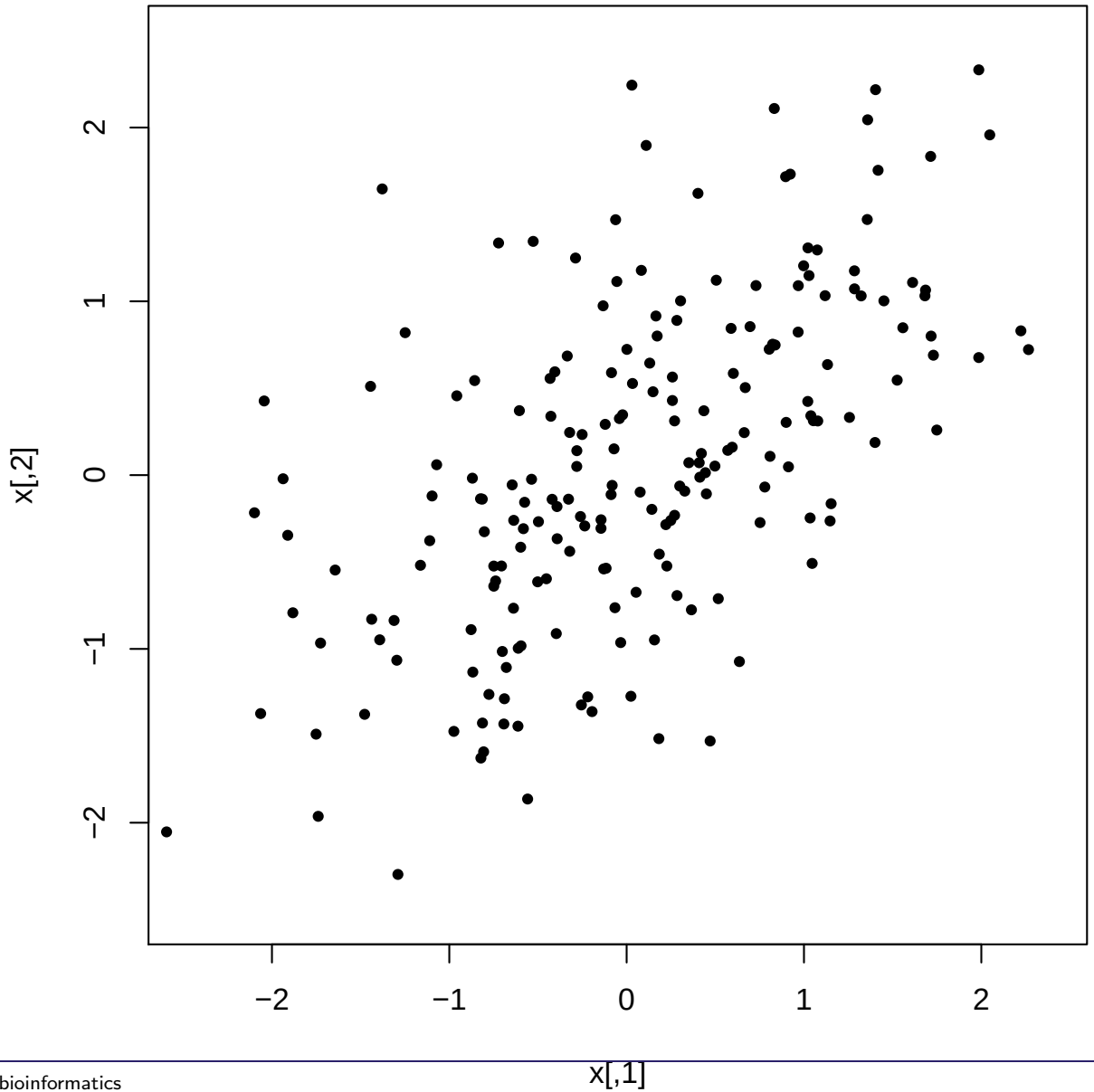
Statistical methods in bioinformatics

Brief introduction, statistical models, dimension reductions.

Claus Thorn Ekstrøm
Biostatistics,
University of Copenhagen
E-mail: ekstrom@sund.ku.dk

Principal component analysis

Dimension reduction of the *covariates*.



Principal component analysis

General algorithm:

1. Compute the covariance matrix of the predictor data set $\mathbf{X}$.
2. Calculate the eigenvalues and corresponding eigenvectors of this covariance matrix
3. The eigenvectors correspond to orthogonal "directions", sort by eigenvalue.

