# Biostatistics 140.754
# Advanced Methods in Biostatistics IV

Jeffrey Leek

Assistant Professor
Department of Biostatistics
jleek@jhsph.edu

Lecture 13

# Tip + Paper

**Tip:** Be a finisher. The key to getting a Ph.D. (other than passing your quals) is the ability to sit down and just power through and get it done. This means sometimes you will have to work late or on a weekend. The people who are the most successful in grad school are the people that just find a way to get it done. If it was easy...anyone would do it.

**Paper of the Day:** "Molecular Structure of Nucleic Acids" http://www.nature.com/physics/looking-back/crick/index.html

# Today's Outline

- Principal components analysis
- Singular value decomposition

# Principal components analysis

- Invented by Karl Pearson
- Generally used for exploratory data analysis
- Intuitive idea: if we have a bunch of variables can we find some smaller set of variables that explain most of the variation.

# Principal components

Suppose we have $n$ measurements on each of $m$ variables $X_j$, $j = 1, \ldots, n$. There are several equivalent components to principal components:

- Produce a derived (and small) set of uncorrelated variables $Z_k = \alpha_k X$, $k = 1, 2, \ldots, q < m$ that are linear combinations of the original variables, and that explain most of the variation in the original data
- Approximate the $n \times m$ matrix $X$ by the best rank-$q$ matrix $\hat{X}_{(q)}$. This is the usual motivation for the SVD.

# Principal components

If $X$ is a random vector with mean 0 and covariance matrix $\Sigma$ then the variance of the linear combination $Z = \alpha^T X$ is given by:

$$\mathrm{Var}(Z) = \alpha^T \Sigma \alpha$$

We are seeking an $\alpha$ such that $\mathrm{Var}(Z)$ is large; clearly we must impose a scale restriction on $\alpha$. This leads to the principal-component criterion:

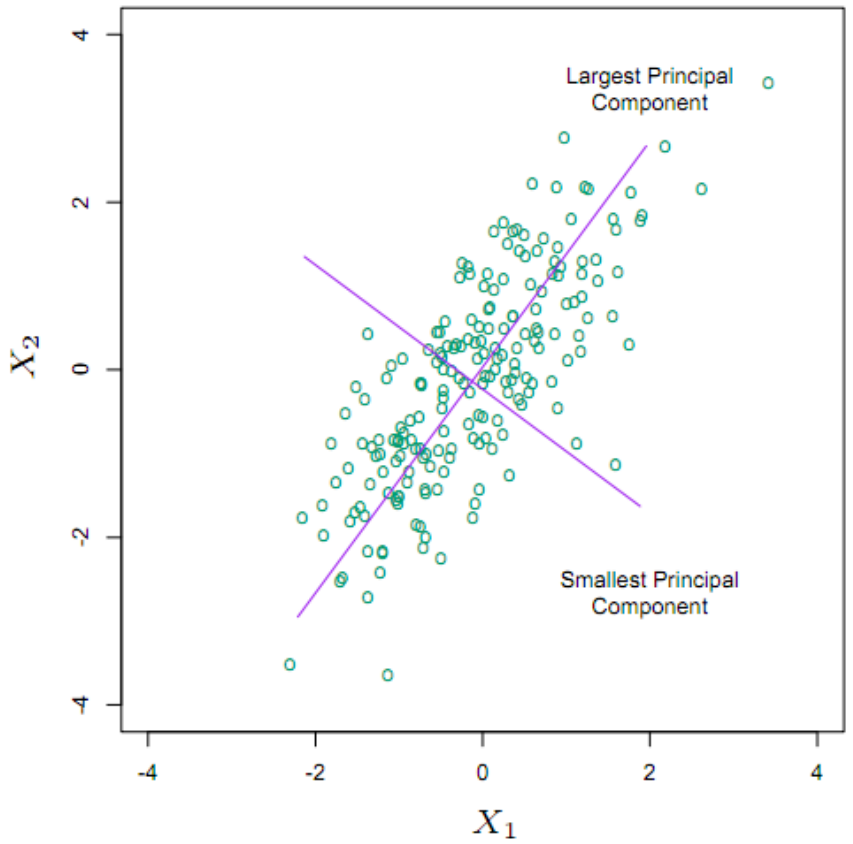
$$\max_\alpha \alpha^T \Sigma \alpha \text{ subject to } ||\alpha|| = 1$$

The solution $\alpha$ is the largest eigenvector of $\Sigma$:

$$\Sigma\alpha = d^2\alpha$$

and $\mathrm{Var}(Z) = \mathrm{Var}(\alpha^T X) = d^2$.

