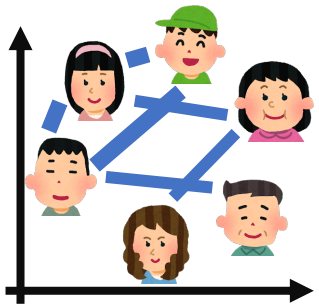
京都大学情報学研究科システム科学専攻
数理システム論分野（シー５）
下平研究室

研究トピックの紹介

- **グラフ埋め込みの理論**
- **グラフ埋め込みの応用**
- **複雑ネットワーク**
- **ブートストラップと選択的推測**
- **高次元小標本の統計学**

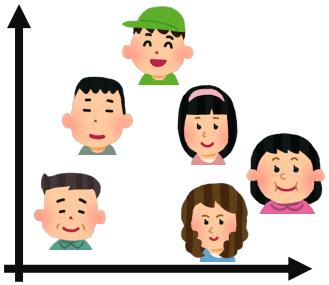
座標とリンク

（リンクだけでも良い）



ニューラルネット

座標

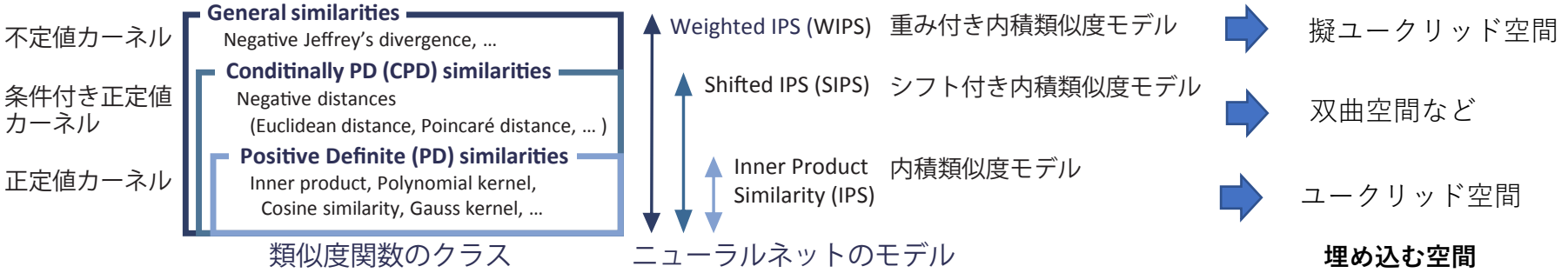


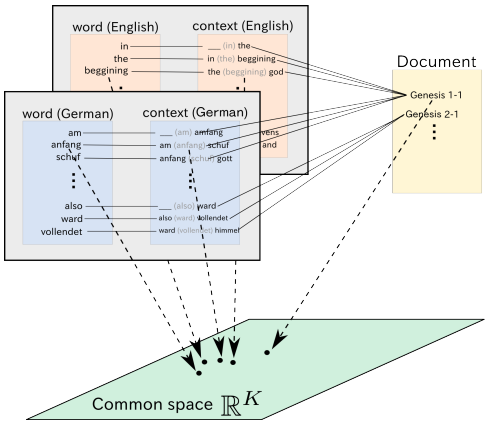
グラフのリンク構造をなるべく保存するようにノードの座標を計算する．統計学の多変量解析の拡張になっていて，機械学習で応用がたくさんある．

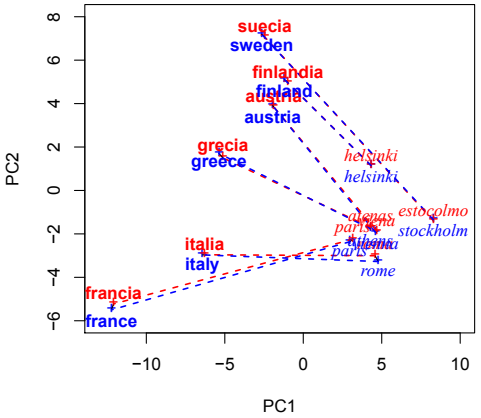
- 一般にノードの類似度をベクトルの内積やコサイン類似度で測る
- 類似度関数を変えたら，もっと性能あがるのでは？

