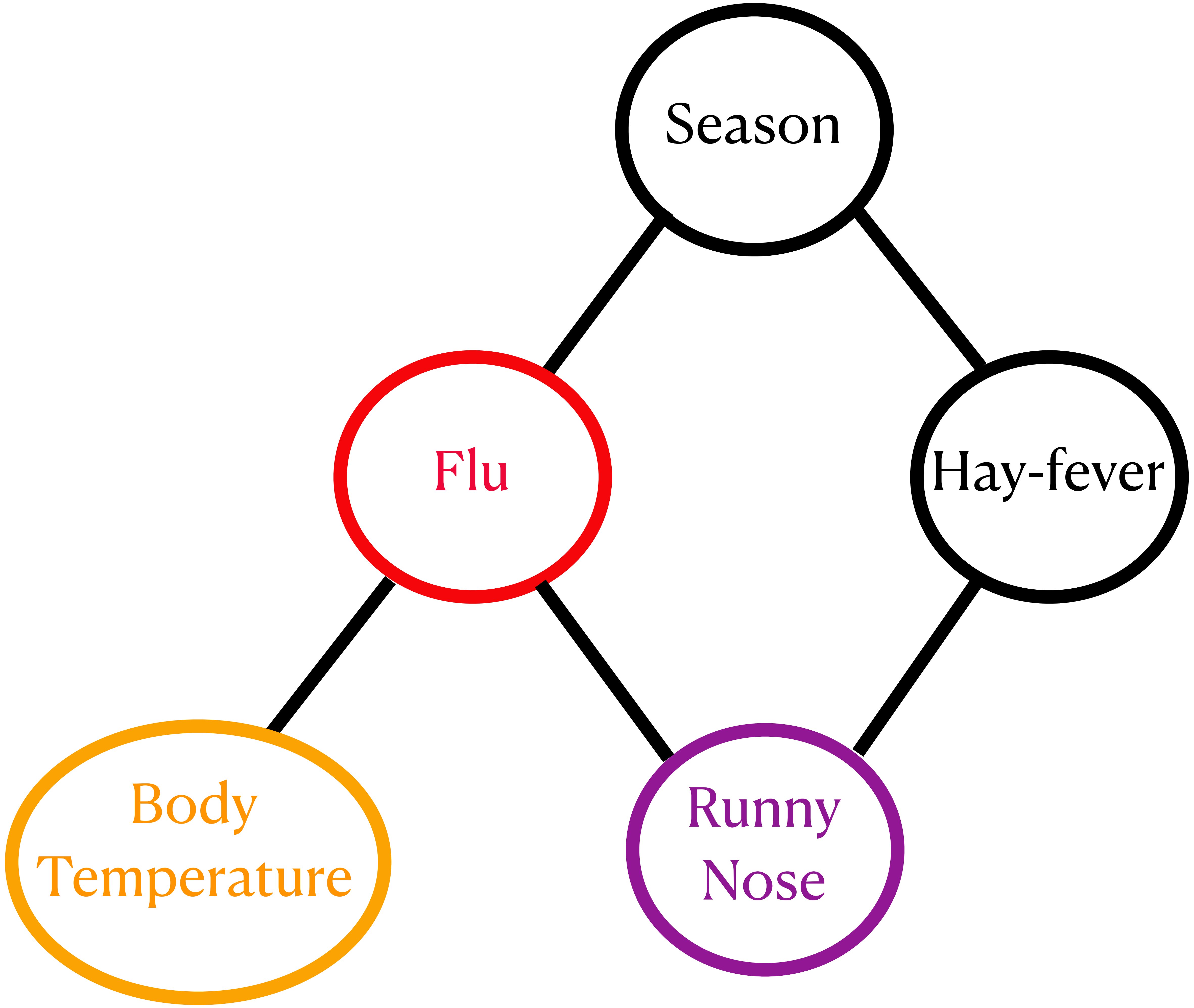
Markov Random Fields

Undirected Graphical Models

- Diagrammatic representations of probability distributions with a Markovian structure

Local Markov Property

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



Pairwise Markov Random Fields

- Consider an undirected graph $G = ([p], E)$.

