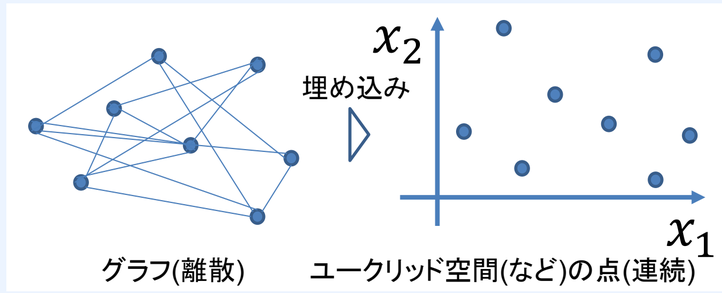
2-21: Dimensionality and Curvature Selection of Graph Embedding using Decomposed Normalized Maximum Likelihood Code-Length

(DNML符号長を用いたグラフ埋め込みの次元・曲率選択; ICDM2023)

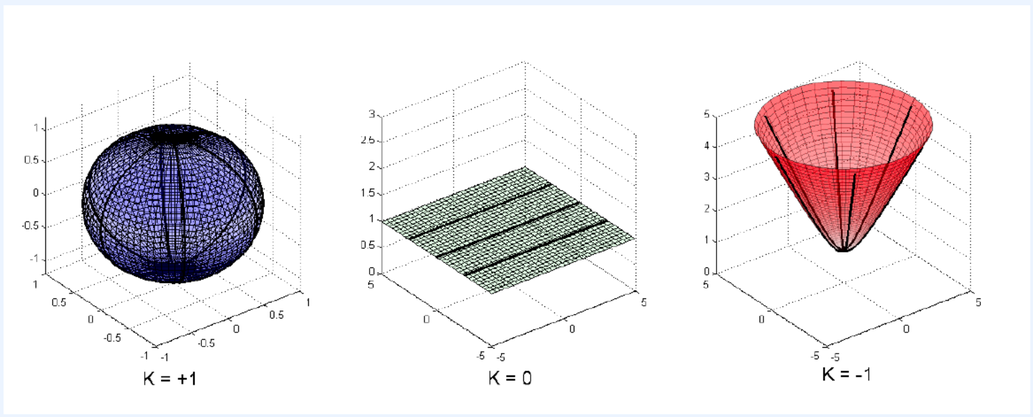
結城 凌 (東京大学); 鈴木 惇 (King's College London); 山西 健司 (東京大学)

研究目的

- グラフ埋め込み研究の発展 [1, 2].



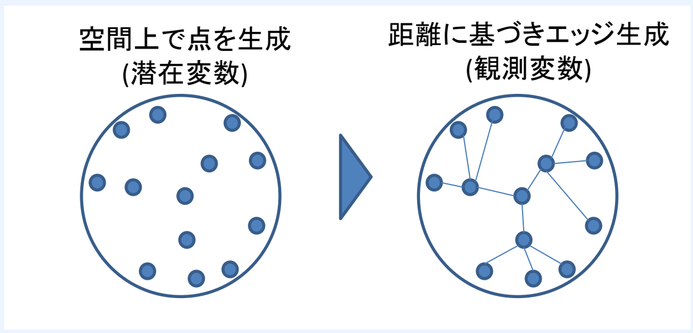
- 次元・空間の選択研究 [3, 4, 5].
  - リンク予測・ノード分類の性能に影響.
- ユークリッド空間・双曲空間・球面.
  - 単連結定曲率リーマン多様体.
  - 埋め込み先空間としてスタンダード.
  - 空間選択→曲率の選択.



本研究の貢献:次元および曲率の同時選択手法の考案.

提案手法

- 潜在変数モデルを設定 [5].
  - Wrapped Normal Distribution[6, 7]を用いる.
  - ガウス分布の多様体への拡張.



- Minimum Description Length (MDL) 原理 [8]を用いる.
  - 統計的モデル選択基準の1つ. 与えられたデータを最も短く符号化するモデルを最良とする.
  - 一致性 [9, 10]・PAC学習での早い収束 [11]など.
- Decomposed Normalized Maximum Likelihood (DNML)符号長 [12].
  - 潜在変数モデルで簡便に符号長を計算可能な手法.

$$L_{\mathrm{DNML}}(y, z) := L_{\mathrm{NML}}(y|z) + L_{\mathrm{NML}}(z),$$

$$L_{\mathrm{NML}}(y|z) := -\log p(y|z; \hat{\beta}(y, z)) + \log \sum_{\bar{y}} p(\bar{y}|z; \hat{\beta}(\bar{y}, z)),$$

$$L_{\mathrm{NML}}(z) := -\log p(z; \hat{\sigma}(z)) + \log \int p(\bar{z}; \hat{\sigma}(\bar{z}))d\bar{z}_{1:D},$$

DNMLを最小化する次元D・曲率K選択.

*画像は[5]から引用.

[1] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "Line: Large-scale information network embedding," in Proceedings of the 24th international conference on world wide web, 2015, pp. 1067–1077.
[2] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–
864.
[3] Z. Yin and Y. Shen, "On the dimensionality of word embedding," in Proceedings of the 32nd International Conference on Neural Information
Processing Systems, 2018, pp. 895–906.

実験および結果

1. 人工データでの次元・曲率同定実験.

Table 4.1: Accuracy of curvature sign estimation.

