Data 451
Principal Components Analysis

Dennis Sun

January 11, 2017

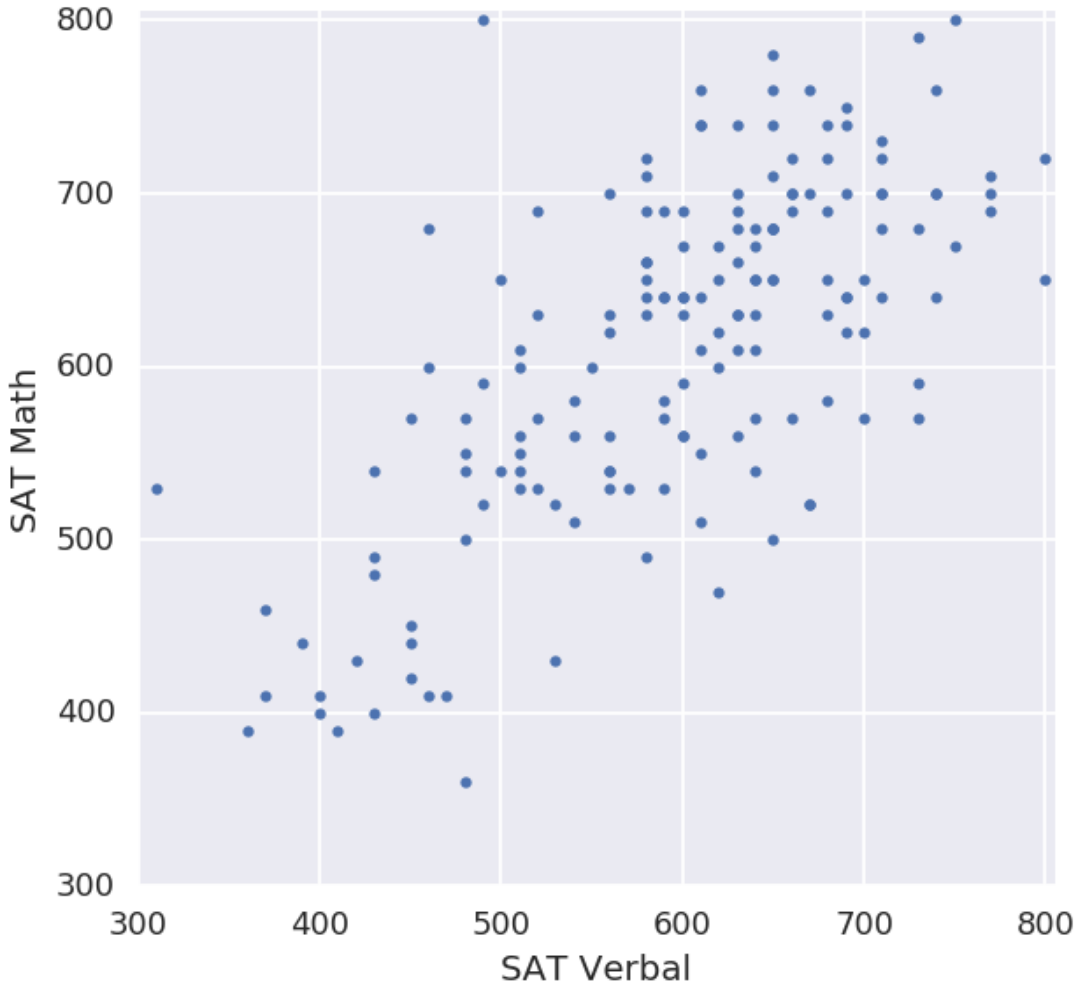
1. Prelude: Intelligence and the g Factor
2. The Math Behind PCA
3. Back to the g factor
4. Scores and Dimensionality Reduction

1. Prelude: Intelligence and the g Factor
2. The Math Behind PCA
3. Back to the g factor
4. Scores and Dimensionality Reduction