# PC: Derived Variables




$Z_1 = \alpha_1^T X$ is the projection of the data onto the longest direction, and has the largest variance amongst all such normalized projections. $\alpha_1$ is the largest eigenvalue of $\hat{\Sigma}$, the sample covariance matrix of $X$. $Z_2$ and $\alpha_2$ correspond to the second-largest eigenvector.

# PC: Singular Value Decomposition

For any $n \times m$ matrix $X$ (assume $n > m$):

$$X = UDV^T$$

is the SVD of $X$ where

- $U$ is $n \times m$ orthogonal, the left singular vectors
- $V$ is $m \times m$ orthogonal, the right singular vectors
- $D$ is diagonal with $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$, the singular values

The SVD always exists, and is unique (up to signs and ties)

If $X$ is centered (column means zero), then the columns of $V$ are the principal components, and $Z_j = U_j d_j$.

$\frac{d_i^2}{\sum_{j=1}^m d_j^2}$ is the "fraction of variation explained" by the $i$th singular vector .

# SVD: best rank $q$ approximation

Let $D_q$ be $D$ but all but the first $q$ diagonal elements set to zero. Then $\hat{X}_q = UD_qV^T$ solves:

$$\min_{\text{rank}(\hat{X}_q)=q} ||X - \hat{X}_q||_F$$

here $||\cdot||_F$ is the Frobenius norm: $||X||_F = \sqrt{\sum_{i,j} x_{ij}^2}$
$= \sqrt{\text{tr}(X^TX)}$

# SVD: best rank $q$ approximation

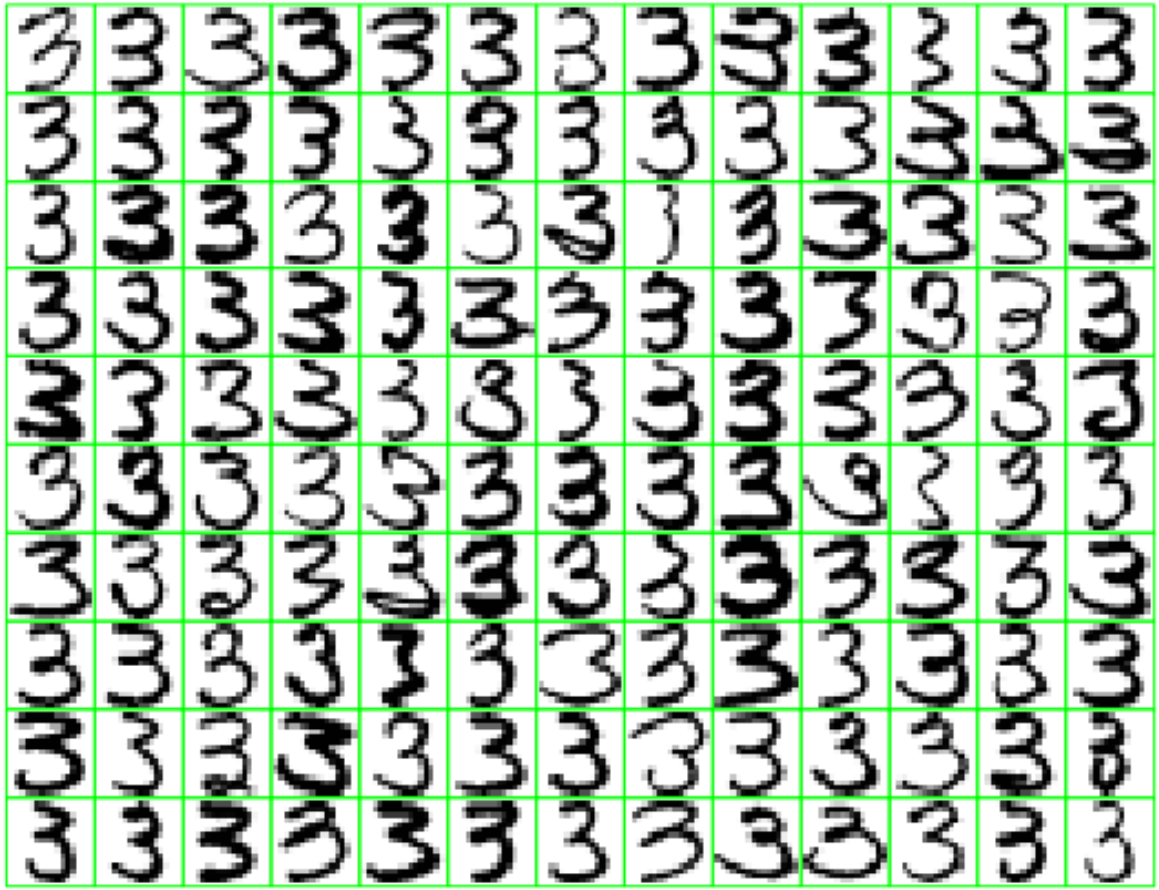
Gabriel 1978 JRSSB showed how to do this with covariates.
Suppose we want to minimize:

$$\min_{\text{rank}(\hat{X}_q)=q} ||X - \beta Z - \hat{X}_q||_F$$

We can do this in three steps:

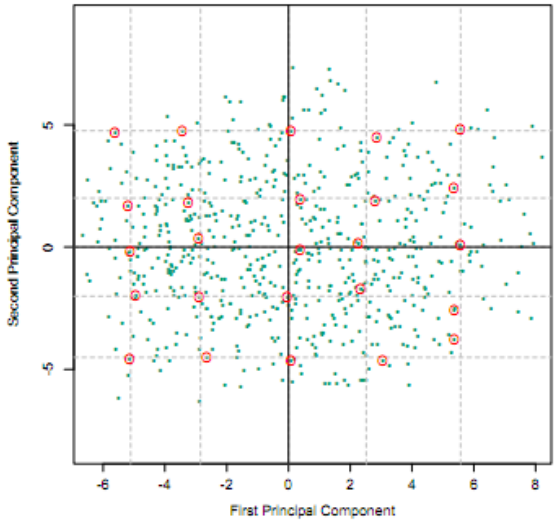
1. Calculate the least squares estimator $\hat{\beta}$ minimizing $||X - \beta Z|_F$
2. Calculate the residuals $R = X - \hat{\beta} Z$
3. Take the singular value decomposition of $R = UDV^T$. Set all but the first $q$ diagonal elements of $D$ to zero.

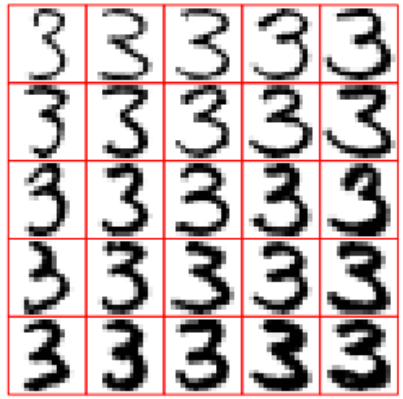
# PC: Exampled – Digit Data



130 threes, a subset of 638 such threes and part of the handwritten digit dataset used in the Elements of Statistical learning. Each three is a $16 \times 16$ greyscale image and the variables $X_j$ $j = 1, \ldots 256$ are the grey scale values for each pixel.

