MIRI
Data Mining & Business Intelligence

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Session 3:
Extensions of CCA.
The Inter Batteries Analysis

Sessions on Multivariate Modeling.

Course on Kernel Based Learning and Multivariate Modeling

Tomàs Aluja-Banet

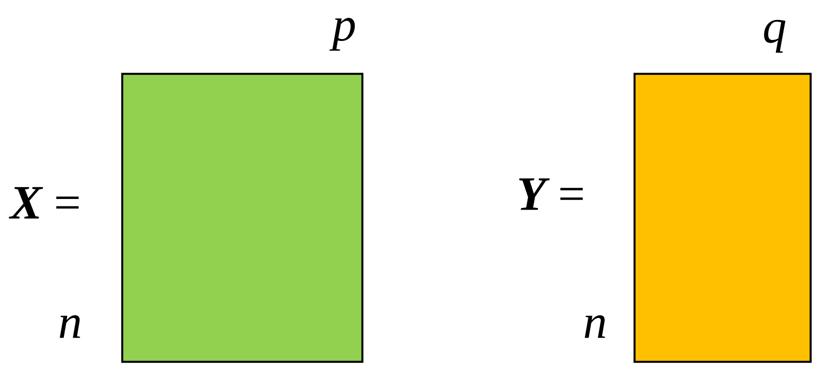
tomas.aluja@upc.edu

INFORMATION MODELING & PROCESSING
information matters

The problem

We continue having two vectors, measured on n individuals (Tucker, 1958)

$$(x_1, x_2, \cdots, x_p) \qquad (y_1, y_2, \cdots, y_q)$$



The goal is the same, to measure the relationship between both multivariate vectors by components derived from the original variables

$$t_h = Xa_h \qquad u_h = Yb_h$$

Canonical components perform well to explain the other group of variables, but not their own group → Hence instead of maximizing the cor(t_h, u_h) ….

The Inter-batteries analysis

What if we relieve the assumption that components t_h and u_h have to be of unit length.

$$Max \left\langle t_h', u_h \right\rangle_N = t_h' N u_h = a'_h X' NYb_h = \gamma_h = \mathrm{cov}(t_h, u_h)$$

$$t_h = Xa_h$$
$$u_h = Yb_h$$

but then we will need to constraint a_h and b_h vectors.

$$a_h' a_h = 1; \quad b_h' b_h = 1$$

Restriction on R^p and R^q

$$\ell = a_h' X' NYb_h - \lambda(a_h' a_h - 1) - \mu(b_h' b_h - 1)$$

$$X'NYb_h - 2\lambda a_h = 0$$
$$Y'NXa_h - 2\mu b_h = 0$$

$$a_h' X' NYb_h = 2\lambda a_h' a_h$$
$$b_h' Y' NXa_h = 2\mu b_h' b_h$$

$$2\lambda = 2\mu = \gamma_h$$

$$\boxed{\begin{aligned} X'NYb_h &= \gamma_h a_h \\ Y'NXa_h &= \gamma_h b_h \end{aligned}}$$ *

Transition relations

Here it is enough to consider X and Y centered, but usually they will be scaled, then matrices of covariances become correlations. If X, Y centered, results depend on the scale of variables.

The solution

$$V_{YX} a_h = \gamma_h b_h$$
$$V_{XY} b_h = \gamma_h a_h$$

rang(V_{XY}) = rang(V_{YX}) = s

Maximal number of solutions s

Analysis in R^p and R^q

$$\boxed{\begin{aligned} V_{XY} V_{YX} a_h &= \gamma_h^2 a_h \\ V_{YX} V_{XY} b_h &= \gamma_h^2 b_h \end{aligned}}$$

$$A'A = I$$
$$B'B = I$$

$$\boxed{\begin{aligned} T &= XA \\ U &= YB \end{aligned}}$$

Analysis in R^n

$$XX'NYY'Nt_h = \gamma_h^2 t_h \qquad t_h = Xa_h$$
$$YY'NXX'Nu_h = \gamma_h^2 u_h \qquad u_h = Yb_h$$

