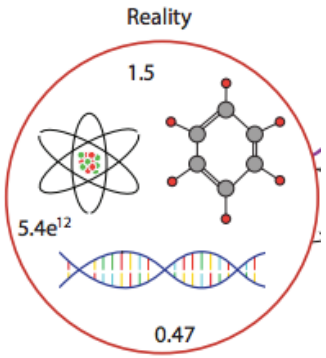
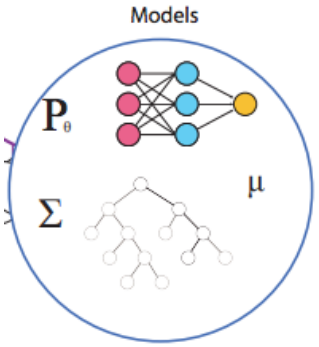




Statistics 154, Spring 2019

Modern Statistical Prediction and Machine Learning

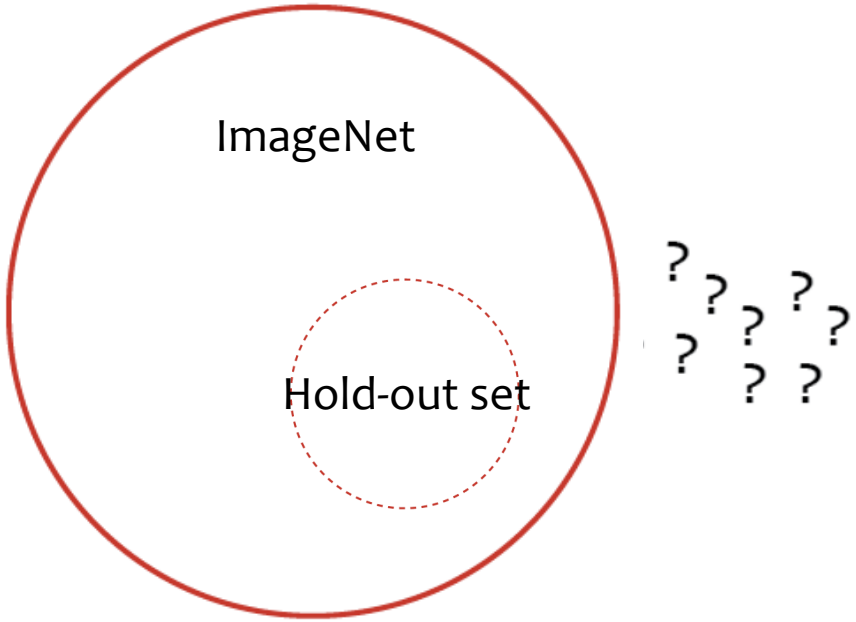
Lecture 5: Principal Component Analysis (PCA)- a popular unsupervised learning tool

Instructor: Bin Yu
(binyu@berkeley.edu); office hours: Tu: 9:30-10:30 am; Wed:**1:30-2:30 pm (change)**
office: 409 Evans

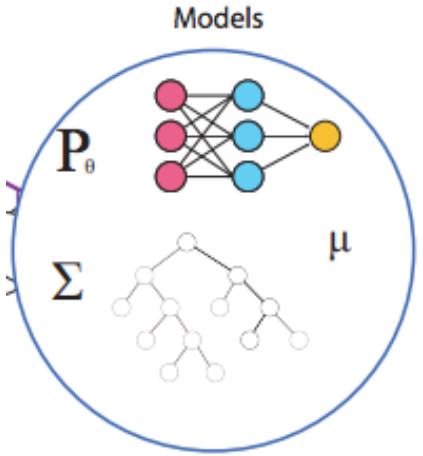
GSIs: Yuansi Chen (Mon: 10-12; 4-6); Raaz Dwivedi (Mon: 12-2; 2-4)
yuansi.chen@berkeley.edu; raaz.rsk@berkeley.edu
(Yuansi: Tuesday 1-3; Raaz: Monday 10:30-11:30, Thurs. 9:30-10:30)

Internal validity is a minimal requirement ??

Training data set



Stats/ML algorithms

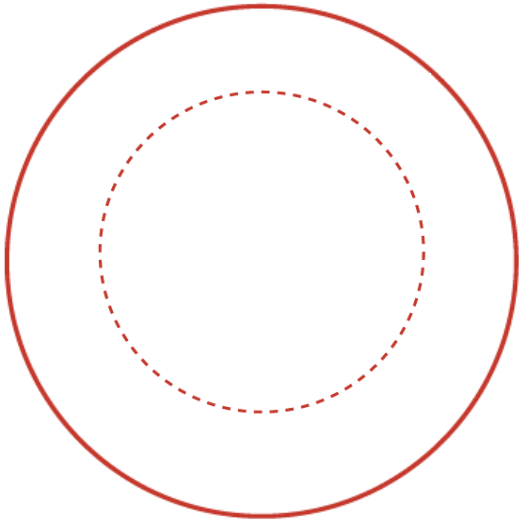


Internal validity: algorithm predicts well on hold-out data from the red circle

External validity is much harder and requires info on the future reality

A simulated prediction problem

- Joe has data as a random sample of 100 from the entire sale price data set, and he wants to choose between mean or median as the ball-park number to give to Jane as a prediction for the next random draw (as a proxy to the next house getting on the market) from the entire data set, with squared error as a prediction performance metric



Translated in math terms

- Joe has data as iid random samples $X_1, \ldots, X_n$ (n=100), he wants to predict Y, indep and identically distributed as $X_1, \ldots, X_n$

- He wants to use data to predict Y with expected squared error as a performance metric $E(\hat{Y} - Y)^2$, which is minimized by the expected value of Y = EY= μ

- What is we use expected absolute error?