Dataset	# of nodes	DNML	AIC	BIC
E-8	400	**0.08**	0.00	0.00
	800	0.67	**0.75**	0.00
	1600	**1.00**	**1.00**	0.67
	3200	**1.00**	**1.00**	**1.00**
H-8	400	**1.00**	**1.00**	**1.00**
	800	0.92	**1.00**	**1.00**
	1600	0.83	**1.00**	**1.00**
	3200	**1.00**	**1.00**	**1.00**
S-8	400	**1.00**	0.33	0.33
	800	**1.00**	**0.67**	0.33
	1600	**1.00**	0.58	0.67
	3200	**1.00**	**1.00**	0.67

Table 4.4: Average Maps of each criterion (Average estimated dimensionalities in parentheses).

Dataset	# of nodes	DNML	AIC	BIC
E-8	400	**1.000**(8.0)	0.625(14.0)	0.625(5.0)
	800	0.750(12.0)	0.556(14.7	**0.833**(6.7)
	1600	**0.871**(9.3)	0.583(14.7)	0.486(3.7)
	3200	**1.000**(8.0)	**1.000**(8.0)	0.670(4.0)
H-8	400	**0.333**(2.0)	**0.333**(3.2)	**0.333**(2.0)
	800	0.333(3.5)	**0.513**(4.3)	0.333(2.0)
	1600	0.431(4.0)	**0.958**(8.7)	0.333(2.0)
	3200	**0.833**(6.7)	0.375(17.3)	0.333(3.8)
S-8	400	0.304(15.5)	**0.333**(3.5)	**0.333**(2.0)
	800	**0.424**(13.3)	**0.444**(3.83)	0.333(2.0)
	1600	0.544(14.3)	**0.625**(5.3)	0.333(2.5)
	3200	**0.708**(16.7)	**0.736**(7.3)	0.333(3.3)

ノードが多い場合，提案手法は高い精度(MAP)で真の次元と曲率を同定.

2. 論文共著関係ネットワーク [13]でのリンク予測性能および推定次元の関係.

Table 4.6: Selected dimensionality and curvature of each method.

Network	DNML	AIC	BIC
AstroPh	$\mathbb{H}^{12}_{-0.25}$	$\mathbb{H}^{15}_{-0.25}$	$\mathbb{H}^{8}_{-0.27}$
HepPh	$\mathbb{E}^{16}$	$\mathbb{H}^{14}_{-0.32}$	$\mathbb{H}^{6}_{-0.36}$
Airport	$\mathbb{E}^{7}$	$\mathbb{E}^{7}$	$\mathbb{H}^{3}_{-0.73}$
WN-mammal	$\mathbb{H}^{4}_{-0.69}$	$\mathbb{H}^{4}_{-0.71}$	$\mathbb{H}^{3}_{-0.72}$
WN-solid	$\mathbb{H}^{4}_{-0.69}$	$\mathbb{H}^{5}_{-0.70}$	$\mathbb{H}^{3}_{-0.71}$

Table 4.7: Average conciseness of each method.

Network	$\epsilon_{\max}$	DNML	AIC	BIC
AstroPh	0.05	**0.484**	**0.467**	0.362
	0.10	**0.532**	0.484	**0.540**
HepPh	0.05	**0.496**	0.398	0.351
	0.10	0.518	0.436	**0.544**
Airport	0.05	**0.390**	**0.405**	0.000
	0.10	**0.549**	**0.550**	0.425
WN-mammal	0.05	0.000	0.000	0.000
	0.10	**0.416**	0.385	0.341
WN-solid	0.05	0.398	**0.549**	0.080
	0.10	0.637	**0.676**	0.521

提案手法はリンク予測のAUCを保ちつつ低次元を選択(conciseness).

研究のポイント

- 次元と曲率の同時選択手法構築.
  - 先行研究の少なさ [14].
- 潜在変数モデルおよびDNMLの利用.
  - 潜在変数モデルを用いない場合，Non-identifiability Problemにより符号長の解析計算が非常に困難.
- 人工データ・実データでの性能検証.

[4] P. T. Hung and K. Yamanishi, "Word2vec skip-gram dimensionality selection via sequential normalized maximum likelihood," Entropy, vol. 23, no. 8, p. 997, 2021.
[5] Skopek, Ondrej, Octavian-Eugen Ganea, and Gary Bécigneul. "Mixed-curvature variational autoencoders." *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. International Conference on Learning Representations, 2020.
[6] Nagano, Yoshihiro, et al. "A wrapped normal distribution on hyperbolic space for gradient-based learning." *International Conference on Machine Learning*. PMLR, 2019.
[7] Skopek, Ondrej, Octavian-Eugen Ganea, and Gary Bécigneul. "Mixed-curvature Variational Autoencoders." *International Conference on Learning Representations*. 2019.
[8] Rissanen, Jorma. "Modeling by shortest data description." *Automatica* 14.5 (1978): 465-471.
[9] Rissanen, Jorma. *Optimal estimation of parameters*. Cambridge University Press, 2012.
[10] Yamanishi, Kenji. *Learning with the Minimum Description Length Principle*. Springer Nature, 2023.
[11] K. Yamanishi, "A learning criterion for stochastic rules," Machine Learning, vol. 9, no. 2-3, pp. 165–203, 1992.
[12] K. Yamanishi, T. Wu, S. Sugawara, and M. Okada, "The decomposed normalized maximum likelihood code-length criterion for selecting hierarchical latent variable models," Data Mining and Knowledge Discovery, vol. 33, no. 4, pp. 1017–1058, 2019.
[13] J. Leskovec and R. Sosi˘c, "Snap: A general-purpose network analysis and graph-mining library," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 8, no. 1, p. 1, 2016.
[14] P. Almagro, M. Boguna, and M. Serrano, "Detecting the ultra low dimensionality of real networks," arXiv preprint arXiv:2110.14507, 2021.