$$t_h' N t_h = a_h' V_X a_h$$
$$u_h' N u_h = b_h' V_X b_h$$

Components t_h and u_h neither normalized, nor orthogonal

IBA in practice

usually $q < p$

$$V_{YX}V_{XY}b_h = \gamma_h^2 b_h$$

$$a_h = \frac{1}{\gamma_h} V_{XY} b_h$$

$$t_h = Xa_h$$

$$u_h = Yb_h$$

number of solutions = $rank(V_{XY})$

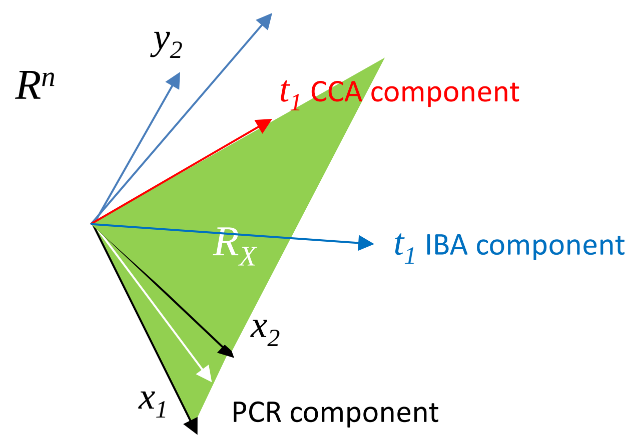
$$t_h^{test} = X^{test} a_h$$

Properties of IBA components

IBA is a compromise between the CCA and the PCA of both blocks

$$\underbrace{\mathrm{cov}(t_h, u_h)}_{\text{IBA criterion}} = \underbrace{cor(t_h, u_h)}_{\text{CCA criterion}} \underbrace{\sqrt{\mathrm{var}(Xa_h)}\sqrt{\mathrm{var}(Yb_h)}}_{\text{PCA criterion}}$$

The IBA components will compromise the explanation of each own group and the prediction between groups



Properties of IBA components

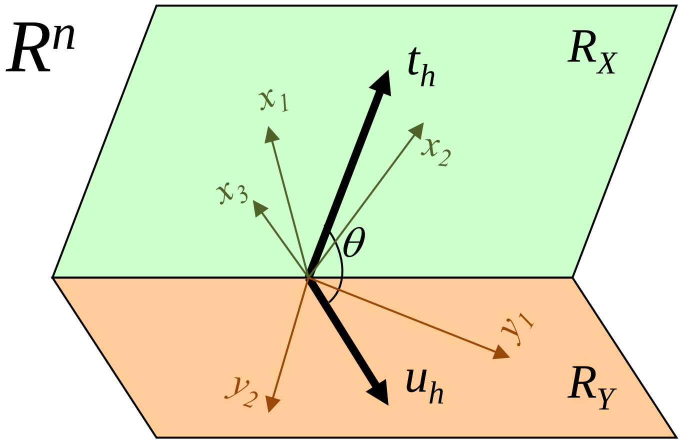
$$\|t_h\|_N^2 = \langle t_h, t_h \rangle_N = a_h' X' N X a_h$$

$$\|u_h\|_N^2 = u_h' N u_h = b_h' Y' N Y b_h$$

$$\text{cov}(t_h, t_l) = a_h' X' N X a_l \neq 0, \quad \text{cov}(u_h, u_l) = b_h' Y' N Y b_l \neq 0$$

IBA components are not orthogonal

Properties of IBA components



* $$X'NYb_h = X'Nu_h = \gamma_h a_h$$
$$Y'NXa_h = Y'Nt_h = \gamma_h b_h$$

$$a_h \text{ colinear to } \frac{1}{\gamma_h}\left(X'NYb_h = X'Nu_h = \left(\text{cov}(x_j, u_h)\right)\right)$$

$$b_h \text{ colinear to } \frac{1}{\gamma_h}\left(Y'NXa_h = Y'Nt_h = \left(\text{cov}(y_k, t_h)\right)\right)$$

a_h and b_h are proportional to the covariance of each variable with the IBA component of the other group

