

多次元尺度構成法

「MDSを使って使って使い倒す！」

MDS入門から非対称MDS実習まで」

2010年3月27日～28日

日本行動計量学会 第13回春の合宿セミナー A2コース

中山 厚穂（長崎大学）

横山 暁（慶應義塾大学）

3月27日（土）午後1(13:00～15:00)

○ 非計量MDSの「R」による
分析方法の紹介

非計量MDS

- Kruskalの方法が最も一般的
 - 2つの対象間の (非) 類似度関係を分析するための多次元尺度構成法である
- 類似度は少なくとも順位尺度であれば良い
- 類似度と距離の関係は単調関係（順序関係）を満たせばよく，線形関係を満たす必要性はない
- 低次元での当てはまりが良く，視覚的理解がしやすい
- 様々な距離に適用可能
- 欠測値に対する許容性ある

Kruskalの方法

- 対象間の関係を，類似度と対象間の距離とが単調関係を満たすように対象を点として多次元空間内に位置付けることで表現
 - 類似度の大きい対象同士はそれぞれの対象を表現する点間の距離は小さく
 - 類似度の小さい対象同士はそれぞれの対象を表現する点間の距離が大きく
- 次のような単調関係をできるだけ満たすように多次元空間内の点の座標を決定

$$\delta_{ij} > \delta_{rs} \Rightarrow d_{ij} \geq d_{rs}$$

ただし， δ_{ij} は対象*i*と*j*の非類似度， d_{ij} は対象*i*と*j*の距離

Kruskalの方法

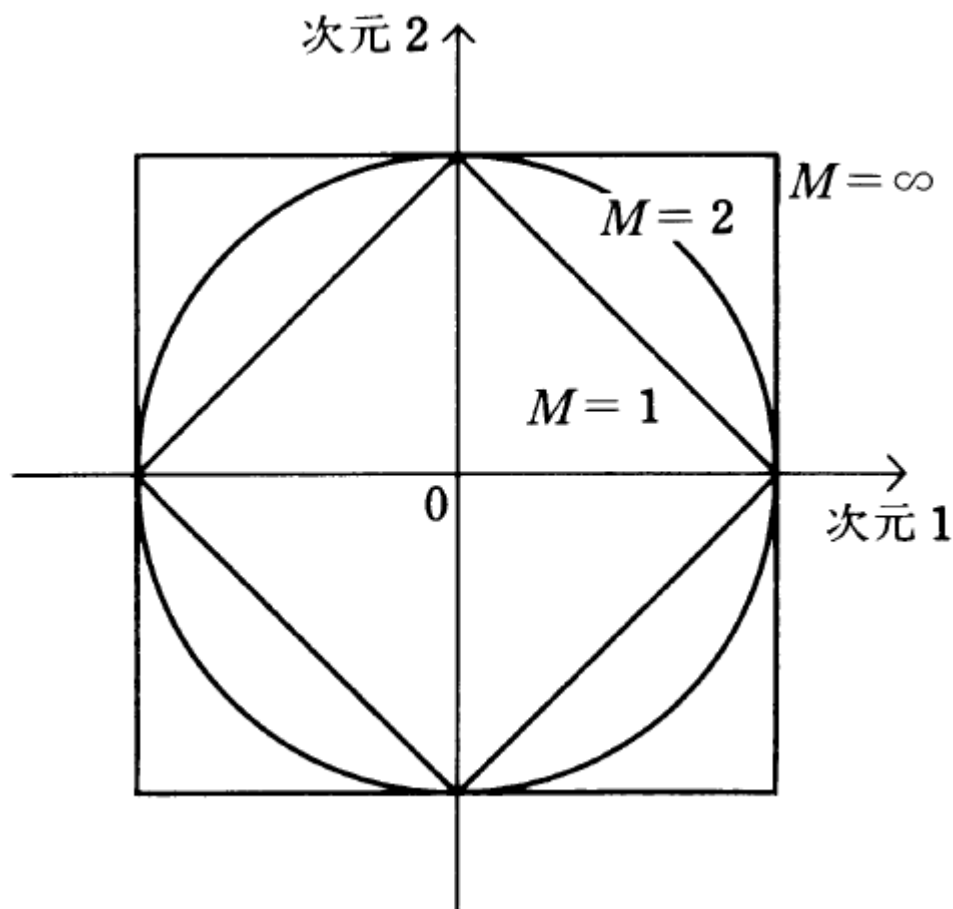
- 布置では、対象間の類似度の大小の関係が、各対象の点間の距離によって視覚的に把握可能
 - 対象を多次元空間内に位置づけた図を布置と呼ぶ
- 対象*i*の次元*t*での座標を x_{it} , p を多次元空間の次元数としたときユークリッド空間での点間距離

