主成分分析の理論とその適用

id;:infinity_th4

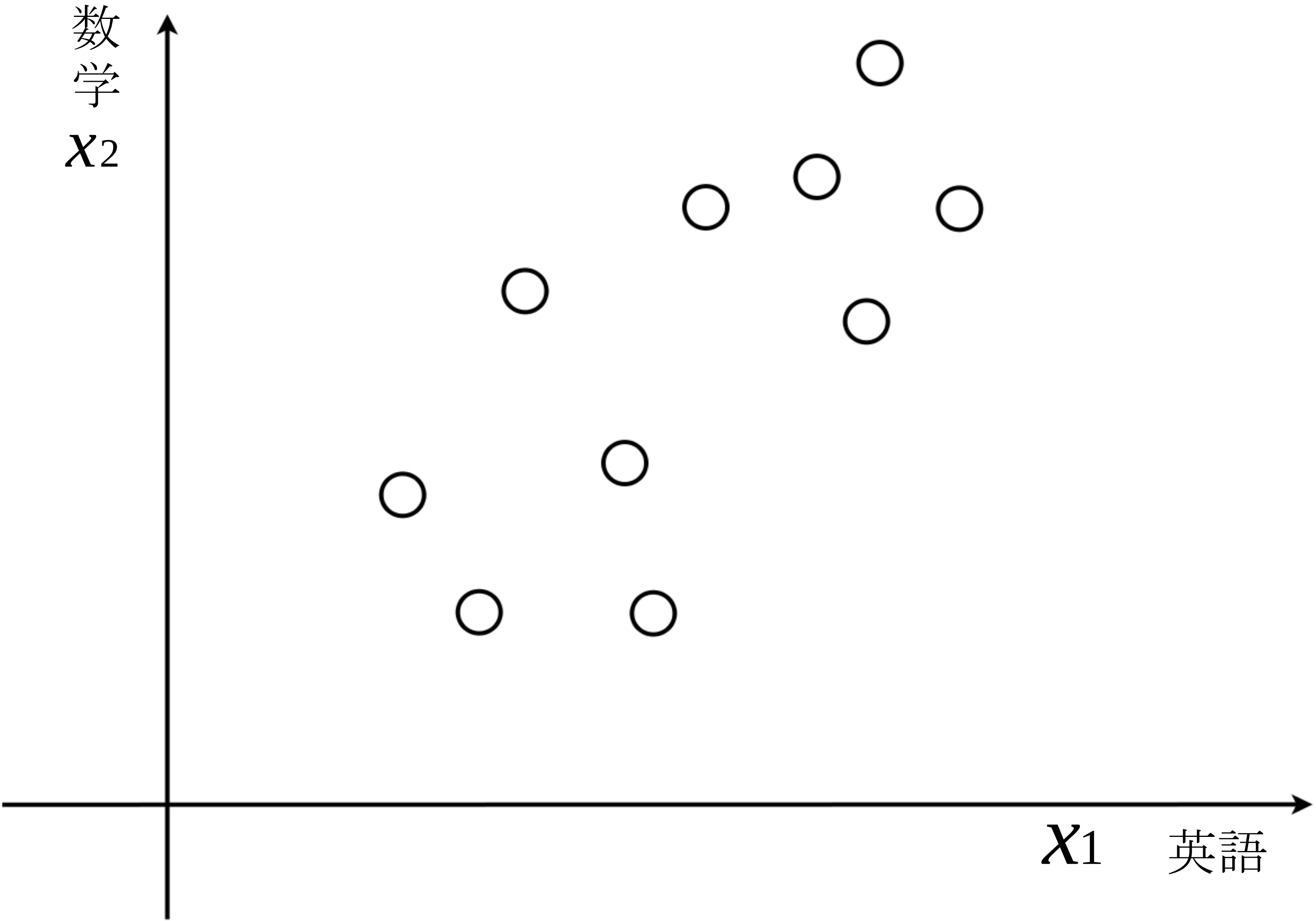
1. 主成分分析とは？

主成分分析とは,互いに相関のある変数について観測された多次元データのもつ情報を,できるだけ失うこと無く,元の変数の線形結合で表される新たな変数を構成し,より少数個の変数に要約して次元の縮小を行う手法.