- 内積だけでも任意の正定値カーネルが表現できる！
- シフト項をいれるだけで，双曲空間への埋め込みが表現できる！
- 重み付き内積モデルにすれば，もっと表現できる！

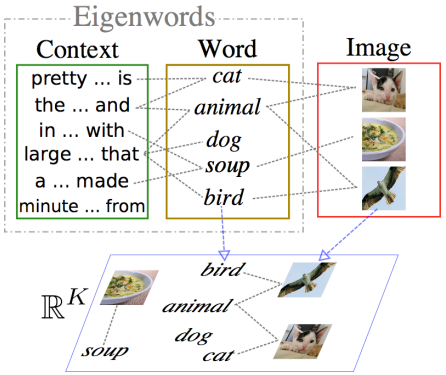
エッシャーの絵は双曲幾何の例
（ポアンカレ埋め込み）







- 対訳コーパスのリンクだけから単語埋め込み
- 辞書を教えなくても，２言語の翻訳単語の座標は近くなる



- 文書データ，ソーシャルメディアの写真とタグから，単語と画像を同時埋め込み
- 画像と単語の相互検索や，意味の加減算による検索ができる

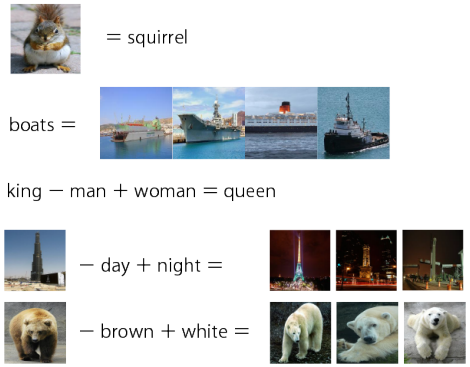


図3. Flickrデータの可視化

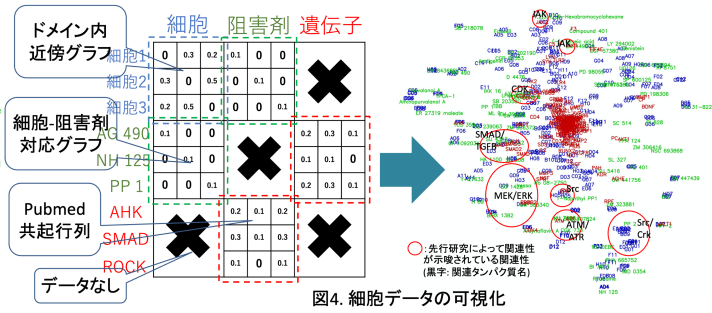
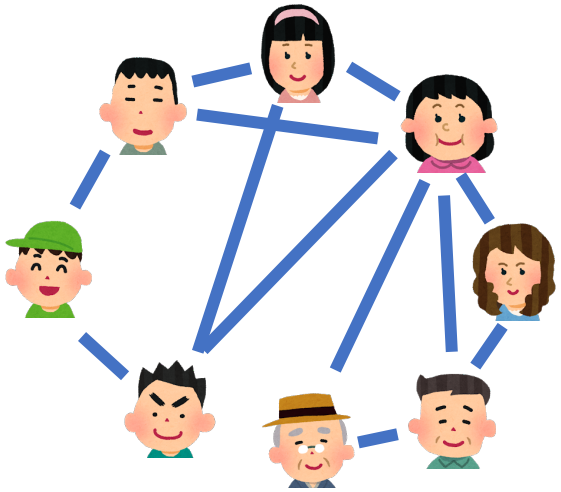


