THE UNIVERSITY OF TOKYO

複数情報源に対する主成分分析

諏訪恭平／冨岡亮太／矢入健久／○鹿島久嗣

東京大学

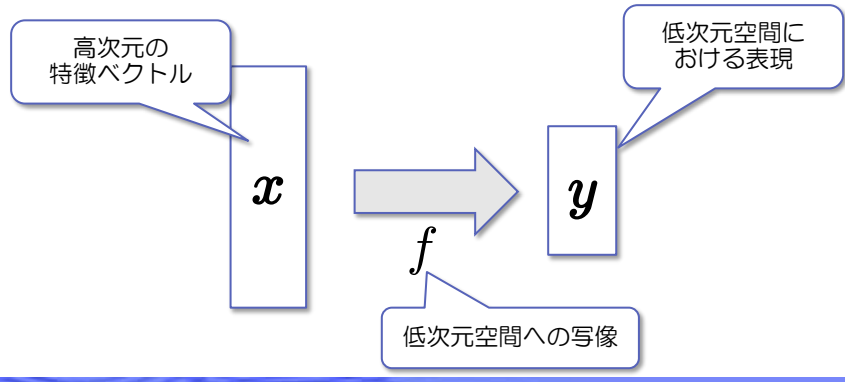
DEPARTMENT OF MATHEMATICAL INFORMATICS

複数種類の情報源から得られたデータに対する主成分分析法を提案します

- はじめての複数情報源に対する主成分分析法
  - 重要な情報源を自動的に選択する機構をもつ
- みどころ：
  - グループラッソ制約を用いることで情報選択機構をもつ
  - 主成分ごとに異なる情報源を選択する
  - グループラッソから複数カーネル学習の導出レシピ
- 観測データからの物体配置図復元タスクで有効性を検証
  - 相補的な情報を適切に組み合わせる
  - 不要な情報源を切り捨て、有用なもののみを残す
- いまのところ実験的にはいまひとつ…

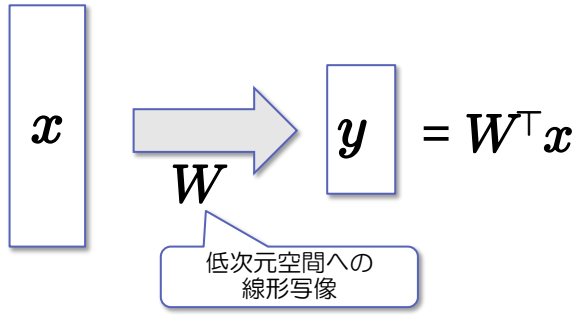
次元削減は 高次元空間のデータを 低次元空間に落として扱いやすくするための技術 です

- 多くの分析対象は高次元の実数値特徴ベクトル $\boldsymbol{x}$ として表現される
- もとのデータ $\boldsymbol{x}$ の性質を保ったまま、低次元の空間における $\boldsymbol{y}$ を得るための写像 f を求めるのが次元削減である
- 視覚化や前処理に利用される



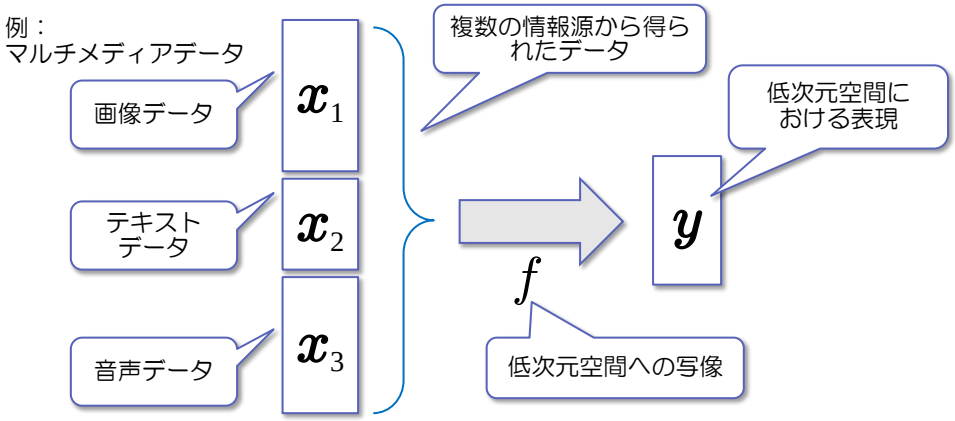
たとえば 主成分分析は 代表的な次元削減法のひとつ です

- 射影行列 $\boldsymbol{W}$（$D \times d$ 行列 ）を用いてD次元から d ($\ll D$)次元に射影
  - $\boldsymbol{W}$ ： 正規直交基底（$\boldsymbol{W}^{\top}\boldsymbol{W} = \boldsymbol{I}$）



