共同研究「経営実務データを用いたデータサイエンティスト育成方法の研究」報告書

2019 年 3 月 31 日

国立大学法人滋賀大学データサイエンス教育研究センター長

竹村　彰通

１．共同研究の趣旨

本共同研究の研究題目は「経営実務データを用いたデータサイエンティスト育成方法の研究」であり、昨年度に引き続き、本共同研究の目的は企業の経営実務データを用いて販売予測や適性な在庫管理など、これからのデータサイエンティストに求められる能力を育成するための方法を研究することである。本共同研究の期間は、平成 30 年 10 月 1 日から平成 31 年 3 月 31 日までである。

特定非営利活動法人ビュー・コミュニケーションズは実際の販売データを教育目的のために適切に処理をおこなった上で国立大学法人滋賀大学に提供し、滋賀大学データサイエンス教育研究センターはデータをプロジェクト型の演習の形に整備する作業を担うことで、データサイエンティスト育成方法に関する共同研究をおこなった。

本年度はビュー・コミュニケーションズから、ホームセンターにおけるシャンプー商品 145 品とペットフード商品 100 品の販売実績の大規模なデータが提供された。このデータは滋賀大学に平成 29 年 4 月 1 日に開設された日本初のデータサイエンス学部の学生にとって、ビッグデータの一端を実感できる実務データであり、特に回帰分析、多変量解析、時系列解析の各種手法を学んだばかりの２年次生の演習のために非常に有用である。その点で本共同研究の意味は大きい。

２．滋賀大学データサイエンス学部の育成人材像と演習の重視

滋賀大学データサイエンス学部では、データサイエンスの基礎要素技術としてのデータエンジニアリング(情報工学)及びデータアナリシス(統計学)に加えて、データからの価値創造の能力の育成を重視している。

この価値創造の能力を育成するためには、実際のデータを用いた演習の中で、試行錯誤や成功体験が必要でありさらに教科書的な理論と実際のデータとの乖離についても経験を積む必要がある。ビュー・コミュニケーションズから提供されたデータは、データ分析における前処理及び試行錯誤を伴う分析作業を経験するためにも、また理論と実際との乖離を経験するためにも有用である。理論と実際の乖離については、ビュー・コミュニケーションズ副理事長の小松秀樹氏の著書「なぜあなたの予測は外れるのか－AI が起こすデータサイエンス革命」においても多くの例とともに示されており、データサイエンス教育において実際のデータを扱うことが非常に重要であることがわかる。小松秀樹氏には 2019 年 10 月 12 日に滋賀大学データサイエンス学部 1 年生向けの講義を担当いただき、実務データの見方や扱い方に関するノウハウを伝えていただいた。

滋賀大学は、2016 年 12 月に「数理及びデータサイエンスに係る教育強化」の６拠点校の一校として、北海道大学、東京大学、京都大学、大阪大学、九州大学とともに文部科学省より選定を受けた。これは滋賀大学のデータサイエンス教育の先進性が高く評価されたためであると考えられる。特に、ビュー・コミュニケーションズから提供されたデータを用いた演習教材のような、価値創造につながる教育コンテンツの開発は拠点校として滋賀

大学が求められている活動である。

３．滋賀大学データサイエンス学部における演習の設計

滋賀大学データサイエンス学部における演習の設計は「PPDAC サイクルを回す」という考え方に基づいている。PPDAC サイクルとは、問題解決における各段階を Problem（問題）、Plan（調査の計画）、Data（データ）、Analysis（分析）、Conclusion（結論）に分けるという考え方である。ビュー・コミュニケーションズ提供のデータに基づく演習においても、データをそのまま学生に示すのではなく、例えば販売と仕入れの関係をどのように考えるか(Problem)、その問題を考えるときにどのようなデータが必要か(Plan)、計画した問題を解決するためにどのようなデータが必要か、またどのように加工すればよいか(Data)なども考えさせることとしている。これについては以下の５節に例示する。

データの分析手法や結論の導き方については、2 年次で履修する回帰分析、多変量解析入門、時系列解析入門だけではなく、3 年次で履修する様々な分析手法を学んだ後のほうが、深い分析が可能となる。しかしながら、滋賀大学データサイエンス学部では、まず手法を学ぶのではなく、1, 2 年次から様々な分野のデータを用いた PPDAC サイクルの繰り返しを重視し、データから出発して様々な分析を実施し、その結果に基づき次に何が必要かを経験させることとしている。本年度ビュー・コミュニケーションズから提供されたデータは 2 年次の演習に適切と考えられるため、本共同研究では、2 年生向けの演習であるデータサイエンスフィールドワーク演習で活用した。