図4. 細胞データの可視化

- 文書，写真，遺伝子データなど混在した関連性データを扱える
- t-SNEなどの可視化を拡張



- 成長するネットワーク（グラフ）の確率モデル
- 新しいリンクの確率が高いのは？

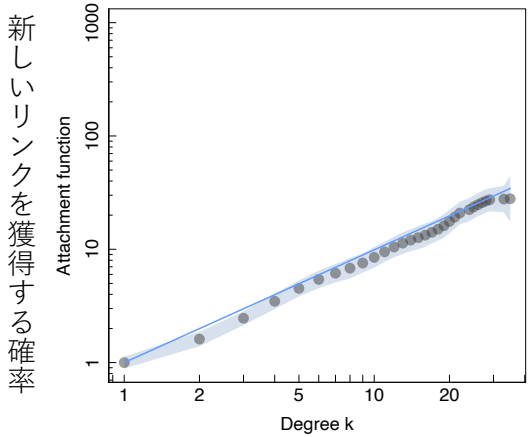
1. リンク数の多い人

2. ポテンシャルの高い人

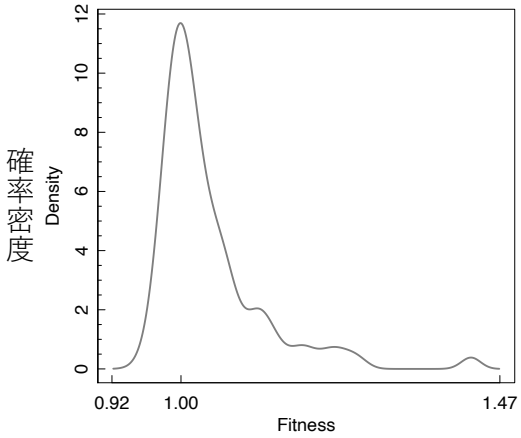
（実は新人なのにすでに２人）

3. 共通する友達の多いペア

これらの効果を分離して推定する統計手法とソフトウエア(PAFit, FoFaF)の開発



現在のリンク数



新しいリンクを獲得する潜在的能力

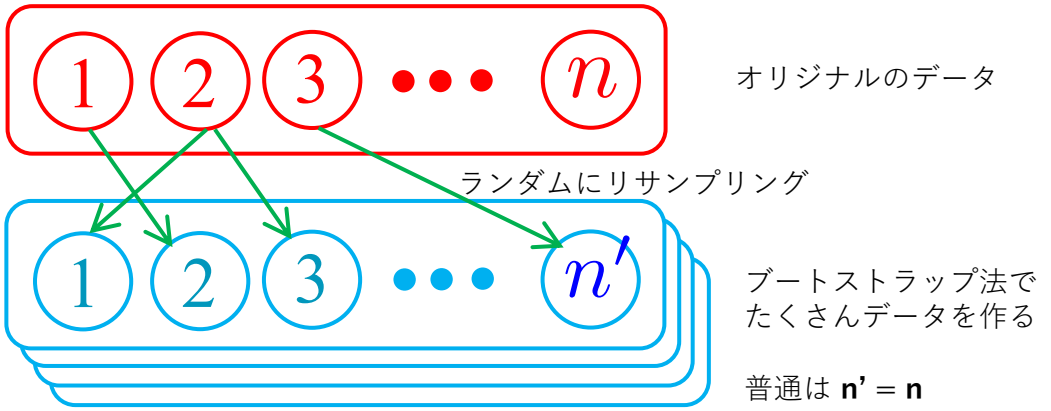
Rank	Estimated fitness	Name
1	1.42	BARABASI, A
2	1.35	NEWMAN, M
3	1.26	JEONG, H
4	1.25	LATORA, V
5	1.24	ALON, U
6	1.23	OLTVAI, Z
7	1.23	YOUNG, M
8	1.22	WANG, B
9	1.21	SOLE, R
10	1.21	BOCCALETTI, S

「コラボする潜在能力」の高い研究者トップ10

Pham, Sheridan, Shimodaira (2020)