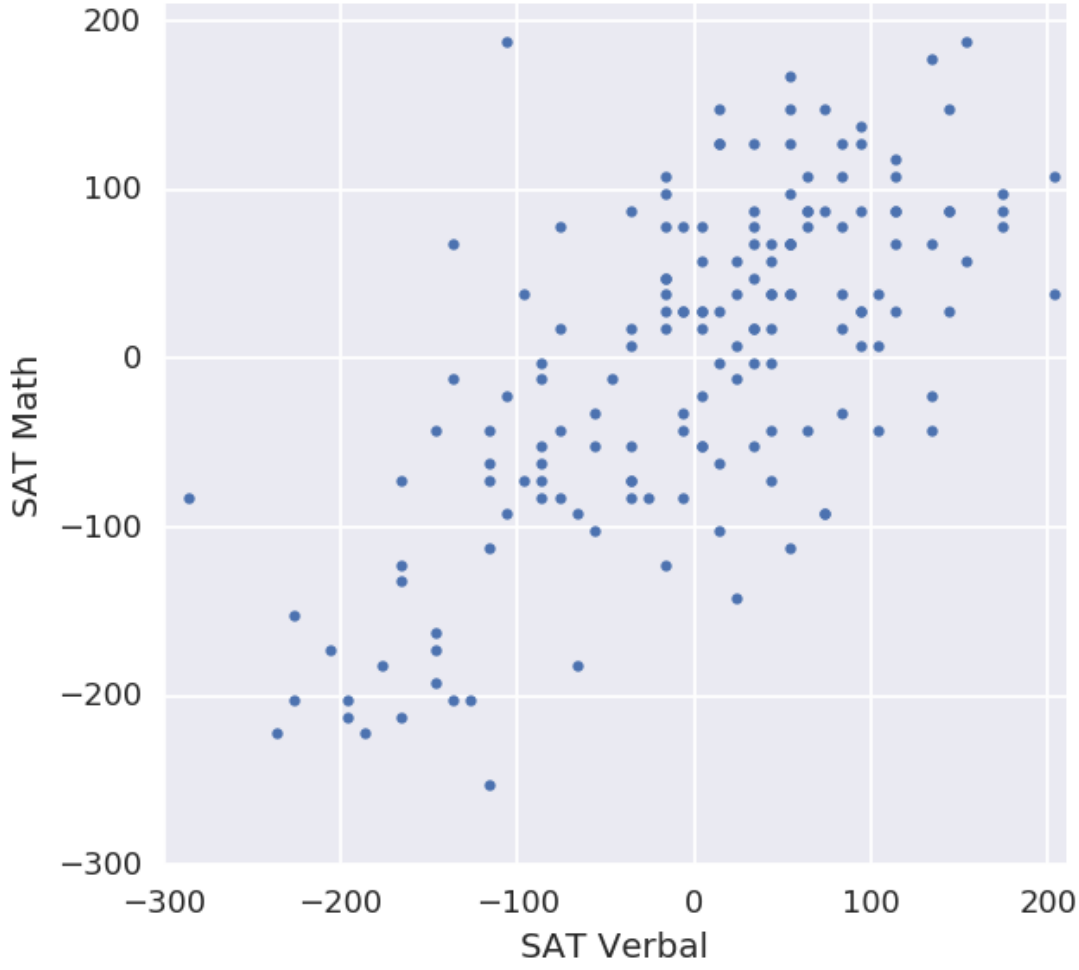
Where does IQ come from?

- In 1904, Charles Spearman noted that children's performance across unrelated school subjects, like Classics, Math, and Music, were positively correlated.
- He hypothesized that all cognitive ability could be traced to a single "general intelligence" factor, which he called the g factor.
- Later, IQ tests were designed to try to measure this g factor. It attempts to quantify intelligence along a single dimension.

