R研究集会2023
統計数理研究所共同利用研究集会
「データ解析環境Rの整備と利用」

Rによる行列分解のいろいろ

2023年12月16日

安川武彦
株式会社JDSC

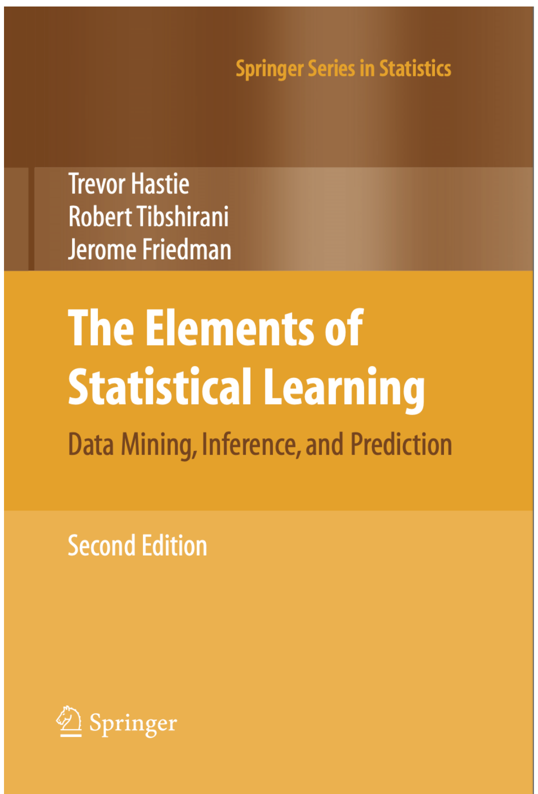
本日の内容

- 行列分解のいろいろ
- Archetypal Analysisの紹介
- 導出アルゴリズムの拡張
- 数値例
- まとめ

行列分解は教師なし学習のアプローチのひとつとして、捉えられ活用されている

Hastie, Tibshirani and Freidman(2009)では、14.6節でNMF, AAが紹介されている

https://hastie.su.domains/ElemStatLearn/



行列分解にはいくつものバリエーションがある

特異値分解・主成分分析とは異なる行列分解の例

手法	分解	
非負行列分解 Lee and Seung(1999,2001)	非負行列を非負の行列で近似	$\left(\underset{n\times m}{\boldsymbol{X}}\right) \approx \left(\underset{n\times p}{\boldsymbol{W}}\right)\left(\underset{p\times m}{\boldsymbol{H}}\right)$
CUR分解 Mahoney and Drineas(2012)	元行列の行と列を選択し近似	$\left(\underset{n\times m}{\boldsymbol{X}}\right) \approx \left(\underset{n\times r}{\boldsymbol{C}}\right)\left(\underset{r\times c}{\boldsymbol{U}}\right)\left(\underset{c\times m}{\boldsymbol{R}}\right)$
本日のメイン **Archetypal Analysis** Cutler and Breiman(1994)	行列を凸結合で近似	$\left(\underset{n\times m}{\boldsymbol{X}}\right) \approx \left(\underset{n\times r}{\boldsymbol{A}}\right)\left(\underset{r\times n}{\boldsymbol{B}}\right)\left(\underset{n\times m}{\boldsymbol{X}}\right)$

非負行列分解（NMF）は、非負制約の元でデータを次元縮約した２つの行列の内積に分解する

非負の要素のみからなる行列を非負制約の元で低次元の二つの行列に分解する。テキスト、音声、画像、推奨システムなどへの応用が行われている

論文

Lee, D. D. and H. S. Seung. (1999). Learning the parts of objects by non-negative matrix factorization, Nature, 401, 788-91.

定式化

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$$

$$\min_{\boldsymbol{W},\boldsymbol{H}} \; J(\boldsymbol{X}||\boldsymbol{W}\boldsymbol{H})$$

$$s.t. \; \boldsymbol{W} \geqslant 0, \boldsymbol{H} \geqslant 0$$

$$\left(\underset{n\times m}{\boldsymbol{X}} \right) \approx \left(\underset{n\times p}{\boldsymbol{W}} \right) \left(\underset{p\times m}{\boldsymbol{H}} \right)$$

