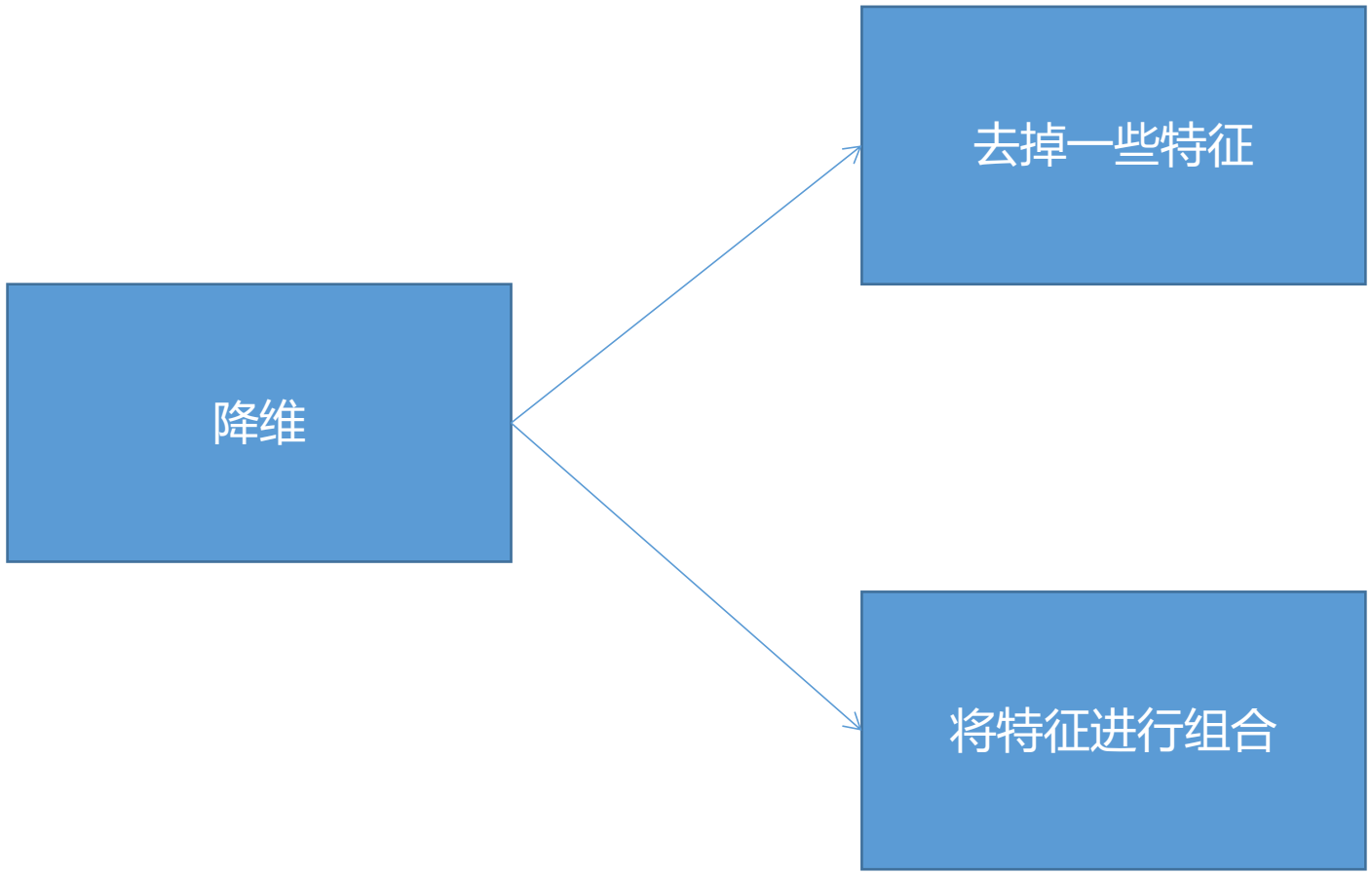
降维（Dimensionality Reduction）

降维的方法



正则化

主成分分析

主成分分析（PCA）

主成分分析与降维

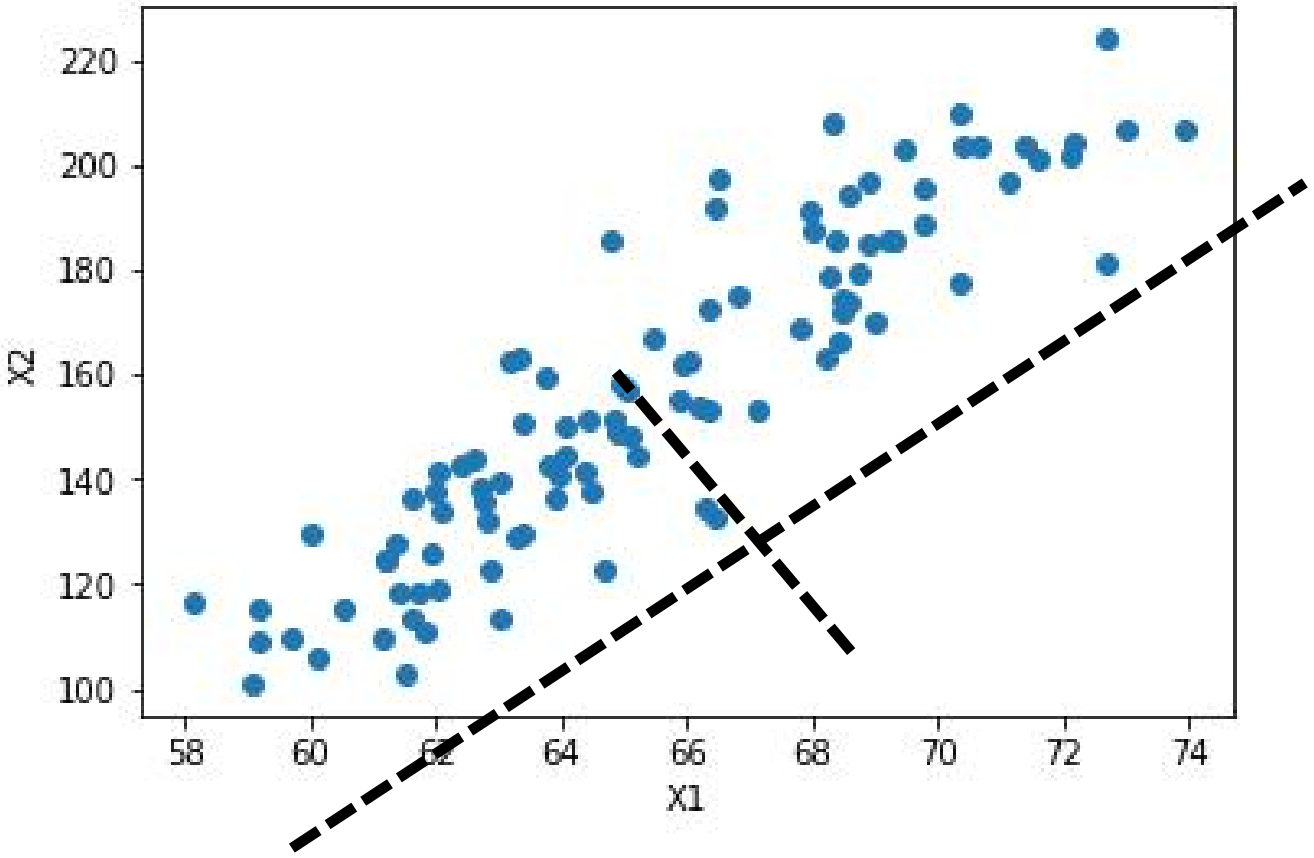
哪个维度反应最大的信息

主成分分析

寻找一个维度，使得样本在这个维度上的投影的方差最大（包含最大信息）

下一个维度，需要与之前的维度均正交，并且使得投影方差最大

寻找主成分



$$\max_{w_1, w_2} \mathrm{Var}\left(w_1 x_1 + w_2 x_2\right)$$

$$\text{s.t. } {w_1}^2 + {w_2}^2 = 1$$

w1 = 0.1083
w2 = 0.9941

k = w2 / w1 = 9.1798

PCA 的求解

PCA 的目标函数

寻找一个主成分

$$X = \begin{bmatrix} | & | & \cdots & | \\ X_1 & X_2 & \cdots & X_d \\ | & | & \cdots & | \end{bmatrix}$$

