Informative Projections

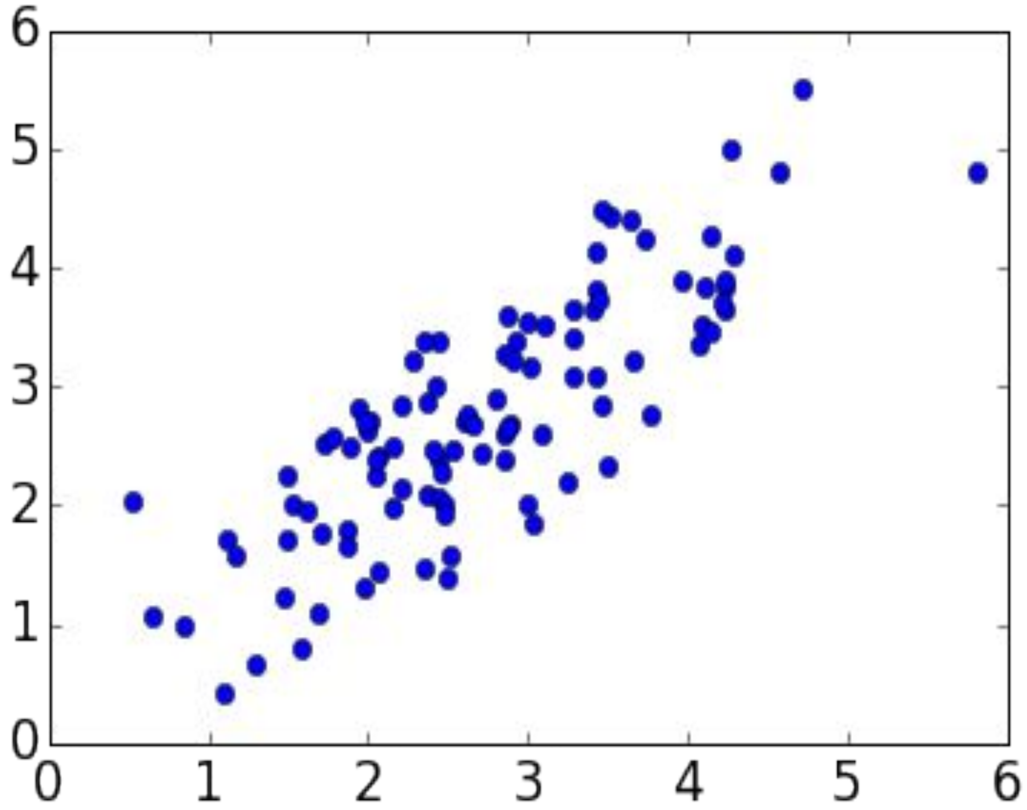
MGTF 495

Class Outline

- Representation Learning
  - k-means
  - EM
  - Agglomerative hierarchical clustering
  - Hands-On
- Informative Projections
  - PCA
  - SVD
  - Latent semantic indexing (LSI)
  - Hands-On

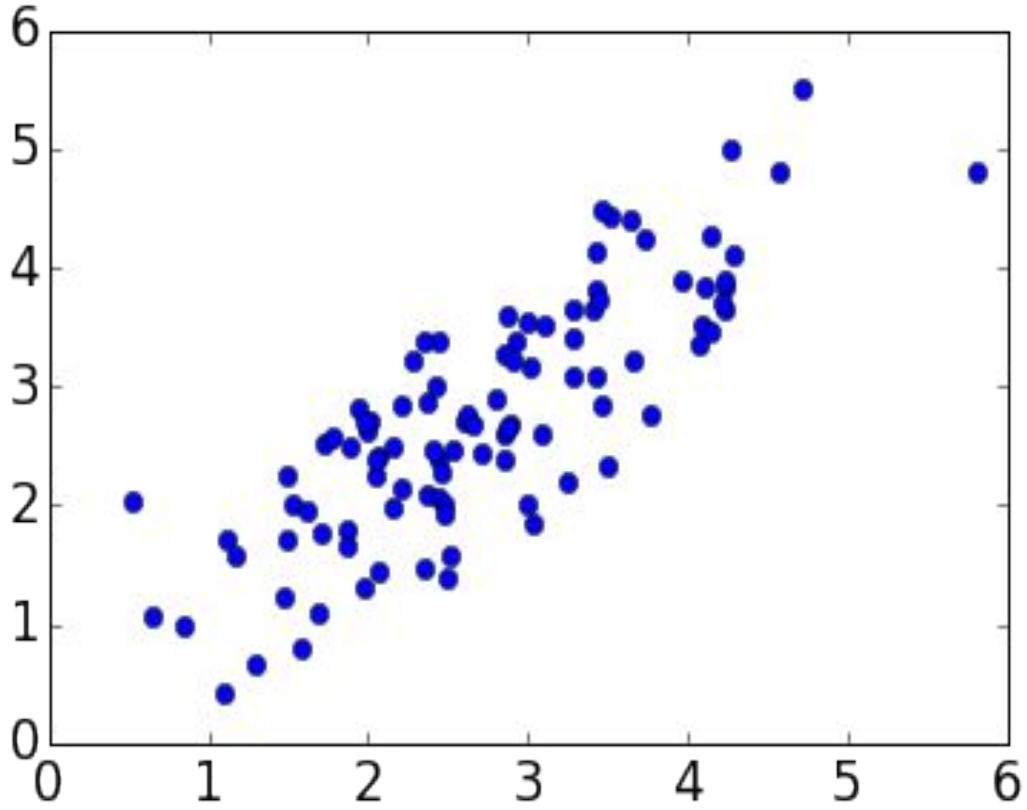
Informative projection

Suppose we wanted just one feature for the following data.



