

多変量解析 <初歩の初歩>に入門

変数のベクトル表現と主成分分析

なかもとけいこ
kenakamoto@nifty.com

今回の目的

- 多変量解析の例を見てみる。
- 変数のベクトル表現をかじってみる
- 主成分分析について簡単に学ぶ。
- 多変量解析の限界を肝に銘じておく。

多変量解析とは?

- その名のとおり, たくさんの変量をまとめて分析するための統計的手法の総称. ものすごく色々ある.

	(主に)量的データ	(主に)質的データ
外部基準あり	(重)回帰分析 正準判別分析 ロジスティック回帰分析	数量化 類 数量化 類
外部基準なし	主成分分析 因子分析 クラスター分析 多次元尺度法	数量化 類(対応分析)

これらのメタモデルとして構造方程式モデリング(SEM: 共分散構造分析)

多変量解析で得すること

- 世の中のことは大抵簡単にはできていない. たくさんの要因が組み合わさって結果が現れる.
 - 一つ一つの要因を順番に見ていっていたのではハッキリした結果は得られない.
 - しかし, ヒトはそんなに賢くないので, 道具なしにたくさんのことを一度に考えることはできない.
- 多変量解析を使うと, すっきり整理して把握することができる!
- 勉強する手間の方が, データを前に悩む手間よりもたぶん少ないです.

得することの例(1)

- 複数の説明変数を使った目的変数の予測.
 - 重回帰分析
 - 決定係数(説明率): それらの説明変数でどれくらい目的変数を予測できるか?
 - 偏回帰係数: 他の変数の影響を除去したときに, 一つの変数がどれくらい目的変数と関連しているか?
 - 比喩の適切さと述部の慣習性による隠喩形式選好の予測

	Mean (SD)	Correlation (N=42)		Multiple regression analysis Standardized weights
		Met. pref.	Aptness	
Metaphor form preference	3.57 (.99)			
Aptness of the comparison	3.27 (.67)	.508**		.439**
Base conventionality	4.88 (.87)	.394**	.239	.289*
				R ² .337

R = .58

Nakamoto & Kusumi (2004: in prep.)

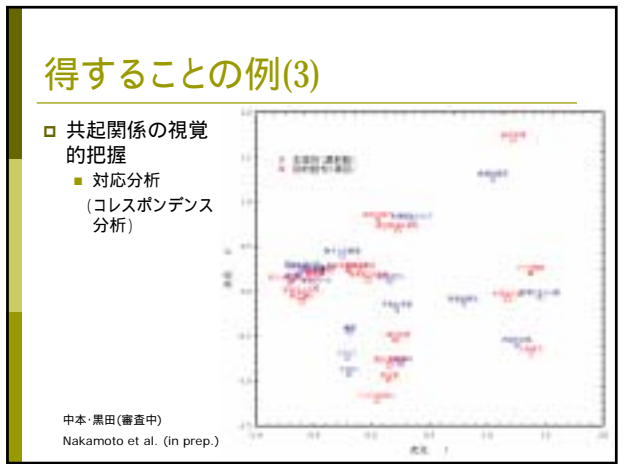
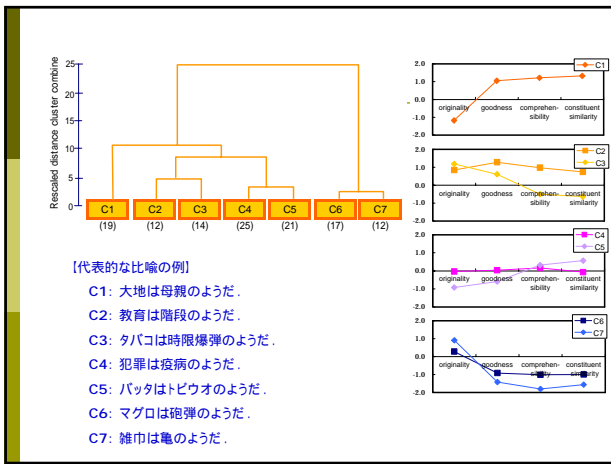
得することの例(2)