$$d_{ij} = \sqrt{\sum_{t=1}^p (x_{it} - x_{jt})^2}$$

ミンコフスキーの一般距離

- 通常はユークリッド距離で分析することが一般的
- 各次元の示す特性が個別に識別できる場合には非ユークリッド距離 (特に市街地距離) を用いる場合もある
 - 何種類かのMを用いて分析
 - 次元数を決定した上で, 最小のストレスが得られたMの結果を解
 - 非ユークリッド距離を用いた分析から得られた布置は次元の方向が一意に定まり, 回転できない

ミンコフスキーの一般距離



- 2次元空間
で, $M=1, M=2, M=\infty$ の場合に
定義される距離において, 原点
から等距離の点を図示
- ユークリッド距離 ($M=2$)
 - 原点を中心とする円周([超]球
面)
- 市街地距離 ($M=1$)
 - 菱形
- $M=1, M=2$ の中間の値で定義
される距離において, 原点から
等距離の点は, $M=1$ の菱形と
 $M=2$ の円周の中間に位置
- $M=2$ と $M=\infty$ の中間の値で定義
される距離において, 原点から
等距離の点は, $M=2$ の円周と
 $M=\infty$ の菱形の中間に位置

ミンコフスキー定数 M が 1 よりも大きいと, 各次元における座標の差の中で絶対値
が大きなものが強調

$M=1$ の場合には, p 個の次元における座標の差の絶対値の最大値が距離 d_{ij} である

Kruskalの方法

- 多次元空間内の対象を表す点の位置を決定するためには、はじめに、ある次元における n 個(分析対象の対象数)の点の仮の位置を決定
- 点間距離が類似度と単調減少関係になるように、多次元空間内での対象の点の座標を徐々に改善
- 最適な位置関係となる各対象の座標を算出
- 対象の座標を求めていく過程
 - 対象間の類似度と対象を表す点間の距離が必ずしも単調関係を満たさない場合もある
- 多次元空間内の点の座標を決定
 - 単調関係をどの程度満たしているのかを示す尺度である不適合度(ストレス)を基準とする

布置の基準化

- ストレスは点間の距離にもとづいて定義されているため、布置の大きさ、原点の位置、次元の方向には影響を受けない

- 布置の重心が原点となるように

$$\sum_{i=1}^n x_{it} = 0 \quad (t = 1, \dots, p)$$

- とし、また、座標値の2乗和が対象の個数に等しくなるように、

$$\sum_{i=1}^n \sum_{t=1}^p x_{it}^2 = n$$

と布置を基準化する必要がある

- 具体的には、平均が0、分散が1となるように得られた布置の座標を基準化したものを次元数の平方根で除す

ストレス S が最小となる座標の導出

- 制約のない最適化問題を解くための勾配法である最急降下法を利用
 - ストレス S が最小となる方向に解 (座標値) を探索
- 次元数 t のもとでストレスを最小化する布置を求めるとすると,
 - 仮の t 次元布置 (初期布置) を求める
 - この布置のストレスが減少するように対象の位置を少しずつ動かして布置を反復的に改善
 - ストレスが一定以上減少しなくなるまで反復を続行

ストレス S が最小となる座標の導出

- このときに得られる布置
 - ストレスがそれ以上改善できない布置(局所極小布置)
 - ストレスが必ずしも最小であるとは限らないことに注意
- 初期の座標値をどのように与えるかに解が依存
- 局所最適解 (局所的な最小値であって, 全体の最小値ではない値) が得られてしまう場合がある
- 様々な初期値を用いて分析を行い, ストレスの最小値が得られている結果を解として採用する必要

ストレス S が最小となる座標の導出

- ストレス S は点間の距離によって定義されるため、あらかじめ布置の次元数を決めておく必要
- 布置の次元数がいくつであるかは通常は分析を行う前には不明
- 何種類かの次元数のもとで分析を行い、各次元数においてストレスを最小にする布置を求める
- 得られたストレスや布置を検討して何次元の布置を解にするかを決定

