CSE 519: Data Science
Steven Skiena
Stony Brook University

Lecture 16: Linear Regression

Singular Value Decomposition

The SVD of an n*m matrix M factors it $M = UDV^T$ where D is diagonal (weighted identity matrix)

Thus UD weights each column of U by D, as does DV^T.

Retaining only the rows/column with large weights permits us to compress m features with relatively little loss.

Reconstruction from SVD

The outer product of vectors yields a matrix

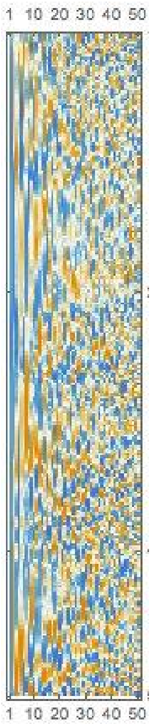
$$P = X \otimes Y \qquad\qquad P[j,k] = X[j]Y[k]$$

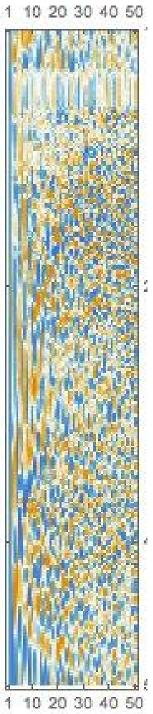
Matrix M can be expressed a sum of outer products from SVD: (UD)_k and (V^T)_k.

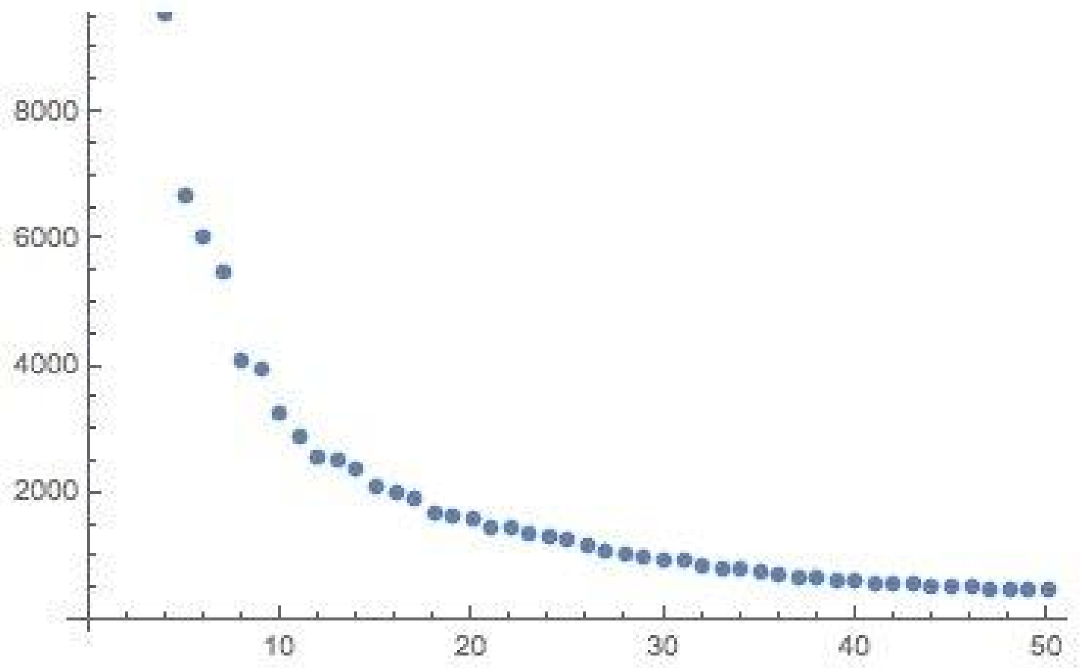
$$C = A \cdot B = \sum_k A_k \otimes B_k^T$$