Translated in math terms

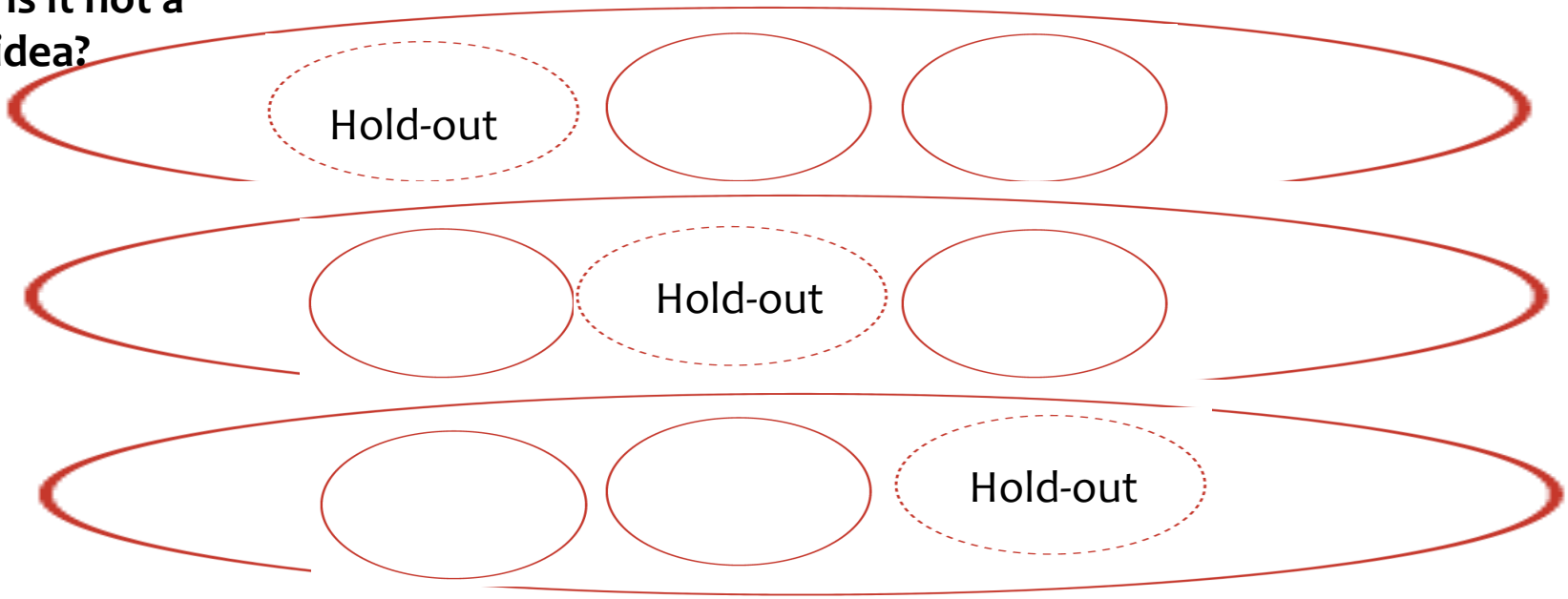
- Q1: given an predictor $\hat{Y}$,
  how to estimate prediction error $E(\hat{Y} - Y)^2$?

- Q2: can we use the estimated prediction error in Q1 to choose between mean and median?

Cross-validation (CV): hold-out sets are re-used and in model fitting, to estimate prediction error within one data set

Given a prediction problem with an "exchangeable" data set, CV creates k "pseudo-replicated" prediction problems or it creates K hold-out sets. K=3 below.

When is it not a good idea?



CV prediction error is the average over K-fold
(not always a good estimate of the pred. error)

CV with K=10 (often used)

- The expected value of the CV prediction error is the prediction error if we only have 90% of the data or 90 data point when n=100

- The prediction error of sample mean with n samples is

  var(Y)(1+1/n) (**black board** derivation)

  compared with

  var(Y)(1+1/(0.9n)) for expected CV prediction error

Blackboard work on math notations for CV

- For K-fold CV, divide the data into K blocks of size m $Z_1, \ldots, Z_K$ indexed by i=1,…, K, where m=n/K is the number of data units in The ith block

$$Z_i = (X_i, Y_i) \in R^{m \times (p+1)}$$

- Denote the (K-1) blocks without the ith block by Z_{-i}

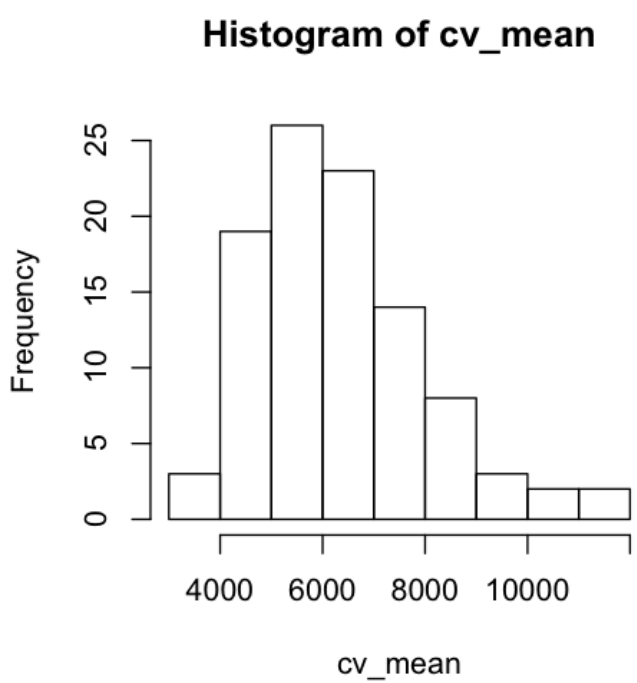
- One can develop a predictor based on Z_{-i} : $\hat{f}(Z_{-i})$ to predict the ith block Z_i , for example, the mean or median of the Y's in Z_{-i}

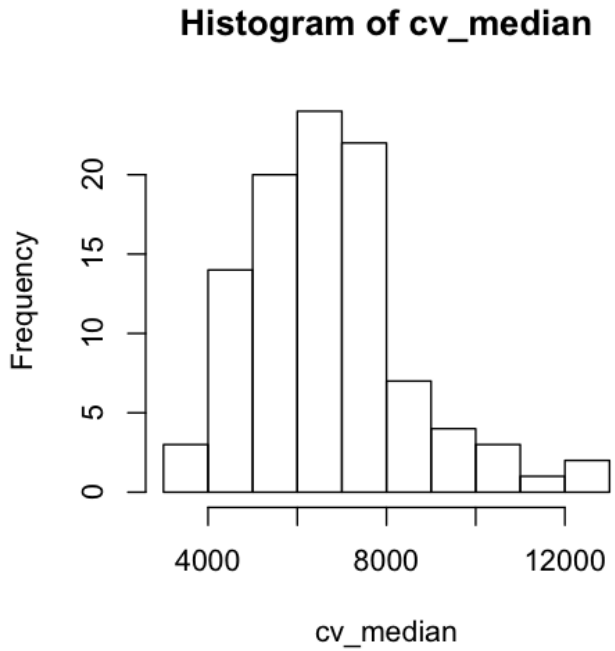
Blackboard work on math notations for CV