分析の流れ

- 分析の最大次元数, 最小次元数, 最大反復数, 反復の停止条件を決定
- 最大次元数から最小次元数までの各次元数においてストレスが最小となる点間距離を類似度と単調関係を満たすように求める
 - 各次元数において最小となったストレスが得られているものを, その次元数における分析結果として採用
- どの次元数の結果を解として採用するのかを決定
 - ストレスの大きさ, 布置の解釈のしやすさ, 肘の基準, 類似度と点間距離の散布図といった解の採用基準と照らし合わせて決定する

関数isoMDSと関数Sammonでのストレスの定義

- 関数isoMDS は以下のように定義したストレスを最小になるように座標を決定(θ は単調増加関数)

$$S = \frac{\sum \sum (\theta(d_{ij}) - \hat{d}_{ij})^2}{\sum \sum \hat{d}_{ij}^2}$$

- 関数Sammonは与えられた t 次元上で以下のように定義した重みつきストレスを最小化するように座標を求める

$$S = \frac{1}{\sum \sum d_{ij}} \frac{\sum \sum (d_{ij} - \hat{d}_{ij})^2}{\sum \sum d_{ij}^2}$$

関数smacofSymでのストレスの定義

- 関数smacofSym は以下のように定義したストレスを最小になるように座標を決定

$$S = \sum w_{ij} (\delta_{ij} - d_{ij})^2$$

- ただし, 一般性を失わないようにするために $\sum w_{ij} \delta_{ij}^2 = n(n-1)/2$ とする

実際の分析では

- 分析の最大次元数と分析の最小次元数を決めて分析を行うのが一般的
 - 分析の最大次元数で分析して布置を求める
 - 分析により得られたその布置の「最大次元数-1」個の次元の座標の値を初期布置として「分析の最大次元数-1」の次元で分析して布置を求める
 - このように、次元数を1つずつ減らしていき、分析の最小次元数まで布置を求めていく
- 布置を改善するための最大反復回数は通常50～200回程度実行

布置の改善の目安

- 最大反復数に達する前にストレスがほぼ0と考えられる値 (目安は, 0.00001程度)まで減少すれば, 反復を終了
- 最大反復数に達していてもさらに布置を改善することができる場合
 - 最大反復数の1回前と最大反復数でのストレスの値を比較
 - その減少の程度が一定の値以上あれば, 布置はさらに改善できる可能性があると判断し, 最大反復数を増やして再分析
 - 減少の程度の目安は0.000001がおおよその基準
- 最大反復数に達していな場合において, それ以上反復して分析を行っても布置は改善できないと判断できる場合
 - 現在の反復と1回前の反復のストレスを比較
 - 減少の程度が一定の値以下であれば, それ以上反復して分析を行っても布置は改善できないと判断して反復を終了
 - 減少の程度の目安は0.000001がおおよその基準

解を採用する基準

- ストレスが小さい布置は適合度が大きいわけなので、ストレスの値は、解の次元数を決定する際の1つの基準
- 対象数や布置の次元数も考慮する必要があるが、およそ0.05から0.5程度までのストレスであればその次元数の結果を解として採用することができるといわれている
- その他
 - 布置の解釈
 - 肘の基準(ストレスの変化の程度)
 - 点間距離と類似度との散布図
 - ...
- いくつかの基準から総合的に判断して解とする次元数を決定するのが望ましい（詳細はテキストに記載）
 - このあと「R」による分析方法の紹介について詳しく説明

◦ **非計量MDSの「R」による
分析方法の紹介**

非計量MDSのプログラム

- 今回はisoMDSとsammonを用いる
 - ストレスの計算方法等が異なる
 - library(MASS)の読み込みが必要
- `isoMDS(d, y = cmdscale(d, k), k = 2, maxit = 50, trace = TRUE, tol = 0.001, p = 2)`
 - `d` : データ（距離形式か非類似度正方行列）
 - `y` : 初期布置（指定なしなら計量MDSの結果を利用）
 - `k` : 次元
 - `maxit` : 最大繰り返し数
 - `trace` : 最適化の過程を出力
 - `tol` : 変化幅の最小限度
 - `p` : ミンコフスキー定数

