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

Ryo Yuki, Atsushi Suzuki, and Kenji Yamanishi

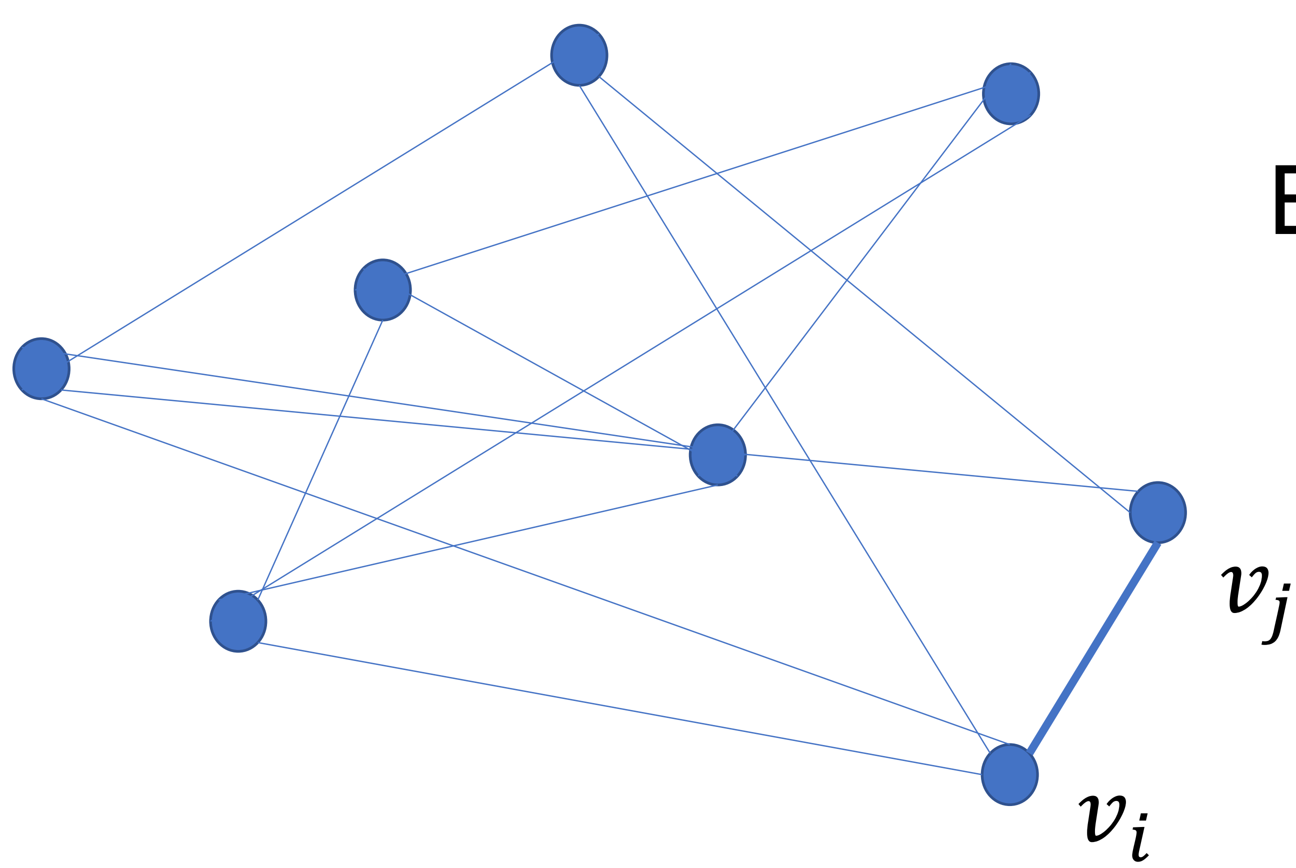
Supervisor: Kenji Yamanishi

Agenda

1. Background

Graph Embedding

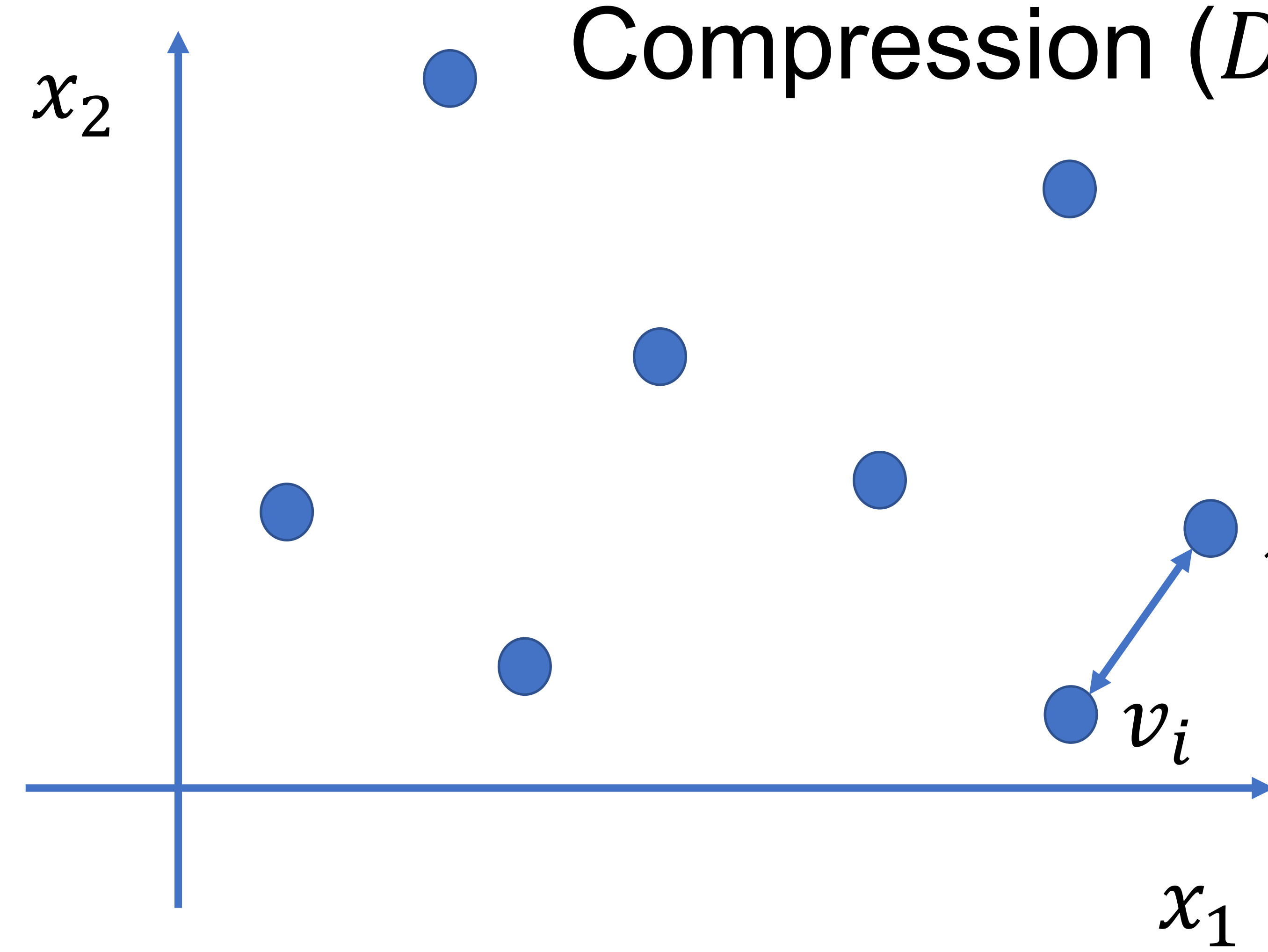
Convert discrete representation to continuous one.



(Attributed/
non-attributed)

Embed

Compression ($D \ll |V|$)



Input

Graphs can represent

- Social networks (Freeman 2000)
- Lexical networks (Ferrer+ 2001)
- Protein-protein interaction networks (Theocharidis+ 2009)

etc...