数据标准化

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$X \xrightarrow{w} X\_proj = Xw = w_1X_1 + w_2X_2 + \cdots + w_dX_d$$

PCA 的目标函数

$$\max_{w} \mathrm{Var}(\mathrm{X\_proj}) = \mathrm{Var}(Xw) = (Xw)^T(Xw) = w^T X^T Xw$$

$$\text{s.t.} \quad w^T w = 1$$

拉格朗日乘子法

$$X^T Xw = \lambda w$$

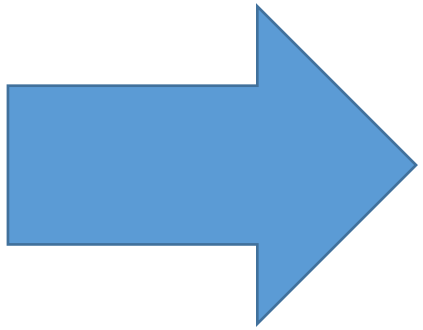
λ、w 分别是矩阵 $X^T X$ 的特征值和特征向量

PCA 的目标函数

寻找 p个主成分

$$\max_{w^{(1)}\dots w^{(p)}} \sum_{i=1}^{p} w^{(i)T} X^T X w^{(i)}$$

$$s.t.\ w^{(i)T} w^{(j)} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$



$$X^T X w^{(i)} = \lambda_i w^{(i)}$$

$w^{(i)}$ 是标准正交特征向量

奇异值分解 SVD

$$\underset{(mxd)}{X} = \underset{(mxm)}{U} \quad \underset{(mxd)}{\Sigma} \quad \underset{(dxd)}{V^T}$$

非零奇异值位于主对角线

$$\Sigma = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_d \end{bmatrix}$$

由特征向量组成的标准正交基

$$V = \begin{bmatrix} w^{(1)} & w^{(2)} & \cdots & w^{(d)} \end{bmatrix}$$

$$U = \begin{bmatrix} \frac{1}{\sigma_1} Xw^{(1)} & \frac{1}{\sigma_2} Xw^{(2)} & \cdots & \frac{1}{\sigma_d} Xw^{(d)} & \text{正交扩展至m维} \end{bmatrix}$$

Proportion of Variance Explained (PVE)

$$\begin{aligned}\mathrm{Var}\left(\mathrm{X\_proj}^{(\mathrm{i})}\right) &= w^{(i)T} X^T X w^{(i)} \\ &= \lambda_{\mathrm{i}} w^{(i)T} w^{(i)} \\ &= \lambda_{\mathrm{i}} \\ &= \sigma_i^{\,2}\end{aligned}$$

如果我们选择了 p 个主成分，则有多少方差被这 p 个主成分解释？

$$\mathrm{PVE} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{p}} \lambda_i}{\mathrm{sum}\left(\lambda_i\right)}$$

