Provable Gaussian Embedding with One Observation

Ming Yu, Zhuoran Yang, Tuo Zhao, Mladen Kolar, Zhaoran Wang

Abstract

- **Fact:** Exponential family embedding is a powerful technique. However, all the existing works are empirical.
- **Contribution:** First theoretical result for exponential family embedding models. We focus on Gaussian embedding and show that the embedding structure can be learned from *one observation*.
- **Assumptions:** Weak dependency among the nodes.
- **Algorithms:** Convex relaxation and non-convex formulation.
- **Theoretical result:** Guaranteed linear convergence up to statistical error for both algorithms.
- **Extension:** The theoretical framework is for *general* exponential family embedding models. For other models, all we need are more complicated probabilistic tools.

Exponential family embedding

- A *known* graph $G = (V, E)$ and the conditional exponential family
- We have m vertices and we observe a p-dimensional vector $x_j \in \mathbb{R}^p$ at vertex j. Let $X = (x_1, \ldots, x_m) \in \mathbb{R}^{p\times m}$ be the data matrix
- Let $c_j = \{k \in V : (j,k) \in E\}$ be the *k*nown context of j
- x_j conditioning on x_{c_j} follows an exponential family distribution

$$x_j | x_{c_j} \sim \texttt{ExponentialFamily}\Big[V \sum_{k\in c_j} V^\top x_k,\ t(x_j)\Big]$$

- The matrix $V \in \mathbb{R}^{p\times r}$ embeds the vector $x_k \in \mathbb{R}^p$ to a lower r-dimensional space with $V^\top x_k \in \mathbb{R}^r$ being the embedding of x_k.
- Examples of context structure:

Figure: Chain

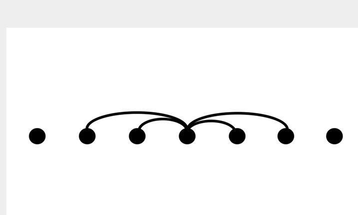

Figure: Nearest neighbor

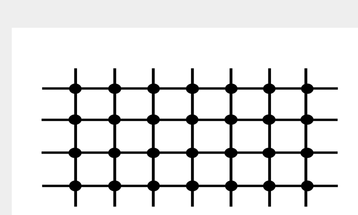

Figure: Lattice

Gaussian embedding

$$x_j | x_{c_j} \sim N\Big(V \sum_{k\in c_j} V^\top x_k, \Sigma_j\Big).$$

Word embedding (cbow)

$$p(x_j | x_{c_j}) = \frac{\exp\left[x_j^\top V \left(\sum_{k\in c_j} V^\top x_k\right)\right]}{\sum_j \exp\left[x_j^\top V \left(\sum_{k\in c_j} V^\top x_k\right)\right]}.$$

Poisson embedding

$$x_j | x_{c_j} \sim \text{Poisson}\Big(\exp\big(V \sum_{k\in c_j} V^\top x_k\big)\Big).$$

Gaussian embedding

- Define $M = VV^\top$, we have

$$x_j | x_{c_j} \sim N\Big(V \sum_{k\in c_j} V^\top x_k, \Sigma_j\Big) = N\Big(M \sum_{k\in c_j} x_k, \Sigma_j\Big).$$

- Let $X_{\text{col}} = [x_1^\top, x_2^\top, ..., x_m^\top]^\top \in \mathbb{R}^{pm\times 1}$ be the column vector obtained by stacking columns of $X \in \mathbb{R}^{p\times m}$.
- Under mild conditions, the conditional distributions are strongly compatible and we have $X_{\text{col}} \sim N(0, \Sigma_{\text{col}})$.
- Let $A \in \mathbb{R}^{m\times m}$ denote the adjacency matrix, with $a_{j,k} = 1$ when there is an edge between nodes j and k and 0 otherwise.
- The Hessian matrix is given by

$$H = \frac{1}{m}\sum_{j=1}^{m}\Big(\sum_{k\in c_j} x_k\Big)\cdot\Big(\sum_{k\in c_j} x_k\Big)^\top = \frac{1}{m} X A A^\top X^\top \in \mathbb{R}^{p\times p}$$

Estimation

- **Loss function**

$$\mathcal{L}(M) = \frac{1}{2m}\sum_{j=1}^{m}\Big\| x_j - M\sum_{k\in c_j} x_k\Big\|^2.$$

- **Convex Relaxation**

$$\min_{M\in\mathbb{R}^{p\times p}, M^\top = M, M\succeq 0} \mathcal{L}(M) + \lambda \|M\|_*.$$

The algorithm is proximal gradient descent method on M.

