

大規模ゲノムコホートデータの人工知能解析

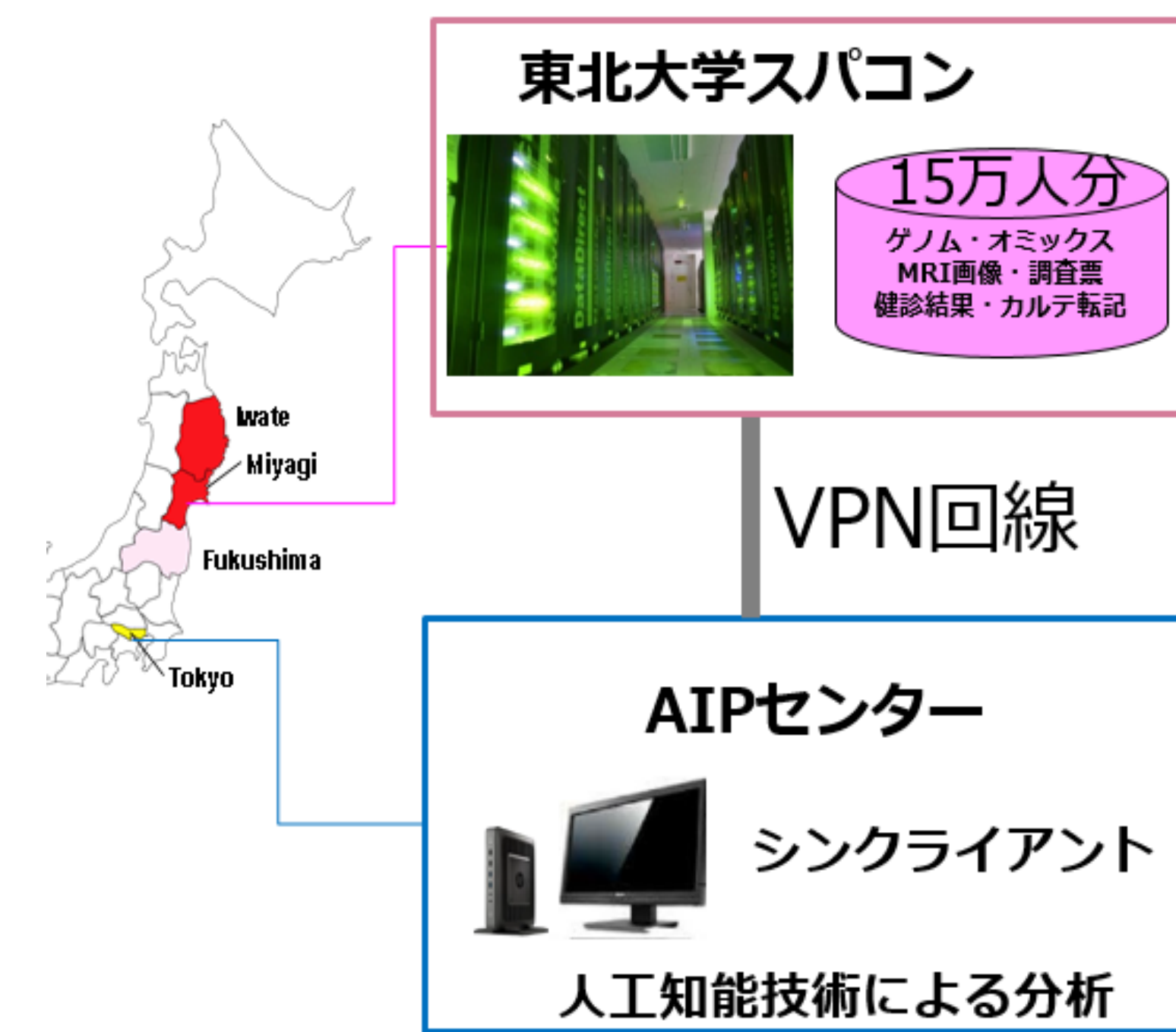
背景: 東日本大震災の被災地域において生じる健康変化の原因を明らかにし、対処することが求められている

