

データサイエンス概論I & II データサイエンス総論I & II

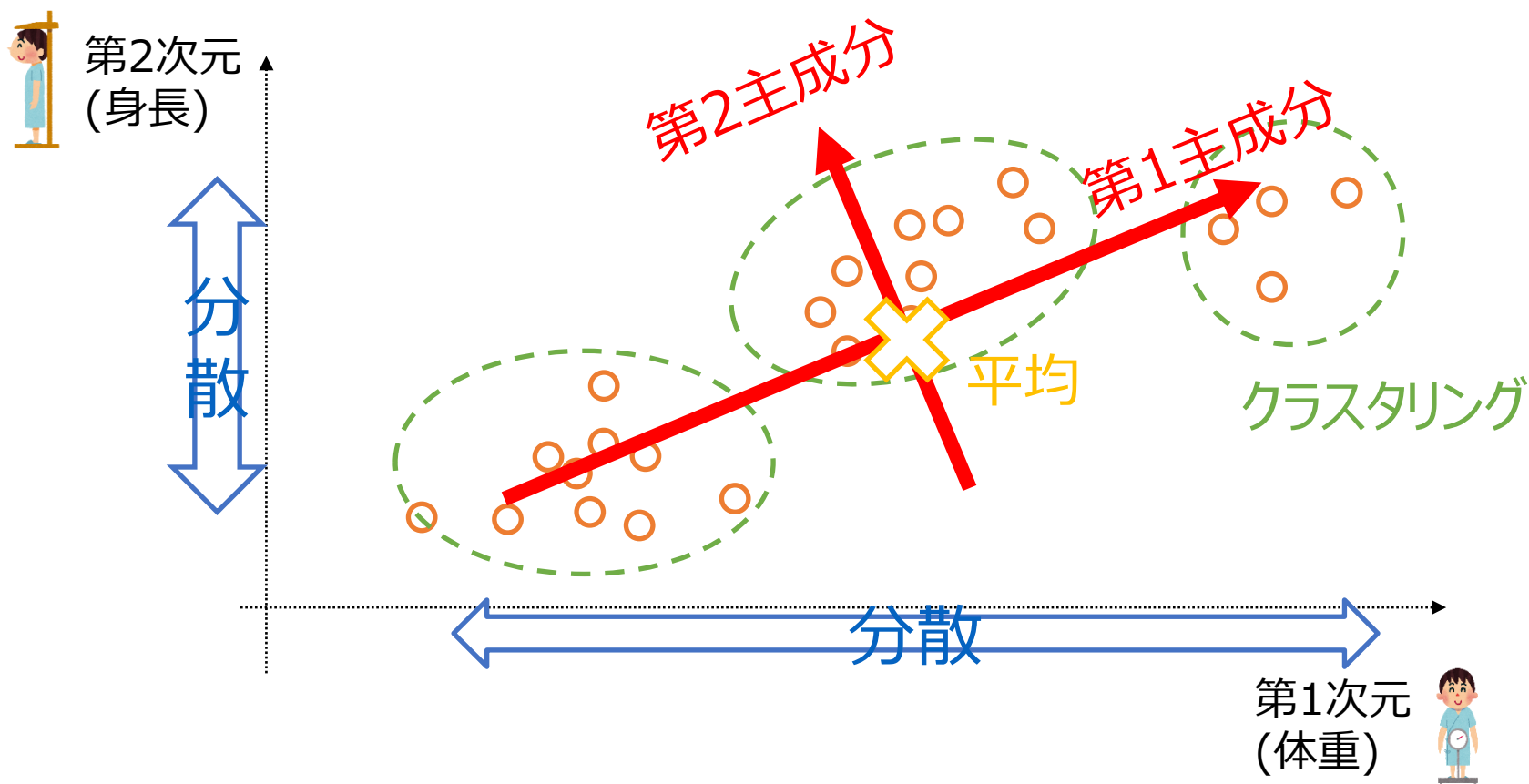
主成分分析

九州大学 数理・データサイエンス教育研究センター

このセクションの話

- 「主成分分析」の考え方を説明
 - いろいろな分野で非常によく使うデータ分析法
- 主成分分析とは，なるべく少ない基底で，データを表現（分析）する方法
 - データをコンパクトに表現できる！
- 「どういう基底を使えば，少なくて済むのか」がポイント
 - あれ，すでにそういう話，あったような…
 - そうなんです，実はもう，主成分分析の大事なところは，説明済みなんです

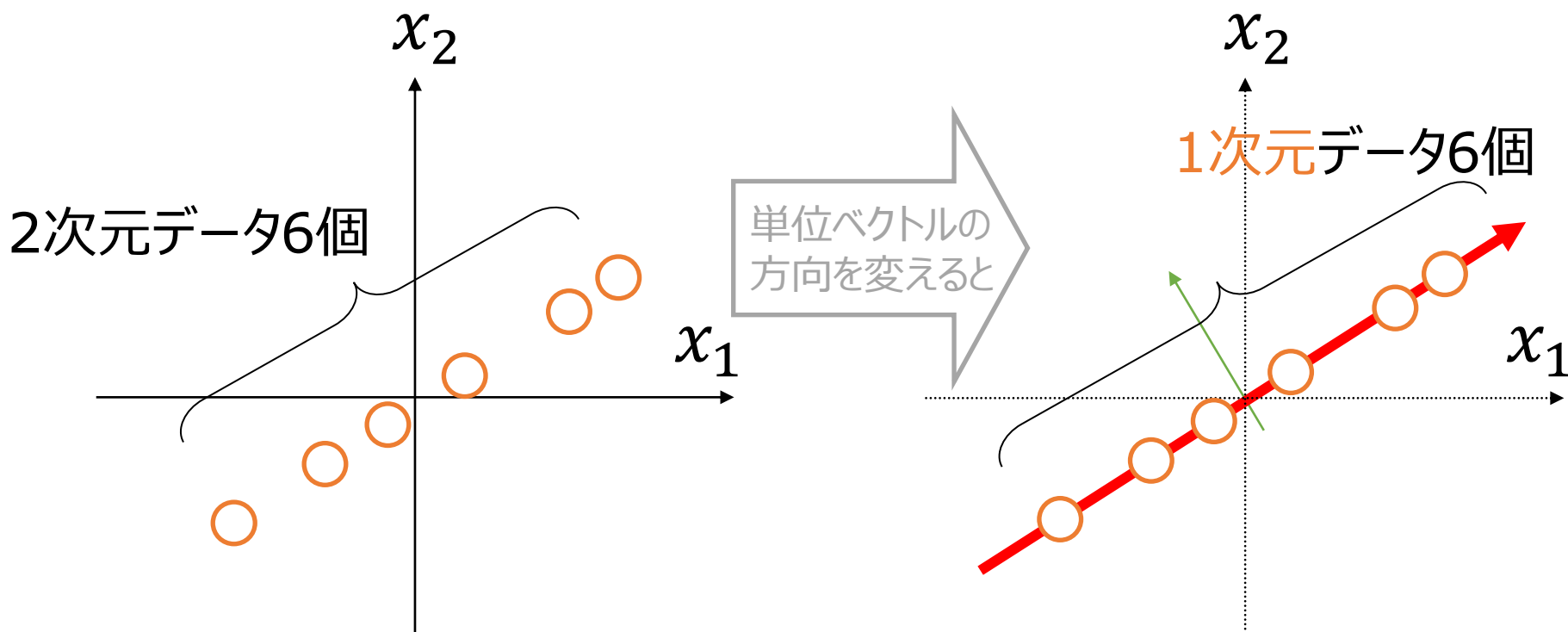
本セクションの最後にこんな図も出てきます



真の次元

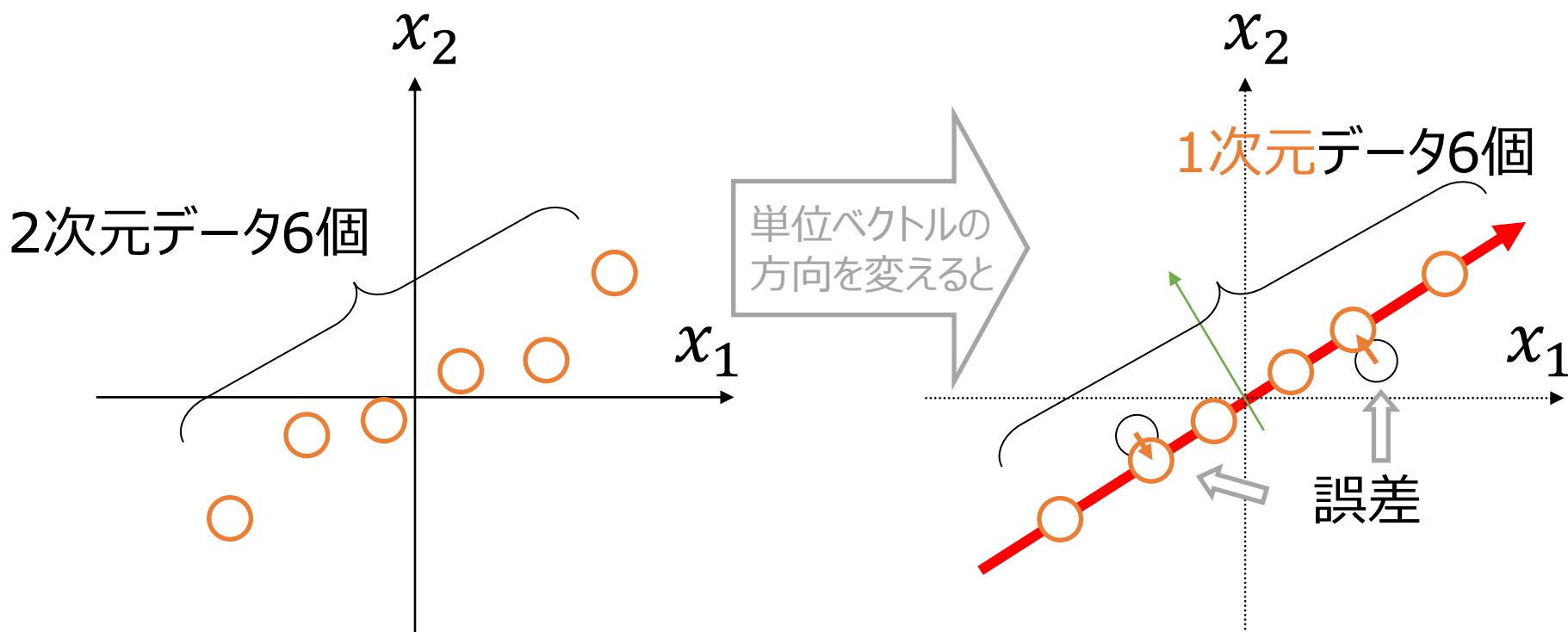
そのベクトルデータの分布は、どれくらいコンパクトなのか？

「真の次元」(1/3)



$\tilde{d} = 1 (< d = 2)$ 次元で表せた！
(真の次元=1)

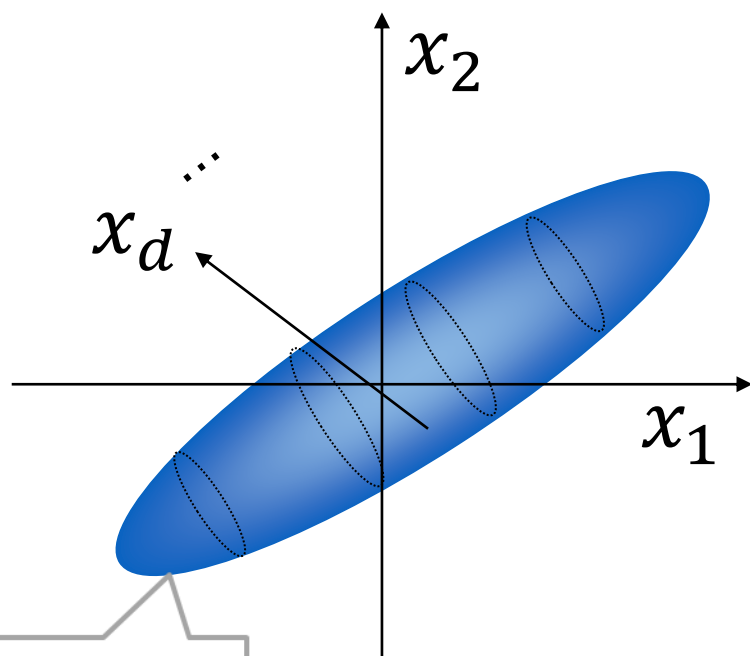
「真の次元」(2/3)



ほぼ $\tilde{d}=1 (< d=2)$ 次元で表せた！
(真の次元 $\doteq 1$)

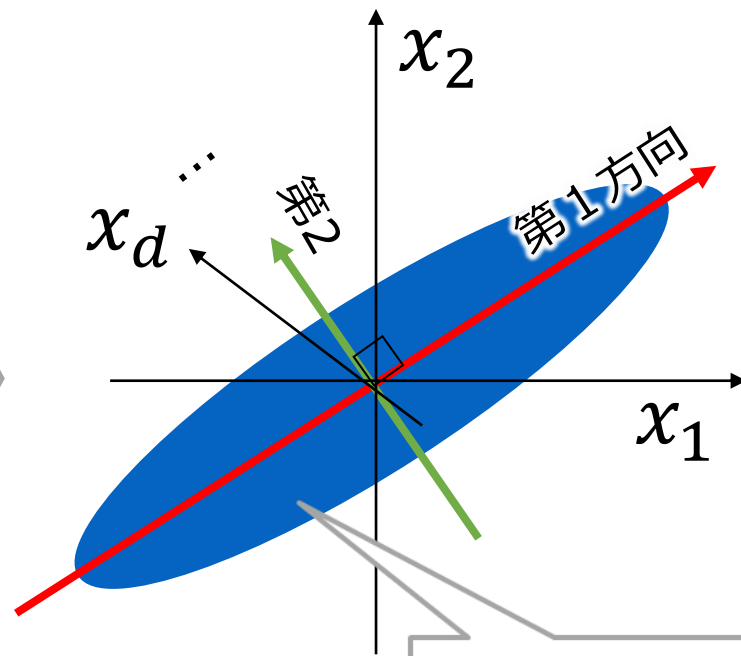
「真の次元」(3/3)

d 次元ベクトルの分布



つぶれた
フランスパンを...