Informative projection

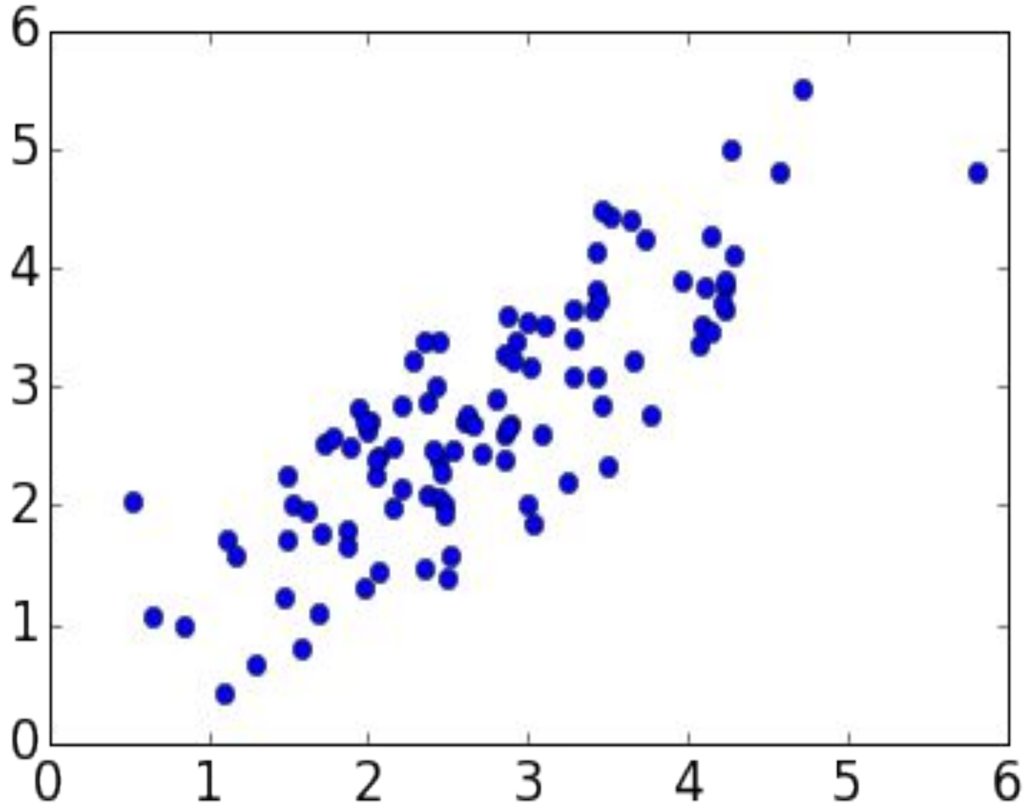
Suppose we wanted just one feature for the following data.



- We could pick a single coordinate.

Informative projection

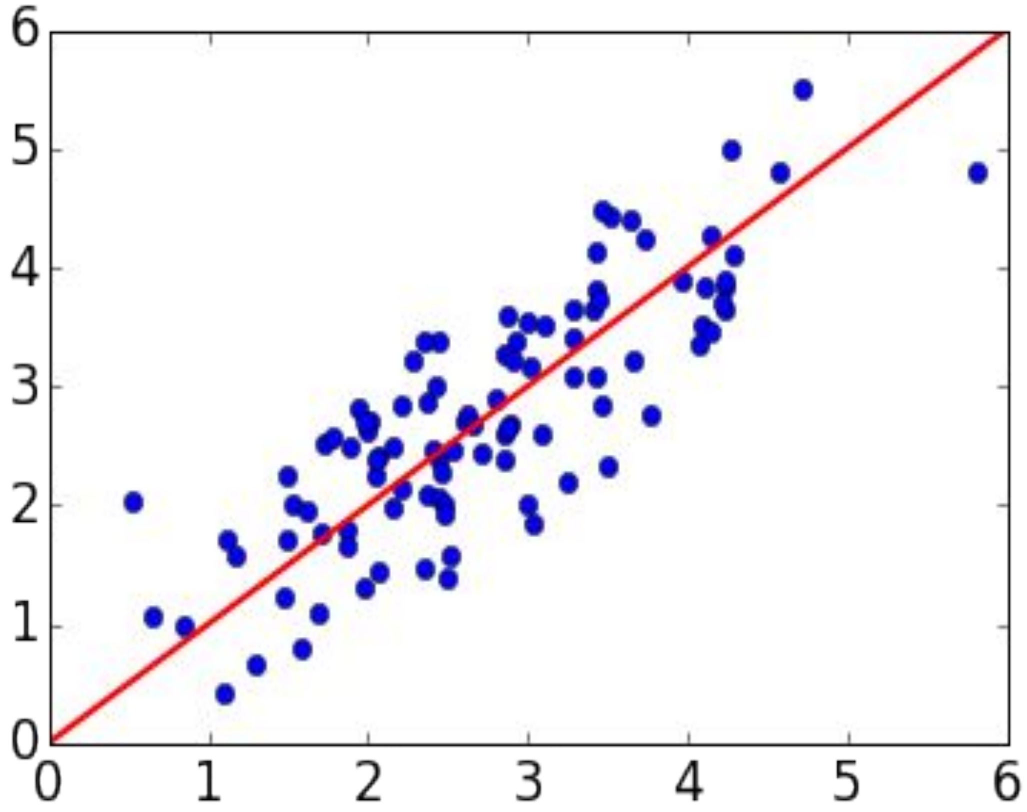
Suppose we wanted just one feature for the following data.



- We could pick a single coordinate.
- Or an arbitrary direction.

Informative projection

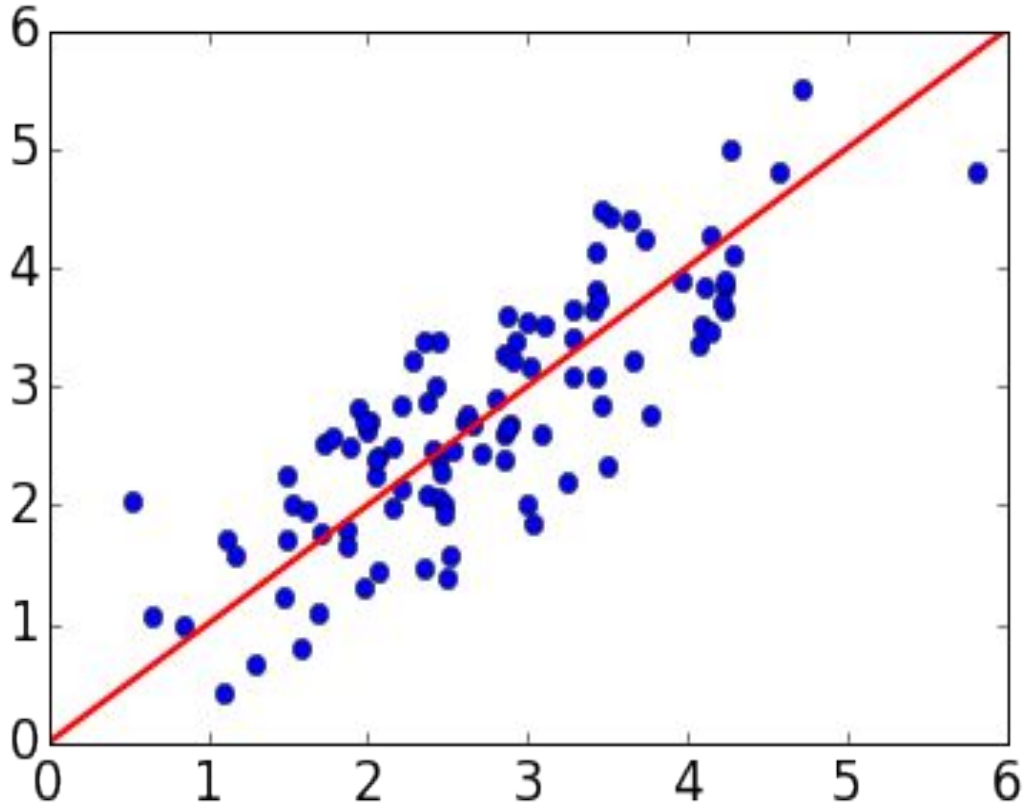
Suppose we wanted just one feature for the following data.