→ **Practically important!**

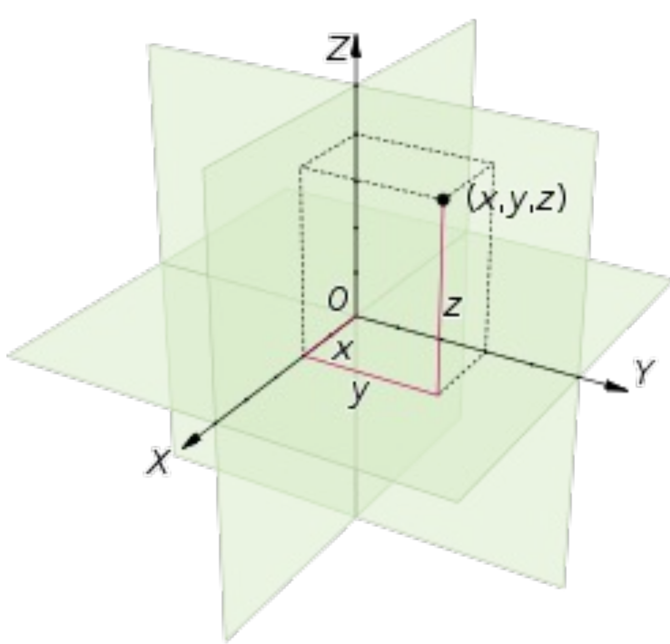
Downstream Tasks

- Node classification
- Link prediction
- Clustering
- Visualization (2-dimensional one)

Graph Embeddings on Riemannian Manifolds

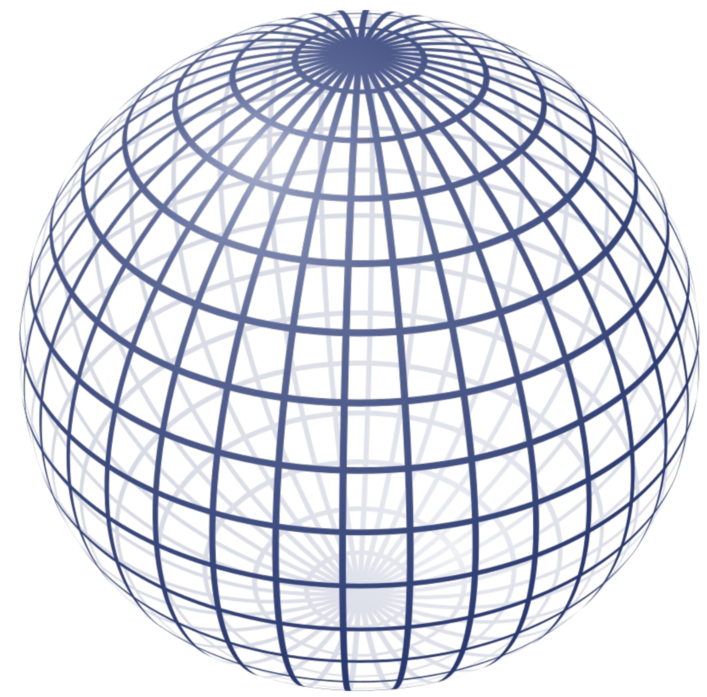
Recent Development of Graph Embeddings on Riemannian Manifolds.

Euclidean Space (Many studies)



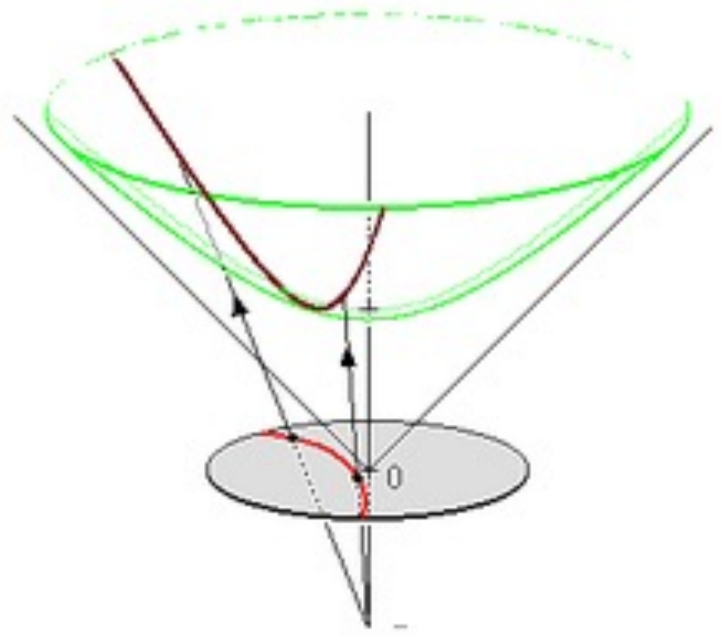
https://en.wikipedia.org/wiki/Euclidean_space

Spherical Space (Gu+ 2019)



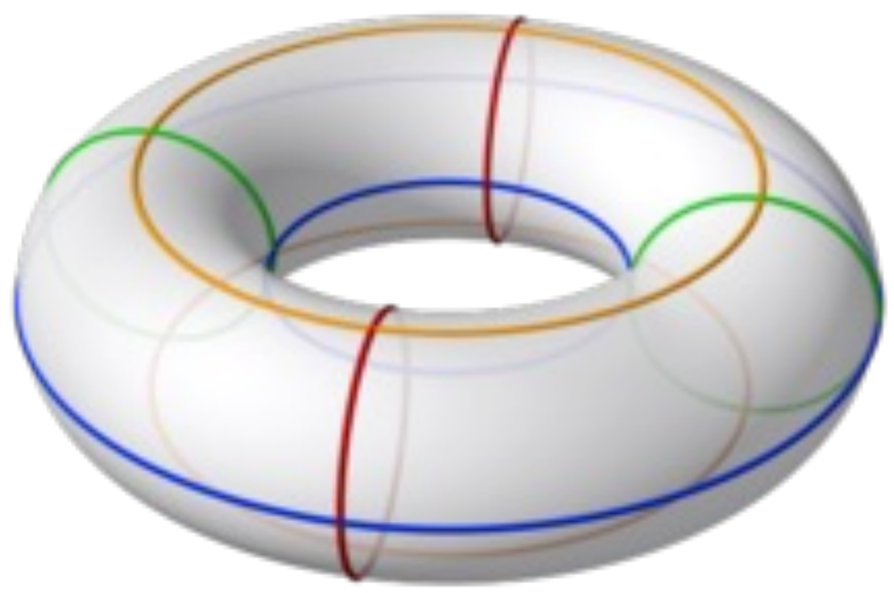
https://en.wikipedia.org/wiki/Sphere

Hyperbolic Space
(Nickel and Kiela 2017, 2018, etc)



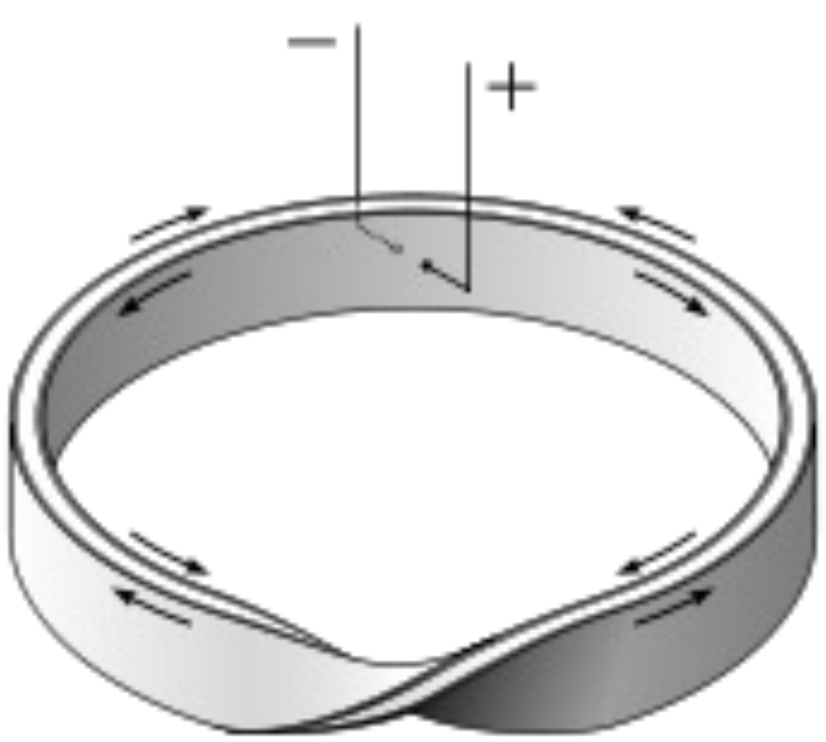
*https://en.wikipedia.org/wiki/Hyperboloid_model

Torus (Ebisu and Ichise 2018)



https://en.wikipedia.org/wiki/Torus

Möbius Ring (Chen+ 2021)