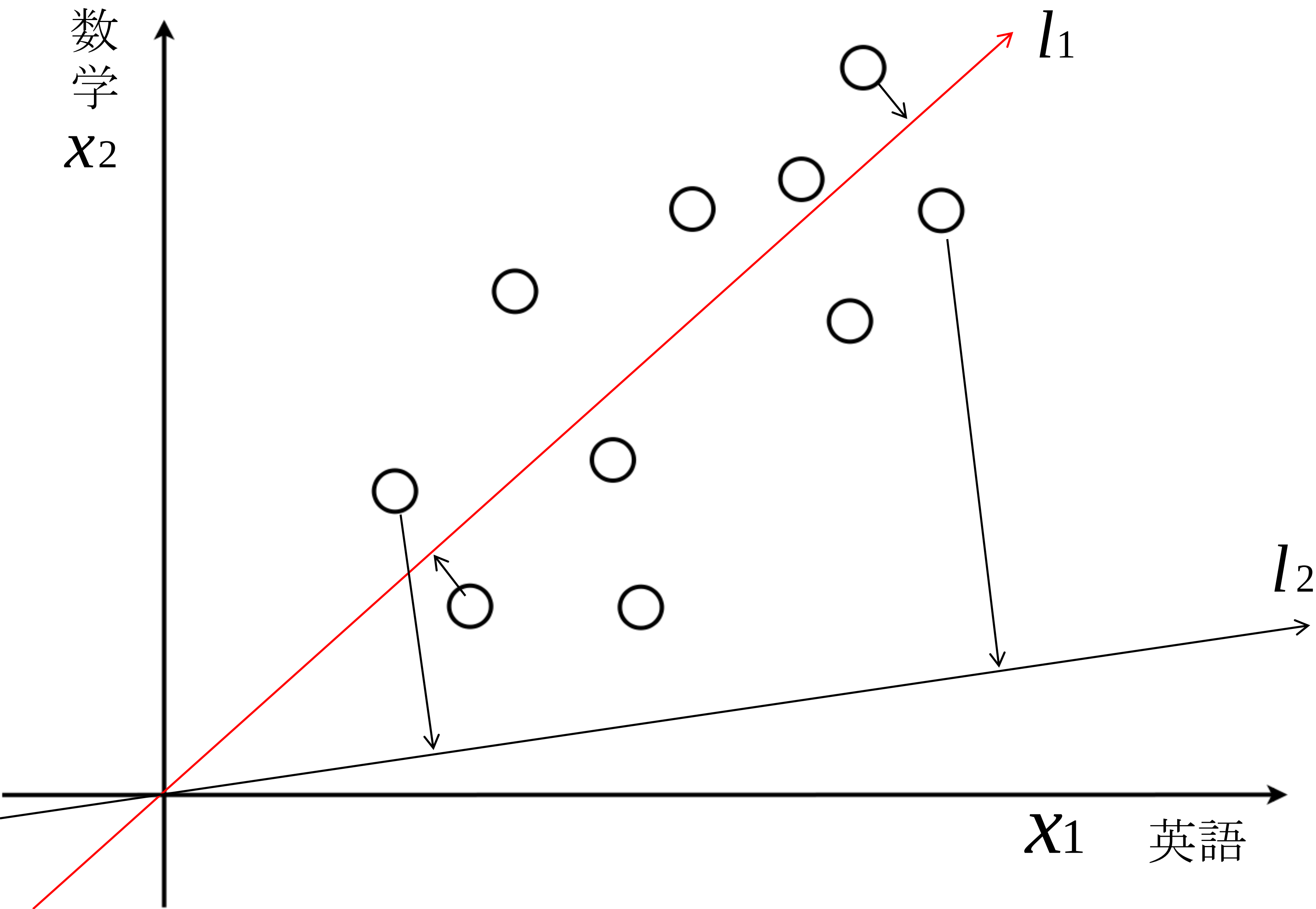
何が嬉しいか？

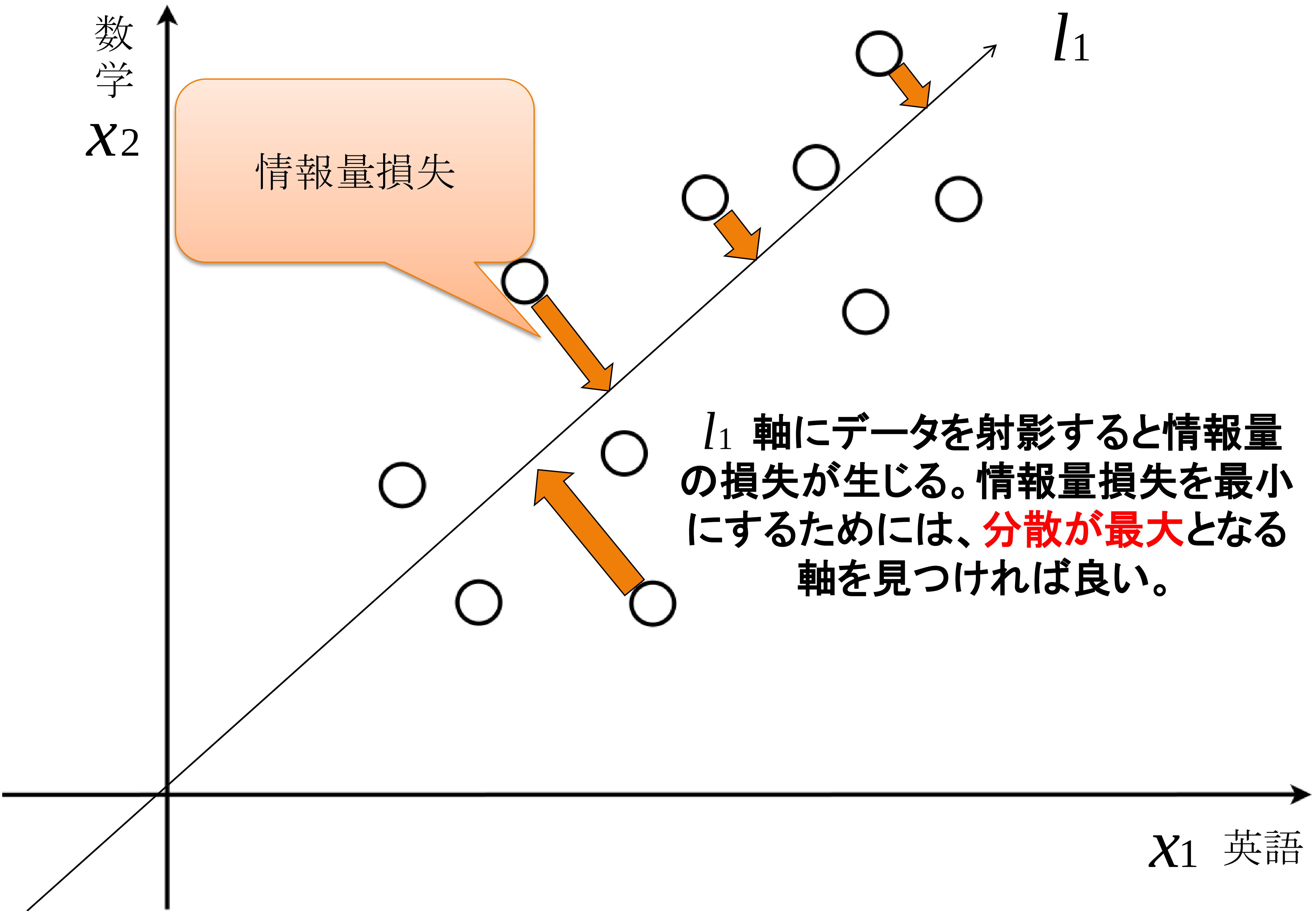
- 個体を特徴付ける複数の変数を融合することによって新たな意味づけを有した変数を生み出し,データの中から有益な情報を得ることができる.
- 高次元データを少数個の変数へ要約して,低次元空間に射影してデータ構造を視覚的に把握できる.