- Any strictly positive distribution in the family of pairwise MRF represented by G factorizes as

$$f_{\mathbf{x}}(\mathbf{x}) \propto \exp\left(\sum_{i\in[p]} \underset{\substack{\Big\downarrow \\ \text{Node Potentials}}}{g_i(x_i)} + \sum_{(i,j)\in E} \underset{\substack{\Big\downarrow \\ \text{Edge Potentials}}}{g_{ij}(x_i, x_j)}\right)$$

Pairwise Markov Random Fields

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

Examples

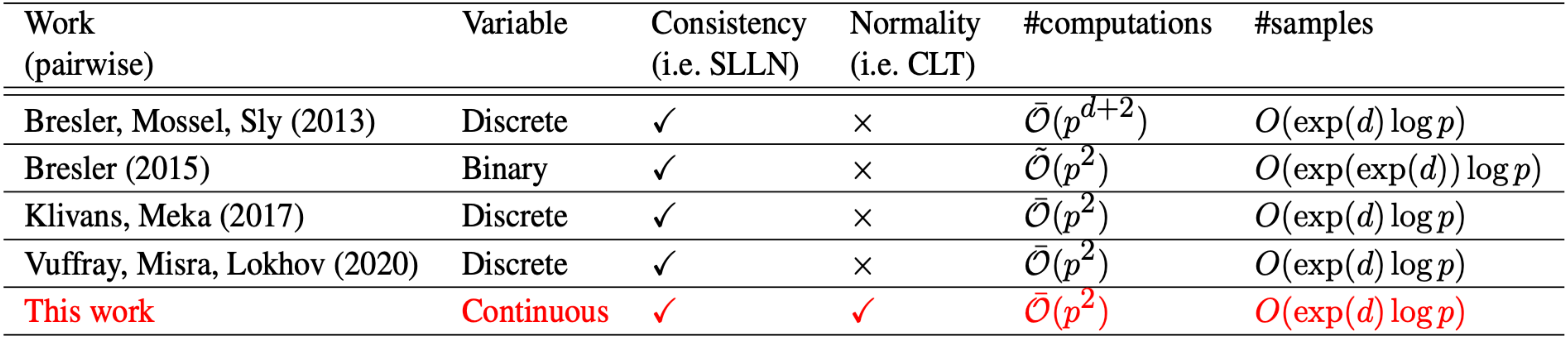
	$g_i(x_i)$	$g_{ij}(x_i, x_j)$
Ising Model	$\theta^{(i)} x_i$	$\theta^{(ij)} x_i x_j$
Discrete Model	$\boldsymbol{\theta}^{(i)}(x_i)$	$\boldsymbol{\theta}^{(ij)}(x_i, x_j)$
Gaussian Model	$\theta_1^{(i)} x_i + \theta_2^{(i)} x_i^2$	$\theta^{(ij)} x_i x_j$

Learning Markov Random Fields

- *Structure recovery* - Given independent samples of $\mathbf{x}$, estimate the underlying graph structure (i.e., the edge set E).
- *Parameter recovery* - Given independent samples of $\mathbf{x}$, estimate all the parameters associated with the joint density.

Comparison with prior works

Binary and Discrete



Work (pairwise)	Variable	Consistency (i.e. SLLN)	Normality (i.e. CLT)	#computations	#samples
Bresler, Mossel, Sly (2013)	Discrete	✓	×	$\bar{\mathcal{O}}(p^{d+2})$	$O(\exp(d)\log p)$
Bresler (2015)	Binary	✓	×	$\tilde{\mathcal{O}}(p^2)$	$O(\exp(\exp(d))\log p)$
Klivans, Meka (2017)	Discrete	✓	×	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$
Vuffray, Misra, Lokhov (2020)	Discrete	✓	×	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$
This work	Continuous	✓	✓	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$