使用 sklearn 进行主成分分析

使用 sklearn 进行主成分分析

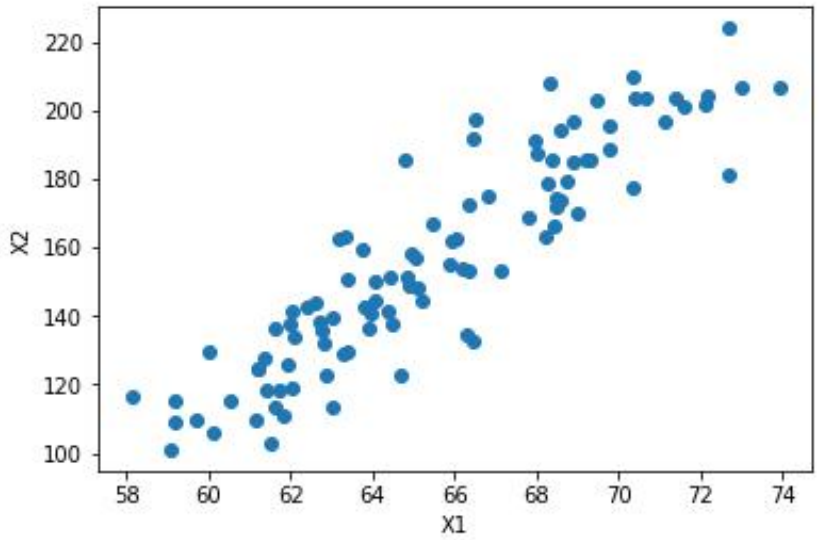
sklearn.decomposition.PCA

关键参数：
-- n_components：主成分个数

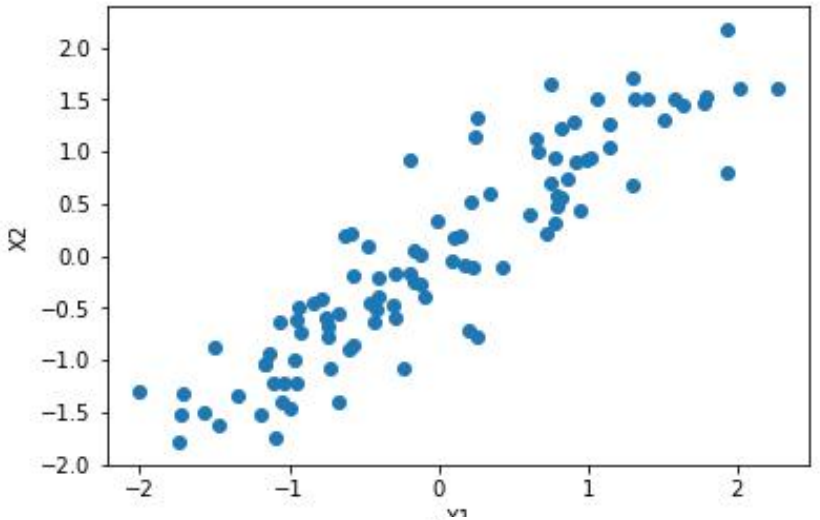
关键属性：
-- components_： w 矩阵

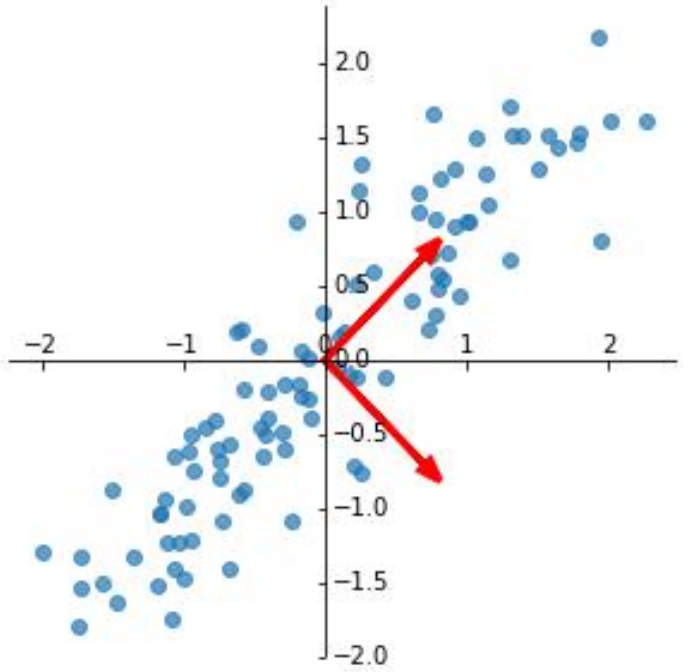
-- explained_variance_ratio_： PVE:

使用 sklearn 进行主成分分析



标准化





PVE： [**0.95451949** 0.04548051]