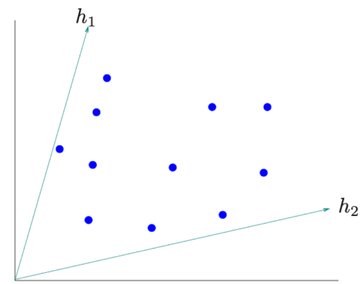
当初はL2ロス、KLダイバージェンスを最小化するアルゴリズムが提案されている

拡張応用

Nicolas Gillis (2021) Nonnegative Matrix Factorization, Society for Industrial & Applied Mathematics,U.S.　参照

R実装

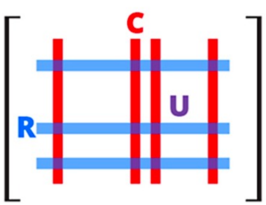
NMF http://renozao.github.io/NMF/
RcppML https://github.com/zdebruine/RcppML



Hastie, et. Al(2009)

CUR分解は、データ行列の行・列を用いて低次元の行列で明示的な低次元近似を行う

データ行列の行と列を用いた低次元近似を行う。PCAやSVDの代替として用例ベースの手法として提案された

論文	M. W. Mahoney and P. Drineas. “CUR matrix decompositions for improved data analysis.” Proc. Natl. Acad. Sci. USA 106 (2009), pp. 697–702.
定式化	$\boldsymbol{X} \approx \boldsymbol{C}\boldsymbol{U}\boldsymbol{R}$ $\left(\underset{n\times m}{\boldsymbol{X}}\right) \approx \left(\underset{n\times r}{\boldsymbol{C}}\right)\left(\underset{r\times c}{\boldsymbol{U}}\right)\left(\underset{c\times m}{\boldsymbol{R}}\right)$ 最初のアイデアは、importance samplingによりCとRを見つけようとするもの。サンプリングの方法は様々な提案がある  Hamm and Huang(2019)
拡張応用	最近では、Tensor分解への拡張が行われている
R実装	rCUR http://cran.nexr.com/web/packages/rCUR/

Archetypal Analysisは、データ行列が特徴的な観測値の組み合わせにより近似する。以降詳しくみていく

極端な観測値（アーキタイプ）の凸結合として行列を近似する。アーキタイプに基礎付けられたデータ構造がある場合に有用

論文

Cutler A, Breiman L (1994). “Archetypal Analysis.” Technometrics, 36(4), 338–347.

定式化

$$\begin{pmatrix} \quad \underset{n\times m}{\boldsymbol{X}} \quad \end{pmatrix} \approx \begin{pmatrix} \underset{n\times r}{\boldsymbol{A}} \end{pmatrix} \begin{pmatrix} \quad \underset{r\times n}{\boldsymbol{B}} \quad \end{pmatrix} \begin{pmatrix} \quad \underset{n\times m}{\boldsymbol{X}} \quad \end{pmatrix}$$

$$\boldsymbol{X}_{n\times m} \approx \boldsymbol{A}_{n\times r}\boldsymbol{B}_{r\times n}\boldsymbol{X}_{n\times m}$$

$$A \geqslant 0 \quad and \quad \sum_r a_{ir} = 1, \quad B \geqslant 0 \quad and \quad \sum_i b_{ri} = 1$$

$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z}$: データ行列をアーキタイプ（$\boldsymbol{Z}$）の凸結合で表現

$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{X}$: アーキタイプをデータの凸結合で表現

拡張 応用

ロバスト、区間データ、アルゴリズムの改善などの拡張が行われている
力学系、画像、テキスト、マーケティングなどへの応用が行われている

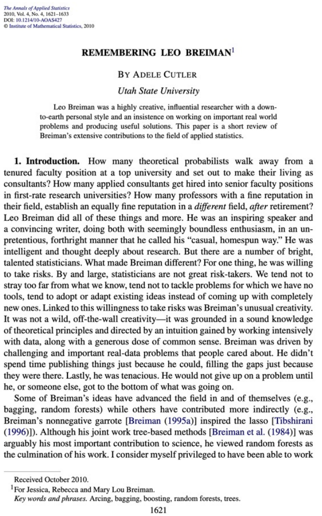
R実装

archetype https://cran.r-project.org/web/packages/archetypes/

Archetypal Analysisのアイデアはオゾン予測から来ている

Cutler(2010) Remembering Leo Breiman, The Annals of Applied Statistics, Vol. 4, No. 4, 1621–1633

