## Principal Components and Singular Factors

- A $p \times n$ data matrix $X$ (tall, skinny): rows are genes, columns are samples/microarrays
- Correlations, relationships, patterns among genes:
  - clustering, similar expression patterns
  - co-regulated genes, (up/down), genetic pathways
- Correlations, relationships, patterns among samples:
  - different tumor types, clinical outcomes, cell cycle positions, tumor or normal, ...
- Correlation is a global measure: decomposes into constituent sources using
  - principal components analysis (PCA), or (equivalently)
  - singular value decomposition (SVD) for singular factor analysis

## Motivating PCA and Factors: Two genes

- Two genes: gene 1, 2, sample $j$, expression levels $x_{1,j}, x_{2,j}$ for $j = 1, \ldots, n$
- Imagine that there are numbers $b_1, b_2$ and $f_1, \ldots, f_n$ such that
  \[
  x_{1,j} \approx b_1 f_j \quad \text{and} \quad x_{2,j} \approx b_2 f_j
  \]
- or, better,
  \[
  x_{1,j} = b_1 f_j + \epsilon_{1,j} \quad \text{and} \quad x_{2,j} = b_2 f_j + \epsilon_{2,j}
  \]
  for “small” $\epsilon$s and the $\epsilon$s are uncorrelated.
- $f_j$ is the value of a factor on sample $j$, and the factor determines all of the correlation (relationship) between levels of expression on genes 1 and 2
- e.g., $b_1 = b_2$, so $x_{1,j} \approx x_{2,j}$ or at least highly correlated
- e.g., $b_1 = 1, b_2 = -1$, so $x_{1,j} \approx -x_{2,j}$, highly negatively correlated
- e.g., $b_1 = 0$, so $x_{1,j}$ and $x_{1,j}$ are uncorrelated
- Linear regression format: the factor variable is a predictor of each $x$ in the two linear regression models.

- Now suppose the $\epsilon$s are not so small, and are themselves related between gene 1 and 2
- Apply the same idea to the $\epsilon$s – introduces a second factor
- Then
  \[
  x_{1,j} = b_{1,1} f_{1,j} + b_{1,2} f_{2,j} + \epsilon_{1,j} \quad \text{and} \quad x_{2,j} \approx b_{2,1} f_{1,j} + b_{2,2} f_{2,j} + \epsilon_{2,j}
  \]
  (relabelling $b_1 \rightarrow b_{1,1}, b_2 \rightarrow b_{2,1}$ and $f_j \rightarrow f_{1,j}$)
- If the $f_{1,j}$ are uncorrelated with the $f_{2,j}$, this describes patterns of dependence between $x$s ‘driven’ by the two separate, unrelated factors
- Linear regression format: the 2 factor variables are predictors of each $x$ in the two linear regression models.

## PCA and Factor Decompositions

- A $p \times n$ data matrix $X$ – $p$ genes, $n$ samples
- Take row (gene) $i$ and sample (microarray) $j$
- Singular value decomposition (SVD) of $X$ can be expressed exactly as
  \[
  x_{i,j} = b_{i,1} f_{1,j} + b_{i,2} f_{2,j} + \cdots + b_{i,n-1} f_{n-1,j} + b_{i,n} f_{n,j}
  \]
for some numbers $b_{i..}$ and $f_{.}$

(this is just linear algebra; no statistics, and no magic).

- Generally, higher order $b_{i,n}$ terms are small, so

$$x_{i,j} = b_{i,1}f_{1,j} + b_{i,2}f_{2,j} + \cdots + b_{i,k}f_{k,j} + \epsilon_{i,j}$$

for some $k < n$ and some ‘small’ terms $\epsilon_{i,j}$ that are uncorrelated across genes $i$ and arrays $j$ (i.e., they are small, residual ‘noise’ terms)

- Linear regression format: the $k$ factor variables are predictors of each of the $p \times$ response variables in $p$ separate, parallel linear regression models.
- The $k$ factor variables explain variability in the expression patterns of the many genes and represent $k$ different aspects of the correlations, structure, patterns exhibited among the genes and across samples
- The regression parameters $b_{i,r}$ for gene $i$ represent different weightings, or loadings, on factor $r$ for this gene – the factors influence/explain the variation in genes differently due to differing values of these loadings.
- The factors have various alternative names: principal components, principal factors, singular factors, among others

