S&DS 355 / 555

Introductory Machine Learning

PCA, Mixtures and Bayes

Tuesday/Thursday, October 8/10

Yale

Checkpoint

- Assignment 3 due tonight
- My office hours today at 4:30 in 17 Hillhouse Room 115
- Assignment 4 out tonight, on PCA and logistic regression
- Practice midterm will go out next Thursday
- Midterm in class on Tuesday, October 15
- Questions?

For Today

- PCA (briefly)
- Latent variables
- Mixtures
- Bayes

PCA

PCA finds the directions of greatest variation in the data, and represents the data in terms of those directions.

PCA: Algorithm

1. Center the data: $x_i \mapsto x_i - \overline{x}$

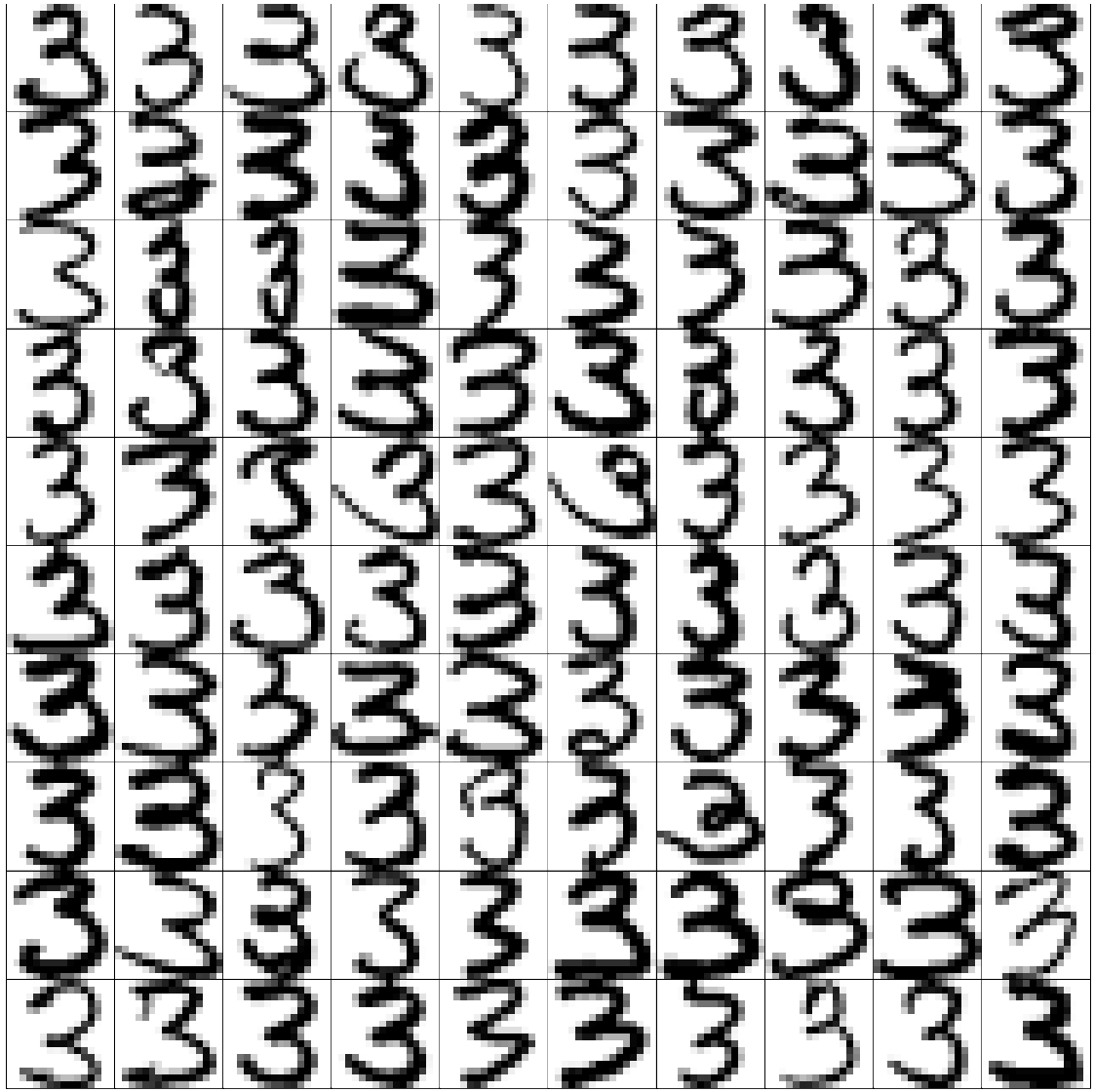
xi is d vector
xiT is 1xd vector
Product is dxd matrix