- Adele Cutlerの博士論文（指導教官はLeo Breiman）
- Breiman先生のコンサルタント時代のアイデアが元
  - 翌日のオゾン予測のために、各日を「極端な」もしくは「典型的な」の混合として表現
  - ほとんどの日は極端な日（アーキタイプ）の中間に位置
  - クラスター分析や主成分分析の代用を想定
- NMFと類似の形式であるが使用目的が異なると思う（感想）



The Annals of Applied Statistics
2010, Vol. 4, No. 4, 1621–1633
DOI: 10.1214/10-AOAS427

REMEMBERING LEO BREIMAN[1]

BY ADELE CUTLER

Utah State University

Leo Breiman was a highly creative, influential researcher with a down-to-earth personal style and an insistence on working on important real world problems and producing useful solutions. This paper is a short review of Breiman's extensive contributions to the field of applied statistics.

1. Introduction. How many theoretical probabilists walk away from a tenured faculty position at a top university and set out to make their living as consultants? How many applied consultants get hired into senior faculty positions in first-rate research universities? How many professors with a fine reputation in their field, establish an equally fine reputation in a *different* field, *after* retirement? Leo Breiman did all of these things and more. He was an inspiring speaker and a convincing writer, doing both with seemingly boundless enthusiasm, in an unpretentious, forthright manner that he called his "casual, homespun way." He was intelligent and thought deeply about research. But there are a number of bright, talented statisticians. What made Breiman different? For one thing, he was willing to take risks. By and large, statisticians are not great risk-takers. We tend not to stray too far from what we know, tend not to tackle problems for which we have no tools, tend to adopt or adapt existing ideas instead of coming up with completely new ones. Linked to this willingness to take risks was Breiman's unusual creativity. It was not a wild, off-the-wall creativity—it was grounded in a sound knowledge of theoretical principles and directed by an intuition gained by working intensively with data, along with a generous dose of common sense. Breiman was driven by challenging and important real-data problems that people cared about. He didn't spend time publishing things just because he could, filling the gaps just because they were there. Lastly, he was tenacious. He would not give up on a problem until he, or someone else, got to the bottom of what was going on.

Some of Breiman's ideas have advanced the field in and of themselves (e.g., bagging, random forests) while others have contributed more indirectly (e.g., Breiman's nonnegative garrote [Breiman (1995a)] inspired the lasso [Tibshirani (1996)]). Although his joint work tree-based methods [Breiman et al. (1984)] was arguably his most important contribution to science, he viewed random forests as the culmination of his work. I consider myself privileged to have been able to work

Received October 2010.

[1]For Jessica, Rebecca and Mary Lou Breiman.

Key words and phrases. Arcing, bagging, boosting, random forests, trees.

Cutler and Breiman(1994)におけるオゾンデータの分析結果

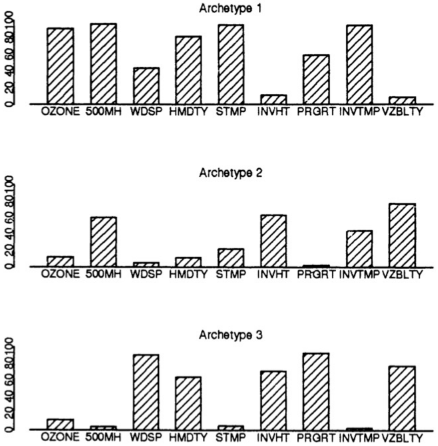


Figure 4. Percentile Profiles of Air-Pollution Archetypes.

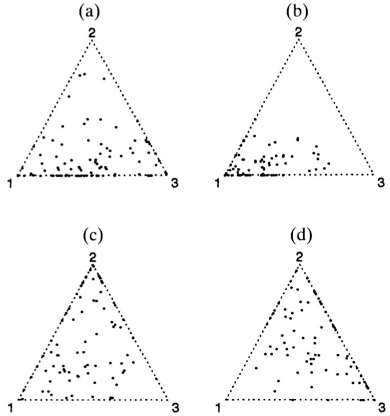
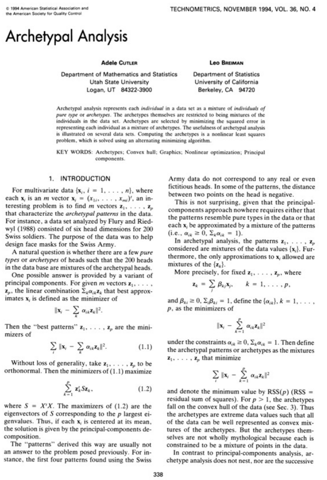


Figure 5. Mixture Plots for Air-Pollution Archetypes. (a) Spring; (b) Summer; (c) Fall; (d) Winter.



