Basic examples of regression, classification and feature extraction

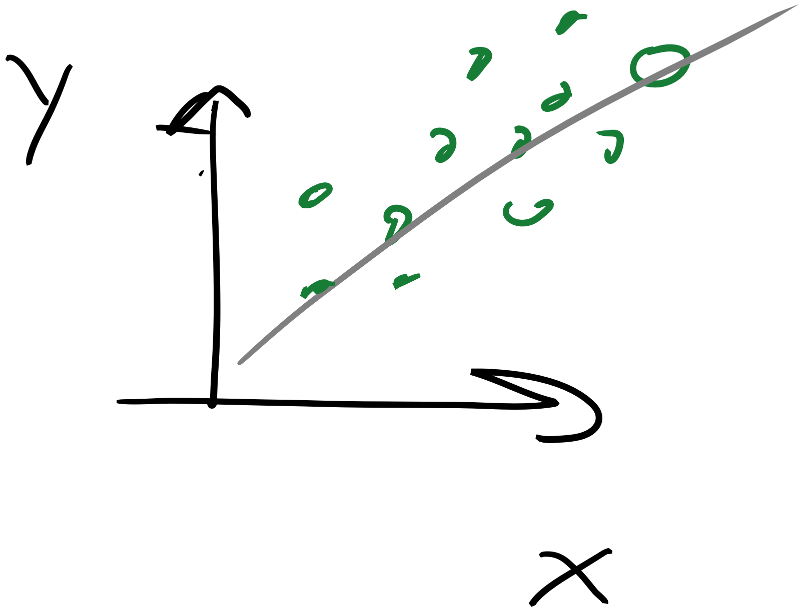
Joachim Mathiesen, Niels Bohr Institute

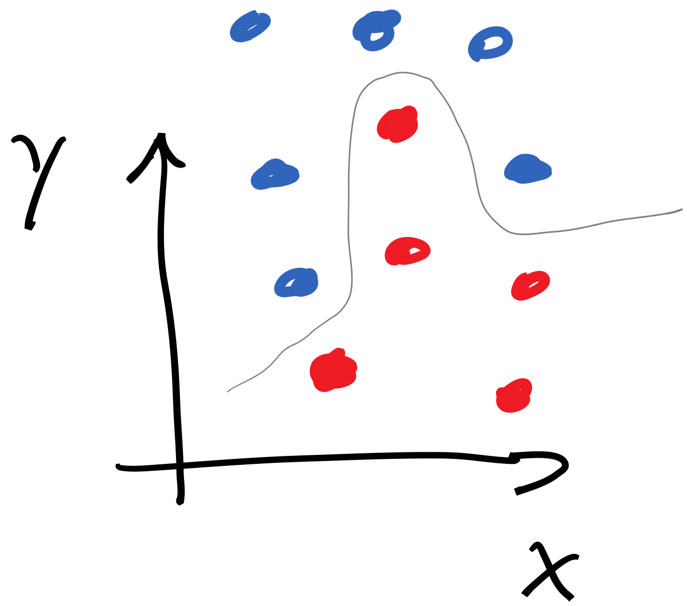
Regression and classification

Classification	Regression
Qualitative variables Gender Sick / not sick Species Brand Property type (apartment or house) …	*Quantitative variables* Value of a house Temperature Income Size Age …

Separate methods are often used for the two problem types. However, many methods can be extended to do both, such as k-nearest neighbors, linear classifiers or regression, support vector machines, kernel estimation, etc.

Regression and Classification

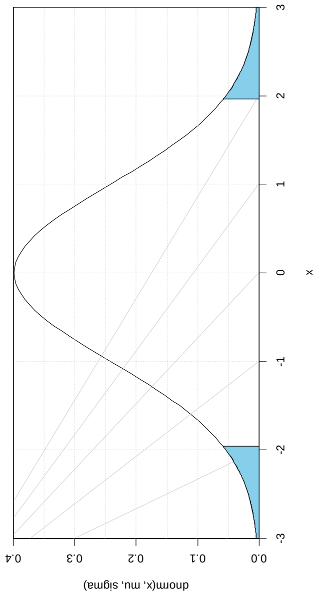




Supervised vs unsupervised learning

Supervised	Unsupervised
During model training, we have for each set of predictor variables an associated target value, i.e. we have sets consisting of pairs (predictors, response) *Examples* (age, height) (job, income) (sqm, value of house) …	We have measurements without an associated response. It will not be possible to fit a basic regression model. Instead, we may infer relationships between our measurements/observations e.g. through clustering in distinct groups. *Examples* Communities in social networks, classification based on personality traits, gene expression data, etc.

Significance



The value for which P = ·05, or 1 in 20, is 1·96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a false indication only once in 22 trials, even if the statistics were the only guide available.

Sir R.A. Fisher, *Statistical Methods for Research Workers*, 1925

P hacking

No shame in searching for patterns in data. However, shame on you, should you apply a significance test to patterns found during your search.

