主成分分析

はじめに

主成分分析（Principal component analysis, PCA）は、データの次元圧縮を行う際に使用される統計的分析手法です。最も多くの分散を含むように無相関の線形結合を求めることで分析における変数の数を削減します。次元圧縮に加えて、データ構造を把握するために PCA から求められる固有ベクトルを分析することもあります。

PCA の目的は、分散を最大化するような変数の線形結合を求めることです。第一主成分の分散は、最大となります。第二主成分の分散は、第一主成分と無相関な全ての線形結合において最大となります。最後の主成分の分散は変数の線形結合において最小となります。全ての主成分を合わせると元の変数と同じ情報をもっていますが、各主成分は直行しており、順位の高い主成分ほど多く情報を保持しています。このように、PCA はデータの線形変換といえます。

PCA はデータが間隔尺度であることを前提としますが、特定の統計モデルを満たすことを想定していません。PCA はスケールに依存するため、共分散行列の主成分と相関行列の主成分とは結果が異なります。応用研究において、共分散行列の PCA は変数が共通の尺度である場合にのみ有効です。

例題 1: 聴力測定データの主成分分析

ここではサンプルデータとして、39 歳の方、100 名の聴力の測定データを使用します。測定値は、左耳と右耳について 4 つの異なる周波数で認識できる最小の強度です。 変数 lft1000 は、1,000Hz の左耳を指します。

```
. use https://www.stata-press.com/data/r17/audiometric
. correlate lft* rght*
```

	lft500	lft1000	lft2000	lft4000	rght500	rght1000	rght2000	rght4000
lft500	1.0000							
lft1000	0.7775	1.0000						
lft2000	0.4012	0.5366	1.0000					
lft4000	0.2554	0.2749	0.4250	1.0000				
rght500	0.6963	0.5515	0.2391	0.1790	1.0000			
rght1000	0.6416	0.7070	0.4460	0.2632	0.6634	1.0000		
rght2000	0.2372	0.3597	0.7011	0.3165	0.1589	0.4142	1.0000	
rght4000	0.2041	0.2169	0.3262	0.7097	0.1321	0.2201	0.3746	1.0000

上記の出力結果より、同じの耳の測定値は左右の耳の測定値より高い相関を持ち、左右の耳においても同じ周波数では高い相関を持つことがわかります。これらの変数は共通の単位のため、変数の共分散行列を分析することは理論的に意義があります。しかし、測定値の分散は大きく異なっています。

```
. summarize lft* rght*, sep(4)
```

Variable	Obs	Mean	Std. dev.	Min	Max
lft500	100	-2.8	6.408643	-10	15
lft1000	100	-.5	7.571211	-10	20
lft2000	100	2	10.94061	-10	45
lft4000	100	21.35	19.61569	-10	70
rght500	100	-2.6	7.123726	-10	25
rght1000	100	-.7	6.396811	-10	20
rght2000	100	1.6	9.289942	-10	35
rght4000	100	21.35	19.33039	-10	75

分散を用いて分析を行う場合、高い周波数の測定値が結果を左右することになります。このような影響に対する臨床的な理由はありません。そこで、相関行列を分析します。

```
. pca lft* rght*
```

Principal components/correlation

Number of obs	=	100
Number of comp.	=	8
Trace	=	8
Rho	=	1.0000

Rotation: (unrotated = principal)

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.92901	2.31068	0.4911	0.4911
Comp2	1.61832	.642997	0.2023	0.6934
Comp3	.975325	.508543	0.1219	0.8153
Comp4	.466782	.126692	0.0583	0.8737
Comp5	.34009	.0241988	0.0425	0.9162
Comp6	.315891	.11578	0.0395	0.9557
Comp7	.200111	.0456375	0.0250	0.9807
Comp8	.154474	.	0.0193	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8
lft500	0.4011	-0.3170	0.1582	-0.3278	0.0231	0.4459	0.3293	-0.5463
lft1000	0.4210	-0.2255	-0.0520	-0.4816	-0.3792	-0.0675	-0.0331	0.6227
lft2000	0.3664	0.2386	-0.4703	-0.2824	0.4392	-0.0638	-0.5255	-0.1863
lft4000	0.2809	0.4742	0.4295	-0.1611	0.3503	-0.4169	0.4269	0.0839
rght500	0.3433	-0.3860	0.2593	0.4876	0.4975	0.1948	-0.1594	0.3425
rght1000	0.4114	-0.2318	-0.0289	0.3723	-0.3513	-0.6136	-0.0837	-0.3614
rght2000	0.3115	0.3171	-0.5629	0.3914	-0.1108	0.2650	0.4778	0.1466
rght4000	0.2542	0.5135	0.4262	0.1591	-0.3960	0.3660	-0.4139	-0.0508

