周期時系列の統計解析

(14) 教科点数の主成分分析

nino

2020 年 1 月 12 日

前報（参考文献１）では，周期時系列データにおける主成分分析の意味について検討した．ここでは，横断面データの具体例として３教科の点数に主成分分析を適用し，その主成分分析結果について検討するとともに，時系列との関係性を調べた．

３教科の点数データ

用いたデータは，参考文献２に記載された５教科の点数のうち，英語と国語と数学の３教科の点数（20人）である．３教科の点数に相関行列による主成分分析を適用した．５教科の点数の主成分分析については参考文献２に詳しく解説されているので，そちらを参照されたい．

表１に３教科の点数を，図１に点数の積み上げ棒グラフを示した．また，表２に相関係数を示した．

表１

生徒ID	英語	国語	数学	合計点
1	88	72	80	240
2	62	53	88	203
3	50	44	25	119
4	43	29	34	106
5	66	66	29	161
6	43	51	52	146
7	75	73	38	186
8	52	69	71	192
9	58	35	65	158
10	46	42	28	116
11	38	57	25	120
12	40	55	30	125
13	66	49	61	176
14	81	73	95	249
15	74	63	36	173
16	60	50	41	151
17	55	42	71	168
18	53	57	55	165
19	78	81	47	206
20	49	66	45	160
平均値	58.9	56.4	50.8	166.0

表２

	英語	国語	数学
英語	1		
国語	0.615	1	
数学	0.476	0.202	1

図１

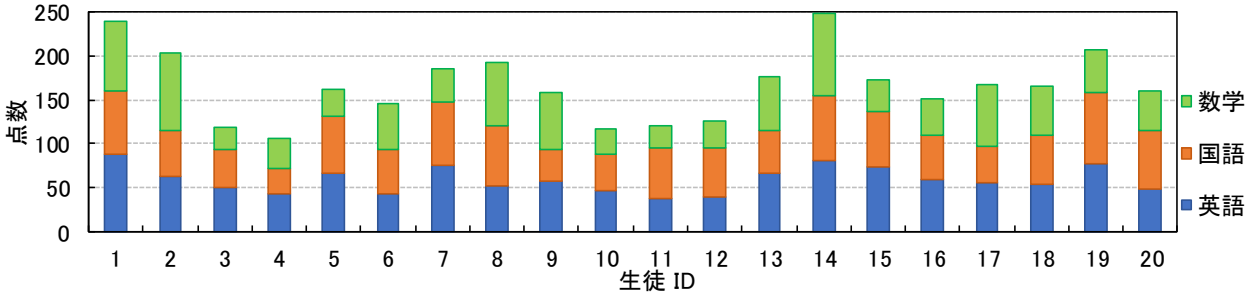


図１から，生徒の合計点はID-14で最も高く，ID-4で最も低かった．また，３教科の点数は生徒ごとに高低の違いがみられた．
表２によると，相関係数は英語と国語が0.615と最も高く，次いで，英語と数学が0.476，そして，国語と数学が0.202と最も低かった．

教科点数の主成分分析

表３に教科点数の主成分分析結果，固有ベクトルおよび因子負荷量を示した．また，図２には主成分別に各教科の固有ベクトルを示した．

表３

主成分結果	z1	z2	z3
固有値	1.882	0.805	0.312
寄与率	0.627	0.268	0.104
累積寄与率	0.627	0.896	1.000
固有ベクトル			
英語	0.661	−0.074	−0.747
国語	0.572	−0.596	0.565
数学	0.487	0.800	0.351
因子負荷量			
英語	0.906	−0.067	−0.417
国語	0.784	−0.534	0.315
数学	0.668	0.718	0.196

図２

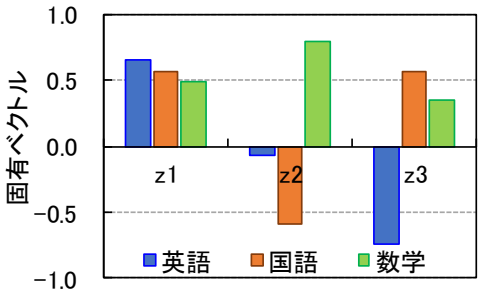


表３の主成分分析結果によると，主成分は３つ形成された．固有値はz1が1.882で１以上の値を示したが，z2とz3はそれぞれ0.805と0.312であった．
固有ベクトル（図２）については，z1は英語，国語，数学の順に少しずつ小さくなるが，３教科ともに正数であることから，z1は合計点的な主成分であると示唆された（参考文献２）．z2は英語がほぼ0であるが，国語は負数，数学は正数を示した．したがって，z2は国語に比べて相対的に数学が得意であるという特徴を示す主成分と考えられた．z3の固有ベクトルは英語で大きな負数を示すが，国語と数学は正数であることから，z3は国語や数学に比べて相対的に英語がかなり不得意であることを示す主成分といえる．
このように，各主成分は３教科の合計点的あるいは得意や不得意を表していることから，各主成分の特徴を各教科の点数を用いて検討した．