# PC: Exampled – Digit Data



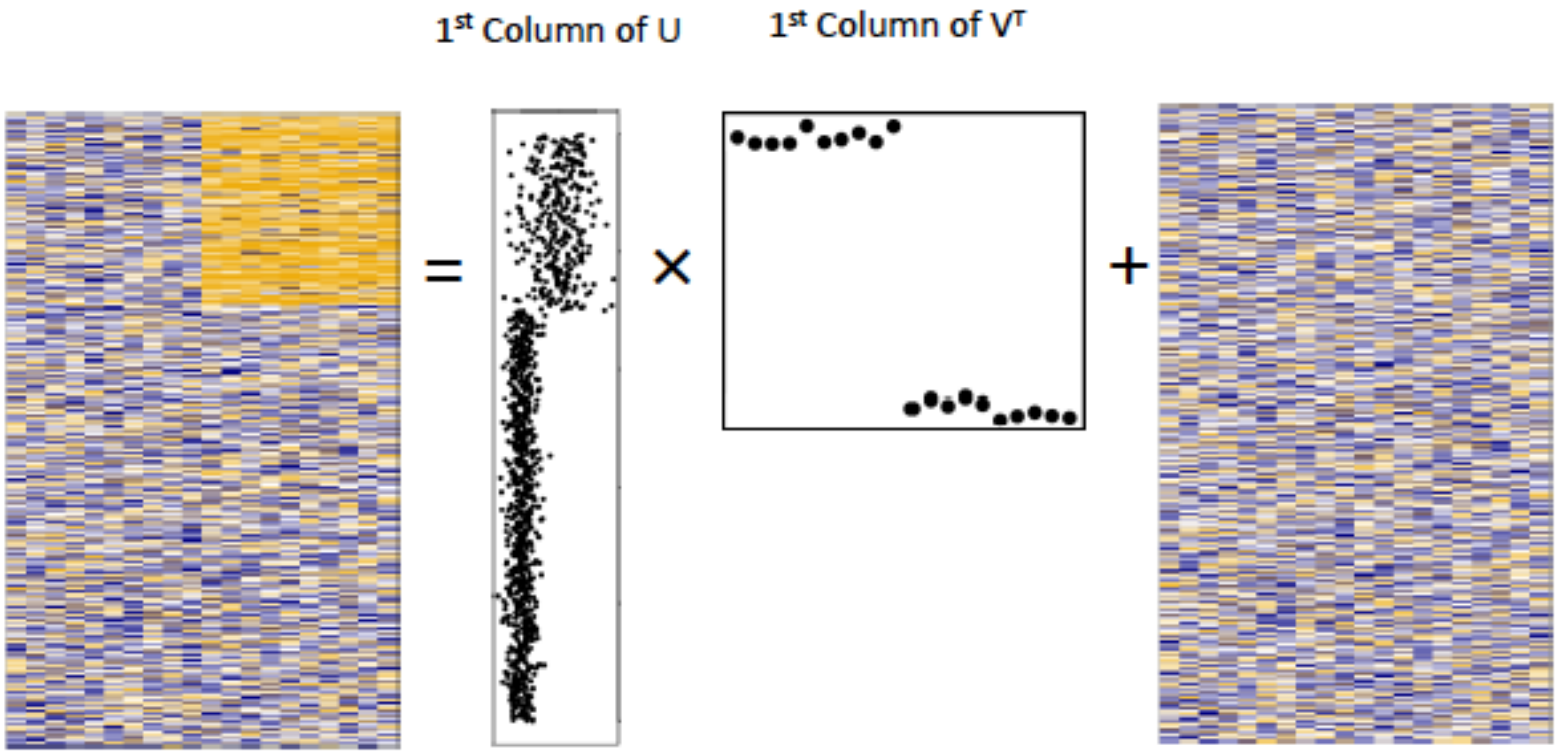
130 threes, a subset of 638 such threes and part of the handwritten digit dataset used in the Elements of Statistical learning. Each three is a $16 \times 16$ greyscale image and the variables $X_j$ $j = 1, \ldots 256$ are the grey scale values for each pixel.

Two-component model has the form:

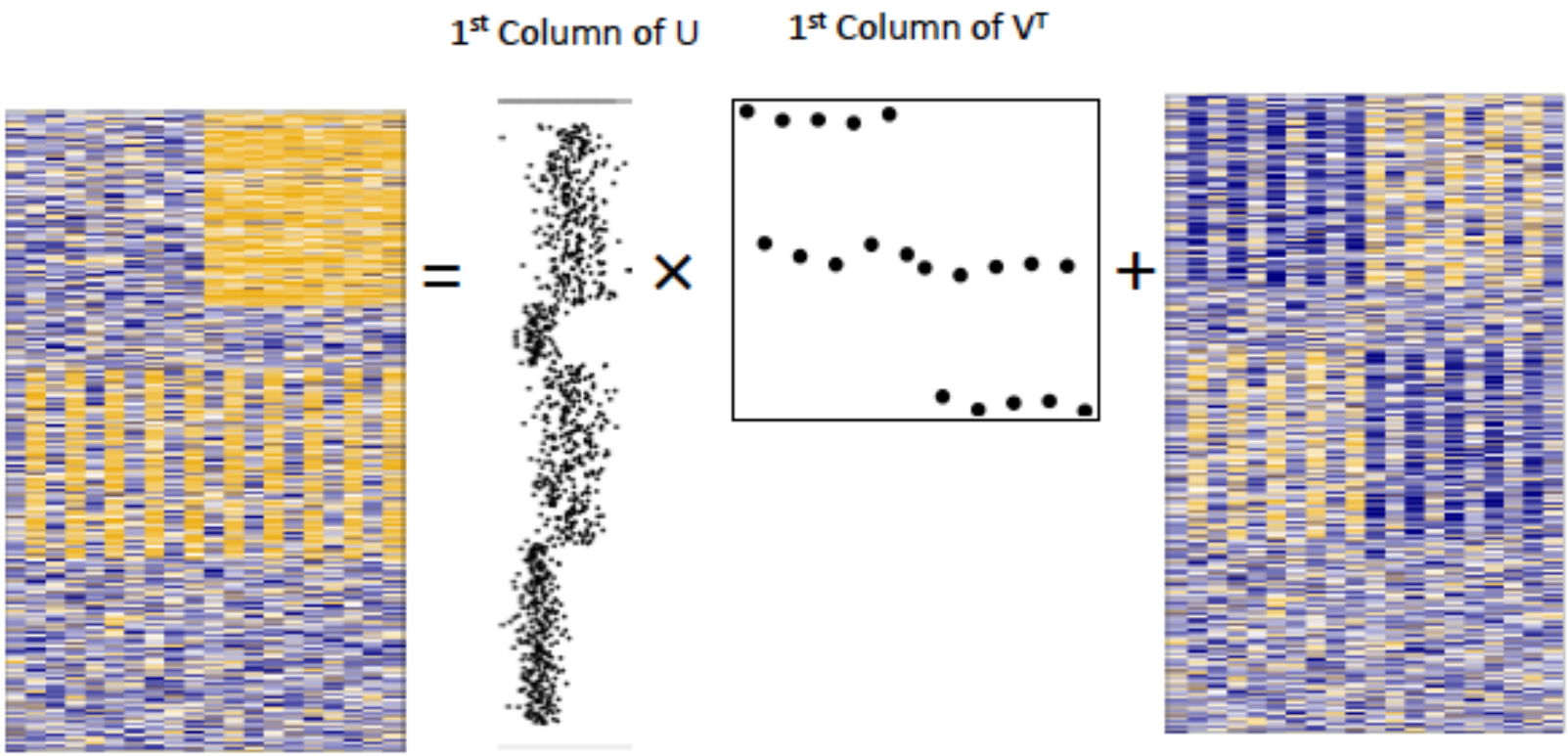
$$\hat{f}(\lambda) = \bar{X} + \lambda_1 v_1 + \lambda_2 v_2$$