目標: 15万規模の大規模ゲノムコホートデータに人工知能技術を適用して、うつ病や心疾患などの健康変化を生じさせる原因を明らかにし、今後のリスク予測を可能にする

エラー自動検出

これまでの成果: 15万人のデータ中には予測不可能な多種のエラーが存在し、予測精度を下げる要因となっていたため、エラーを自動検出する人工知能アルゴリズムを開発した(kurPCA法[1])

実績: kurPCAは東北メディカルメガバンク機構での実データクリーニングに利用されている



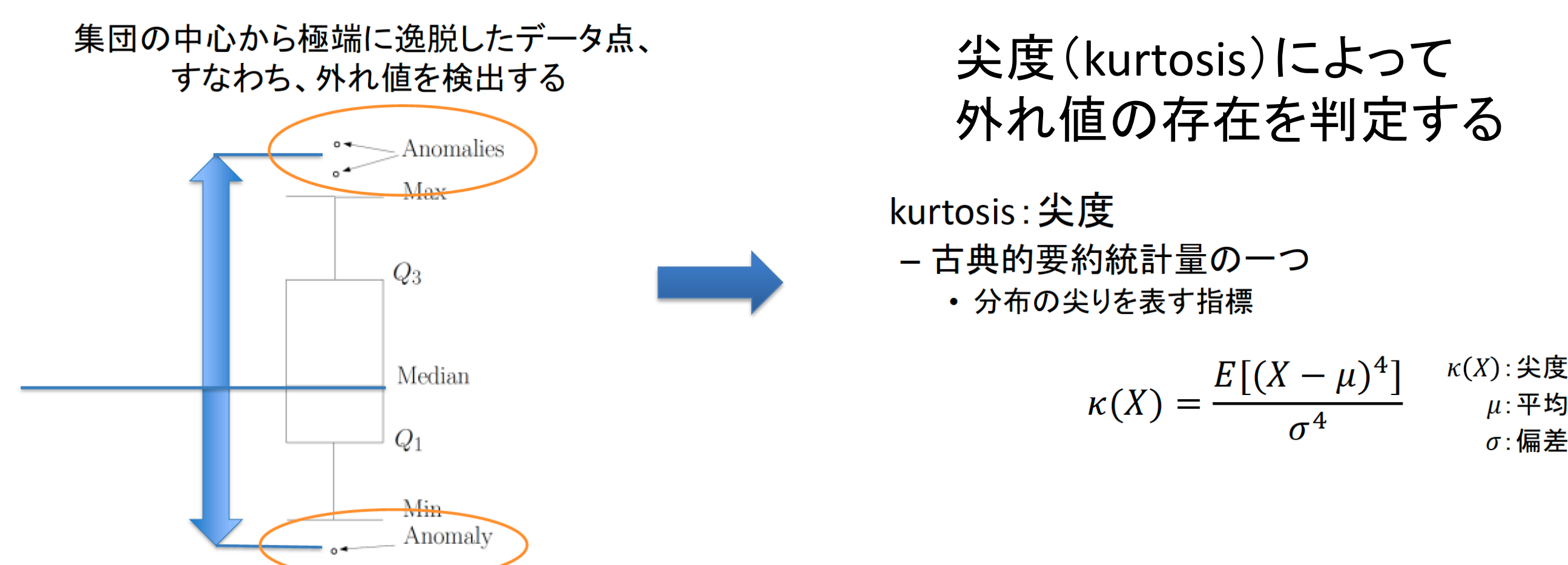
方法

検出対象のエラータイプ

- 一変数上で外れ値となっているデータ
- 多変数上で外れ値となっている個体のデータ
- 未回答であるべき質問に対する回答データ(および、その逆)