Identifying the g factor in data

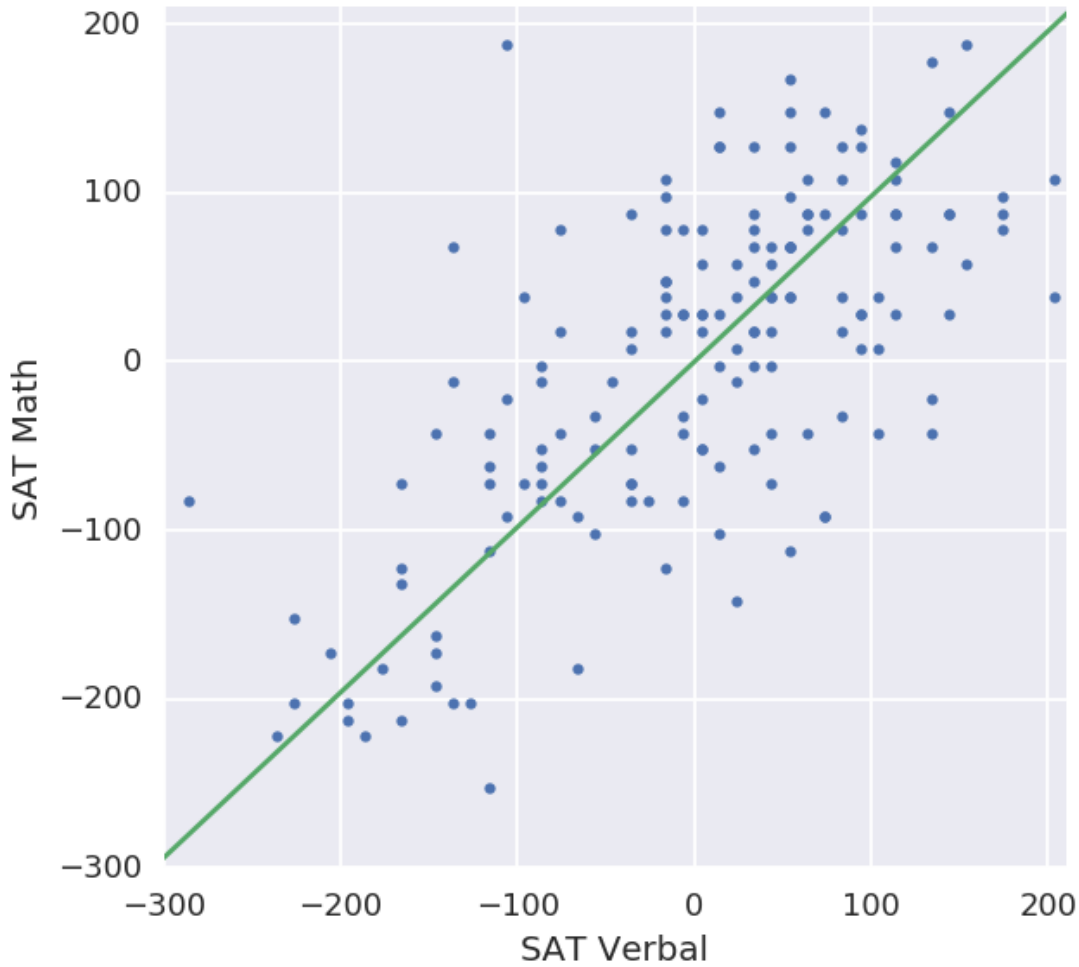


Identifying the g factor in data



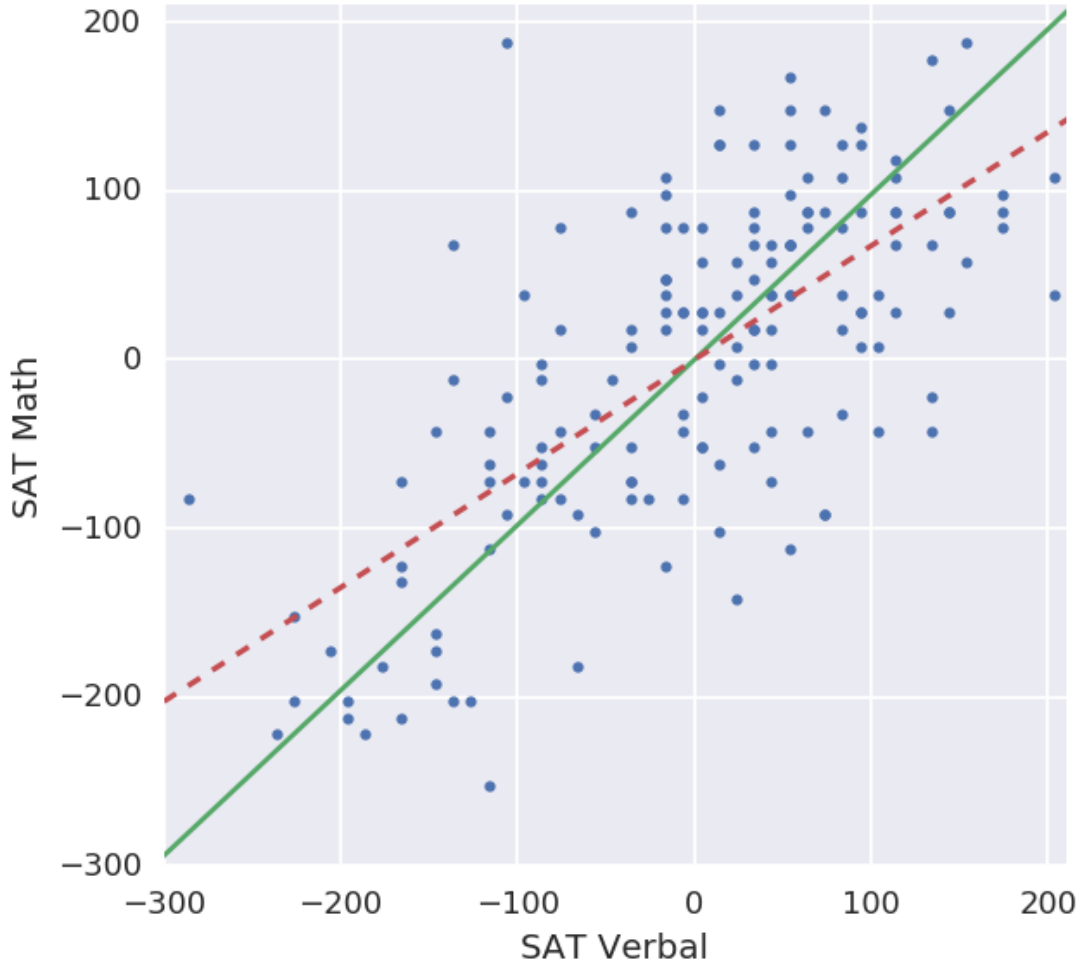
It is typical to first **center** the variables so that they have mean 0.