- We could pick a single coordinate.
- Or an arbitrary direction.

Informative projection

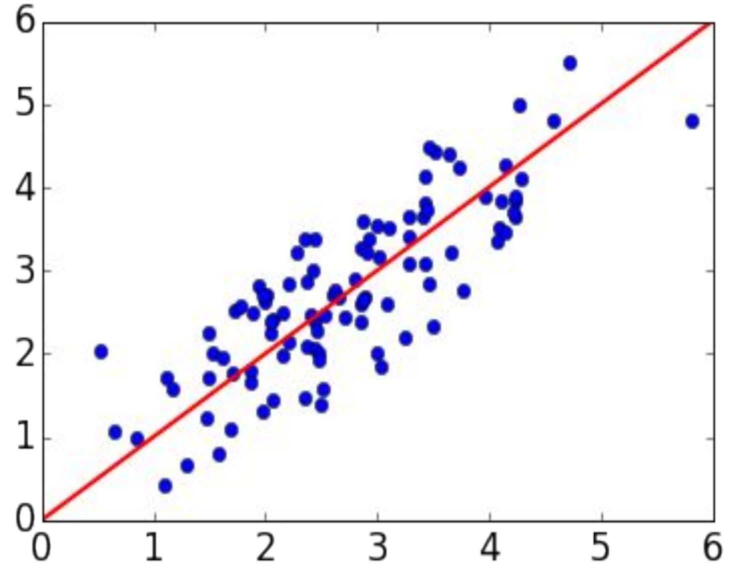
Suppose we wanted just one feature for the following data.



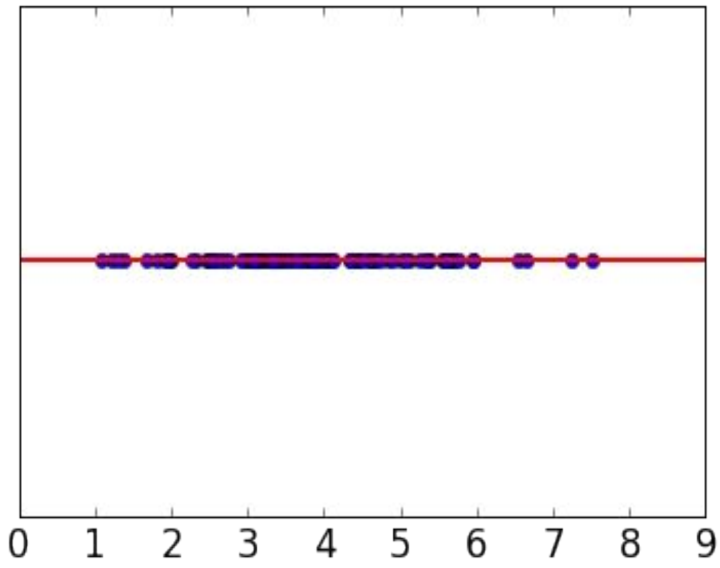
- We could pick a single coordinate.
- Or an arbitrary direction.

A good choice: the **direction of maximum variance**.

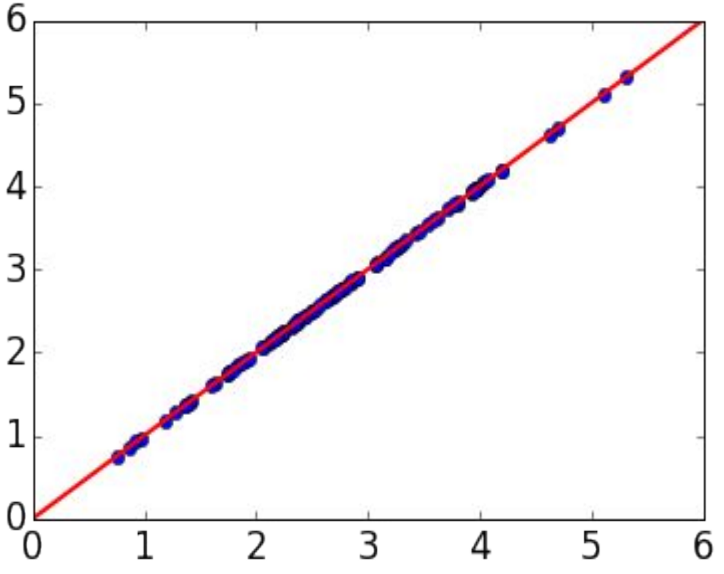
Two types of projection



Projection onto R^1:

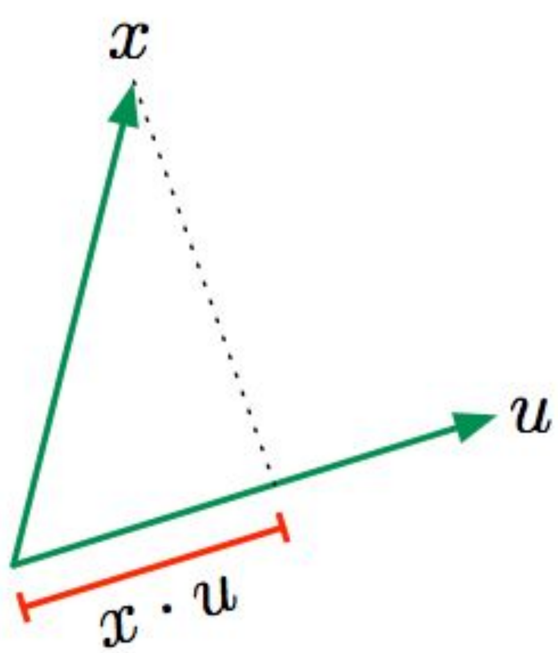


Projection onto a 1-d line in R^2



Projection: formally

What is the projection of $x \in \mathbb{R}^p$ onto direction $u \in \mathbb{R}^p$ (where $\|u\| = 1$)?



