Mathematical Statistics Team
Hidetoshi Shimodaira
数理統計学チーム　下平 英寿

RIKEN Center for Advanced Intelligence Project

数理統計学チーム@京大クラスタ

下平・本多研究室（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），井上雅章（大学院生リサーチ・アソシエイト），土屋平（パートタイマー），Cao Ruixing（パートタイマー），内藤雅博（パートタイマー），山際宏明（パートタイマー），木原健太（パートタイマー），大山百々勢（パートタイマー），橋本竜馬（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），横井 祥（客員研究員），奥野彰文（客員研究員），本多淳也（客員研究員）

概要・研究の意義

統計的推測の方法論の研究を行う．主なテーマは以下の通り．(1) **帰納的推論の原理の探求**：とくに多数の仮説について統計推測するときの信頼度計算の方法論．多重検定，多重比較，モデル選択，選択的推測などの手法を開発する理論研究．系統樹推定や発現解析など生命科学における標準手法になっている．今後，AIによる「発見」（仮説生成と検証）を実用化するときの基盤技術になる．(2) **情報統合の多変量解析と適切な情報表現の探求**：とくにマルチモーダル関連性データのグラフ埋め込みによる表現学習．画像認識や自然言語処理での実践や擬ユークリッド空間への埋め込みや加法構成性の理論など．今後，AIによる「高度な思考」を実装するときの基盤技術になる．

これまでの代表的な研究成果

・**マルチスケール・ブートストラップ法による信頼度計算**：複雑な機械学習による予測や推定値の信頼区間や仮説のp値でも使える汎用手法を弊研究室で開発し，遺伝子発現解析等の標準手法となっている（4論文の被引用数＞11000，50近くの国際特許でも利用）．しかし選択バイアスの問題があったが，選択的推測として理論的にほぼ解決に成功した．

・**グラフ埋め込みによる次元削減**：マルチモーダルデータの関連性を深層学習によるグラフ埋め込みとして定式化し，従来の多変量解析などを一般化．表現可能な類似度の関数クラスを明らかにし，従来の内積よりも擬ユークリッド空間，双極空間への埋め込みは表現力を飛躍的に拡大し，実装も容易であることを証明した．

目指すゴール，今後の展開

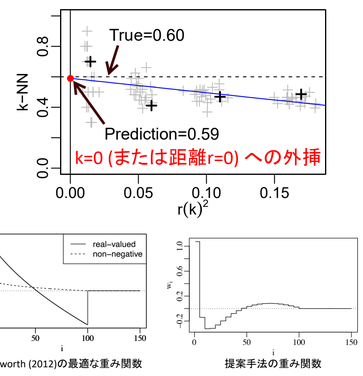
・頻度論，ベイズ，情報論的方法など従来の統計学・機械学習で帰納的推論の方法論が議論されていたが，「p値の誤用」など様々な問題が指摘されている．AIに限らずあらゆる分野で重要な「データからの推論」のより良い原理を探求する．とくに選択的推論の手法開発と応用を行う．

・単語ベクトルで確認されている意味の演算（king − man + woman = queen）など，「構成性」に関連してNeurIPS（2019/12）でも注目されているが，このための理論にとりくみ，高度な思考を実現するステップとしたい．

k-近傍法 (k-NN) のバイアスをゼロにする

- Okuno, Shimodaira (**NeurIPS 2020**), Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate
- 操，田中，奥野，下平（人工知能学会全国大会 2021）マルチスケール k-近傍法における回帰関数および損失関数の検討
- 田中，奥野，下平（人工知能学会全国大会 2021）マルチスケールk-近傍法による画像の Extreme Multi-Label 分類

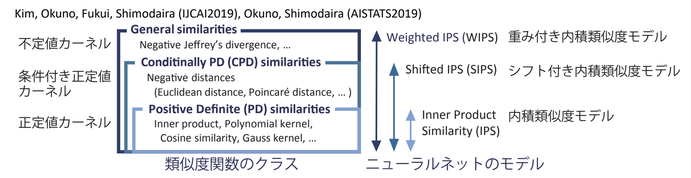
- k-NNは最も単純な判別，分類の手法．特徴量をみて最も近い過去事例をk個選び，そのラベルの平均値を出力する
- 小さいkはバイアスが小さく，大きいkは分散が小さくなる．両者の影響は分類誤差でトレードオフの関係にある
- 理論上はk=0とすればバイアスがゼロになる．そこで複数のkの値でk-NNを実行し，それをk=0へ外挿する．分散をおさえつつ，現実には存在しない架空の「0-NN」を計算する「マルチスケール k-近傍法」を提案
- Flickrの画像からタグ推定で有効性を確認
- 提案手法の収束レートは理論的な最適レートを達成