Identifying the g factor in data



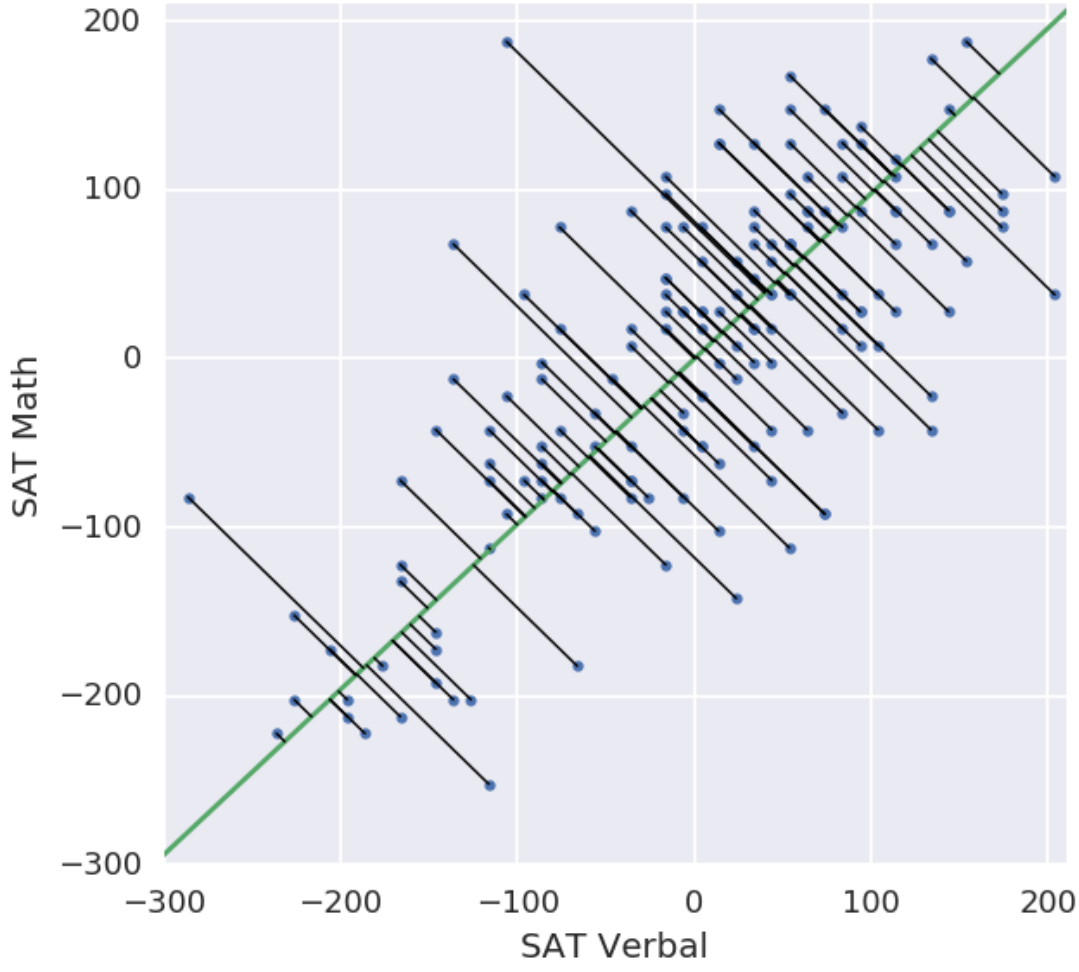
The g factor is a combination of math and verbal skills. It is the direction in the data of greatest variability. This is called the **first principal component**.