\tilde{d} ($< d$)次元ベクトルの分布



楕円形の板と
みなした感じ

ほぼ \tilde{d} ($< d$)次元で表せた！
(真の次元 $\doteq \tilde{d}$)

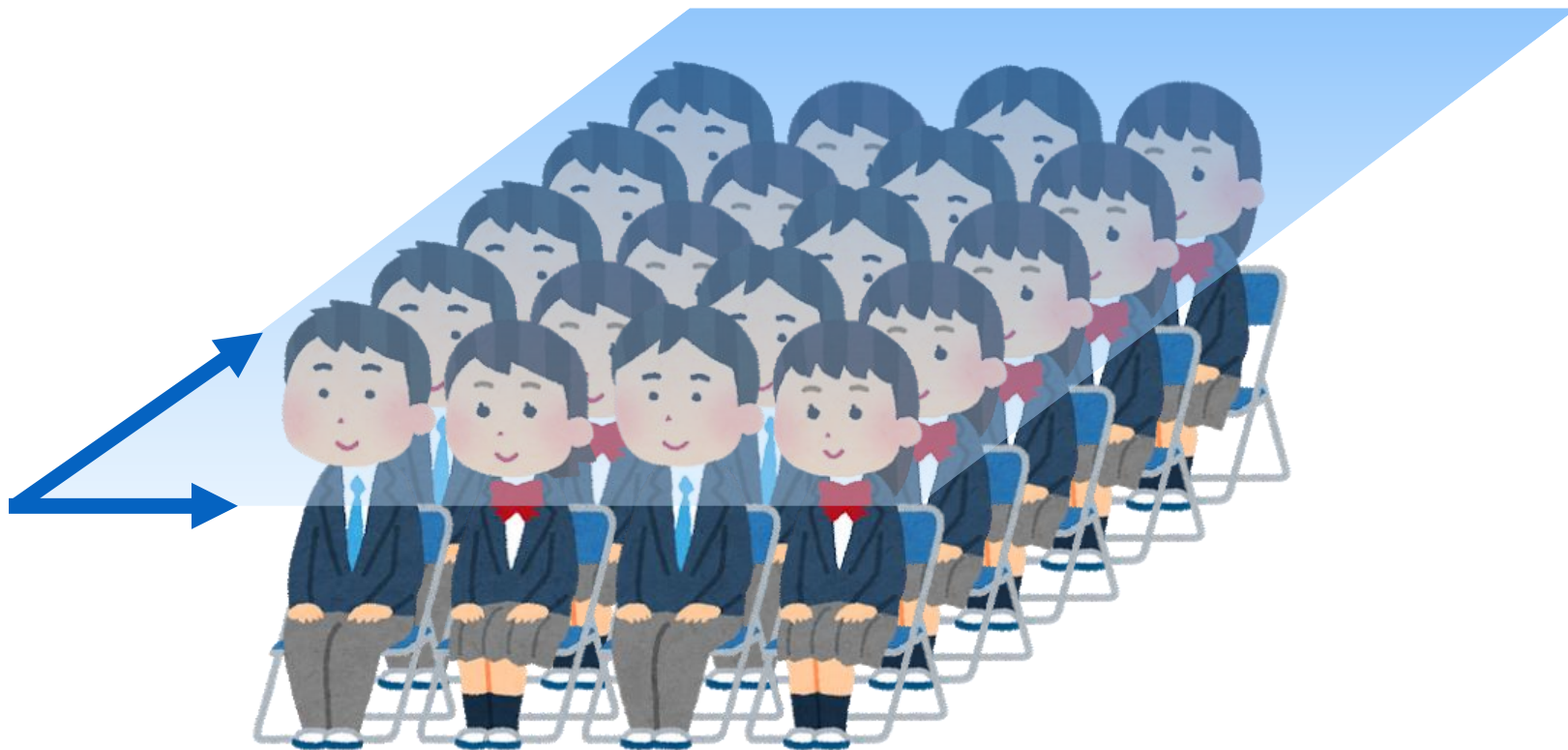
真の次元の分かりやすい例：皆さんの頭の分布

- 講義室の皆さんの頭の位置は $d = 3$ 次元ベクトル (x, y, z) で表現可能
- 多くの学生さんがいれば、 $d = 3$ 次元上に点（データ）が分布



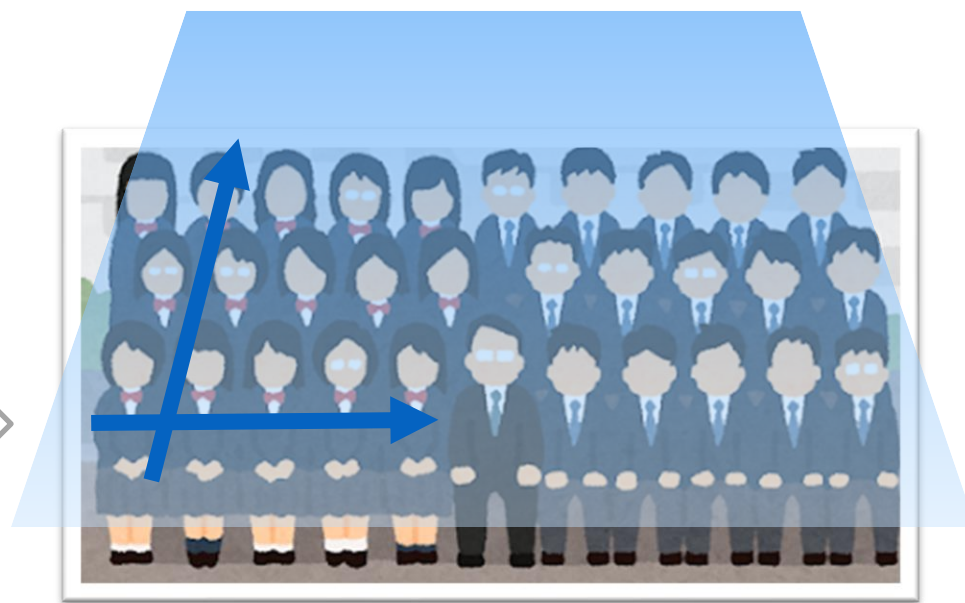
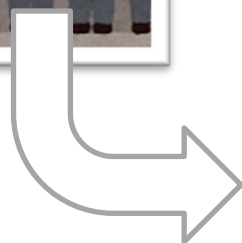
真の次元の分かりやすい例：皆さんの頭の分布

- 座高があまり変わらないとすれば、真の次元 \tilde{d} は ≈ 2 次元！



真の次元の分かりやすい例：皆さんの頭の分布

- 「ひな壇」に並んで立っているとすれば，傾いた平面に
 - 依然として真の次元 \tilde{d} は $\asymp 2$ 次元



真の次元の分かりやすい例：皆さんの頭の分布

- 立ったり座ったり好き勝手だとすると、「真の次元 \tilde{d} は2次元」とは言いづらい
 - 真の次元 \tilde{d} はもとの次元 d と同じで3次元



真の次元の分かりやすい例：皆さんの頭の分布

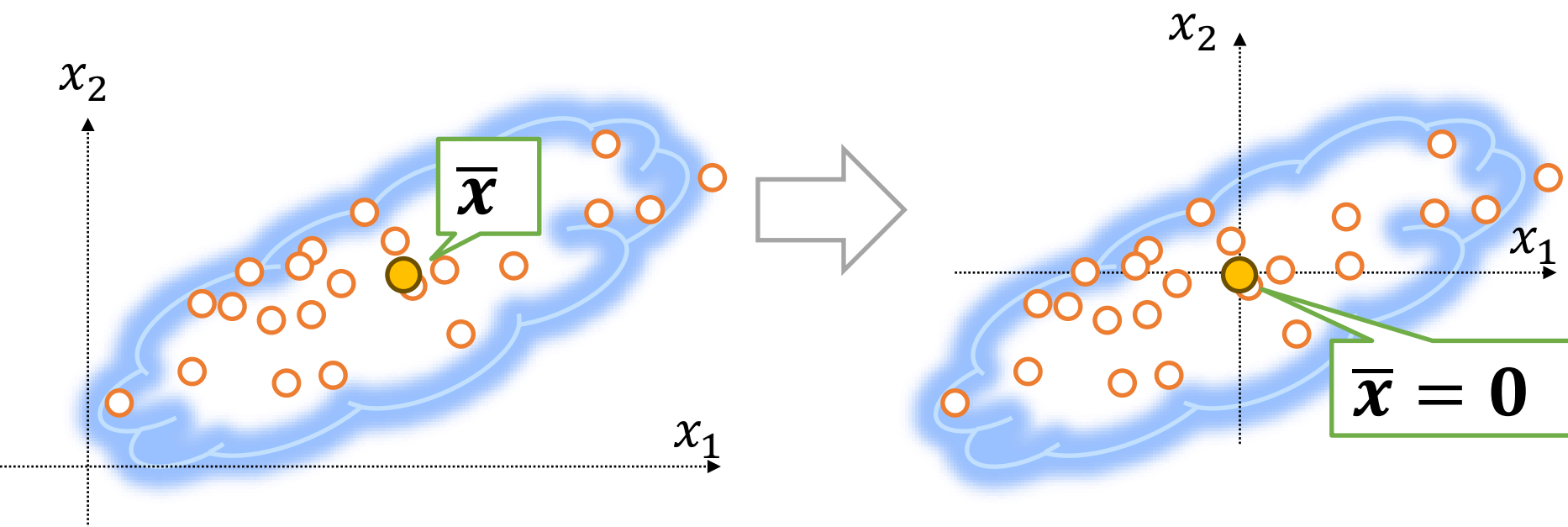
- 横一列に並んでいるとすれば，ほぼ直線
- 真の次元 \tilde{d} は ≈ 1 次元に！



主成分分析の原理

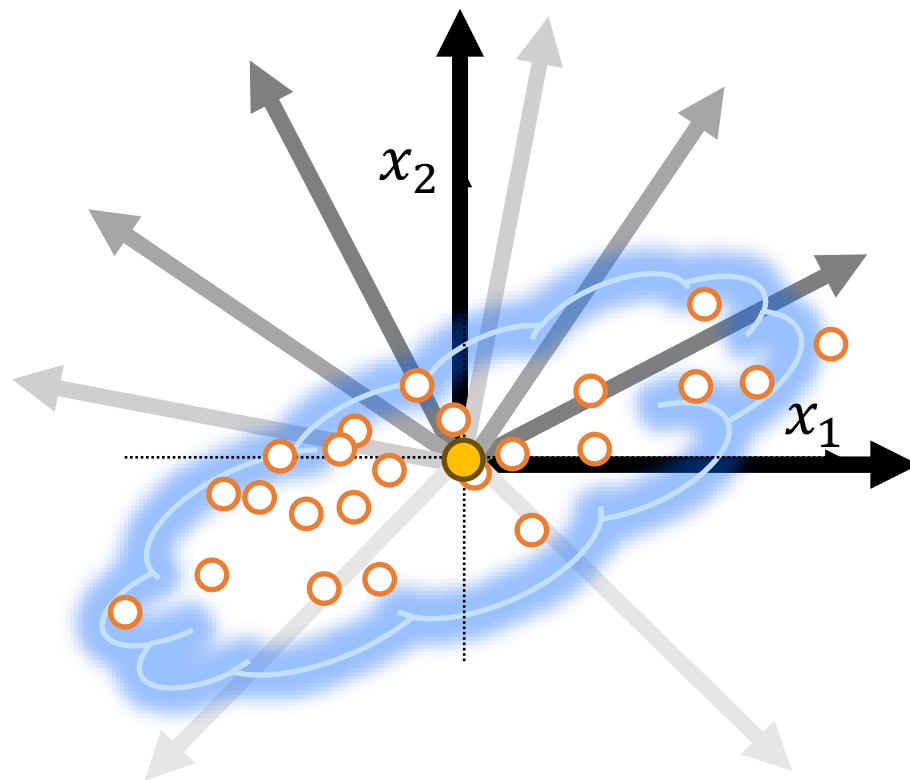
コンパクトな基底の求め方

わかりやすさのため, しばらく
「データ集合の平均 = 0」としましょう

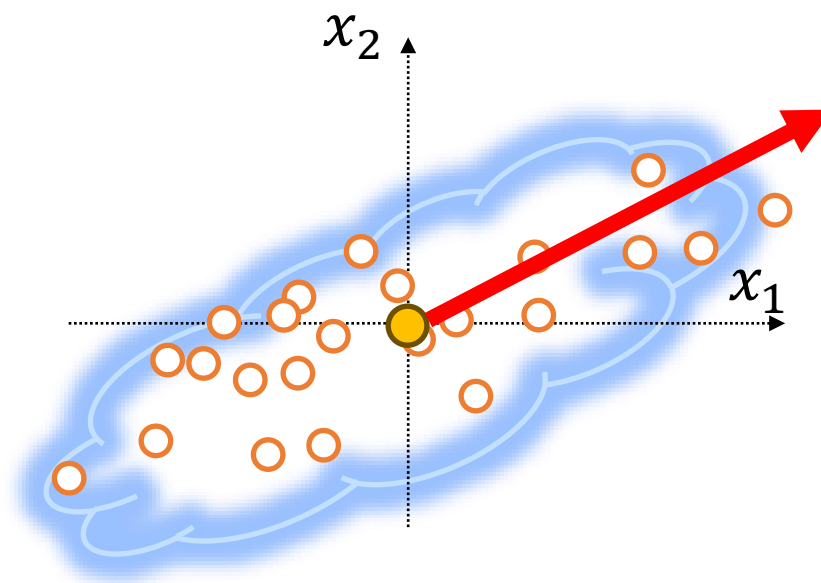


※単に平行移動でずらしただけ

分析(=分布把握)に最も適した基底を考える：
どれが最も「コンパクト」に分布を表現？

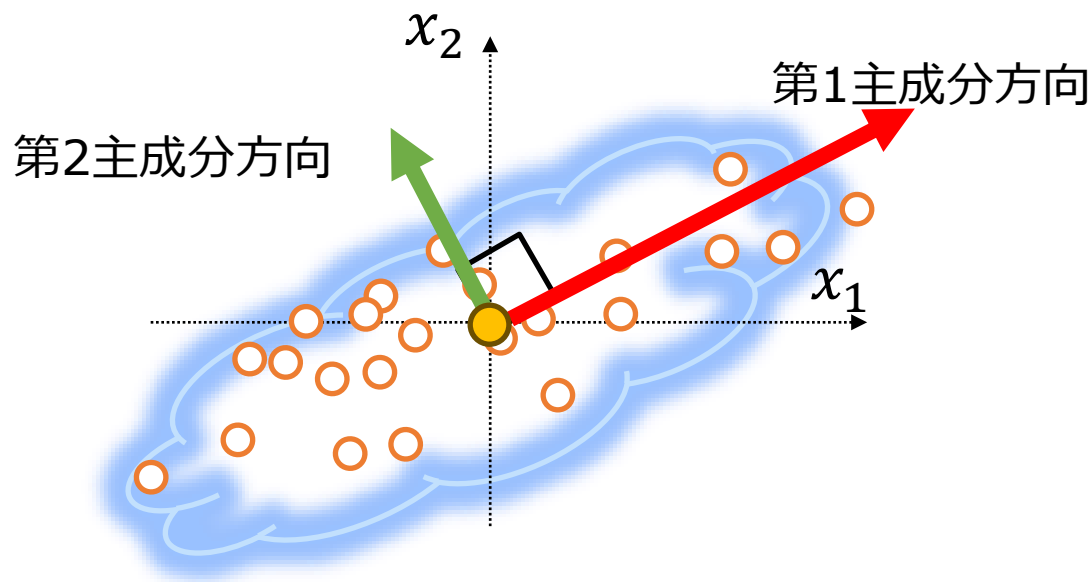


まずは最も広がった方向でしょう！
(「ソコソコ戻る」可能性が最も高い＝分布を最もよく表現)