４．本年度ビュー・コミュニケーションズより提供されたデータ

本年度ビュー・コミュニケーションズから提供されたデータは、あるホームセンターのシャンプー商品（ボトルタイプ、詰替えタイプ、ボディシャンプーも含む）145 品とペットフード商品（犬用、猫用、缶詰タイプ、袋タイプ）100 品についての 2014 年 12 月 22 日から 2017 年 12 月 17 日までの週次販売数である。また各商品について、次のような情報も付加されていた。

●シャンプー商品
匿名化されたメーカー名
ボトルタイプか詰替えタイプか
（分析段階で値段情報も追加）
●ペットフード商品
匿名化されたメーカー名
犬用か猫用か
缶詰タイプか袋タイプか
内容量

本学部 2 年後期で実施されるデータサイエンスフィールドワーク演習において、履修者に本データの分析を実施させた。本データは次のような xlsx 形式のファイルで提供した。

シャンプーデータ

	A	B	C	D	E	F	G	H	I	J
1	分類	値段	20141222	20141229	20150105	20150112	20150119	20150126	20150202	20150209
2	Q社_商品1_詰替	2592	0	0	0	0	0	0	0	0
3	A社_商品1_詰替	1140	8	13	3	3	13	9	13	20
4	A社_商品2_詰替	428	9	4	4	12	12	5	19	7
5	B社_商品1_詰替	348	0	0	0	0	0	0	0	0
6	C社_商品1_詰替	702	0	0	0	0	0	0	0	0
7	B社_商品2_詰替	570	0	0	0	0	0	0	0	0
8	B社_商品3_詰替	819	0	0	0	0	0	0	0	0
9	A社_商品3_詰替	428	4	4	5	4	6	4	9	3
10	C社_商品2_詰替	886	3	1	3	5	2	2	5	3

ペットフードデータ

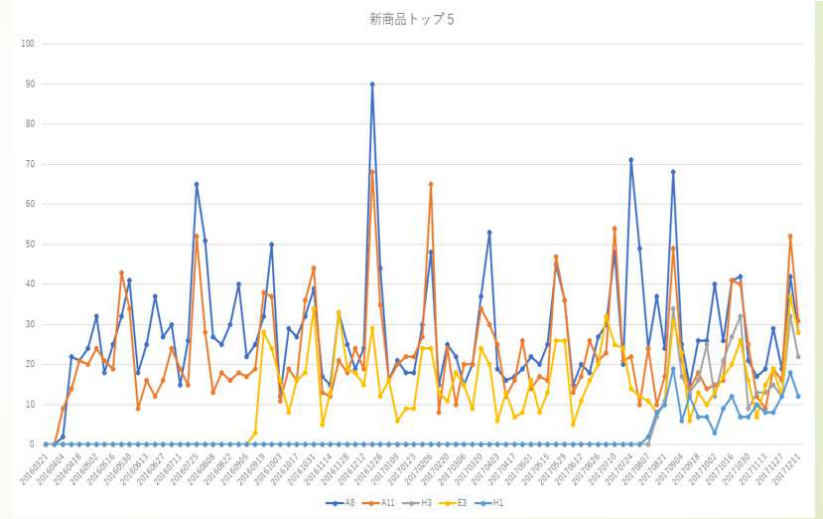
	A	B	C	D	E	F	G	H	I
1		20141222	20141229	20150105	20150112	20150119	20150126	20150202	20150209
2	A社_商品1_犬用_缶_400g	34	39	77	13	37	20	17	86
3	B社_商品1_犬用_缶_400g	0	0	0	0	0	0	0	0
4	A社_商品2_犬用_缶_400g	36	39	3	9	5	70	23	30
5	C社_商品1_犬用_缶_100g	7	2	5	4	1	1	26	5
6	C社_商品2_犬用_缶_100g	7	5	11	20	3	8	7	2
7	C社_商品3_犬用_缶_100g	23	12	10	19	5	24	30	32
8	D社_商品1_猫用_缶_70g	32	10	29	17	19	31	11	28
9	B社_商品2_犬用_缶_400g	0	0	0	0	0	0	0	0
10	D社_商品2_猫用_缶_70g	31	25	31	20	47	19	29	51