Variable	Unexplained
lft500	0
lft1000	0
lft2000	0
lft4000	0
rght500	0
rght1000	0
rght2000	0
rght4000	0

pca コマンドは 2 つの表を表示します。1 つ目の表は、相関行列の固有値の一覧であり降順で表示されます。各固有値に対応する固有ベクトルが、2 つ目の表に一覧表示されます。これらは主成分であり、負荷量を 2 乗して列和をとると 1 となります。(例えば、$0.4011^2 + 0.4210^2 \ + \cdots + 0.2542^2 = 1$)

備考：因子分析と連動して主成分を扱うソフトウェアでは、関連する固有値に正規化された主成分を表示する傾向にあります。正規化については、推定後コマンド（estat loadings）で計算できます。

固有値は、変数の分散の合計になります。相関行列を分析しているため、変数は標準化されており、分散の合計は 8 になります。また各固有値は、主成分の分散です。第一主成分の分散は 3.93 で、全体の 49%(3.93/8)を説明しています。第二主成分の分散は 1.62 で、全体の 20%(1.62/8)を説明しています。主成分同士は無相関です。例えば、

$0.4011 \times (-0.3170) + 0.4210 \times (-0.2255) + \cdots + 0.2542 \times 0.5135 = 0$となります。

結果的に 2 つの主成分により全体の分散の 69%を説明できると言えます。成分が相関する場合、部分的に同じ情報を説明しており、その重複した情報のために全体の分散とは等しくなりません。8 つの主成分を結合することで全ての変数における分散を説明します。したがって、2 つ目の表における説明されなかった分散の一覧は全て 0 であり、1 つ目の表に示される通り Rho=1.00 となります。

分散の 85%以上を説明するには、4 つの主成分を参照する必要があります。オプション componets(4)を指定して、これらの成分の一覧を表示できます。

```
. pca lft* rght*, components(4)
```

```
Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          4
                                                  Trace            =          8
    Rotation: (unrotated = principal)             Rho              =     0.8737
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.92901	2.31068	0.4911	0.4911
Comp2	1.61832	.642997	0.2023	0.6934
Comp3	.975325	.508543	0.1219	0.8153
Comp4	.466782	.126692	0.0583	0.8737
Comp5	.34009	.0241988	0.0425	0.9162
Comp6	.315891	.11578	0.0395	0.9557
Comp7	.200111	.0456375	0.0250	0.9807
Comp8	.154474	.	0.0193	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
lft500	0.4011	-0.3170	0.1582	-0.3278	.1308
lft1000	0.4210	-0.2255	-0.0520	-0.4816	.1105
lft2000	0.3664	0.2386	-0.4703	-0.2824	.1275
lft4000	0.2809	0.4742	0.4295	-0.1611	.1342
rght500	0.3433	-0.3860	0.2593	0.4876	.1194
rght1000	0.4114	-0.2318	-0.0289	0.3723	.1825
rght2000	0.3115	0.3171	-0.5629	0.3914	.07537
rght4000	0.2542	0.5135	0.4262	0.1591	.1303

表示された 1 つ目の表は、先ほど表示したものと変わりありません。2 つ目の表には、4 つの主成分が一覧表示されています。これらの 4 つの成分ではデータに含まれる全ての情報を説明できていません。言い換えると、変数における分散の一部は考慮されていません。この考慮されていない分散は、削除された成分の負荷量（関連する固有値により重みづけされた値）の 2 乗和に等しくなります。説明されていない分散の平均は、説明されていない分散の 13%(1-0.87)に等しくなります。

主成分をより詳しく見てみましょう。第一成分は全ての変数においておおよそ同程度の大きさで正の値となっています。人の耳の全体的な感度として捉えることができます。第二成分は、両耳において高周波数は正の値をとり、低い周波数は負の値をとっています。したがって、第二成分は周波数の高低を識別しています。第三成分は中程度の周波数とそれ以外の周波数を識別しています。最後に、第四成分は左耳が負の値、右耳が正の値となっており、左右の耳を識別しています。

第一成分は 8 つの全ての変数において同程度の負荷量を持っていると述べました。これに関してデータが多変量正規分布に従っていることを想定して検定することができます。今回の場合、pca コマンドは標準誤差と関連する統計量を推定します。

紙面の都合上、第二成分までの結果のみを表示します。また、オプション vce(normal) を指定します。