主成分と教科点数との関係

相関行列の主成分分析における主成分スコアは各教科点数の基準値に重み係数としての固有ベクトル（表３）を掛けた値を足し合わせたものである．このことを参考にして，主成分スコアと教科点数の関係を調べた．

z1は合計点的な主成分であり，合計点は各教科点数に同じ重み係数１を掛けた値を足し合わせたものである．また，z2は国語に比べて数学が得意な主成分であるから，国語の重み係数を－1，数学の重み係数を＋１とした値［数学－国語］に対応していると考えられる．z3は国語や数学に比べて相対的に英語がかなり不得意であることから，例えば，英語の重み係数を－１，国語と数学の重み係数を0.5とした値［－英語＋0.5国語＋0.5数学］にほぼ対応していると考えられる．

図３にz1主成分スコアと合計点の関係，図４にz2主成分スコアと［数学－国語］の関係，そして図５にz3主成分スコアと［－英語＋0.5国語＋0.5数学］の関係をそれぞれ示した．

図３

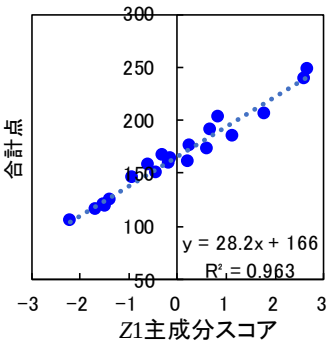


図４

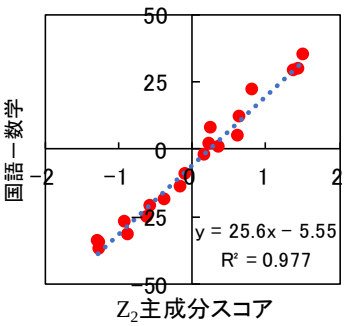
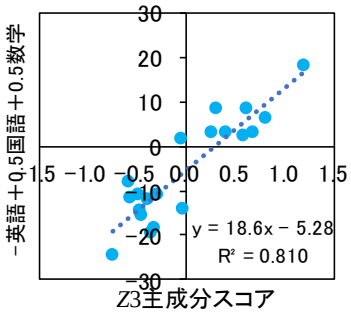


図５



z1主成分スコアと合計点およびZ2主成分スコアと［国語－数学］はいずれも高い相関を示しており，よく対応していた（図３，４）．しかし，z3主成分スコアと［－1英語＋0.5国語＋0.5数学］の相関は低かった（図５）．この理由として，重み係数の違いやz3の固有値は0.312と低く情報量が少ないことなどが影響したためと考えられた．