- We have a loss function $\ell(\cdot, \cdot)$ to measure the prediction error, then we get prediction error (PE) on the ith block.
- For $u, w \in R^m$, define $\ell(u, w) := \sum_{j=1}^{m} \ell(u_j, w_j)$
- Let $PE_i = \frac{1}{m} \ell(\hat{f}(Z_{-i}), Y_i)$
- The CV (estimated) prediction error is $CV_{\hat{f}}(\ell) = \frac{1}{K} \sum_{i=1}^{K} PE_i(\ell)$
- When the loss is squared error, we get $CV_{\hat{f}}(MSE) = \frac{1}{K} \sum_{i=1}^{K} MSE_i$

CV K=10, correct prediction error is around 6,000, but CV error could be as small as 4,000 or as large as 10,000 or 12,000

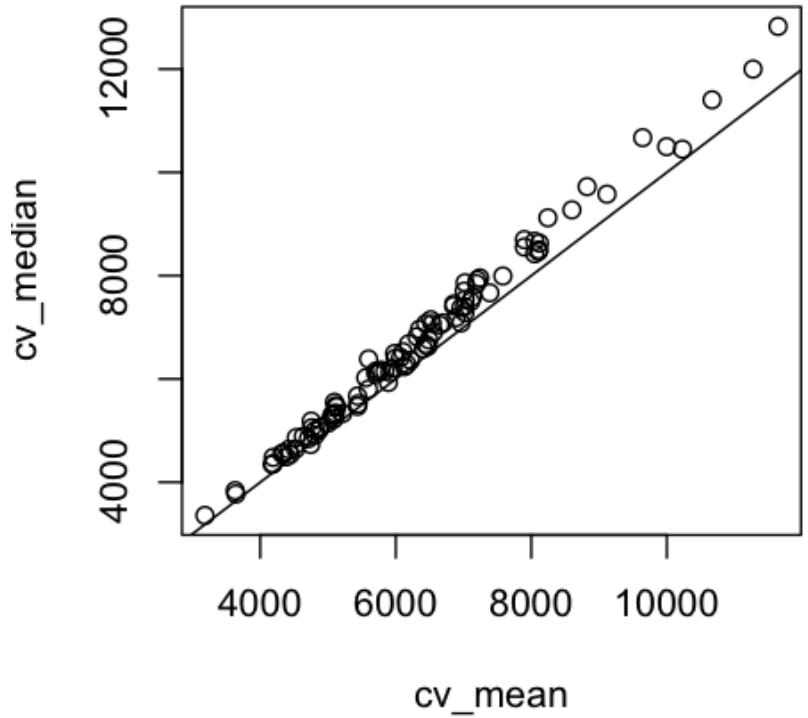
- When we have different samples of data, the CV prediction error has quite a big variability, worse for median





CV K=10: with a diagonal line via abline(0,1)

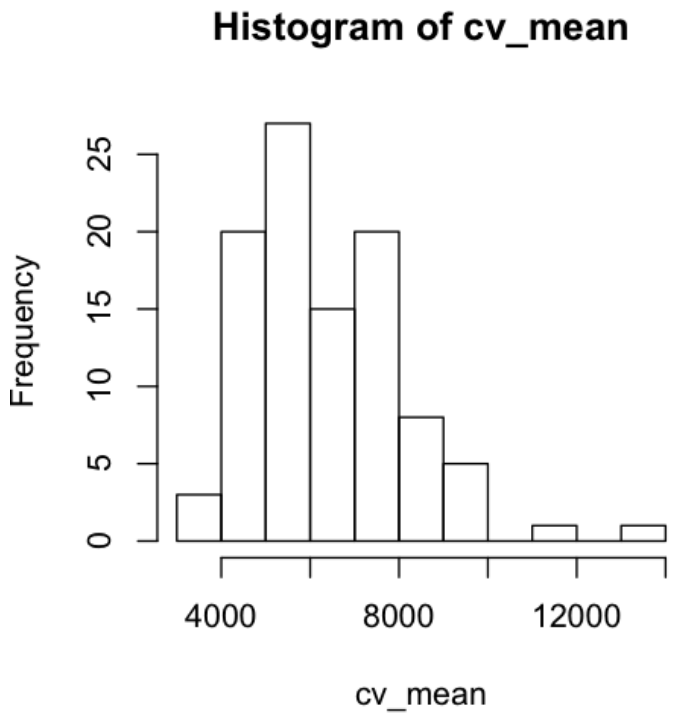
- For each run (or each 100 samples), CV CAN help us decide on mean vs. median almost all the time since the points are almost all above the line, which means that cv_median is larger than cv_mean – implying that mean is chosen by CV with squared error.



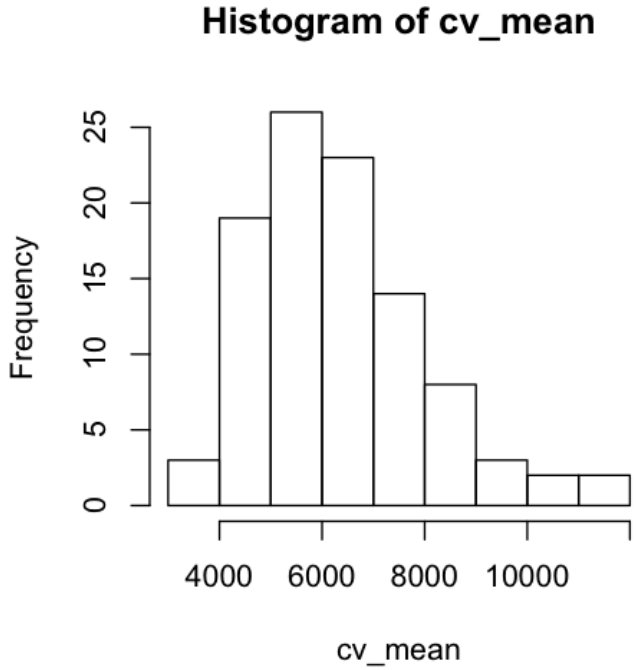
Similarly for Absoluate value Error.