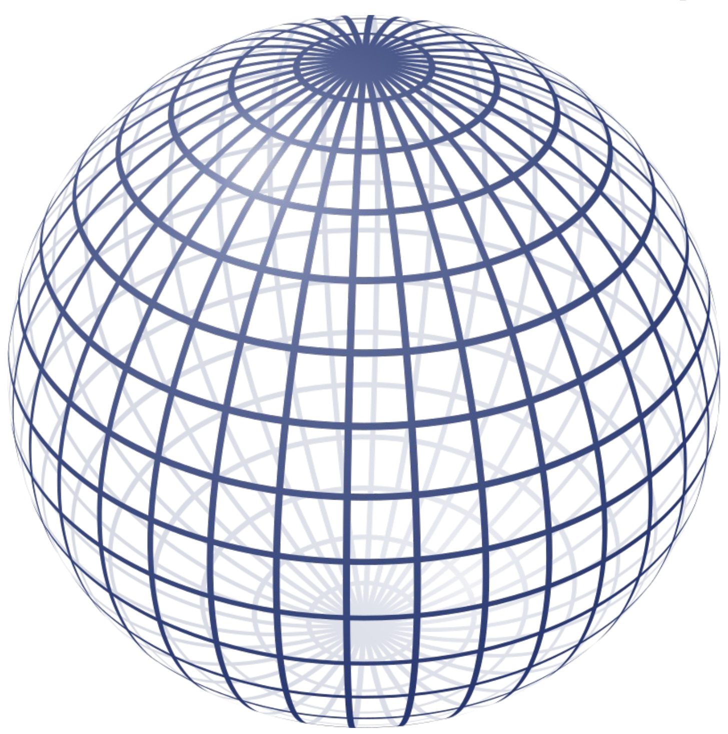


→How should we choose the best Riemmanian manifold associated with a given graph?

Model Spaces

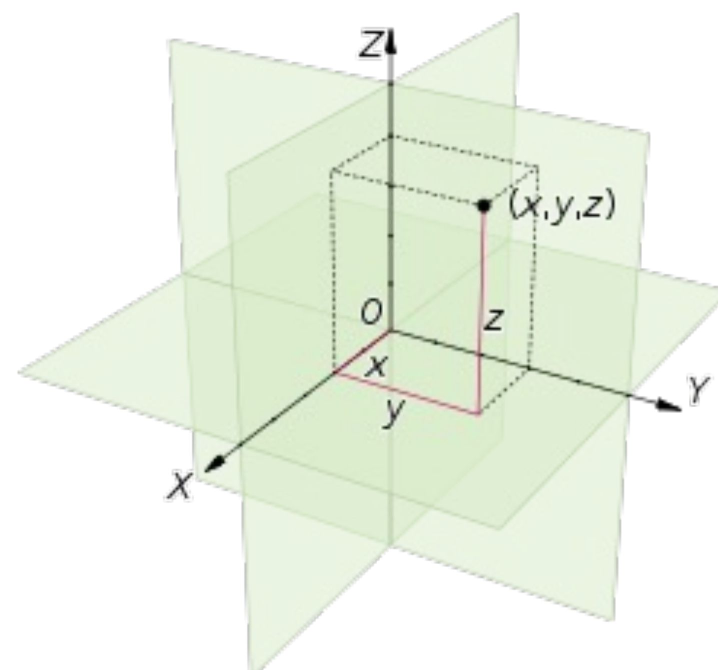
Spherical, Euclidean, and Hyperbolic spaces are chosen.

Spherical Space (K>0)



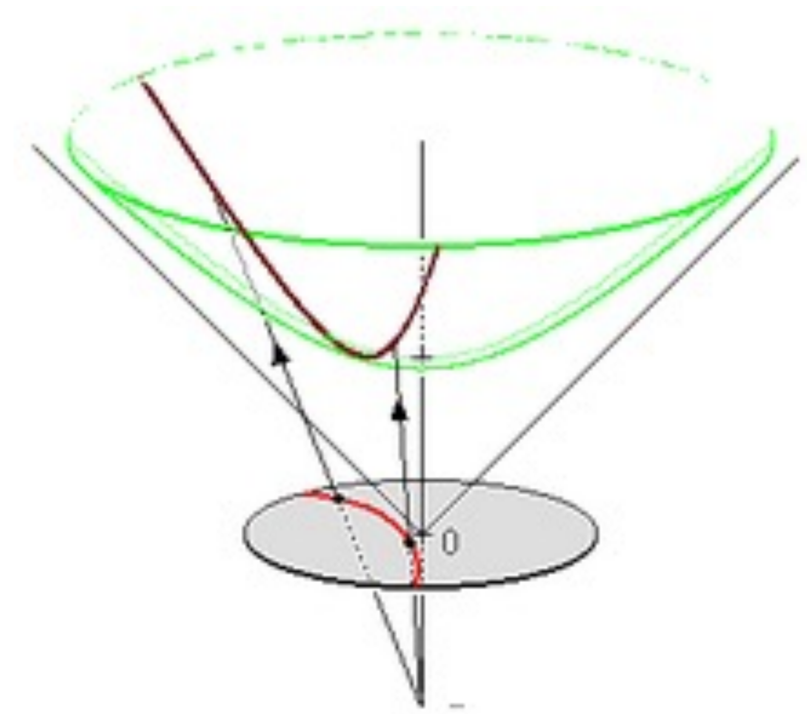
https://en.wikipedia.org/wiki/Sphere

Euclidean Space (K=0)



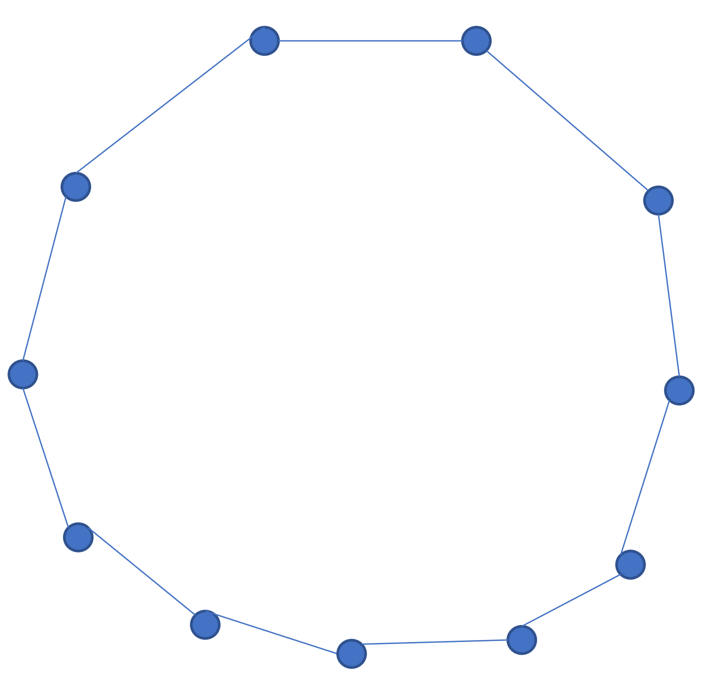
https://en.wikipedia.org/wiki/Euclidean_space

Hyperbolic Space (K<0)

