Partial Least Squares Regression

Y Centered, X_i has mean(X_i)=0, Var(X_i)=1 for all i.

1. $\hat{\varphi}_{1j} = \langle \mathbf{x}_j, \mathbf{y} \rangle$: regressing $\mathbf{y}$ on each $\mathbf{x}_j$
2. $z_1 = \sum \hat{\varphi}_{1j} \mathbf{x}_j$
3. $\hat{\theta}_1 = \langle \mathbf{z}_1, \mathbf{y} \rangle / \langle \mathbf{z}_1, \mathbf{z}_1 \rangle$ coefficient of regressing $\mathbf{y}$ on $\mathbf{z}_1$,
4. Update the $\mathbf{x}_i$'s by orthogonalizing them w/r $\mathbf{z}_1$.
5. Update $\mathbf{y}$ by the residuals of the previous linear fit.

Iterate these 5 steps

This produces a sequence of orthogonal vectors $\{\mathbf{z}_i\}$ and a sequence of estimators $\hat{\beta}_j^{PLS}$

R program for PLS

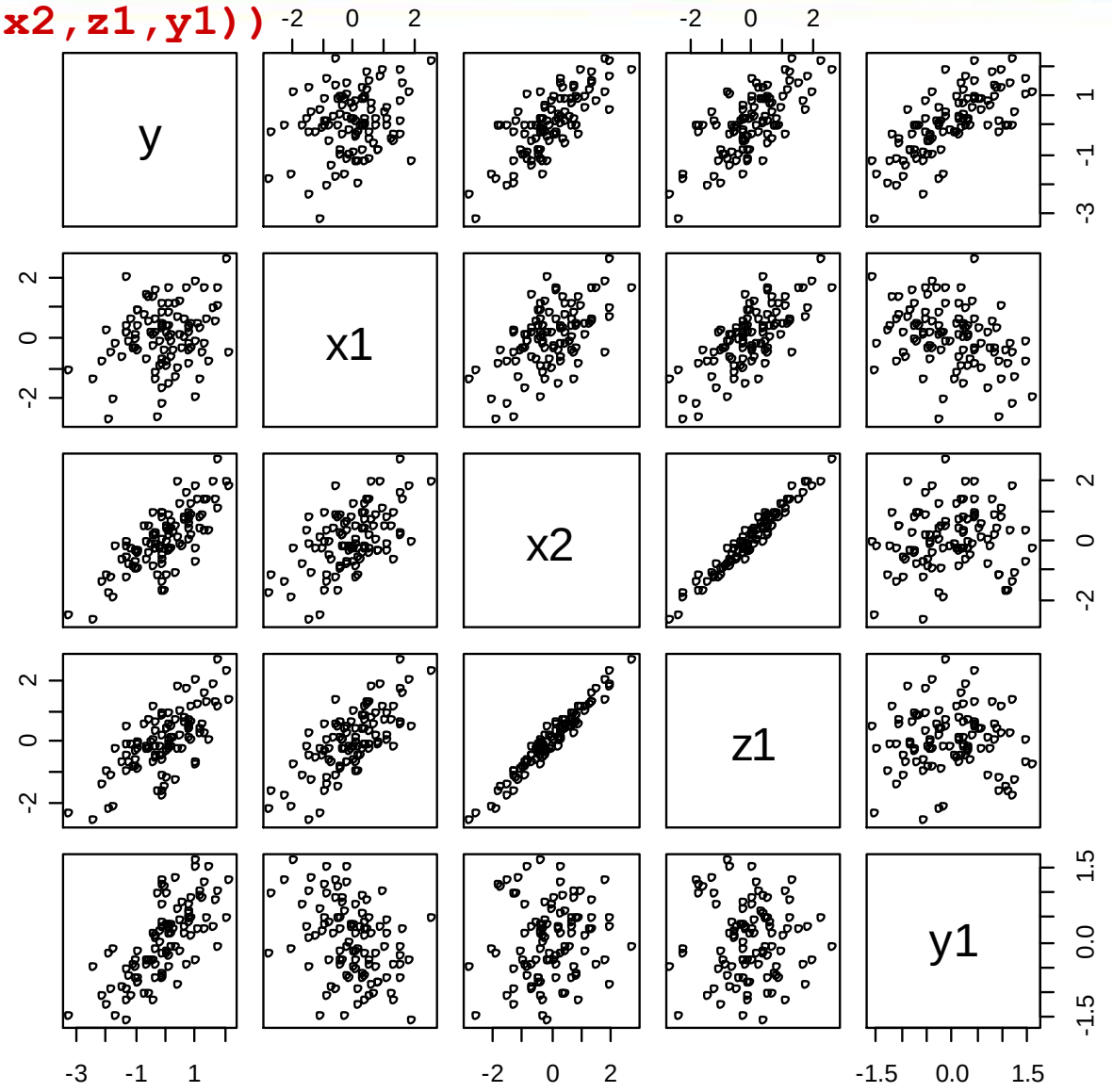
```
# Generate some data
y0 = rnorm(100)        # y: response
y = y0 - mean(y0)      # y: centered - not strictly needed
x1 = rnorm(100)        # define x1 the first predictor
x1 = (x1 - mean(x1))/sd(x1)   # x1 standardized
x2 = y+x1+rnorm(100)          # define x2 the second predictor
x2 = (x2 - mean(x2))/sd(x2)

# We have defined the data: 3 variables y x1 x2

# Start First iteration
pi1 = sum(y*x1)               # define the coef of the 1st PLS
pi2 = sum(y*x2)               #
z1 = pi1*x1 + pi2*x2          # z1 is first PLS
z1 = (z1 - mean(z1))/sd(z1)   # z1 standardized
th1 = lsfit(z1,y,int=F)$coef # calculate reg coef of z1
```

Scatter Matrix of intermediate vars

```
pairs(cbind(y,x1,x2,z1,y1))
```



R program (cont.)

```
# Finish first iteration
y1 = y - th1*z1                          # calculate new responses
x11 = x1 - sum(x1*z1)*z1/sum(z1*z1) # orthogonal to z1
x21 = x2 - sum(x2*z1)*z1/sum(z1*z1) # orthogonal to z1

# Now we do the second iteration.
   phi1 = sum(y1*x11)
   phi2 = sum(y1*x21)
   z2 = phi1*x11 + phi2*x21
   z2 = (z2 - mean(z2))/sd(z2)
   th2 = lsfit(z2,y1,int=F)$coef

    y2 = y1 - th2*z2
# Another way to calculate z2:
   z2 = (x11-mean(x11))/sd(x11)
   pairs(cbind(y1,x11,x21,z1,z2))
```

R program (cont.)

```
# write a function that does it

fpls = function(x,y,k) {
  x1 = x
  z = x[,1:k]*0
  theta = NULL
  phi = array(NA, dim=c(k,ncol(x)) )
  for(i in 1:k) {
  #  start by standardizing the variables
    y1 = y - mean(y)
  for( j in 1:ncol(x))  x1[,j] = (x1[,j] -
   mean(x1[,j]))/sd(x1[,j])
  phi[i,] = apply(x1*y1,2,sum)
.
.
}
```