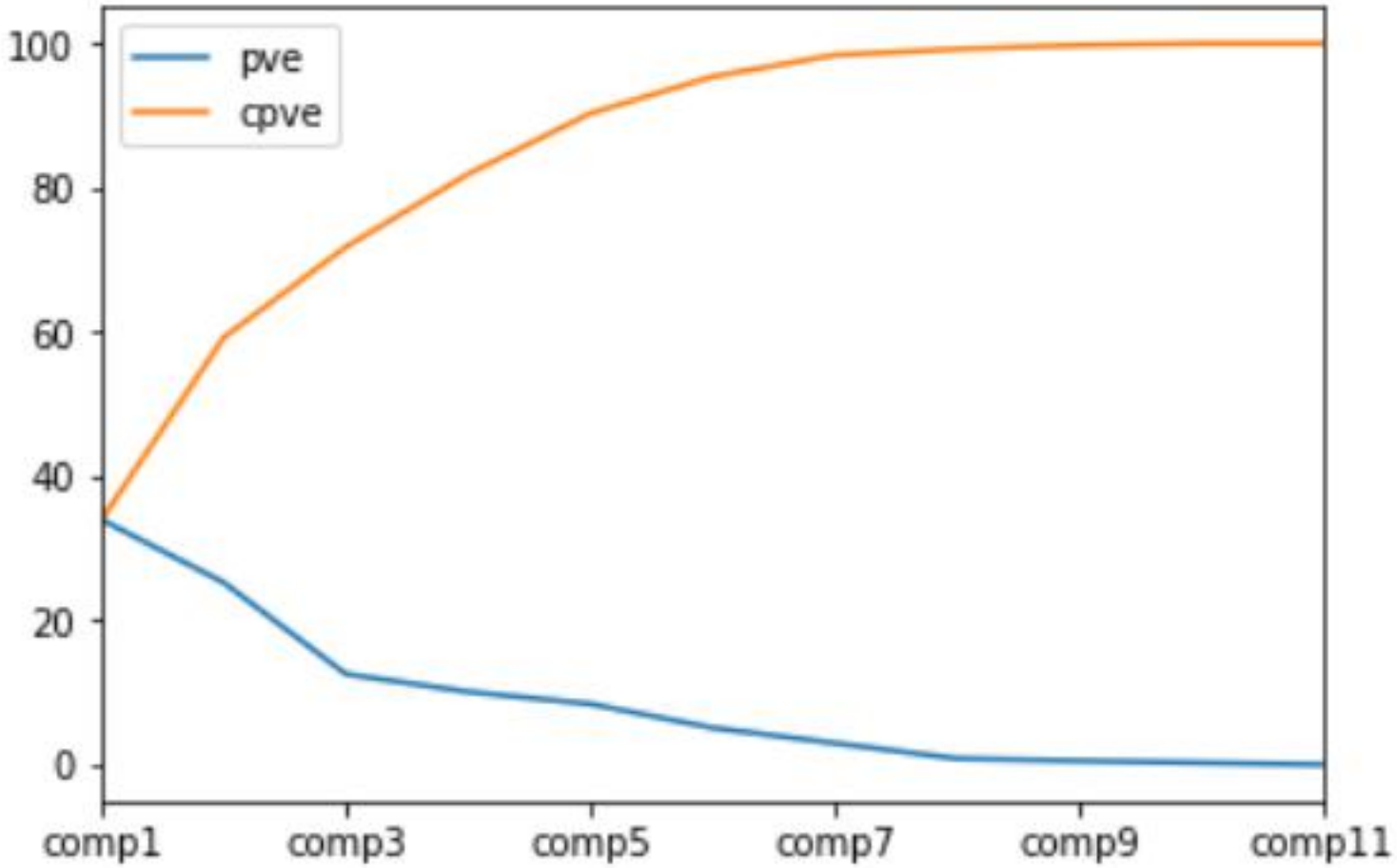
第一个主成分解释了 95% 的样本方差

使用 PCA 进行降维分析

饮料行业 2018年财务指标

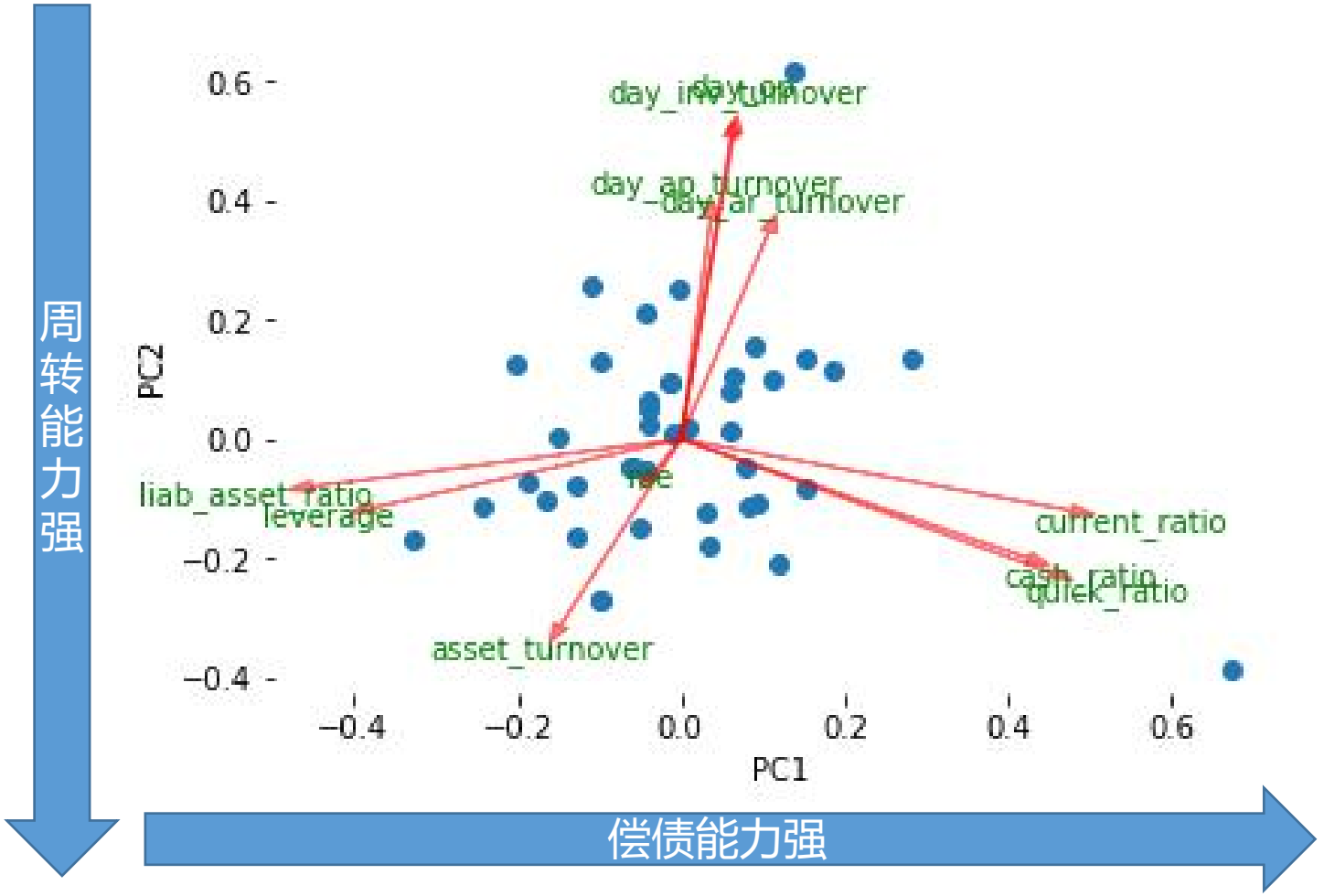
code	name	roe	leverage	asset_turnover	current_ratio	quick_ratio	cash_ratio	day_inv_turnover	day_ar_turnover	day_ap_turnover	day_op	liab_asset_ratio
000019.SZ	深粮控股	12.05	1.47	2.85	2.89	1	0.43	55.1	9.22	9.22	64.31	32.95
000568.SZ	泸州老窖	21.69	1.32	0.62	2.86	2.26	2.17	370.68	0.25	124.71	370.93	24.25
000596.SZ	古井贡酒	24.11	1.61	0.77	2.1	1.54	0.71	416.57	1.08	85.76	417.65	35.81
000729.SZ	燕京啤酒	1.39	1.39	0.63	1.63	0.63	0.5	197.67	6.69	55.74	204.36	22.24
000752.SZ	*ST西发	-70.85	2.25	0.25	1	0.95	0.72	34.12	3.1	45.63	37.21	45.33
000799.SZ	酒鬼酒	10.64	1.29	0.44	3.08	1.87	1.83	1146.5	1.87	101.31	1148.36	24.06