以上の結果を踏まえて，主成分ごとの特徴を検討した．

z1主成分スコアと合計点

z1主成分スコアが高い順に生徒IDを並べ替え，合計点と比較して検討した（図６）．

図６

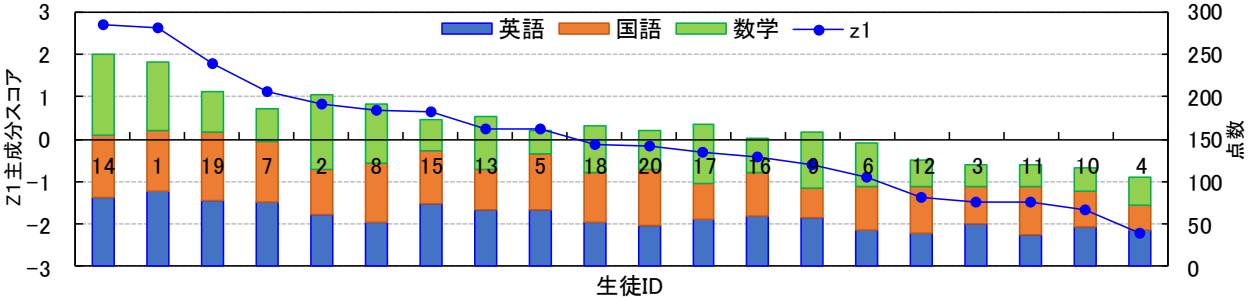


図６から，z1主成分スコアは合計点と良く対応しているが，一部生徒間ではそれらの高低が逆転していた．例えば，ID-7の合計点はID-2の合計点より低いが，z1主成分スコアはID-7のほうがID-2より高かった．これは，固有ベクトル（表３）は大きい順に英語，国語，数学であるが，ID-7は英語と国語の点数ほうが数学の点数より高いのに対して，ID-2は数学の点数ほうが英語と国語の点数よりも高いためである．

このように，z1主成分スコアは分散を最大化するように調整された固有ベクトルを用いていることに留意する必要がある．

偏差値について

z2とz3については，それぞれ各教科の相対的な得意・不得意を示す主成分なので，教科点数よりは偏差値に基づく値を用いたほうが主成分の特徴を調べるのに適している．

偏差値yは，得点xと平均値Aとの差を標準偏差σで割った値を10倍して，50を加えた値で

あり，平均が50，標準偏差が10の値となる．

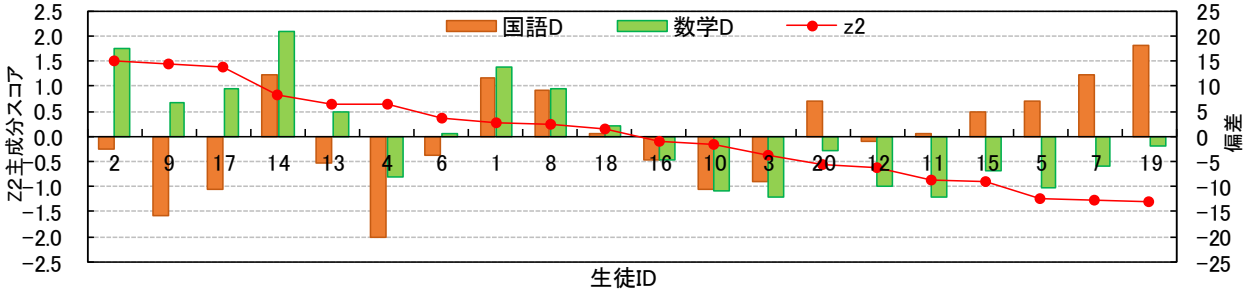
$$y = 10(x - A)/\sigma + 50 \qquad (1)$$

評価の指標として，式(1)の右辺の第一項（以降，偏差という）を用いる．偏差は，平均点からの点数のばらつきを示す．以降，各教科の偏差の名称は添え字Dを加えたもので表わす，例えば，英語の偏差は「英語D」とした．

z2主成分スコアと偏差による評価

z2主成分スコアが高い順に生徒IDを並べ替え，各生徒の［数学D－国語D］の違いを調べた（図７）．

図７

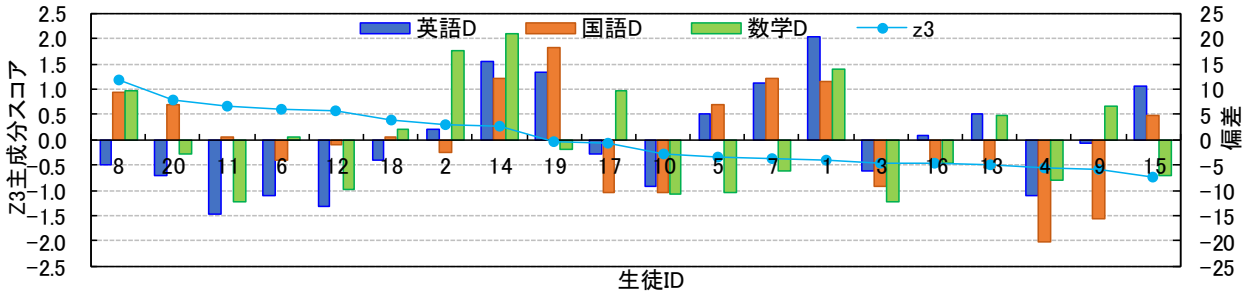


z2主成分スコアが大きいほど［数学D－国語D］は大きい傾向を示すが，数学Dと国語Dの値は生徒ごとに異なった．例えば，z2主成分スコアはID-2とID-9とID-17でいずれも1.5程度を示した．［数学D－国語D］がほぼ同じだからである．しかし，３人とも数学Dは平均値以上で，国語Dは平均以下であるが，それぞれ高低に違いがみられる．一方，z2主成分スコアが負数の場合は国語が得意で数学が不得意であることを示している．特に，ID-5とID-7とID-19はz2主成分スコアがほとんど同じだが，国語Dと数学Dがそれぞれ異なっていた．このように，z2の固有値は１以下（0.802）であったが，その特徴をよく表していた．