10人の英語と数学の得点のデータをプロット

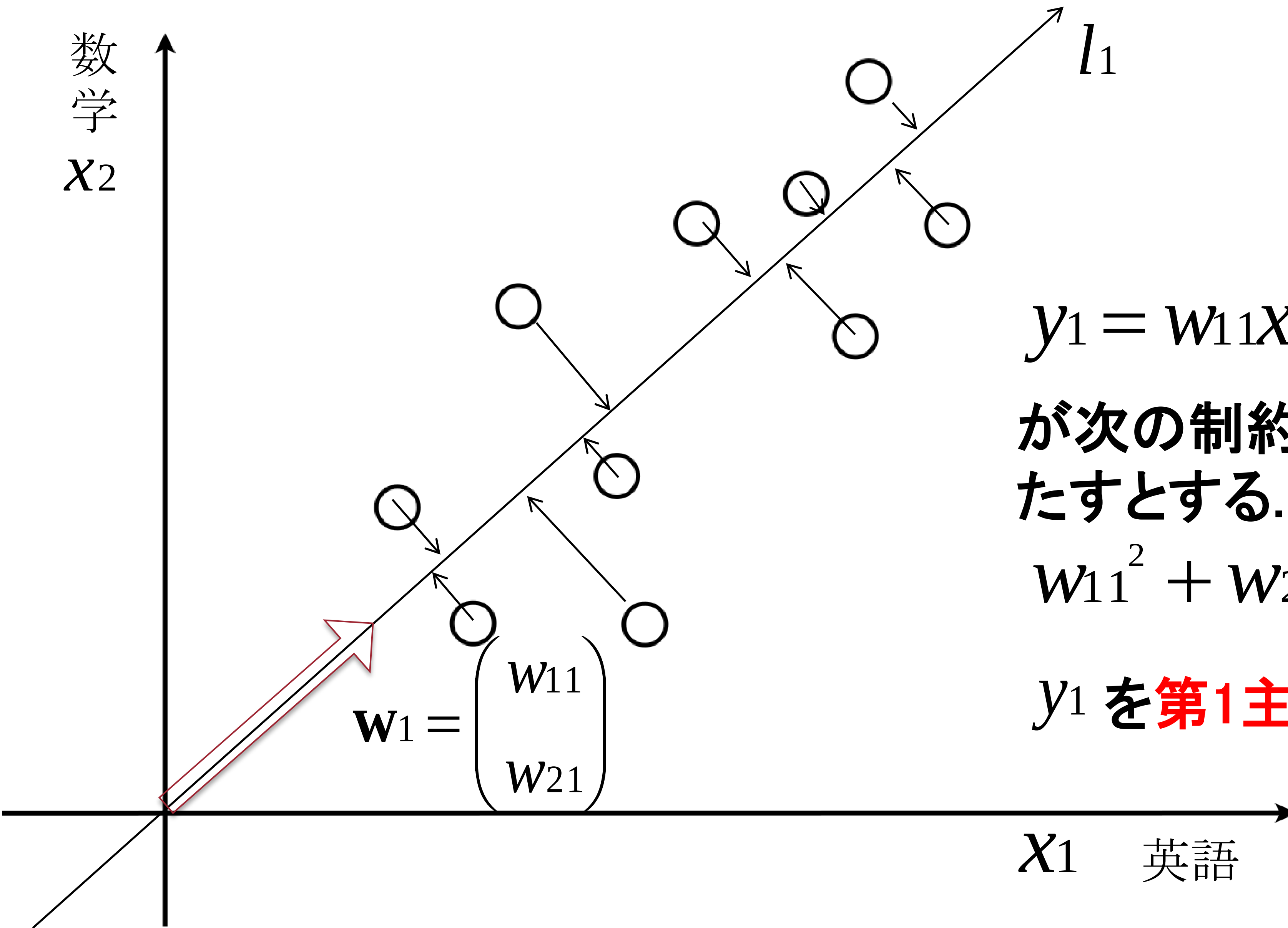


l_2 より, l_1 軸のほうが,データの傾向を表現できる.




数学
x_2
l_1
情報量損失
l_1 軸にデータを射影すると情報量の損失が生じる。情報量損失を最小にするためには、分散が最大となる軸を見つければ良い。
x_1 英語