- これを「第1主成分」と呼ぶ

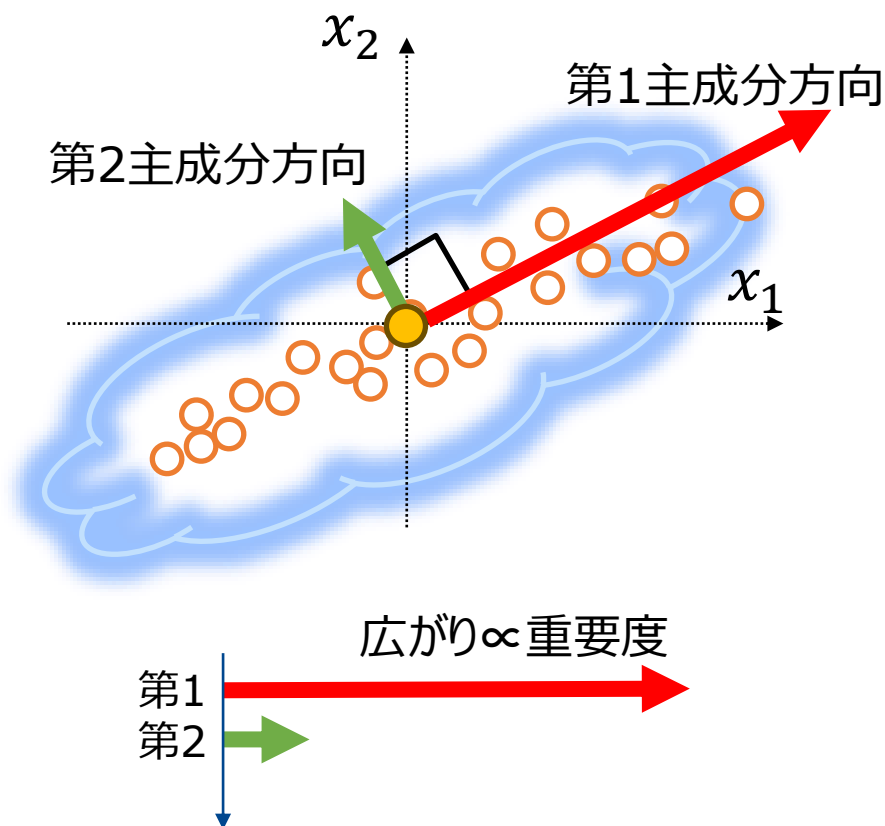
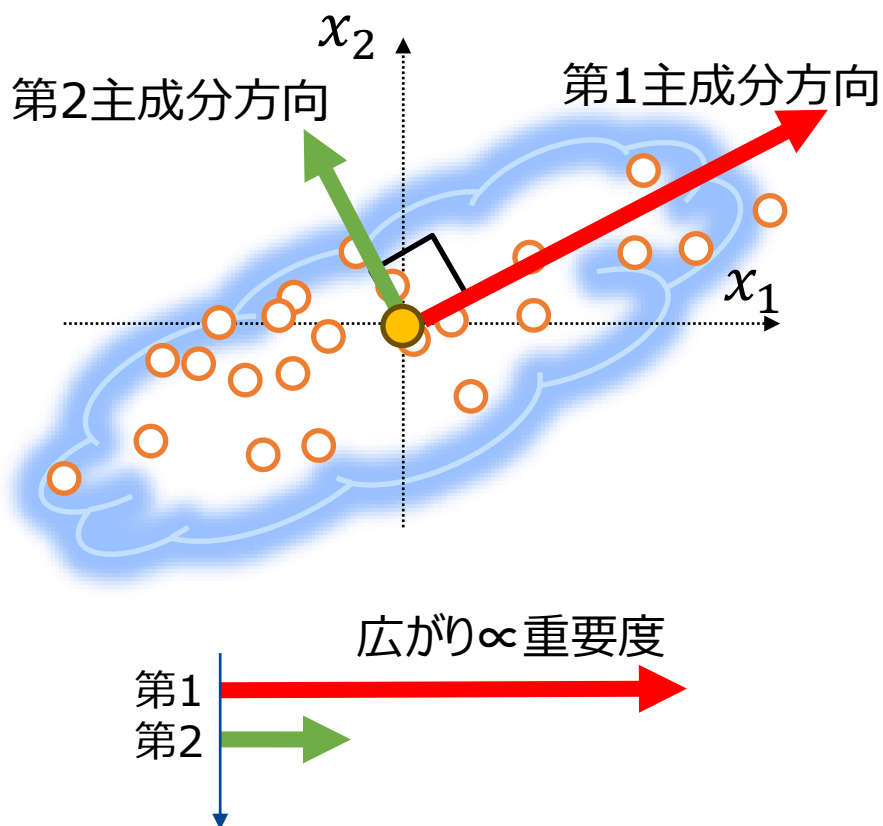
2番目は、第1主成分に直交する方向



- これを「第2主成分」と呼ぶ
- 元々が2次元($d = 2$)の場合はこれらを基底に選らんで終了

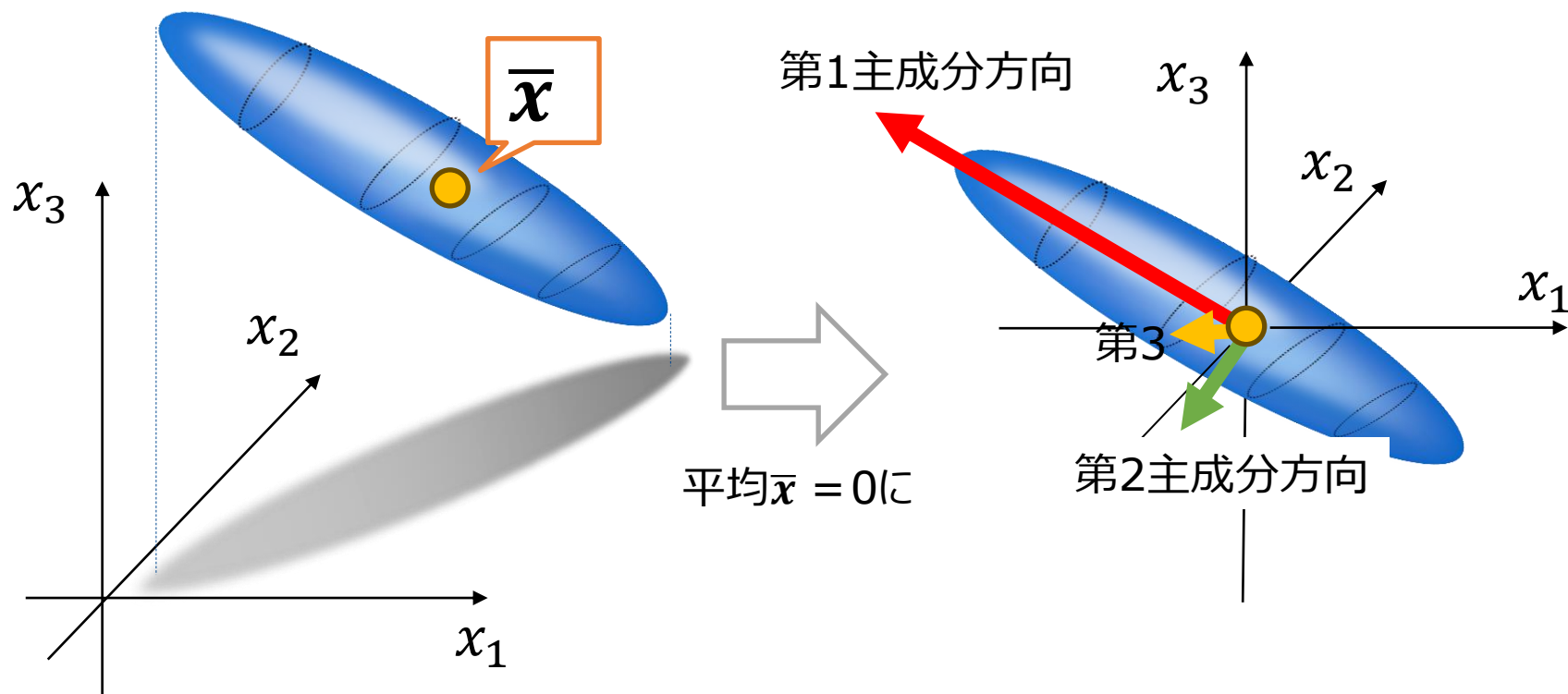
主成分ベクトルの長さ = 広がり具合 \propto 重要度

- 主成分は「方向」だけでなく、「重要度」も持っています
- その方向にデータがどれくらい広がっているか(分散) = 重要度



$d(\geq 3)$ 次元の場合は...

- 「直交し、かつ広がり大きい方向から」、第1主成分、第2主成分、 \dots 、第 d 主成分

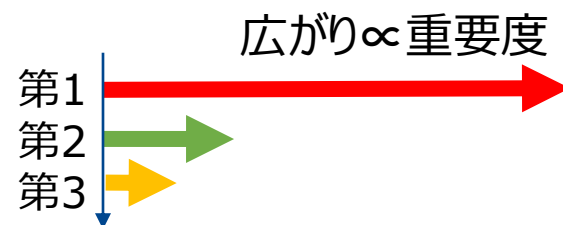
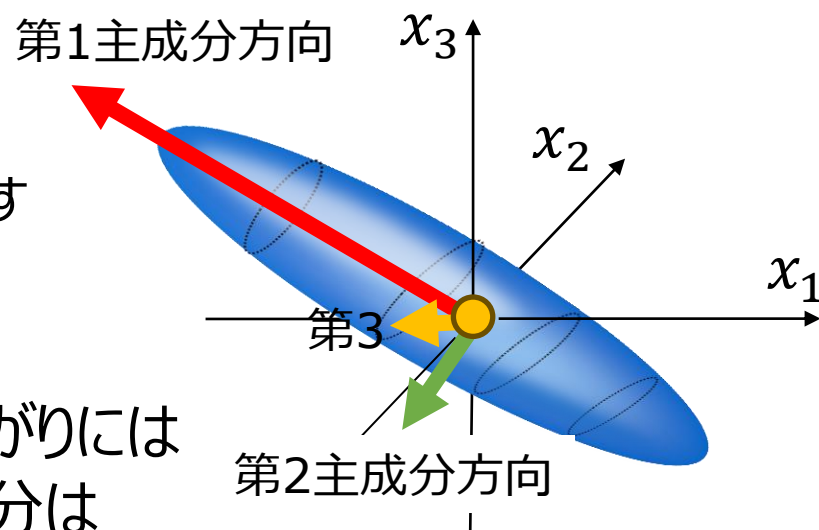


以上が「主成分分析の基本的考え方

主成分はいくつ求まる？

- d 次元ベクトル集合については、
 d 個の主成分が求まる
 - = 直交基底が自動的に求まるということです

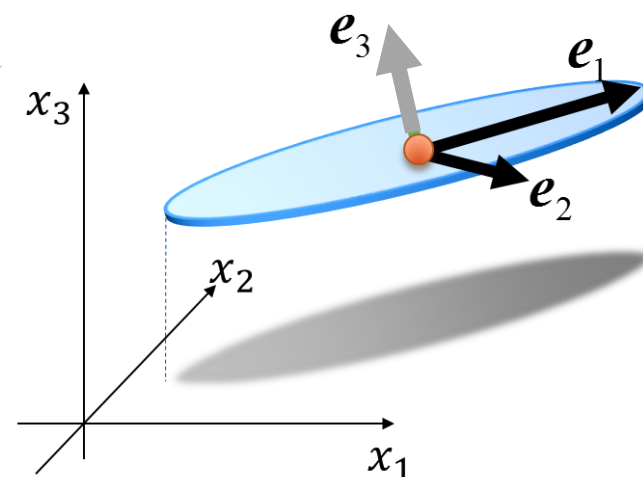
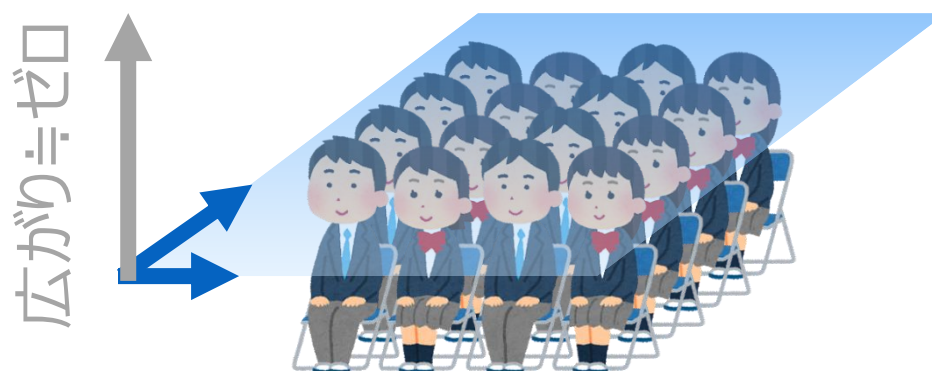
- ただし、データ(=ベクトル集合)の広がりには
偏りがあるので、重要度の高い主成分は
 d よりも少ない
 - = 「真の次元」の数ぐらいしか、重要な主成分は
存在しない



- よって、上位 \tilde{d} ($< d$)個の主成分のみを利用する場合が多い

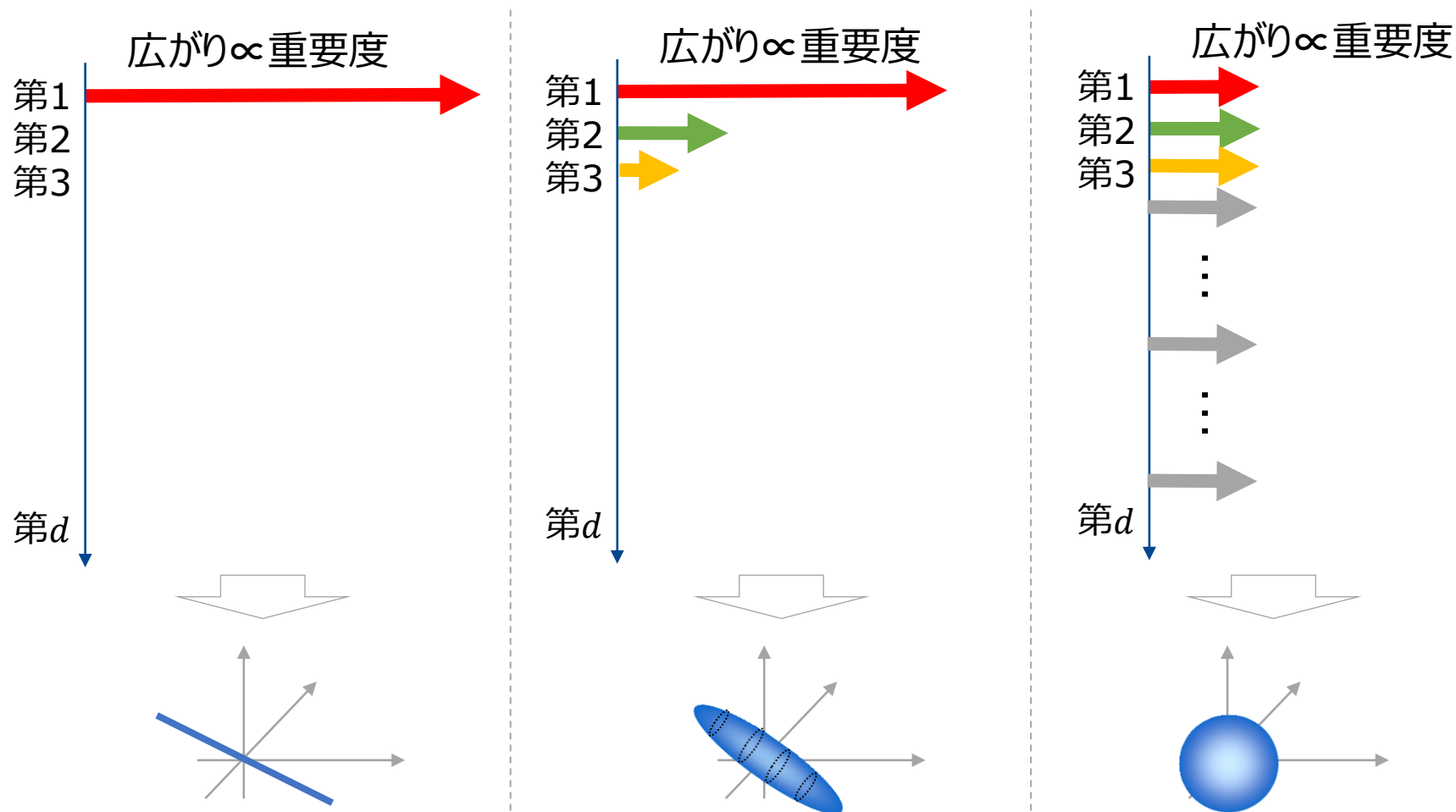
各主成分の重要度を見ると「分布の姿」や「真の次元」がわかる(1/2)