非計量MDSのプログラム

- `sammon(d, y = cmdscale(d, k), k = 2, niter = 100, trace = TRUE, magic = 0.2, tol = 1e-04)`
 - `d`: データ（距離形式か非類似度正方行列）
 - `y`: 初期布置（指定なしなら計量MDSの結果を利用）
 - `k`: 次元
 - `niter`: 最大繰り返し数
 - `trace`: 最適化の過程を出力
 - `magic`: ステップサイズの初期値
 - `tol`: 変化幅の最小限度

準備

- データの読み込み
 - `tourism<-read.table("tourism.dat", header=T, sep="¥t", row.names=1)`
- 類似度データなので非類似度データに
 - `tourism.dis<-8-tourism`
- データを距離行列に変換する
 - `tourism.dist<-as.dist(tourism.dis)`
 - 変換しなくても分析可能だが初期布置が必要になる

	Greece	Hawaii	West Coast	Hong Kong	London	Paris	East America
Hawaii		5					
West Coast		3	1				
Hong Kong		7	4	6			
London		2	7	4	7		
Paris		3	6	2	5	1	
East America		2	3	2	6	4	5

isoMDSの実行

➤ `tourism.iso<-isoMDS(tourism.dist, k=2)`

initial value 11.510834

iter 5 value 2.153903

iter 10 value 1.845324

iter 15 value 1.734993

iter 15 value 1.733410

iter 15 value 1.733410

final value 1.733410

converged

と出力

isoMDSの実行

➤ tourism.isoで結果が見れる

\$points

	[,1]	[,2]
Greece	2.2442891	1.089808
Hawaii	-3.0727925	1.358510
West Coast	-0.3686213	0.497304
Hong Kong	-4.1223052	-2.955376
LondonParis	3.8157995	-1.130995
East America	1.0741955	-1.976969
Australia	0.4294350	3.117718

\$stress

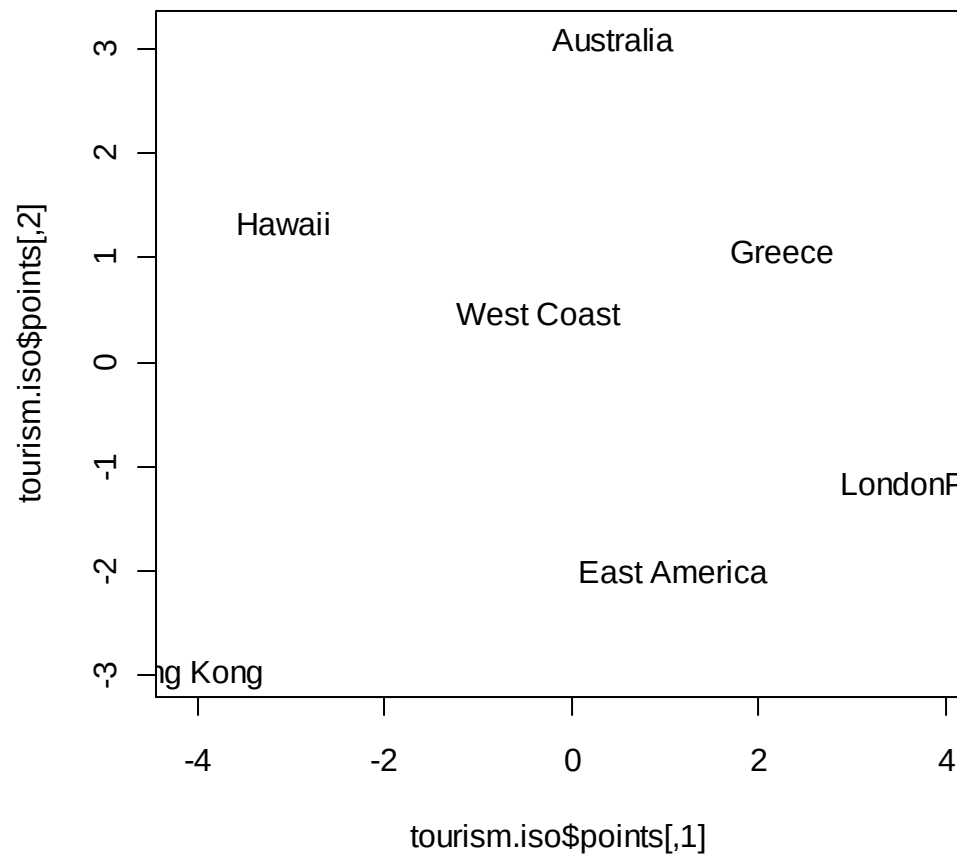
[1] 1.73341

isoMDSの実行

- `$points`が布置, `$stress`がストレス
- `Tourism.iso$points`, `tourism.iso$stress`で直接値を出力できる
- 布置を描く
 - `plot(tourism.iso$points)`
 - ラベル付きなら
 - `plot(tourism.iso$points, type="n")`
 - `text(tourism.iso$points, rownames(tourism.iso$points))`