軸に射影したデータの**分散が最大となる軸**を求める



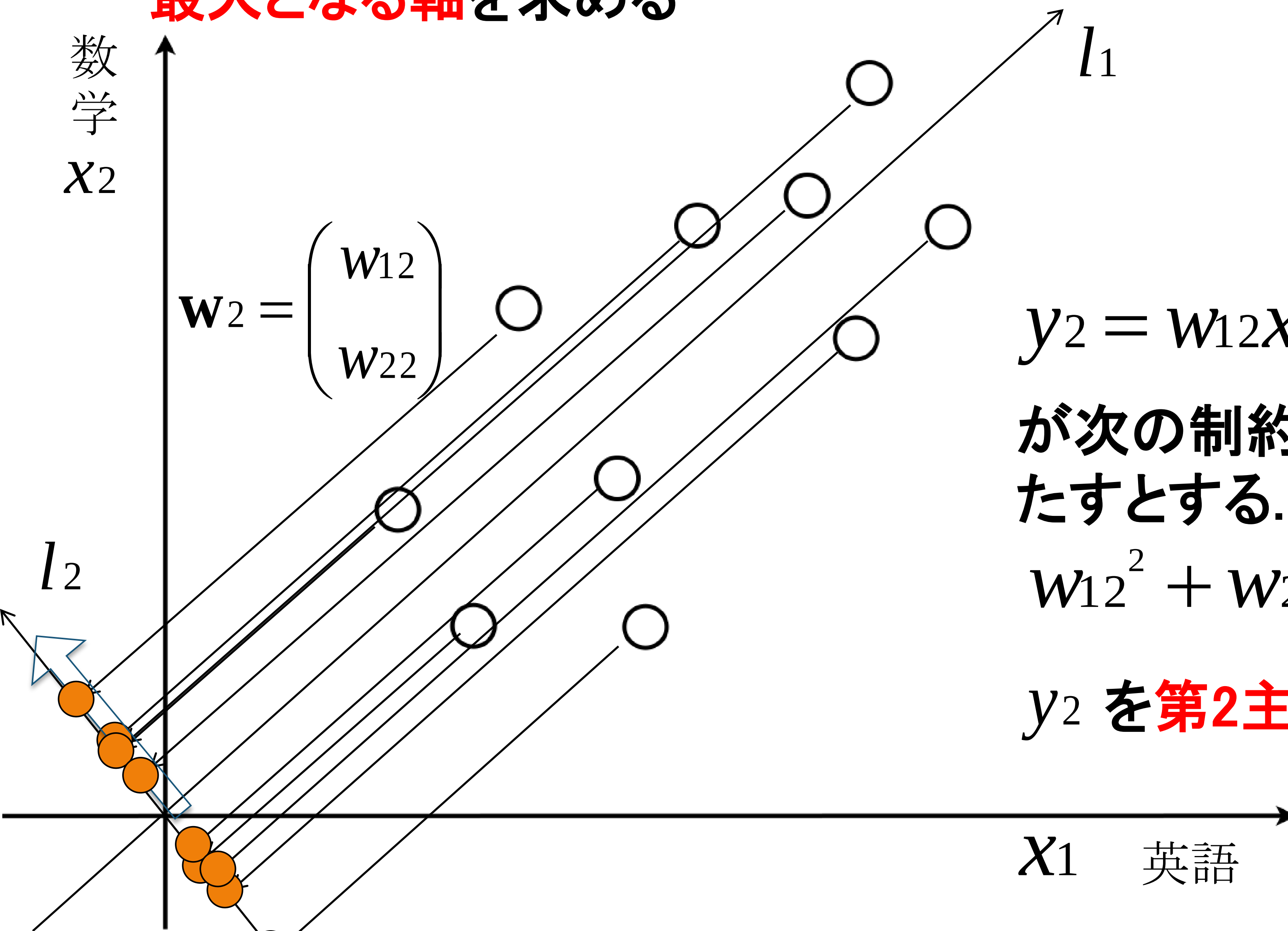
$$y_1 = w_{11}x_1 + w_{21}x_2$$

が次の制約条件を満たすとする.

$$w_{11}^2 + w_{21}^2 = 1$$

y_1 を**第1主成分**という.

次に l_1 軸に直交する軸で,その軸に射影した時に分散が最大となる軸を求める



$$\mathbf{w}_2 = \begin{pmatrix} w_{12} \\ w_{22} \end{pmatrix}$$

$$y_2 = w_{12}x_1 + w_{22}x_2$$

が次の制約条件を満たすとする.

$$w_{12}^2 + w_{22}^2 = 1$$

y_2 を第2主成分という.

データ行列

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{pmatrix}$$

行は,各個体の特徴を表すp個の変数

列は,各個体のp個の変数に関するデータ

1.1 主成分導出のプロセス

1. 標本分散共分散行列 $S = (s_{jk}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ を求める.

S の要素は $s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ji} - \bar{x_j})(x_{ki} - \bar{x_k})$ である.

(ここで, $\bar{\mathbf{x}}$ のp次元標本平均ベクトルは $\bar{\mathbf{x}} = (\bar{x_1} \quad \bar{x_2} \quad \cdots \quad \bar{x_p}), \bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$ とする)

2. $|S - \lambda I_p| = 0$ の解である固有値を, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ とする.

