数理統計学チーム
Mathematical Statistics Team

下平英寿
Hidetoshi Shimodaira

数理統計学チーム@京大クラスタ

下平研究室（情報学研究科，京都大学）

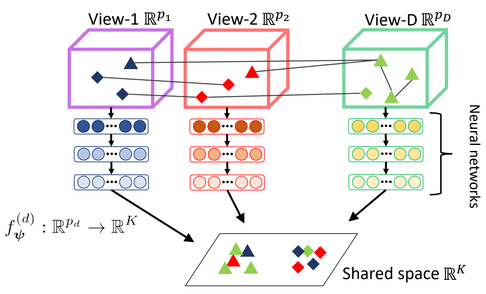
下平英寿（チームリーダー），Thong Pham（特別研究員），福井一輝（パートタイマー），奥野彰文（パートタイマー），井上雅章（パートタイマー），KIM GEEWOOK（パートタイマー），田中卓磨（パートタイマー），岩田具治（客員研究員），寺田吉壱（客員研究員），伊森晋平（客員研究員），廣瀬慧（客員研究員），白石友一（客員研究員），劉言（客員研究員）

統計学と機械学習

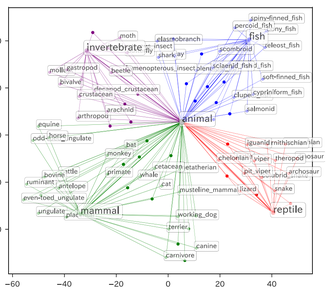
- 統計的仮説検定，領域の検定，選択的推測（selective inference）の理論
- マルチスケール・ブートストラップ法などリサンプリング法
- 汎化誤差を共変量シフト，不完全観測などの設定で推定する情報量規準
- マルチドメインデータの多変量解析，ニューラルネットによるグラフ埋め込み
- 単語埋め込みなど自然言語処理，画像と単語の埋め込みと相互検索
- 成長ネットワークの統計的推測
- 系統樹，遺伝子発現データ，ゲノムデータの解析

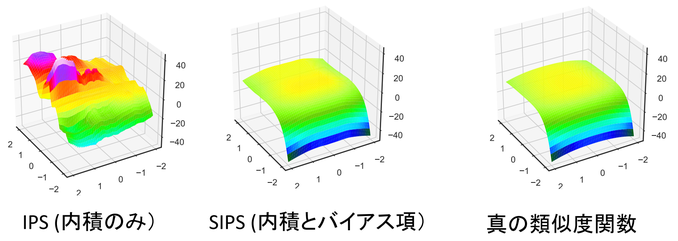
高い表現力をもつ類似度関数の学習理論

- Okuno, Hada, Shimodaira (**ICML 2018**) A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks
- 内積だけでも十分に大きなニューラルネットワークを併用することで広いクラスの類似度関数を学習できることを数学的に証明（Universal Approximation定理とMercer定理，推定量の漸近的一致性の証明）
- グラフ埋め込みでの学習アルゴリズム（mini-batch SGD)
- 画像，単語などマルチドメインの多変量解析への拡張
- 自然言語処理（単語埋め込み），画像検索などの応用

