Statistical Inference for Pairwise Graphical Models Using Score Matching

Ming Yu[1], Varun Gupta[1], Mladen Kolar[1]
[1]University of Chicago Booth School of Business

Introduction Undirected probabilistic graphical models are widely used to explore and represent dependencies between random variables. We consider pairwise interaction graphical models with densities belonging to an exponential family and write

$$\log p_\theta(x) = \theta^T t(x) - \Psi(\theta) + h(x) \tag{1}$$

The main focus of the paper is on construction of an asymptotically normal estimator for parameters in (1) and performing (asymptotic) inference for them. We illustrate a procedure for construction of valid confidence intervals and propose a statistical test for existence of edges. Our inference results are robust to model selection mistakes, which commonly occur in ultra-high dimensional setting.

Assume we are interested in the edge between node a and b. Let $\theta^{ab} \in \mathbb{R}^{2p-1}$ denote the vector

$$\theta^{ab} = (\theta_{ab}, \underbrace{\theta_{a1}, ..., \theta_{ap}}_{\text{index for } a}, \underbrace{\theta_{1b}, ..., \theta_{pb}}_{\text{index for } b})^T.$$

We use Hyvärinen scoring rule to estimate θ^{ab}, as in [2]. Compared to previous work on high-dimensional inference in graphical models, this is the first work on inference in models where computing the normalizing constant is intractable.

Score Matching Let $X \in \mathcal{X}$ be a random variable, and let $\mathcal{P}$ be a family of distributions over $\mathcal{X}$. A scoring rule $S(x, Q)$ is a function that quantifies accuracy of $Q \in \mathcal{P}$, introduced in [1]. One finds optimal score estimator $\widehat{Q} \in \mathcal{P}$ that minimizes the empirical score

$$\widehat{Q} = \arg\min_{Q \in \mathcal{P}} \mathbb{E}_n \left[S(x_i, Q)\right].$$

with the scoring rule

$$S(x, Q) = (1/2)||\nabla \log q(x)||_2^2 + \Delta \log q(x).$$

In **exponential family**, for fixed indices (a, b), let $q_\theta^{ab}(x)$ be the conditional density of (X_a, X_b) given $X_{-ab} = x_{-ab}$. We have

$$\log q_\theta^{ab}(x) = \langle \theta^{ab}, \varphi(x)\rangle - \Psi^{ab}(\theta, x_{-ab}) + h^{ab}(x)$$

We then have the following scoring rule

$$S^{ab}(x, \theta) = (1/2)\theta^T \Gamma(x)\theta + \theta^T g(x), \tag{2}$$

where $\Gamma = \varphi_1\varphi_1^T + \varphi_2\varphi_2^T$, with $\varphi_1 = (\partial/\partial x_a)\varphi$, $\varphi_2 = (\partial/\partial x_b)\varphi$, $g = \varphi_1 h_1^{ab} + \varphi_2 h_2^{ab} + \Delta_{ab}\varphi(x)$, $h_1^{ab} = (\partial/\partial x_a)h^{ab}$, and $h_2^{ab} = (\partial/\partial x_b)h^{ab}$.

The scoring rule can be easily extended to non-negative data with different formulas.

Methodology Our three steps procedure for estimating θ_{ab}

Step 1: Find pilot estimator of θ^{ab} by solving the following problem and let $\widehat{M}_1 = M(\widehat{\theta}^{ab}) := \{(c, d) \mid \widehat{\theta}_{cd}^{ab} \neq 0\}$.

$$\widehat{\theta}^{ab} = \arg\min_{\theta \in \mathbb{R}^{s'}} \mathbb{E}_n \left[S^{ab}(x_i, \theta)\right] + \lambda||\theta||_1 \tag{3}$$

Step 2: Let $\widehat{\gamma}^{ab}$ be a minimizer of

$$\frac{1}{2}\mathbb{E}_n[(\varphi_{1,ab}(x_i) - \varphi_{1,-ab}(x_i)^T\gamma)^2 + (\varphi_{2,ab}(x_i) - \varphi_{2,-ab}(x_i)^T\gamma)^2] + \lambda||\gamma||_1.$$

Step 3: Let $\widetilde{M} = \{(a, b)\} \cup \widehat{M}_1 \cup M(\widehat{\gamma}^{ab})$. Our estimator is

$$\widetilde{\theta}^{ab} = \arg\min \mathbb{E}_n \left[S^{ab}(x_i, \theta)\right] \quad \text{s.t.} \quad M(\theta) \subseteq \widetilde{M}. \tag{4}$$

This is an extended abstract related to the existing publication at NIPS 2016. The full paper website is here.