As a one-dimensional value:

$$x \cdot u = u \cdot x = u^T x = \sum_{i=1}^{p} u_i x_i.$$

As a p-dimensional vector:

$$(x \cdot u)u = uu^T x$$

"Move $x \cdot u$ units in direction u"

Quick quiz

What is the projection of $x = \binom{2}{3}$ onto the following directions?

Give, first, a one-dimensional value and, then, a two-dimensional vector.

1. The coordinate direction e_1?

Quick quiz

What is the projection of $x = \binom{2}{3}$ onto the following directions?

Give, first, a one-dimensional value and, then, a two-dimensional vector.

1. The coordinate direction e_1? $2, \binom{2}{0}$

Quick quiz

What is the projection of $x = \binom{2}{3}$ onto the following directions?

Give, first, a one-dimensional value and, then, a two-dimensional vector.

1. The coordinate direction e_1? $2, \begin{pmatrix} 2 \\ 0 \end{pmatrix}$
2. The direction of $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$?

Quick quiz

What is the projection of $x = \binom{2}{3}$ onto the following directions?

Give, first, a one-dimensional value and, then, a two-dimensional vector.

1. The coordinate direction e_1? $2, \binom{2}{0}$
2. The direction of $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$? $-1/\sqrt{2}, \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$

Class Outline

- Representation Learning
  - k-means
  - EM
  - Agglomerative hierarchical clustering
  - Hands-On
- Informative Projections
  - PCA
  - SVD
  - Latent semantic indexing (LSI)
  - Hands-On

Projection onto multiple directions

Want to project $x \in \mathbb{R}^p$ into the k-dimensional subspace defined by vectors $u_1, \ldots, u_k \in \mathbb{R}^p$.

This is easiest when the u_i's are **orthonormal**:

- They each have length one.
- They are at right angles to each other: $u_i \cdot u_j = 0$ whenever $i \neq j$

Then the projection, as a k-dimensional vector, is

$$(x \cdot u_1, x \cdot u_2, \ldots, x \cdot u_k) = \underbrace{\begin{pmatrix} \longleftarrow & u_1 & \longrightarrow \\ \longleftarrow & u_2 & \longrightarrow \\ & \vdots & \\ \longleftarrow & u_k & \longrightarrow \end{pmatrix}}_{\text{call this } U^T} \begin{pmatrix} \uparrow \\ x \\ \downarrow \end{pmatrix}$$

As a p-dimensional vector, the projection is

$$(x \cdot u_1)u_1 + (x \cdot u_2)u_2 + \cdots + (x \cdot u_k)u_k = UU^T x.$$

Projection onto multiple directions: example

Suppose data are in $\mathbb{R}^4$ and we want to project onto the first two coordinates.

Take vectors $\quad u_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad u_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \quad$ (notice: orthonormal)

Then write $\quad U^T = \begin{pmatrix} \longleftarrow & u_1 & \longrightarrow \\ \longleftarrow & u_2 & \longrightarrow \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$

The projection of $x \in \mathbb{R}^4$, as a 2-d vector, is

$$U^T x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

The projection of x as a 4-d vector is

$$UU^T x = \begin{pmatrix} x_1 \\ x_2 \\ 0 \\ 0 \end{pmatrix}$$

But we'll generally project along non-coordinate directions.

The best single direction

Suppose we need to map our data $x \in \mathbb{R}^p$ into just **one** dimension:

$$x \mapsto u \cdot x \quad \text{for some unit direction } u \in \mathbb{R}^p$$

What is the direction u of maximum variance?

Theorem: Let Σ be the $p \times p$ covariance matrix of X. The variance of X in direction u is given by $u^T \Sigma u$.

- Suppose the mean of X is $\mu \in \mathbb{R}^p$. The projection $u^T X$ has mean

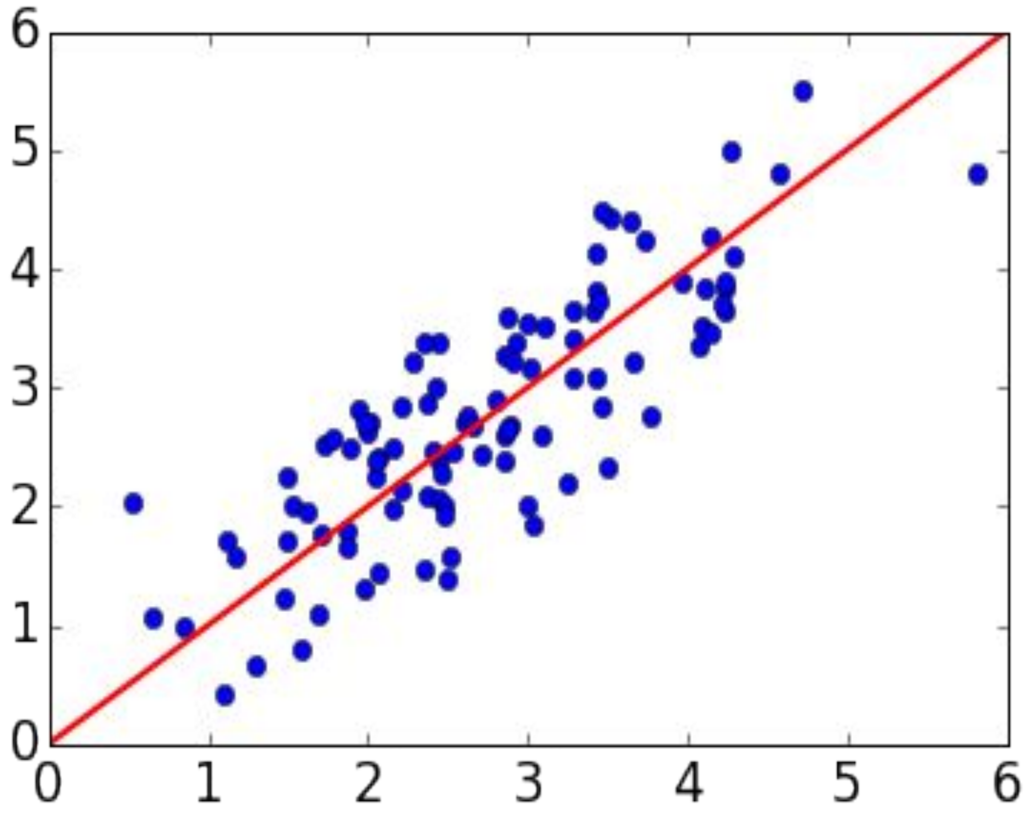
$$\mathbb{E}(u^T X) = u^T \mathbb{E} X = u^T \mu.$$

- The variance of $u^T X$ is

$$\begin{aligned} \text{var}(u^T X) &= \mathbb{E}(u^T X - u^T \mu)^2 = \mathbb{E}(u^T (X - \mu)(X - \mu)^T u) \\ &= u^T \mathbb{E}(X - \mu)(X - \mu)^T u = u^T \Sigma u. \end{aligned}$$

Another theorem: $u^T \Sigma u$ is maximized by setting u to the first **eigenvector** of Σ. The maximum value is the corresponding **eigenvalue**.