CV with K=n: leave-one-out CV (mean)

K=n has CV prediction error
closest to the n-sample prediction error
but with a larger variance, when compared to K=10



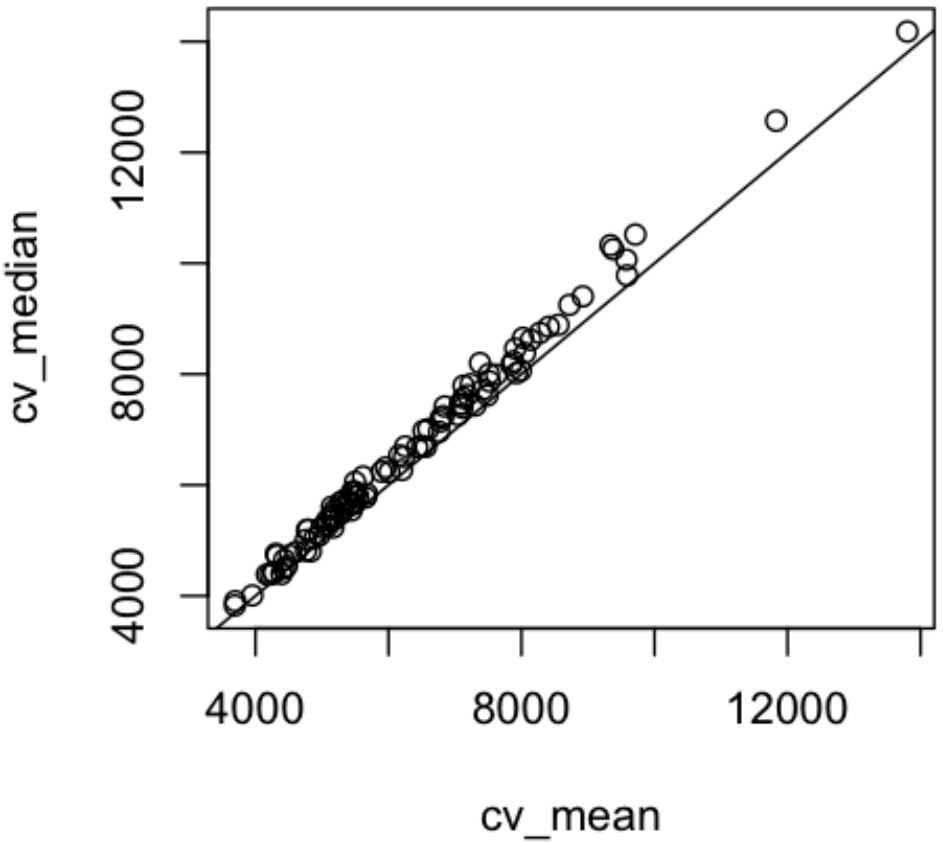
K=n



K=10

But CV with L2 still chooses mean almost all the time (similarly chooses median for L1)

- K=n, leave-one-out CV



A good practice is 3-partition of data

- For exchangeable data and approx. not dependent
- First thing, set aside a test set (20-30%): that is set aside not used in model fitting and comparisons -- only for estimating the prediction error or confirming discoveries at the very end
- Divide training data into two parts (20-30% and the rest)
  validation set (set aside to get prediction error and used many times)
  fitting set (run CV on)

Summary on CV

- Con: **CV prediction error is NOT the real prediction error** – it could be quite away from it (both ways: over or under-estimation)

- Pro: CV works well for comparing methods or selecting regularization parameter (e.g. bandwidth in kernel density or histogram)

- Prediction error metric (e.g. L1 vs. L2) matters, esp. in high dim

The above holds not just for the toy simulation that we did, but in general. However, most people often don't realize the **con** part…

Taxonomy

of Machine Learning/Statistics

Labeled Data

Indirect (reward)

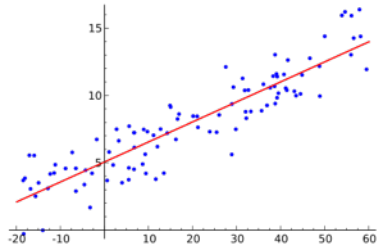
Unlabeled Data

Supervised Learning

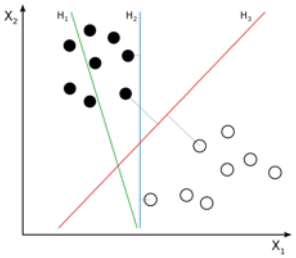
Reinforcement & Bandit Learning

Unsupervised Learning

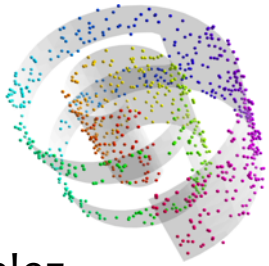
Regression



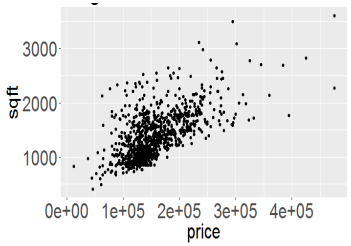
Classification



Dimensionality Reduction



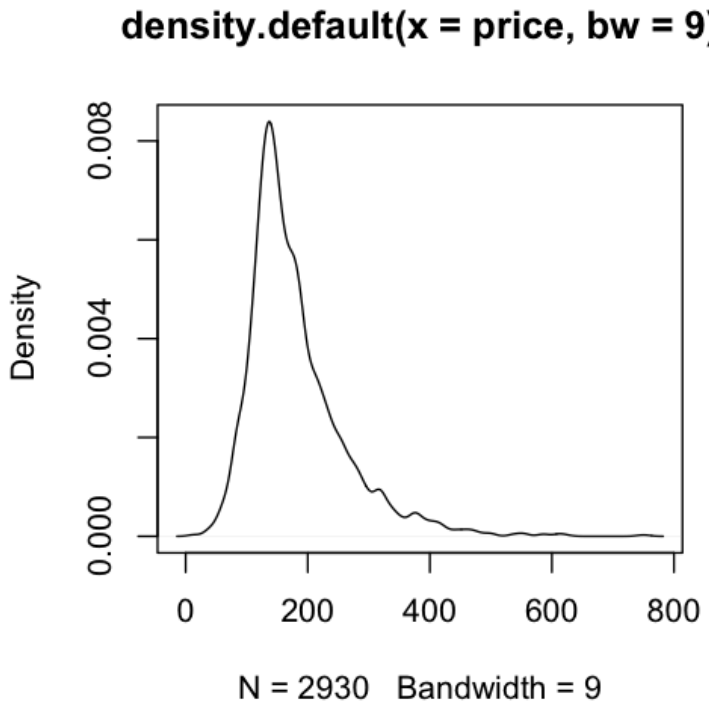
Clustering



Thanks to J. Gonzalez

Unsupervised learning

At a high level, unsupervised learning is about finding interesting patterns in data **without** a response variable