- ケースの類似性を利用した分類(クラスタリング)
 - (階層的)クラスタ分析

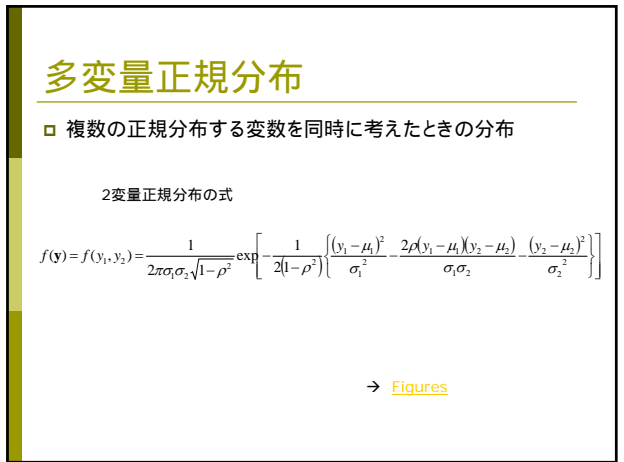
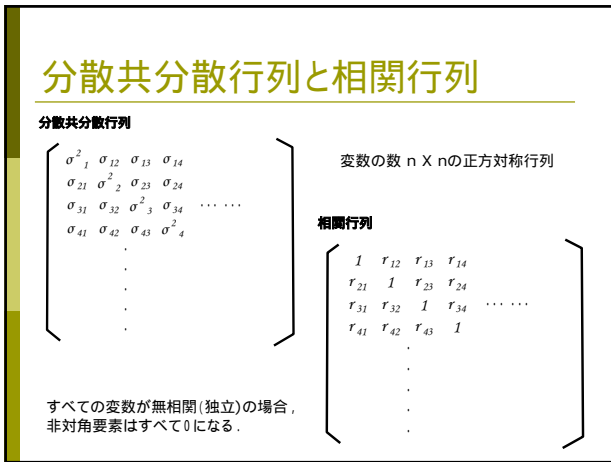
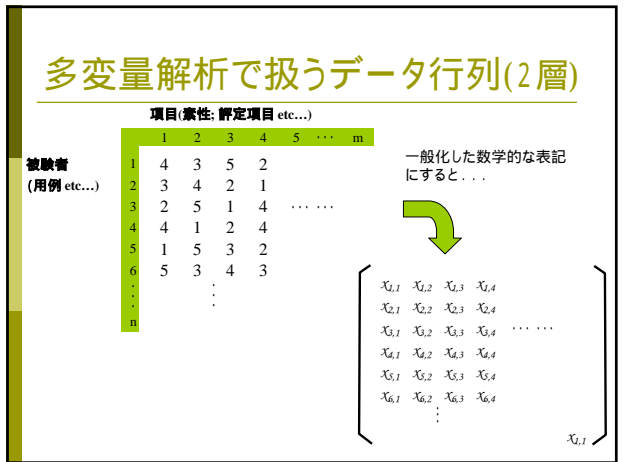
- 比喩の分類