Best single direction: example



This direction is the **first eigenvector** of the 2×2 covariance matrix of the data.

The best *k*-dimensional projection

Let Σ be the $p \times p$ covariance matrix of X. Its **eigen-decomposition** can be computed in $O(p^3)$ time and consists of:

- real **eigenvalues** $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$
- corresponding **eigenvectors** $u_1, \ldots, u_p \in \mathbb{R}^p$ that are orthonormal: that is, each u_i has unit length and $u_i \cdot u_j = 0$ whenever $i \neq j$.

Theorem: Suppose we want to map data $X \in \mathbb{R}^p$ to just k dimensions, while capturing as much of the variance of X as possible. The best choice of projection is:

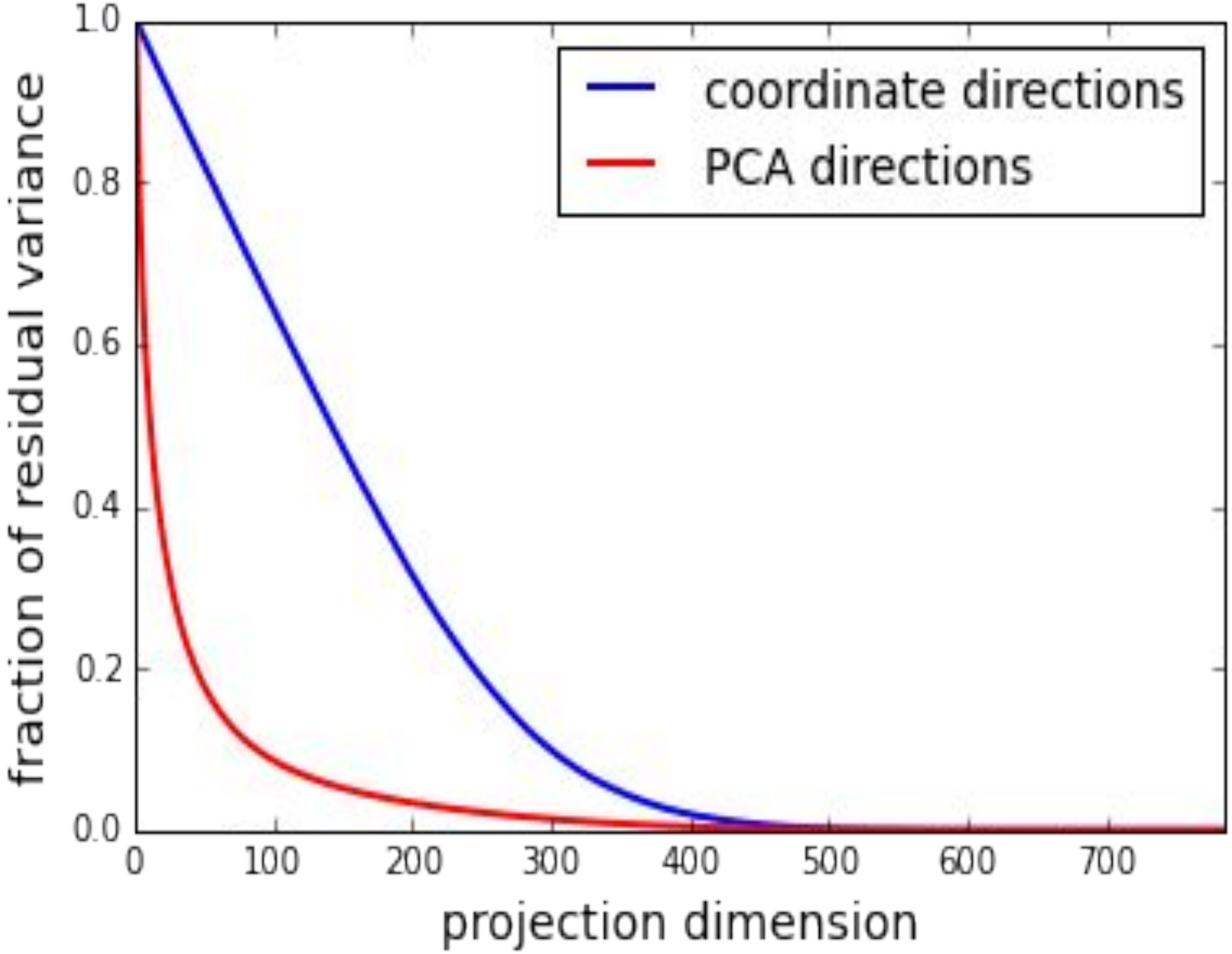
$$x \mapsto (u_1 \cdot x, u_2 \cdot x, \ldots, u_k \cdot x),$$

where u_i are the eigenvectors described above.

Projecting the data in this way is **principal component analysis** (PCA).

Example: MNIST

Contrast coordinate projections with PCA:



MNIST: image reconstruction

Reconstruct this original image from its PCA projection to k dimensions.

$k = 200$

$k = 150$

$k = 100$

$k = 50$

Q: What are these reconstructions exactly?

A: Image X is reconstructed as UU^Tx, where U is a $p \times k$ matrix whose columns are the top k eigenvectors of Σ.

Review: eigenvalues and eigenvectors

There are several steps to understanding these.

1. Any matrix M defines a function (or **transformation**) $x \mapsto Mx$.

2. If M is a $p \times q$ matrix, then this transformation maps vector $x \in \mathbb{R}^q$ to vector $Mx \in \mathbb{R}^p$.

3. We call it a **linear transformation** because $M(x + x') = Mx + Mx'$.

4. We'd like to understand the nature of these transformations. The easiest case is when M is **diagonal**:

$$\underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 10 \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} 2x_1 \\ -x_2 \\ 10x_3 \end{pmatrix}}_{Mx}$$

5. What about more general matrices that are symmetric but not necessarily diagonal? They also just scale coordinates separately, but in a **different coordinate system**.

Review: eigenvalues and eigenvectors

Let M be a $p \times p$ matrix.
We say $u \in \mathbb{R}^p$ is an **eigenvector** if M maps u onto the same direction, that is,

$$Mu = \lambda u$$

for some scaling constant λ. This λ is the **eigenvalue** associated with u.

