PCA OF THREE VARIABLES IN XLISP-STAT

JAN DE LEEUW

Suppose $u_i = (x_i, y_i, z_i)$ are n points in $\mathbb{R}^3$. We want to find direction cosines $w = (w_1, w_2, w_3)$ and an intercept $a = (a_1, a_2, a_3)$ such that the line $\mathcal{L} = \{y \mid y = a + \lambda w\}$ approximates the u_i as closely as possible.

The loss function we use is

$$\sigma(a, w) = \sum_{i=1}^{n} \min_{\lambda_i} \ (u_i - a - \lambda_i w)'(u_i - a - \lambda_i w).$$

This means that we project the u_i perpendicularly on the line, and measure the sum of squared distances of the u_i and the projections $\hat{u}_i$. Then choose a and w, with $w'w = 1$, such that $\sigma(a, w)$ is minimized.

It is clear that this problem is identical to the problem of minimizing

$$\sigma(a, w, \lambda) = \sum_{i=1}^{n} (u_i - a - \lambda_i w)'(u_i - a - \lambda_i w)$$

over all three sets of variables, with constraint $w'w = 1$. This show that we may also require, without loss of generality, that the λ_i sum to zero.

We first minimize of a for given w and λ. This gives $\hat{a} = u_\bullet$, with superscript $\bullet$ indicating mean. Thus

$$\min_a \ \sigma(a, w, \lambda) = \sum_{i=1}^{n} (\tilde{u}_i - \lambda_i w)'(\tilde{u}_i - \lambda_i w),$$

with tildes over symbols indicating deviations from the mean. Now minimize of the λ_i, which must add up to zero. The solution is $\hat{\lambda}_i = w'\tilde{u}_i$, which indeed adds up to zero. We now find

$$\min_\lambda \min_a \ \sigma(a, w, \lambda) = \sum_{i=1}^{n} \{\tilde{u}_i'\tilde{u}_i - \hat{\lambda}_i^2\}.$$

In the final step of our minimization problem, we *maximize* the sum of squares of the $\hat{\lambda}_i$ over w with $w'w = 1$. But

$$\sum_{i=1}^{n} \hat{\lambda}_i^2 = w'\{\sum_{i=1}^{n} \tilde{u}_i\tilde{u}_i'\}w = w'Cw,$$

Date: February 1, 2021.

where $C = \tilde{U}'\tilde{U}$ is the 3×3 cross-product matrix of the $\tilde{u}_i$. It follows that $\hat{w}$ is the normalized eigenvector corresponding with the dominant eigenvalue of C, or equivalently of the covariance matrix of the u_i. Suppose this dominant eigenvalue is ω. We then draw the line between $\hat{a} - \frac{1}{2}\sqrt{\omega}\hat{w}$ and $\hat{a} + \frac{1}{2}\sqrt{\omega}\hat{w}$, which has length ω.

This procedure can be repeated for the other two eigenvalues and eigenvectors, that capture the other dimensions of variation.

In Xlisp-Stat we simply add a PCA method to the spin-proto. Sending an instance of the spin-proto the PCA message will draw the three principal axis, with squared length equal to the eigenvalues. The code is given below

```
(defmeth spin-proto :pca ()
(let* ((n (send self :num-points))
       (x (send self :point-coordinate 0 (iseq n)))
       (y (send self :point-coordinate 1 (iseq n)))
       (z (send self :point-coordinate 2 (iseq n)))
       (c (* (1- n) (covariance-matrix x y z)))
       (m (list (mean x) (mean y) (mean z)))
       (g (eigen c))
       (e (second g))
       (f (sqrt (first g))))
(send self :dircos m (first e) (/ (elt  f 0) 2))
(send self :dircos m (second e) (/ (elt f 1) 2))
(send self :dircos m (third e) (/ (elt f 2) 2))
(print e)
(print f)
))

(defmeth spin-proto :dircos (m w u)
(send self :abline (+ m (* u w)) (- m (* u w)) )
)

(defmeth spin-proto :abline (a b)
    (send self :add-lines (make-pairs a b))
)

(defun make-pairs (x y)
  (let ((n (length x)))
     (mapcar #'(lambda (z) (list (elt x z) (elt y z))) (iseq n))
))
```