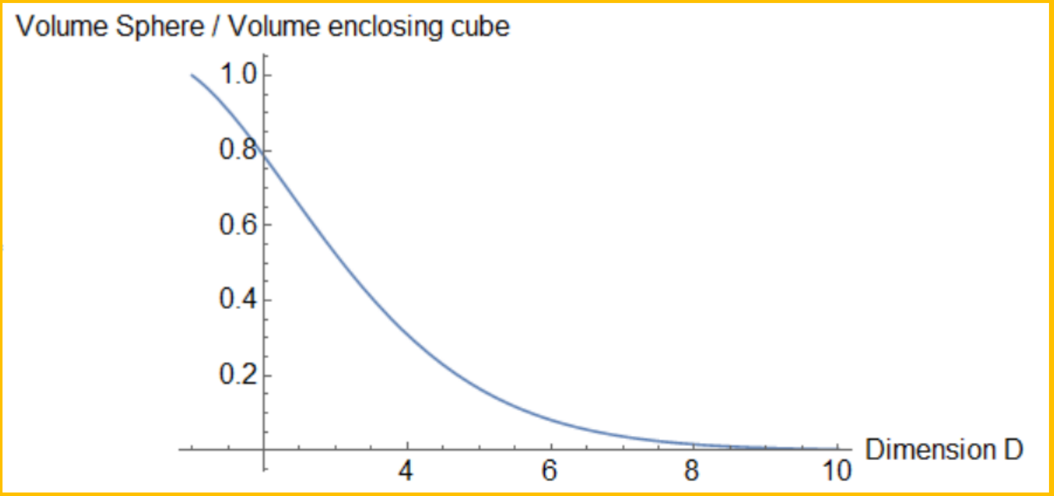
P-hacking (or data dredging) is the blind search for patterns that can be presented as significant without being part of a prior hypothesis.

Curse of dimensionality

Th
sp
di

Al
sp
sp

If data is clustered inside a sphere, the relative fraction of space covered approaches zero as the number of dimensions increase.



Curse of dimensionality

Under fairly broad assumptions (for basic norms in vector space), there will for all $\epsilon > 0$ exist a number of dimensions D such that

$$P(\, d_{min} > (1-\epsilon) d_{max}\,)=1$$

The distance between a data point and its closest neighbours will as the dimensionality increases approach the distance to the most remote data points in feature space.

Contrast disappears!

Outline

I. Basic examples
   - Linear models
   - Cross-validation

II. Feature Extraction
   - Principal Component Analysis
   - Application to questionnaire data on personality and in gender prediction

III. Support vector machines
   - Introduction, hyper planes, binary classification, regression
   - application to real estate data

IV. Quick introduction to deep learning using Keras
   - Convolutional neural networks
   - Neural networks for gender prediction

Linear Models – Why Bother?

- Despite the lack of sex appeal, it must be emphasized that linear models are often most useful.
- Simple meaningful interpretation of model parameters + a well developed framework for model validation.
- More exotic methods build upon insights from linear models.
- For low quality data, a linear regression is often as good or better than anything else.

Linear Regression

In linear regression, the learning part is to establish the approximate linear relationship between the target variable(s) and the N predictor variables

$$Y \approx a_0 + \sum_{i=1}^{N} a_i X_i + \xi,$$

where ξ is assumed to be normal distributed noise. For a given observation (labelled k)

$$\{x_i^k\}_{i=1}^N$$

the prediction $\hat{y}^k$ is computed by

$$\hat{y} = a_0 + \sum_{i=1}^{N} \hat{a}_i x_i$$

where the Model parameters, $\hat{a}_i$, are estimated by minimizing the sum of squared residuals over a set of M observations of the target variable y^k and predictor variables x_i^k,

$$RSS = \sum_{k=0}^{M} (\hat{y}^k - y^k)^2$$

Example: Property Sales Price

Consider the relation between the sales price y_i and the size in square meters of appartments x_i.

$$y_i = a_0 + a_1 \times x_i + \xi_i$$

the standard error of the predicted model parameters are computed from

$$SE(\hat{a}_0) \approx \sigma^2(\xi) \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{k=1}^{M} (x^k - \bar{x})^2} \right), \qquad SE(\hat{a}_1) \approx \frac{\sigma^2(\xi)}{\sum_{k=1}^{M} (x^k - \bar{x})^2}$$

where

$$\sigma^2(\xi) \approx \sqrt{\frac{RSS}{M-2}}$$

We now perform a hypothesis test on our paramters a_i by testing the null hypothesis $a_i = 0$ against the alternative hypothesis $a_i \neq 0$. For that purpose we use the t-distribution with $M - 2$ degrees of freedom on

$$t = \frac{\hat{a}_i}{SE(\hat{a}_i)}$$

A p-value is then obtained by computing the probability of observing a value larger than or equal to $|t|$.