2. Compute the $d \times d$ sample covariance $S = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$

3. Find the first k eigenvectors of S

4. Project the data onto those k vectors

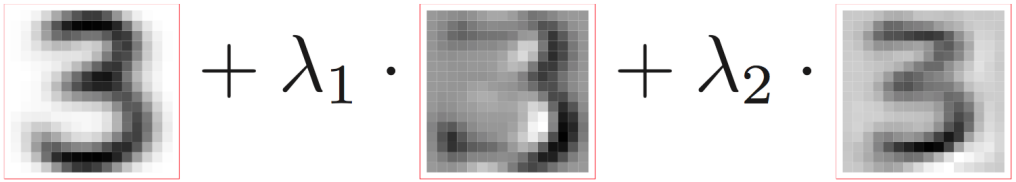
Handwritten Digits (3s)



Handwritten Digits (3s)

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

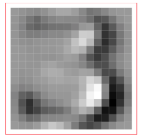
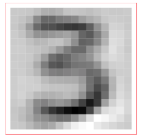
$$= \bar{x} + \lambda_1 \cdot v_1 + \lambda_2 \cdot v_2.$$

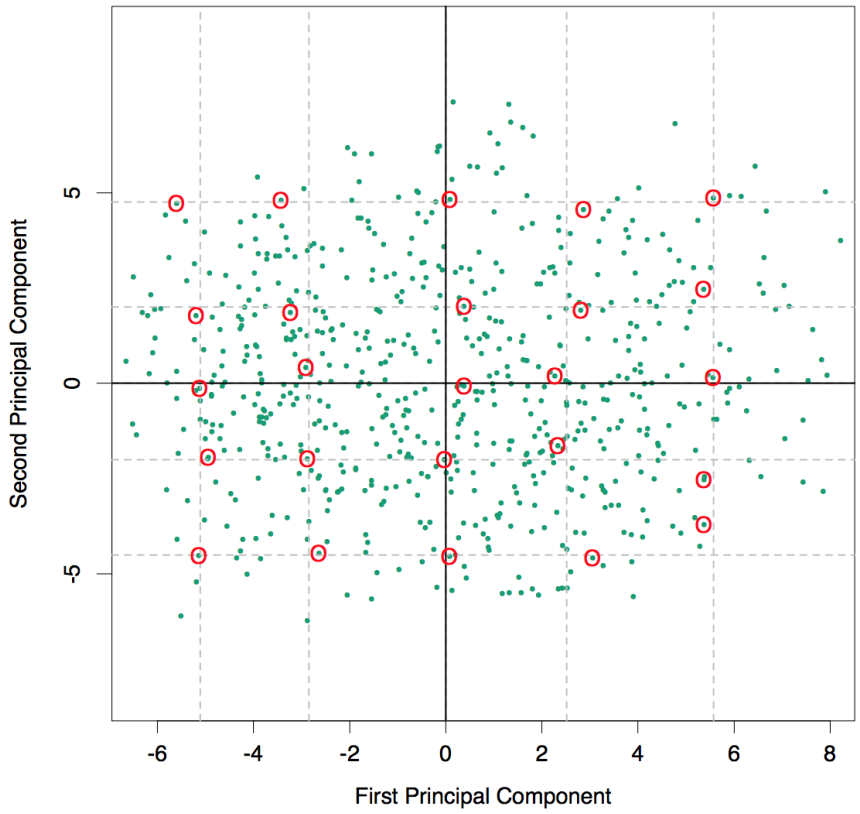


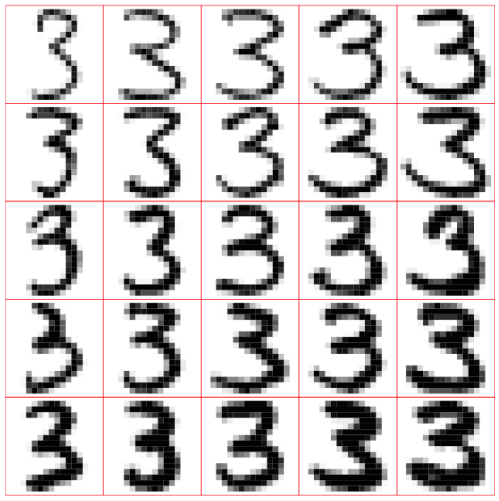
Handwritten Digits (3s) – Top 2 components