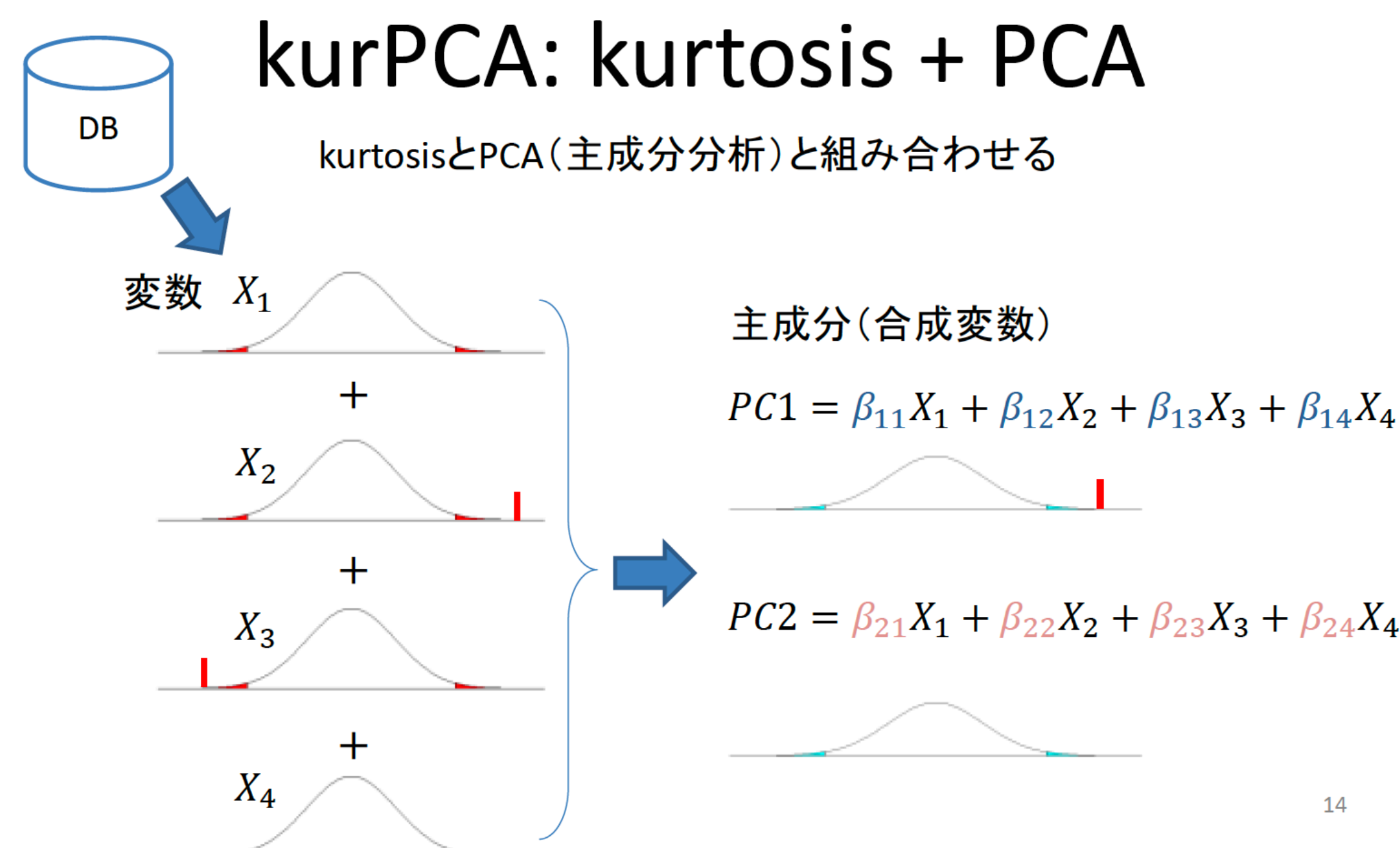
1. 一変数上で外れ値となっているデータ

kurtosis(尖度)は、確率変数の四次モーメントで表され、二次モーメントの分散よりも平均値から大きく離れた値(外れ値)に対する感受性が高い[2]ことから、外れ値の存在有無の判定にkurtosisを用いることを提案した。