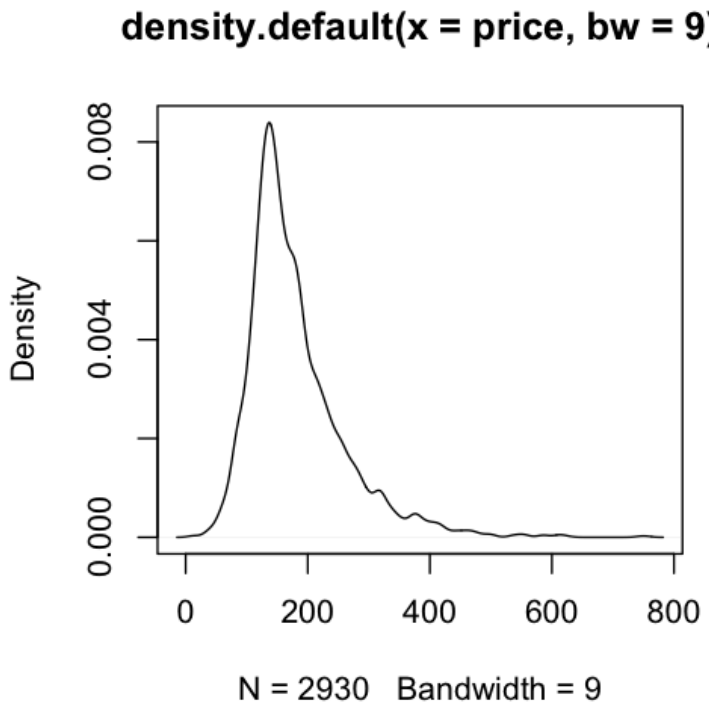


What is an interesting pattern about this data set?

Ames house sales price from 2006-2010

Supervised learning

At a high level, supervised learning is about finding interesting patterns in data **without** a response variable



Ames house sales price from 2006-2010

Interesting pattern:

one mode or peak
at about 150K, which
means that within a fixed
price interval, there are
more houses around 150K;
or one main cluster around 150K

How do we visualize beyond 4 dimensions? Ames data has 80 features

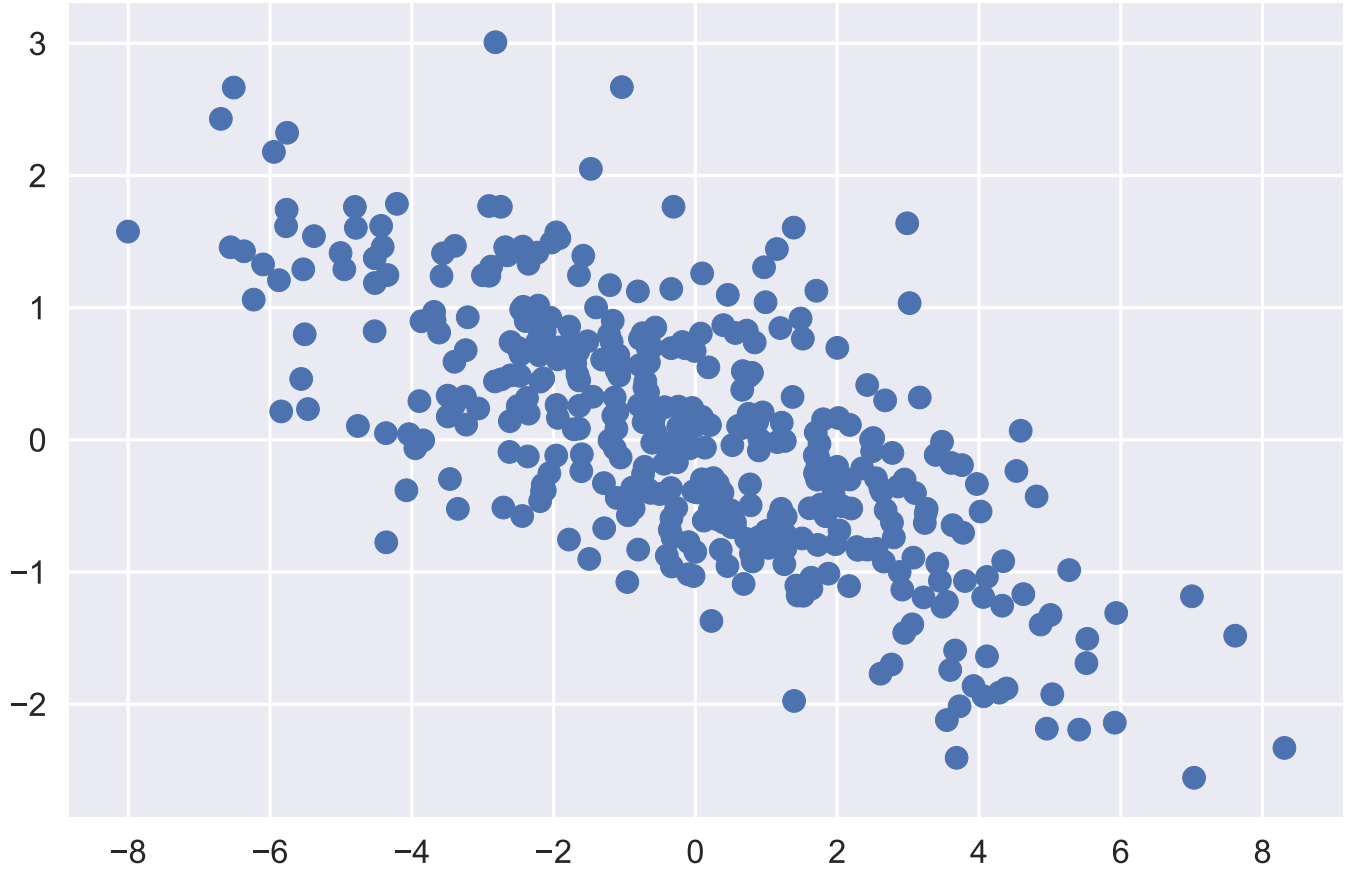
- We can at most "look" at 3-dim data (or 4-dim with movies of data)
- Visual cortex takes up about 30% of our brain so we want to see projections of data into low dim spaces (2d or 3d, for example)
- How to project?
- Random projection? Or?