z3主成分スコアと偏差による評価

z3主成分スコアが高い順に生徒IDを並べ替えて，z3と各偏差との関係を調べた（図８）．

図８



z3主成分スコアの特徴は明確には認められない．z3の場合はz2の２教科の場合とは異なり３教科が影響し合うため，また，固有値が小さい（情報量が少ない）ため，成績評価に利用するのは難しいと考えられた．

気象データと教科点数の因子負荷量の比較

これ以降は，教科点数（横断面データ）から時系列データを推定する方法等について検討した．推定した時系列データから各教科の特性について何らかの知見が得られるかもしれないと考えたからである．その準備として，まず，前報（参考文献１）における気象時系列の主成分分析結果を併用して検討する．

前報では，周期時系列の実例として４年間における毎月の気象データ（日射量，気温，気圧）に相関行列による主成分分析を適用し検討した．結果の概要は，以下の通りである．

① 周期は，日射量と気温が12か月，気圧は12か月と6か月の合成と考えられた．

② 日射量と気温とは約2か月の位相差があり，気圧の12か月周期成分は日射量や気温とはほぼ逆位相の関係にあった．

③ 主成分分析の結果，主として12か月周期の第一主成分z1と，それに直交（90°＝3か月の位相差）した第二主成分z2，および6か月周期の第三主成分z3が抽出された．

これらの結果を念頭に置いて，気象データの因子負荷量（表４）と，教科点数の因子負荷量（表５）を比較検討した．

表４

	z1	z2	z3
日射量	0.828	−0.508	0.237
気温	0.827	0.513	0.232
気圧	−0.903	0.003	0.430

表５

	z1	z2	z3
英語	0.906	−0.067	−0.417
国語	0.784	−0.534	0.315
数学	0.668	0.718	0.196

まず，表４の日射量と表５の国語を比較してみると，各主成分の因子負荷量の符号は両者で一致し，因子負荷量の値もおおよそ一致している．気温と数学についても同様である．したがって，横断面データの国語と数学はそれぞれ時系列データの日射量と気温に対応すると考えられた．

一方，気圧と英語については，z2との因子負荷量はとも0に近い値を示し，また，英語における各主成分の因子負荷量の符号は気圧における各主成分の因子負荷量にマイナスをかけた符号と一致した．気圧は日射量や気温とは逆位相であったから，英語は国語や数学とはほぼ順位相であるとみなすことができる．

このように，符号の違いはあるものの，国語と数学は12か月周期，英語は12か月と6か月の２つの周期の合成であると仮定できる．この仮定を前提として，横断面データの教科点数から時系列データを推定することを試みた．

相関係数による各教科の時系列モデルの推定

因子負荷量から各教科の時系列を推定することはできないが，各教科点数の相関係数から時系列のパラメータ（位相差等）を推定することは可能である．各教科点数の相関係数（表２）はそれらの位相差のコサインで表されることなどを利用する．

各教科の時系列を下記のコサイン関数モデル（国語$N(t)$，数学$M(t)$，英語$E(t)$）で表した．また，以降の検討で使用する期間と間隔は，気象時系列（参考文献１）の場合と同様に，

４年間における毎月の値(t=1～48)とした．この場合，１か月＝30°に相当する．

$$N(t)=H_{N12}\cos[2\pi\,(t+\alpha_{N12})/12] \quad (1)$$

$$M(t)=H_{M12}\cos[2\pi\,(t+\alpha_{M12})/12] \quad (2)$$

$$E(t)=H_{E12}\cos[2\pi\,(t+\alpha_{E12})/12]+H_{E6}\cos[2\pi\,(t+\alpha_{E6})/6] \quad (3)$$