これらのデータは分類情報が文字列となっていること、また販売開始時期の異なるデータが混在していることから、データ分析を行う前にデータの前処理も必要となり、この部分に時間が費やされることが想定される。また、予測や分類を行うには各種手法が必要となり、これらのスキルの向上を目的とした。

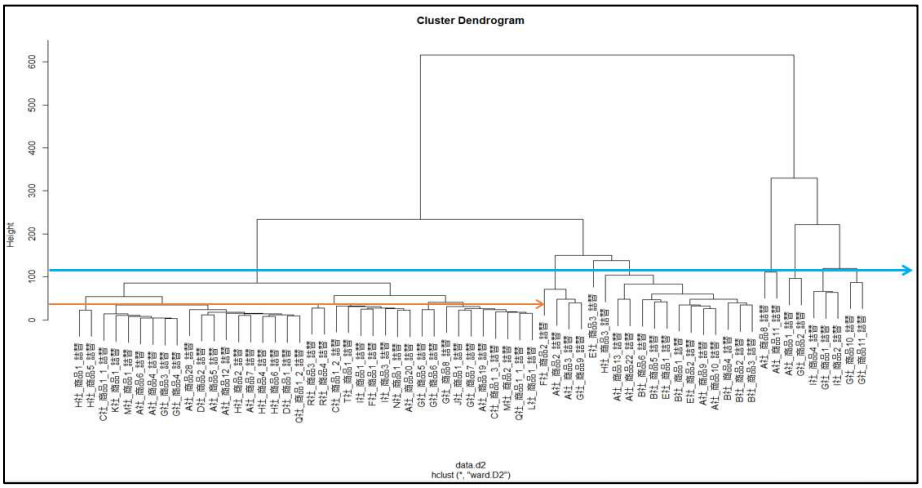
５．提供データを用いた演習課題の例

本データに基づく分析を 5 グループ（1 グループ 5 名程度）が実施し、販売傾向の抽出や季節の影響など様々な観点からの分析が行われた。

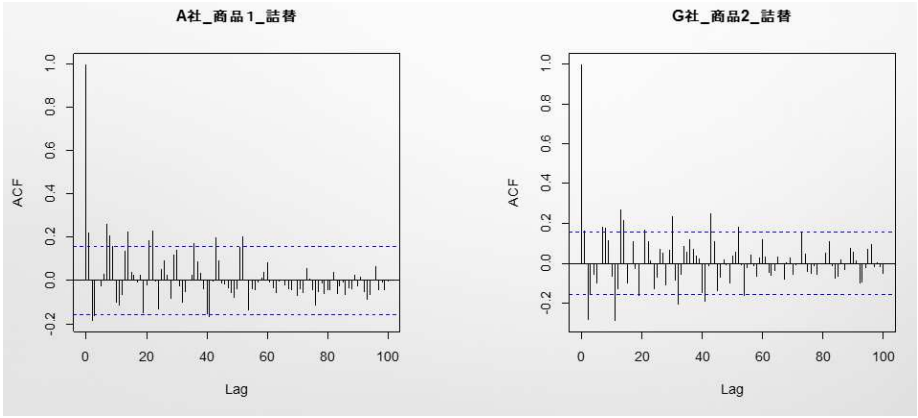
例１：時系列ｸﾞﾗﾌによる販売傾向の把握



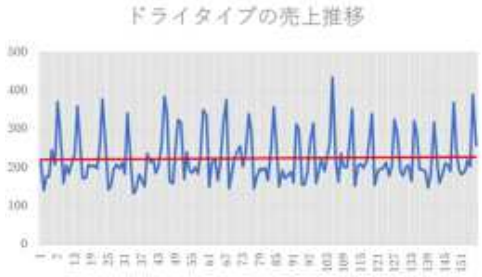
例２：ｸﾗｽﾀｰ分析による商品分類



例３：コレログラムによる周期性の抽出



例４：回帰分析による売上傾向の分析



６．まとめ

想定された通り、データの前処理にかなりの時間を要していた。また、自由な発想力を見るために問題設定も自由としたが、どのような問題を設定するかという点についても苦労したようである。また、実データの分析にまだ慣れていないこともあり、仮説に対する適切な手法の選択、各種手法の分析結果の解釈、商品知識に関する知識不足等、様々な課題があったが、本データ分析を通して、データの前処理の重要性、これまで学んだプログラムスキル（R または Python のスキル）の向上、各種手法の実践的活用、手法の理解が深まったと考えられる。最終発表においても、誤った分析も若干含まれていたが、今回のような実データの分析は座学だけでは身につきにくいスキルを向上させるのに有効であった。