- 「広がり」がほぼゼロの主成分がたくさんあれば、それだけ真の次元は低いということ



- 重要度の変化を見れば、より細かいことまでわかる
 - 次スライド

各主成分の重要度を見ると「分布の姿」や「真の次元」がわかる(2/2)

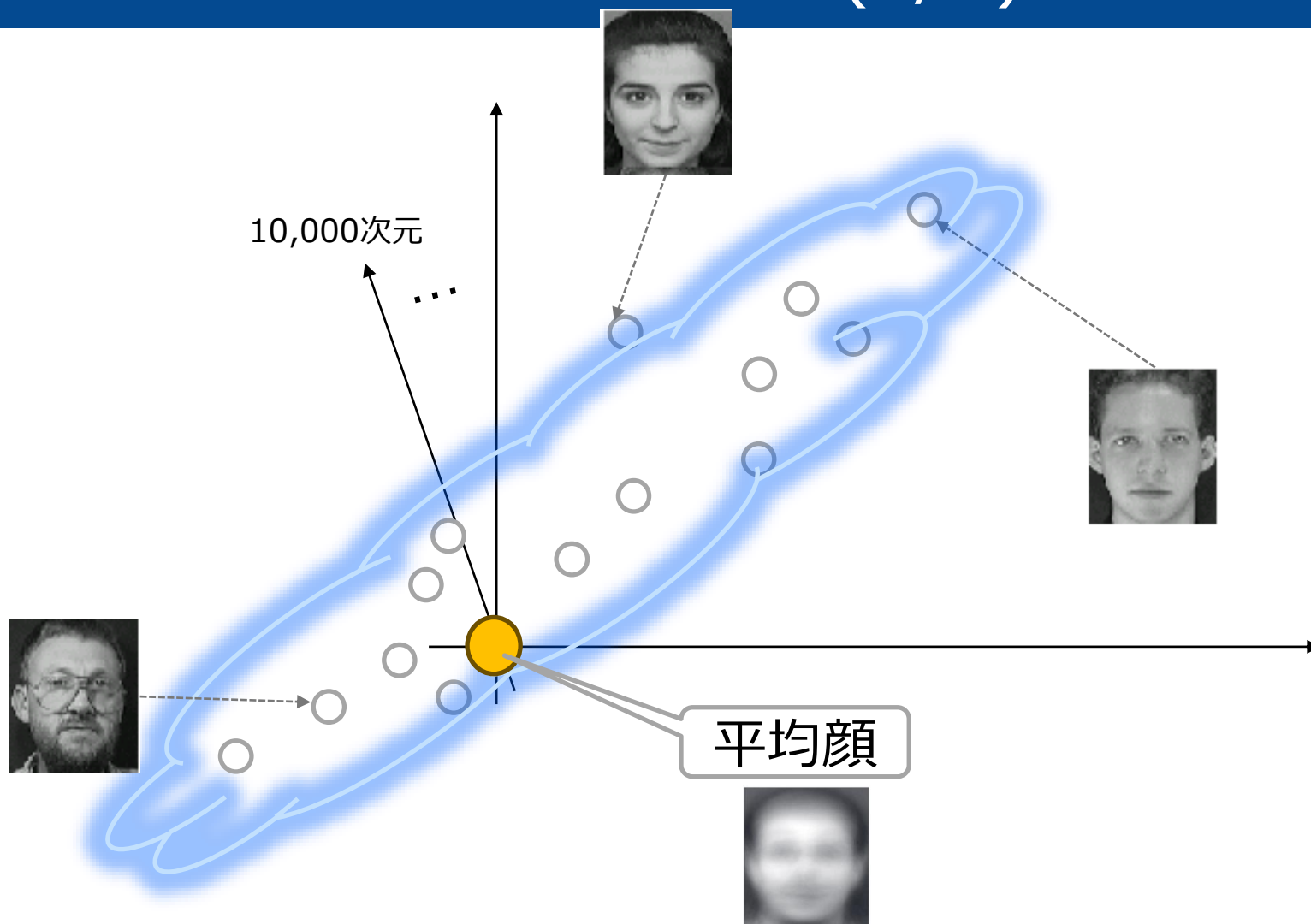


高次元データの広がりの様子を把握する重要なツール！

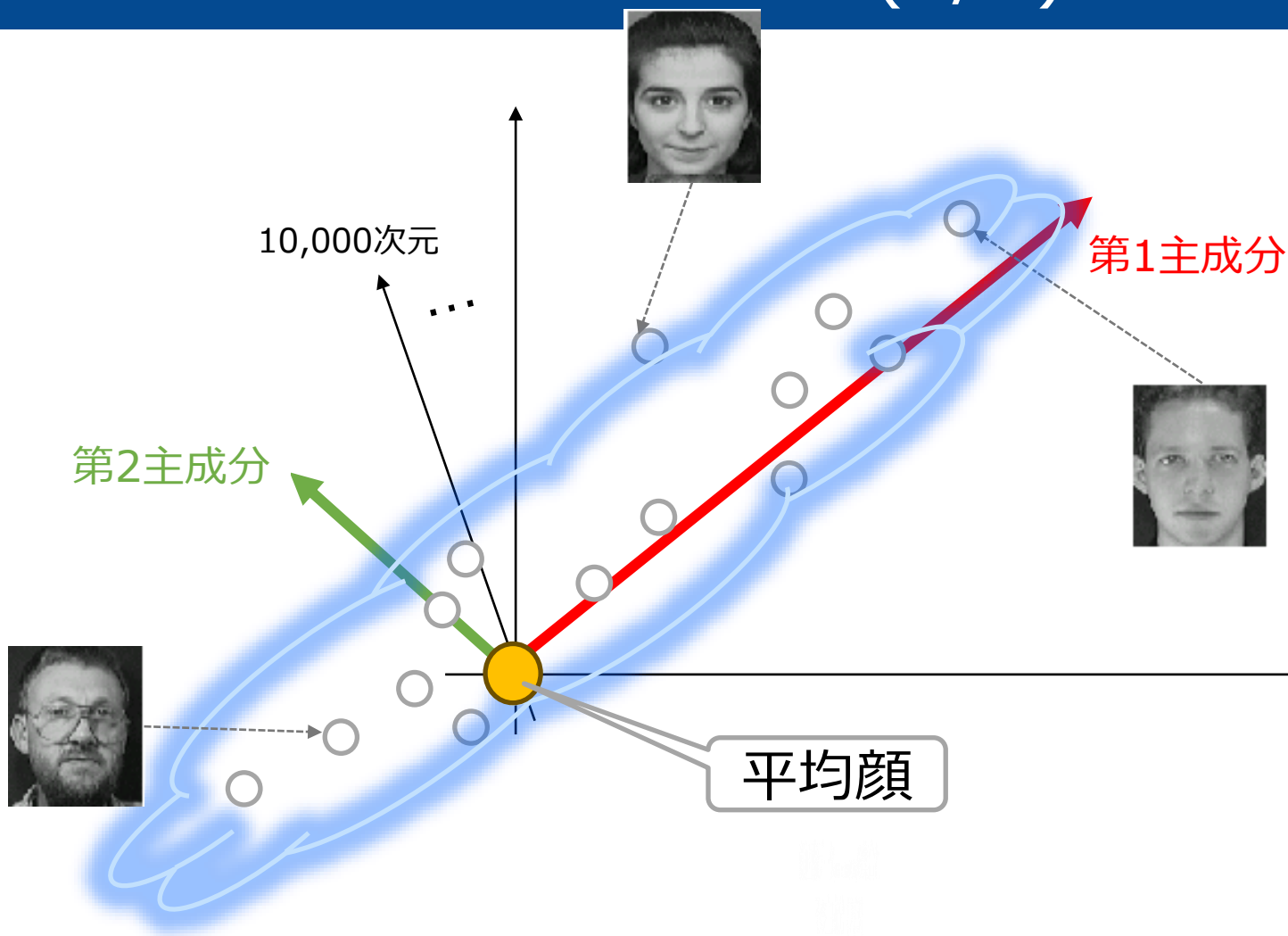
顔画像データ集合を例に
(高次元ベクトルを対象とした)
主成分分析の挙動を理解する

画像もベクトルなので主成分分析可能

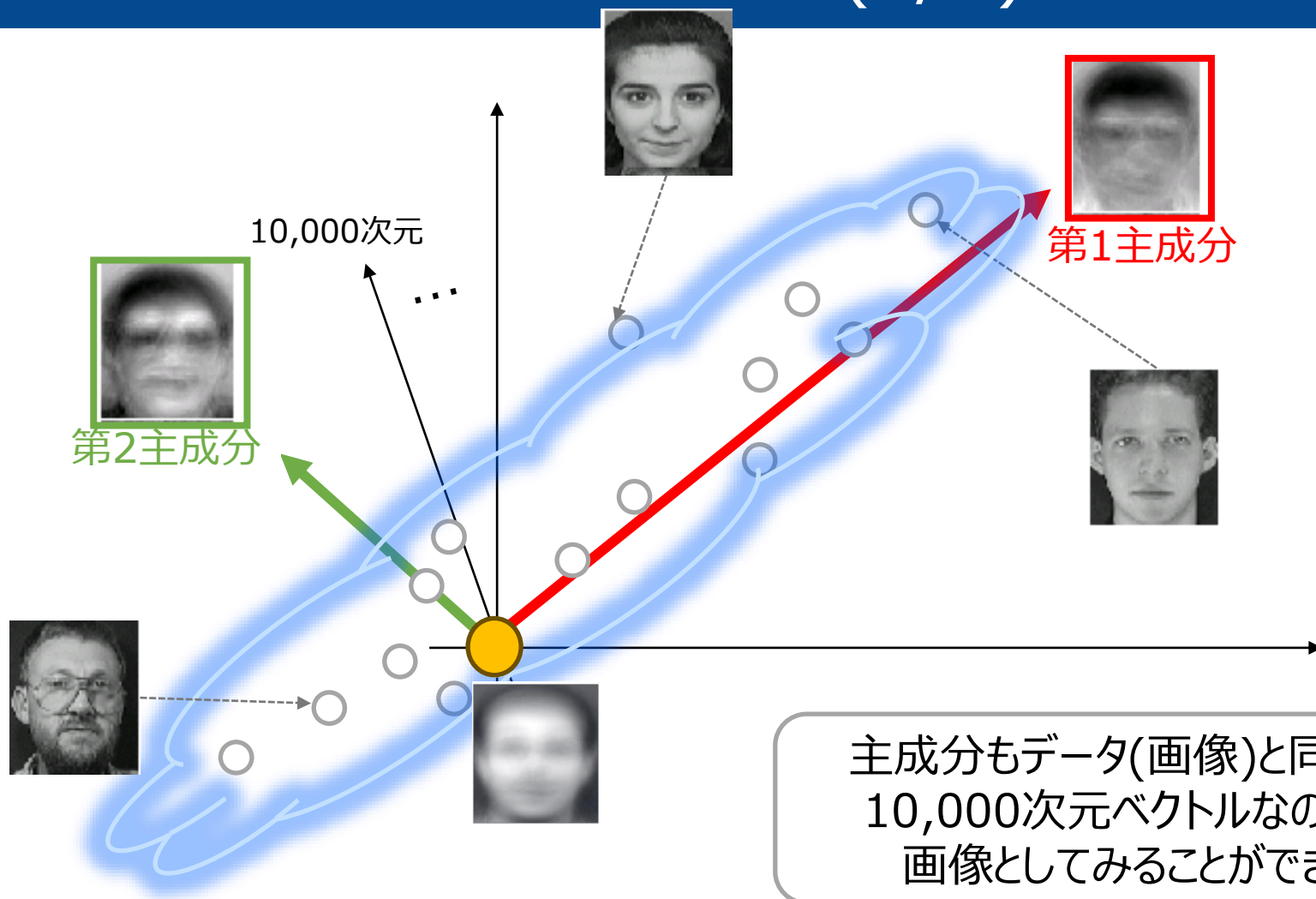
顔画像データ_(100 × 100画素)の集合と その主成分 (1/3)



顔画像データ_(100 × 100画素)の集合と その主成分 (2/3)



顔画像データ_(100 × 100画素)の集合と その主成分 (3/3)



顔画像に対する主成分



データ集合
(一部)



得られた主成分

上位
(重要)



中位



下位



徐々に主成分を加えていくと...

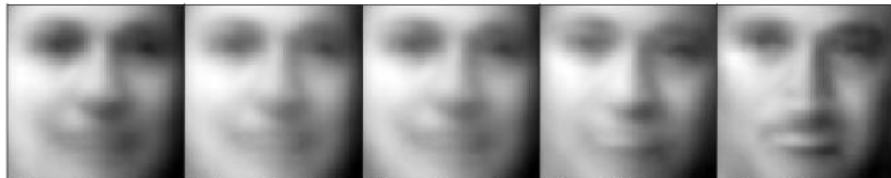
Sandipan Dey (UMBC)

#efaces=0, res=56.852



<https://sandipanweb.wordpress.com/2018/01/06/eigenfaces-and-a-simple-face-detector-with-pca-svd-in-python/>

#efaces=1, res=29.769 #efaces=2, res=27.586 #efaces=5, res=27.347 #efaces=10, res=23.01 #efaces=20, res=18.755



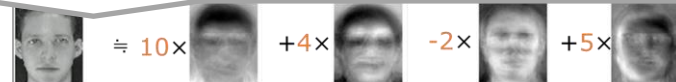
#efaces=40, res=15.416 #efaces=60, res=13.066 #efaces=80, res=11.821 #efaces=100, res=10.342 #efaces=150, res=8.813



#efaces=200, res=7.924 #efaces=300, res=6.626 #efaces=400, res=5.454 #efaces=1000, res=1.963 #efaces=1071, res=1.617



例に出した4つでは元には戻らなそう…
(ただし分類するためには、元に戻る必要はない)



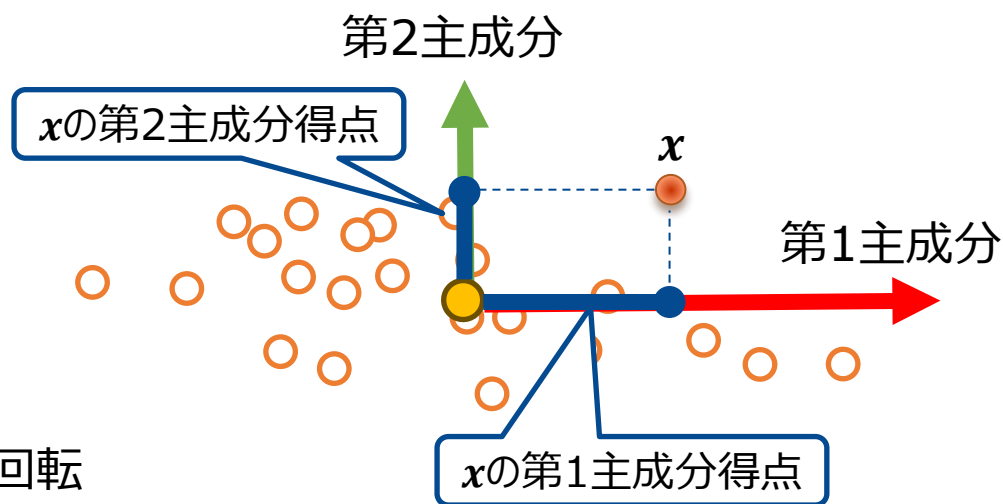
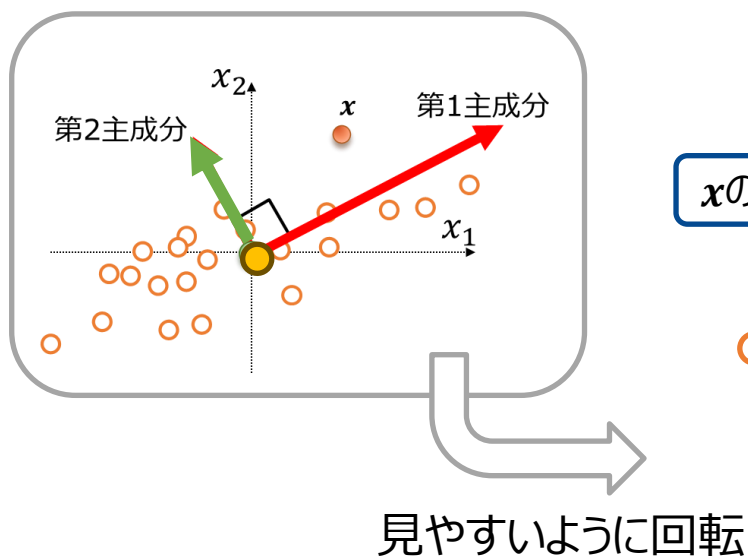
の特徴量 = (10, 4, -2, 5)^T