```
. pca l* r*, comp(2) vce(normal)
```

Principal components/correlation — Number of obs = 100
Number of comp. = 2
Trace = 8
Rho = 0.6934
SEs assume multivariate normality — SE(Rho) = 0.0273

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
Eigenvalues						
Comp1	3.929005	.5556453	7.07	0.000	2.839961	5.01805
Comp2	1.618322	.2288653	7.07	0.000	1.169754	2.066889
Comp1						
lft500	.4010948	.0429963	9.33	0.000	.3168236	.485366
lft1000	.4209908	.0359372	11.71	0.000	.3505551	.4914264
lft2000	.3663748	.0463297	7.91	0.000	.2755702	.4571794
lft4000	.2808559	.0626577	4.48	0.000	.1580491	.4036628
rght500	.343251	.0528285	6.50	0.000	.2397091	.446793
rght1000	.4114209	.0374312	10.99	0.000	.3380571	.4847846
rght2000	.3115483	.0551475	5.65	0.000	.2034612	.4196354
rght4000	.2542212	.066068	3.85	0.000	.1247303	.3837121
Comp2						
lft500	-.3169638	.067871	-4.67	0.000	-.4499885	-.1839391
lft1000	-.225464	.0669887	-3.37	0.001	-.3567595	-.0941686
lft2000	.2385933	.1079073	2.21	0.027	.0270989	.4500877
lft4000	.4741545	.0967918	4.90	0.000	.284446	.6638629
rght500	-.3860197	.0803155	-4.81	0.000	-.5434352	-.2286042
rght1000	-.2317725	.0674639	-3.44	0.001	-.3639994	-.0995456
rght2000	.317059	.1215412	2.61	0.009	.0788427	.5552752
rght4000	.5135121	.0951842	5.39	0.000	.3269544	.7000697

LR test for independence: chi2(28) = 448.21 Prob > chi2 = 0.0000
LR test for sphericity: chi2(35) = 451.11 Prob > chi2 = 0.0000

Explained variance by components

Components	Eigenvalue	Proportion	SE_Prop	Cumulative	SE_Cum	Bias
Comp1	3.929005	0.4911	0.0394	0.4911	0.0394	.056663
Comp2	1.618322	0.2023	0.0271	0.6934	0.0273	.015812
Comp3	.9753248	0.1219	0.0178	0.8153	0.0175	-.014322
Comp4	.4667822	0.0583	0.0090	0.8737	0.0127	.007304
Comp5	.34009	0.0425	0.0066	0.9162	0.0092	.026307
Comp6	.3158912	0.0395	0.0062	0.9557	0.0055	-.057717
Comp7	.2001111	0.0250	0.0040	0.9807	0.0031	-.013961
Comp8	.1544736	0.0193	0.0031	1.0000	0.0000	-.020087

pca コマンドは、ここでは推定コマンドのような働きをしています。出力は異なる方程式から生成されています。一つ目の式は固有値を含んでいます。Comp1 と表示されている二つ目の式は第一主成分です。pca コマンドは、固有値の標準誤差を表示しています。応用研究において、固有値の値を検定することは稀ですが、結果を解釈する際に安定性を考慮することができます。表示された結果から 1 つ目の固有値は、3.929 であり、標準誤差は 0.56 であることがわかります。

また、pca コマンドは主成分の標準誤差も表示しています。また、共分散も推定できます。