Comparison with prior works

Binary and Discrete

Work (pairwise)	Variable	Consistency (i.e. SLLN)	Normality (i.e. CLT)	#computations	#samples
Bresler, Mossel, Sly (2013)	Discrete	✓	×	$\bar{\mathcal{O}}(p^{d+2})$	$O(\exp(d)\log p)$
Bresler (2015)	Binary	✓	×	$\tilde{\mathcal{O}}(p^2)$	$O(\exp(\exp(d))\log p)$
Klivans, Meka (2017)	Discrete	✓	×	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$
Vuffray, Misra, Lokhov (2020)	Discrete	✓	×	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$
This work	Continuous	✓	✓	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$

Generalized Interaction Screening Objective (GISO)

Vuffray, Misra, Lokhov — NeurIPS 2020

- GISO is a node specific convex objective function.
- The GISO can recover the graph structure and the 'edge' parameters in discrete graphical models.
- Suppose $f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\theta}) \propto \exp\Big(g(\boldsymbol{\theta}, \mathbf{x})\Big)$.
- Then, population GISO for node i is : $\mathbb{E}\Big[\exp\big(-g(\boldsymbol{\theta}, \mathbf{x})\big)\Big]$.

Continuous Markov Random Fields

Beyond the Gaussian case

- Most of the existing methods work with the following extension of the Ising model -

$$f_{\mathbf{x}}(\mathbf{x}) \propto \exp\left(\sum_{i\in[p]} \theta^{(i)} x_i + \sum_{(i,j)\in E} \theta^{(ij)} x_i x_j\right).$$

Our method is applicable to a large class of distributions beyond this.

- 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.

Problem Formulation

Continuous Random Variables

- *Bounded domain*
- *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 - Polynomial basis, Harmonic basis

- *Bounded parameters*
- *Sparsity* : Maximum degree of the underlying graph is at-most d.

Algorithm

Overview

1. First, we recover the graph structure and the associated edge parameters —
    1.1. Extend the GISO to the continuous setting
2. Second, we recover the node structure —
    2.1. Transform the problem of learning node parameters to a sparse linear regression
    2.2. Use a robust variation of lasso, and knowledge of the learned edge parameters

Algorithm

Learning edge parameters

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

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

  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.

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

- The finite sample GISO can recover the 'edge' parameters in continuous graphical models as well!

Algorithm

Learning node parameters

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

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) \propto \exp\left(\boldsymbol{\lambda}^{*T}(x_{-i})\boldsymbol{\phi}(x_i)\right)$$

  where $\boldsymbol{\lambda}^*(x_{-i})$, the canonical parameter, is linear function of node and edge parameters.

- Let $\boldsymbol{\mu}^*(x_{-i}) = \mathbb{E}[\boldsymbol{\phi}(\mathsf{x}_i)|X_{-i} = x_{-i}]$.
- By duality of exponential family, if we know $\boldsymbol{\mu}^*(x_{-i})$, we can recover $\boldsymbol{\lambda}^*(x_{-i})$.
- Learning $\boldsymbol{\mu}^*(x_{-i})$, can be viewed as a traditional regression problem.
- $\boldsymbol{\mu}^*(\cdot)$ is Lipschitz ⟶ approximately *linearize* it ⟶ sparse linear regression.

Main results

GISO - KL Divergence

- The population GISO is equivalent to a “local” MLE!
- Theorem 1. Consider $i \in [p]$. Then, with $D(\cdot \parallel \cdot)$ representing a node-specific KL-divergence,

$$\arg\min \text{Population GISO} = \arg\min D(\cdot \parallel \cdot)$$

Main results

Consistency and Normality

- Theorem 2. Consider $i \in [p]$. Then,
    A. The finite sample estimate of GISO is asymptotically consistent!
    B. Under some mild conditions, the finite sample estimate of GISO is asymptotically normal!

↓

- Even though the traditional MLE is intractable, this ‘local’ M-estimation is tractable.
- However, unlike traditional MLE, this is not asymptotically efficient.

Main results

Finite Sample Guarantees

- Theorem 3. Structure recovery can be achieved with $\Omega\big(\exp(d)\log(p)\big)$ samples
- Theorem 4. Parameter recovery can be achieved with $\Omega\big(\exp(d)\log(p)\big)$ samples

Thank you!