比喩	出典	隠喩可能性 no=0	構造類似性 no=0	比喩性 no=0	面白さ no=1	判別 分析結果
11 警備-弦	yt	6.87 (2.42)	5.95 (2.06)	4.33 (2.28)	5.23 (1.92)	
90 寒波-鉄壁	ku	4.95 (2.59)	3.38 (2.10)	4.70 (1.79)	4.13 (2.04)	
62 微風-吐息	ku	6.33 (2.57)	4.78 (2.42)	4.45 (2.16)	4.51 (2.07)	
78 寝息-華笛	ku	4.95 (2.97)	3.23 (1.98)	4.52 (2.05)	4.16 (2.13)	
89 愛-季節	tk	6.78 (2.43)	4.60 (2.35)	4.98 (2.38)	5.18 (2.44)	
22 議論-戦争	ku	6.62 (2.06)	5.37 (2.22)	4.23 (1.87)	4.44 (2.16)	(1)
80 煙草-時限爆弾	ao	4.16 (2.69)	3.25 (2.18)	4.70 (2.30)	4.52 (2.51)	
50 動悸-早鐘	ku	5.57 (2.80)	4.63 (2.69)	4.05 (1.74)	4.25 (2.05)	
5 激怒-噴火	ku	8.00 (1.51)	7.23 (1.94)	4.02 (2.77)	4.69 (2.41)	
21 治療-修理	ku	7.13 (2.04)	6.83 (1.67)	3.64 (2.17)	3.97 (2.24)	(1)
34 孤独-砂漠	ku	6.75 (2.40)	6.00 (1.90)	5.09 (2.24)	4.84 (2.30)	
71 雲配-足音	ku	5.63 (2.38)	5.97 (1.94)	4.30 (2.35)	3.87 (2.22)	
44 革命-地震	ku	6.03 (2.18)	4.65 (2.17)	4.58 (2.12)	4.31 (1.99)	
58 芋毛-雲崩	au	5.10 (2.45)	3.13 (2.28)	4.64 (1.91)	3.85 (1.99)	
29 記憶-倉庫	ku	6.56 (4.22)	5.35 (2.45)	5.06 (2.22)	4.34 (2.32)	
45 教育-階段	ao	6.79 (2.21)	4.83 (2.35)	4.77 (2.10)	4.90 (2.05)	
15 殺し木-枯らし	ai	7.57 (1.99)	6.72 (2.00)	4.34 (2.64)	5.03 (2.37)	
70 マグロ-榴弾	an	3.90 (2.76)	2.55 (2.08)	4.67 (2.42)	3.97 (2.06)	

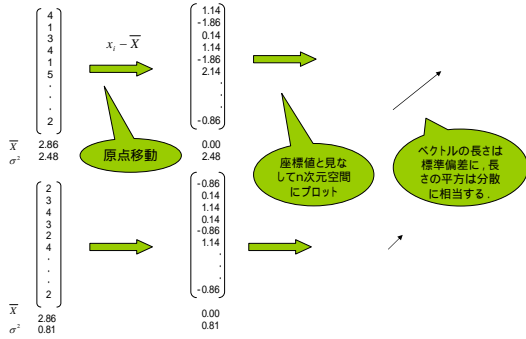
中本・橋見(2004)