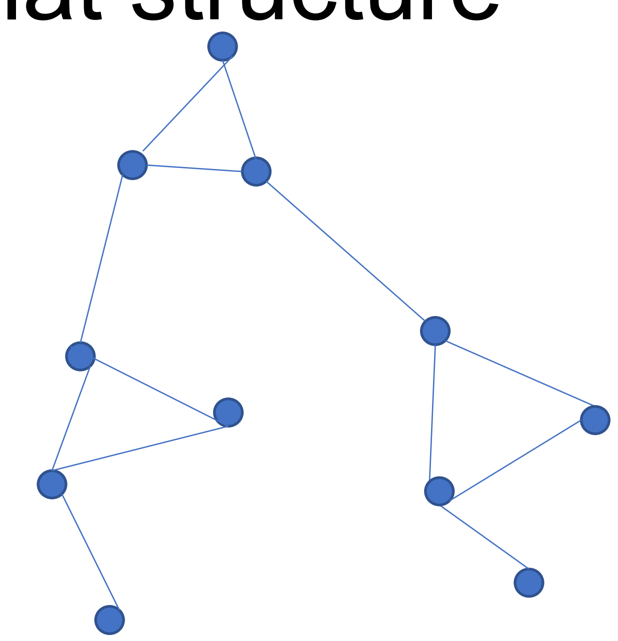


*https://en.wikipedia.org/wiki/Hyperboloid_model

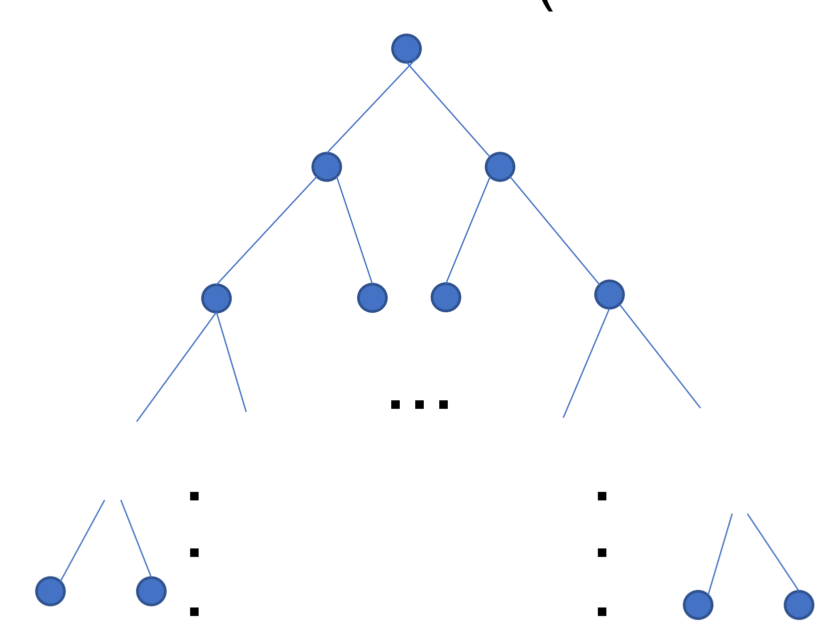
Cyclic structure (Gu+ 2019)



Flat structure



Tree-like structure (Krioukov+ 2010 etc)



Select D and K for the given graph.

MDL Principle (Rissanen 1978)

Select the model that minimizes the code-length.

Minimum Description Length (MDL) Principle

For data $x = x_1, \dots, x_n$ and model M, the MDL criterion is given by

$$MDL(x|M) = L(x|M) + L(M),$$

where $L(x|M)$ and $L(M)$ is encoding functions, and the best model is given by

$$\hat{M} = argmin_M \, MDL(x|M).$$

Example: Normalized Maximum Likelihood (NML) Code-Length (Shtarkov 1987)

For a parametric class of probability distributions $\mathcal{P}_M = \{p(x;\theta, M): \theta \in \Theta_M\}$, the NML code-length is given by

$$L_{NML}(x|M) = -\log p\left(x; \hat{\theta}(x), M\right) + \log C_n(M),$$

$C_n(M) = \sum_y p(y; \hat{\theta}(y), M)$: parametric complexity.

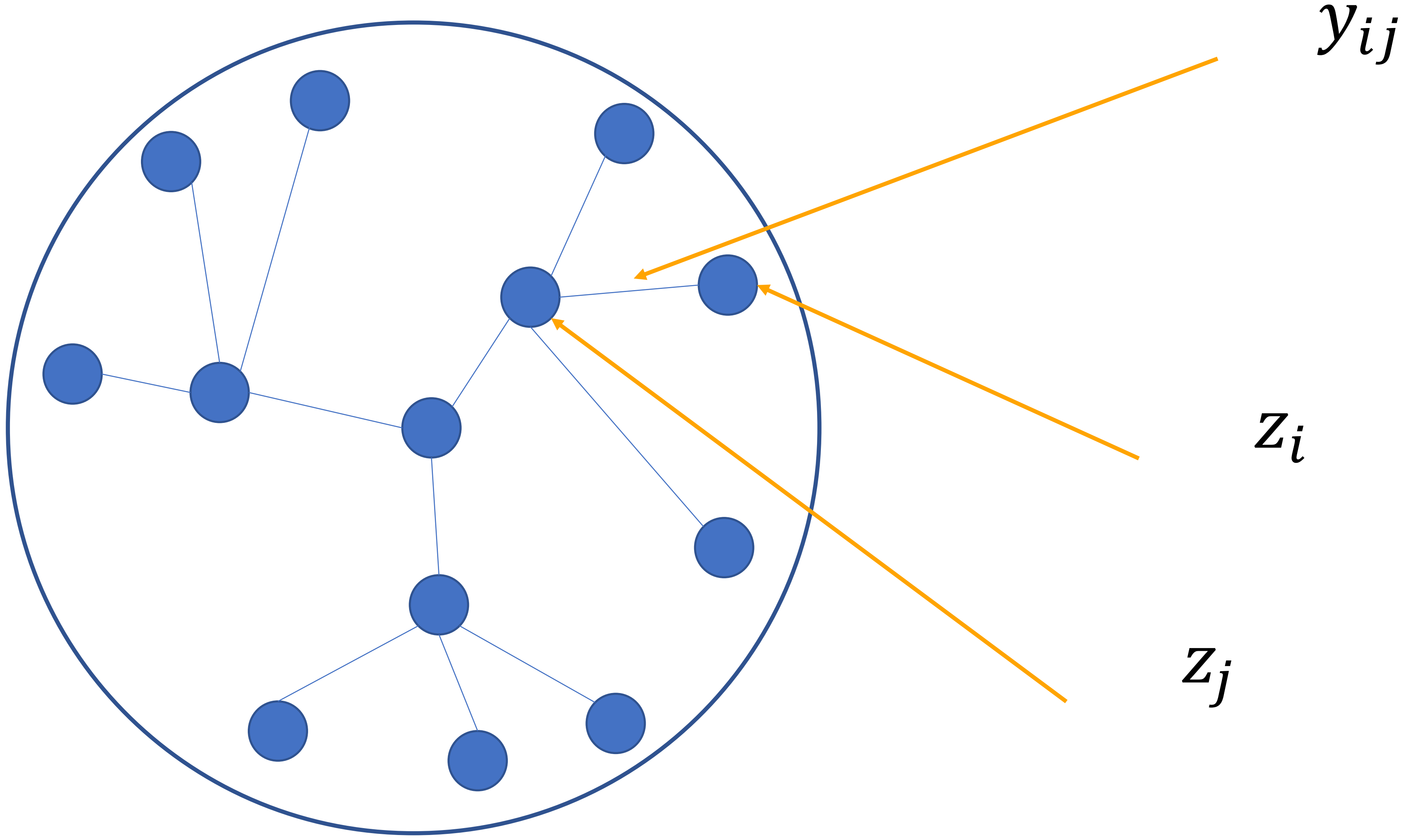
MDL model selection by regarding the dimensionality and similarity as a model.

Agenda

2. Formulation of Graph Embedding

Conventional Formulation of Graph Embedding (Nickel and Kiela 2018)

Connect points with logistic function.



$$p_K(\boldsymbol{y}; \boldsymbol{z}, \gamma) := \prod_{(i,j)\in\Lambda_n} p_K(y_{ij}; z_i, z_j, \gamma),$$

$$p_K(y_{ij}; z_i, z_j, \gamma) := \begin{cases} \frac{1}{1+\exp(d^K_{z_i z_j} - \gamma)} & (y_{ij} = 1), \\ 1 - \frac{1}{1+\exp(d^K_{z_i z_j} - \gamma)} & (y_{ij} = 0). \end{cases}$$

$\gamma > 0$

Logistic function.

Non-Identifiability Problem

Non-identifiability problem

Non-identifiability refers to a situation where there is no one-to-one correspondence between parameters and probability distributions. For all x, there exist $\theta_1 \neq \theta_2$ such that the following equation holds.

$$p(x;\theta_1) = p(x;\theta_2).$$

- The asymptotic normality for the maximum likelihood estimate does not hold.
- Conventional information criteria such as AIC (Akaike 1974), BIC (Schwarz 1978), etc… do not guarantee their rationales because their derivation depend on asymptotic theory.
- Calculation of the NML code-length is also difficult.