Feature Extraction – Alleviating the Curse of Dimensionality

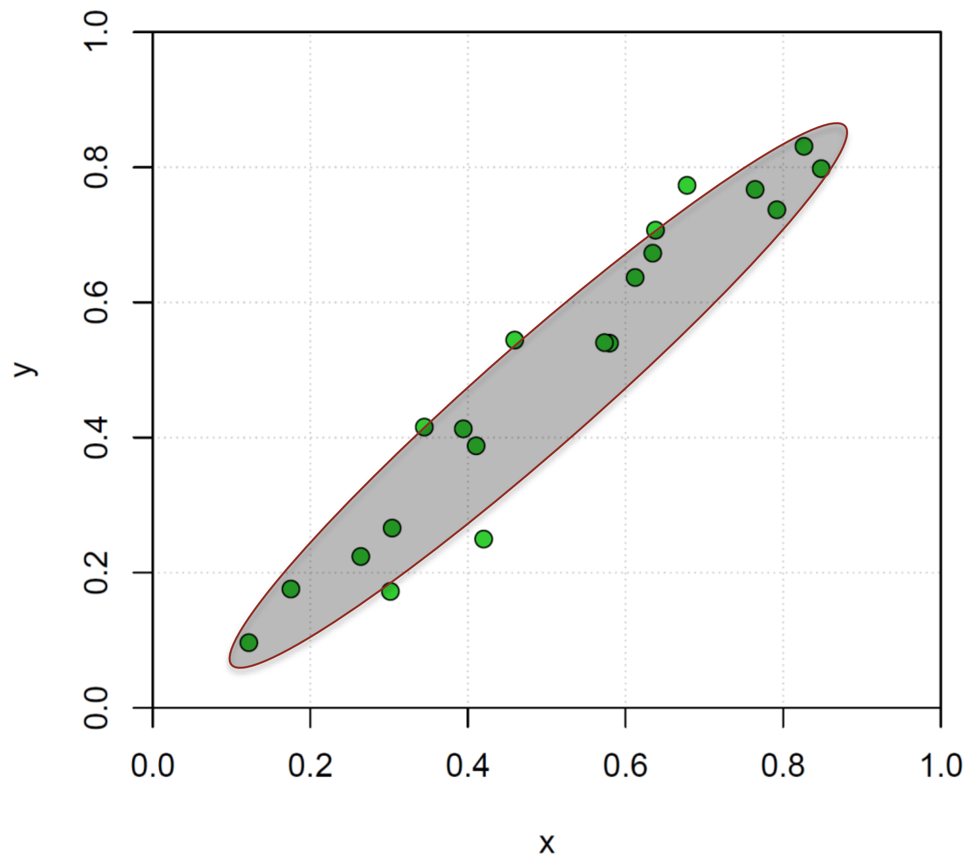
Feature Extraction – the art of keeping only the relevant information and discard everything else. In other words, the task is to perform a sensible dimensionality reduction of data through a linear or non-linear transformation of data.

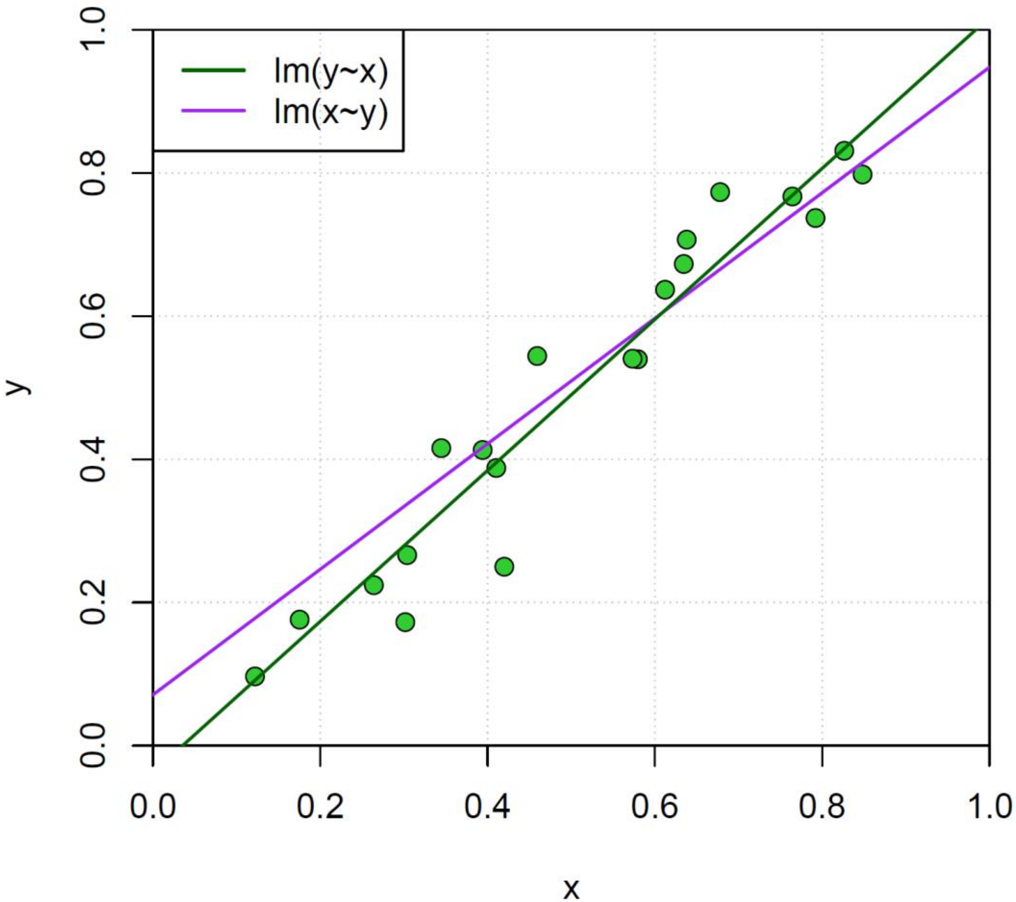
Often the feature extraction is a significant part of the overall process