3. 固有値に対応する長さ1に正規化した固有ベクトルを,

$$\mathbf{w}_1 = \begin{pmatrix} w_{11} \\ w_{12} \\ \vdots \\ w_{1p} \end{pmatrix}, \mathbf{w}_2 = \begin{pmatrix} w_{21} \\ w_{22} \\ \vdots \\ w_{2p} \end{pmatrix}, \cdots, \mathbf{w}_p = \begin{pmatrix} w_{p1} \\ w_{p2} \\ \vdots \\ w_{pp} \end{pmatrix}$$ とする.

今, $\mathbf{w}_i{}'\mathbf{w}_i = 1, \mathbf{w}_i{}'\mathbf{w}_j = 0, (i \neq j)$ が成り立つ.

4. 元の変数の線形結合で表されるp個の**主成分**とその分散は,
次で与えられる.

第 i 主成分 : $y_i = \mathbf{w}_i{}'\mathbf{x}$　分散 : $\mathrm{V}(y_i) = \lambda_i, (i = 1, 2, \cdots, p)$

1.2 多次元データの基準化と標本相関係数

基準化について:

例えば,気温(華氏),住民数(人数),降水量(インチ)という異なる単位をもつデータに主成分分析を実行したいとする.

このとき,単位が異なっているので,観測データの基準化が必要.

多次元データの基準化と標本相関係数

n個のp次元データに基づく標本平均ベクトル $\bar{\mathbf{x}}$ と

標本分散共分散行列 S を求める.

$\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots, x_{pi})', (i = 1,2,\cdots,n)$ を次のように基準化する.

$$\mathbf{z}_i = (z_{1i}, z_{2i}, \cdots, z_{pi})', z_{ji} = \frac{x_{ji} - \bar{x}_j}{\sqrt{s_{jj}}}, (j = 1,2,\cdots,p)$$

$\mathbf{z}_i$ に基づく標本分散共分散行列を計算する.
$j,k=1,2,\cdots,p$ に対して,

$$Cov(z_j,z_k)=\frac{1}{n}\sum_{i=1}^{n}(z_{ji}-\overline{z_j})(z_{ki}-\overline{z_k})=\frac{Cov(x_j,x_k)}{\sqrt{s_{jj}}\sqrt{s_{kk}}}\equiv r_{jk}$$

となるので,標本分散共分散行列は,以下のようになる.

$$R=(r_{ik})=\begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

これを**標本相関行列**という.

データへの主成分分析の適用

米国の41都市における大気汚染に関する以下の7つの変数をもつデータを用いる.

SO2:	大気中の二酸化硫黄の含有量(マイクログラム/立方メートル)
N.Temp:	年間平均気温(華氏)(気温にマイナスを掛けたもの)
Manu:	20人以上を雇用する製造業者の数
Pop:	住民数(1970年の国政調査に基づく)(千人単位)
Wind:	年間平均風速(マイル/時間)
Precip:	年間平均降水量(インチ)
Days:	降水のあった日数の年間平均


Winnipeg
バンクーバー
Vancouver
シアトル
Seattle
スポケーン
Spokane
ワシントン
ポートランド
Portland
オレゴン
モンタナ
ノースダコタ
サウスダコタ
アイダホ
ワイオミング
ミネソタ
ミネアポリス
Minneapolis
ウィスコンシン
Milwaukee
ミシガン
アイオワ
シカゴ
Chicago
イリノイ
トレド
Detroit
トロント
Toronto
オタワ
Ottawa
モントリオール
Montréal
ケベック
Québec
ニューブランズウィック
メイン
バーモント
ニューハンプシャー
マサチューセッツ
ロードアイランド
コネチカット
ニュージャージー
デラウェア
メリーランド
コロンビア特別区
ペンシルバニア
Philadelphia
ウェストバージニア
バージニア
ネブラスカ
ソルトレイクシティ
Salt Lake City
ネバダ
ユタ
コロラド
アメリカ合衆国
カンザスシティ
ミズーリ
St. Louis
サクラメント
サンフランシスコ
San Francisco
カリフォルニア
ウィチタ
Wichita
タルサ
Tulsa
オクラホマ
アーカンソー
テネシー
ラスベガス
Las Vegas
アルバカーキ
ロサンゼルス
Los Angeles
ニューメキシコ
メサ
Mesa
ダラス
ミシシッピ
アトランタ
アラバマ
ノースカロライナ
サウスカロライナ
ティフアナ
Tijuana
エンセナダ
Ensenada
メヒカリ
Mexicali
シウダーフアレス
Juárez
テキサス
サンアントニオ
San Antonio
ヒューストン
Houston
ルイジアナ
モービル
Mobile
ジャクソンビル
Jacksonville
エルモシージョ
Hermosillo
チワワ
Chihuahua
カリフォルニア湾
クリアカン
Culiacan
サルティーヨ
Saltillo
レイノサ
Reynosa
マタモロス
Matamoros
モンテレイ
Monterrey
メキシコ湾
マイアミ
Miami
ラパス