TECHNOMETRICS, NOVEMBER 1994, VOL. 36, NO. 4

Archetypal Analysis

Adele CUTLER
Department of Mathematics and Statistics
Utah State University
Logan, UT 84322-3900

Leo BREIMAN
Department of Statistics
University of California
Berkeley, CA 94720

Archetypal analysis represents each *individual* in a data set as a mixture of *individuals of pure type* or *archetypes*. The archetypes themselves are restricted to being mixtures of the individuals in the data set. Archetypes are selected by minimizing the squared error in representing each individual as a mixture of archetypes. The usefulness of archetypal analysis is illustrated on several data sets. Computing the archetypes is a nonlinear least squares problem, which is solved using an alternating minimizing algorithm.

KEY WORDS: Archetypes; Convex hull; Graphics; Nonlinear optimization; Principal components.

1. INTRODUCTION

For multivariate data $\{\mathbf{x}_i, i = 1, \ldots, n\}$, where each $\mathbf{x}_i$ is an m vector $\mathbf{x}_i = (x_{1i}, \ldots, x_{mi})'$, an interesting problem is to find m vectors $\mathbf{z}_1, \ldots, \mathbf{z}_p$ that characterize the *archetypal patterns* in the data. For instance, a data set analyzed by Flury and Riedwyl (1988) consisted of six head dimensions for 200

Archetypal Analysisの問題設定

L2損失関数を最小化する問題として交互最小２乗アルゴリズムを用いて、制約付最適化問題を解いている(Cutler and Breiman, 1994)

定式化	$$\boldsymbol{X}_{n\times m} \approx \boldsymbol{A}_{n\times r}\boldsymbol{B}_{r\times n}\boldsymbol{X}_{n\times m}$$ $$= \boldsymbol{A}_{n\times r}\boldsymbol{Z}_{r\times m}$$ 混合係数　アーキタイプ
損失関数	$$\mathcal{J}(\boldsymbol{A},\boldsymbol{B}) = \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{B}\boldsymbol{X}\|_F^2$$ $$= \sum_{i=1}^{n}\left\|\boldsymbol{x}_i - \sum_{r=1}^{r} a_{ik}\sum_{j=1}^{n} b_{kj}\boldsymbol{x}_j\right\|^2$$
最適化問題	$$\min \mathcal{J}(\boldsymbol{A},\boldsymbol{B})$$ $$\text{s.t.}\quad \boldsymbol{A} \geqslant 0, \sum_k a_{ik} = 1, \quad \boldsymbol{B} \geqslant 0, \sum_i b_{ki} = 1$$

$$\begin{pmatrix} \\ \underset{n\times m}{\boldsymbol{X}} \\ \\ \end{pmatrix} \approx \begin{pmatrix} \\ \underset{n\times r}{\boldsymbol{A}} \\ \\ \end{pmatrix}\begin{pmatrix} \underset{r\times n}{\boldsymbol{B}} \end{pmatrix}\begin{pmatrix} \\ \underset{n\times m}{\boldsymbol{X}} \\ \\ \end{pmatrix}$$

Cutler and Breiman(1994)の交互最小２乗アルゴリズムでは、AとZを交互に最適化している

- Zを固定した上で損失関数Jを最小にするAを求め、Aを既知として損失関数Jを最小にするBを求める
- 損失関数最小化にあたっては制約条件を反映させるために、凸最小 2 乗法を罰則付きに拡張したアルゴリズムが用いられている

Algorithm 2.1 algorithm for archetypal analysis

```
given the number of archetypes r
initialize matrices Z, A and B, and compute J_{t=0}
repeat
    arg min_{a_i} ||x_i − Σ_{k=1}^{p} a_{ik} z_k||^2 s.t. a_i ≥ 0 and Σ_{k=1}^{p} a_k = 1
    calculate z̃_i solveing x_i = Σ_i a_{ik} z̃_i
    arg min_{b_i} ||z̃_i − Σ_{k=1}^{p} b_{ik} x_i||^2 s.t. b_i ≥ 0 and Σ_{i=1}^{n} b_i = 1
    caluculate z_i = Σ_{i=1}^{n} b_{ki} x_i
    caluculate J_t
until |J_{t+1} − J_t| ≤ ε
return A, B
```

given the number of archetypes r

initialize matrices Z, A and B, and compute $\mathcal{J}_{t=0}$

repeat

$\quad \underset{\boldsymbol{a}_i}{\arg\min} \|x_i - \sum_{k=1}^{p} a_{ik} z_k\|^2$ s.t. $\boldsymbol{a}_i \geq 0$ and $\sum_{k=1}^{p} \boldsymbol{a}_k = 1$