Properties of IBA components

$$\gamma_h = \operatorname{cov}(t_h, u_h) = a_h' X'NYb_h = \langle a_h, X'NYb_h \rangle = \cos(a_h, X'Nu_h)\left\|X'Nu_h\right\| = \left\|X'Nu_h\right\|$$

$$Max\,\gamma_h \Rightarrow \gamma_h^2 = \left\|X'Nu_h\right\|^2 = \sum_{j=1}^{p} \operatorname{cov}^2(x_j, u_h) \quad = \sum_{k=1}^{q} \operatorname{cov}^2(y_k, t_h)$$

The IBA components maximize the sum of covariances with the variables of the other group

$$\operatorname{cov}(t_h, u_h) = \gamma_h$$

$$\operatorname{cov}(t_h, u_l) = 0 \quad a_h' X'NYb_l = a_h' V_{XY} b_l = \gamma_l a_h' a_l = 0$$

Covariance with homonymous components is maximized, otherwise is zero

Singular Value Decomposition

$$\left.\begin{aligned} V_{YX} a_h &= \gamma_h b_h & V_{YX} A &= B\Lambda \\ V_{XY} b_h &= \gamma_h a_h & V_{XY} B &= A\Lambda \end{aligned}\right\} \quad V_{XY} = A\Lambda B' \qquad \Lambda = \begin{bmatrix} \ddots & & \\ & \gamma_h & \\ & & \ddots \end{bmatrix}$$

$$V_{XY} = \sum_{h=1}^{s} \gamma_h a_h b'_h$$

$$\mathrm{cov}(x_j, y_k) = \sum_{h=1}^{s} \frac{1}{\gamma_h} \begin{pmatrix} \vdots \\ \mathrm{cov}(x_j, u_h) \\ \vdots \end{pmatrix} \begin{pmatrix} \cdots & \mathrm{cov}(y_k, t_h) & \cdots \end{pmatrix}$$

$$cor(x_j, y_k) = \sum_{h=1}^{s} \frac{cor(x_j, u_h) \times cor(y_k, t_h)}{cor(t_h, u_h)}$$

Variable displays

Variables

basis $t_1, t_2, \ldots$
(non orthogonal)

$$\psi_h^X = \begin{pmatrix} \vdots \\ cor(x_j, t_h) \\ \vdots \end{pmatrix} \qquad \psi_h^Y = \begin{pmatrix} \vdots \\ cor(y_k, t_h) \\ \vdots \end{pmatrix}$$

basis $u_1, u_2, \ldots$
(non orthogonal)

$$\varphi_h^X = \begin{pmatrix} \vdots \\ cor(x_j, u_h) \\ \vdots \end{pmatrix} \qquad \varphi_h^Y = \begin{pmatrix} \vdots \\ cor(y_k, u_h) \\ \vdots \end{pmatrix}$$

They are not biplots, and difficult to interpret since dimensions are not orthogonal

Individual displays

Individuals

2 displays $Xa_h = t_h$ $Yb_h = u_h$

Display of individuals in the (t_h, u_h) basis.
To reveal the strength of the liaison

$$(t_1, u_1), (t_2, u_2), \ldots$$

Relation with the orthogonal procrustean rotation

Orthogonal procrustean rotation

$Min \left\| X - YR \right\|^2$ R rotation matrix

$X'Y = nV_{XY} = A\Lambda B', \quad \Rightarrow R = BA'$

We are assuming X and Y of the same dimensions, otherwise we complete with columns of zeroes.

$$\min \left(\left\| X - YR \right\|^2 = tr\left((X - YR)'(X - YR) \right) = tr(X'X + Y'Y - 2X'YR) \right)$$

$$\Rightarrow \max \left(tr(X'YR) = tr(A\Lambda B'R) = tr(\Lambda B'RA) = tr(\Lambda Q) = \sum \gamma_h q_{hh} \right)$$

$$\left| q_{hh} \right| \leq 1 \quad \Rightarrow \quad Q = I \quad \Rightarrow \quad R = BA'$$

$$\Rightarrow \quad Max\, tr(X'YR) = \sum \gamma_h$$

$$\left\| X - YBA' \right\|^2 = tr\left((X - YBA')(X' - AB'Y') \right) =$$

$$tr\left((X - YBA')AA'(X' - AB'Y') \right) = tr(XA - YB)(A'X' - B'Y') = \left\| T - U \right\|^2 \qquad T = XA \quad U = YB$$