Identifying the g factor in data



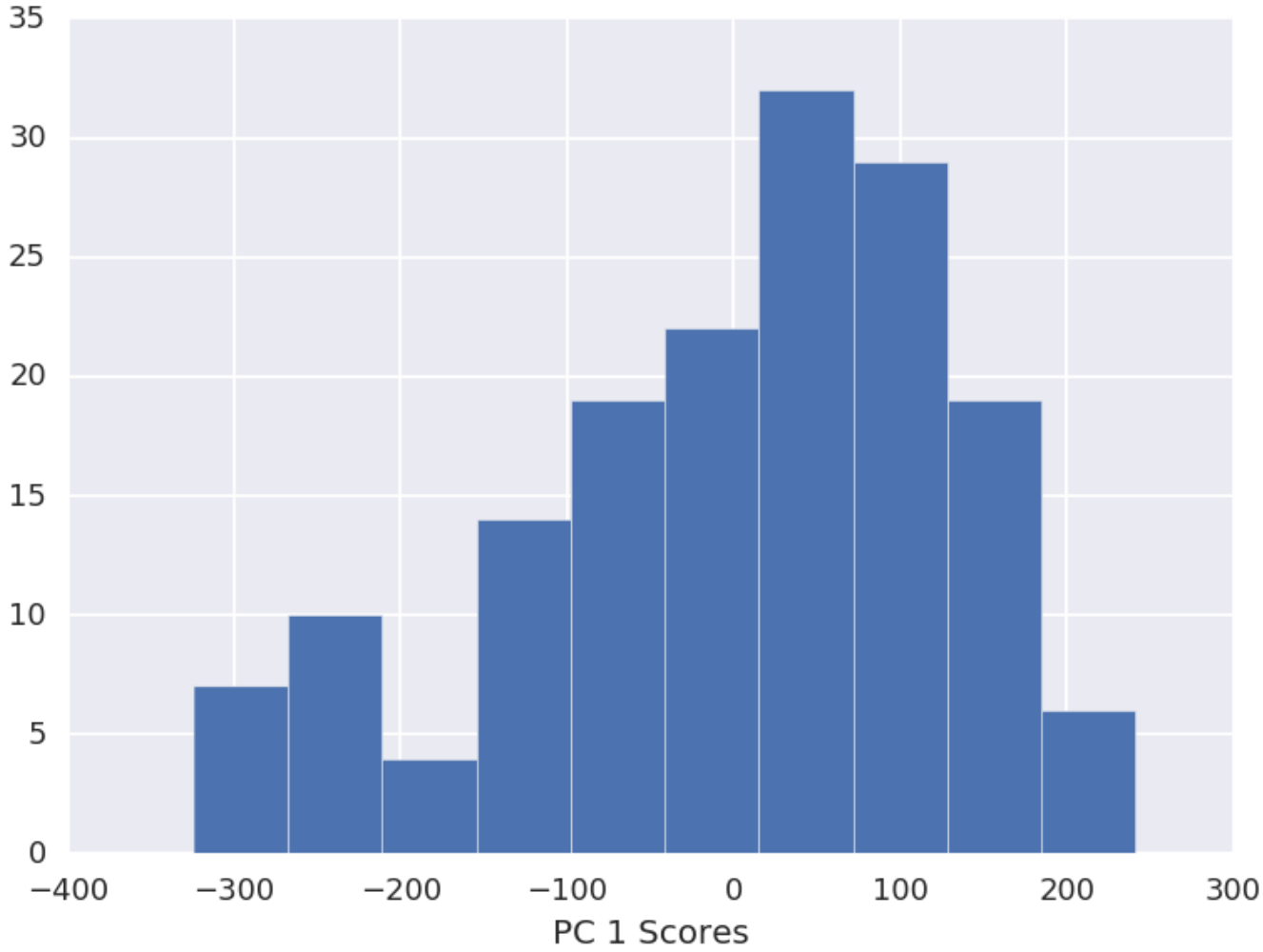
The direction of greatest variability is *different* from the linear regression line.