Question: What are the eigenvectors and eigenvalues of:

$$M = \begin{pmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 10 \end{pmatrix} \; ?$$

Answer: Eigenvectors e_1, e_2, e_3, with corresponding eigenvalues $2, -1, 10$.

Notice that these eigenvectors form an orthonormal basis.

Eigenvectors of a real symmetric matrix

Theorem. Let M be any real symmetric $p \times p$ matrix. Then M has

- p eigenvalues $\lambda_1, \ldots, \lambda_p$
- corresponding eigenvectors $u_1, \ldots, u_p \in R^p$ that are orthonormal

We can think of $u_1, \ldots, u_p$ as being the axes of the natural coordinate system for understanding M.

Example: consider the matrix

$$M = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$$

It has eigenvectors

$$u_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad u_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

- Are these eigenvectors orthonormal?
- What are the corresponding eigenvalues? 2, 4

Spectral decomposition

Theorem. Let M be any real symmetric $p \times p$ matrix. Then M has

- p eigenvalues $\lambda_1, \ldots, \lambda_p$
- corresponding eigenvectors $u_1, \ldots, u_p \in R^p$ that are orthonormal

Spectral decomposition: Here is another way to write M:

$$M = \underbrace{\begin{pmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \cdots & u_p \\ \downarrow & \downarrow & & \downarrow \end{pmatrix}}_{U\text{: columns are eigenvectors}} \underbrace{\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}}_{\Lambda\text{: eigenvalues on diagonal}} \underbrace{\begin{pmatrix} \longleftarrow & u_1 & \longrightarrow \\ \longleftarrow & u_2 & \longrightarrow \\ & \vdots & \\ \longleftarrow & u_p & \longrightarrow \end{pmatrix}}_{U^T}$$

Thus $Mx = U\Lambda U^T x$, which can be interpreted as follows:

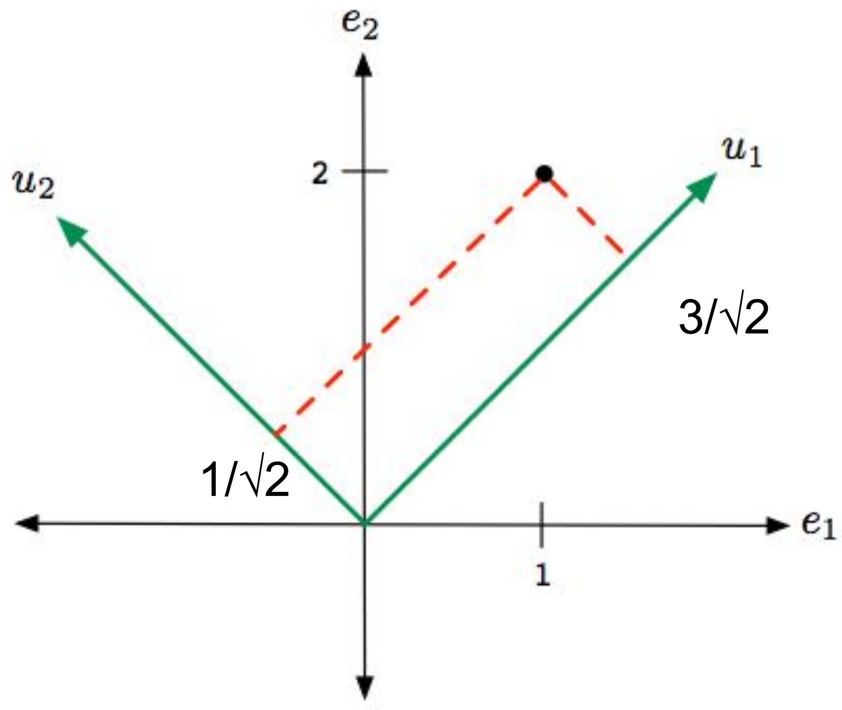
- U^T rewrites x in the $\{u_i\}$ coordinate system
- Λ is a simple coordinate scaling in that basis
- U then sends the scaled vector back into the usual coordinate basis

Spectral

Apply spectral decomposition to the matrix M we saw earlier:

$$M = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} = \underbrace{\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}}_{U} \underbrace{\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}}_{\Lambda} \underbrace{\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}}_{U^T}$$

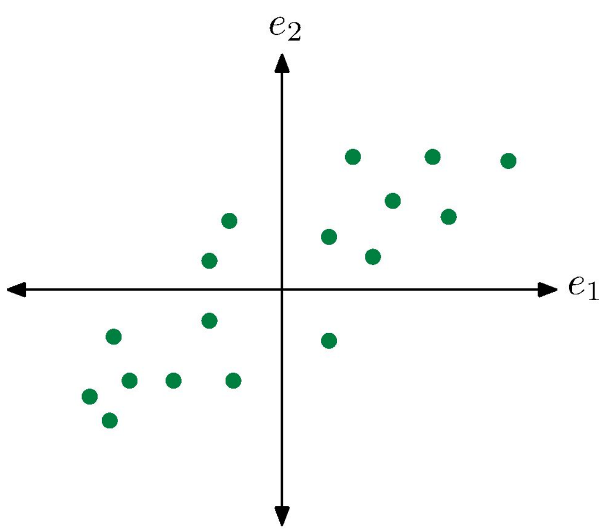
$$\begin{aligned} M\begin{pmatrix} 1 \\ 2 \end{pmatrix} &= ??? \\ &= U \;\; \Lambda \;\; U^T \;\; \begin{pmatrix} 1 \\ 2 \end{pmatrix} \\ &= U \;\; \Lambda \;\; \frac{1}{\sqrt{2}}\begin{pmatrix} 3 \\ 1 \end{pmatrix} \\ &= U \;\; \frac{1}{\sqrt{2}}\begin{pmatrix} 12 \\ 2 \end{pmatrix} \\ &= \begin{pmatrix} 5 \\ 7 \end{pmatrix} \end{aligned}$$



Principal component analysis: recap

Consider data vectors $X \in \mathbb{R}^p$

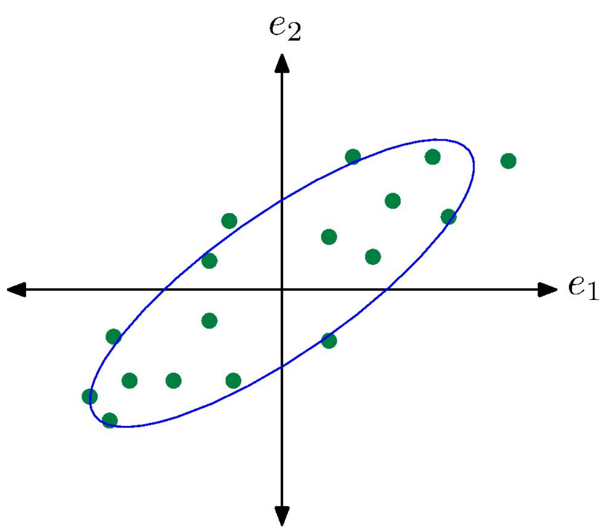
- The covariance matrix Σ is a $p \times p$ symmetric matrix.
- Get eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, eigenvectors $u_1, \ldots, u_p$.
- $u_1, \ldots, u_p$ is an alternative basis in which to represent the data.
- The variance of X in direction u_i is λ_i.
- To project to k dimensions while losing as little as possible of the overall variance, use $x \mapsto (x \cdot u_1, \ldots, x \cdot u_k)$.