Several general methods exist, for example

- Independent Component Analysis
- t-Distributed Stochastic Neighbor Embedding
- Factor Analysis
- Principal Component Analysis
- Non-linear reduction methods

Principal Component Analysis



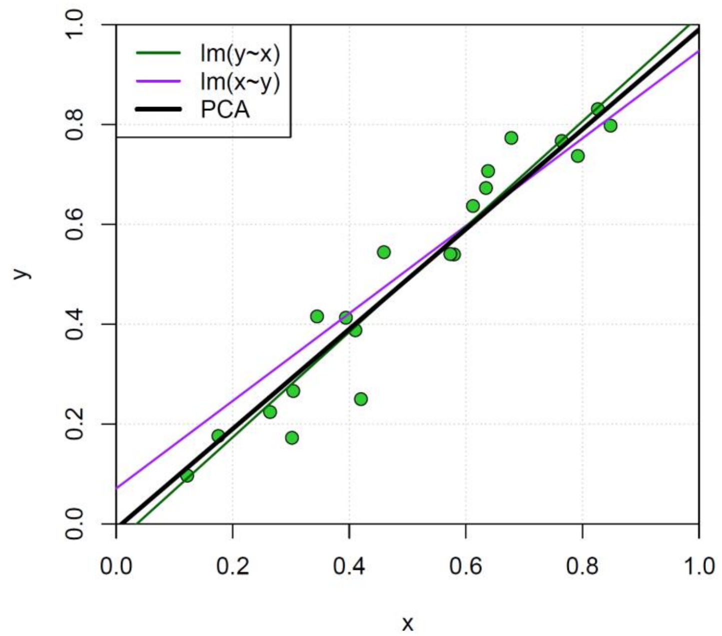


The linear model lm(x~y) minimizes the least squares error horizontally, whereas

lm(y~x) minimizes the vertical error.

The two fits will in general be different.

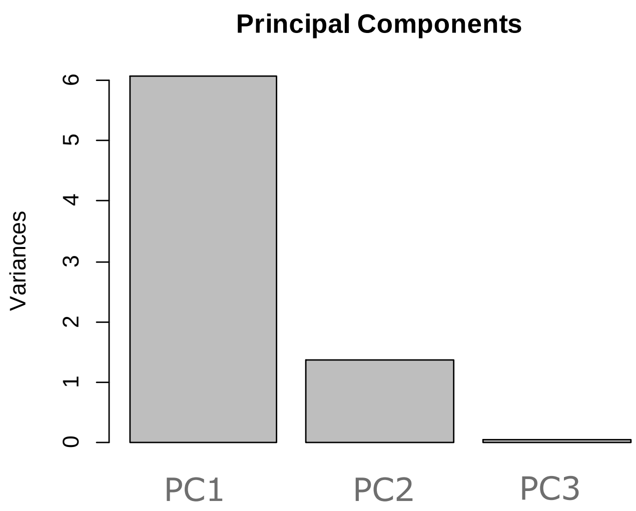
```
#perform the PCA
pcs=prcomp(cbind(x,y),scale=T)
pcs$rotation

##         PC1        PC2
## x 0.7071068  0.7071068
## y 0.7071068 -0.7071068

pcs$sdev

## [1] 1.4006318 0.1955262
```

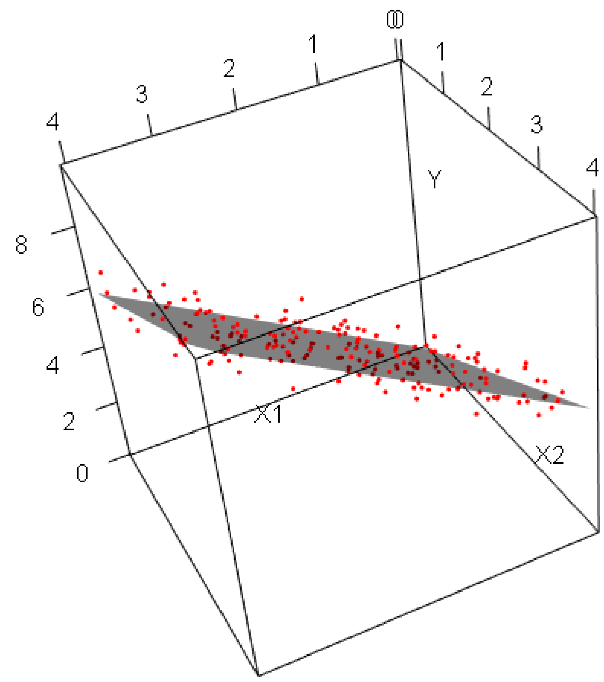


The leading principal component is in the direction of most variance.