```
. estat vce
```

分散共分散行列に含まれる大量の情報を表示すること自体は有用ではありません。しかし、推定することにより、主成分の特性を検定することができます。第一主成分の負荷量が同じ大きさであると言えるでしょうか。testparm コマンドに 2 つのオプションを指定して実行します。オプション eq(Comp1)は、Comp1 式（すなわち、第一主成分）に対する係数の検定を指定しており、オプション equal は、係数が 0 であるという検定の代わりに、係数が互いに等しいことを検定します。これは、主成分が 1 に正規化されていることを考慮した適切な仮説です。

```
. testparm lft* rght*, equal eq(Comp1)
```

```
( 1)  - [Comp1]lft500 + [Comp1]lft1000 = 0
( 2)  - [Comp1]lft500 + [Comp1]lft2000 = 0
( 3)  - [Comp1]lft500 + [Comp1]lft4000 = 0
( 4)  - [Comp1]lft500 + [Comp1]rght500 = 0
( 5)  - [Comp1]lft500 + [Comp1]rght1000 = 0
( 6)  - [Comp1]lft500 + [Comp1]rght2000 = 0
( 7)  - [Comp1]lft500 + [Comp1]rght4000 = 0

           chi2(  7) =    7.56
         Prob > chi2 =    0.3729
```

負荷量が等しいという帰無仮説を棄却することができません。したがって、第一成分への解釈はデータと矛盾しないようです。

また、pca コマンドは主成分によって説明された分散の比率の標準誤差も表示します。この情報は主に統計量に対する仮説検定というよりは式の強度を示しています。また、研

究を比較するのにも便利です。ある研究では 2 つの主成分で分散の 70%を説明できたとします。一方で追試的な研究によると 80%を説明できました。サンプルの変動によって重要な差が生じているでしょうか。

pca コマンドは、regress や xtlogit のように推定コマンドであるため、pca と入力することで結果を再表示できます。オプション vce(normal)とともに使用すれば、オプション novce を使用して標準的な pca の出力を表示できます。

```
. pca, novce
```

```
Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          2
                                                  Trace            =          8
    Rotation: (unrotated = principal)             Rho              =     0.6934
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.92901	2.31068	0.4911	0.4911
Comp2	1.61832	.642997	0.2023	0.6934
Comp3	.975325	.508543	0.1219	0.8153
Comp4	.466782	.126692	0.0583	0.8737
Comp5	.34009	.0241988	0.0425	0.9162
Comp6	.315891	.11578	0.0395	0.9557
Comp7	.200111	.0456375	0.0250	0.9807
Comp8	.154474	.	0.0193	1.0000

```
Principal components (eigenvectors)
```

Variable	Comp1	Comp2	Unexplained
lft500	0.4011	-0.3170	.2053
lft1000	0.4210	-0.2255	.2214
lft2000	0.3664	0.2386	.3805
lft4000	0.2809	0.4742	.3262
rght500	0.3433	-0.3860	.2959
rght1000	0.4114	-0.2318	.248
rght2000	0.3115	0.3171	.456
rght4000	0.2542	0.5135	.3193

例題 2：共分散行列を用いた分析

一般的に共分散行列の主成分と相関行列の主成分は異なります。pca コマンドは、初期設定で相関行列を用いて主成分分析を実行する設定になっています。共分散行列を用いた分析結果を表示するには、オプション covariance を指定します。

```
. pca l* r*, comp(4) covariance
```

```
Principal components/covariance                 Number of obs    =        100
                                                Number of comp.  =          4
                                                Trace            =     1154.5
    Rotation: (unrotated = principal)           Rho              =     0.9396
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	706.795	527.076	0.6122	0.6122
Comp2	179.719	68.3524	0.1557	0.7679
Comp3	111.366	24.5162	0.0965	0.8643
Comp4	86.8501	57.4842	0.0752	0.9396
Comp5	29.366	9.53428	0.0254	0.9650
Comp6	19.8317	6.67383	0.0172	0.9822
Comp7	13.1578	5.74352	0.0114	0.9936
Comp8	7.41432	.	0.0064	1.0000