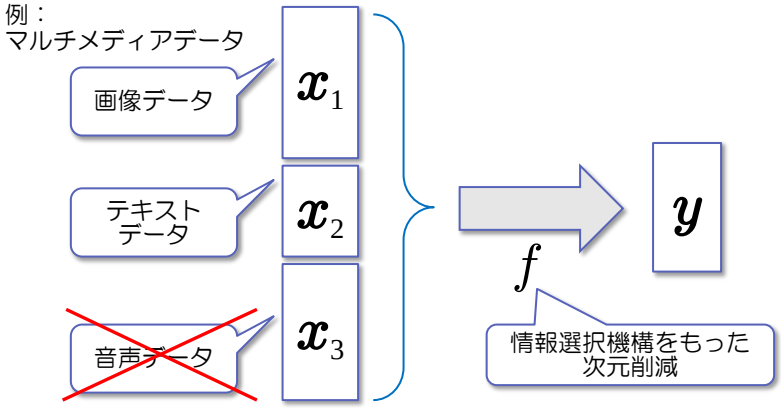
本研究では 複数の情報源からデータが得られるような場合 の次元削減を考えます

- 多くの場合、一つの対象は複数の情報源からのデータをもつ
  - マルチメディアデータ：画像、テキスト、音声、…
  - 医療データ： 検査項目、SNP、遺伝子発現、…



本研究では 複数の情報源の取捨選択/重みづけを自動的に行う次元削減法を考えます

- データには特徴的な構造をもったものとそうでないものがある
- これらを自動的に取捨選択または重みづけする次元削減法がほしい
- 具体的には主成分分析をベースに考える



主成分分析の復習

主成分分析の復習：射影の分散を最大にするような写像を求めます

- 1次元に落とす場合を考える
  - $\boldsymbol{X}$：データのデザイン行列（$\boldsymbol{X}^\top$の列が各特徴ベクトル）
  - $\boldsymbol{w}$：射影ベクトル
- 射影後の分散を最大にするような最適化問題を解くことで、射影ベクトル $\boldsymbol{w}$ が求まる

$$\text{maximize}_{\boldsymbol{w}} \ \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} \quad \text{（射影後の分散）}$$

$$\text{s.t.} \ \|\boldsymbol{w}\|_2^2 \leq 1 \quad \text{（射影ベクトルの大きさを制限）}$$

主成分分析の復習：
最適な射影は固有値問題を解くことで得られます

- 最適解は固有値問題を解くことによって求まる

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} = \lambda \boldsymbol{w}$$

  - 最大固有ベクトルを求める
- 2次元以上に落とす場合には：
  - 方法１：固有値の大きいほうから固有ベクトルを複数とる
  - 方法２：d 次元目までの寄与分を取り除いたデザイン行列

$$\boldsymbol{X}^{(d+1)} \equiv \boldsymbol{X}^{(d)} - \boldsymbol{y}^{(d)}\boldsymbol{w}^{(d)\top}$$

に対し固有値問題の最大固有ベクトル $\boldsymbol{w}^{(d+1)}$ を求める操作を繰り返す

提案手法：複数情報源の主成分分析

提案手法：主成分分析の目的関数における2-ノルム制約をグループラッソの制約でおきかえます

- 主成分分析を 情報源がM 個ある場合に拡張する
  - $\boldsymbol{X} \equiv (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$ ：M 個の情報源データのデザイン行列
  - $\boldsymbol{w} \equiv (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M)$ ：射影ベクトル
- 最適化問題における2ノルム制約を、グループラッソで用いられる制約で置き換える

$$\text{maximize}_{\boldsymbol{w}} \ \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} \quad \text{（射影後の分散）}$$

$$\text{s.t.} \ \sum_m \|\boldsymbol{w}_m\|_2 \leq 1 \quad \text{（射影ベクトルの大きさを制限）}$$