滋賀大学データサイエンス学部としては、今後もさらに興味深い演習課題をビュー・コミュニケーションズとの共同研究の形で開発して行きたいと考えている。また、数理及びデータサイエンスに係る教育強化の拠点校としては、このような教育コンテンツを全国展開していくことも重要な課題である。

（追加報告）

事例研究：時系列データへの主成分分析適用トラップ

2019 年 3 月 31 日

特定非営利活動法人　ビュー・コミュニケーションズ

副理事長　小松　秀樹

データサイエンス学部の大学2年生が行った、実務上の時系列データに対する分析・解釈を、教育教材研究の視点から評価してみた。多変量解析の基礎学習を行った学生が、それらをどの様に適用し、どの程度有用な結論を導き出せるのか評価することは、学生に限らず広く実業界における人材育成に対しても多くの気づきを与えることが期待される。

（学生の演習概要）

1.提供データ　　小売業A社のﾍﾟｯﾄﾌｰﾄﾞ100商品の週次売上数量(156週､16社分)

2.データ期間　　2014年12/22～2017年12/11

3.学生グループ数　　5グループ（1グループ　5～6名）

4.演習期間　　2018年10月～11月

本稿で取り上げるのは、5グループ中2グループが行った主成分分析の事例である。1930 年代にハロルド・ホテリングが名付けた主成分分析は、多次元データを少次元に縮約する手法として知られている。ここで、主成分分析の概要を振り返っておく。

一般的に述べると、P個の変量　$X=(X_1、X_2・・・、Xp)^T$が与えられたとき、Xを線形結合させたZの分散を最大化させるベクトルCを求める問題すなわち、共分散行列の固有値問題と捉えられる。

$Z=C_1X_1+C_2X_2+・・・+CpXp$　（ただし、$C^TC=1$）

$Max\ C^T\ \Sigma_x\ C=\ e^T\ \Sigma_x\ e=\lambda$（$\Sigma_x$は共分散行列、eは1に基準化された固有ベクトル）

この分析方法は、そもそも多次元を少次元に縮約する必要があるのか、線形結合で表現されるZに有意な解釈が与えられるのか、変量データの異常値が相関を歪めてはいないか等々多くの問題をはらむ手法でもある。

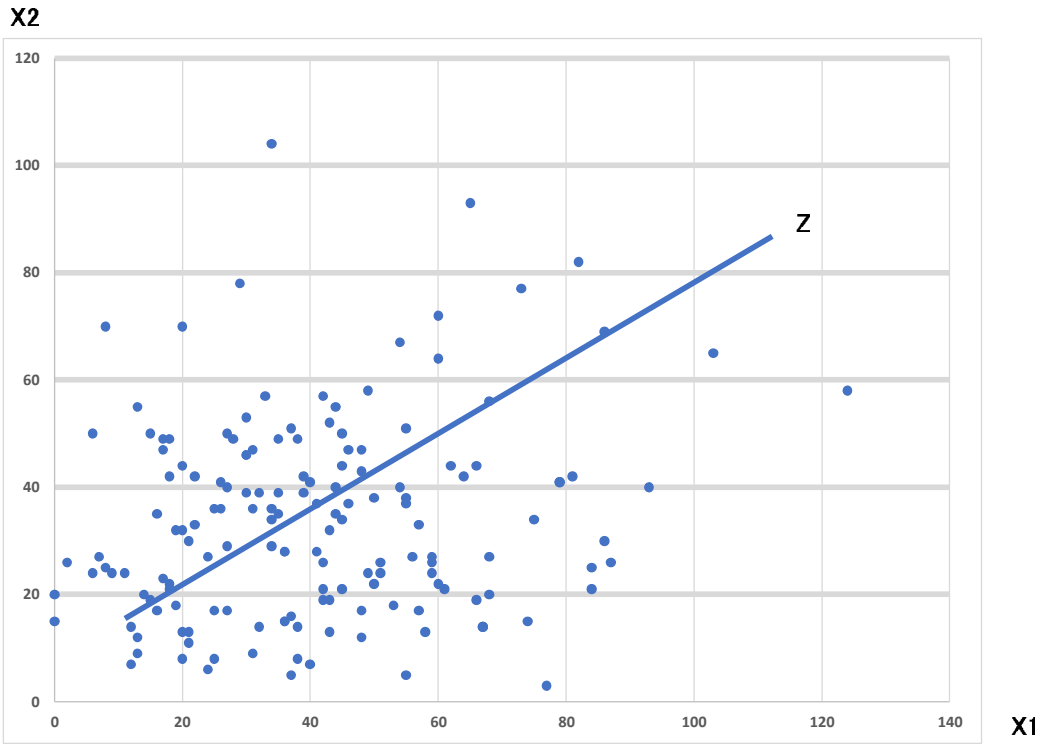
さて、この分析方法を時系列データに適用した学生グループが2グループあった。学生グループの分析・解釈を追いかけるために、仮想的にその過程を設定し、分析・解釈結果がどの様に出されたものかを推定することとした。