$$= \quad \boxed{3} + \lambda_1 \cdot \boxed{3} + \lambda_2 \cdot \boxed{3}.$$

# Interpretability of Singular Vectors/Principal Components

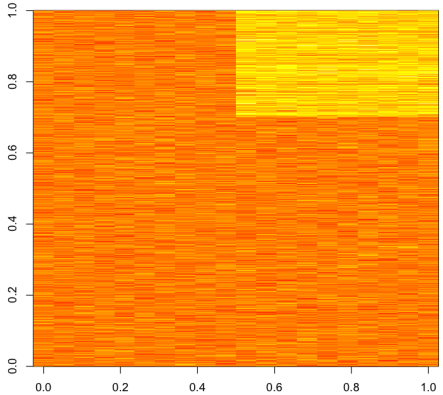


# Interpretability of Singular Vectors/Principal Components



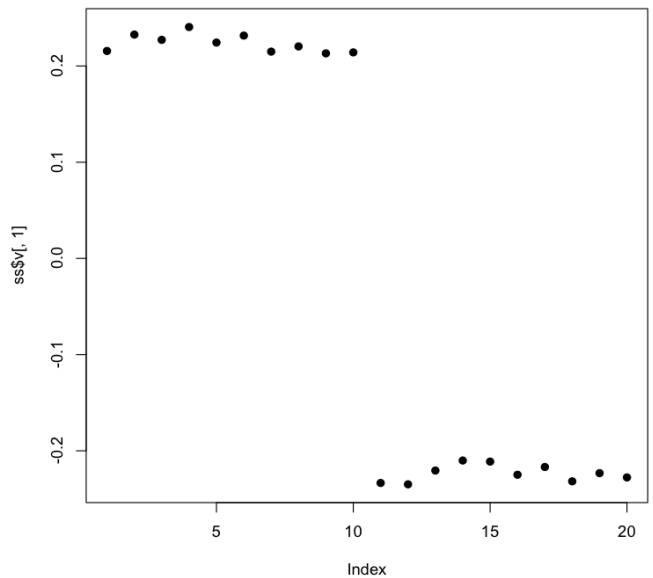
# SVD/PCA in R

```
> dat = matrix(rnorm(1000*20),nrow=1000)
> dat = dat + c(rnorm(300,mean=3),rep(0,700)) %*% t(rep(c(0,1),each=10))
> image(t(dat)[,nrow(dat):1])
```