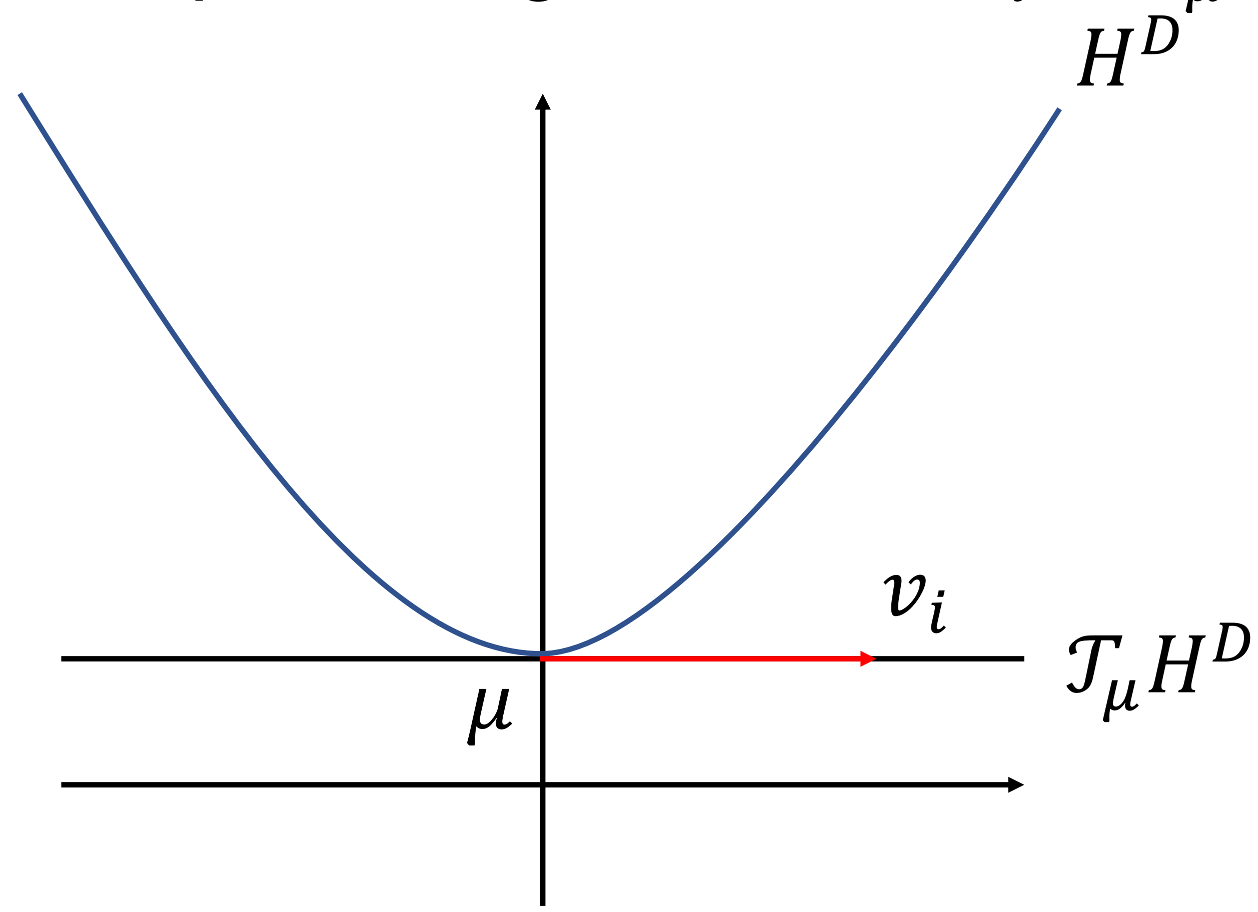
The conventional formulation of hyperbolic embedding is non-identifiable. Thus, we use latent variable models.

Agenda

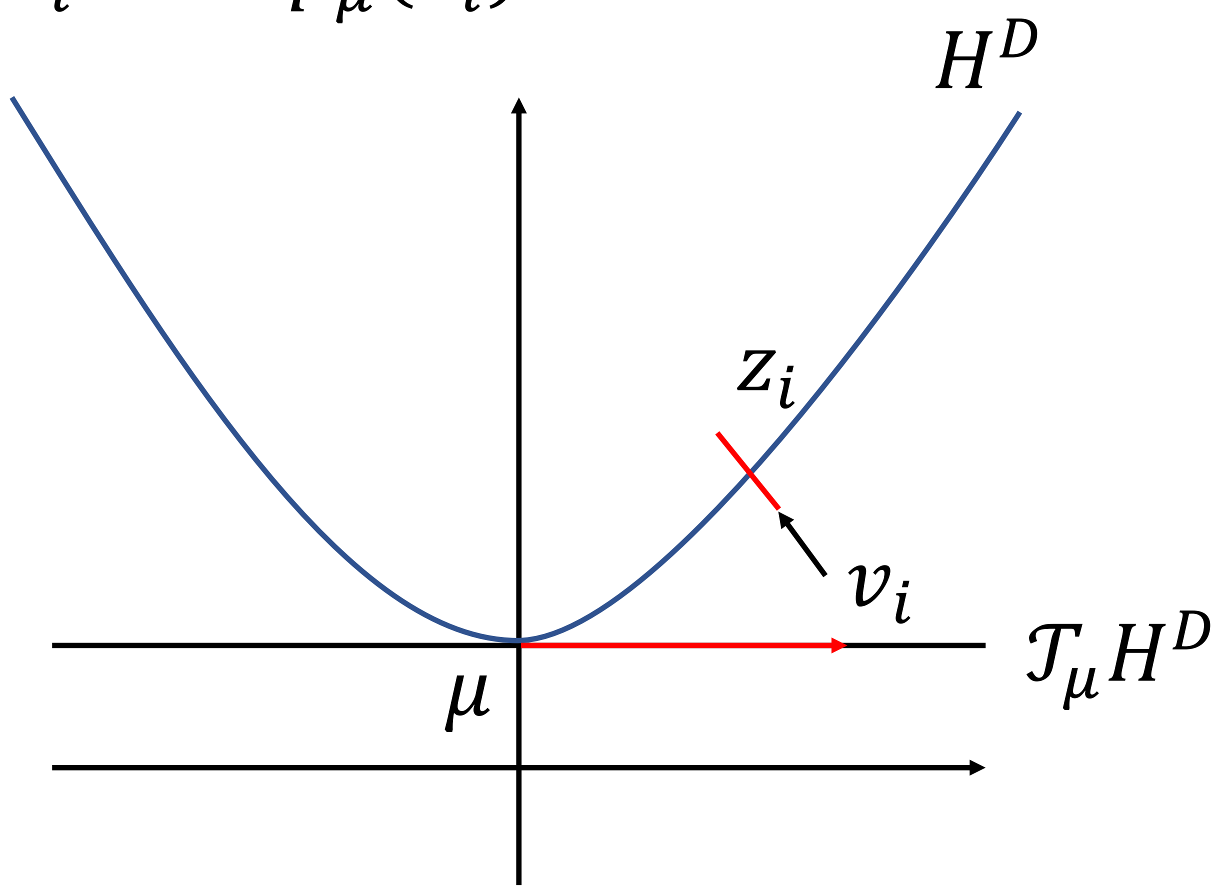
3. Dimensionality and Curvature Selection using DNML code-length

Wrapped Normal Distribution on Hyperbolic Space (Nagano+ 2019)

1. Sample a tangent vector v_i in $\mathcal{T}_\mu H^D$.



2. $z_i = Exp_\mu(v_i)$.



$$p(\boldsymbol{v}; \Sigma) := \prod_{i \in [n]} p(v_i; \Sigma),$$

$$p(v_i; \Sigma) := \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} v_i^\top \Sigma^{-1} v_i\right).$$