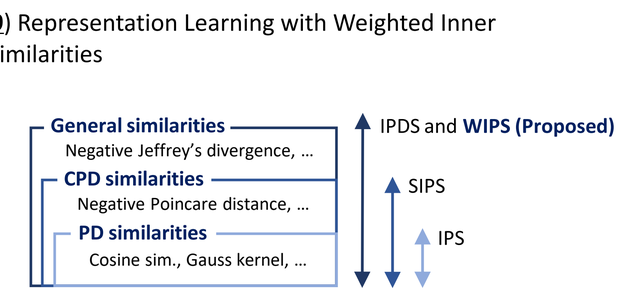


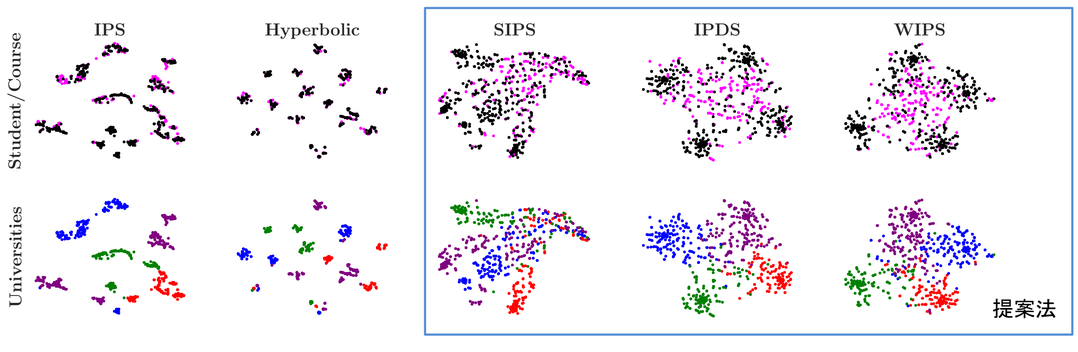
- Okuno, Shimodaira (**ICML workshop Theoretical Foundations and Applications of Deep Generative Models 2018**) On representation power of neural network-based graph embedding and beyond
- Okuno, Kim, Shimodaira (to appear **AISTATS 2019**) Graph Embedding with Shifted Inner Product Similarity and Its Improved Approximation Capability
- 内積類似度 (IPS) が表現できるのは 正定値関数（positive definite, PD)だけである (例) コサイン類似度
- 内積だけでなくバイアス項も学習するモデルShifted Inner Product Similarity (SIPS)の提案
- SIPSは任意の条件付き正定値関数 (conditionally PD, CPD)を学習できることを数学的に証明
- CPDは十分に広いクラスである (例) ユークリッド空間の距離，Poincare埋め込み，双曲空間の距離，（一部の）Wasserstein 距離など，機械学習でよく利用される距離がたいてい表現できる
- さらに，2つの内積の差を学習するモデル Inner Product Difference Similarity (IPDS)も提案
- IPDSは不定値 (indefinite)カーネルを含む類似度関数を学習できることを数学的に証明



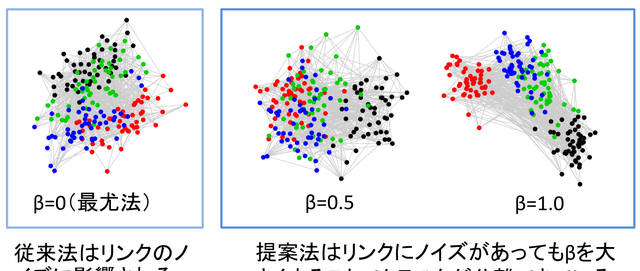


- Kim, Okuno, Fukui, Shimodaira (**arXiv:1902.10409**) Representation Learning with Weighted Inner Product for Universal Approximation of General Similarities
- 重み付き内積を学習するモデル Weighted Inner Product Similarity (WIPS)の提案と高速でスケーラブルな学習法の実装
- 重みの値として負も許して学習することにより，不定値 (indefinite)カーネルを含む類似度関数を学習できることを数学的および実験で証明





- Okuno, Shimodaira (to appear **AISTATS 2019**) Robust Graph Embedding with Noisy Link Weights
- 類似度関数学習やグラフ埋め込みがノイズに強くなるようにロバスト化する手法の提案
- 提案手法 (β-グラフ埋め込み)は新規に考案した損失関数 (経験積率 β-スコア）の最小化を実行
- ロバスト統計学におけるdensity power divergenceの理論を発展させたもの