000848.SZ	承德露露	21.07	1.44	0.75	2.67	2.34	2.3	93.27	0.04	52.23	93.31	31.18
000858.SZ	五粮液	22.91	1.34	0.51	3.77	3.2	3.14	383.67	1.07	107.97	384.74	24.36
000860.SZ	顺鑫农业	10.06	2.59	0.63	1.75	0.81	0.73	398.01	2.48	13.34	400.49	61.07
000869.SZ	张裕A	11.26	1.39	0.4	1.83	0.83	0.64	492.07	17.71	130.63	509.78	24.6
000929.SZ	兰州黄河	-9.66	2.03	0.36	3.63	2.86	2.67	243.29	15.68	49.72	258.97	20.91
002304.SZ	洋河股份	25.7	1.47	0.52	2.3	1.41	0.25	758.05	0.1	67.22	758.16	32.16
002461.SZ	珠江啤酒	4.53	1.45	0.34	2.88	2.7	1.17	80.48	4.16	69.8	84.64	32.21
002568.SZ	百润股份	6.68	1.28	0.52	2.09	1.92	1.43	71.65	28.4	168.87	100.06	20.69
002646.SZ	青青稞酒	4.58	1.19	0.48	2.85	1.39	0.91	583.85	8.13	166.76	591.98	19.14
200019.SZ	深粮B	12.05	1.47	2.85	2.89	1	0.43	55.1	9.22	9.22	64.31	32.95
200596.SZ	古井贡B	24.11	1.61	0.77	2.1	1.54	0.71	416.57	1.08	85.76	417.65	35.81
200869.SZ	张裕B	11.26	1.39	0.4	1.83	0.83	0.64	492.07	17.71	130.63	509.78	24.6
600059.SH	古越龙山	4.28	1.17	0.36	4.65	1.88	1.02	653.12	27.62	136.08	680.74	14.63
600084.SH	*ST中葡	-6.94	1.27	0.12	3.52	1.24	0.97	2346.81	56.06	106.17	2402.86	22.45
600132.SH	重庆啤酒	35	2.95	1.02	0.85	0.66	0.56	83.68	4.87	80.2	88.55	64.77
600189.SH	吉林森工	1.54	2.33	0.24	1.04	0.57	0.33	435.68	20.18	209.38	455.86	53.76
600197.SH	伊力特	18.39	1.37	0.66	3.3	2.26	2.11	261.69	1.48	64.93	263.16	23.43
600199.SH	金种子酒	4.44	1.38	0.42	2.89	1.86	1.4	353.15	31.83	158.21	384.98	26.52
600238.SH	ST椰岛	5.21	2.18	0.42	1.57	0.93	0.43	329.49	29.28	61.58	358.77	45.42