- **Non-convex Optimization**

$$\min_{V\in\mathbb{R}^{p\times r}} \mathcal{L}(VV^\top).$$

The algorithm is gradient descent on V:

$$V^{(t+1)} = V^{(t)} - \eta \cdot \nabla_V \mathcal{L}(VV^\top)\big|_{V=V^{(t)}}.$$

Assumptions

- **Assumption EC. [Eigenvalue]** The minimum and maximum eigenvalues of $\mathbb{E}H$ are bounded from below and from above:

$$0 < c_{\min} \le \sigma_{\min}(\mathbb{E}H) \le \sigma_{\max}(\mathbb{E}H) \le c_{\max} < \infty.$$

- **Assumption SC. [Weak dependence]** There exists a constant ρ_0 such that

$$\max\big\{\|A\|_2, \|\Sigma_{\text{col}}^{1/2}\|_2\big\} \le \rho_0.$$

Theoretical result

- **Lemma.** Suppose the assumptions (EC) and (SC) are satisfied. Then for $m \ge c_0 p$ we have

$$\frac{1}{2}c_{\min} \le \sigma_{\min}(H) \le \sigma_{\max}(H) \le 2c_{\max}$$

with high probability. Therefore we have

$$\kappa_\mu \cdot \|\Delta\|_F^2 \le \delta\mathcal{L}(\Delta) \le \kappa_L \cdot \|\Delta\|_F^2$$

for any $\Delta \in \mathbb{R}^{p\times p}$ where

$$\delta\mathcal{L}(\Delta) = \mathcal{L}(M^* + \Delta) - \mathcal{L}(M^*) - \langle \nabla\mathcal{L}(M^*), \Delta\rangle.$$

- **Theorem. [Convex]** Suppose the assumptions (EC) and (SC) are satisfied. Taking $\lambda = \mathcal{O}\big(\sqrt{p/m}\big)$, we have

$$\|\widehat{M} - M^*\|_F = \mathcal{O}_P\Big(\frac{1}{\kappa_\mu}\sqrt{\frac{pr}{m}}\Big).$$

- **Theorem. [Non-convex]** Suppose the assumptions (EC) and (SC) are satisfied. After T iterations we have

$$d^2\Big(V^{(T)}, V^*\Big) \le \beta^T d^2\Big(V^{(0)}, V^*\Big) + \frac{C}{\kappa_\mu^2}\cdot e_{\text{stat}}^2,$$

for some constant $\beta < 1$ and a constant C.

The subspace distance is defined as

$$d^2(V, V^*) = \min_{OO^\top = O^\top O = I} \|V - V^* O\|_F^2.$$

The statistical error is defined as

$$e_{\text{stat}} = \sup_{\Delta\in\Omega} \big\langle \nabla\mathcal{L}(M^*), \Delta\big\rangle,$$

where

$$\Omega = \big\{\Delta : \Delta \in \mathbb{R}^{p\times p}, \Delta = \Delta^\top, \text{rank}(\Delta) = 2r, \|\Delta\|_F = 1\big\}.$$

Experiment

- We generate the data according to the conditional distribution using Gibbs Sampling.
- We set $p = 100, r = 5$ and vary the number of nodes m. We set $\Sigma_j = \Sigma$ to be a Toeplitz matrix with $\Sigma_{i\ell} = \rho^{|i-\ell|}$ with $\rho = 0.3$.
- The metric is the estimation error $\|\widehat{M} - M^*\|_F / \|M^*\|_F$.

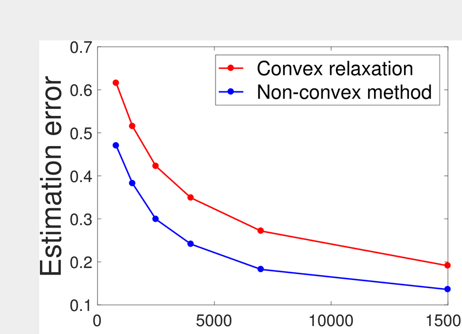

Figure: Chain

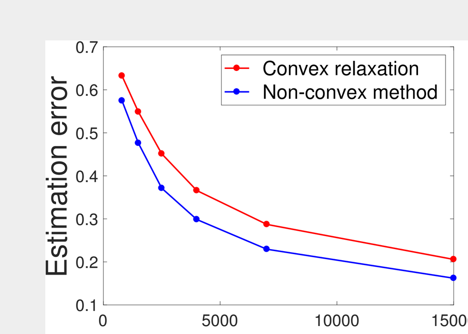

Figure: Nearest neighbor

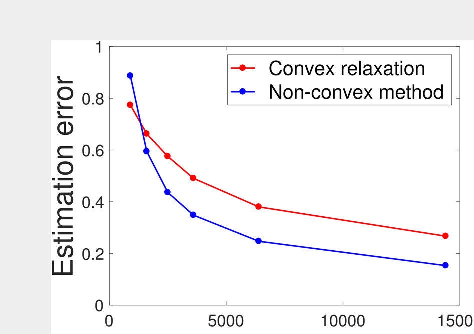

Figure: Lattice

ming93@uchicago.edu