IBA components reconstitutes the closest configuration between both blocks

Optimal approximation of V_{XY}

The IBA biplot is not as easy to interpret as the CCA, but it gets optimal approximation of the true covariances

Reconstitution of r rang

$$\sum_{h=1}^{s}\gamma_h^2 = tr\left(V_{XY}V_{YX}\right) = \left\|V_{XY}\right\|^2 = \left\|\hat{V}_{XY}^r\right\|^2 + \left\|V_{XY} - \hat{V}_{XY}^r\right\|^2$$

$$\hat{V}_{XY}^r = \sum_{h=1}^{r}\gamma_h a_h b_h'$$

$$\left\|\hat{V}_{XY}^r\right\|^2 = \sum_{h=1}^{r}\gamma_h^2$$

$$\left\|V_{XY} - \hat{V}_{XY}^r\right\|^2 = \sum_{h=r+1}^{s}\gamma_h^2$$

Significant components (with the usual assumptions and standardized data) (Tucker, 1958)

$$\gamma_1 = \gamma_2 = \cdots = \gamma_s = 0 \qquad r = 0 \rightarrow n\sum_{h=1}^{s}\gamma_h^2 \sim \chi^2_{p*q}$$

$$\gamma_1 > \gamma_r > \gamma_{r+1} = \cdots = \gamma_s = 0 \qquad r > 0 \rightarrow \frac{n(p-r)(q-r)\sum_{h=r+1}^{s}\gamma_h^2}{(p-\sum_{h=1}^{r}\mathrm{var}(t_h))(q-\sum_{h=1}^{r}\mathrm{var}(u_h))} \sim \chi^2_{(p-r)(q-r)}$$

IBA Regression

IBA Regression in practice:
We use the t_1, t_2, … significant interbattery components as explanatory latent components of the y_j variables.

$$Y = T_{(r)}^{IBA}\tilde{B}_{(r)} + \varepsilon_Y \qquad \tilde{B}_{(r)} = (T_{(r)}'^{IBA} N T_{(r)}^{IBA})^{-1} T_{(r)}'^{IBA} N Y$$

Expressing the model as function of the original variables

$$Y = XA_{(r)}^{IBA}\tilde{B}_{(r)} + \varepsilon_Y = XB + \varepsilon_Y \qquad B = A_{(r)}^{IBA}(T_{(r)}'^{IBA} N T_{(r)}^{IBA})^{-1} T_{(r)}'^{IBA} N Y$$

Limitations of IBA

- Number max. of components = $rank(V_{XY}) \leq min(p,q,n-1)$
- Solution depends on scale of variables, if data just centered
- Non orthogonal components
- Joint representations are not biplots

Advantages of IBA

- Components are good for prediction and altogether good factors for the own group.
- Deals with collinearity

Running IBA

```
> library(plsdepot)
> iba <- interbat(X, Y, scaled = TRUE)

> iba$values
[1] 1.272426 0.005657 0.001106
> iba$x.scores
 t1        t2         t3
1  -0.6429 -0.07471 -0.764316
2  -0.7697 -0.15463 -0.366123
3  -0.9074  0.20078  0.452996
4   0.6884 -0.09726  0.808580
5  -0.4867 -0.24373 -1.363398
6   …………………
> iba$y.scores
 u1        u2        u3
1   0.37145 -0.05444 -0.82290
2   1.34032  0.19638 -0.71715
3   0.08235  0.58493  0.86557
4   0.35497 -0.62863  0.74383
5  -0.46312 -0.39857  0.39749
6   …….. .
```

max. covariances

= T

= U

```
> iba$x.wgs
            t1      t2      t3
Weight -0.5899  0.7721 -0.2364
Waist  -0.7713 -0.4522  0.4478
Pulse   0.2389  0.4465  0.8623

> iba$y.wgs
               u1      u2       u3
Tractions -0.6133 -0.2140  0.76029
Push.ups  -0.7470 -0.1556 -0.64638
Jumps     -0.2567  0.9643  0.06443

> iba$cor.xt
            t1     t2       t3
Weight -0.9476 0.4049 -0.19478
Waist  -0.9620 0.1169 -0.07474
Pulse   0.5108 0.6088  0.95034

> iba$cor.yu
               u1     u2       u3
Tractions -0.8802 0.1946  0.61617
Push.ups  -0.9397 0.4257 -0.13368
Jumps     -0.7407 0.9420  0.01581
```