Dimension reduction for more interpretable results of high-dim data

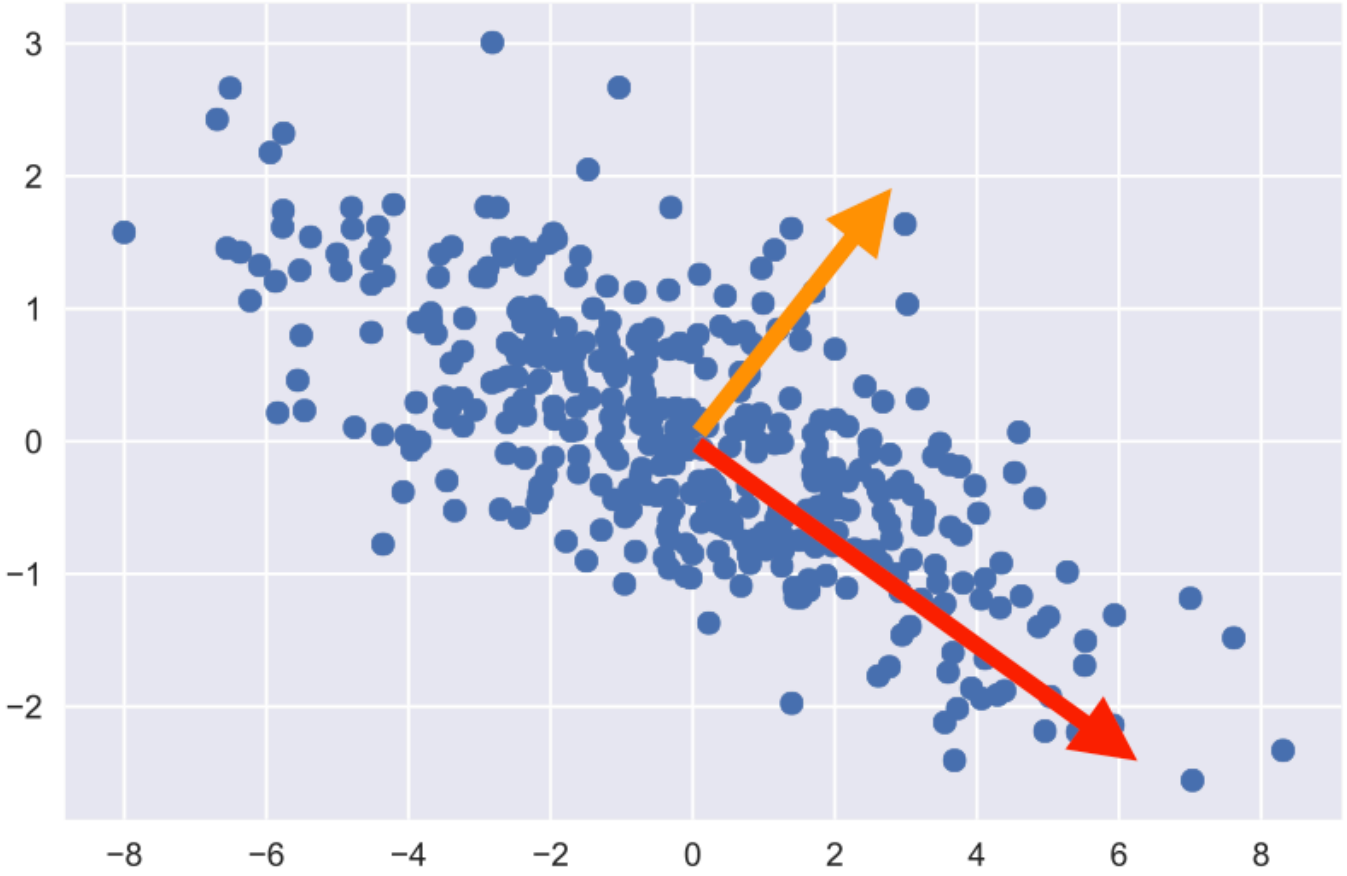
Dimensionality reduction is needed for

- Visualization
- Fast computation
- Smaller storage and faster communication
- One form for regularization: Simpler models (to reduce variance of estimation)

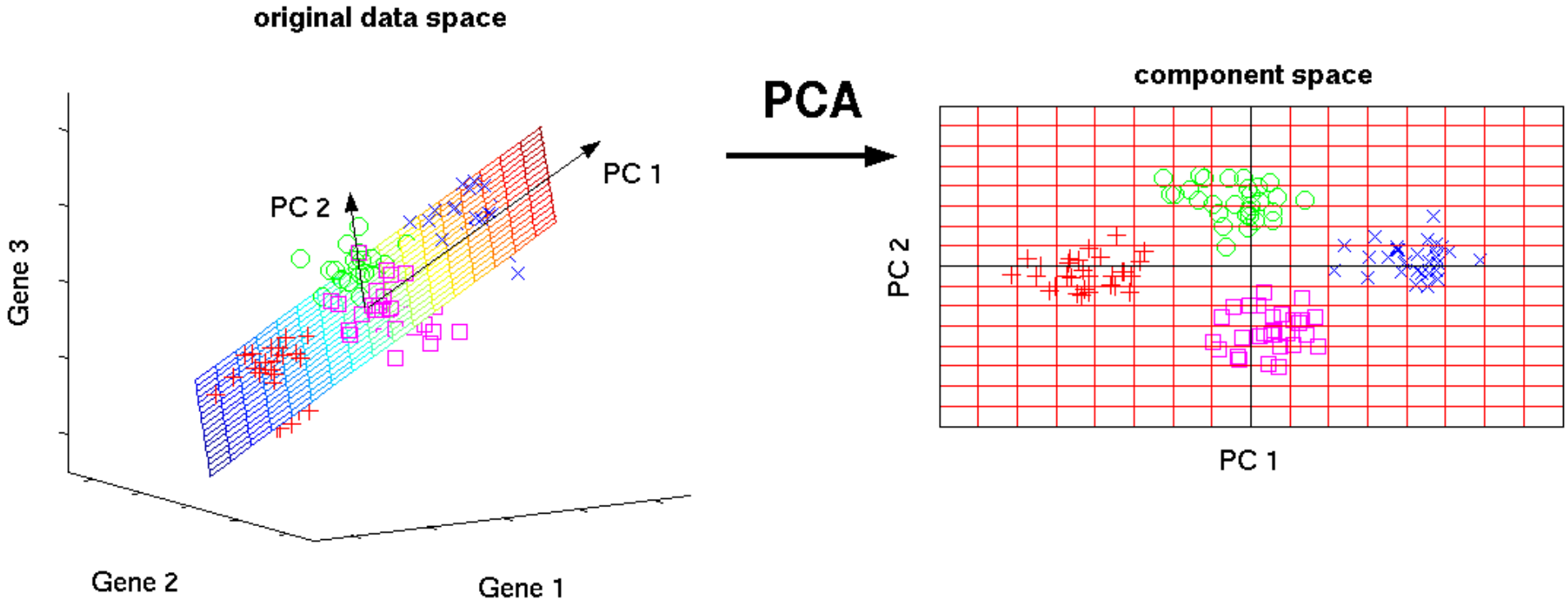
Principal component analysis (PCA) in 2-dim



Principal component analysis (PCA) in 2-dim



PCA in 3-dim to 2-dim: 1-dim lost
when is it not a bad idea?