そこで、学生グループが、A社ペットフード商品A_1、A_2の156週次売上数量X_1、X_2（次表）を分析対象として、主成分分析を紙と鉛筆で解いていくと想定してみる（安易に既存の統計ソフトを使わないとする）。

	X1	X2
20141222	34	36
20141229	39	39
20150105	77	3
20150112	13	9
20150119	37	5
20150126	20	70
20150202	17	23
20150209	86	30
20150216	19	32
20150223	20	8
20150302	49	24
20150309	0	20
20150316	51	24
20150323	22	42
20150330	25	8
20150406	18	21
20150413	43	13
20150420	21	13
20150427	34	34
20150504	45	21
20150511	74	15
20150518	12	7
20150525	13	12
20150601	31	36
20150608	6	24
20150615	53	18
20150622	84	25
20150629	27	17
20150706	21	30
20150713	40	7
20150720	57	17
20150727	81	42
20150803	28	49
20150810	30	39
20150817	19	18
20150824	20	44
20150831	32	39
20150907	31	9
20150914	24	6
20150921	35	39
20150928	24	27
20151005	55	5
20151012	7	27
20151019	93	40
20151026	45	34
20151102	48	12
20151109	41	28
20151116	36	15
20151123	42	21
20151130	61	21
20151207	66	44
20151214	50	38
20151221	0	15
20151228	44	55
20160104	29	78
20160111	39	42
20160118	43	19
20160125	25	17
20160201	103	65
20160208	6	50
20160215	32	14
20160222	42	26
20160229	66	19
20160307	12	14
20160314	59	26
20160321	55	38
20160328	20	13
20160404	48	47
20160411	43	52
20160418	68	20
20160425	22	33
20160502	20	32
20160509	84	21
20160516	30	46
20160523	62	44
20160530	60	22
20160606	14	20
20160613	59	24
20160620	30	53
20160627	50	22
20160704	58	13
20160711	2	26
20160718	56	27
20160725	38	49
20160801	48	17
20160808	36	28
20160815	38	14
20160822	46	47
20160829	38	8
20160905	18	22
20160912	18	42
20160919	87	26
20160926	15	50
20161003	51	26
20161010	8	25
20161017	45	50
20161024	68	27
20161031	18	49
20161107	46	37
20161114	48	43
20161121	35	35
20161128	79	41
20161205	35	49
20161212	44	35
20161219	124	58
20161226	54	40
20170102	27	50
20170109	54	67
20170116	67	14
20170123	67	14
20170130	40	41
20170206	79	41
20170213	26	36
20170220	42	19
20170227	57	33
20170306	27	29
20170313	36	28
20170320	68	56
20170327	60	72
20170403	13	55
20170410	41	37
20170417	55	51
20170424	21	11
20170501	34	29
20170508	9	24
20170515	26	41
20170522	64	42
20170529	33	57
20170605	59	27
20170612	44	40
20170619	17	47
20170626	43	32
20170703	75	34
20170710	34	104
20170717	86	69
20170724	36	15
20170731	16	17
20170807	11	24
20170814	82	82
20170821	16	35
20170828	49	58
20170904	45	44
20170911	8	70
20170918	27	40
20170925	42	57
20171002	25	36
20171009	37	16
20171016	65	93
20171023	37	51
20171030	15	19
20171106	17	49
20171113	60	64
20171120	30	46
20171127	31	47
20171204	73	77
20171211	55	37