	F10a	F10b	F10c	F10d	F10e	F10f	F10g	F10h	F10i	F10j	F10k	F10l	F10m	F10n	F10o	F10p	F10q	F10r	F10s	F10t	F10u	F10v	F10w	F10x	F10y	F10z
F10a 最大の人数	15	9	3	19	5	11	6	10	1	6	1	4	1	2	4	2										
F10b 二人の差	7	6	2	3	11	7	7	1	1	7	0	2	0	0	0	0										
F10c 結果にたいして	1.29	1.44	0.71	0.97	0.81	0.99	0.81	0.78	0.69	0.69	0.69	0.69	0.69	0.69	0.69	0.69										
F10d 三人の人数	7	12	6	14	9	10	12	11	14	9	3	0	0	0	0	0										
F10e ストーリー	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
F10f 演り方	4	6	1	2	10	18	12	11	13	13	11	10	10	10	10	10										
F10g ライオン	2	2	1	6	0	2	4	20	2	5	0	0	0	0	0	0										
F10h イラスト	0.24	1.23	0.23	0.38	0.25	1.05	1.16	0.91	0.29	0.41	1.00	0.00	1.00	0.04	1.27	0.23										
F10i 動物の顔	2	1	2	6	2	7	7	0	13	9	0	0	0	0	0	0										
F10j 動物の手	6	5	0	4	4	10	9	0	1	12	0	0	0	0	0	0										
F10k 大抵の場合	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0										
F10l 土地柄	1	0	1	1	1	2	1	1	10	4	0	14	2	0	0	0										
F10m 新築の家	1	3	11	0	3	0	2	1	1	1	0	1	14	0	1	7										
F10n 積立貯蓄	1.17	0.86	0.84	1.26	0.70	1.01	0.76	1.18	0.45	1.09	0.74	0.26	0.78	0.79	1.37	0.81										
F10o 天候のせい	1.42	1.27	0.96	1.16	1.17	1.21	1.00	0.87	1.15	1.04	0.71	1.27	0.96	1.04	1.00	1.00										
F10p 動物の尻	1	6	4	0	3	0	0	1	1	1	2	1	4	0	0	0										
F10q 動物の首	1.25	0.71	0.81	1.20	0.79	1.00	0.85	0.81	0.57	1.21	1.07	0.87	0.96	0.81	1.25	0.81										
F10r 手帳	1	3	4	0	0	0	4	0	0	0	0	0	0	0	0	0										
F10s 手帳の書き方	1.71	1.04	1.13	1.17	0.80	1.19	0.78	0.77	1.07	0.71	0.73	1.27	0.81	0.81	1.00	1.00										
F10t 手帳の書き方	7	8	2	1	6	0	1	1	2	4	2	1	0	1	0	1										
F10u 日記	0.71	0.88	1.00	0.84	0.81	1.00	0.77	0.81	0.87	0.78	0.81	0.81	0.81	0.81	0.81	0.81										
F10v	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0										
F10w	1.12	1.47	0.88	0.78	0.73	0.81	0.75	0.89	0.89	0.77	1.29	0.11	2.14	4.66	1.81	0.79										

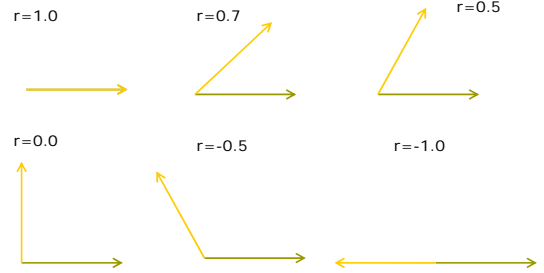


変数のベクトル表現



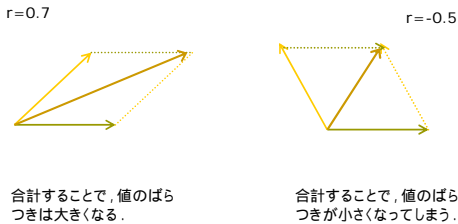
相関(共分散)のベクトル表現(1)

相関(共分散)はベクトルの角度 $\cos(\theta) = \frac{x_1 \text{と} x_2 \text{の内積}}{x_1 \text{の長さ} \times x_2 \text{の長さ}} = \text{相関係数}$ (の余弦)に相当する。



相関(共分散)のベクトル表現(2)

合計点を表すベクトル



相関(共分散)のベクトル表現(3)



3次元ベクトル以上のが3つ以上ある場合、うまい面を発見してやると、2次元上に全部のベクトルがキレイに乗っかり、普通は3次元以上を使わないと表現しきれない。

この面を探すために解くのが相関行列の固有値, 固有ベクトル

固有値と固有ベクトル(1)

多変量解析の本を読んでも、とにかくしゅっちゅう出てくるコトバ。

- 主成分分析 (因子分析)
- 判別分析
- 数量化, 類 など...