単語埋め込みの加法構成性と論理演算

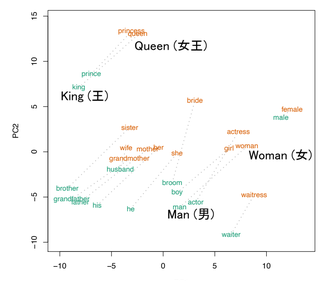
- Kim, 横井，下平，単語埋め込みの二種類の加法構成性，言語処理学会 第26回年次大会（NLP2020）
- 横井，下平，単語埋め込みの確率的等方化，言語処理学会 第27回年次大会（NLP2021）
- 内藤，横井，下平，単語埋め込みによる論理演算，言語処理学会 第27回年次大会（NLP2021）
- Naito, Yokoi, Kim, Shimodaira, Revisiting Additive Compositionality: AND, OR and NOT Operations with Word Embeddings, ACL-IJCNLP 2021 Student Research Workshop.
- 内藤，横井，下平，単語埋め込みの加法構成性の精緻化と論理演算，2021年度統計関連学会連合大会

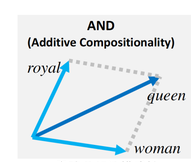
- 加法構成性の観点から表現学習の理論研究．たとえば自然言語処理の単語埋め込みにおいて，Queen = King − Man + Womanの演算を分散表現におけるベクトルの加減算として実現するような特徴空間に関する基礎研究を行っている．パターン認識を超えた「思考」を実現する知能をめざす．
- 成果1．シンプルな重み付き内積の学習によって擬ユークリッド空間の埋め込みが実現できることを理論と実験で示した．たとえば階層構造を学習できると言われる双曲空間の埋め込みもこれに含まれる．
- 成果2．いわゆる加法構成性（概念ベクトルの和によって意味が構成されること）はANDに相当しており，Queen = King − Man + Womanのようなアナロジー計算が単語ベクトルの加減算を用いて実現できることが既存研究で知られている．これを発展させるため，ORに相当する演算や，さらに集合のしぼりこみに相当する演算を導出した．これらからNOTに相当する演算（つまり部分集合の補集合）を導出した．いずれもシンプルなベクトルの演算として実現できる．



ベクトル加減算でアナロジー計算ができる （既存研究でよく知られた結果）

$$\boldsymbol{v}_{queen} \approx \boldsymbol{v}_{king} - \boldsymbol{v}_{man} + \boldsymbol{v}_{woman}$$

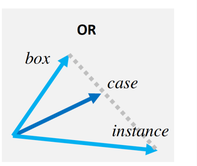




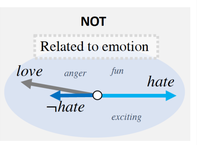
$$\boldsymbol{v}_{queen} \approx \boldsymbol{v}_{royal} + \boldsymbol{v}_{woman}$$
$$\boldsymbol{v}_{king} \approx \boldsymbol{v}_{royal} + \boldsymbol{v}_{man}$$

$$\forall c \in V, \quad p(w|c) = p(w_1|c) \cdots p(w_s|c) \qquad \boldsymbol{v}_w = \boldsymbol{v}_{w_1} + \cdots + \boldsymbol{v}_{w_s}$$

$$\forall c \in V, \quad p(w|c) = p(w_1|c) + \cdots + p(w_s|c)$$
$$\boldsymbol{v}_w \approx \sum_{i=1}^{s} \frac{p(w_i)}{p(w)} \boldsymbol{v}_{w_i}$$ ORは頻度重みの重心