主成分分析でわかること

どうしてこんな大変な思いをしなくてはならないのか？

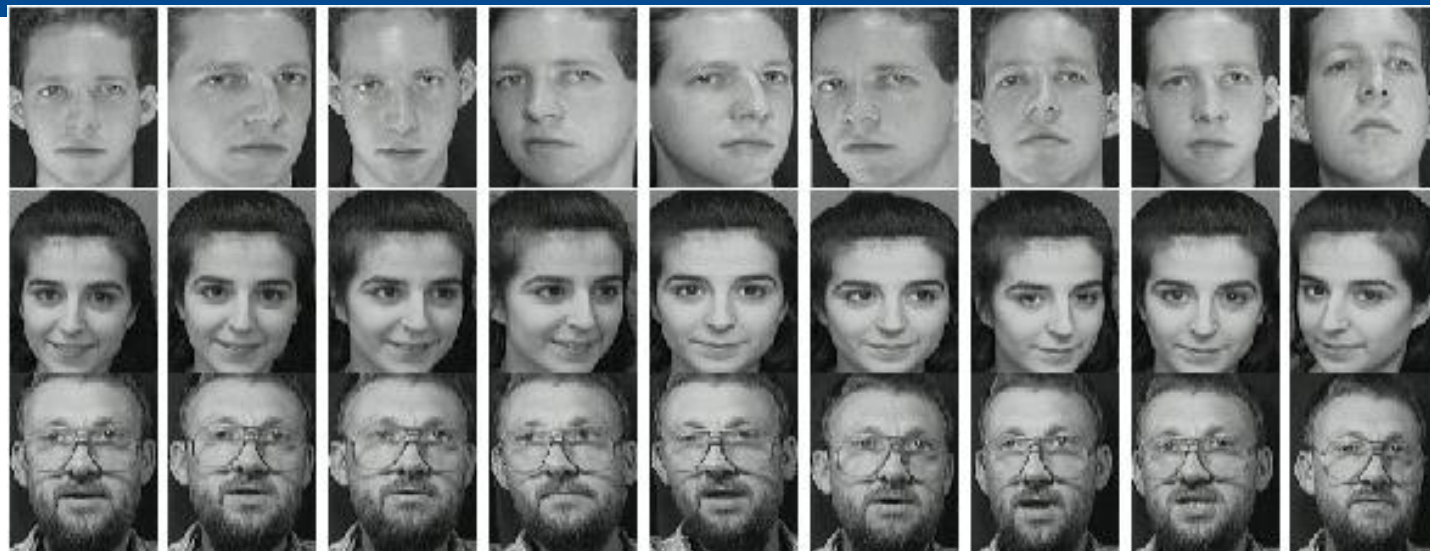
主成分得点

- 各データが、各主成分方向にどの程度成分量を持つか？
 - = 主成分を材料とみた場合の、各データの「レシピ」



【再掲】顔画像に対する主成分分析の結果

データ集合
(一部)



得られた主成分

上位
(重要)



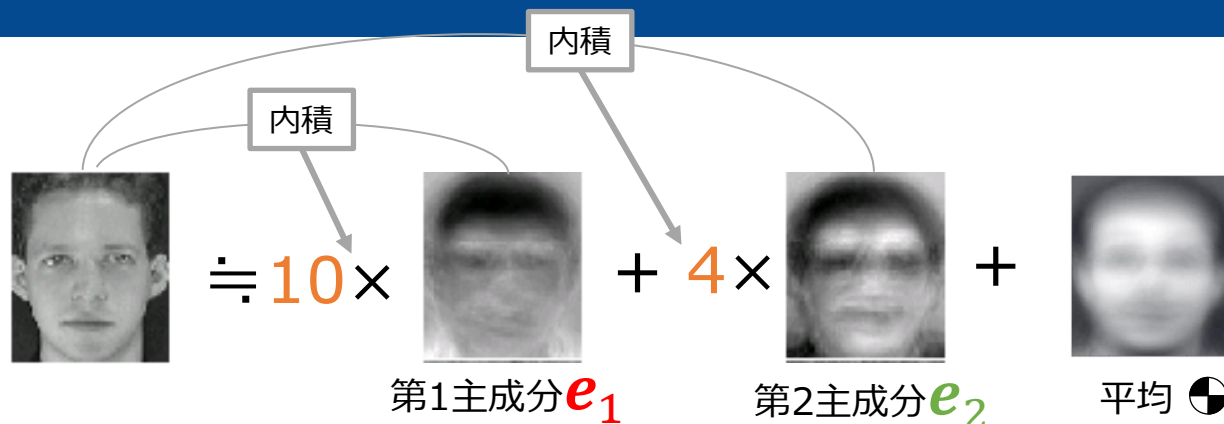
中位



下位

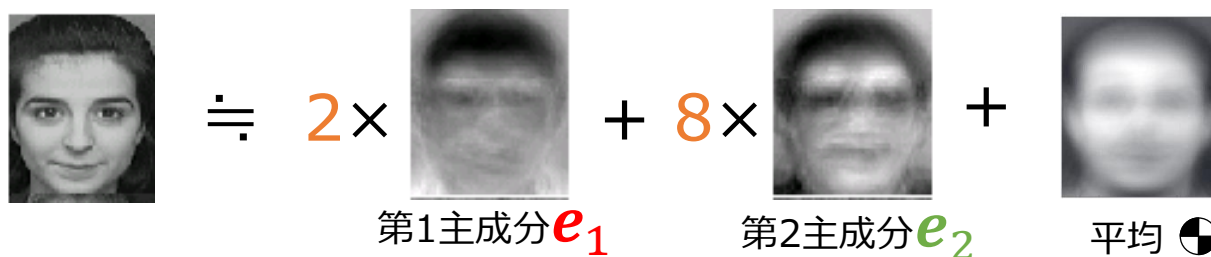


主成分得点で各データを表現



の主成分得点 = (10, 4)

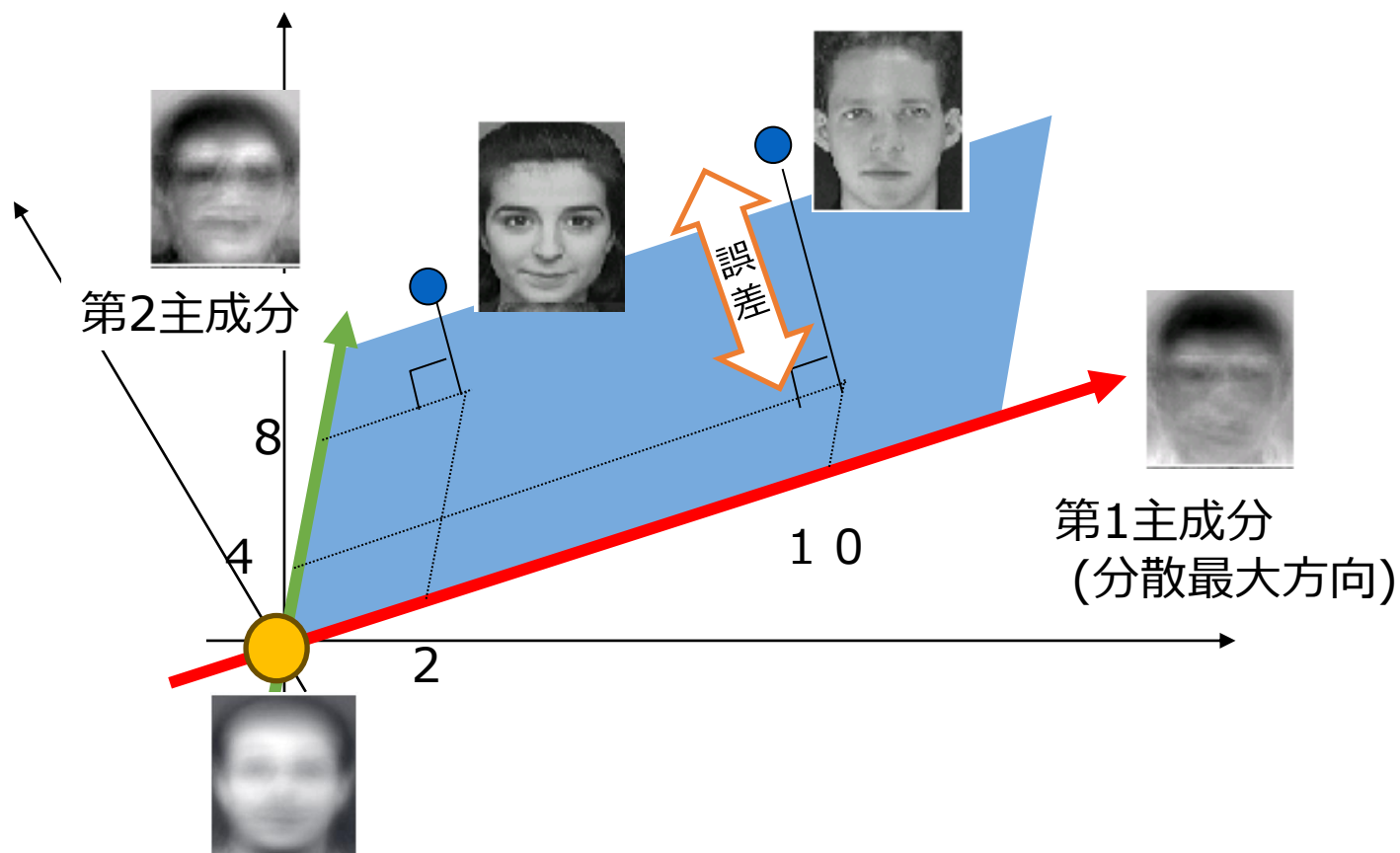
画像を2次元で
表現！



の主成分得点 = (2, 8)

画像を2次元で
表現！

こんな感じになってます



主成分を増やせば「誤差」を減らせます

$$\begin{aligned}
 & \text{顔画像} \doteq 10 \times \text{第1主成分} \mathbf{e}_1 + 4 \times \text{第2主成分} \mathbf{e}_2 - 2 \times \text{第3主成分} \mathbf{e}_3 + 5 \times \text{第4主成分} \mathbf{e}_4 + \text{平均}
 \end{aligned}$$

⇒ 顔画像の主成分得点 = (10, 4, -2, 5)

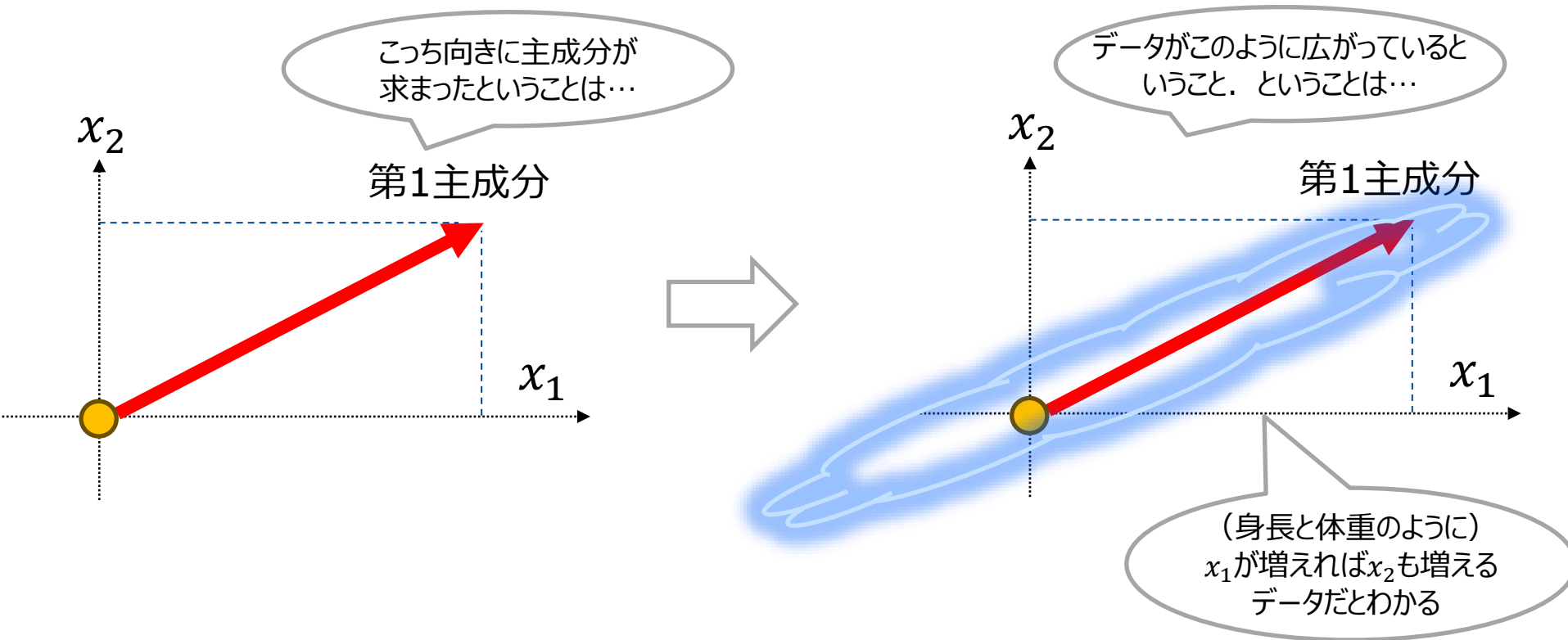
画像を4次元で表現！

$$\begin{aligned}
 & \text{顔画像} \doteq 2 \times \text{第1主成分} \mathbf{e}_1 + 8 \times \text{第2主成分} \mathbf{e}_2 - 11 \times \text{第3主成分} \mathbf{e}_3 - 8 \times \text{第4主成分} \mathbf{e}_4 + \text{平均}
 \end{aligned}$$

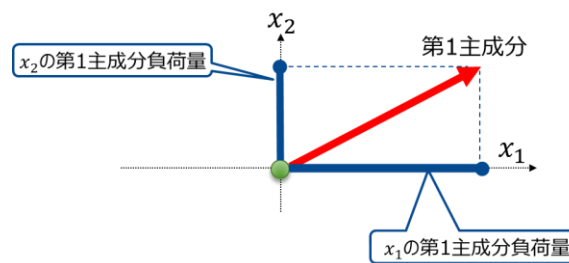
⇒ 顔画像の主成分得点 = (2, 8, -11, -8)