Summing only the largest matrix products produces an approximation of M

Error Declines with Dimensionality

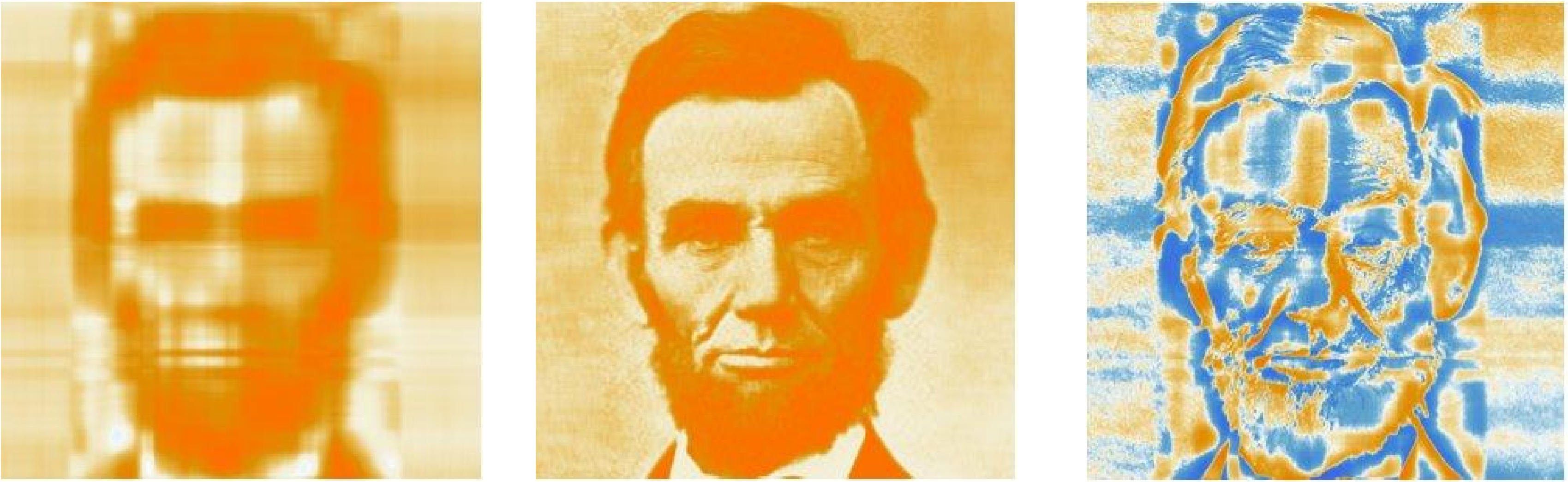






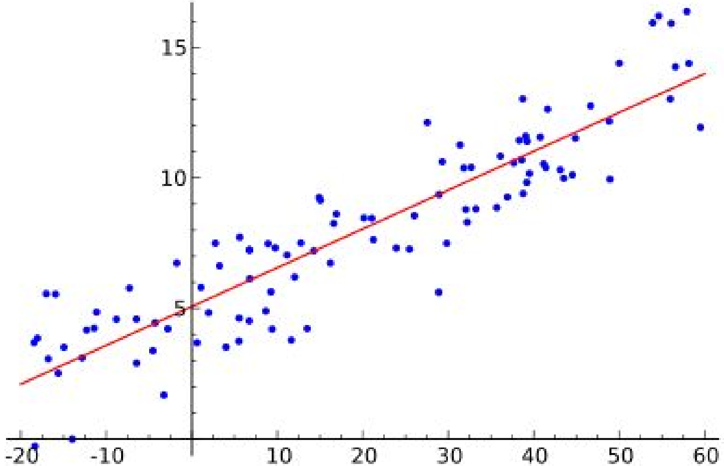
Reconstructing Lincoln

Lincoln’s face from 5 and 50 singular values, a substantial compression of the original matrix.



Linear Regression

Given a collection of *n* points, find the line which best approximates or fits the points.



Why Linear Functions?

Linear relationships are easy to understand, and *grossly* appropriate as a default model:

- Income grows linearly with time worked.
- Housing prices grow linearly with area.
- Weight increases linearly with food eaten.

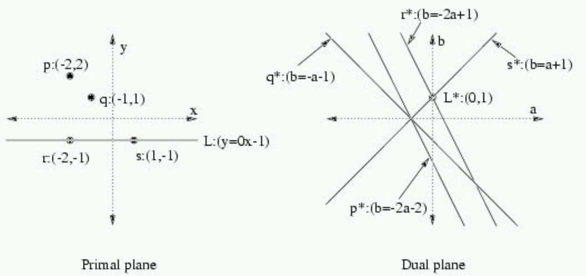
Statistician's rule: If you want a function to be linear, measure it at only two points.