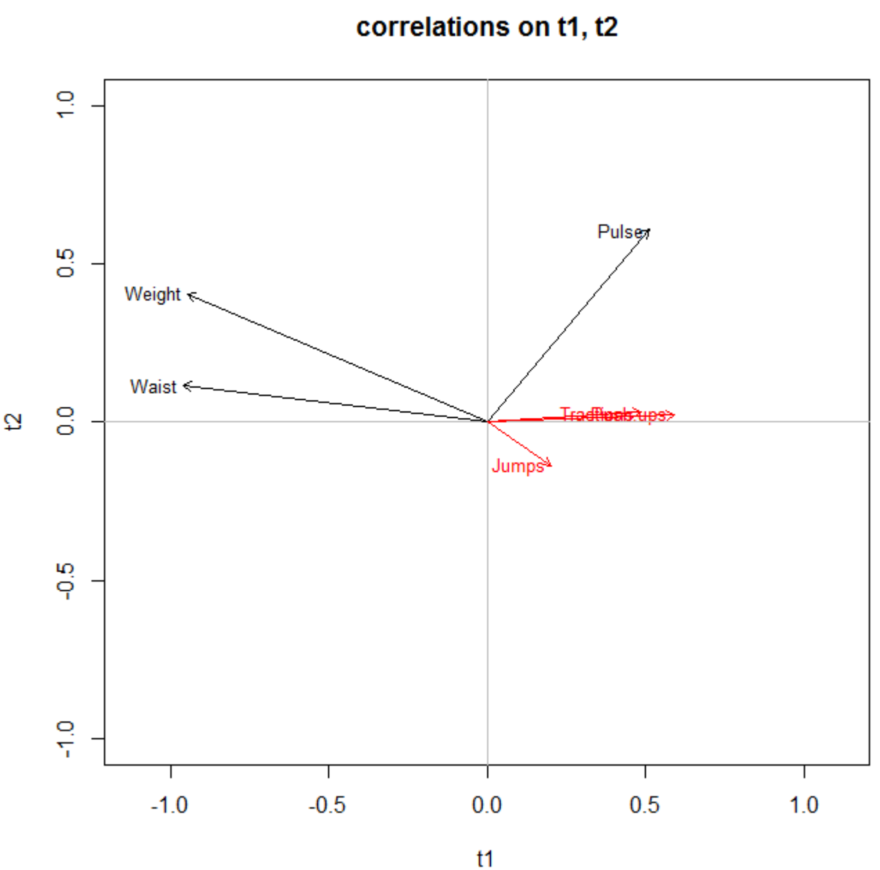
= A

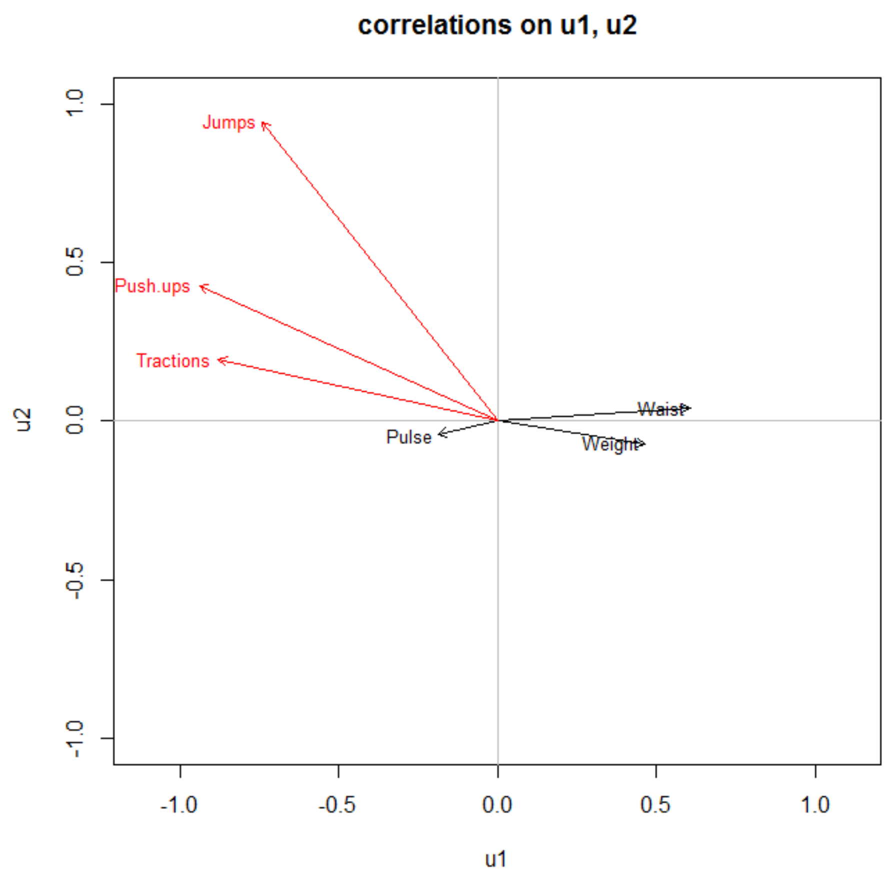
= B

Running IBA

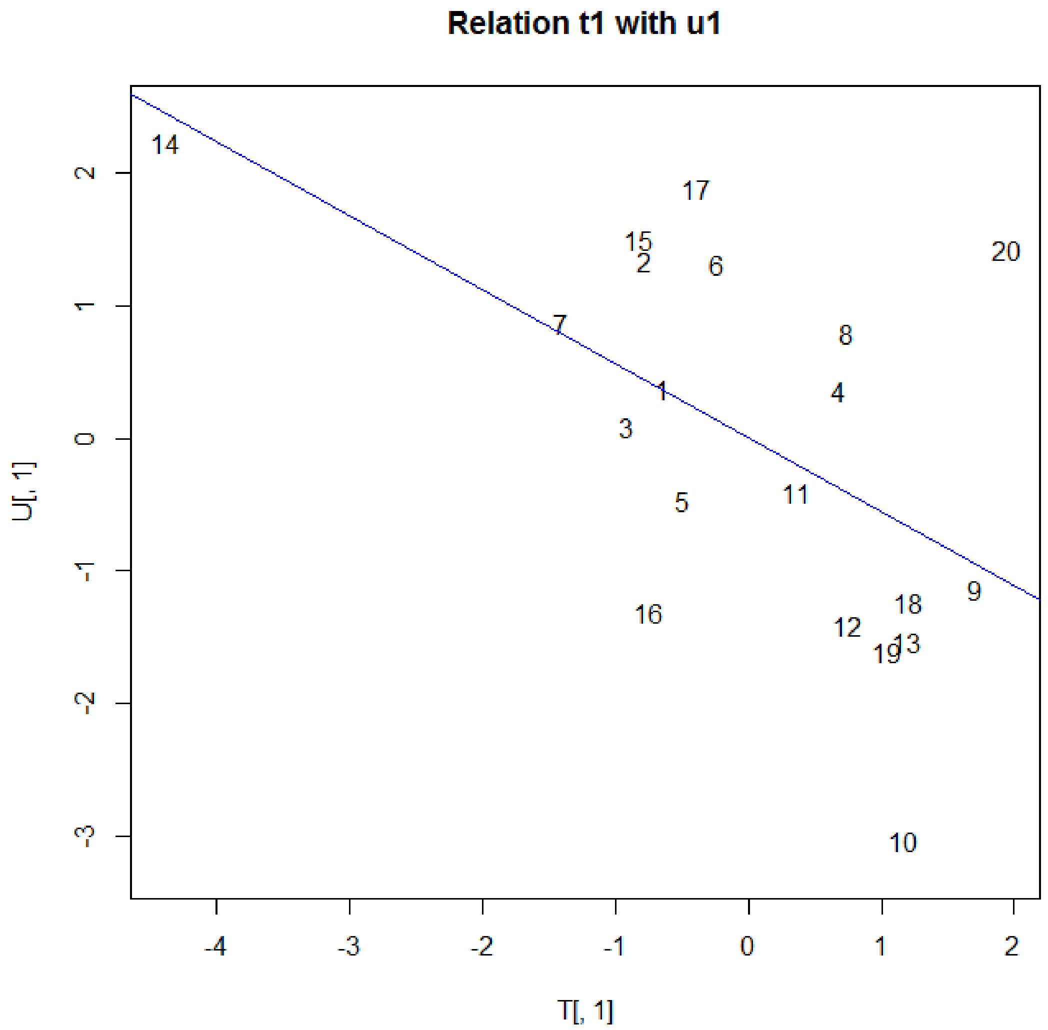
```
> iba$cor.xu
             u1       u2       u3
Weight  0.4647 -0.07254  0.01414
Waist   0.6077  0.04249 -0.02679
Pulse  -0.1882 -0.04195 -0.05158

> iba$cor.yt
               t1       t2        t3
Tractions 0.4862  0.03028 -0.030384
Push.ups  0.5921  0.02202  0.025832
Jumps     0.2035 -0.13643 -0.002575

> iba$cor.tu
           t1         t2         t3         u1         u2         u3
t1  1.000e+00 -1.290e-01  2.808e-01 -5.536e-01 -6.165e-17  4.072e-16
t2 -1.290e-01  1.000e+00  5.789e-01  1.682e-16 -1.767e-01 -2.827e-15
t3  2.808e-01  5.789e-01  1.000e+00 -1.433e-16  2.595e-15 -7.189e-02
u1 -5.536e-01  1.682e-16 -1.433e-16  1.000e+00 -4.743e-01 -1.970e-01
u2 -6.165e-17 -1.767e-01  2.595e-15 -4.743e-01  1.000e+00 -1.197e-01
u3  4.072e-16 -2.827e-15 -7.189e-02 -1.970e-01 -1.197e-01  1.000e+00
```