Identifying the g factor in data



We can **project** the points onto this direction to obtain **scores**.

Identifying the g factor in data



We've now reduced our two-dimensional data to just a single dimension.

The Math Behind PCA

$$X = \begin{pmatrix} \text{——} \mathbf{x}_1 \text{——} \\ \text{——} \mathbf{x}_2 \text{——} \\ \vdots \\ \text{——} \mathbf{x}_n \text{——} \end{pmatrix}$$

Assume the variables have been **centered** to all have mean 0.

We want to find the direction $\mathbf{v}$ that maximizes variability in the scores $s_i = \mathbf{x}_i \cdot \mathbf{v}$.

In other words, we want $\mathbf{v}$ that maximizes:

$$\sum_{i=1}^{n} s_i^2 = \sum_{i=1}^{n} (\mathbf{x}_i \cdot \mathbf{v})^2 = \mathbf{v}^T X^T X \mathbf{v}.$$

Problem: The scores can be made arbitrarily large by choosing $\mathbf{v}$ to be as large as possible.

The Math Behind PCA

$$X = \begin{pmatrix} \text{——} \mathbf{x}_1 \text{——} \\ \text{——} \mathbf{x}_2 \text{——} \\ \vdots \\ \text{——} \mathbf{x}_n \text{——} \end{pmatrix}$$

So we also require that $\mathbf{v}$ have length 1, i.e., $\sum_{i=1}^{n} v_i^2 = 1$.

The optimization problem is:

$$\underset{\mathbf{v}}{\text{maximize}} \quad \mathbf{v}^T X^T X \mathbf{v} \qquad\qquad \text{subject to} \quad \mathbf{v}^T \mathbf{v} = 1$$

This is a job for Lagrange multipliers!

$$\mathcal{L}(\mathbf{v}, \lambda) = \mathbf{v}^T X^T X \mathbf{v} + \lambda(1 - \mathbf{v}^T \mathbf{v}).$$

Now take the derivative with respect to $\mathbf{v}$ and set it equal to zero:

$$\nabla_{\mathbf{v}} \mathcal{L} = 2X^T X \mathbf{v} - 2\lambda \mathbf{v} = 0$$

We see that $\mathbf{v}$ must satisfy $X^T X \mathbf{v} = \lambda \mathbf{v}$.

The Math Behind PCA

$$X = \begin{pmatrix} \text{——} \mathbf{x}_1 \text{——} \\ \text{——} \mathbf{x}_2 \text{——} \\ \vdots \\ \text{——} \mathbf{x}_n \text{——} \end{pmatrix}$$

The first principal component $\mathbf{v}$ satisfies:

$$X^T X \mathbf{v} = \lambda \mathbf{v}$$

for some λ. In other words, it is an eigenvector of $X^T X$.

But which one? Let's plug this back into the objective we are trying to maximize:

$$\mathbf{v}^T X^T X \mathbf{v} = \mathbf{v}^T (\lambda \mathbf{v}) = \lambda (\mathbf{v}^T \mathbf{v}) = \lambda.$$

So $\mathbf{v}$ is the eigenvector with the largest eigenvalue!

The Math Behind PCA

So the 1st principal component $\mathbf{v}_1$ is the eigenvector of $\Sigma = X^T X$ with the largest eigenvalue. The 2nd principal component $\mathbf{v}_2$ is the one with the second largest. And so on.

PCA allows you to take your p-dimensional data and reduce it to the k dimensions of greatest variability.

Will we always be able to find p principal components?

