### ♠ Multiple Linear Regression Model

• Recall the model in matrix form:

$$\mathbf{y} = \mathbf{X}' \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

• Predictor variables in  $p \times n$  matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$  (columns are samples, rows are predictor variables)

## SVD of X

• SVD is

# $\mathbf{X} = \mathbf{BF}$ or $\mathbf{X} = \mathbf{ADF}$

where  $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n]$  is  $n \times n$  matrix of factors (columns represent samples, and rows represent factor variables)

### ♦ SVD Regression

• Combine the SVD with the regression model to get

$$\mathbf{y} = \mathbf{F}' \boldsymbol{\theta} + \boldsymbol{\epsilon}$$

with

• 
$$\theta = \mathbf{B}'\boldsymbol{\beta} \text{ or } \theta = \mathbf{D}\mathbf{A}'\boldsymbol{\beta}$$

- Multiple regression on the factor variables themselves as predictors
- n predictor variables, not p
- Regression parameter vector  $\pmb{\theta}$  to estimate
- Dimension reduction of inference/estimation problem when p > n, as is the case in gene expression analyses

#### ▲ Bayesian Analysis and Stochastic Regularisation

- If p > n we end up with n parameters to be estimated with n observations
- Least squares and other standard methods inapplicable: exact fit to observed data, no predictive value ("over-fitting")
- Generally, remove some factors that do not vary or contribute much to the SVD (small values of the singular values in the **D** matrix)
- More useful and formal solutions lie in Bayesian analysis that involves "stochastic regularisation" of the estimation problem estimate  $\theta$  with some partial constraints on values imposed probabilistically (Insert two semesters of statistics in here please!).
- Typically, reduce to a smaller number of factors and then apply Bayesian analysis to the rest
- Corresponding estimation of  $\beta$  via  $\beta = AD^{-1}\theta$

#### ♠ Software, Computation and Summary

- Point estimate analysis: iterative computation of estimates of  $\theta$  that are Bayesian posterior modes (EM algorithms, MAP estimation)
- Full Bayesian analysis using stochastic simulation methods (Markov chain Monte Carlo simulation, Gibbs sampling): see discussion in the *binary regression* context