画像を4次元で表現！

主成分の向きは 要素間の相関を表す(1/3)

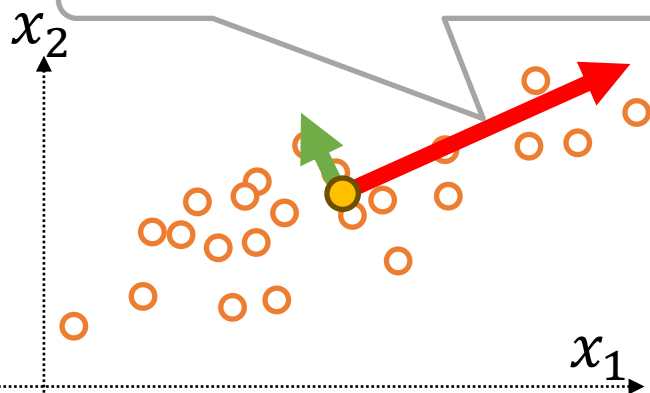


- 参考: 「主成分負荷量」という用語があります

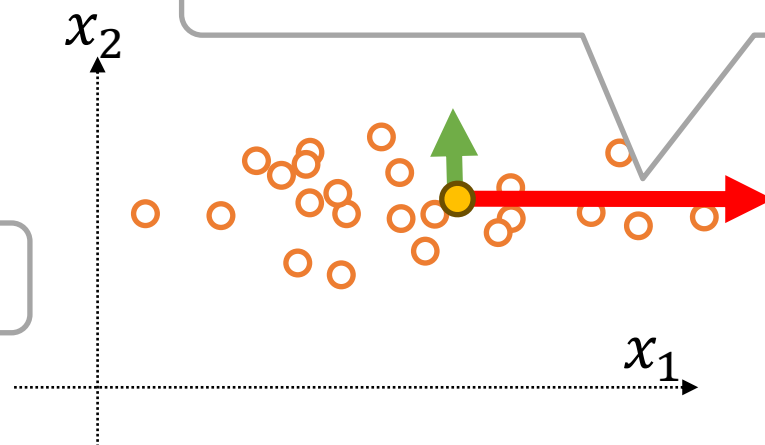


主成分の向きは 要素間の相関を表す(2/3)

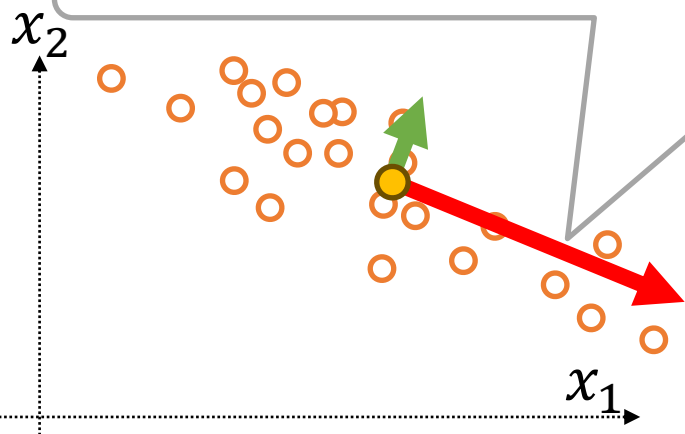
主成分が x_1 と x_2 の「正の相関」を示している



主成分が無相関を示している



主成分が x_1 と x_2 の「負の相関」を示している

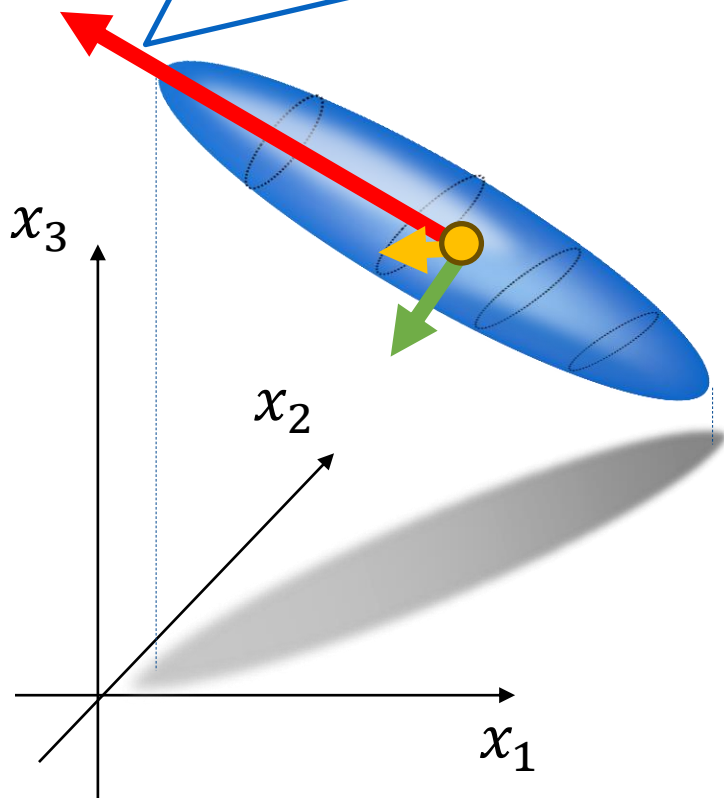


主成分の向きは 要素間の相関を表す(3/3)

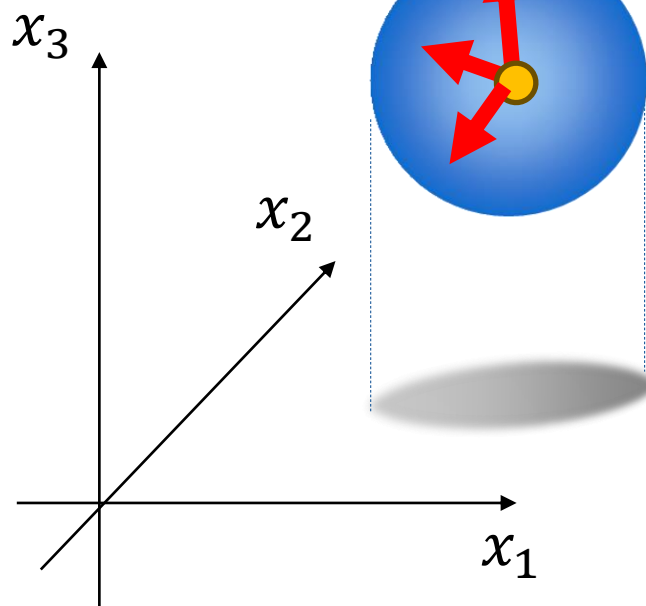
非常に広がっている方向があり(=偏りがあり),

$x_1 \rightarrow \text{大}, x_2 \rightarrow \text{大}, x_3 \rightarrow \text{小}$

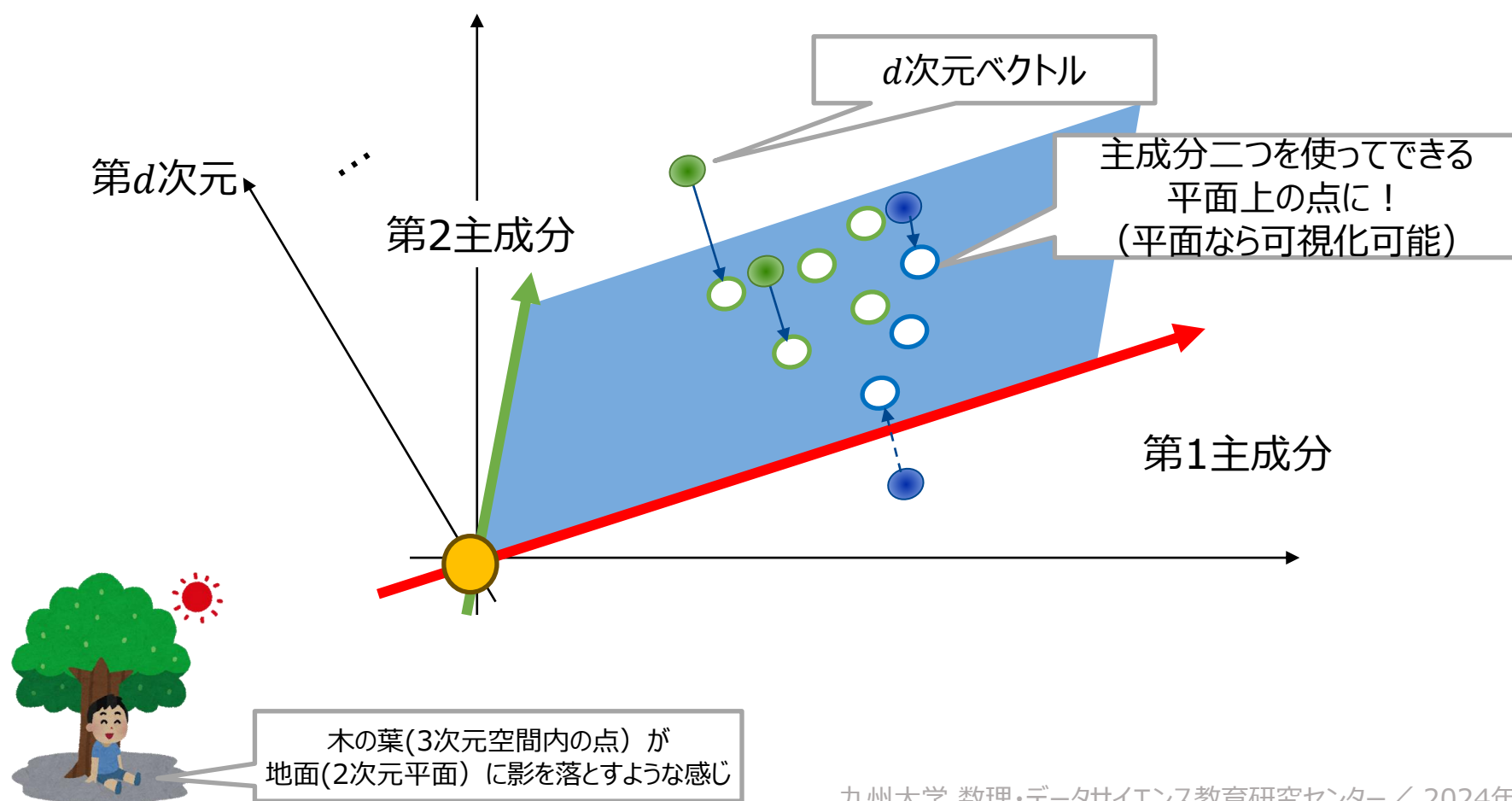
($x_1 \rightarrow \text{小}, x_2 \rightarrow \text{小}, x_3 \rightarrow \text{大}$, と等価)