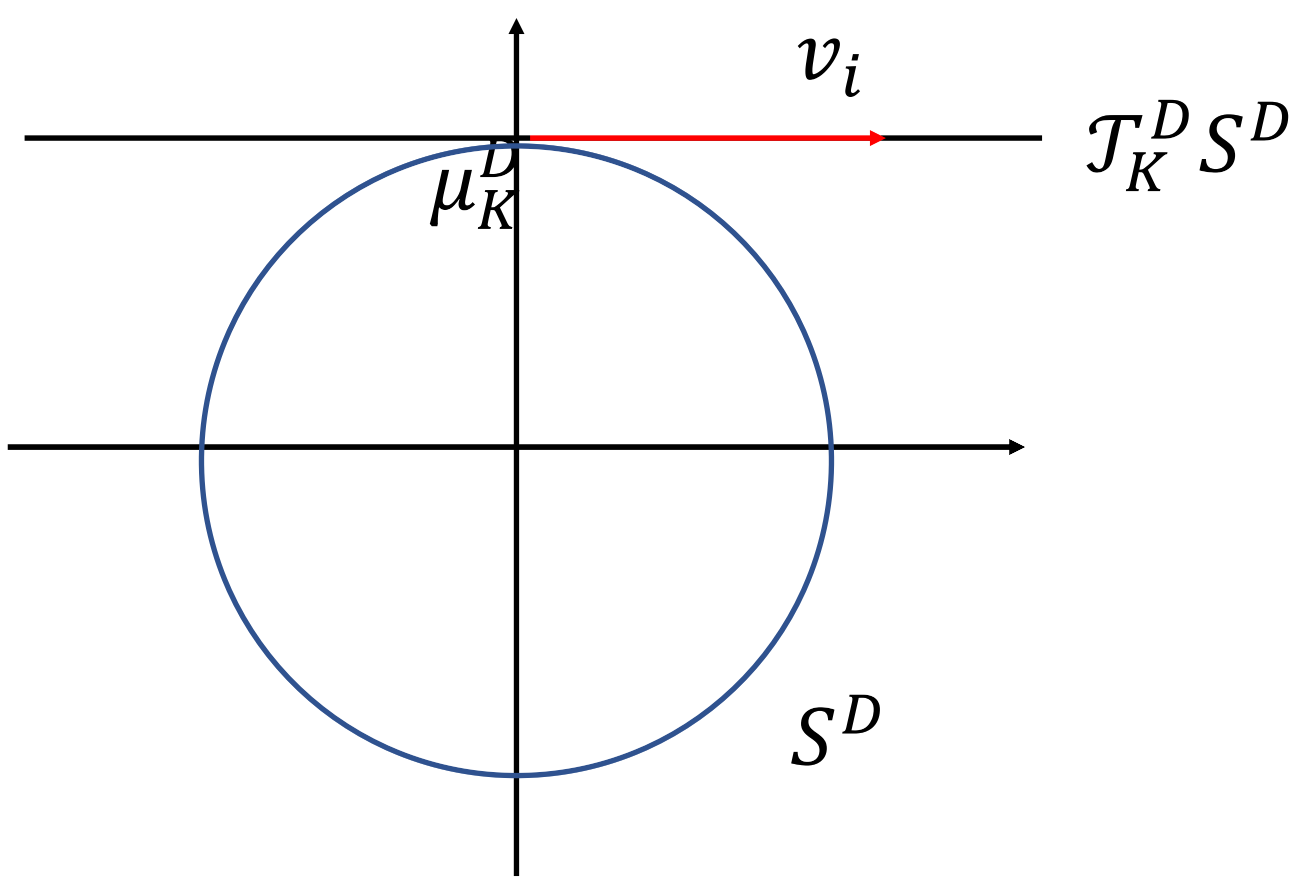
Gaussian distribution
for each tangent vector

$$p(z_i; \Sigma) := \underbrace{\frac{1}{J(z_{i,1:D} : v_i)}}_{\text{Jacobian}} \underbrace{p(v_i, \Sigma)}_{\text{Probability density of each tangent vector.}},$$

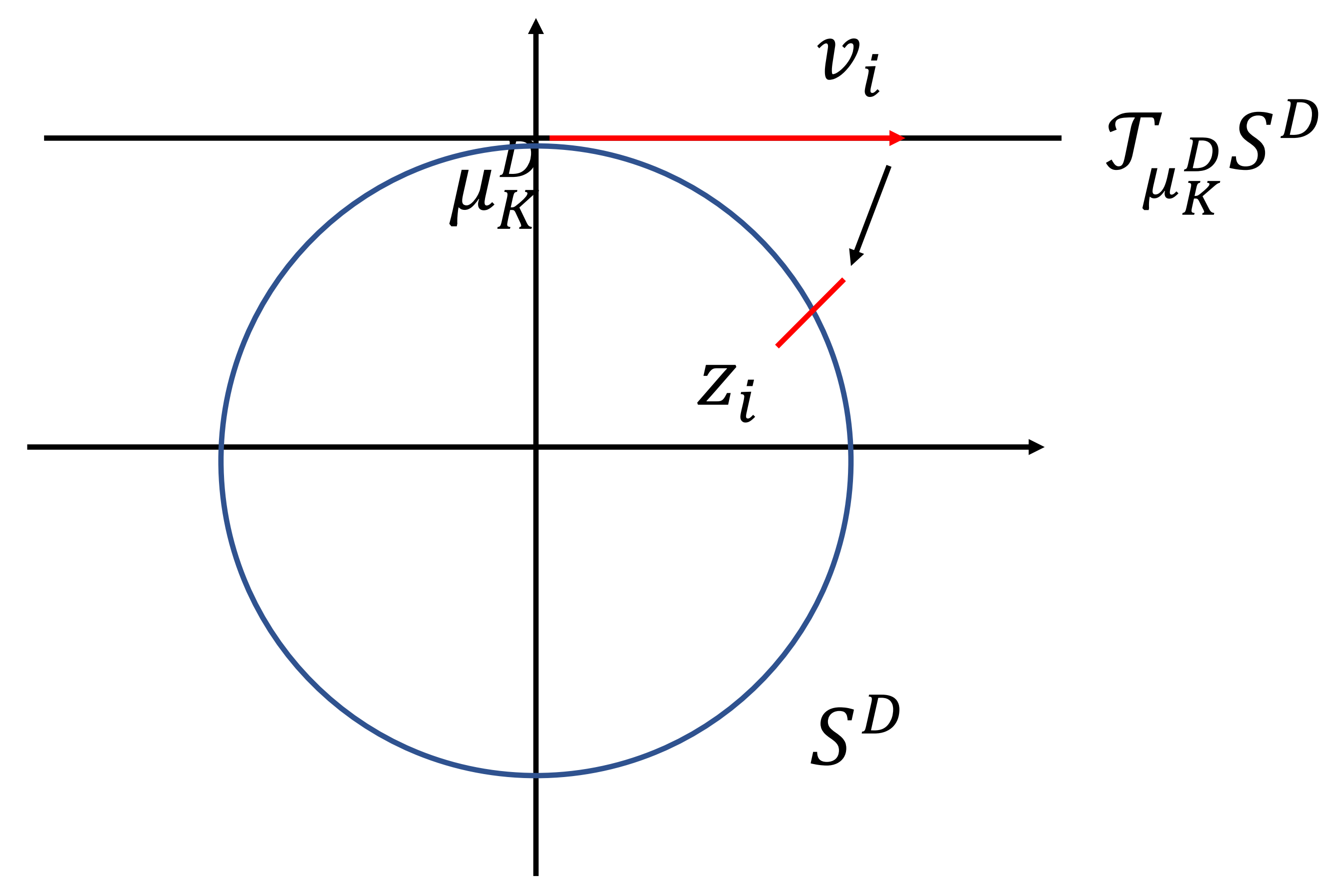
$$J(z_{i,1:D} : v_i) = \left\{ \frac{\sinh \|v_i\|_{\mathcal{L}}}{\|v_i\|_{\mathcal{L}}} \right\}^{D-1}.$$

WNDs on Spherical Space

1. Sample a tangent vector v_i in $\mathcal{T}_{\mu_K^D} S^D$.



2. $z_i = Exp_\mu^K(v_i)$.



$$p_K(v_i;\ \Sigma) := \frac{f(\tilde{v}_i;\ \Sigma)}{W_K(\Sigma)}.$$

Multivariate truncated normal dist.

$$W_K(\Sigma) := \int_{\|\tilde{v}'\|_2 \le \frac{\pi}{\sqrt{|K|}}} f(\tilde{v}';\ \Sigma) d\tilde{v}'$$

$$f(\boldsymbol{x};\ \Sigma) := \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \Sigma^{-1} \boldsymbol{x}\right).$$

$$p_K(\boldsymbol{z}_i;\ \Sigma) := \frac{1}{J_K(\boldsymbol{z}_i;\ \boldsymbol{v}_i)}\ p_K(\boldsymbol{v}_i;\ \Sigma),$$

$$J_K(\boldsymbol{z} : \boldsymbol{v}) := \begin{cases} 1 & (K = 0), \\ \left(\frac{|\sin_K(\sqrt{|K|}\|\boldsymbol{v}\|_K)|}{\sqrt{|K|}\|\boldsymbol{v}\|_K}\right)^{D-1} & (K \neq 0). \end{cases}$$

* Wrapped normal distributions for Euclidean space is standard Gaussian distributions.

2. Decomposed Normalized Maximum Likelihood (DNML) Code-Length

Use DNML code-length for LVMs.