$\quad$ calculate $\tilde{\boldsymbol{z}}_i$ solveing $x_i = \sum_i a_{ik} \tilde{\boldsymbol{z}}_i$

$\quad \underset{\boldsymbol{b}_i}{\arg\min} \|\tilde{\boldsymbol{z}}_i - \sum_{k=1}^{p} b_{ik} x_i\|^2$ s.t. $\boldsymbol{b}_i \geq 0$ and $\sum_{i=1}^{n} \boldsymbol{b}_i = 1$

$\quad$ caluculate $z_i = \sum_{i=1}^{n} b_{ki} x_i$

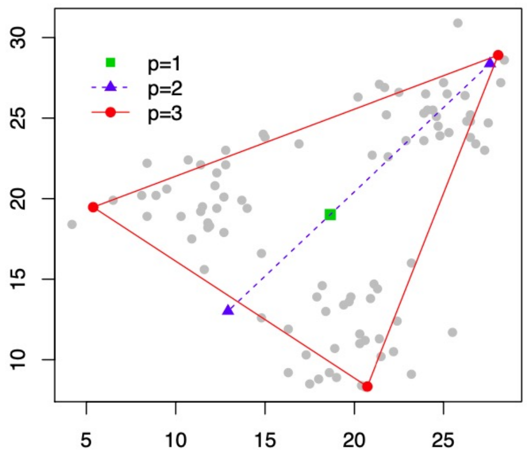
$\quad$ caluculate $\mathcal{J}_t$

until $|\mathcal{J}_{t+1} - \mathcal{J}_t| \leq \epsilon$

return A, B

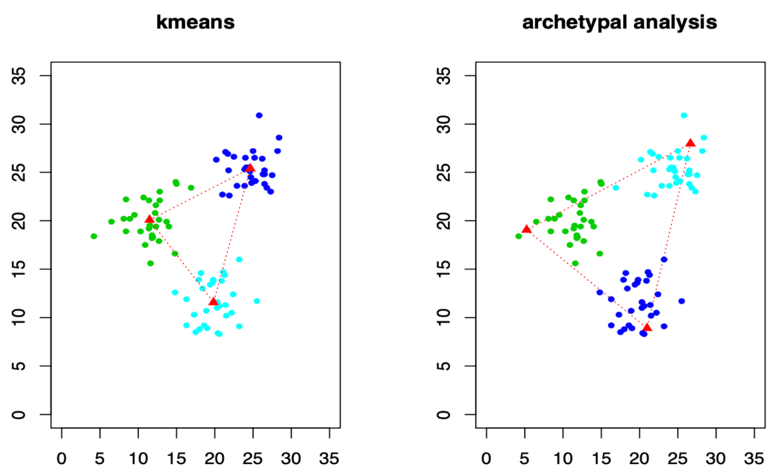
Archetypal Analysisでは、データのアーキタイプを見つけ、凸包でデータを近似する

2次元のシミュレーションデータでアーキタイプ分析を実施



- データを加工点（アーキタイプ）を1,2,3としたときの結果をプロット
- データ全体は、アーキタイプの凸結合として表現される

同じデータにk-meansを適用した場合と比較




```
library(archetypes)

pdat.aa<- archetypes(pdat, 3, maxIterations=3000)

plot(pdat,type="n", xlim=c(0, 35), ylim=c(0,35),xlab="", ylab="")

points(pdat, col="gray", pch=20)

points(pdat.aa[[1]], pch=17, col=2)

lines(pdat.aa[[1]][c(1,2,3,1),], pch=17, col=2, lty=3)

title("archetypal analysis")
```