isoMDSの実行

➤ 布置



isoMDS, sammonの問題点

- 布置

- 布置が基準化されていない

- 次元の指定

- 最大次元数と最小次元数を指定し，次元を減らして分析することは自動的にはできない

布置の問題

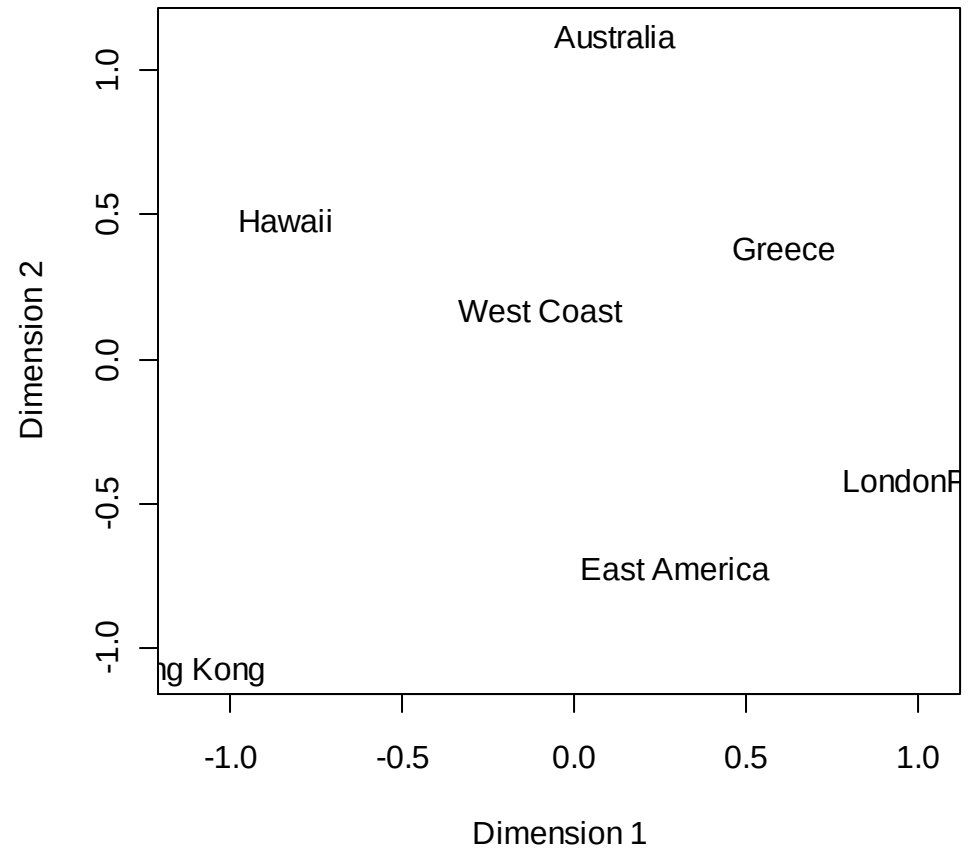
- 事前にDLしてもらった「mdsplot100325.R」を使用する
- normx(x)という関数
 - xには布置を入力
 - `norm.config<-normx(tourism.iso$points)`

Plot

- 「mdsplot100325.R」にある軸ラベルを変更してplotできる関数を利用する
- `mdsplot(d, x=1, y=2, label=T, norm=F)`
 - `d` : 布置
 - `x` : 布置に対する横軸の次元(default=1)
 - `y` : 布置に対する縦軸の次元(default=2)
 - `label` : 座標をラベルにするかどうか(default=T)
 - `norm` : 布置の基準化を行うかどうか(default=F)
- 3次元以上の分析を行い、次元を指定して布置を描くことも可能
 - 3次元Plotを行うこともできるが、ここでは割愛