The variance around PC2 is considerably smaller than around PC1, hence data can be approximated to some extent by a one dimensional line.

Principal Component Analysis





Little variance around the third PC, i.e. data roughly exists in a 2D space.

Applications of Principal Component Analysis

The general paradox of feature extraction is that when you throw away potentially relevant information you may greatly improve the performance of your model.

Feature extraction is important not only as input to statistical models but also as general way to compress data. As an example, we shall here consider the MNIST data of handwritten digits.

THE MNIST DATABASE

of handwritten digits

Yann LeCun, Courant Institute, NYU
Corinna Cortes, Google Labs, New York
Christopher J.C. Burges, Microsoft Research, Redmond

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

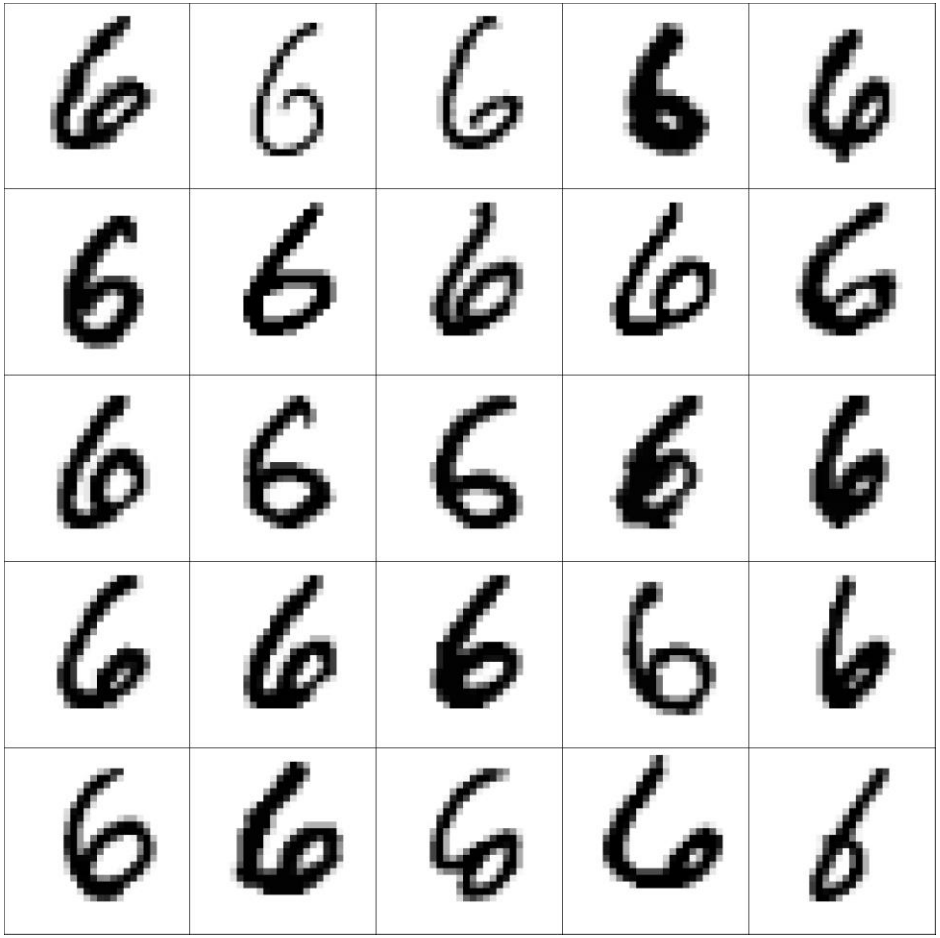
It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Four files are available on this site:

```
train-images-idx3-ubyte.gz:  training set images (9912422 bytes)
train-labels-idx1-ubyte.gz:  training set labels (28881 bytes)
t10k-images-idx3-ubyte.gz:   test set images (1648877 bytes)
t10k-labels-idx1-ubyte.gz:   test set labels (4542 bytes)
```