（学生グループが行う仮想的分析・解釈過程）

（１）　商品 A_1, A_2 の売上数量散布図を作成しながら、Z 軸を設定する



主成分分析を行う変量は X_1, X_2 なので、合成変数 Z（ここでは第一成分のみ考える）として

$$Z = C_1X_1 + C_2X_2$$

その分散が最大となる C_1, C_2 を求める訳であるが、ここで少し準備をしておく。

（２）　各種の定義をする

変量　　$\boldsymbol{x}_1 = \{x_{1\,1}, x_{1\,2}, \cdots, x_{1\,t}\}$　　$\boldsymbol{x}_2 = \{x_{2\,1}, x_{2\,2}, \cdots, x_{2\,t}\}$　（t=1～156 週，$x_{1\,t}$は観測値）

分散　　$V_{x_1} = \frac{1}{n}\sum_{t=1}^{n}(x_{1\,t} - \bar{x}_1)^2$　　$V_{x_2} = \frac{1}{n}\sum_{t=1}^{n}(x_{2\,t} - \bar{x}_2)^2$

共分散　　$S_{12} = \frac{1}{n}\sum_{t=1}^{n}(x_{1\,t} - \bar{x}_1)(x_{2\,t} - \bar{x}_2)$

よって、Zの分散は

$$\begin{aligned} V_z &= \frac{1}{n}\sum_{t=1}^{n}(z_t - \bar{z})^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}\{c_1(x_{1t} - \bar{x}_1) + c_2(x_{2t} - \bar{x}_2)\}^2 \\ &= c_1^2V_{x_1} + 2c_1c_2S_{12} + c_2^2V_{x_2} \end{aligned}$$

また、c_1，c_2をZ軸が$\boldsymbol{x}_1$，$\boldsymbol{x}_2$軸となす角θ_1，θ_2の余弦$(\cos\theta_1, \cos\theta_2)$と考えると、$c_1^2 + c_2^2 = 1$。

（３）　Zの分散最大化を解く。

ラグランジュ未定乗数法より、制約条件$c_1^2 + c_2^2 = 1$のもと、

$$f(c_1, c_2, \lambda) = c_1^2 V_{x_1} + 2c_1 c_2 S_{12} + c_2^2 V_{x_2} - \lambda(c_1^2 + c_2^2 - 1)$$

の最大値を求めればよい。その条件は

$$\begin{aligned} \frac{\partial f}{\partial c_1} &= 2V_{x_1} c_1 + 2S_{12} c_2 - 2\lambda c_1 = 0 \\ \frac{\partial f}{\partial c_2} &= 2S_{12} c_1 + 2V_{x_2} c_2 - 2\lambda c_2 = 0 \\ \frac{\partial f}{\partial \lambda} &= 1 - (c_1^2 + c_2^2) = 0 \end{aligned}$$

以上より、

$$\begin{aligned} V_{x_1} c_1 + S_{12} c_2 &= \lambda c_1 \\ S_{12} c_1 + V_{x_2} c_2 &= \lambda c_2 \end{aligned}$$

書き換えると

$$\begin{pmatrix} V_{x_1} & S_{12} \\ S_{12} & V_{x_2} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \lambda \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

これは分散共分散行列$A = \begin{pmatrix} V_{x_1} & S_{12} \\ S_{12} & V_{x_2} \end{pmatrix}$の固有値問題となる。

したがって、固有方程式$|A - \lambda I| = 0$（I：単位行列）を解けばよいこととなる。

$$\left| \begin{pmatrix} V_{x_1} & S_{12} \\ S_{12} & V_{x_2} \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right| = 0$$

より、$\lambda^2 - (V_{x_1} + V_{x_2})\lambda + V_{x_1} V_{x_2} - S_{12}^2 = 0$の解（固有値）$\lambda_1$，$\lambda_2$が求まる。

（この場合は第１成分z_1のみ考えているので、$\lambda_1 \geq \lambda_2$としてλ_1を求める）

（４）事例の答

商品A_1，A_2の週次売上数量（$\boldsymbol{x_1}$，$\boldsymbol{x_2}$）データより、

$$V_{x_1} = 22.6 \quad V_{x_2} = 18.6 \quad S_{12} = 77.8$$

と計算されるので

$$\begin{vmatrix} 22.6-\lambda & 77.8 \\ 77.8 & 18.6-\lambda \end{vmatrix} = \lambda^2 - (22.6+18.6)\lambda + 22.6 \cdot 18.6 - 77.8^2 = 0$$

より、固有値$\lambda_1 = 98.4$。

固有ベクトルは$22.6c_1 + 77.8c_2 = 98.4c_1$より、$C = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.97 \end{pmatrix}$