```
Principal components (eigenvectors)
```

Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
lft500	0.0835	0.2936	-0.0105	0.3837	7.85
lft1000	0.1091	0.3982	0.0111	0.3162	11.71
lft2000	0.2223	0.5578	0.0558	-0.4474	11.13
lft4000	0.6782	-0.1163	-0.7116	-0.0728	.4024
rght500	0.0662	0.2779	-0.0226	0.4951	12.42
rght1000	0.0891	0.3119	0.0268	0.2758	11.14
rght2000	0.1707	0.3745	0.2721	-0.4496	14.71
rght4000	0.6560	-0.3403	0.6441	0.1550	.4087

予想した通り、分析結果は不明瞭になりました。分析された分散の合計は、1154.5 となり、これは 8 つの変数の分散の合計であり、共分散行列の対角成分の和です。導かれた主成分は分散の比率の大きいものほど影響を与えています。変数間で分散が大きく異なる共分

散行列の場合によく見られます。負荷量を比較することができないため、主成分を解釈することは困難です。

例題 3：行列を直接入力して主成分分析を行う

分析するために保持しているデータがオリジナルのデータではなく、相関行列や共分散行列であるかもしれません。そのような場合に備えて、PCA を実行するために pcamat コマンドが用意されています。データを左耳のみに絞って、使用方法を簡潔に紹介します。

```
. correlate lft*, cov
```

	lft500	lft1000	lft2000	lft4000
lft500	41.0707			
lft1000	37.7273	57.3232		
lft2000	28.1313	44.4444	119.697	
lft4000	32.101	40.8333	91.2121	384.775

手元にはオリジナルのデータが無く、変数に関する共分散のデータのみを持っているとしましょう。correlate コマンドは、r(C)に共分散を格納するため、この行列を使用します。pcamat コマンドに観測数を指定するためのオプション n(100)と変数名を付与するためのオプション names()を指定して実行します。

```
. matrix Cfull = r(C)
. pcamat Cfull, comp(2) n(100) names(lft500 lft1000 lft2000 lft4000)
```

Principal components/correlation　　Number of obs = 100
Number of comp. = 2
Trace = 4
Rotation: (unrotated = principal)　　Rho = 0.8169

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.37181	1.47588	0.5930	0.5930
Comp2	.895925	.366238	0.2240	0.8169
Comp3	.529687	.327106	0.1324	0.9494
Comp4	.202581	.	0.0506	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
lft500	0.5384	-0.4319	.1453
lft1000	0.5730	-0.3499	.1116
lft2000	0.4958	0.2955	.3387
lft4000	0.3687	0.7770	.1367

共分散行列を入力する必要がある場合、対角成分を含む左下三角成分または右上三角成分を入力してください。(相関行列の場合は、対角成分に 1 を入力してください。) 例えば、行ごとの共分散行列の下三角成分を入力し、対角成分を含めて 1 行の行列として入力します。

```
. matrix Clow = (41.0707, 37.7273, 57.3232, 28.1313, 44.4444, ///
>                119.697, 32.101, 40.8333, 91.2121, 384.775)
```

行列 Clow は、1 行 10 列の行列です。構造を見易くするために、以下のように数値を入力する方が好ましいでしょう。

```
. matrix Clow = (41.0707, ///
>                37.7273, 57.3232, ///
>                28.1313, 44.4444, 119.697, ///
>                32.101, 40.8333, 91.2121, 384.775)
```

行ベクトルまたは列ベクトルとして左下三角成分または右上三角成分を指定する場合、変数名を行列の行名または列名として指定することはできません。そのため、pcamat コマンドにおいてオプション names()を使用します。さらにベクトルが右上三角成分ではなく左下三角成分であることを示すためにオプション shape(lower)を指定する必要があります。

```
. pcamat Clow, comp(2) shape(lower) n(100) ///
>                       names(lft500 lft1000 lft2000 lft4000)
```