http://yann.lecun.com/exdb/mnist/

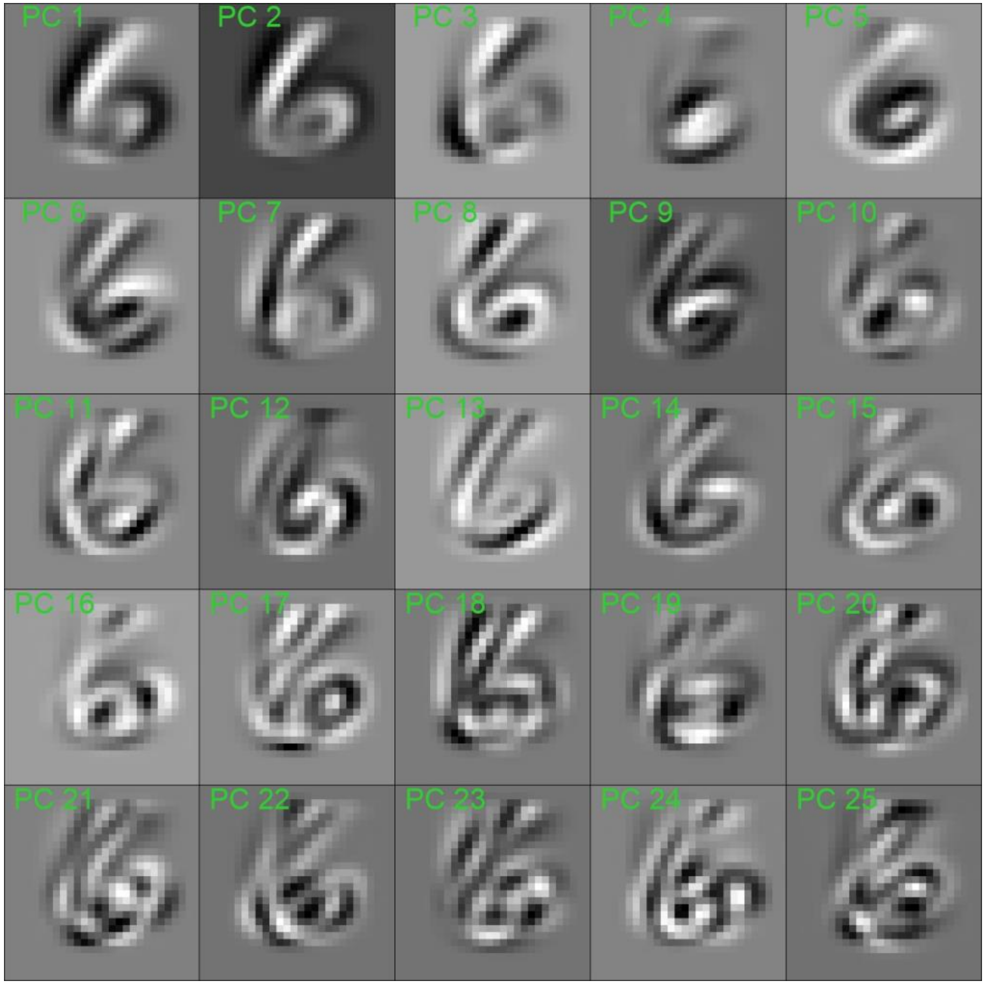
Digit compression



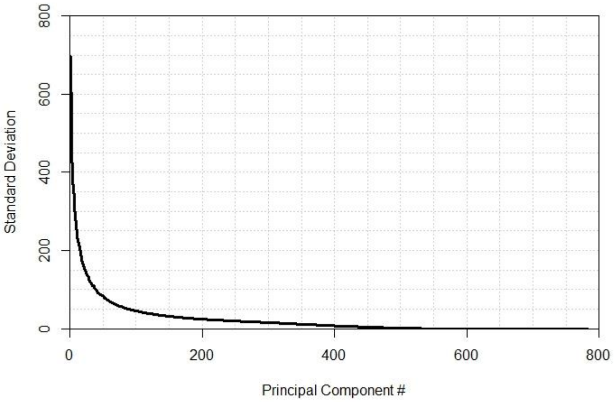
All digits come as 28x28 grey scale values in the range 0-255.

A digit is therefore a vector in 784 dimensional space.