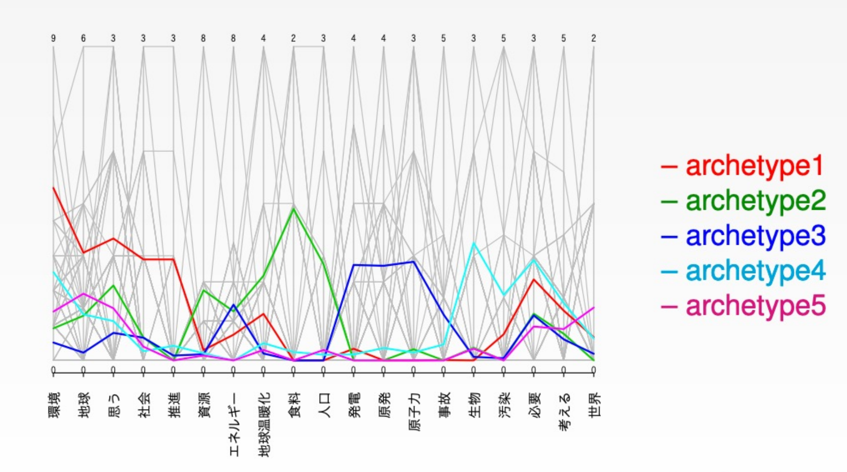
Archetypal Analysisの応用(1/2)

- 旭硝子財団 地球環境アンケート 2011 (http://www.af-info.or.jp/)
- 前処理を行って113 文書 × 121 語の文書単語行列を作成。件数データであるため、KLダイバージェンスを損失関数としてアーキタイプを抽出
- Topicモデルの代替としの利用ができるが、他の手法（LDA、NMF）と比べたアドバンテージは今のところ明確ではない

アーキタイプ(BX)の特徴語の抽出

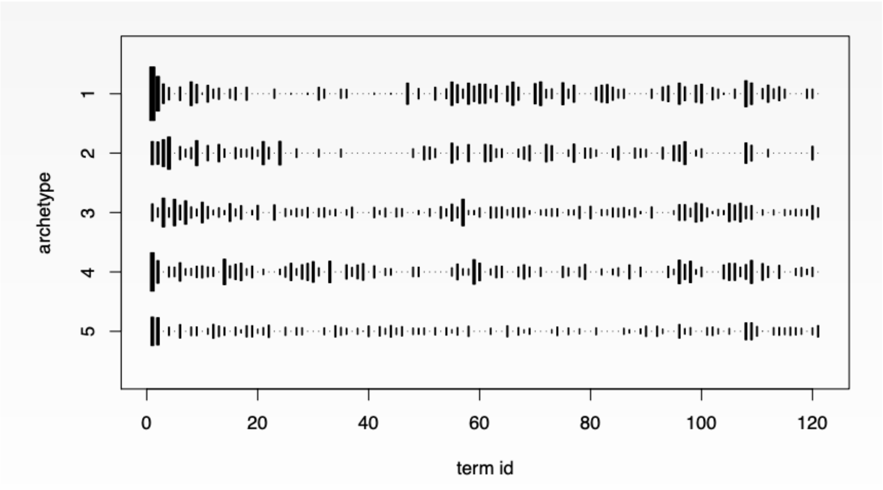
atype1	atype2	atype3	atype4	atype5
環境	資源	エネルギー	環境	環境
地球	エネルギー	発電	生物	地球
思う	地球温暖化	原発	汚染	考える
社会	食料	原子力	必要	思う
推進	人口	事故	考える	世界

アーキタイプ(BX)の特徴語のpcpプロット



アーキタイプ(BX)のhinton plot

推定されたアーキタイプの語の重みを相対化して表示



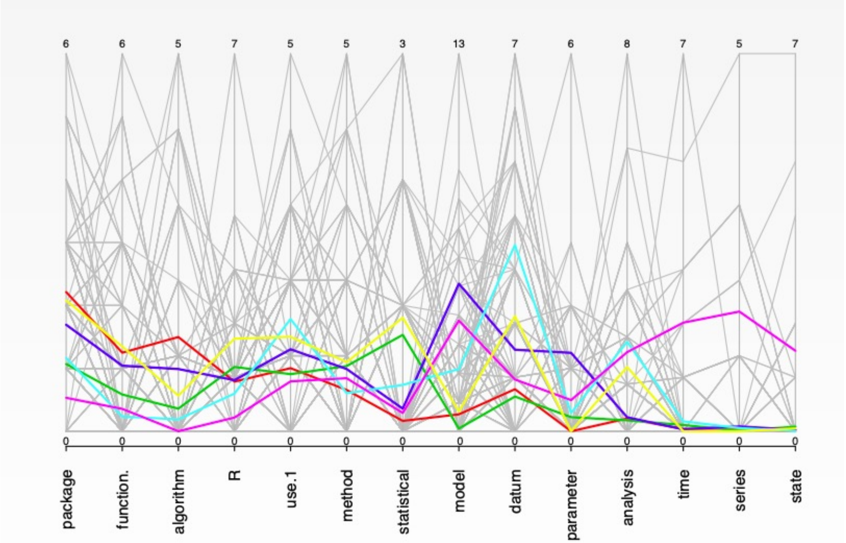
Archetypal Analysisの応用(2/2)