Linear Regression and Duality

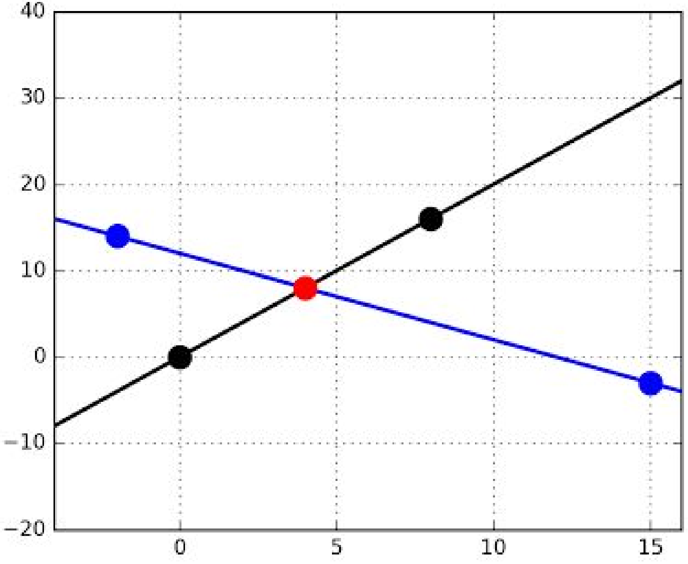
In solving linear systems, given *n* lines we seek the point that lies on all the lines.

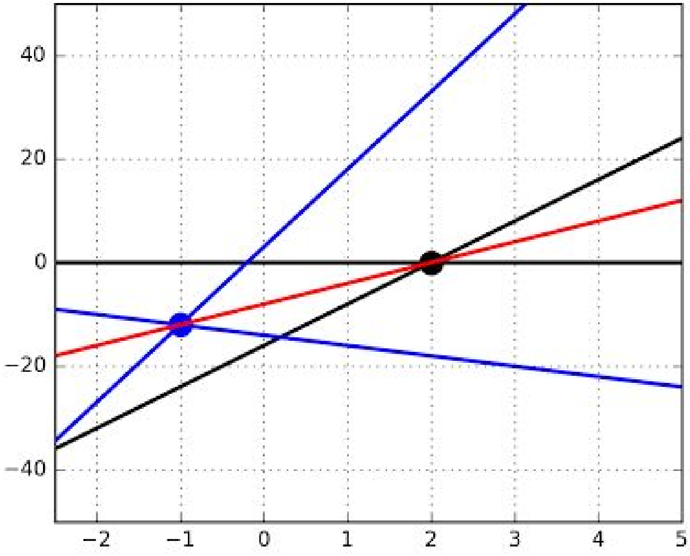
In regression, we seek the line that lies on "all" *n* points.

By the duality transformation (s,t) <-> y= (s)x-t lines are equivalent to points in another space.



Duality Example

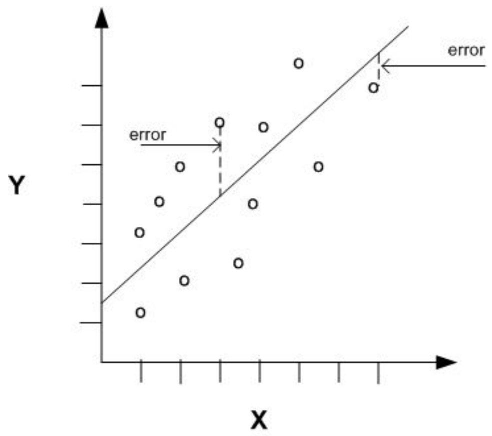




Error in Linear Regression

The residual error is the difference between the predicted and actual values: $r_i = y_i - f(x_i, \beta)$

Least squares regression minimizes the sum of the squares of the residuals of all points.
This metric is chosen because (1) it has a nice closed form and (2) it ignores the sign of the errors.



Solving Linear Regression

Consider the *n*m* system *Ax=b*. The vector *w* of coefficients for the best fitting line is given by:

$$w = (A^T A)^{-1} A^T b$$

Product of *(m*m)*(m*n)*(n*1) (m*1)* is *m*1*

Thus least squares optimization reduces to inversion and multiplying matrices.

Linear Regression in One Variable

We seek the best fitting line $y = w_0 + w_1 x$

The slope of this line is:

$$w_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r_{xy} \frac{\sigma_x}{\sigma_y}$$

The intercept follows since l goes through the x-mean and y-mean.

Connections with Correlation

- If x is uncorrelated with y, w1 should be zero.
- If x,y are perfectly correlated, the slope should depend upon the magnitudes of x,y, as given by t$w = (A^T A)^{-1} A^T b$ions.
- The formula includes correlation-related terms (covariance matrix of variables, and variables against target)

Where Does This Come From?

The error vector *(b-Aw)* must be orthogonal the the vector for each variable, or we could improve the fit by adjusting *w*.