※もともとの制約は $\sum_m \|\boldsymbol{w}_m\|_2^2 \leq 1$

- グループラッソの制約は<u>各情報源の射影ベクトル単位で**0**につぶす</u>
  - 1-ノルムの特徴選択効果を特徴のグループに対して与える

最適化のテクニック：グループラッソ制約を 扱いやすい2ノルム制約に帰着します

- ラグランジアンはそのままは最適化しにくい

2-ノルムの1乗

$$L(\boldsymbol{w}, \lambda) \equiv \boldsymbol{w}^\top \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{w} - \lambda \left(\sum_m \|\boldsymbol{w}_m\|_2 - 1\right)$$

2-ノルムの2乗！（扱いやすい）

- ラグランジアンを扱いやすい形に変形する
  - 相加相乗平均から

$$\|\mathbf{w}_m\|_2 = \min_{\beta_m \geq 0} \frac{1}{2}\left(\frac{\|\mathbf{w}_m\|_2^2}{\beta_m} + \beta_m\right)$$

  - これをもちいて

$$\tilde{L}(\mathbf{w}, \boldsymbol{\beta}, \lambda) \equiv \mathbf{w}\mathbf{X}^\top \mathbf{X}\mathbf{w} - \lambda\left(\sum_{m=1}^{M} \frac{1}{2}\left(\frac{\|\mathbf{w}_m\|_2^2}{\beta_m} + \beta_m\right) - 1\right)$$

$\boldsymbol{\beta}$ が登場

    - 「2ノルムの2乗」の項を「2ノルムの2乗」で置き換えた
    - 但しパラメータ$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)$ が増えていることに注意

新たに導入されたパラメータβ は 各情報源の重要度 として解釈できます

- $1/\beta_m$ はm番目の情報源に対応する $\|w_m\|_2^2$ の係数になっている

$$\tilde{L}(\mathbf{w}, \boldsymbol{\beta}, \lambda) \equiv \mathbf{w}\mathbf{X}^\top\mathbf{X}\mathbf{w} - \lambda\left(\sum_{m=1}^{M}\frac{1}{2}\left(\frac{\|\mathbf{w}_m\|_2^2}{\beta_m} + \beta_m\right) - 1\right)$$

β が大きい⇒
w_mの正則化が弱い

- β_m はm番目の情報源の重要度として解釈できる
  - β_m が大きいほど w_m の正則化が弱まる ⇒ w_mを大きくできる
  - β_m が小さいほど w_m の正則化が強まる ⇒ w_mが0になりやすい

逐次最適化アルゴリズム：射影ベクトル w と重要度パラメータ β を交互に最適化します

- 射影ベクトル$\boldsymbol{w} \equiv (w_1, w_2, \ldots, w_M)$ についての最大化
  - $\boldsymbol{w} \equiv (w_1, w_2, \ldots, w_M)$ と $\boldsymbol{X} \equiv (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_M)$ をスケーリングする

⇨ $\tilde{\mathbf{w}}_m \equiv \frac{\mathbf{w}_m}{\sqrt{\beta_m}}$ ⇨ $\tilde{\mathbf{X}}_m(\beta_m) \equiv \sqrt{\beta_m}\mathbf{X}_m$

  - 通常の主成分分析と同様の形のラグランジアンが得られる

$$\tilde{L}(\tilde{\mathbf{w}}, \lambda) = \tilde{\mathbf{w}}^\top\tilde{\mathbf{X}}(\boldsymbol{\beta})^\top\tilde{\mathbf{X}}(\boldsymbol{\beta})\tilde{\mathbf{w}} - \frac{\lambda}{2}\left(\|\tilde{\mathbf{w}}\|_2^2 + \sum_{m=1}^{M}\beta_m - 2\right)$$

  - 固有値問題に帰着される
    - 最大固有ベクトルを求める

$$\tilde{\mathbf{X}}(\boldsymbol{\beta})^\top\tilde{\mathbf{X}}(\boldsymbol{\beta})\tilde{\mathbf{w}} = \frac{\lambda}{2}\tilde{\mathbf{w}}$$