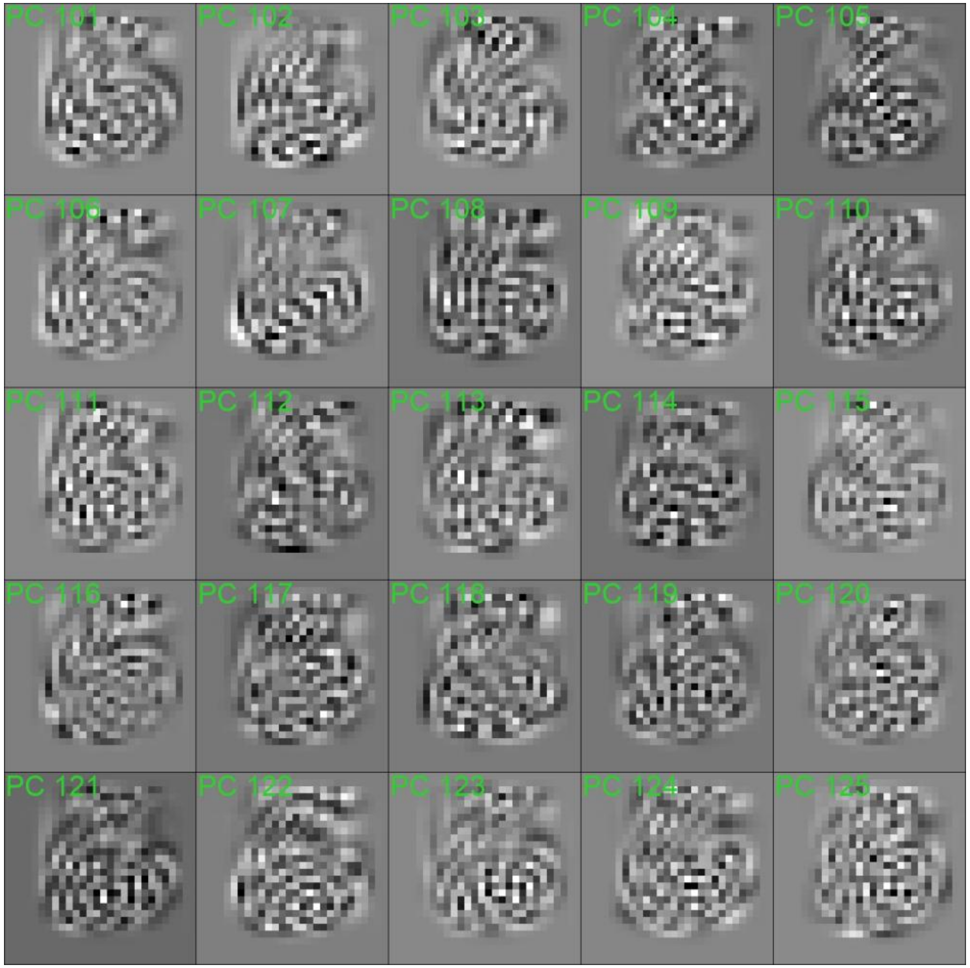
Digit compression



Plot of the 25 leading components achieved from approximately 5000 handwritten digits. 60% of the variance is accounted for in the 10 first components.

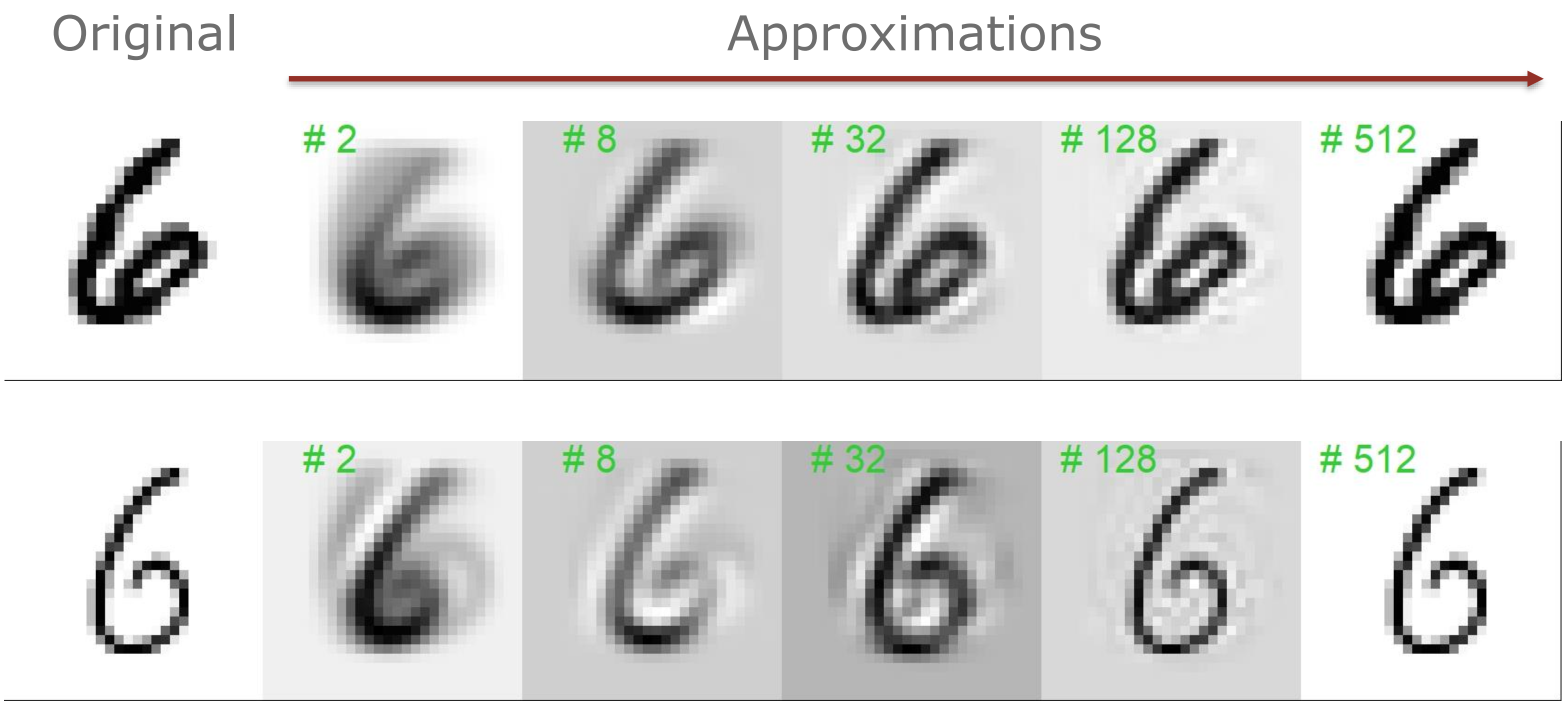


Digit compression



Plot of the 101st to 125th components. Limited information is left.