Correlations





Relation t_1 with u_1



Communalities

```
> for (i in 1:p) {for ( j in 1:a) {com.xt[i,j] <- summary(lm(Xs[,i]~T[,1:j]))$r.squared}}

> rownames(com.xt) <- names(X)

> print(com.xt)
         [,1]   [,2] [,3]
Weight 0.8980 0.9792    1
Waist  0.9255 0.9255    1
Pulse  0.2609 0.7239    1
```

```
> for (i in 1:q) {for ( j in 1:a) {com.yu[i,j] <- summary(lm(Ys[,i]~U[,1:j]))$r.squared}}

> rownames(com.yu) <- names(Y)

> print(com.yu)
            [,1]   [,2] [,3]
Tractions 0.7747 0.8388    1
Push.ups  0.8830 0.8835    1
Jumps     0.5487 0.9988    1
```

Redundancies

```
> for (i in 1:q) {for ( j in 1:a) {red.yt[i,j] <- summary(lm(Ys[,i]~T[,1:j]))$r.squared}}

> rownames(red.yt) <- names(Y)

> print(red.yt)
            [,1]    [,2]   [,3]
Tractions 0.2363 0.24514 0.3396
Push.ups  0.3506 0.36044 0.4365
Jumps     0.0414 0.05374 0.0539
```

```
> for (i in 1:p) {for ( j in 1:a) {red.xu[i,j] <- summary(lm(Xs[,i]~U[,1:j]))$r.squared}}

> rownames(red.xu) <- names(X)

> print(red.xu)
          [,1]    [,2]    [,3]
Weight 0.21597 0.24418 0.26792
Waist  0.36926 0.51037 0.54784
Pulse  0.03542 0.05763 0.07487
```

Number of significant dimensions

```
> iba$statistic
      phi df  p.value
1 25.5838  9 0.002389
2  0.5842  4 0.964800
3  0.1034  1 0.747824
```

$$H_0 \quad \gamma_{1\ldots h} > 0, \gamma_{h+1\ldots s} = 0$$

```
> print(RMPRESS)
  Tractions Push.ups Jumps
1    0.8902   0.8199 0.991
2    0.9046   0.8473 1.054
3    0.9200   0.9124 1.090
>
> print(R2cv)
  Tractions Push.ups    Jumps
1   0.13655  0.26761 -0.07009
2   0.10841  0.21772 -0.21050
3   0.07789  0.09289 -0.29409
```

