

### ♠ Principal Components and Singular Factors

- $p \times n$  data matrix  $\mathbf{X}$  (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  $j = 1, \dots, n$
- Imagine that there are numbers  $b_1, b_2$  and  $f_1, \dots, f_n$  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  $x_{2,j}$  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

$$x_{1,j} = b_{1,1} f_{1,j} + b_{1,2} f_{2,j} + \epsilon_{1,j}^* \quad \text{and} \quad x_{2,j} = b_{2,1} f_{1,j} + b_{2,2} f_{2,j} + \epsilon_{2,j}^*$$

(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

- $p \times n$  data matrix  $\mathbf{X}$  –  $p$  genes,  $n$  samples
- Take row (gene)  $i$  and sample (microarray)  $j$
- Singular value decomposition (SVD) of  $\mathbf{X}$  can be expressed *exactly* as

$$x_{i,j} = b_{i,1} f_{1,j} + b_{i,2} f_{2,j} + \dots + b_{i,n-1} f_{n-1,j} + b_{i,n} f_{n,j}$$

---

for some numbers  $b_{i,1}$  and  $f_{1,j}$ .  
(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  $p \times n$  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  $j = 1, \dots, n$ , with  $\mathbf{x}_j$  = column  $j$  of  $\mathbf{X}$

$$\mathbf{x}_j = \mathbf{b}_1 f_{1,j} + \mathbf{b}_2 f_{2,j} + \cdots + \mathbf{b}_n f_{n,j}$$

or

$$\mathbf{x}_j = \sum_{r=1}^n \mathbf{b}_r f_{r,j}$$

where each  $\mathbf{b}_r$  is a  $p \times 1$  column vector of the loadings for all genes on factor  $r$  as  $r = 1, \dots, n$

- Or,

$$\mathbf{x}_j = \mathbf{B} \mathbf{f}_j$$

with  $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_n]$  and

$$\mathbf{f}_j = \begin{pmatrix} f_{1,j} \\ f_{2,j} \\ \vdots \\ f_{n,j} \end{pmatrix}$$

- Or,

$$\mathbf{X} = \mathbf{B} \mathbf{F}$$

where now  $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n]$

### ♠ Important Mathematical features

- The factor variables are uncorrelated, so represent different underlying sources of covariability in the  $\mathbf{X}$  data
- Sample correlations between any two factor variables is zero
- Sample variance of each factor variable is 1
- Formally,  $\mathbf{F}$  is an *orthogonal* matrix:  $\mathbf{F}'\mathbf{F} = \mathbf{I}$  and  $\mathbf{F}\mathbf{F}' = \mathbf{I}$ , where  $\mathbf{I}$  is the  $n \times n$  identity matrix

•

$$\mathbf{X} = \mathbf{B}\mathbf{F} = \mathbf{A}\mathbf{D}\mathbf{F}$$

with

- $n \times n$  diagonal  $\mathbf{D} = \text{diag}(d_1, \dots, d_n)$  of non-negative values in decreasing order
- the *singular values of  $\mathbf{X}$*
- $\mathbf{A}$  is  $p \times n$  matrix such that  $\mathbf{B} = \mathbf{A}\mathbf{D}$
- $\mathbf{A}$  has *orthonormal columns*:  $\mathbf{A}'\mathbf{A} = \mathbf{I}$  the  $p \times p$  identity matrix

- In terms of  $\mathbf{A}, \mathbf{D}$ ,

$$\mathbf{x}_j = \mathbf{A}\mathbf{D}\mathbf{f}_j$$

or

$$\mathbf{x}_j = \sum_{r=1}^n \mathbf{a}_r d_r f_{r,j}$$

or

$$x_{i,j} = a_{i,1}d_1f_{1,j} + a_{i,2}d_2f_{2,j} + \dots + a_{i,n}d_nf_{n,j}$$

- 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  $\mathbf{a}_r$  – column  $r$  of  $\mathbf{A}$  – describe relationships among genes due to factor  $r$ .  $\mathbf{A}$  is called the *factor loadings matrix*
- Elements in rows  $r$  of  $\mathbf{F}$  describe relationships among samples/microarrays due to factor  $r$

## ♠ Properties and More Interpretation

- The factors (principal components) are themselves linear combinations of the data variables, namely  $\mathbf{f}_j = \mathbf{D}^{-1}\mathbf{A}'\mathbf{x}_j$  for each sample  $j$ , or

$$f_{i,j} = a_{1,i}d_1^{-1}x_{1,j} + a_{2,i}d_2^{-1}x_{2,j} + \dots + a_{p,i}d_n^{-1}x_{p,j}$$

- In fact, among all possible (unit length) 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  $\mathbf{c}$ , such that  $\mathbf{c}'\mathbf{c} = 1$ , compute the  $n$  values  $\mathbf{c}'\mathbf{x}_j$ , ( $j = 1, \dots, n$ ), and then find the sample variance of these  $n$  values; choose another vector  $\mathbf{c}$ , do it again; the largest variance arises when  $\mathbf{c}$  is the first column of  $\mathbf{A}\mathbf{D}^{-1}$  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)
- Relationship to eigen-decomposition of sample variance matrix  $\mathbf{S}$ . Take (with no loss of generality) centred data (sample mean subtracted by rows) so that  $(n-1)\mathbf{S} = \mathbf{X}\mathbf{X}'$ . SVD implies

$$(n-1)\mathbf{S} = \mathbf{A}\mathbf{D}^2\mathbf{A}'$$

- $\mathbf{A}$  is matrix whose columns are eigenvectors of  $\mathbf{S}$
- Squared singular values are eigenvalues of  $\mathbf{S}$  (up to a constant)
- “Variation explained” is sum of squares of eigenvalues
- $p \geq n$  : - at most  $n$  non-zero eigenvalues:  $\mathbf{S}$  is rank deficient, singular (our case with  $p \gg n$  – many more genes than samples, typically)

- 
- $(n - 1)\mathbf{S} = \sum_j d_j^2 \mathbf{a}_j \mathbf{a}_j'$
  - Variance decomposition: for gene  $i$ ,  $s_i = \sum_j a_{i,j}^2 d_j^2$  showing the role of the factor loadings  $a_{i,j}$  of gene  $i$  on factor  $j$  in explaining variance
  - Covariance decomposition: for genes  $i, k$ ,  $s_{i,k} = \sum_j a_{i,j} a_{k,j} d_j^2$  showing the similar role of the factor loadings, and their signs

### ♠ 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 (unless data are first centred)
- 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
- Exploration of columns of  $\mathbf{A}$  to identify factor loadings patterns: genes “related” in terms of sign and magnitude on a given factor
- Use of image (heat map) displays
- Some matlab functions (ShowGene, pairs and bcpairs, scatter3 and sc3, imagesc, imagecf, ...)

### ♠ 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
- Exploratory uses on selected subsets of genes (variables)