Assumptions We provide high-level conditions that allow us to establish properties of each step in our procedure.

1. Model Sparsity: Let

$$\gamma^{ab,*} = \text{argmin } \mathbb{E}[(\varphi_{1,ab}(x_i) - \varphi_{1,-ab}(x_i)^{\mathrm{T}}\gamma)^2 + (\varphi_{2,ab}(x_i) - \varphi_{2,-ab}(x_i)^{\mathrm{T}}\gamma)^2]$$

We have sparsity: $m = |M(\theta^{ab,*})| \vee |M(\gamma^{ab,*})| \ll n$.

2. Sparse Eigenvalue: The following event holds with high probability

$$\mathcal{E}_{\text{SE}} = \{\phi_{\min} \leq \phi_-(m\log n, \mathbb{E}_n[\Gamma(x_i)]) \leq \phi_+(m\log n, \mathbb{E}_n[\Gamma(x_i)]) \leq \phi_{\max}\}$$

3. Finite Moment: Both $\mathbb{E}_{q^{ab}}\left[||\Gamma(X_a, X_b, x_{-ab})\theta^{ab,*}||^2\right]$ and $\mathbb{E}_{q^{ab}}\left[||g(X_a, X_b, x_{-ab})||^2\right]$ are finite.

Theorem Suppose the above assumptions hold. Define w^* with $w_{ab}^* = 1$ and $w_{-ab}^* = -\gamma^{ab,*}$, we have

$$\sqrt{n}\left(\widetilde{\theta}_{ab} - \theta_{ab}^*\right) = -\sigma_n^{-1} \cdot \sqrt{n}\mathbb{E}_n\left[w^{*,\mathrm{T}}\left(\Gamma(x_i)\theta^{ab,*} + g(x_i)\right)\right] + \mathcal{O}\left(\phi_{\max}^2\phi_{\min}^{-4} \cdot \sqrt{n}\lambda^2 m\right)$$

where $\sigma_n = \mathbb{E}_n[\eta_{1i}\varphi_{1,ab}(x_i) + \eta_{2i}\varphi_{2,ab}(x_i)]$.

When $\boldsymbol{(m\log p)^2/n = o(1)}$, we have

$$\boldsymbol{\sqrt{n}(\widetilde{\theta}_{ab} - \theta_{ab}^*) \longrightarrow_D N(0, V) + o_p(1)}$$

where $V = (\mathbb{E}[\sigma_n])^{-2} \cdot \text{Var}\left(w^{*,\mathrm{T}}\left(\Gamma(x_i)\theta^{ab,*} + g(x_i)\right)\right)$. We estimate V using $\widetilde{\theta}^{ab}$ and $\widetilde{\gamma}^{ab}$. Using this estimate, we have

$$\lim_{n\to\infty} \sup_{\theta^* \in \Theta} \mathbb{P}_{\theta^*}\left(\theta_{ab}^* \in \widetilde{\theta}_{ab} \pm z_{\alpha/2} \cdot \sqrt{\widehat{V}/n}\right) = \alpha + o(1).$$

Experimental results We illustrate finite sample properties of our inference procedure on data simulated from three different Exponential family distributions: Gaussian Graphical Model, Normal Conditionals, and Exponential Graphical Model (non-negative data). In each example, we report the mean coverage rate of 95% confidence intervals for several coefficients averaged over 500 independent simulation runs.

Table 1: Empirical Coverage for the 3 Models

	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$	$w_{1,10}$
Gaussian Graphical Model	94.6%	92.4%	92.6%	94.0%
Normal Conditionals	93.2%	93.4%	94.6%	95.0%
Exponential Graphical Model	92.6%	92.0%	92.2%	92.4%

In general, non-negative score matching is harder than regular score matching [2]. We can see that our method works quite well on the first two models; while for Exponential Graphical Model the empirical coverage rate tends to be about 92%, rather than the designed 95% - still impressive for the not so large sample size.

[1] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6 (Apr):695–709, 2005.

[2] Lina Lin, Mathias Drton, Ali Shojaie, et al. Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854, 2016.