Example data sets – already preprocessed

• Yeast cell-cycle data
  ~679 genes (exhibiting periodic behavior)
  15 time points – cdc time course
  no missing entries
  normalized (to green) log-ratios, from spotted arrays
  regular (15) and row-standardized columns (only first 12)
  Spellman et al. (1998)

• Mouse tissues data
  ~459 genes (present in tissues under consideration, sequence info available)
  25 tissues
  no missing entries.
  logs of signals, normalized to overall average, from affy.
  regular and row-standardized columns
  Su et al. (2002)
**Principal Components** (PCA or equivalently Singular Value Decomposition)

$N$ points in $R^T$ ($N = \# \text{ of genes, } T = \# \text{ of conditions}$)

based on the variability of the data cloud:

- Extract a few basic expression patterns (find a subspace).
- Give a low-dimensional reconstruction of the gene expression profiles (project the points)

- As a “structural summary” of the data
- As a “cleaning” step prior to further analyses

$$P_SX_i = a_{1,i}V_1 + a_{2,i}V_2$$
A $k$-dimensional projection preserving most of the variability structure of the data cloud.

Spectral decompositions of the covariance matrix

\[ S = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})' = \sum_{j=1}^{T} \lambda_j v_j v_j' \]

\( \{v_1 \ldots v_T\} \) Eigenvectors, orthogonal directions ranked in terms of...

\( \lambda_1 \geq \lambda_2 \ldots \geq \lambda_T \) Eigenvectors, variability

\( a_{1,i} = W_{i,1} = v_1' X_i \ldots a_{T,i} = W_{i,T} = v_T' X_i \)

\( S = \text{Span} (v_1 \ldots v_k) \) $k$-dimensional projection preserving most of the variability structure of the data cloud.

\( P_S X_i = U_i = W_{i,1} v_1 + \ldots + W_{i,k} v_k \)
For both data sets, row (gene) **centering and standardization:**

- eliminate average expression and variation magnitude effects
- restrict analysis to “pure shapes” of gene expression profiles.

Before: cloud variability dominated by average expression magnitude

After: we have “created” a shape; points are on a \((T-1)\)-hyperball surface
How complex is the data? (dimension)

yeast cell cycle data
PC(1,2) ~ 63% ; PC(1,2,3) ~ 74%

mouse tissue data
PC(1,2,3) ~ 47% ; PC(1,2,3,4,5,6) ~ 62%

Basic patterns

(eigenvectors)

Basic patterns

plus another 3 at least
Low dimensional representations (visualization)

(yeast cell cycle data)

(mouse tissue data)

(times)

((the W's)

((the V's))
Identifying genes that “drive” patterns (ranking on projections)

(functions of the $W$’s)

Genes closest to “pure” cycling behavior?  

YLR190W  top ORFs
YOR391C
YKR037C
YML058W
YHR005C
YDR191W
YKL185W
YNL058C
YGR042W
YLR326W ...

Genes relevant to brain&spine?  

NM_013670  top RefSeqs
NM_019634
NM_019675
NM_053076
NM_020012
NM_053076
NM_024287
NM_018794
NM_019999
NM_023429 ...
Principal components for reducing noise and artifacts

(Yeast cell cycle data: emiliorate dampening in amplitude, and trend…
• dis-synchronization
• expression reaction to synchronization drugs, “crowd” effects?)