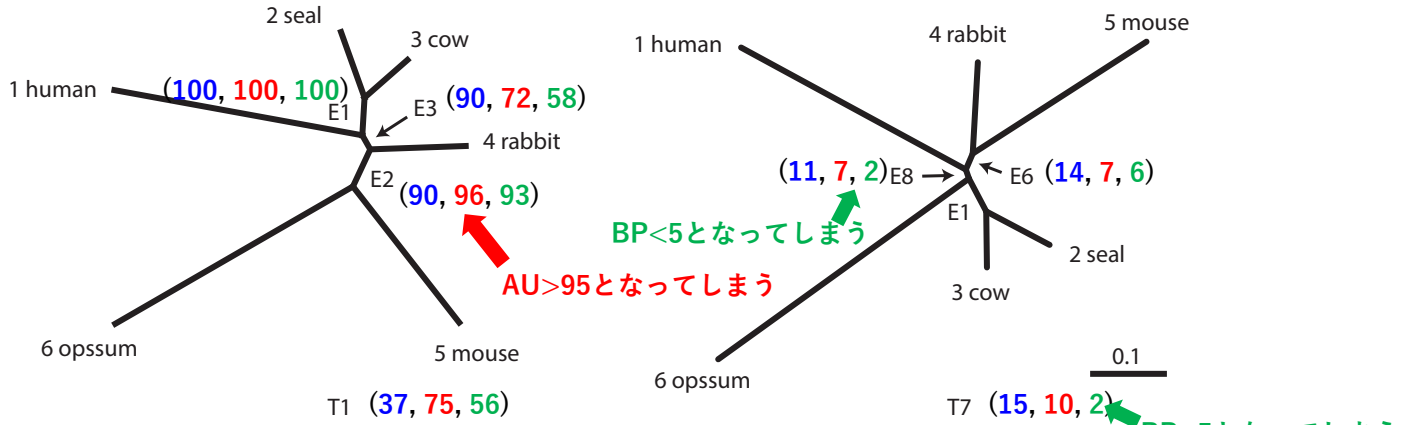
論文共著者ネットワークデータ（複雑ネットワーク分野）からの推定結果

- ブートストラップ法は，データからリサンプリング（復元抽出）
- 擬似的にたくさんデータができるので，何回もデータ解析すれば，そのバラツキもわかる
- 簡単なアルゴリズムだが，背後には確率分布の空間における距離とか曲率など高度な数理統計学



- 同じ結果が出た回数を素朴に数える(BP)は，「ニセの発見」になりやすい！
- マルチスケール・ブートストラップ法(AU)は，**n’ = -n** としてバイアスゼロ！
- じつは「膨大な仮説集合から仮説を選んでるバイアス」を直してなかった
- そのバイアスも新たに開発したばかりの選択的推測(SI)で解決できた！

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）

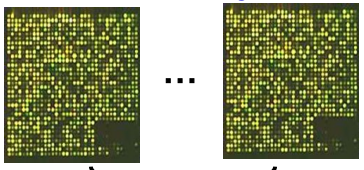


系統樹推定の例．T1はデータが支持する系統樹だが，実際にはT7が真実の系統樹と考えられる．

- 従来の統計理論では，次元dを固定して，サンプルサイズnが大きい極限を考えてる
- ところが，nは小さいのに，dが大きいデータを分析することが増えている
- 新しい統計理論が必要！

High-dimension, low-sample-size (HDLSS) data

Microarray



n

例. 白血病や乳がん等
d (遺伝子数) ≈ 10000
n (患者数) ≈ 100

$d \gg n$

~~大標本理論~~

標本共分散行列の逆行列が不安定
各種次元の呪い

$d \to \infty,\ d/n \to \infty$
$n \to \infty$ or n: fixed

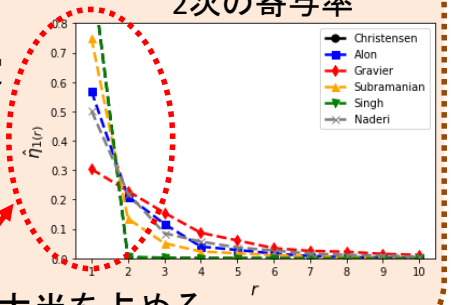
新たな理論・方法論の構築が必要

高次元小標本漸近理論

潜在空間の解析

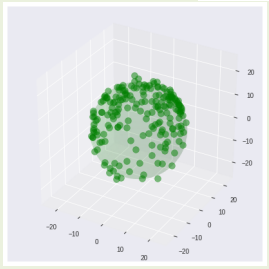
固有解析
・固有値・固有ベクトル推定
・固有値は高次元で
　不一致性を持つ
→要修正

2次の寄与率



上位の固有空間が大半を占める

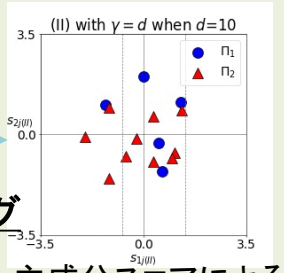
ノイズ空間の解析



球面集中現象

応用

判別分析
クラスタリング



主成分スコアによる
クラスタリング