ここで，H_{N12}，H_{M12}，H_{E12}はそれぞれ12か月周期の国語，数学，英語の振幅，H_{E6}は6か月周期の英語の振幅である．また，α_{N12}，α_{M12}，α_{E12}はそれぞれ12か月周期の国語，数学，英語の位相（月），α_{E6}は6か月周期の英語の位相（月）である．

① 国語と数学はともに12か月周期である．同一周期モデルにおける位相差は相対的なものなので，どちらかの教科の位相を基準に定めれば良い．そこで，国語の位相α_{N12}＝0月とした．また，この場合，振幅は相関係数に影響しないので，$H_{N12}=H_{M12}=1$とした．そうすると，次の国語のモデル式が得られる．

$$N(t)=\cos[2\pi\,t/12] \quad (4)$$

② 国語の位相α_{N12}＝0月としたので，数学の位相α_{M12}は数学と国語の位相差に相当する．両教科の相関係数は0.202（表２）であるから，$\alpha_{M12}=12\arccos(0.202)/2\pi\fallingdotseq2.61$か月が得られた．したがって，数学のモデル式として次式が得られる．

$$M(t)=\cos[2\pi\,(t+2.16)/12] \quad (5)$$

国語と数学のモデル式により，時間t（1～48）におけるそれぞれの値が確定する．

③ 英語は12か月と6か月の合成周期であるから，パラメータは位相だけではなく，振幅も関わってくる．式(3)より，英語のパラメータは４つ存在するが，振幅については，両周期の振幅の比率が同じならば，相関係数は変わらない．そこで，12か月周期の振幅をH_{M12}＝１とした．なお，位相についても同様だが，ここでは考慮しない．理由は後述する．したがって，３つのパラメータを未知数とした英語のモデル式が得られる．

$$E(t)=\cos[2\pi\,(t+\alpha_{E12})/12]+H_{E6}\cos[2\pi\,(t+\alpha_{E6})/6] \quad (6)$$

式(6)において，３つのパラメータは英語と国語および英語と数学の２つの相関係数（表２）を同時に満たす必要がある．

英語のパラメータの求めるためにExcelのソルバーを用いた．その概要を述べる．

準備として，国語の値と英語の式(6)および数学の値と英語の式(6)をそれぞれ別個にExcelに作成しておく．それぞれについて，相関係数の関数式および英語の式(6)の３つのパラメータをそれぞれ別のセルに作成する．

まずに，国語と英語については，ソルバーにおける「目的セル」に相関係数の関数式のセルを，「目標値」に国語と英語の相関係数(0.615)を，そして，「変数セル」に３つのパラメータのセルを設定し，３つのパラメータのセルに初期値を入力したのち，計算を実行する．

次に，求められた３つのパラメータの値を，数学と英語の３つのパラメータのセルにコピーし，「目標値」に数学と英語の相関係数(0.476)を設定したのち，計算を実行する．同じ操作を相互に繰り返し，目標値が国語と英語の相関係数および数学と英語の相関係数の両方を満たすまで計算を行う．なお，繰り返し計算回数は，初期値をα_{E12}