$z_1 = x_1 + 0.97x_2$と、第１主成分が求められた。

(5) Ｚ軸の解釈

変量X_1，X_2に同じような値の係数（因子負荷量でも同様)、1、0.97 を乗じた形がＺなので、X_1，X_2を足したものがＺと読める。あえて言えば、商品A_1，A_2の販売合計Ｚは総合販売力軸と解釈されよう。

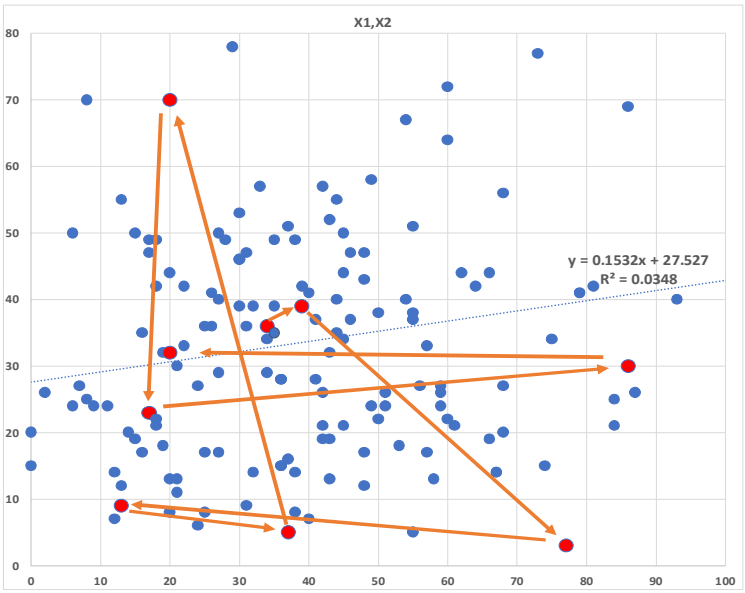
実際の学生グループの解釈も、「第一主成分は総合の売り上げ量を表していると思われる」としている。仮想グループ、実際のグループとも解釈結果*に大差ない。

しかし、ペットフードメーカーの開発現場やこれら商品を扱う小売現場で、各商品の販売合計をもって総合販売力と解釈してみせると、「そんな答えなら分析などいらない、自明のことだ」などかなり冷ややかな反応が得られるだろう。

*実際の学生グループは、カリキュラム上、統計ソフトも学習対象となっている為か、既存の統計ソフト（R）を用いて分析作業を行っており、紙と鉛筆で式を解いておらず、分析方法理解度に関する評価は困難であった。以下のような評価が得られると、実用データの演習用教材としての与え方がより効果的なものになろう。