広がり一様→球状分布
→無相関



データ分布の可視化にも使える

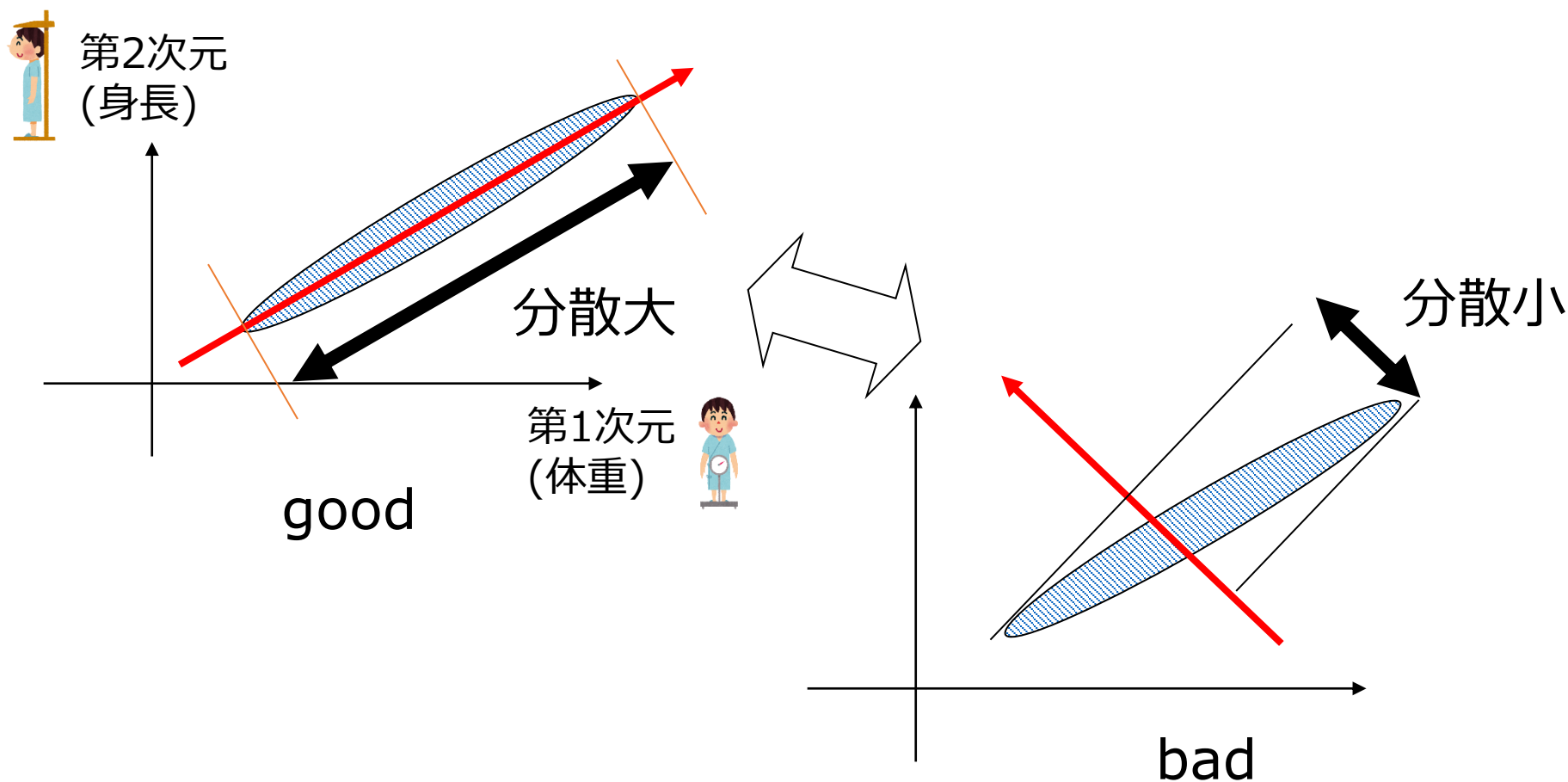


主成分を求める実際の方法

詳細略

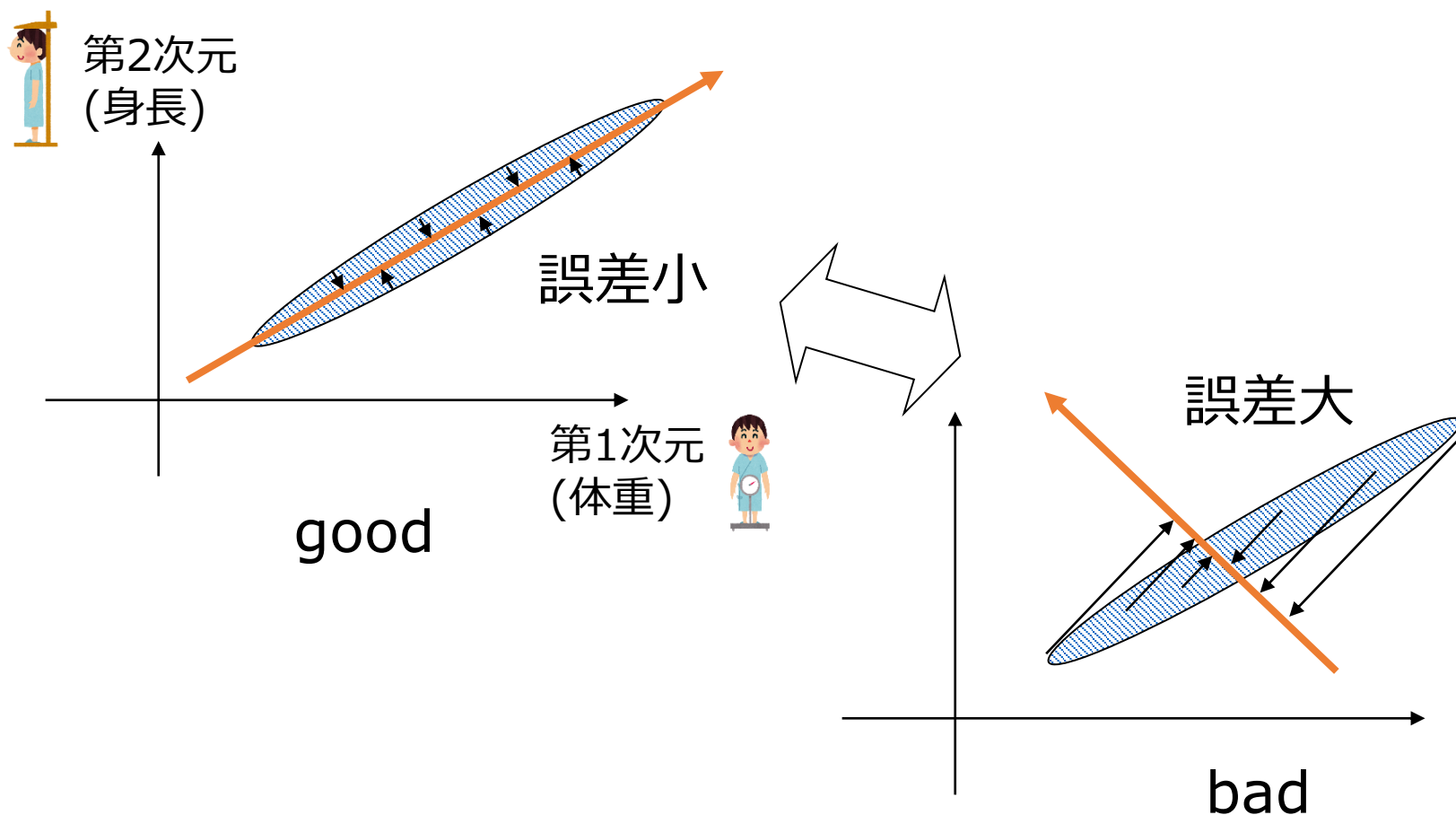
広がった方向を求めるための2つの基準(1): 分散最大基準

- なるべく分散の大きくなる方向に主成分を求めたい



広がった方向を求めるための2つの基準(2): 最小二乗誤差基準

- 「なるべく誤差が小さくなる方向に主成分を求めたい」

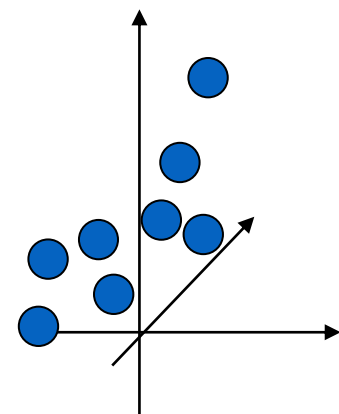


主成分分析の解法

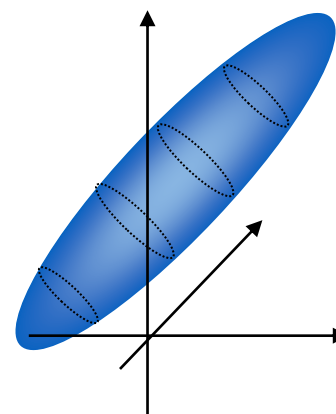
- 分散・最小誤差，どちらの基準でも，次のように解ける
 - まったく同じ主成分が求まる
 - というか，まったく同じ式に帰着するので，答え（＝主成分）も同じ
- 【参考】実際には以下の3ステップで求まる
 1. データ集合(各々 d 次元ベクトル)から共分散行列 Σ を求める
 2. Σ の固有値と固有ベクトルを求める
 3. 固有値の大きなものから \tilde{d} 個の固有ベクトルを主成分とする（ \tilde{d} は適当に決定）



解法の雰囲気を図解する



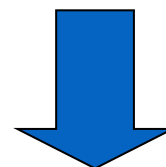
共分散行列
 Σ を求める



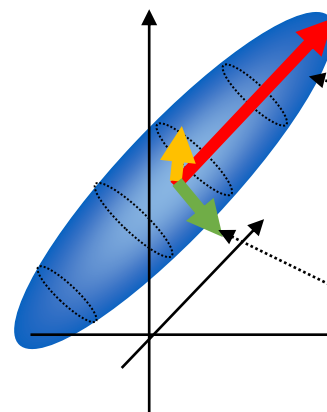
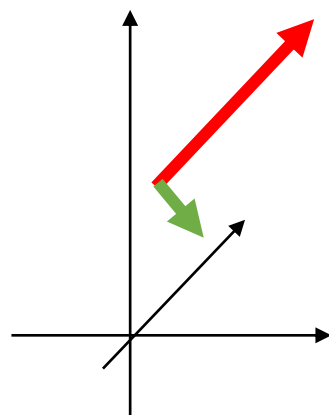
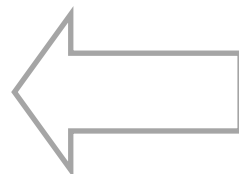
楕円球と
見てみる



固有値
固有ベクトル



上位 \tilde{d} 個だけ残す

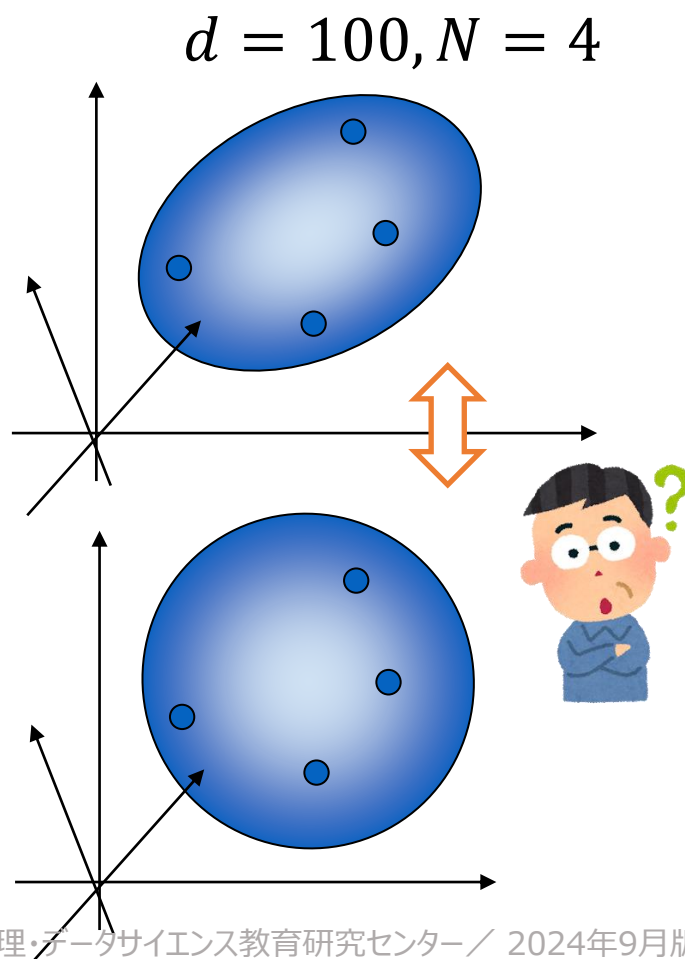
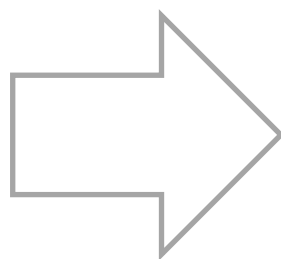
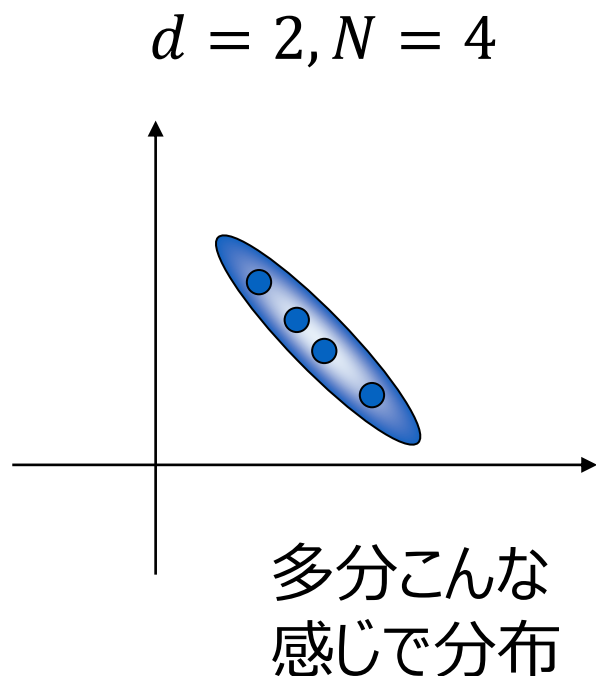


第1固有ベクトル (方向)
第1固有値 (広がり=重要度)

第2 "

主成分分析で注意すべき点

- 次元が高いのにデータが少ないと、分布の形がわからないので、求まる主成分も怪しい！
 - 可能なら次元数 d の10倍は欲しいところ...



分布状況の解析手段について, これまでのまとめ

色々な方法で分布状況を解析してきました

- 平均

- 分布の中心

- (各軸の)分散

- 各要素(各座標軸)での広がり具合

- クラスタリング

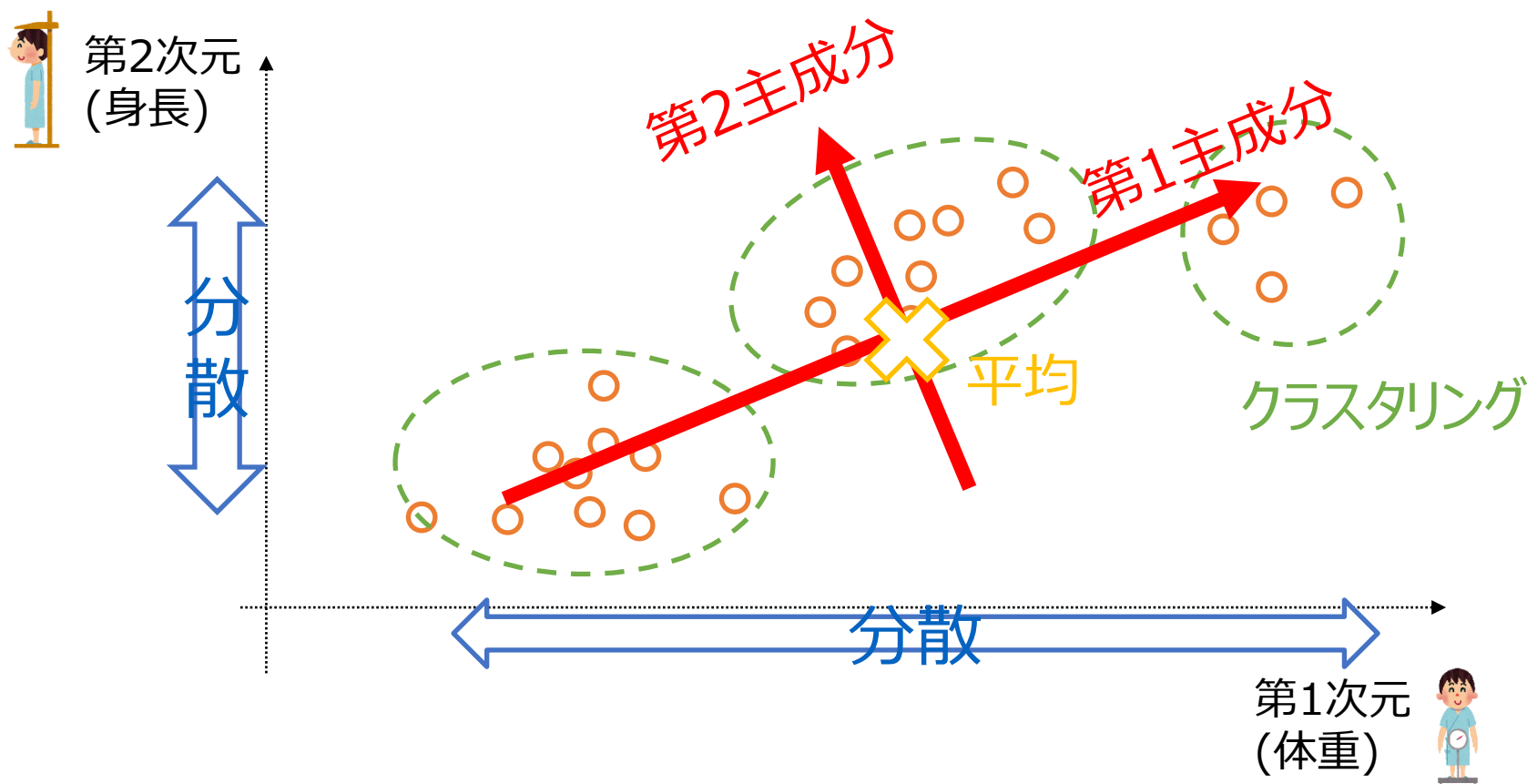
- 分布全体をグループに分ける

- 主成分

- 分布が最も広がっている方向 = 第一主成分
- 第一主成分に直交しつつ, 次に最も広がっている部分 = 第二
- 分布の「真の次元」もわかる

どれがいいとか
悪いとかではない.
みんな違って
みんないい.