標本相関行列

データの標本相関行列は、次のようになる。

	SO2	N.Temp	Manuf	Pop	Wind	Precip	Days
SO2	1.00000000	0.43360020	0.64476873	0.49377958	0.09469045	0.05429434	0.36956363
N.Temp	0.43360020	1.00000000	0.19004216	0.06267813	0.34973963	-0.38625342	0.43024212
Manuf	0.64476873	0.19004216	1.00000000	0.95526935	0.23794683	-0.03241688	0.13182930
Pop	0.49377958	0.06267813	0.95526935	1.00000000	0.21264375	-0.02611873	0.04208319
Wind	0.09469045	0.34973963	0.23794683	0.21264375	1.00000000	-0.01299438	0.16410559
Precip	0.05429434	-0.38625342	-0.03241688	-0.02611873	-0.01299438	1.00000000	0.49609671
Days	0.36956363	0.43024212	0.13182930	0.04208319	0.16410559	0.49609671	1.00000000

標本相関行列の固有値と固有ベクトル

固有値:

PC1	PC2	PC3	PC4	PC5	PC6	PC7
2.72811	1.51233	1.39497	0.89199	0.34677	0.10028	0.02551

固有ベクトル:

	[PC1]	[PC2]	[PC3]	[PC4]	[PC5]	[PC6]	[PC7]
SO2	-0.48969	0.08457	0.01435	0.40421	0.73039	-0.18334	-0.149529
N.Temp	-0.31537	0.08863	-0.67713	0.18522	-0.16246	0.61066	-0.02366
Manuf	-0.54116	-0.22588	0.26715	-0.02627	-0.16410	0.04273	0.74518
Pop	-0.48758	-0.28200	0.34483	-0.11340	-0.34910	0.08786	-0.64912
Wind	-0.24987	0.05547	-0.31126	-0.86190	0.26825	-0.15005	-0.01576
Precip	-0.000187	0.62587	0.49203	-0.18393	0.16059	0.55357	0.01031
Days	-0.26017	0.67796	-0.10957	0.10976	-0.43996	-0.50494	-0.00821

第1主成分は,「居住性」

	[PC1]
SO2	**−0.48969**
N.Temp	**−0.31537**
Manuf	**−0.54116**
Pop	**−0.48758**
Wind	−0.24987
Precip	−0.000187
Days	−0.26017

第1主成分は,「居住性」を表している.

値が大きければ,大気汚染がなく,工業地帯でない環境と解釈できる.

第2主成分は,「雨量」

	[PC2]
SO2	0.08457
N.Temp	0.08863
Manuf	−0.22588
Pop	−0.28200
Wind	0.05547
Precip	**0.62587**
Days	**0.67796**

第2主成分は,「雨量」を表している.(PrecipとDaysの値が**0.62**と**0.67**に注目する)