- Journal of Statistical Softwareの要旨(http://datacube.wu.ac.at/)
  - 2000年 – 2009年

アーキタイプ(BX)の特徴語

atype1	atype2	atype3	atype4	atype5	atype6
package	R	model	datum	model	datum
function	package	package	model	time	package
algorithm	method	datum	analysis	analysis	R
R	statistical	parameter	use.1	series	analysis
use.1	use.1	use.1	package	state	function

アーキタイプ(BX)の特徴語のプロット



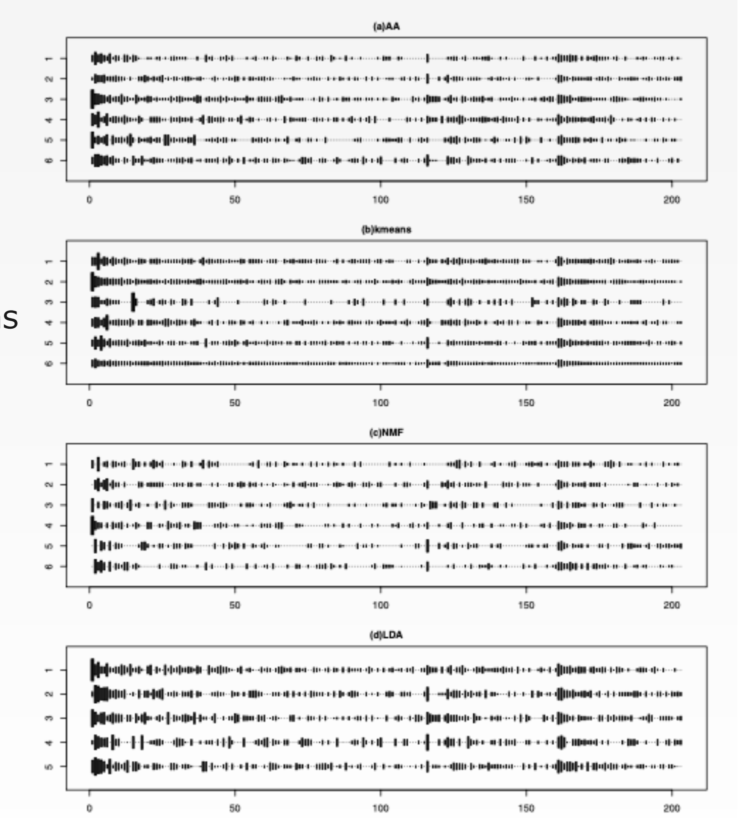
他のアルゴリズムよる語の重み

AA

kemans

NMF

LDA



Archetypal Analysisの応用・拡張

応用	• 力学系 (Stone, 1996; Cutler, 1997; Stone and Olson, 1999; Stone, 2002; Luttman et al., 2007) • マーケティング (Li et al., 2003; D’Esposito et al., 2006; Lee et al., 2010) • スポーツデータ (Eugster, 2012) • 遺伝子情報の解析 (Huggins et al., 2007;Mohammadi et al., 2018) • 画像処理 (Marinetti et al., 2007) • 非破壊検査 (Marinetti et al., 2006)
拡張	• 区間データ (D’Esposito et al. (2006); D’Esposito and Palumbo (2011, 2012); Corsaro and Marino (2010)) • ロバストアーキタイプ 分析 (Eugster and Leisch, 2011) • 拡張モデルのRへの実装（Eugster and Leisch, 2009） • Archetypoids（Vinue et al. 2015） • 大規模データのための近似アルゴリズム（Han et al., 2022） • Deep Archetypal Analysis(Keller, 2019)

まとめ

本日の内容

- 特異値分解に拠らない行列分解の各種アルゴリズムのR実装を紹介

- 特に、Archetypal Analysisの概要・拡張・応用例を解説

今後の見通し

- それぞれの分解の特徴を踏まえた適切な応用例を探す