PCA: Core Idea

The central idea of PCA is :

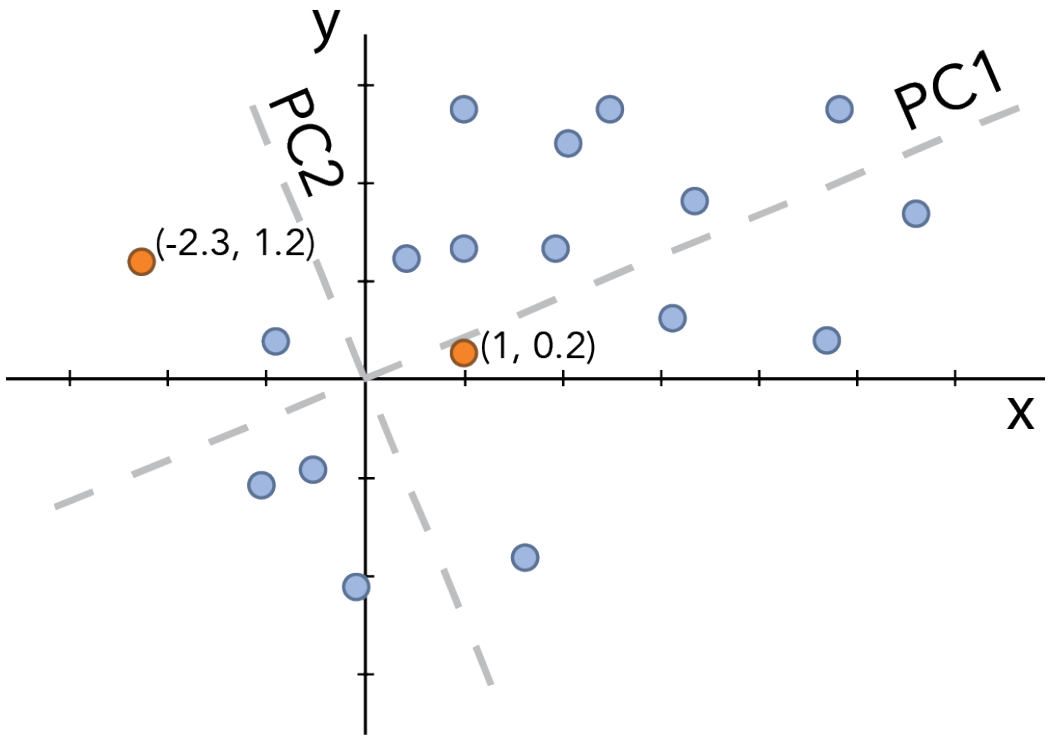
- **to reduce the dimensionality** of a data set that has a large number of **interrelated** variables,
- while **retaining as much as possible of the variation** present in the data set.

This is achieved by **transforming** to a new set of variables (PCs)

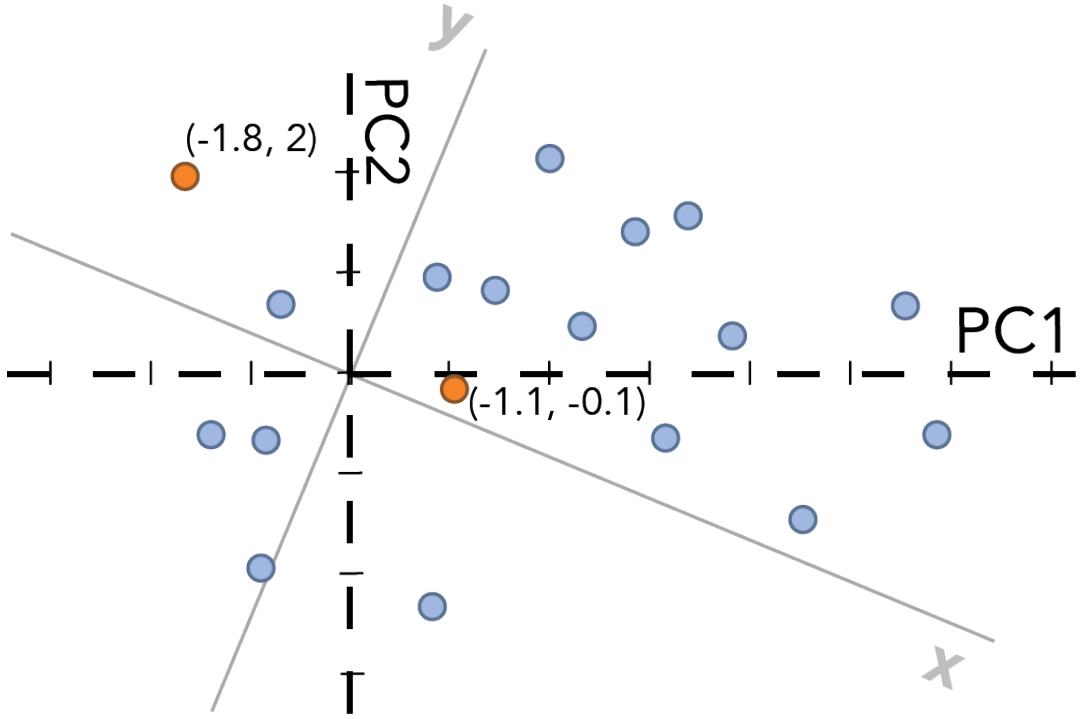
- which are **uncorrelated (orthogonal)**, and
- which are **ordered** so that **(hopefully)** the first few retain most of the variation present in all of the original variables.