固有値問題

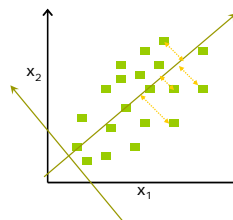
- 正方行列 A ($n \times n$) が与えられたとき,

$$Ax = \lambda x$$

- を満たす λ を固有値, x を固有ベクトルという。
- n 次の正方行列の固有値は n 個ある。
- 第1固有値が一番大きく、だんだん小さくなっていく (cf., → スクリーンプロット)

固有値と固有ベクトル(2)

うまい合成変量を見つけるために使う=主成分分析



データのちらばりを最も効率よく表現できる軸を探します。

→ その軸は, x_1, x_2, \dots, x_n の合成変量
→ $f_1 = a_1x_1 + a_2x_2 + \dots + a_nx_n$ として表せる。

この例だと二つの変量 (X_1 と X_2) を一つの (合成) 変量でほとんど表せるようになる。

固有値 = 新しい軸で表せる分散の大きさに対応。

固有ベクトル = 合成得点を求めるための重みに相当

実際の例: 黒田ほか(2005)から抜粋

F02	侵略	ある国の戦車部隊が<ポーランド>を襲った
F02	侵略	ある国の戦車部隊が<スコ>を襲った
F03	強盗	二人組の強盗が<都内の銀行>を襲った
F03	強盗	二人組の強盗が<スネ>を襲った
F05	虐待	通り魔が<下校中の小学生>を襲った
F05	虐待	通り魔が<ユチ>を襲った
F06	捕食	ライオンが<インバラの群れ>を襲った
F06	捕食	ライオンが<ルエ>を襲った
F09	気象(大)	大型の台風が<日本列島>を襲った
F09	気象(大)	大型の台風が<ホヌ>を襲った
F10	気象(小)	強烈な突風が<出動途中のOL>を襲った
F10	気象(小)	強烈な突風が<ロウ>を襲った
F11	疫病	インフルエンザの流行が<アジア>を襲った
F11	疫病	インフルエンザの流行が<ツサ>を襲った
F13	発病	悪性のガンが<働き盛りの男性>を襲った
F13	発病	悪性のガンが<ミフ>を襲った
F15	悪感情	言いようのない不安が<やり手の部長>を襲った
F15	悪感情	言いようのない不安が<レウ>を襲った
ns	ns	ユチが<ラヨ>を襲った

Xは生き物である
Xは場所である
Xは人間である
Xの規模は個人/個体よりも大きい
Xは人または動物の集まりである
Xは施設である
Xは自然現象である
Xは意図を持っている

×

実際の例: 黒田ほか(2005)から抜粋

□ 相関行列

- やはり、あちこちの項目で相関あり.
- できれば、うまくまとめて全体像を把握したい...
- というわけで、PCA を実行

	x1	x2	x3	x4	x5	x6	x7	x8	
x1	Xは生き物である	1.00							
x2	Xは場所である	-0.96	1.00						
x3	Xは人間である	0.77	-0.75	1.00					
x4	Xの規模は個人/個体よりも大きい	-0.86	0.88	-0.92	1.00				
x5	Xは人または動物の集まりである	-0.39	0.50	-0.72	0.72	1.00			
x6	Xは施設である	-0.59	0.70	-0.36	0.53	0.29	1.00		
x7	Xは自然現象である	-0.12	0.08	-0.48	0.34	0.32	-0.46	1.00	
x8	Xは意図を持っている	0.82	-0.89	0.67	-0.78	-0.63	-0.65	0.02	1.00

実際の例: 黒田ほか(2005)から抜粋

□ では、実際にJMPで計算してみましょう。

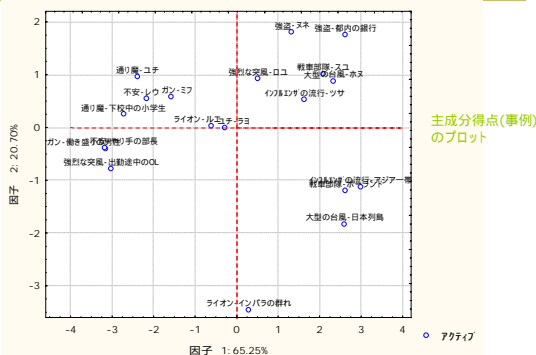
- Sample dataは、
- PCA-sample-data-kuroda-et-al(2005).xls