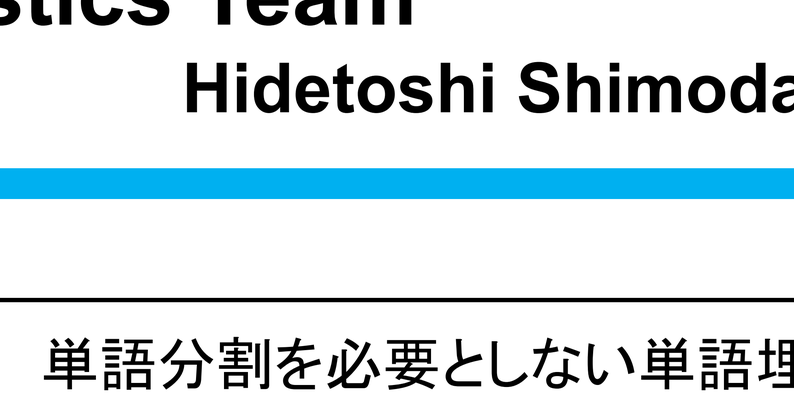


従来法はリンクのノイズに影響される

提案法はリンクにノイズがあってもβを大きくすることでクラスタが分離できている

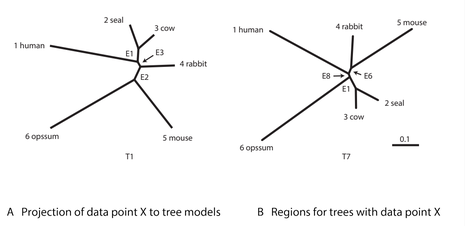
単語分割を必要としない単語埋め込み

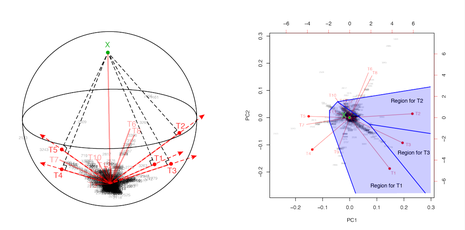
- Kim, Fukui, Shimodaira (**EMNLP Workshop on Noisy User-generated Text W-NUT 2018**), Word-like character *n*-gram embedding
- Kim, Fukui, Shimodaira (to appear **NAACL-HLT 2019**), Segmentation-free compositional *n*-gram embedding
- 既存の単語埋め込み法では，文書データを前処理によって分割して，単語の列を生成する必要があった．英語などスペースで区切られた言語では問題ないが，日本語，中国語，韓国語では形態素解析による前処理が必要であり，これが障害となることがあった
- とくにソーシャルメディアでは表記ゆれ，新語の問題がある
- そこで「単語分割を全く行わない単語埋め込み法」の提案
- 単語，フレーズ，文に対して連続的に適用できる
- コーパスのすべての部分文字列に対して埋め込みを効率的に計算