Principal component analysis: recap

Consider data vectors $X \in \mathbb{R}^p$

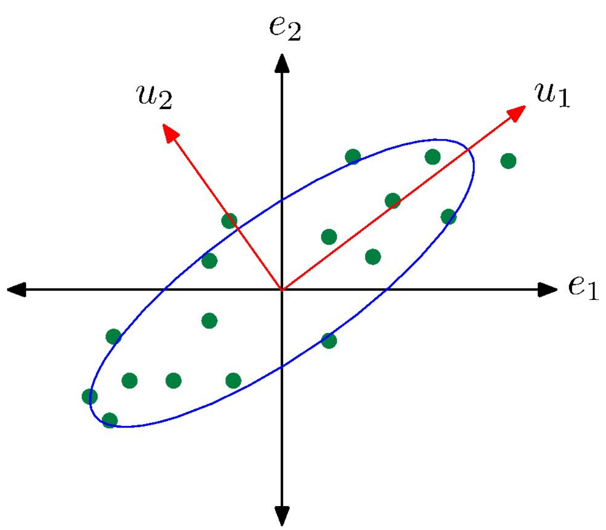
- The covariance matrix Σ is a $p \times p$ symmetric matrix.
- Get eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, eigenvectors $u_1, \ldots, u_p$.
- $u_1, \ldots, u_p$ is an alternative basis in which to represent the data.
- The variance of X in direction u_i is λ_i.
- To project to k dimensions while losing as little as possible of the overall variance, use $x \mapsto (x \cdot u_1, \ldots, x \cdot u_k)$.



Principal component analysis: recap

Consider data vectors $X \in \mathbb{R}^p$

- The covariance matrix Σ is a $p \times p$ symmetric matrix.
- Get eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, eigenvectors $u_1, \ldots, u_p$.
- $u_1, \ldots, u_p$ is an alternative basis in which to represent the data.
- The variance of X in direction u_i is λ_i.
- To project to k dimensions while losing as little as possible of the overall variance, use $x \mapsto (x \cdot u_1, \ldots, x \cdot u_k)$.



Example: personality assessment

What are the dimensions along which personalities differ?

- *Lexical hypothesis*: most important personality characteristics have become encoded in natural language.
- Allport and Odbert (1936): sat down with the English dictionary and extracted all terms that could be used to distinguish one person's behavior from another's. Roughly 18000 words, of which 4500 could be described as personality traits.
- Step: group these words into (approximate) synonyms. This is done by manual clustering. e.g. Norman (1967):

Spirit	Jolly, merry, witty, lively, peppy
Talkativeness	Talkative, articulate, verbose, gossipy
Sociability	Companionable, social, outgoing
Spontaneity	Impulsive, carefree, playful, zany
Boisterousness	Mischievous, rowdy, loud, prankish
Adventure	Brave, venturous, fearless, reckless
Energy	Active, assertive, dominant, energetic
Conceit	Boastful, conceited, egotistical
Vanity	Affected, vain, chic, dapper, jaunty
Indiscretion	Nosey, snoopy, indiscreet, meddlesome
Sensuality	Sexy, passionate, sensual, flirtatious

- Data collection: Ask a variety of subjects to what extent each of these words describes them.

Personality assessment: the data

Matrix of data (1 = strongly disagree, 5 = strongly agree)

	shy	merry	tense	boastful	forgiving	quiet
Person 1	4	1	1	2	5	5
Person 2	1	4	4	5	2	1
Person 3	2	4	5	4	2	2
		⋮				

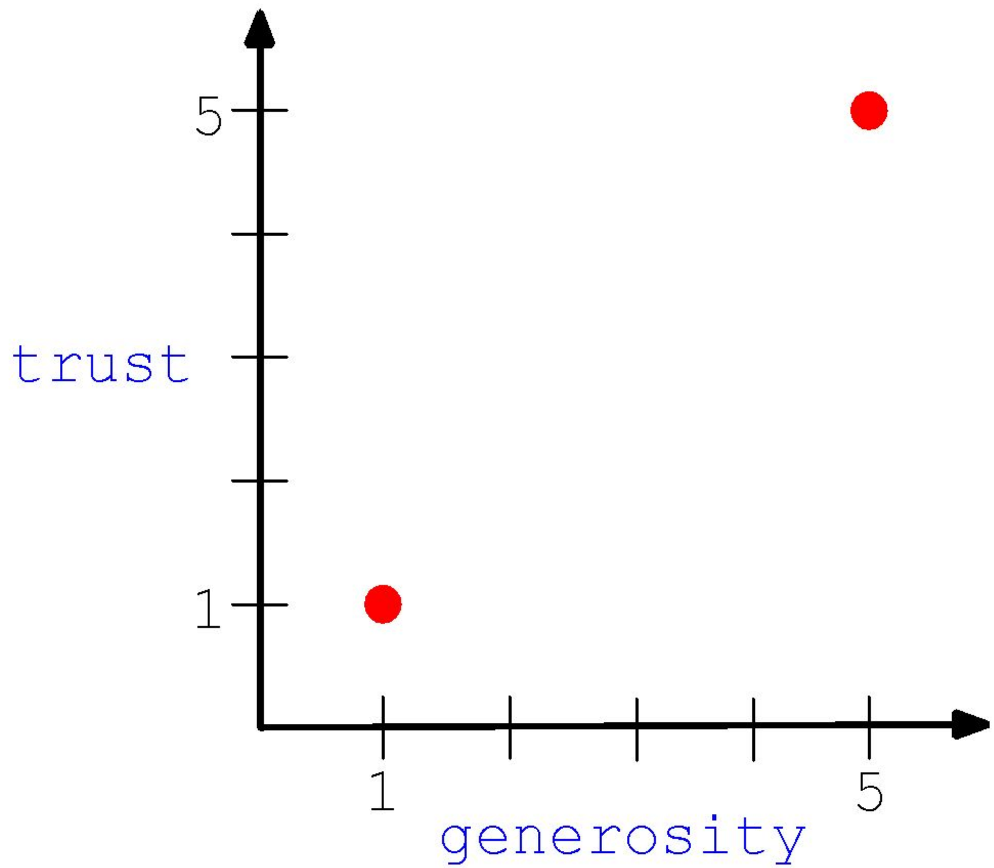
How to extract important directions?