実際の例: 黒田ほか(2005)から抜粋

□ 結果(ただし、この例ではStatisticaを使用)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
固有値と寄与率	5.22	1.66	0.61	0.26	0.14	0.07	0.03	0.01
固有値の合計	5.22	6.88	7.49	7.75	7.89	7.97	7.99	8.00
寄与率	65.25	20.70	7.65	3.29	1.77	0.91	0.34	0.08
累積寄与率	65.25	85.95	93.61	96.89	98.67	99.58	99.92	100.00
固有ベクトル								
Xは生き物である	-0.399	-0.092	-0.463	-0.142	-0.273	0.158	-0.381	0.594
Xは場所である	0.416	0.133	0.261	0.131	-0.229	-0.287	0.308	0.703
Xは人間である	-0.388	0.272	-0.015	0.333	-0.531	-0.557	-0.030	-0.266
Xの規模は個人/個体よりも大きい	0.421	-0.128	0.022	-0.235	0.071	-0.498	-0.702	-0.062
Xは人または動物の集まりである	0.310	-0.252	-0.790	0.099	0.025	-0.248	0.362	-0.113
Xは施設である	0.283	0.519	-0.146	-0.584	-0.442	0.221	0.082	-0.192
Xは自然現象である	0.088	-0.725	0.243	-0.099	-0.590	0.149	0.074	-0.150
Xは意図を持っている	-0.394	-0.159	0.110	-0.660	0.203	-0.449	0.350	0.075
主成分負荷量(変数とPCの相関係数に一致)								
Xは生き物である	-0.912	-0.118	-0.362	-0.073	-0.103	0.043	-0.063	0.049
Xは場所である	0.952	0.171	0.204	0.067	-0.086	-0.077	0.051	0.058
Xは人間である	-0.886	0.351	-0.012	0.171	-0.200	-0.150	-0.005	-0.022
Xの規模は個人/個体よりも大きい	0.962	-0.165	0.017	-0.121	0.027	-0.134	-0.116	-0.005
Xは人または動物の集まりである	0.708	-0.325	-0.618	0.051	0.010	-0.067	0.060	-0.009
Xは施設である	0.648	0.668	-0.115	-0.299	-0.167	0.060	0.013	-0.016
Xは自然現象である	0.202	-0.932	0.190	-0.051	-0.222	0.040	0.012	-0.012
Xは意図を持っている	-0.901	-0.205	0.086	-0.339	0.077	-0.121	0.058	0.006

実際の例: 黒田ほか(2005)から抜粋



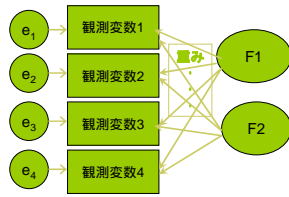
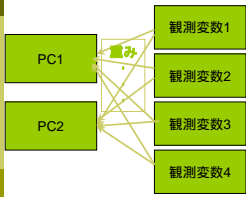
主成分分析のまとめ

- 相関のある変数群をまとめて、新しい合成変数を創り出す手法
 - 数学的にはそんなにややくくない.
 - モデル依存的な仮定の少ない記述統計的手法
- 主成分は、分散説明率(寄与率)の大きいものから、順次、独立=直交するように抽出される.
 - 主成分間には相関がないので、クラスター分析などに用いる前段階の処理として使ってもよい(かもしれない)
- 因子分析(今回割愛)とは次スライドのような違いがある.

主成分分析 PCA と因子分析 FA

PCAの模式図

FAの模式図



・ほとどの観測変数の分散を少数の次元で最もよく反映する座標軸 (=PC) を引き直す。
 ・主成分得点は、観測変数の重み付き和
 ・顕在変数から顕在変数の推定 = 記述的統計
 ・尺度不変、回転不変ではない。

・個々の観測変数に対して、複数の因子がどのようにか寄与しているかを推定。
 ・因子得点は潜在変数であり観測不可能。
 ・種々の仮定を含むモデル依存的分析
 ・尺度不変、回転不変である。

多変量解析の限界 (1)