These zero dot products mean $A^T(b - Aw) = 0$

Simple algebra then gives

$$w = (A^T A)^{-1} A^T b$$

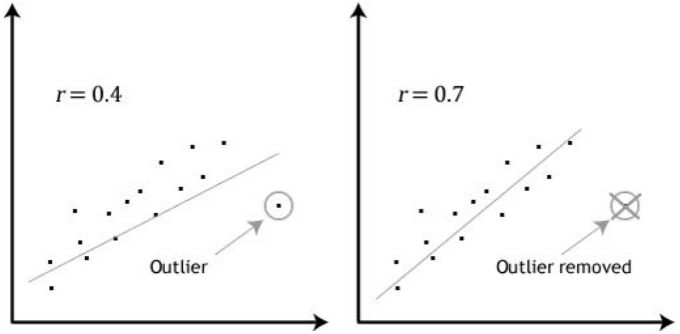
Better Regression Models

Proper treatment of variables yiels better models:

- Removing outliers
- Fitting non-linear functions
- Feature/target scaling
- Collapsing highly correlated variables

Outliers and Linear Regression

Because of the quadratic weight of residuals, outlying points can greatly affect the fit.



Identifying outlying points and removing them in a principled way can yield a more robust fit.

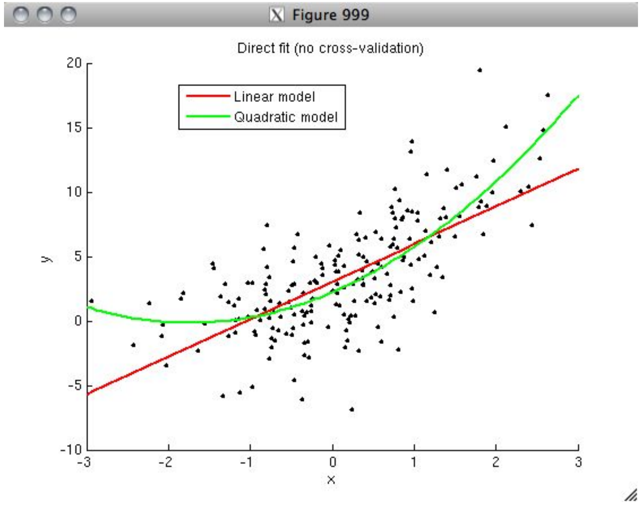
Fitting Non-Linear Functions

Linear regression fits lines, not high-order curves!

But we can fit quadratics by creating another variable with the value x^2 to our data matrix.

We can fit arbitrary polynomials (including square roots) and exponentials/logarithms by explicitly including the component variables in our data matrix: *sqrt(x)*, *lg(x)*, *x^3*, *1/x*.

However explicit inclusion of all possible non-linear terms quickly becomes intractable.



Feature Scaling: Z-scores

Features over wide numerical ranges (say national population vs. fractions) require coefficients over wide scales to bring together.

$$V = c1 * 300{,}000{,}000 + c2 * 0.02$$

Fixed learning rates (step size) will over/under shoot over such a range, in gradient descent.

Scale the features in your matrix to Z-scores!

Dominance of Power Law Features

Consider a linear model for years of education, which ranges from 0 to 12+4+5=19.

$$Y = c1 * income + c2$$

No such model can gives sensible answers for both my kids and Bill Gates' kids.

Z-scores of such power law variables don't help because they are just a linear transformation.

Feature Scaling: Sublinear Functions

An enormous gap between the largest/smallest and median values means no coefficient can use the feature without blowup on big values.

The key is to replace/augment such features *x* with sublinear functions like log(x) and sqrt(x).

Z-scores of these variables will prove much more meaningful.

Small Coefficients Need Small Targets

Trying to predict income from Z-scored variables will *need* large coefficients: how can you get to $100,000 from functions of -3 to +3?

If your features are normally distributed, you can only do a good job regressing to a similarly distributed target.

Taking logs of large targets give better models.

Avoid Highly Correlated Features

Suppose you have two perfectly-correlated features (e.g. height in feet, height in meters).

This is confusing (how should weight be distributed between them?) but worse…

The rows in the covariance matrix are dependent (r1 = c*r2) so $w = (A^T A)^{-1} A^T b$

requires inverting a singular matrix!

Punting Highly Correlated Features

Perfectly correlated features provide no additional information for modeling.

Identify them by computing the covariance matrix: either one can go with little loss.

This motivates the problem of dimension reduction: e.g singular value decomposition, principal component analysis.