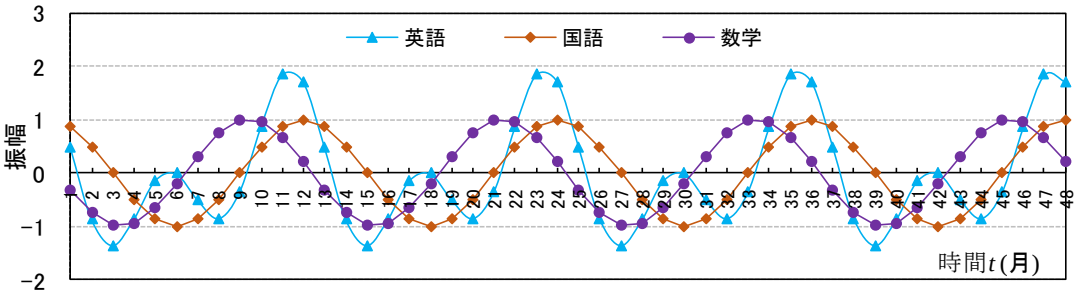
＝1，H_{E6}＝1，α_{E6}＝0.5とした場合，合計６回であった．

３教科のモデル式のパラメータをまとめて表６に，それらの時系列を図9に示した．

表６

	国語	数学	英語	
周期(月)	12	12	12	6
振幅	1	1	1	0.987
位相(月)	0	2.61	1.01	0.500

図９



数学は国語よりも2.61か月進んでいる．英語の振幅は，6か月周期が0.987を示し，12か月周期とほとんど同じであった．両周期の影響はほぼ等しいが，時系列図では，両周期がほぼ重なっているため，12か月周期に相当する振幅が6か月周期単独の振幅より高い．その12か月周期の最大振幅は国語と数学の最大振幅の間にあった．一方，英語の位相は，12か月周期で約1.01か月，6か月周期では約0.5か月を示した．ただし，実際は，6か月周期の位相はどんな値を入力しても，相関係数はほとんど変わらない．その位相を0.5か月としたのは，散布図が各教科の横断面データとその時系列の横断面データの分布が比較的類似しているからである．このように，英語は特徴的な周期特性を示している．

各教科のコサイン関数モデルの主成分分析

３教科のコサイン関数モデルに主成分分析を適用して得られた各主成分スコアの時系列を図10に示した．なお，このモデルでは３教科の点数の場合と同じ相関行列を用いているので，主成分分析結果は表３と一致する．

図10

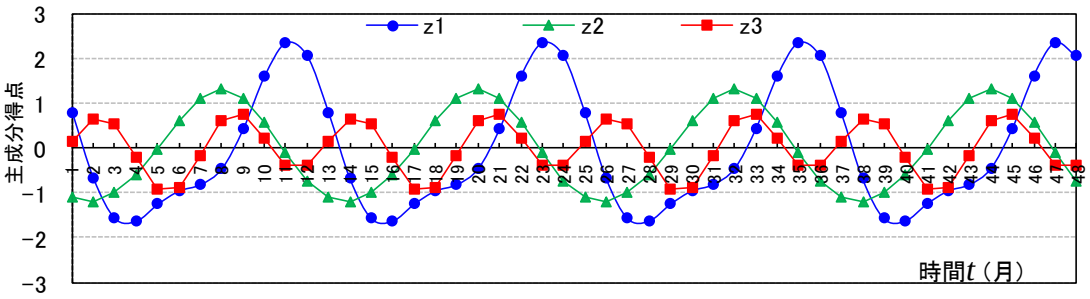


図10によると，z1は12か月周期が主体であるが，6か月周期の影響もみられた．最小二乗法を用いてz1を両周期成分のコサイン関数モデルに分解すると，6か月周期成分の振幅は12か月周期成分の振幅の約1/3であった．z1の12か月周期に相当する最大振幅の位相は３教科の位相のほぼ平均（図９）にあった．したがって，z1が合計点的な特徴を示すのは，

主として12か月周期成分が寄与しており，z1と各教科との位相差が各教科の違いを表していると考えられた．
z2はほぼ12か月周期のみで形成されていた．z2の位相はz1の12か月周期成分の位相より約３か月（90°）進んでおり，両者は直交していた．したがって，z2が国語より数学が得意であるという特徴を示すのは，同一周期にある国語と数学がz2との位相差によって両教科の違いを表していると考えられた．
z3については，12か月周期の影響もみられるが，6か月周期が主体となっている．最小二乗法を用いて両周期成分のコサイン関数モデルに分解すると，6か月周期成分の振幅は12か月周期成分の振幅の約2.7倍であった．6か月周期主体のz3が12か月周期だけから成る国語と数学や，両周期の合成である英語と関連付けるのは無理がある．そのため，z3の特徴をよく表現することができなかったと考えられる．

これまで，想像力を働かせて（働かせすぎて），３教科の相関係数から時系列モデルを推定し，それらの主成分分析結果について考察した．周期を次元に置き換えてみると，国語と数学は共通の次元のなかで両教科が対の関係（文系と理系？）にあり，英語は国語や数学とは異なる次元も含まれていると考えることができる．では，この場合の次元や対の関係とは何を意味するのだろうか？さらに多くの教科点数データについて検討を重ねていく必要がある．

謝辞
３教科の点数データの使用を快諾して頂いた我楽多頓陳館の杉本氏に感謝いたします．

参考文献
１．物理のかぎしっぽ，周期時系列の統計解析（13)周期時系列における主成分分析の意味
http://hooktail.sub.jp/contributions/shuki13.pdf
２．我楽多頓陳館，雑学の部屋，雑学コーナー，統計学入門，第16章　主成分分析
http://www.snap-tck.com/room04/c01/stat/stat16/stat1601.html