DNML Code-Length (Yamanishi+ 2019)

Assume that observable variable y and latent variable z follow

$$p(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}, M) := p(\boldsymbol{y} \mid \boldsymbol{z}; \boldsymbol{\theta}_1, M)p(\boldsymbol{z}; \boldsymbol{\theta}_2, M),$$

Then, DNML code-length is defined as

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

where

Negative logarithm of the maximum likelihood

$$L_{\mathrm{NML}}(\boldsymbol{y} \mid \boldsymbol{z}) := \boxed{-\log p(\boldsymbol{y} \mid \boldsymbol{z}; \hat{\boldsymbol{\theta}}_1(\boldsymbol{y}, \boldsymbol{z}))} + \boxed{\log \sum_{\boldsymbol{y}'} p(\boldsymbol{y}' \mid \boldsymbol{z}; \hat{\boldsymbol{\theta}}_1(\boldsymbol{y}', \boldsymbol{z})),}$$

$$L_{\mathrm{NML}}(\boldsymbol{z}) := \boxed{-\log p(\boldsymbol{z}; \hat{\boldsymbol{\theta}}_2(\boldsymbol{z}))} + \boxed{\log \sum_{\boldsymbol{z}'} p(\boldsymbol{z}'; \hat{\boldsymbol{\theta}}_2(\boldsymbol{z}')).}$$

Penalty term (parametric complexity).

Our contribution: derived an explicit form of the approximation of each penalty term.

(Yamanishi+ 2019).

→Derive $L_{NML}(y \mid z)$ and $L_{NML}(z)$ for two priors.

1. Derivation of $L_{NML}(y \mid z)$

NML is approximated using Fisher information.
(Rissanen 1996, Grunwald et al., 2015)

1. Derivation of $L_{NML}(y \mid z)$

$L_{NML}(y \mid z)$ is approximated by

$$L_{\mathrm{NML}}(\boldsymbol{y} \mid \boldsymbol{z}, D, K) \approx -\log p(\boldsymbol{y} \mid \boldsymbol{z}; \hat{\gamma}(\boldsymbol{y}, \boldsymbol{z})) + \log \frac{n(n-1)}{4\pi} + \underline{\log \int_{\gamma_{\min}}^{\gamma_{\max}} \sqrt{|I_n(\gamma)|}\, d\gamma},$$

Integral over the parameter domain.

$$I_n(\gamma) = E_\gamma \left[-\frac{2}{n(n-1)} \underline{\frac{\partial^2 \log p(\boldsymbol{y} \mid \boldsymbol{z}; \gamma)}{\partial \gamma^2}} \right]$$

Fisher information

Likelihood

With some calculation, we have

$$I_n(\gamma) = \frac{2}{n(n-1)} \sum_{(i,j) \in \Lambda_n} \frac{1}{4} \cosh\left(\frac{d^K_{z_i z_j} - \gamma}{2} \right)^{-2}.$$

Numerical integration with the Gaussian quadrature (Vetterling+ 1992).

2. Derivation of $L_{NML}(z)$

2. Derivation of $L_{NML}(z)$

For **Euclidean and Hyperbolic** cases, $L_{NML}(z)$ is given by

$$L_{\mathrm{NML}}(\boldsymbol{z} \mid D, K) = -\log p_K(\boldsymbol{z}; \hat{\boldsymbol{\sigma}}(\boldsymbol{z})) + D \log \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{n}{2e}\right)^{\frac{n}{2}} + \sum_{i \in [D]} \log \log \frac{\sigma_i^{\max}}{\sigma_i^{\min}}.$$

The derivation mainly depends on Rissanen's g-function (Rissanen 2012).

For **spherical** case, parametric complexity of multivariate truncated normal distribution is not trivial to obtain. Thus, we use the importance sampling:

$$\begin{aligned} \int p_K(\boldsymbol{v};\ \hat{\boldsymbol{\sigma}}(\boldsymbol{v})) d\boldsymbol{v} &= \int \frac{p_K(\boldsymbol{v};\ \hat{\boldsymbol{\sigma}}(\boldsymbol{v}))}{q(\boldsymbol{v})} q(\boldsymbol{v}) d\boldsymbol{v} \\ &= E\left[\frac{p_K(\boldsymbol{v};\ \hat{\boldsymbol{\sigma}}(\boldsymbol{v}))}{q(\boldsymbol{v})}\right] \\ &\approx \sum_{\boldsymbol{v}} \frac{p_K(\boldsymbol{v};\ \hat{\boldsymbol{\sigma}}(\boldsymbol{v}))}{q(\boldsymbol{v})}, \end{aligned}$$

where $q(\boldsymbol{v})$ is a sampling distribution of $\boldsymbol{v}$.

Agenda

4. Experimental Results

Experimental Results

DNML can identify the curvature sign and dimensionality with sufficient amount of data, whereas accurate estimation of curvature is still challenging.

Table 5.1. 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 5.4. 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)

Table 5.2. Results of average estimated curvature for latent variable models with the true curvatures.

# of nodes	−1.25	−1.00	−0.75	0.10	0.20	0.30
400	−0.22	−0.21	−0.20	0.21	0.31	0.36
800	−0.18	−0.16	−0.16	0.18	0.27	0.30
1600	−0.12	−0.11	−0.10	0.13	0.20	0.21
3200	−0.09	−0.08	−0.07	0.10	0.12	0.12

→Accurate estimation of curvature is still challenging!

2. Experiment with Real-World Networks

DNML performs high conciseness with sufficient amount of nodes.

Table 5.6. Selected dimensionality and curvature of each method.

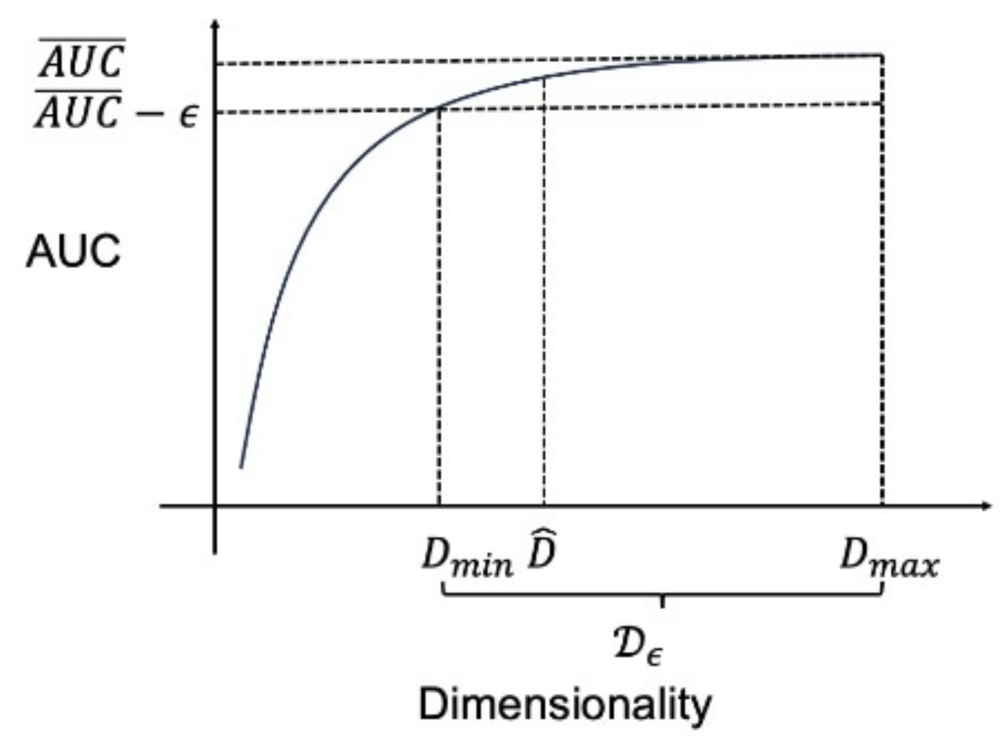
Network	DNML	AIC	BIC
AstroPh	$\mathbb{H}^{12}_{-0.25}$	$\mathbb{H}^{15}_{-0.25}$	$\mathbb{H}^{8}_{-0.27}$
HepPh	$\mathbb{E}^{16}$	$\mathbb{H}^{14}_{-0.32}$	$\mathbb{H}^{6}_{-0.36}$
Airport	$\mathbb{E}^{7}$	$\mathbb{E}^{7}$	$\mathbb{H}^{3}_{-0.73}$
WN-mammal	$\mathbb{H}^{4}_{-0.69}$	$\mathbb{H}^{4}_{-0.71}$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

Table 5.5. Statistics of real-world datsets.