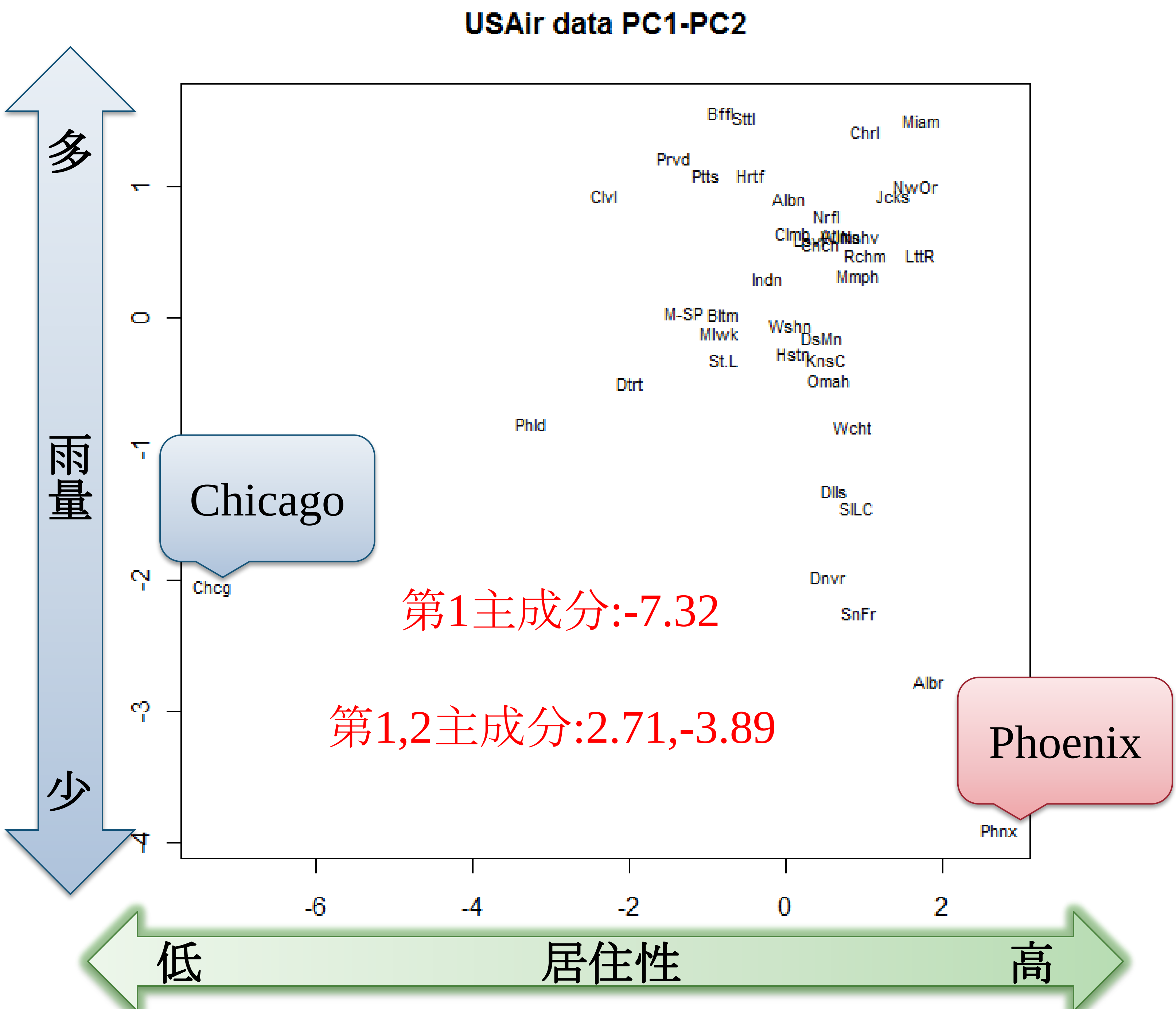
値が大きければ,雨の多い環境と解釈できる.

第3主成分は,「気候のタイプ」

	[PC3]
SO2	0.01435
N.Temp	−0.67713
Manuf	0.26715
Pop	0.34483
Wind	−0.31126
Precip	0.49203
Days	−0.10957

第3主成分は,「気候のタイプ」を表している.(N.TempとPrecipの値が−0.67と0.49であるから、対比している)

値が大きくなるほど,気温が高く,降水量の多い都市を表す.


USAir data PC1-PC2
多
雨量
少
低
居住性
高
Chicago
Phoenix
第1主成分:-7.32
第1,2主成分:2.71,-3.89
Chcg
Phnx
Albr
SnFr
Dnvr
SlLC
Dlls
Wcht
Phld
Dtrt
Omah
KnsC
Hstn
St.L
DsMn
Wshn
Mlwk
M-SP
Bltm
Indn
Mmph
Rchm
LttR
Clvl
Albn
Nrfl
Jcks
NwOr
Prvd
Ptts
Hrtf
Bffl
Sttl
Chrl
Miam
-6
-4
-2
0
2
1
0
-1
-2
-3
-4