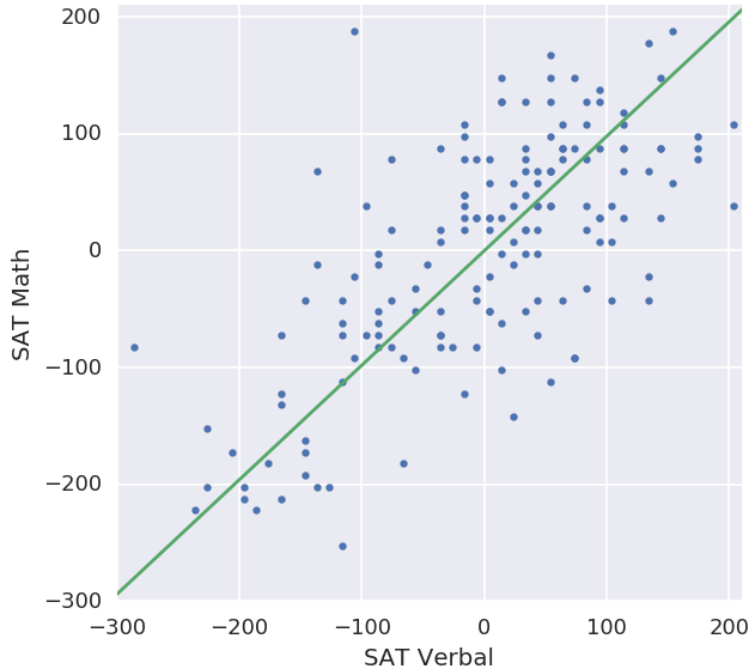
The answer is **yes**, by the **Spectral Theorem of Linear Algebra**:

Theorem

If A is a symmetric matrix, then there exists an orthonormal basis of eigenvectors with real eigenvalues.

The matrix $X^T X$ is symmetric!

PCA and the SAT Data



The 1st principal component is $\mathbf{v}_1 = \begin{pmatrix} .714 \\ .700 \end{pmatrix}$.

This means the score of any observation $\mathbf{x}_i$ is:

$$s_{i1} = \mathbf{x}_i \cdot \mathbf{v}_1 = .714x_{i1} + .700x_{i2}.$$

In other words, we have constructed a new variable

$$g\text{-factor} = .714 \cdot (\text{SAT Verbal}) + .700 \cdot (\text{SAT Math}).$$

Spearman's Data

Spearman studied the performance of school children in 6 subjects and obtained the following correlation matrix:

	Classics	French	English	Math	Pitch	Music
Classics	1.00	.83	.78	.70	.66	.63
French	.83	1.00	.67	.67	.65	.57
English	.78	.67	1.00	.64	.54	.51
Math	.70	.67	.64	1.00	.45	.51
Pitch	.66	.65	.54	.45	1.00	.40
Music	.63	.57	.51	.51	.40	1.00

.

Note that this is not the original data. Why is this data sufficient to calculate the principal components?

What is the dimension of the first principal component $\mathbf{v}_1$?

Spearman found the first principal component to be

$$\mathbf{v}_1 = \begin{pmatrix} .958 & .882 & .803 & .750 & .673 & .646 \end{pmatrix}^T .$$

What does this tell you about his g-factor?

1 Prelude: Intelligence and the g Factor

2 The Math Behind PCA

3 Back to the g factor

4 Scores and Dimensionality Reduction

Dimensionality Reduction

Usually the goal of PCA is to reduce the dimensionality of our data. Instead of carrying around p variables, we might only want to carry around 2 or 3 variables (e.g., so that we can visualize our data).

So we usually don't care so much about the loadings $\mathbf{v}$ of a principal component. We just want the scores of each observation $\mathbf{x}_i$ along the direction $\mathbf{v}$:

$$s_i = \mathbf{x}_i \cdot \mathbf{v}.$$

In matrix notation, the vector of scores along the jth principal component is:

$$\mathbf{s}_j = X\mathbf{v}_j.$$

If we take the top 2 principal components, we can actually make a scatterplot of the scores $\mathbf{s}_1$ and $\mathbf{s}_2$.

Another Way to Compute Scores

However, calculating the scores by $\mathbf{s}_j = X\mathbf{v}_j$ is not efficient because it first requires finding $\mathbf{v}_j$.

Each $\mathbf{v}_j$ is an eigenvector of X^TX, a $p \times p$ matrix. If p is very large, this may be infeasible.

To find another way to calculate the scores, let's hit both sides of this equation on the left by X^T.

$$X^T\mathbf{s}_j = X^TX\mathbf{v}_j = \lambda_j\mathbf{v}_j.$$

Now let's hit both sides of this equation by X.

$$XX^T\mathbf{s}_j = \lambda_jX\mathbf{v}_j = \lambda_j\mathbf{s}_j.$$

So the vector of scores $\mathbf{s}_j$ is an eigenvector of XX^T, which is an $n \times n$ matrix. When $n \ll p$, this is an easier to calculate!