- 大量の数値データを**縮約**し、(場合によっては誤差成分をコントロールし) **可視化**するための技術。
 - 得られる結果は、データの縮約的記述であり、それ自体が何か説明を与えるわけではない。
 - 解釈は**実質科学的見地** (心理学や言語学の専門家としての知識とカン) からなされるべき。
- 結果をどこまで一般化するかは**研究上の観点**による。客観的基準で決められるわけではない。
 - 解析結果を外挿する場合は特に**慎重な判断**が必要。

多変量解析の限界 (2)

- 一般的に使用される多変量解析は複合系ではあるが、**複雑系ではない**。
 - 基本的には**線形モデル**にのっとっている。
 - 確率的な変動はモデルに組み込まれているが、その挙動は**足し算かけ算に限られる**。
 - 非線形(複雑系)の現象を線形で近似的に記述していることを自覚すること。
 - 現象の全体ではなく一部を扱っているがゆえに、線形で近似できるケース
 - 非線形な現象に対し、研究者/被験者/解析者のいずれかが線形な解釈を行った結果がデータとして得られているケース

自分で出来るようになるために

- コンピュータに慣れる!
 - 多くは手計算で出来るようなものではないし、たとえ出来たとしても時間の無駄です。
 - Excel(表計算ソフト)を上手く使えるようになる。
 - 得意な統計ソフトを一つ作る。
- がんばって本を読むクセをつける!
 - 分からないから数式を飛ばす/ 記号や添え字を確認しないで流し読みする等して、数学系の本が理解できるわけがない。
- 自分の扱っているデータの構造をイメージできるようになる!
 - これが出来ると、手法を自分で選択したり、効率の良いデータ入力フォーマットを作ったりできるようになる。

今回のまとめ

- 今回は下記のことを学びました。
 - 多変量解析には色々あります。
 - 複数の変数ベクトルがくっついた行列を扱います。
 - 直接扱うのは相関行列であることも多いです。
 - 多変量正規分布という分布があります。
 - 変数はベクトルで、相関係数はベクトル間の角度(の余弦)で表せます。
 - 変数に重み付けして足してやれば、新しい合成変数が得られます。これに関係する数学的手法が固有値問題であり、主成分分析です。
 - 多変量解析にも色々な点で限界があります。
- 結局は自分の努力で身につけねばならない部分が多々あります。

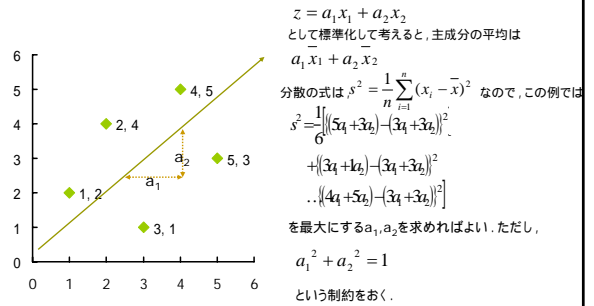
おまけ: 代表的な統計ソフトウェア

- SPSS (Amos, Clementine)
 - たぶん、日本の心理学界では最もユーザが多い。
 - ツールバーからコマンド選択で大抵のことは出来るのでらくちん。
 - でも、少し凝ったことをしようとしたり、一般的な使用頻度の低い分析(e.g., LOGLINEAR)をしようとすると、スクリプトを書く必要がある。
- SAS
 - スクリプトを書いて実行する方式
 - 非常に多くの分析手法をサポートしている。参考書も、和文、英文とも色々出ている。
 - スクリプトをsasウィンドウで書いていると、妙に重い...
- Statistica
 - コマンド選択式で解析可能。らくちん。
 - グラフの出力が非常にウツクシイ。
 - ユーザーズガイド(ソフトウェアに付属)が充実している。
 - けれども3.0Jは焦って日本語版をリリースしたらしくあまりデキはよくない。
- R
 - ナカモトが乗り換えを考えているソフト。ただし、私はまだ使ったことはありません。
 - なんといっても、無料なところがすごい!!!
 - Open Developmentなので、新しい解析手法が実装されるのが早い。たとえば、とっくの昔にSOM(自己組織化マップ)、ICA(独立成分分析)が実装されているし、多重比較も色々な手法を取りそろえている。
 - 最近、本もたくさん出ているので、興味のある人は一緒に勉強しましょう。