Jolliffe, Principal Component Analysis, 2nd edition

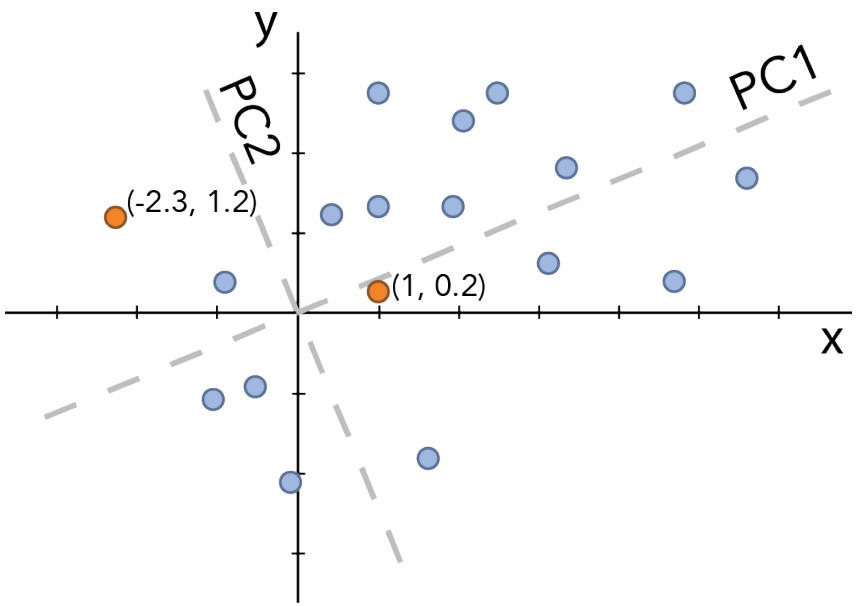
Two dimensional example:
Coordinate Axes

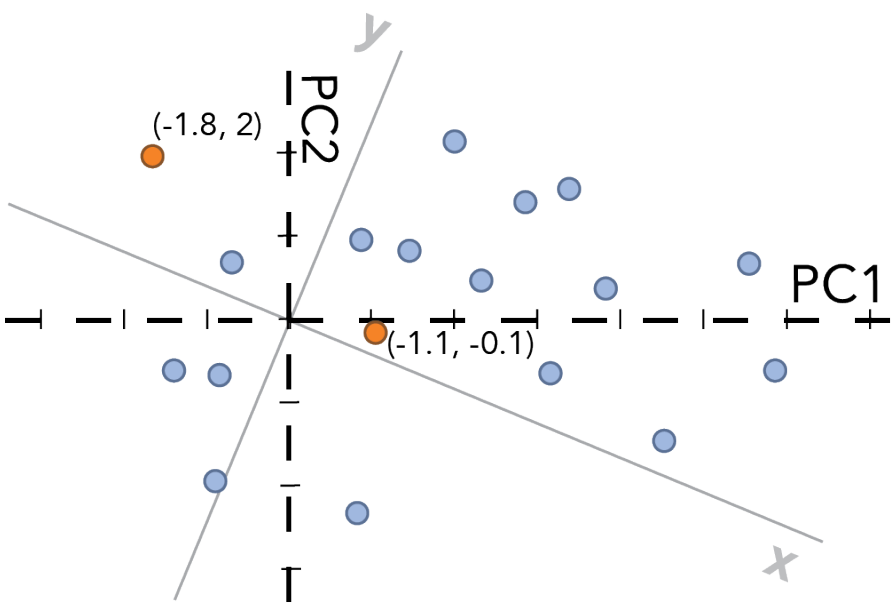


Two dimensional Example: PC Axes through rotation U in eigen decomposition of G =X'X

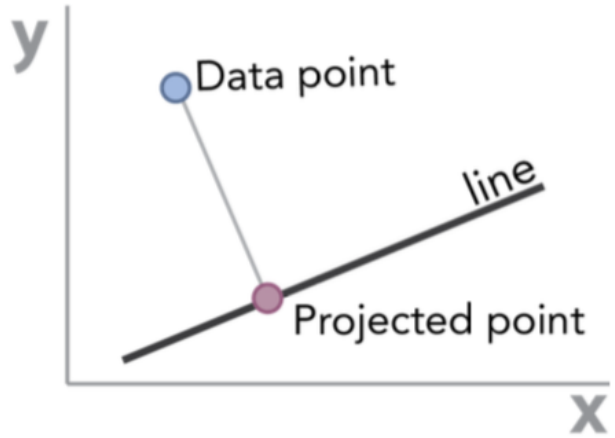


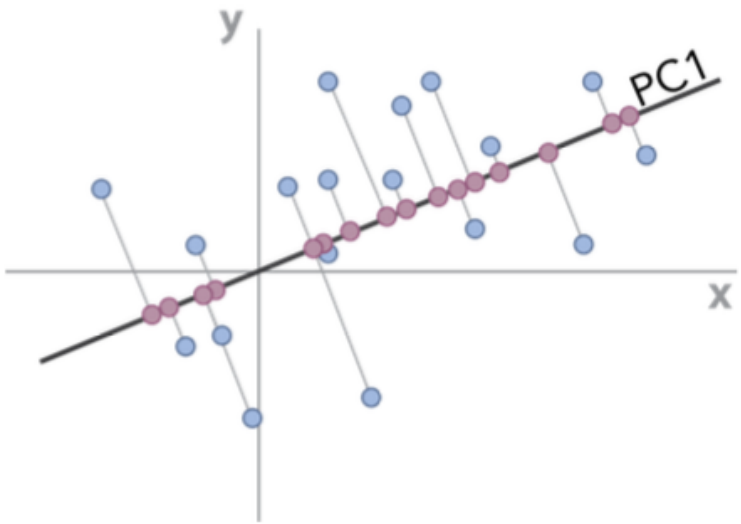
Two dimensional Example: together

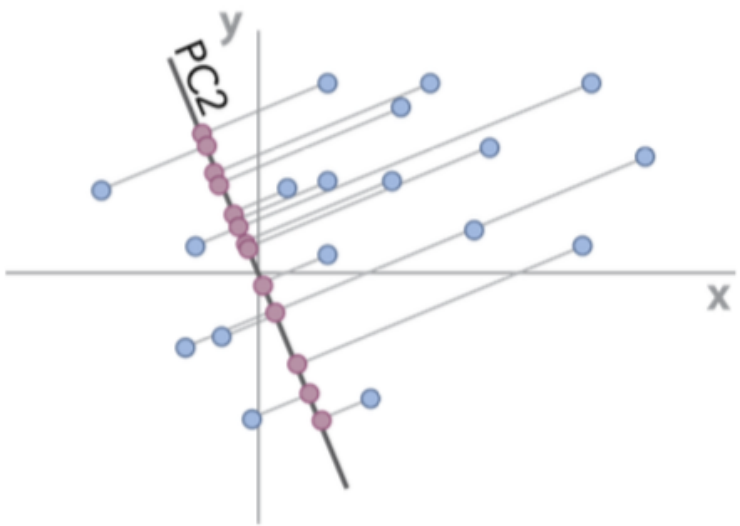




Projection on Principal Components







Math behind PCA

Blackboard work

For a sketch, see pca19.pdf

Reading assignments

- Review of eigen-value decomposition (SVD too)

- James et al chapter on PCA?