Principal component analysis

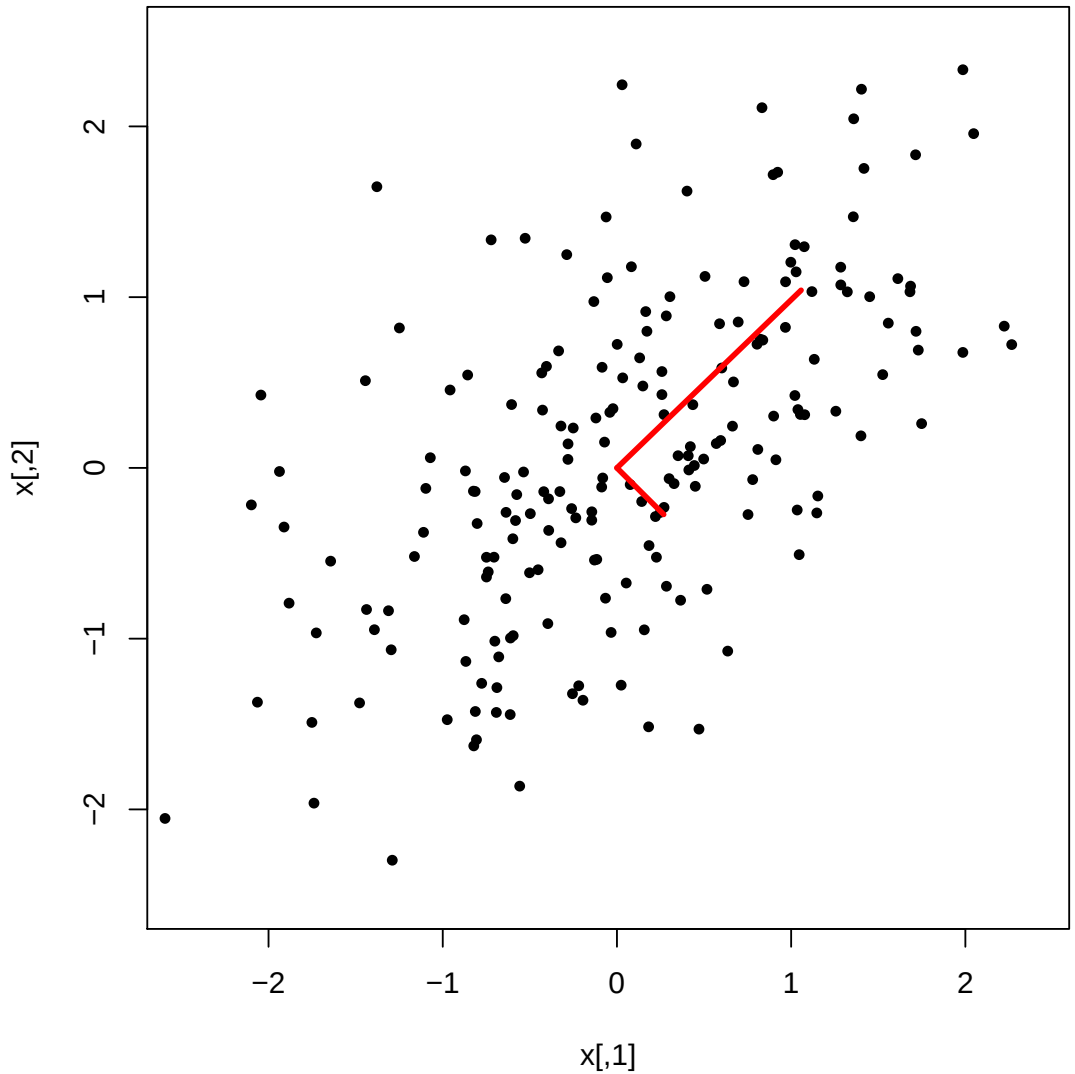
General algorithm:

1. Compute the covariance matrix of the predictor data set $\mathbf{X}$.
2. Calculate the eigenvalues and corresponding eigenvectors of this covariance matrix
3. The eigenvectors correspond to orthogonal "directions", sort by eigenvalue.

Reduce dimensionality so pick a unit vector u, and replace each data point with its projection $u^t x$.
These new data points have variance $u^t \Sigma u$ if Σ was the variance of x. Find u s.t. $u^t \Sigma u$ is maximized which is exactly the eigenvector with the largest eigenvalue.

Principal component analysis

Dimension reduction of the covariates.



Principal component regression

- Instead of smoothly shrinking the coordinates on the principal components, PCR either does not shrink a coordinate at all or shrinks it to zero.
- Keep the k largest eigenvalue components and use the k projection on them as input to a GLM.
- Discrete shrinkage effect compared to ridge regression.
- Ridge regression shrinks the coefficients of the principal components, with relatively more shrinkage applied to the smaller components than the larger; principal components regression discards the $p - k$ smallest eigenvalue components.

Example — PCR

5 components

```
  Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.9500     1.3270  13.527 3.05e-15 ***
prPC1         0.3544     0.3245   1.092   0.2824
prPC2         0.1961     0.3363   0.583   0.5637
prPC3        -0.1120     0.3397  -0.330   0.7436
prPC4        -0.6515     0.3486  -1.869   0.0702 .
prPC5         0.1130     0.3526   0.320   0.7506

Residual standard error: 8.393 on 34 degrees of freedom
Multiple R-squared:  0.1335,Adjusted R-squared:  0.00606
F-statistic: 1.048 on 5 and 34 DF,  p-value: 0.4061
```