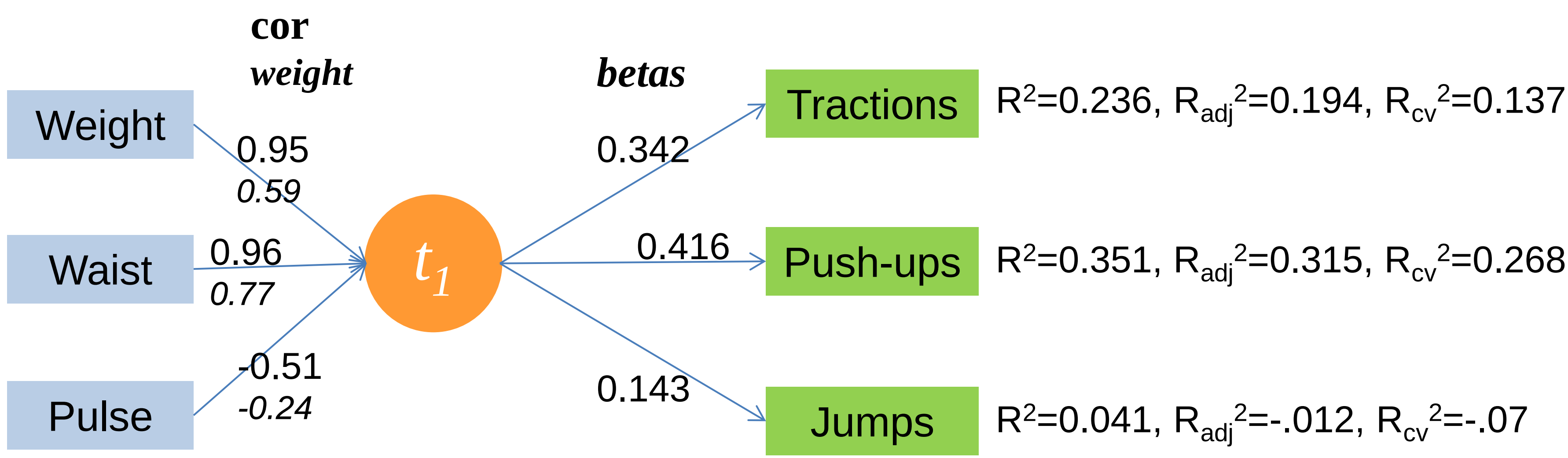
Running IBA

```
> lmY <- lm(Ys~T[,1]-1)
> summary(lmY)
Response Tractions :
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
T[, 1]    0.342      0.141    2.42    0.025 *
Residual standard error: 0.874 on 19 degrees of freedom
Multiple R-squared:  0.236,     Adjusted R-squared:  0.196
F-statistic: 5.88 on 1 and 19 DF,  p-value: 0.0254

Response Push.ups :
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
T[, 1]    0.416      0.130     3.2   0.0047 **
Residual standard error: 0.806 on 19 degrees of freedom
Multiple R-squared:  0.351,     Adjusted R-squared:  0.316
F-statistic: 10.3 on 1 and 19 DF,  p-value: 0.00469

Response Jumps :
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
T[, 1]    0.143      0.158    0.91     0.38
Residual standard error: 0.979 on 19 degrees of freedom
Multiple R-squared:  0.0414,    Adjusted R-squared:  -0.00905
F-statistic: 0.821 on 1 and 19 DF,  p-value: 0.376

> summary(manova(lmY))
          Df Pillai approx F num Df den Df Pr(>F)
T[, 1]     1  0.432     4.31      3     17   0.02 *
Residuals 19
```

Running IBA

```
> # coefficients as functions of the original variables
>
> b.coef <- A[,1:nd] %*% lmY$coefficients
> rownames(b.coef) <- names(X)

> print(b.coef,digits=4)
       Tractions Push.ups    Jumps
Weight  -0.20152 -0.24545 -0.08434
Waist   -0.26351 -0.32094 -0.11029
Pulse    0.08161  0.09939  0.03415

> cor(X,Y)
       Tractions Push.ups    Jumps
Weight   -0.3897  -0.4931 -0.22630
Waist    -0.5522  -0.6456 -0.19150
Pulse     0.1506   0.2250  0.03493
```

Summary of IBA results



Com = 0.69 Red = 0.21= mean(R^2) mean(R_{cv}^2) = 0.11