♠ PCA and Factor Decompositions: Matrix/vector form

- Sample $j = 1, \ldots, n$, with $x_j = \text{column } j$ of $X$

$$x_j = b_1f_{1,j} + b_2f_{2,j} + \cdots + b_nf_{n,j}$$

or

$$x_j = \sum_{r=1}^n b_rf_{r,j}$$

where each $b_r$ is a $p \times 1$ column vector of the loadings for all genes on factor $r$ as $r = 1, \ldots, n$
- Or,

$$x_j = Bf_j$$

with $B = [b_1, \ldots, b_n]$ and

$$f_j = \begin{pmatrix} f_{1,j} \\ f_{2,j} \\ \vdots \\ f_{n,j} \end{pmatrix}$$

- Or,

$$X = BF$$

where now $F = [f_1, f_2, \ldots, f_n]$

♠ Important Mathematical features

- The factor variables are uncorrelated, so represent different underlying sources of covariability in the $X$ data
- Sample correlations between any two factor variables is zero
- Sample variance of each factor variable is 1
- Formally, $F$ is an orthogonal matrix: $F'F = I$ and $FF' = I$, where $I$ is the $p \times p$ identity matrix
with
- \( n \times n \) diagonal \( D = \text{diag}(d_1, \ldots, d_n) \) of non-negative values in decreasing order
- the singular values of \( X \)
- \( A \) is \( p \times n \) matrix such that \( B = AD \)
- \( A \) has orthonormal columns: \( A' A = I \) the \( p \times p \) identity matrix

\[ X = BF = ADF \]

- In terms of \( A, D, x_j = ADF_j \) or
  \[ x_{i,j} = a_{i,1}d_1f_{1,j} + a_{i,2}d_2f_{2,j} + \cdots + a_{i,n}d_nf_{n,j} \]

- Singular values describe relative importance of factors in describing relationships and variability in data matrix
- Percent “total variation explained” by factor \( j \) is \( 100d_j^2/\sum_{i=1}^n d_i^2 \)
- Elements in column \( r \) of \( A \) describe relationships among genes due to factor \( r \)
- Elements in rows \( r \) of \( F \) describe relationships among samples/microarrays due to factor \( r \)

\[ \star \] Properties and More Interpretation

- The factors (principal components) are themselves linear combinations of the data variables, namely \( f_j = D^{-1}A'x_j \) for each sample \( j \), or
  \[ f_{i,j} = a_{1,i}d_1^{-1}x_{1,j} + a_{2,i}d_2^{-1}x_{2,j} + \cdots + a_{p,i}d_p^{-1}x_{p,j} \]

- In fact, among all possible linear combinations of the data variables, the factors are those that explain the most “variability” in the \( x \) data, in the sense that
  - The first factor is the linear combination of the data that has the largest sample variance. (For any vector \( c \), compute the \( n \) values \( c'x_j \), \( j = 1, \ldots, n \), and then find the sample variance of these \( n \) values; choose another vector \( c \), do it again; the largest variance arises when \( c \) is the first column of \( AD^{-1} \) so that the linear combination is the first factor.)
  - The second factor is the linear combination of the data that has the the largest sample variance once corrected for the first factor (and subject to being orthogonal)
  - Many other properties (e.g., see Ripley section 9.1)

\[ \star \] Plotting Data and Factors

- Often informative displays are achieved by plotting factors against sample number and scatter plotting data on pairs of factors
- Useful for discrimination of samples: finding patterns and structure in the \( n \) samples that may be related to a biological state or features (e.g., tumor versus normals)
- First factor often represents average levels of genes in each sample
- Clustering methods can be applied, often most usefully, to factors rather than the full data set – computational efficiencies
- Higher order factors can represent small, idiosyncratic features in data
- Factors can be most useful in regression models as predictors of outcomes

\[ \star \] Practical consideration

- PCA/SVD depends on scale of measurement of variables
- Gene expression on a standard scale – same for all genes
  - best on some kind of log scale
  - require normalisation off all arrays to a standard scale