Network	# Nodes	# Edges
AstroPh	18, 772	198, 080
ca-HepPh	12, 008	118, 505
Airport	3, 188	18, 630
WN-mammal	1, 180	6, 540
WN-solid	1, 232	5, 696

Table 5.7. 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



$$\text{conciseness}(\hat{D}, \epsilon_{\max}) := \frac{1}{\epsilon_{\max} P} \sum_{i=0,1,\ldots,P} c\left(\hat{D}, \frac{i}{P}\epsilon_{\max}\right),$$

$$c(\hat{D}, \epsilon) := \begin{cases} 1 - \frac{\log_2 \hat{D} - \log_2 D_{\min}}{\log_2 D_{\max} - \log_2 D_{\min}} & (\hat{D} \in \mathcal{D}_\epsilon), \\ 0 & (\hat{D} \notin \mathcal{D}_\epsilon), \end{cases}$$

Summary

Research question

- How can we determine the dimensionality and curvature of graph embedding?

Solution

- Latent variable models for graph embedding.
  - Universal latent variable models over all curvature using wrapped normal distributions.
- Apply decomposed normalized maximum likelihood (DNML) code-length to the model.

Contribution

- Derivation of the explicit formula of DNML code-length.
- Empirical validation of our proposed method.

Reference #1

[Freeman 2000] Freeman, Linton C. "Visualizing social networks." *Journal of social structure* 1.1 (2000): 4.

[Ferrer+ 2001] Ramon Ferrer I Cancho and Richard V Sol´e. The small world of human language. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1482):2261–2265, 2001

[Theocharidis+ 2009] Athanasios Theocharidis, Stjin Van Dongen, Anton J Enright, and Tom C Freeman. Network visualization and analysis of gene expression data using biolayout express 3d. Nature protocols, 4(10):1535–1550, 2009.

[Tang+2015] Tang, Jian, et al. "Line: Large-scale information network embedding." *Proceedings of the 24th international conference on world wide web*. 2015.

[Nickel and Kiela 2017] Maximillian Nickel and Douwe Kiela. Poincar´e embeddings for learning hierarchical representations. In Advances in Neural Information Processing Systems, volume 30, 59 2017.

[Nickel and Kiela 2018] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In Proceedings of the 35th International Conference on Machine Learning, volume 80, pages 3779–3788, 2018.

[Gu+ 2019] Albert Gu, Frederic Sala, Beliz Gunel, and Christopher R´e. Learning mixed-curvature representations in product spaces. In International Conference on Learning Representations, 2019.

[Rissanen 1978] Jorma Rissanen. Modeling by shortest data description. Automatica, 14(5):465–471, 1978.

[Shtarkov 1987] Yurii Mikhailovich Shtar'kov. Universal sequential coding of single messages. Problemy Peredachi Informatsii, 23(3):3–17, 1987.

[Akaike 1974] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[Schwarz 1978] Gideon Schwarz. Estimating the Dimension of a Model. The Annals of Statistics, 6(2):461 – 464, 1978.

[Yamanishi 2023] Kenji Yamanishi. Learning with the Minimum Description Length Principle. Springer Nature, 2023.

[Yin and Shen 2018] Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, page 895–906, 2018.

[Gu+ 2021] Weiwei Gu, Aditya Tandon, Yong-Yeol Ahn, and Filippo Radicchi. Principled approach to the selection of the embedding dimension of networks. Nature Communications, 12(1):1–10, 2021.

[Wang 2019] Yu Wang. Single training dimension selection for word embedding with PCA. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3597–3602, 2019.

[Almagro and Boguna 2022] Pedro Almagro, Mari´an Bogu˜n´a, and M ´Angeles Serrano. Detecting the ultra low dimensionality of real networks. Nature communications, 13(1):6096, 2022.

Reference #2

Luo+ 2019] Gongxu Luo, Jianxin Li, Hao Peng, Carl Yang, Lichao Sun, Philip S. Yu, and Lifang He. Graph entropy guided node embedding dimension selection for graph neural networks. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pages 2767–2774, 2021.

[Hung and Yamanishi 2021] Pham Thuc Hung and Kenji Yamanishi. Word2vec skip-gram dimensionality selection via sequential normalized maximum likelihood. Entropy, 23(8), 2021.

[Okuno+2018] Akifumi Okuno, Tetsuya Hada, and Hidetoshi Shimodaira. A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks. In Proceedings of the 35th International Conference on Machine Learning, pages 3888–3897. PMLR, 10–15 Jul 2018.

[Okuno+ 2019] Akifumi Okuno, Geewook Kim, and Hidetoshi Shimodaira. Graph embedding with shifted inner product similarity and its improved approximation capability. In Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, volume 89, pages 644–653. PMLR, 16–18 Apr 2019.

[Kim+ 2019] Geewook Kim, Akifumi Okuno, Kazuki Fukui, and Hidetoshi Shimodaira. Representation learning with weighted inner product for universal approximation of general similarities. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, pages 5031–5038, 7 2019.

[Prokhorenkova+ 2019] Liudmila Prokhorenkova, Egor Samosvat, and Pim van der Hoorn. Global graph curvature. In Algorithms and Models for the Web Graph, pages 16–35. Springer International Publishing, 2020.

[Krioukov+ 2010] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Mari´an Bogu˜n´a. Hyperbolic geometry of complex networks. Physics Review E, 82:036106, Sep 2010.

[Yang and Rideout 2020] Weihua Yang and David Rideout. High dimensional hyperbolic geometry of complex networks. Mathematics, 8(11):1861, 2020

[Nagano+ 2019] Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In Proceedings of the 36th International Conference on Machine Learning, volume 97, pages 4693–4702, 2019.

[Barabási 2013] Barabási, Albert-László. "The new science of networks." *Cambridge MA. Perseus* (2002).

[Yamanishi+ 2019] Kenji Yamanishi, Tianyi Wu, Shinya Sugawara, and Makoto Okada. The decomposed normalized maximum likelihood code-length criterion for selecting hierarchical latent variable models. Data Mining and Knowledge Discovery, 33(4):1017–1058, 2019.

[Vetterling+ 1992] William T Vetterling, William T Vetterling, William H Press, William H Press, Saul A Teukolsky, Brian P Flannery, and Brian P Flannery. Numerical recipes: example book c. 1992.

[Rissanen 2012] Jorma Rissanen. Optimal estimation of parameters. 2012.

[Mathai 1997] A M Mathai. Jacobians of Matrix Transformations and Functions of Matrix Arguments. WORLD SCIENTIFIC, 1997.

[Bonnabel 2013] Silv`ere Bonnabel. Stochastic gradient descent on riemannian manifolds. IEEE Transactions on Automatic Control, 58(9):2217–2229, 2013.

[Fellbaum 1998] Christiane Fellbaum. WordNet: An Electronic Lexical Database. The MIT Press, 1998.

[Ebisu and Ichise 2018] Ebisu, Takuma, and Ryutaro Ichise. "Toruse: Knowledge graph embedding on a lie group." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.

Reference #3

[Chen+ 2021] Chen, Yao, et al. "MöbiusE: Knowledge graph embedding on möbius ring." *Knowledge-Based Systems* 227 (2021): 107181.

[Yuan and Lin 2006] Yuan, Ming, and Yi Lin. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68.1 (2006): 49-67.

[Grünwald 2007] Peter D Grunwald. The minimum description length principle. MIT press, 2007.

[Miyaguchi and Yamanishi 2018] Kohei Miyaguchi and Kenji Yamanishi. High-dimensional penalty selection via minimum description length principle. Machine Learning, 107:1283–1302, 2018.

[Grünwald and Mehta 2019] Peter D Gr¨unwald and Nishant A Mehta. A tight excess risk bound via a unified pac-bayesian–rademacher–shtarkov–mdl complexity. In Algorithmic Learning Theory, pages 433–465. PMLR, 2019.

[Reddi+ 2016] Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. Advances in neural information processing systems, 29, 2016.

[Mosci+ 2010] Sofia Mosci, Lorenzo Rosasco, Matteo Santoro, Alessandro Verri, and Silvia Villa. Solving structured sparsity regularization with proximal methods. In Machine Learning and Knowledge Discovery in Databases: European Conference, pages 418–433. Springer, 2010.

[Xiao+ 2019] Han Xiao, Minlie Huang, and Xiaoyan Zhu. Transg: A generative model for knowledge graph embedding. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2316–2325, 2016.