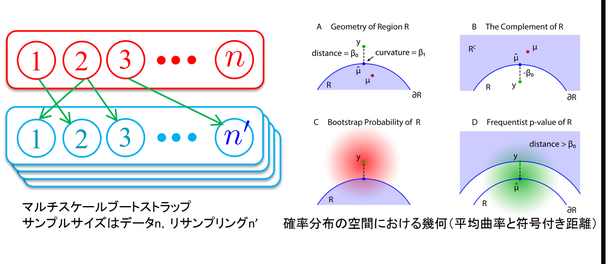
$$p(W = w \mid W \in A, c) = p(W = w \mid W \in A) \exp(\boldsymbol{v}_{w|A}^{\top} \boldsymbol{u}_c)$$
$$\boldsymbol{v}_{w|A} \approx \boldsymbol{v}_w - \boldsymbol{v}_A \qquad v_A = \sum_{w \in A} \frac{p(w)}{p(A)} \boldsymbol{v}_w$$
部分集合への限定操作は，その集合の重心を使った中心化



$$\boldsymbol{v}_{\neg w|A} \approx \boldsymbol{v}_{A \setminus \{w\}} - \boldsymbol{v}_A \qquad \text{NOTは補集合の計算}$$
$$= -\frac{p(W = w \mid W \in A)}{1 - p(W = w \mid W \in A)} \boldsymbol{v}_{w|A}$$

マルチスケール・ブートストラップと選択的推測

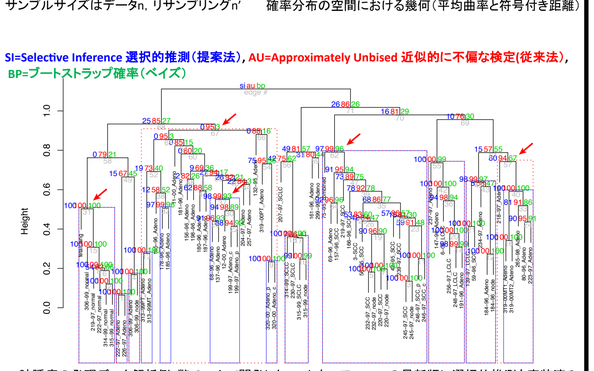
- 従来の統計的仮説検定では，事前に仮説を定める必要がある．ところが医学，科学全般で必ずしもこれが守られず，実際にはデータを見てから仮説を設定し，その同じデータを再び用いて仮説検定を実行するため，偽の発見となりやすくその危険性が指摘されてきた．頻度論のp値の代わりにベイズ事後確率を用いてもこの問題は解決しない
- 「データを見てから仮説を選択すること」を設定に組み込んだ選択的推測（selective inference），選択後の推測（post selection inference）の統計手法や理論が近年，構築されつつある
- 弊チームの先行研究（Terada, Shimodaira **arXiv:1711.00949v2**）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．確率分布の空間における仮説領域の曲率や距離といった幾何学にもとづいてアルゴリズムが構成される
- この手法をもとに発現解析や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 Shimodaira, Terada (**Frontiers in Ecology and Evolution 2019**).
- さまざまな特徴量から判別や予測を行うとき，すべての特徴量を使わずに，Lassoなどの手法で有用な特徴量だけを選択するが，選択した特徴量の係数について信頼区間を計算すると，係数＝0という帰無仮説が有意に棄却されやすくなるバイアスがある．これは選択バイアスの影響であり，本来は重要でない特徴量にもかかわらず，誤って重要と判断されてしまう（偽陽性）
- 従来法よりLassoにおける検出力が向上し，本来は重要な特徴量が誤って重要でないと判断されてしまう（偽陰性）をへらすことができた．また，従来法では困難であった非凸の正則化項（SCAD，MCP）に初めて適用できた Terada, Shimodaira (**arXiv:1905.10573v3**).