2. 多変数上で外れ値となっている個体のデータ

多変数上で同時に外れ値となるデータは、正規化した変数間で強い相関を持つ値をとるデータであり、主成分軸上でより外れ具合が強調されるため、主成分軸上でkurtosisにより外れ値の有無を判定し、外れ値がなくなるまで反復的に外れ値を抽出するアルゴリズムを開発した(kurPCA)。また主成分において、離散変数は線形で表され、連続変数に近づくことが期待され、連続変数と同様のアルゴリズムが適用できる。よって、離散変数における稀な組み合わせの検出も可能である。



従来法との比較

シミュレーションにより、従来法[3,4]との比較を行った結果、

- 連続変数において、従来法と引けを取らなかった
 - 離散(二値)変数において、従来法では計算不能となる場合でもkurPCAでは問題なく計算できた
- 詳細は、[1]を参照のこと。

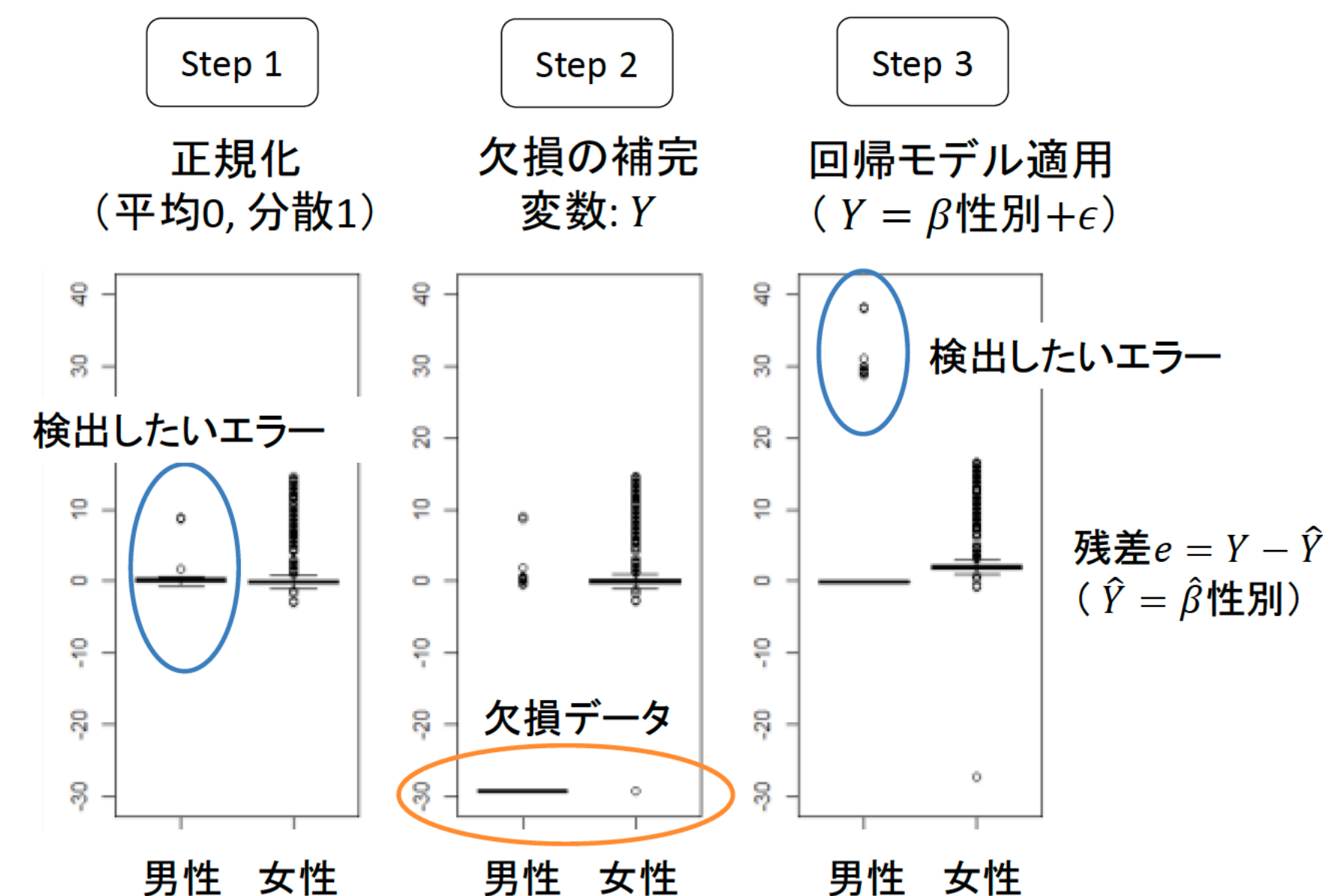
考察

本研究では、未知のエラーに対し、kurtosis, PCA, 回帰モデルを組み合わせることで自動検出を可能とするツールを開発した。本研究で用いた指標は変数の分布型に依存せずに定義したが、分布型が既知の場合、固定値が与えられる(正規分布ならば、kurtosisの閾値は3である)。さらに、本研究で適用したデータは質問票であり、欠損が多く含まれたが、これらの欠損は質問の設計者が予め回答者を限定(女性のみ、喫煙者のみ等)した結果生じたものであった。我々は事前に質問内容の情報を持たなくても、前処理(RAMP)を施すことにより、自動で欠損パターンを検出し、その上で設計の意図から外れたデータを抽出することに成功した。

3. 未回答であるべき質問に対する回答データ

欠損パターンの調整(regression adjustment of missing patterns:RAMP)

質問票の質問が回答者を限定して作られた場合、欠損がパターンとしてデータに潜在する。そこで、kurPCAの前処理として、回帰モデルを用いて依存関係がある変数のペアを探索し、パターンを調整した上で外れ値検出を行う方法を提案した。