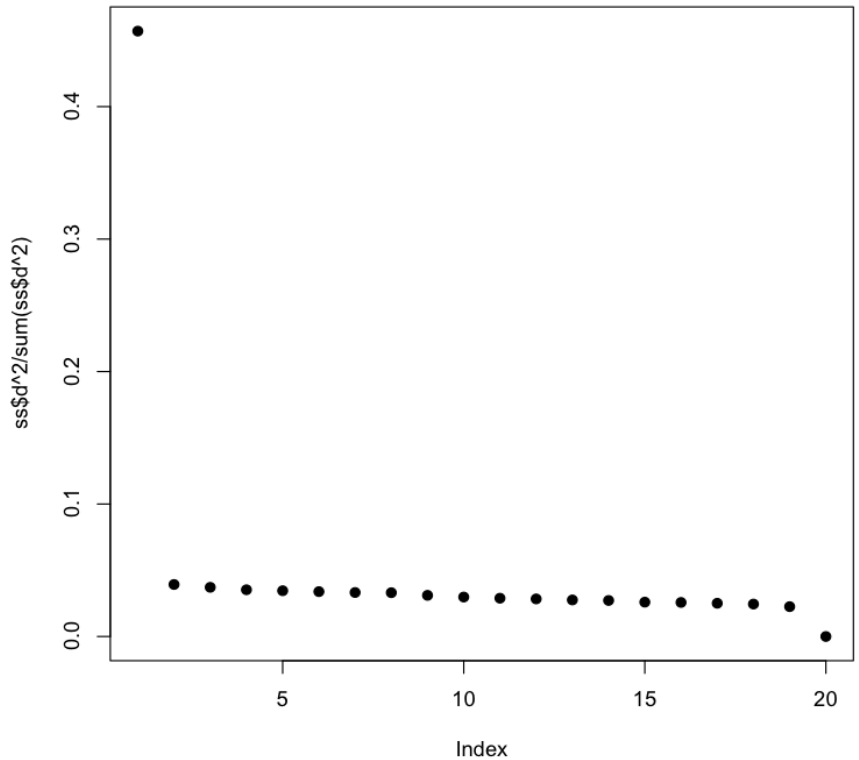

```
> ss = svd(dat - rowMeans(dat))

> plot(ss$v[,1],pch=19)
```



# SVD/PCA in R (Scree plot)

```
> plot(ss$d^2/sum(ss$d^2),pch=19)
```




It's rarely this obvious!

# Choosing the Number of Components

If $m$ is large:

BIOMETRICS DOI: 10.1111/j.1541-0420.2010.01455.x

**Asymptotic Conditional Singular Value Decomposition for High-Dimensional Genomic Data**

**Jeffrey T. Leek**
Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205-2179, U.S.A.
*email:* jleek@jhsph.edu

Otherwise:

- Calculate observed statistics $s_i = \frac{d_i^2}{\sum_{j=1}^m d_j^2}$, $i = 1, \ldots, m$
- Permute each row or column of the data separately to get null matrices $X^{0b}$, $b = 1, \ldots, B$
- Recalculate the SVD and get null statistics $s_i^0 = \frac{d_i^{02}}{\sum_{j=1}^m d_j^{02}}$, $i = 1, \ldots, m$
- Calculate $p$-values for each component:

$$p_i = \frac{1 + \sum_{b=1}^B I(s_i^0 > s_i)}{B + 1}$$

Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis.

# SVD/PCA in R

You might also look in the `corpcor` package if you are dealing with big data sets in only one dimension.

It uses this trick:

$$X = UDV^T$$

So

$$X^T X = VDU^T UDV^T = VD^2 V^T$$

$$XX^T = UDV^T VDU^T = UD^2 U^T$$

The eigenvectors of $X^T X$ are the columns of $V$ and the eigenvectors of $XX^T$ are the columns of $U$. So if $m$ is much smaller than $n$, or vice versa, one of $X^T X$ and $XX^T$ will have small dimension.

# How to Get a Highly Cited Paper

**Singular value decomposition for genome-wide expression data processing and modeling**

**Orly Alter*†, Patrick O. Brown‡, and David Botstein***

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

Both have over 1,000 citations!

# Summary

- SVD and PCA estimate the same thing, but have different motivations
- They are both good for finding patterns in multivariate data
- The components may not always be interpretable (some people like other decompositions like ICA)
- They get used a lot!

# Course Summary

- Focus on the scientific problem first - make sure your models make sense
- Parametric models are not the One True Path...
- Minimizing bias or variance may not be the one true path
- In applied statistics there really is no One True Path
- So be sensible/sane statisticians
- Applied statistics is fun/cool/exciting!