Digit compression

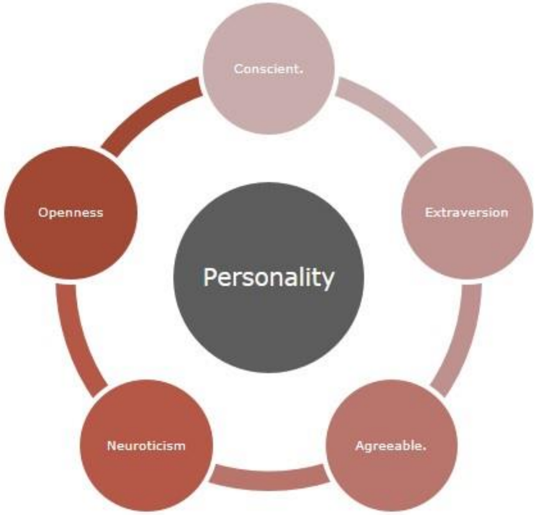


Useful for digit recognition?

CLASSIFIER	PREPROCESSING	TEST ERROR RATE (%)	Reference
Non-Linear Classifiers			
40 PCA + quadratic classifier	none	3.3	LeCun et al. 1998
1000 RBF + linear classifier	none	3.6	LeCun et al. 1998
SVMs			
SVM, Gaussian Kernel	none	1.4	
SVM deg 4 polynomial	deskewing	1.1	LeCun et al. 1998
Reduced Set SVM deg 5 polynomial	deskewing	1.0	LeCun et al. 1998
Virtual SVM deg-9 poly [distortions]	none	0.8	LeCun et al. 1998
Virtual SVM, deg-9 poly, 1-pixel jittered	none	0.68	DeCoste and Scholkopf, MLJ 2002
Virtual SVM, deg-9 poly, 1-pixel jittered	deskewing	0.68	DeCoste and Scholkopf, MLJ 2002
Virtual SVM, deg-9 poly, 2-pixel jittered	deskewing	0.56	DeCoste and Scholkopf, MLJ 2002

http://yann.lecun.com/exdb/mnist/

Exercise


Conscient.
Extraversion
Agreeable.
Neuroticism
Openness
Personality

The data

43 answers to questions about personality.

Answers indicate how well a person agrees with a given statement and are encoded as values in the range 0 to 4.

```
dat1=read.table("RGender.dat")
#Dimensions of data frame
dim(dat1)

## [1]  44 940

#Row or column names
rownames(dat1) #colnames(dat1)

##  [1] "gender"                        "bfi_worry.answer"
##  [3] "bfi_stable.answer"             "bfi_few_art.answer"
##  [5] "bfi_caring.answer"             "bfi_helpfull.answer"
##  [7] "bfi_distract.answer"           "bfi_inventive.answer"
##  [9] "bfi_tense.answer"              "bfi_talk.answer"
## [11] "bfi_reliable.answer"           "bfi_cold.answer"
## [13] "bfi_rutine.answer"             "bfi_disorderly.answer"
## [15] "bfi_social.answer"             "bfi_fight.answer"
## [17] "bfi_energi.answer"             "bfi_reserved.answer"
## [19] "bfi_imagination.answer"        "bfi_quiet.answer"
## [21] "bfi_creative.answer"           "bfi_confident.answer"
## [23] "bfi_effective.answer"          "bfi_strong_personality.answe
## [25] "bfi_enthusiasm.answer"         "bfi_rude.answer"
## [27] "bfi_unbalanced.answer"         "bfi_taste_art.answer"
## [29] "bfi_original.answer"           "bfi_error.answer"
## [31] "bfi_coorporation.answer"       "bfi_hold_on.answer"
## [33] "bfi_currious.answer"           "bfi_relaxed.answer"
## [35] "bfi_work.answer"               "bfi_play.answer"
## [37] "bfi_lazy.answer"               "bfi_shy.answer"
## [39] "bfi_depressed.answer"          "bfi_art.answer"
## [41] "bfi_careless.answer"           "bfi_nervous.answer"
## [43] "bfi_forgive.answer"            "bfi_calm.answer"
```

Personality traits

We consider the following traits, “the Big Five” or the “Newtonian Mechanics” of psychology:

- Conscientiousness - A tendency to be organized and dependable
- Agreeableness - A tendency to be compassionate and cooperative
- Neuroticism - the tendency to experience unpleasant emotions easily
- Openness - reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has.
- Extraversion – outgoing and energetic

44 questions to determine personality traits

I see Myself as Someone Who...

____1. Is talkative

____2. Tends to find fault with others

____3. Does a thorough job

____4. Is depressed, blue

____5. Is original, comes up with new ideas

____6. Is reserved

____23. Tends to be lazy

____24. Is emotionally stable, not easily upset

____25. Is inventive

____26. Has an assertive personality

____27. Can be cold and aloof

____28. Perseveres until the task is finished

Disagree strongly	Disagree a little	Neither agree nor disagree	Agree a little	Agree Strongly
1	2	3	4	5

Dimensionality of feature space

From the principal component analysis of the data, does it seem reasonable to claim that personality can be mapped by five dimensions?

Gender prediction

Use the principal components to learn a statistical model and use the model to assess how well one can predict gender using as input the Big Five personality items. Try with different models, e.g. GLM and SVM and compute the corresponding ROC curves.