PVE



Index	pve	cpve
comp1	33.9666	33.9666
comp2	25.2488	59.2155
comp3	12.5579	71.7734
comp4	10.093	81.8664
comp5	8.41438	90.2808
comp6	5.10441	95.3852
comp7	2.98475	98.3699
comp8	0.835789	99.2057
comp9	0.533529	99.7392
comp10	0.260773	100
comp11	4.38953e-10	100

biplot

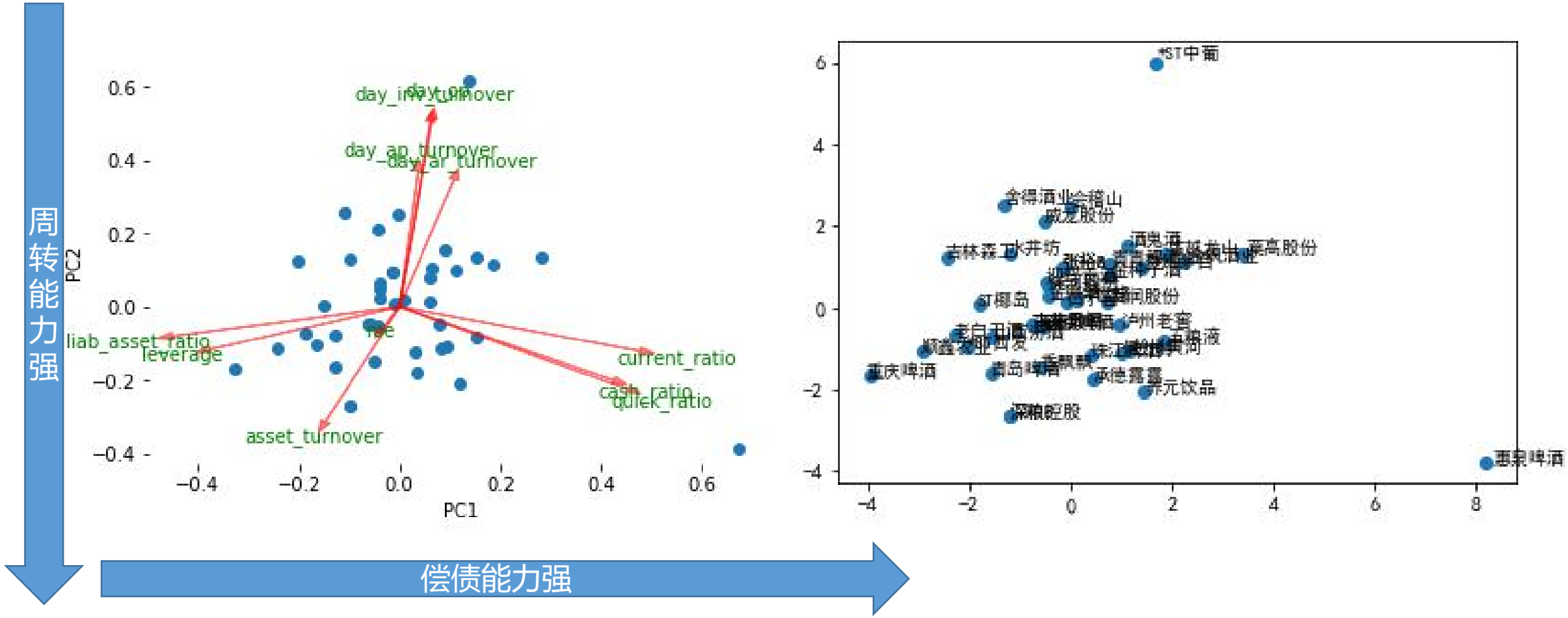


Index	comp1	comp2
roe	-0.0340927	-0.0577597
leverage	-0.376129	-0.114494
asset_turnov…	-0.147216	-0.312118
current_ratio	0.476676	-0.119022
quick_ratio	0.452253	-0.223054
cash_ratio	0.424344	-0.19884
day_inv_turn…	0.0615657	0.50537
day_ar_turno…	0.107693	0.347962
day_ap_turno…	0.0362467	0.370204
day_op	0.064935	0.512709
liab_asset_r…	-0.450503	-0.0792625

第一个主成分主要关注偿债能力，comp1 值越高偿债能力越强

第二个主成分主要关注周转能力，comp2 值越高周转能力越低

biplot


周转能力强
偿债能力强
day_inv_turnover
day_ap_turnover
day_ar_turnover
current_ratio
cash_ratio
quick_ratio
asset_turnover
liab_asset_ratio
leverage
roe
PC1
PC2
*ST中葡
重庆啤酒