・観測値、変量、確率変数の概念理解度
・平均、分散、共分散、相関係数などの基本統計量の扱い習熟度と意味理解度
・固有値問題の意味理解度　etc

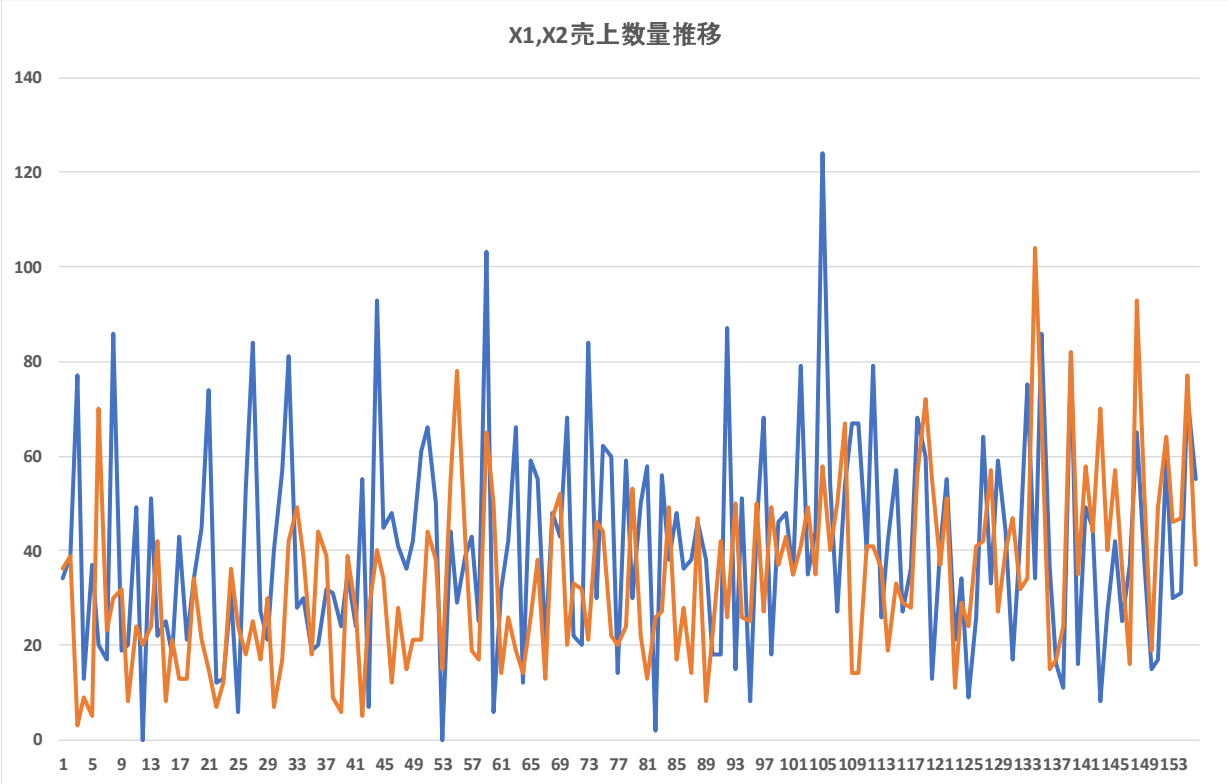
最大の問題は、X_1、X_2の時系列データを時間を無視したX_1、X_2の散布図から主成分分析を始めようとした点にある。時間を無視した X_1、X_2の散布図でX_1とX_2の相関をとると、相関係数（R）は 0.187 で無相関となる。無相関な変量を結合させた Z 軸に有意な解釈を与えようとするのは、別の与件がない限り困難である。また、図 B のように座標（X_1, X_2）の時間遷移を見ても、時間に対して無関係な動きをしている様子が分かる。



時系列データに主成分分析を行うことの危うさは、これまでも指摘されてきた。主成分分析を含む多変量解析の枠組みでは、観測データをある分布からランダムサンプリングされたものと捉えている。つまり、抽出順序は結果に影響を与えないと考えて分析が行われる。そのため、時間軸における順序とそれに伴う相関が重要となる時系列データには不向きと言える。無論、時間を超えて変量間の相関が存在する場合もあり（時間スケールが問題とはなるが）、主成分分析を時系列データに適用しようとする試みも考えられる（ex. 景気指数インデックス）。

学生が行うべき分析の第一歩は、時間順序とともに変量X_1, X_2の売上数量がどの様な変動特性を示していたかを推移グラフで可視化させることである。平均、分散、共分散等基本統計分析、多変量解析、時系列解析に取りかかる前に、可視化作業が重要な手順となる。

実業においても、目標値（前年比的に与えられることが多い）・実績値と言った数値は多変量・時系列データの為、データ量が膨大になってしまう。そこで、可視化する作業がおろそかにされることが多い。



図Cの X_1, X_2 の売上数量推移グラフを見て気ずく事は多々ある。X_1 は 156 週を通じて明確な増加傾向は見られない。不定期的に売上数量が多い週が見られる。X_2 は明確な増加傾向が見られ、2017 年には X_1 を上回る売上になっている。・・・

この様に、可視化することで重要な情報が得られることが多い。時系列解析の第一歩は、可視化作業にあり、この作業の重要性は学生、実業とも変わらない。

次のステップとしては、X_1, X_2 の分散、季節変動性等を扱うとよい（売上の安定化や季節対応は、重要な戦略である）。これらのステップを経て本格的な時系列解析に進むべきである。

異なる時点の売上数量の相関性（自己相関）は時系列解析の入口であるが、多変量時系列解析に至るまでの過程は実業では飛ばした方が経営組織戦略的に望ましい。解析方法のイメージ共有がほぼ不可能に近いからである。しかし、専門的

学習を目的にする学生の場合は、ここからが学習しがいがあるところである。ＡＲＭＡモデル、モデリングと予測、長期記憶モデル、ＶＡＲモデル、Granger 因果関係、カルマンフィルタ、分散不均一モデル（ＡＲＣＨモデル）、スペクトル解析との関係などがテーマとなる。

本稿では、主成分分析と時系列分析の観点から、大学 2 年生が陥りやすい罠を見てきたが、実業の世界でも同様のことが多い事も合わせて指摘した。