- 重要度パラメータ$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)$ についての最大化
  - こちらは閉じた形で解が求まる $\beta_m^{new} = \sqrt{\beta_m^{old}}\|\tilde{\mathbf{w}}_m\|_2$

主成分の逐次抽出：第 d 主成分までの寄与分をデータから取り除き、第 d+1 主成分を求めます

- 繰り返しによって、主成分が1軸求まる
  - 射影ベクトル$\boldsymbol{w}$ についての最大化（最大固有ベクトル）
  - 重要度パラメータβ についての最大化（閉じた形の解）
- さらにこれを繰り返すことで、1軸づつ主成分を求める
  - 第 d 主成分は $\mathbf{y}^{(d)} \equiv \sum_{m=1}^{M} \mathbf{y}_m^{(d)}$ $\quad \mathbf{y}_m^{(d)} \equiv \mathbf{X}_m \mathbf{w}_m^{(d)}$
  - 第 d 主成分の寄与分をデータから取り除く

$$\mathbf{X}_m^{(d+1)} \equiv \mathbf{X}_m^{(d)} - \mathbf{y}^{(d)} \mathbf{w}_m^{(d)\top}$$

- 主成分ごとに異なった情報源重要度が得られる
  - 既存の情報源選択法と異なる点

マルチカーネル化：2ノルム制約が現れたことによりカーネル関数による非線形化が可能になります

- 目的関数の変形によって2ノルムの2乗が登場した

$$\tilde{L}(\tilde{\mathbf{w}}, \lambda) = \tilde{\mathbf{w}}^\top \tilde{\mathbf{X}}(\boldsymbol{\beta})^\top \tilde{\mathbf{X}}(\boldsymbol{\beta}) \tilde{\mathbf{w}} - \frac{\lambda}{2} \left(\|\tilde{\mathbf{w}}\|_2^2 + \sum_{m=1}^{M} \beta_m - 2 \right)$$

- カーネル法による非線形化が可能
  - パラメータをおきかえる $\tilde{\mathbf{w}}_m \equiv \tilde{\mathbf{X}}_m(\boldsymbol{\beta})^\top \tilde{\boldsymbol{\alpha}}$
  - 固有値問題は：

情報源ごとのカーネル関数（マルチカーネル）

$$\tilde{\mathbf{K}}(\boldsymbol{\beta}) \tilde{\boldsymbol{\alpha}} = \frac{\lambda}{2} \tilde{\boldsymbol{\alpha}} \quad \text{ただし} \quad \tilde{\mathbf{K}}(\boldsymbol{\beta}) \equiv \sum_{m=1}^{M} \beta_m \mathbf{K}_m$$

$$\mathbf{K}_m \equiv \mathbf{X}_m \mathbf{X}_m^\top$$

数値実験：物体配置図の復元タスク

物体配置図の復元タスクを題材に その復元精度で性能評価します

- 複数の物体の配置図を複数の観測点での観測データから復元する
  - 観測点と観測対象の座標はともに不明
- 例えば、各観測点からの各物体への距離や方角などが観測される
- 次元削減によって2次元に落とすことで配置図を復元する