- Treat each column as a data point, find tight clusters
- Treat each row as a data point, apply PCA
- Other ideas: factor analysis, independent component analysis, ...

Many of these yield similar results

What does PCA accomplish?

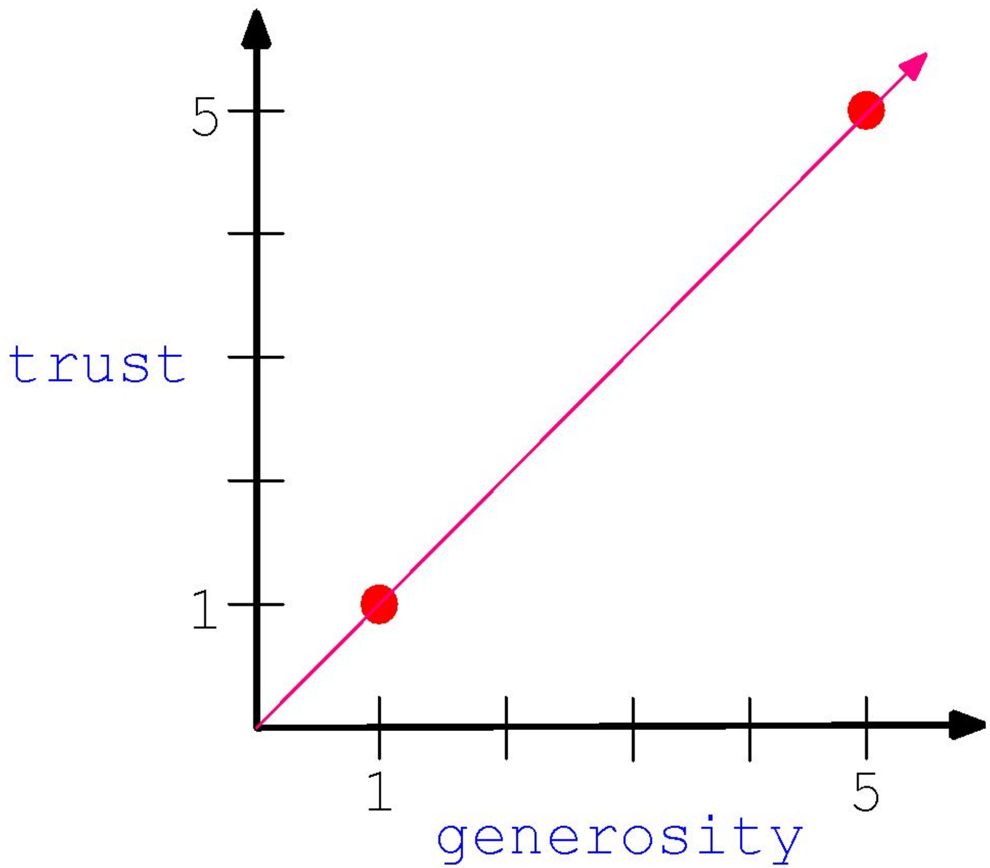
Example: suppose two traits (generosity, trust) are highly correlated, to the point where each person either answers "1" to both or "5" to both.



This single PCA dimension entirely accounts for the two traits.

What does PCA accomplish?

Example: suppose two traits (generosity, trust) are highly correlated, to the point where each person either answers “1” to both or “5” to both.



This single PCA dimension entirely accounts for the two traits.

The “Big Five” taxonomy

Extraversion		Agreeableness		Conscientiousness		Neuroticism		Oppenness/Intellect	
Low	High	Low	High	Low	High	Low	High	Low	High
-.83 Quiet	.85 Talkative	-.52 Fault-finding	.87 Sympathetic	-.58 Careless	.80 Organized	-.39 Stable*	.73 Tense	-.74 Commonplace	.76 Wide interests
-.80 Reserved	.83 Assertive	-.48 Cold	.85 Kind	-.53 Disorderly	.80 Thorough	-.35 Calm*	.72 Anxious	-.73 Narrow interests	.76 Imaginative
-.75 Shy	.82 Active	-.45 Unfriendly	.85 Appreciative	-.50 Frivolous	.78 Planful	-.21 Contented*	.72 Nervous	-.67 Simple	.72 Intelligent
-.71 Silent	.82 Energetic	-.45 Quarrelsome	.84 Affectionate	-.49 Irresponsible	.78 Efficient	.14 Unemotional*	.71 Moody	-.55 Shallow	.73 Original
-.67 Withdrawn	.82 Outgoing	-.45 Hard-hearted	.84 Soft-hearted	-.40 Slipshot	.73 Responsible		.71 Worrying	-.47 Unintelligent	.68 Insightful
-.66 Retiring	.80 Outspoken	-.38 Unkind	.82 Warm	-.39 Undependable	.72 Reliable		.68 Touchy		.64 Curious
	.79 Dominant	-.33 Cruel	.81 Generous	-.37 Forgetful	.70 Dependable		.64 Fearful		.59 Sophisticated
	.73 Forceful	-.31 Stern*	.78 Trusting		.68 Conscientious		.63 High-strung		.59 Artistic
	.73 Enthusiastic	-.28 Thankless	.77 Helpful		.66 Precise		.63 Self-pitying		.59 Clever
	.68 Show-off	-.24 Stingy*	.77 Forgiving		.66 Practical		.60 Temperamental		.58 Inventive
	.68 Sociable		.74 Pleasant		.65 Deliberate		.59 Unstable		.56 Sharp-witted
	.64 Spunky		.73 Good-natured		.46 Painstaking		.58 Self-punishing		.55 Ingenious
	.64 Adventurous		.73 Friendly		.26 Cautious*		.54 Despondent		.45 Witty*
	.62 Noisy		.72 Cooperative				.51 Emotional		.45 Resourceful*
	.58 Bossy		.67 Gentle						.37 Wise
			.66 Unselfish						.33 Logical*
			.56 Praising						.29 Civilized*
			.51 Sensitive						.22 Foresighted*
									.21 Polished*
									.20 Dignified*

Many applications, such as online match-making.