Plot

- `mdsplot(tourism.iso$points, x=1, y=2, label=T, norm=T)`



次元の指定の問題

- 最大次元数と最小次元数を指定して分析を行うことができない
 - `isoMDS(data, k=4)`
 - `isoMDS(data, k=3)`
 - `isoMDS(data, k=2)`と分析していけばいいが、結果の比較が困難
- `MDSexe20100327.R`
 - `mdsplot100325.R`も含む（若干オプション指定を変更してある）

MDSexe20100327.R

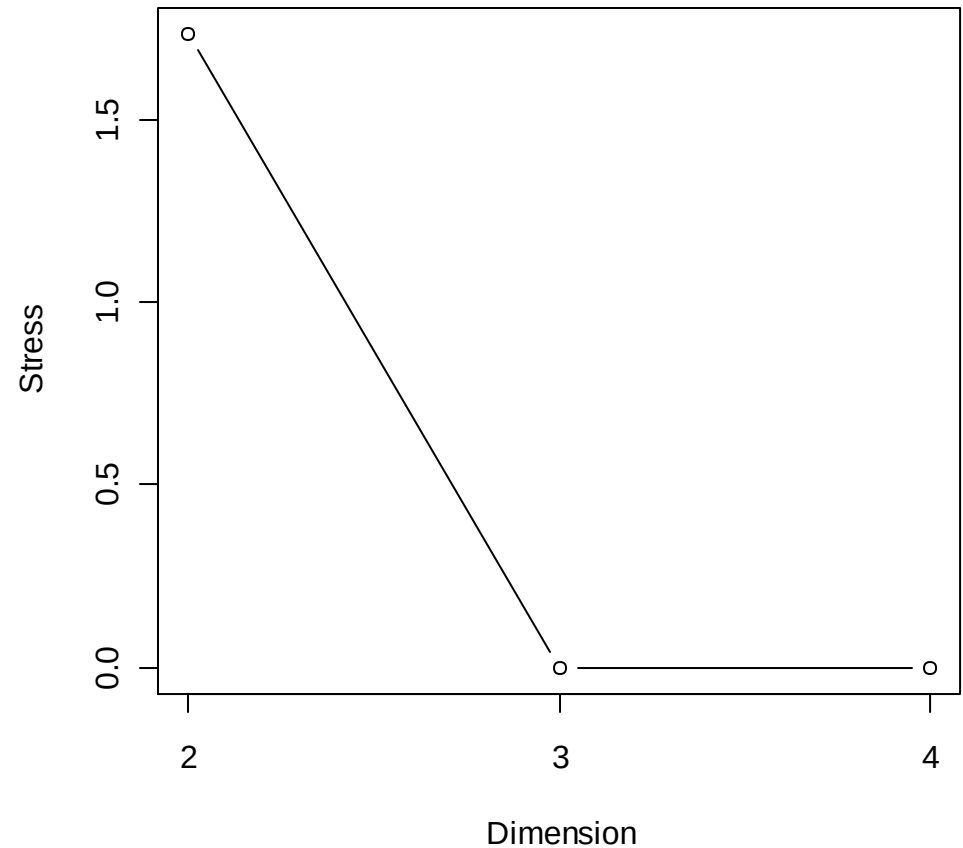
- MDSexe(data, maxdim, mindim=1, model=1, norm=T)
 - 最大次元数から最小次元数まで次元を減らしながら分析するスクリプト.
 - cmdscale, isoMDS, sammonが利用できる.
- data : 距離行列, 非類似度対象行列 (cmdscale, isoMDS, sammonで分析できるデータ形式)
- maxdim : 分析の最大次元数
- mindim : 分析の最小次元数(default=1)
- model : cmdscaleは1, isoMDSは2, sammonは3を指定(default=1)
- norm : isoMDS, sammonの時に布置の基準化を行うかどうか(default=T)

MDSexe20100327.R

- isoMDSやsammonの分析における初期布置
 - 最大次元数の時は合理的初期布置（cmdscaleを利用）,
 - 最大次元数 - 1 ~ 最小次元数の時は, 1つ上の次元での布置(*)において, 次元毎の2乗和の最も小さい次元を省いた布置を初期布置とする.
 - (*)ただし, isoMDSにおいてnorm=Tを指定した場合には, 基準化された布置

MDSexe20100327.R

- `tourism.result<-MDSexe(tourism.dist, 4, 2, 2, norm=T)`
- 自動的に次元xストレスのスクリープロットが出力



MDSexe20100327.R

- MDSexeplot(tourism.result, dim=2, x=1, y=2)
- 2次元の結果を出力

