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 existing works are empirical.
- **Contribution**: First theoretical result for exponential family embedding models. We focus on Gaussian embedding and show that the theoretical 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 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 $\ell_j \in \{k \in V : (j, k) / E\}$ be the known context of $j$.
- $x_j$ conditioning on $x_{\ell_j}$ follows an exponential family distribution $x_j | x_{\ell_j} \sim \text{ExponentialFamily}(V \sum_{k \in \ell_j} V^\top x_k, t(x_j))$.

Gaussian embedding

- Define $M = VV^\top$, we have $x_j | x_{\ell_j} \sim N(\sum_{k \in \ell_j} V^\top x_k, \Sigma_j) = N(M \sum_{k \in \ell_j} x_k, \Sigma_j)$.
- Let $X_{\ell_j} = [x^\top_1, x^\top_2, \ldots, x^\top_m]^\top \in \mathbb{R}^{m \times 1}$ 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 $X_{\ell_j} \sim N(0, \Sigma_{\ell_j})$.
- 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 $H = \frac{1}{m} \sum_{j=1}^m \left( \sum_{k \in \ell_j} x_k \right) \left( \sum_{k \in \ell_j} x_k \right)^\top = \frac{1}{m} X A A^\top X^\top \in \mathbb{R}^{p \times p}$.

Estimation

- **Loss function**: $\mathcal{L}(M) = \frac{1}{2m} \sum_{j=1}^m \| x_j - M \sum_{k \in \ell_j} x_k \|^2$. The algorithm is proximal gradient descent method on $M$.

- **Convex Relaxation**: $\min_{M \in \mathbb{R}^{p \times p}, M^\top = M, \lambda \geq 0} \mathcal{L}(M) + \lambda \| M \|_1$.

- **Non-convex Optimization**: $\min_{M \in \mathbb{R}^{p \times p}} \mathcal{L}(VV^\top)$.

Assumptions

- **Assumption EC. [Eigenvalue]** The minimum and maximum eigenvalues of $EH$ are bounded from below and from above: $0 < \sigma_{\text{min}}(EH) \leq \sigma_{\text{max}}(EH) \leq c_{\text{max}} < \infty$.
- **Assumption SC. [Weak dependence]** There exists a constant $\rho_0$ such that $\max \left\{ \| A \|_2, \| \sum_{j=1}^m V^\top x_j \|_2 \right\} \leq \rho_0$.

Theoretical result

- **Lemma**: Suppose the assumptions (EC) and (SC) are satisfied. Then for $m \geq \bar{m}$ we have $\frac{1}{\rho_0} \frac{\sigma_{\text{min}}(H)}{\sigma_{\text{max}}(H)} \leq \sigma_{\text{min}}(H) \leq \sigma_{\text{max}}(H) \leq 2c_{\text{max}}$ with high probability. Therefore we have $\kappa_p \cdot \| \Delta \|_2 \leq \kappa_k \cdot \| \Delta \|_2$ for any $\Delta \in \mathbb{R}^{p \times p}$.

- **Theorem. [Convex]** Suppose the assumptions (EC) and (SC) are satisfied. Taking $\lambda = \mathcal{O}(\sqrt{p/m})$, we have $\| \tilde{M} - M^* \|_F = \mathcal{O}(1/\kappa_p \sqrt{p/m})$.

- **Theorem. [Non-convex]** Suppose the assumptions (EC) and (SC) are satisfied. After $T$ iterations we have $\mathcal{O}(\sqrt{\kappa_p^2 / \kappa_k^2 \epsilon_{\text{stat}}}^2)$.

Experiment

- **Assumption** EC and SC. After $T$ iterations we have $\mathcal{O}(\sqrt{\kappa_p^2 / \kappa_k^2 \epsilon_{\text{stat}}}^2)$.

- **Theorem. [Convex]** Suppose the assumptions (EC) and (SC) are satisfied. Taking $\lambda = \mathcal{O}(\sqrt{p/m})$, we have $\| \tilde{M} - M^* \|_F = \mathcal{O}(1/\kappa_p \sqrt{p/m})$.

- **Theorem. [Non-convex]** Suppose the assumptions (EC) and (SC) are satisfied. After $T$ iterations we have $\mathcal{O}(\sqrt{\kappa_p^2 / \kappa_k^2 \epsilon_{\text{stat}}}^2)$.