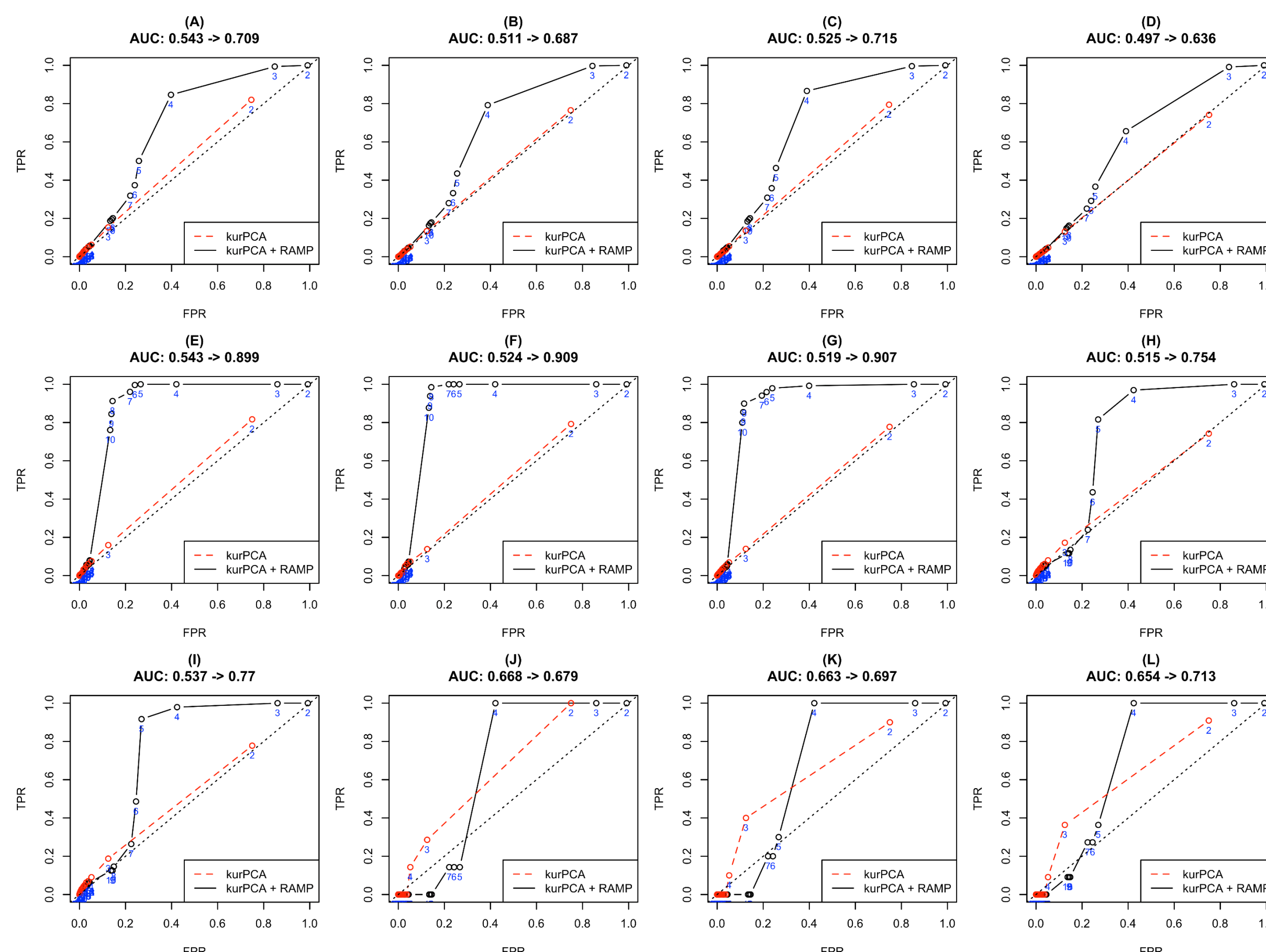


RAMPの性能

RAMPで検出されたペア

Index	Adjusted variable	Adjustment variable
(A)	ゆっくり歩く頻度	1回あたりの時間
(B)	早足で歩く頻度	1回あたりの時間
(C)	軽・中程度の運動の頻度	1回あたりの時間
(D)	激しい運動の頻度	1回あたりの時間
(E)	習慣的に飲み始めるようになった年齢を教えてください。	あなたはアルコール類を飲みますか？
(F)	平均してアルコール類を飲む頻度はどのくらいですか？	あなたはアルコール類を飲みますか？
(G)	現在もタバコを吸っていますか？	生まれてからこれまでに 100 本以上のタバコを吸いましたか？
(H)	現在タバコを吸っている人にお聞きします。何歳から吸い始めましたか？	現在もタバコを吸っていますか？
(I)	現在タバコを吸っている人にお聞きします。1日何本吸いましたか？	現在もタバコを吸っていますか？
(J)	タバコを止めた人にお聞きします。何歳の時にタバコを止めましたか？	現在もタバコを吸っていますか？
(K)	タバコを止めた人にお聞きします。何歳からタバコを始めましたか？	現在もタバコを吸っていますか？
(L)	タバコを止めた人にお聞きします。1日何本吸っていましたか？	現在もタバコを吸っていますか？

kurPCA と RAMP+kurPCA の比較



参照

- [1] Sakurai R, et al. *Int J Epidemiol* 2019;doi: 10.1093/ije/dyz012
- [2] Westfall HP. *Am Stat* 2014;68:191–5.
- [3] Hubert M, Van Der Veeken S. *J Chemom* 2008;22:235–46.
- [4] Hubert M, Engelen S. *Bioinformatics* 2004;20:1728–36.