Multiple Linear Regression Model

- Recall the model in matrix form:
  \[ y = X'\beta + \epsilon \]
- Predictor variables in \( p \times n \) matrix \( X = [x_1, x_2, \ldots, x_n] \) (columns are samples, rows are predictor variables)

SVD of \( X \)

- SVD is
  \[ X = BF \quad \text{or} \quad X = ADF \]
  where \( F = [f_1, f_2, \ldots, f_n] \) 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
  \[ y = F'\theta + \epsilon \]

  with
  \[ \theta = B'\beta \quad \text{or} \quad \theta = DA'\beta \]
- Multiple regression on the factor variables themselves as predictors
- \( n \) predictor variables, not \( p \)
- Regression parameter vector \( \theta \) 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