それぞれを図示すると...
(図は2次元ですが、高次元でもできます)



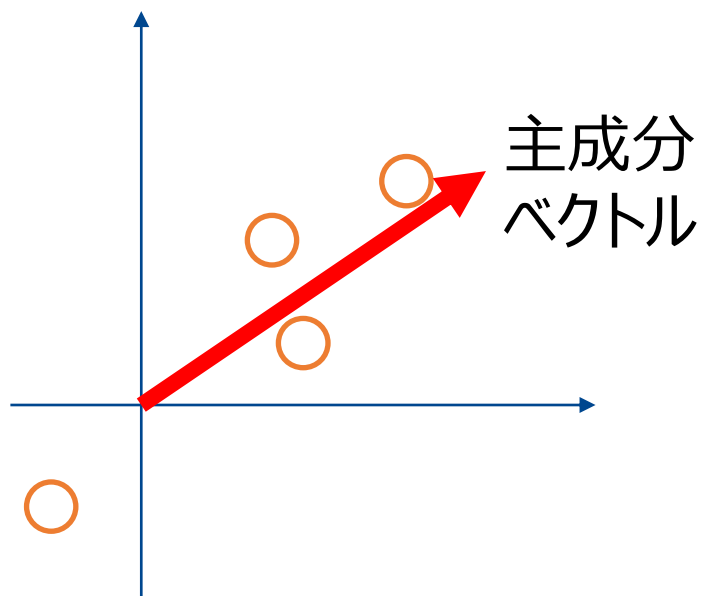
【付録 1】 データ集合の因子分析

主成分分析とは、似て非なるもの

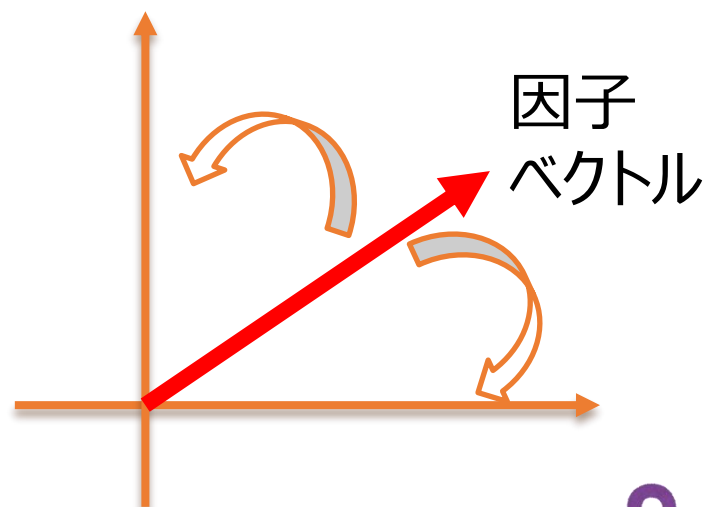
主成分分析 vs 因子分析

混同されがちだが、目的からして結構違う...

主成分分析 =
各データを主成分で
うまく表現するのが仕事



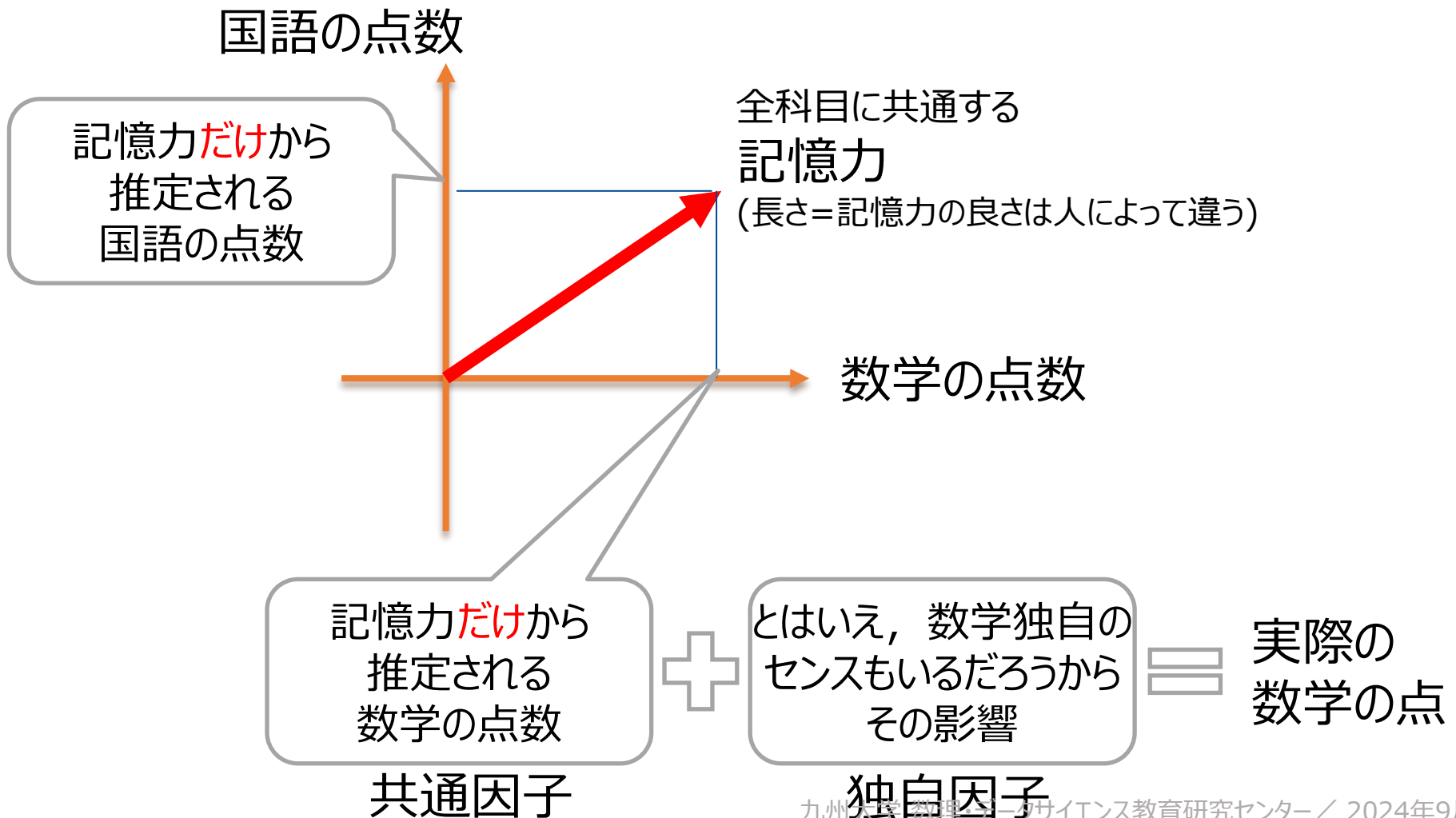
因子分析 =
各座標軸(要素)を因子で
うまく表現するのが仕事



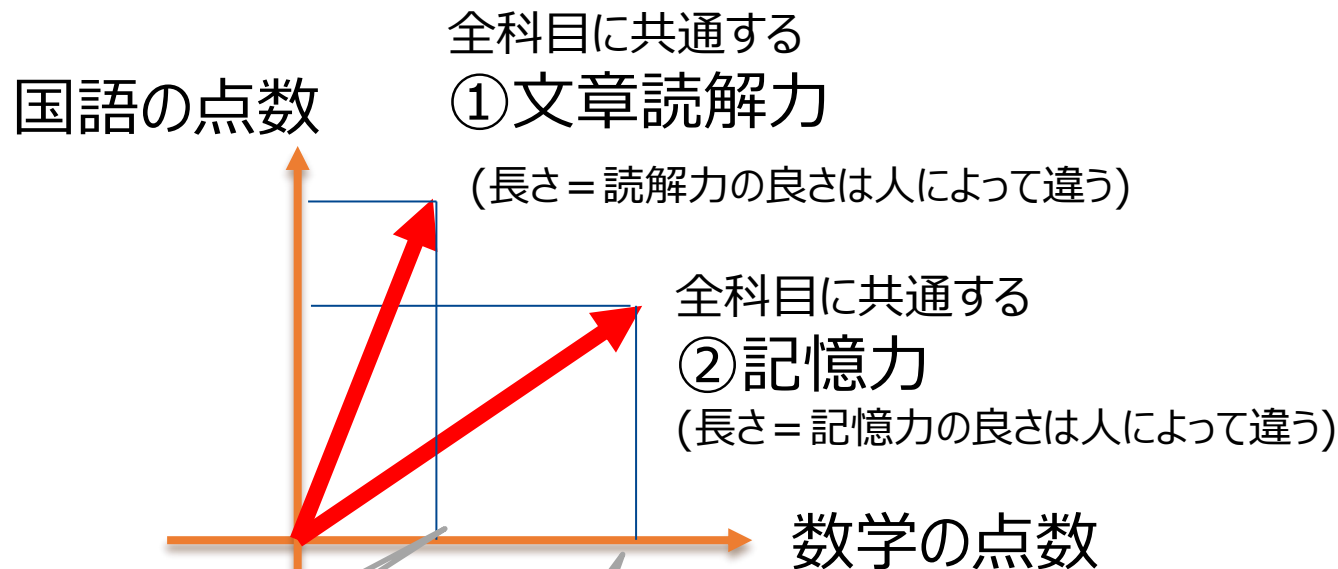
ん？わかったような
わからんような...



因子分析，よく出てくる例で説明



もしかしたら複数の共通因子があるかも



読解力^{だけ}から
推定される
数学の点数

共通因子①



記憶力^{だけ}から
推定される
数学の点数

共通因子②



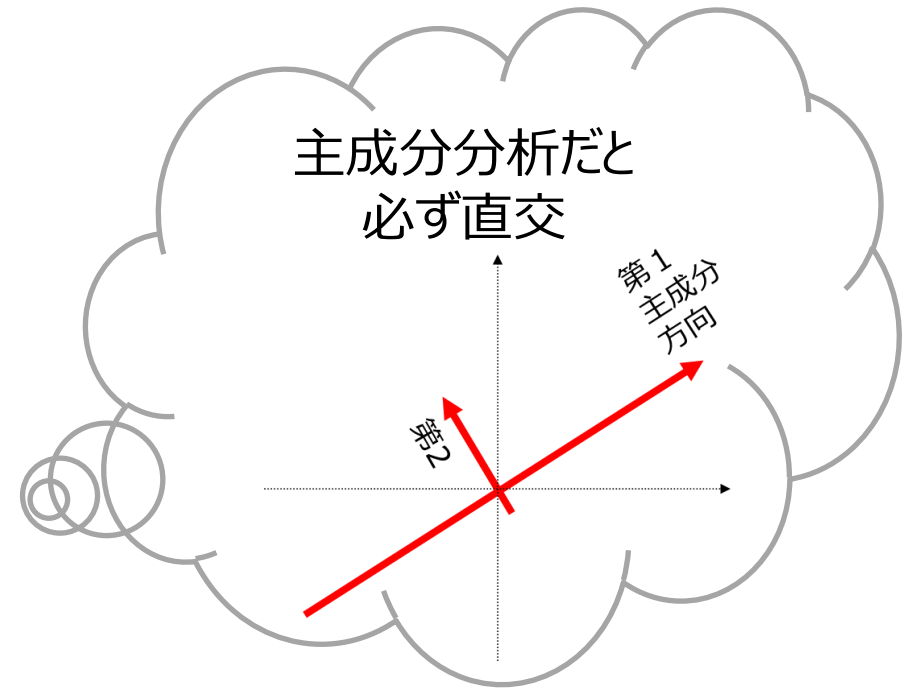
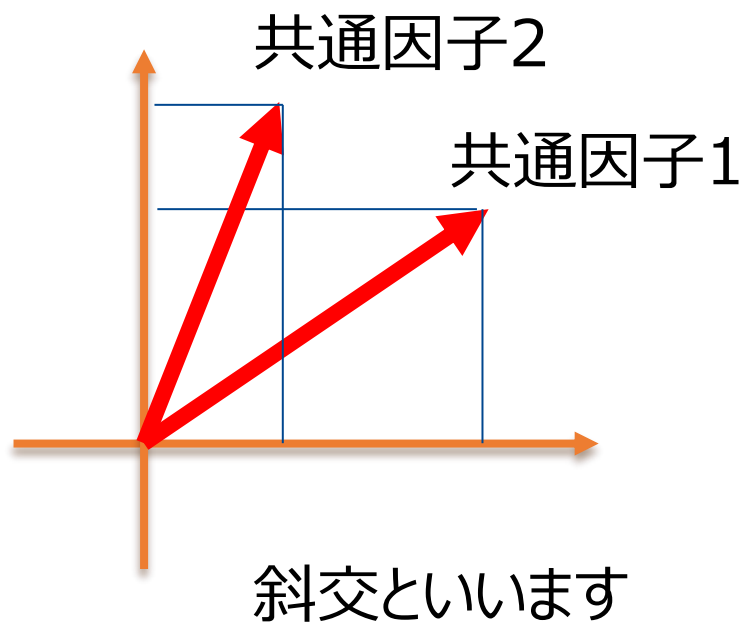
とはいえ、数学独自の
センスもいるだろうから
その影響

独自因子

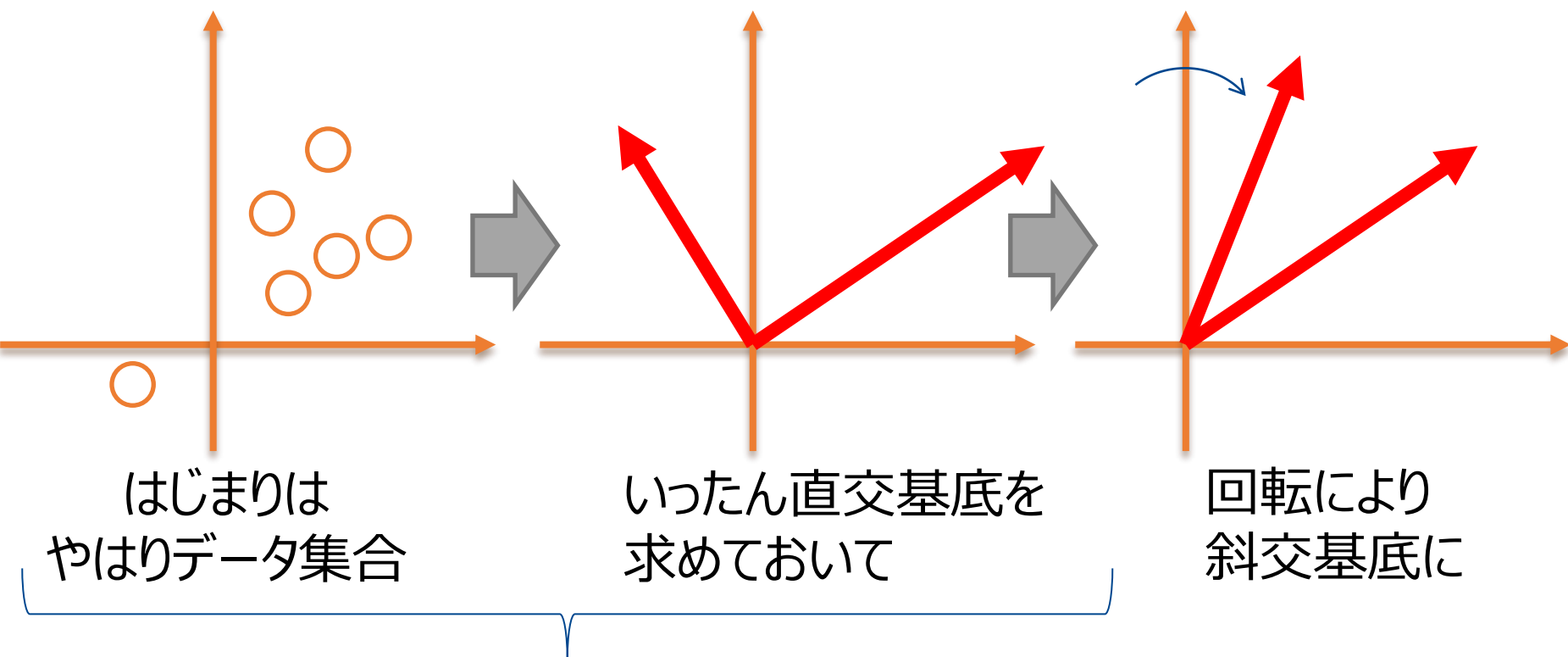


実際の
数学の点

共通因子は直交しなくてもよい



共通因子の求め方の例



この部分が主成分分析に似ている
(ただし因子分析では独自因子も考えるので、
実際には主成分分析とは違う)

参考：より違いを深く知りたい方へ

- <http://www.sigmath.es.osaka-u.ac.jp/~kano/research/seminar/30BSJ/kano.pdf>

主成分分析は因子分析ではない！

○狩野 裕

大阪大学 大学院人間科学研究科