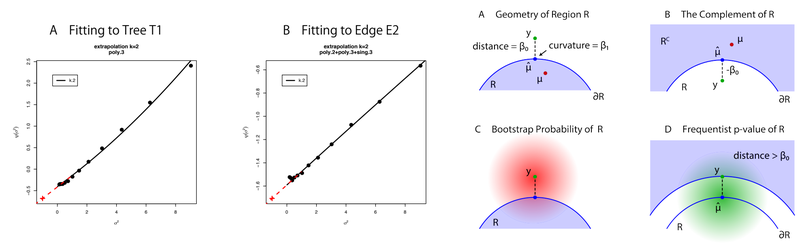


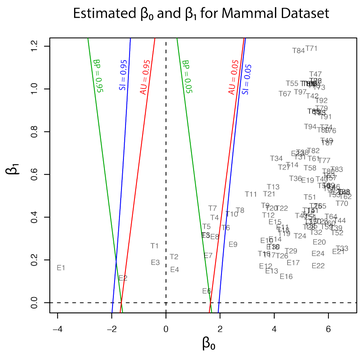
選択的推測の理論とその系統樹推定への応用

- Shimodaira, Terada (**arXiv:1902.04964**), Selective Inference for Testing Trees and Edges in Phylogenetics
- 従来の統計的仮説検定では，事前に仮説を定める必要がある．ところが医学，科学全般で必ずしもこれが守られず，実際にはデータを見てから仮説を設定し，その同じデータを再び用いて仮説検定を実行するため，偽の発見となりやすくその危険性が指摘されてきた．頻度論のp値の代わりにベイズ事後確率を用いてもこの問題は解決しない
- 「データを見てから仮説を選択すること」を設定に組み込んだ選択的推測 (selective inference) の理論が構築されつつある
- 弊チームの先行研究 (Terada, Shimodaira **arXiv:1711.00949**)ではマルチスケール・ブートストラップ法 (Shimodaira 2002, 2004, 2008)によってブートストラップ確率のスケーリング則から選択的推測のp-値を計算する数理統計理論を与えている
- 本研究ではその手法をもとに分子進化系統樹を推定する問題へ実際に応用し，クラスタリングや系統樹のクレードのp-値において選択的推測による調整の重要性を示した



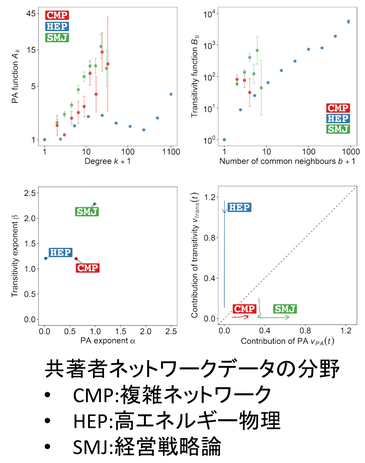






複雑ネットワーク成長モデルの統計推測

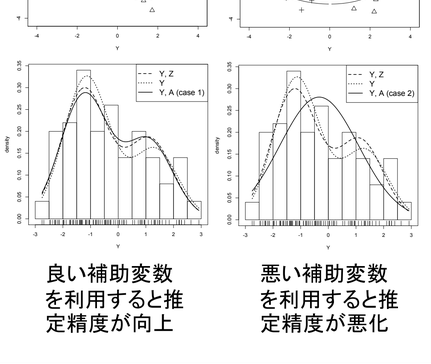
- Pham, Sheridan, Shimodaira (to appear **Journal of Statistical Software 2019**), PAFit: an R Package for Estimating Preferential Attachment and Node Fitness in Temporal Complex Networks
- Inoue, Pham, Shimodaira (**Unifying Themes in Complex Systems IX. ICCS 2018. Springer Proceedings in Complexity. Springer**), Transitivity vs Preferential Attachment: Determining the Driving Force Behind the Evolution of Scientific Co-Authorship Networks
- 複雑ネットワークの成長モデルをネットワーク時系列データから統計推測する手法とソフトエア(PAFit)の開発
- 優先的選択（知名度）と適応度（能力）の効果または優先的選択（知名度）と推移度（共通の友人)の効果のノンパラメトリック同時推定
- 共通する論文共著者が一人でもいれば，次に共著する確率は急増



共著者ネットワークデータの分野
- CMP:複雑ネットワーク
- HEP:高エネルギー物理
- SMJ:経営戦略論

補助変数選択の情報量規準

- Imori, Shimodaira (**arXiv:1902.07954**), An information criterion for auxiliary variable selection in incomplete data analysis
- 主要変数に加えて学習データで補助変数が観測できるとき，補助変数も利用したほうが主要変数のモデル推定精度が高くなる事がある
- 主要変数に無関係な補助変数はノイズとなり有害
- 主要変数の一部は観測できない
- このような設定で汎化誤差の推定量である情報量規準を初めて導出
- 補助変数や主要変数の選択，もっと一般にモデル選択を行える
- 既存の赤池情報量規準（AIC）などを特殊な場合として含む一般化



良い補助変数を利用すると推定精度が向上

悪い補助変数を利用すると推定精度が悪化