マルチスケールブートストラップ
サンプルサイズはデータn，リサンプリングn′

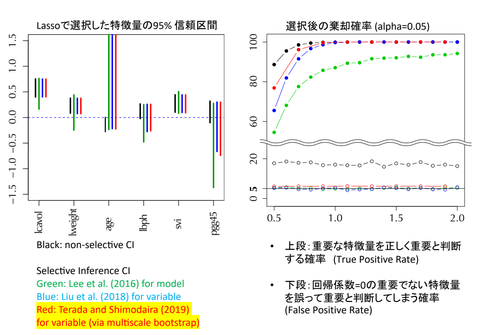
確率分布の空間における幾何（平均曲率と符号付き距離）

SI=Selective Inference 選択的推測（提案法），AU=Approximately Unbised 近似的に不偏な検定(従来法)，BP=ブートストラップ確率（ベイズ）

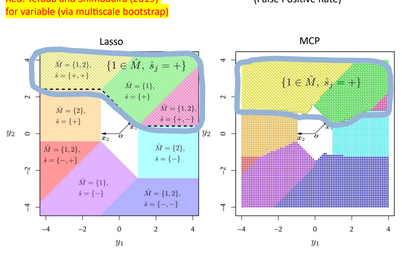


肺腫瘍の発現データ解析例．弊チームで開発したソフトウエアpvclustの最新版に選択的推測も実装済み

Distance: correlation
Cluster method: average



- 上段：重要な特徴量を正しく重要と判断する確率 (True Positive Rate)
- 下段：回帰係数=0の重要でない特徴量を誤って重要と判断してしまう確率 (False Positive Rate)



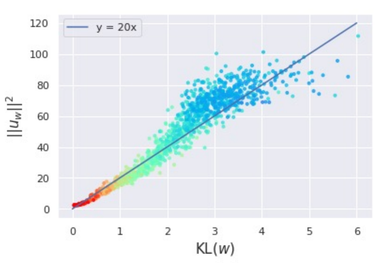
LassoとMCPの特徴量空間における選択領域のイメージ図

単語ベクトルの長さは意味の強さを表す

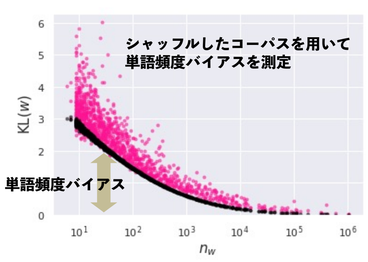
- 大山，横井，下平，単語ベクトルの長さは意味の強さを表す，言語処理学会 第28回年次大会（NLP2022）
- 単語の「意味の強さ」をKL情報量を用いて定式化した．単語の周辺単語の頻度分布は単語の意味を表す（分布仮説）が，周辺単語分布がユニグラム分布からどれだけ異なるかをKL情報量で測定したものが意味の強さを表すという定式化である．このKL情報量は単語ベクトルの長さの2乗に相当することを実験と理論で示した．

$$\mathrm{KL}\big(p(\cdot\,|w) \,\|\, p(\cdot)\big) = \sum_{w' \in V} p(w'|w) \log \frac{p(w'|w)}{p(w)}$$

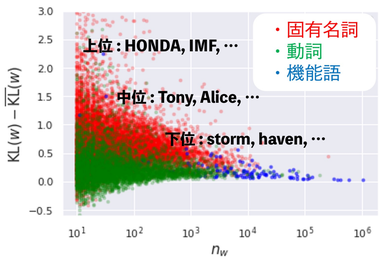
$p(\cdot\,|w)$ 単語 w の周辺単語分布　　$p(\cdot)$ ユニグラム分布



単語ベクトルの長さの2乗はKL情報量と強い相関がある



KL情報量には，単語出現頻度によるバイアスがある



KL情報量をバイアス補正したもの

バンディット問題と強化学習

- Hayashi, Honda, Kashima, Bayesian optimization with partially specified queries, Machine Learning（2022）
- Matsuura, Honda, El Hanafi, Sozu, Sakamaki, Optimal adaptive allocation using deep reinforcement learning in a dose-response study, Statistics in Medicine（2022）

- クラウドソーシングのワーカーのように入力の特徴量を完全には指定できず一部が確率的に決定されるような新たなベイズ最適化の枠組みを提案した．また，この設定においてトンプソン抽出を一般化したアルゴリズムを提案し，その性能保証を与えた．
- 新薬の用量決定に関する治験において強化学習を用いた動的な患者割り振り方策を構成した．本研究で扱った第2相試験はバンディット問題としては純粋探索問題に対応し，これは累積報酬最大化に比べて強化学習アルゴリズムによる方策学習が難しい設定であったが，治験の分野で知られていた推定手法を適切に組み込んだ学習を行うことで従来の漸近論に基づく方策に比べて性能を大きく改善した．