数式アレルギーのない人は、これ以降のおまけにチャレンジしてください。

おまけ: 主成分と固有値 (簡単な例)

□ 主成分を求めるとは、 $a_1x_1 + a_2x_2$ の分散を最大化するように、 a_1, a_2 を決めること



さっきの式を一生涯整理すると、

$$s^2 = 2a_1^2 + 1.2a_1a_2 + 2a_2^2 = V(a_1, a_2)$$

ここでラグランジュの乗数法を採用。未定乗数を λ とする。

$$F(a_1, a_2, \lambda) = V(a_1, a_2) - \lambda(a_1 + a_2 - 1)$$

で、 (a_1, a_2, λ) で F を偏微分すると、

$$\frac{\partial F}{\partial a_1} = 2(2a_1 + 0.6a_2 - \lambda a_1)$$

$$\frac{\partial F}{\partial a_2} = 2(0.6a_1 + 2a_2 - \lambda a_2)$$

$$\frac{\partial F}{\partial \lambda} = -(a_1^2 + a_2^2 - 1)$$

なので、以下の連立方程式をときばいいことになる。

$$\begin{cases} 2a_1 + 0.6a_2 - \lambda a_1 = 0 & \dots \\ 0.6a_1 + 2a_2 - \lambda a_2 = 0 & \dots \\ -(a_1^2 + a_2^2 - 1) = 0 \end{cases}$$

ここで連立方程式を行列の形で表すと、

$$\begin{bmatrix} 2.0 & 0.6 \\ 0.6 & 2.0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \lambda \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

分散の式の整理を思い出すと、この式の左側の行列は分散共分散行列になっていることが分かる。ここで、固有値、固有ベクトルとは、

$$Ax = \lambda x$$

を満たす λ と x であることを思い出すと、 λ は分散共分散行列の固有値、 $[a_1, a_2]^T$ は分散共分散行列の固有ベクトルになっていることが分かる。さらに、全スライド 式に a_1, a_2 をかけてみると、

$$2a_1^2 + 0.6a_1a_2 - \lambda a_1^2$$

$$2a_2^2 + 0.6a_1a_2 - \lambda a_2^2$$

2式を足しあわせると、

$$2a_1^2 + 1.2a_1a_2 + 2a_2^2 - \lambda(a_1^2 + a_2^2) = 0$$

適当に移行してやったり、 $a_1^2 + a_2^2 = 1$ にしたことを思い出すと、

$$2a_1^2 + 1.2a_1a_2 + 2a_2^2 = V(a_1, a_2) = \lambda$$

つまり、主成分の係数(うまい合成変量をつくるための重み)とは、分散共分散行列の最大固有値に属する固有ベクトルであるなので、あとは解くだけです。

おまけのおまけ: ラグランジュの乗数法

ラグランジュの乗数法

関数 $U(a_2, a_1, a_0)$ が条件 $a_2^2 + a_1^2 - 1 = 0$ のもとに、点 (α, β, γ) で極値をとるならば、関数

$F(a_2, a_1, a_0, \lambda)$ を

$$F(a_2, a_1, a_0, \lambda) = U(a_2, a_1, a_0) - \lambda(a_2^2 + a_1^2 - 1)$$

とおいたとき、極値をとる点 (α, β, γ) は連立方程式

$$\frac{\partial F}{\partial a_2} = 0, \frac{\partial F}{\partial a_1} = 0, \frac{\partial F}{\partial a_0} = 0, a_2^2 + a_1^2 - 1 = 0$$

の解になる。

(有馬・石村(1987)からまる写し。これ以上のことは自分で勉強してください)