	物体 1	物体 2	...	物体 N
観測点 1				
観測点 2				
...				
観測点 D				

観測点と物体の距離

	物体 1	物体 2	...	物体 N
観測点 1				
観測点 2				
...				
観測点 D				

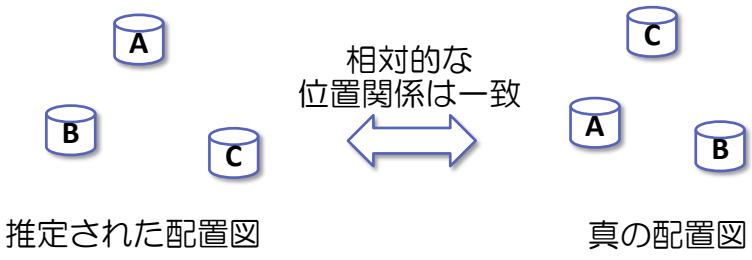
観測点からみた物体の方角

次元削減

	物体 1	物体 2	...	物体 N
X軸				
Y軸				

評価方法：
配置図の復元精度は任意の3点の相対位置の一致度で測ります

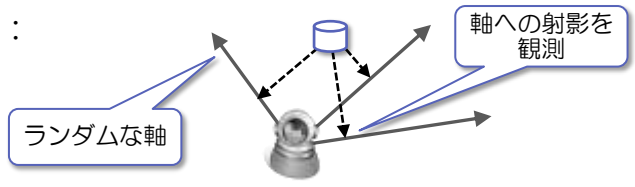
- 教師なし次元削減では絶対的な座標は復元されない
  - わかるのは対的な位置関係のみ
- 復元精度は任意の3点の相対位置が一致する割合で評価する
  （＝ 時計回りの順が一致する割合）



推定された配置図　　真の配置図

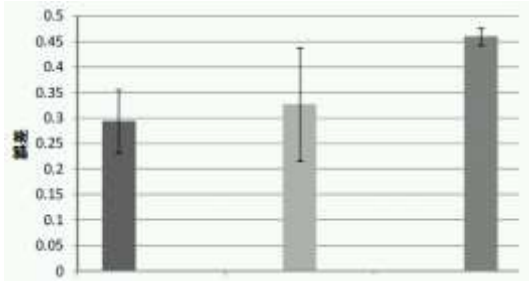
実験１：相補的な情報源の統合

- ランダムな3方向に物体を射影したときの距離を3つの情報源とする
  - 相補的に働く
- これらを組み合わせることで配置図の復元精度が上がるか？
- 比較手法：
  - 提案手法
  - 通常の主成分分析
  - 各情報源で主成分分析を行ったもののうち最も精度の高いもの
- ランダムにデータを生成：
  - 物体数：50
  - 観測点数：500



実験１：提案手法は相補的な情報をうまく組み合わせることができます

- 比較手法
  - 提案手法
  - 通常の主成分分析
  - 各情報源で主成分分析を行ったもののうち最も精度の高いもの
- 提案手法が低い誤差と標準偏差を達成する



実験２：有用でない情報源の切り捨て

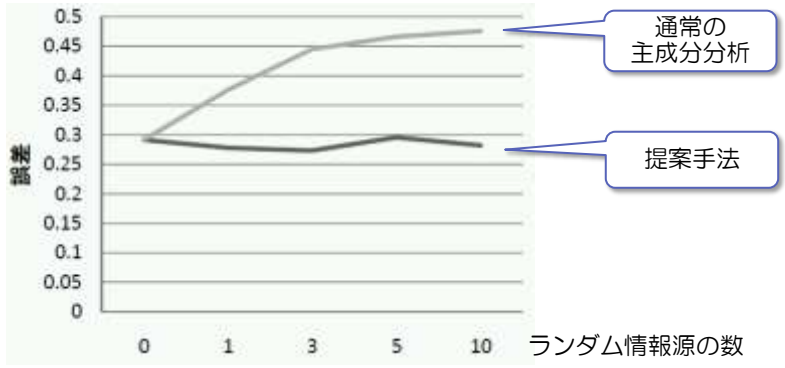
- ただ一つの有用な情報源に、役に立たない（ランダムな）情報源を加えて行った場合の復元精度への影響を調べる
- 有用な情報源として：
  観測点から一定距離内に物体がある（1）かない（0）か

	物体 1	物体 2	…	物体 N
観測点 1	1	0	…	1
観測点 2	1	1	…	0
…	…	…	…	…
観測点 D	0	0	…	1

各観測点から一定距離内にその物体があるかどうか

実験２：提案手法はノイズに影響されず有用な情報源のみを選び出すことができます

- ランダム情報源の数を増やしていくと：
  - 通常の主成分分析では精度が悪化していく
  - 提案手法では悪化しない



まとめ: 複数情報源から得られたデータに対する主成分分析法を提案しました

- 複数情報源に対する主成分分析法を提案した
- ポイント：
  - グループラッソ制約を主成分分析に用いることで情報源選択機構を実現する
  - 情報源の重要度を導入することで問題を解きやすくする
  - 主成分ごとに異なる情報源を選択する
- 提案手法は：
  - 相補的に働く情報源を適切に組み合わせることができる
  - 不要な情報源を取り除くことができる

  ことを物体配置図の復元タスクで検証した