Lambda: xTv
Lambda * v = xT v * v

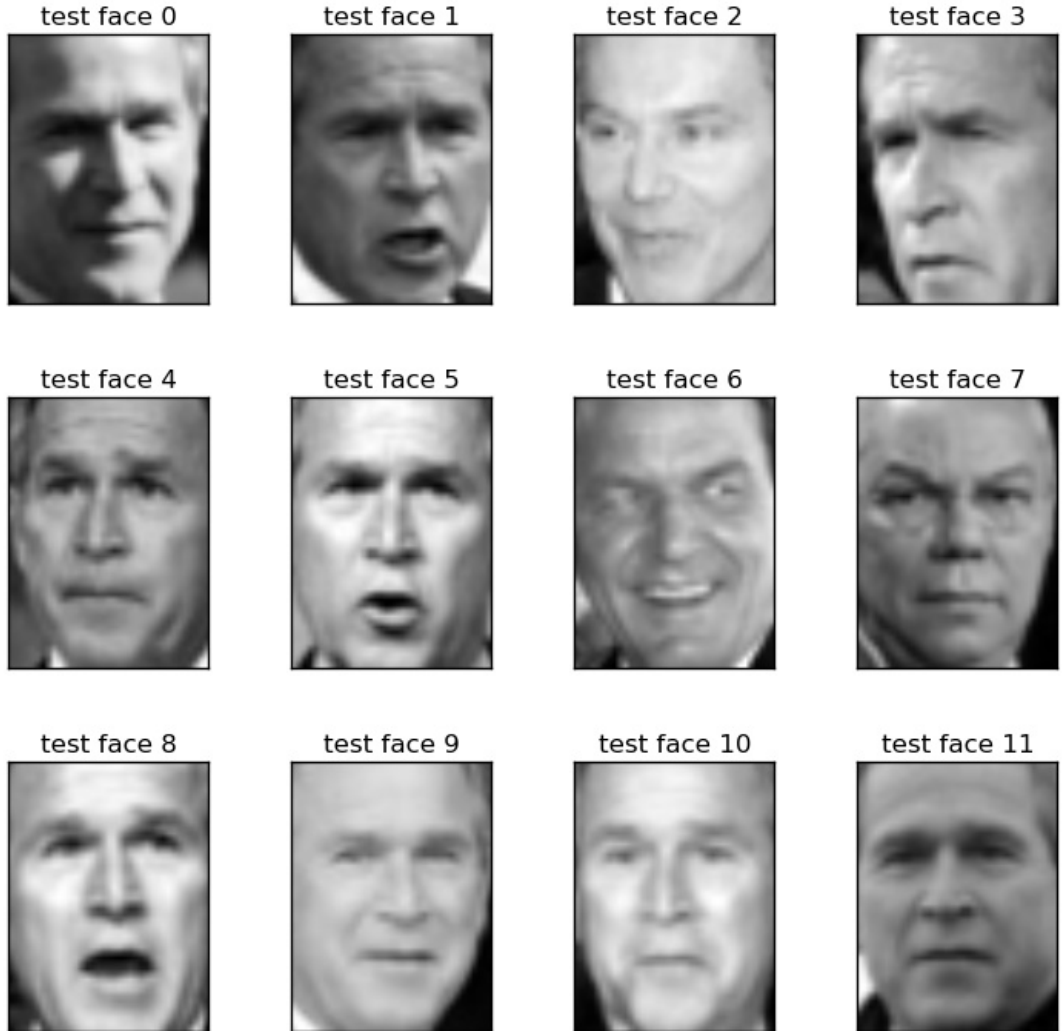
$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$$= \text{  } + \lambda_1 \cdot \text{  } + \lambda_2 \cdot \text{  } .$$

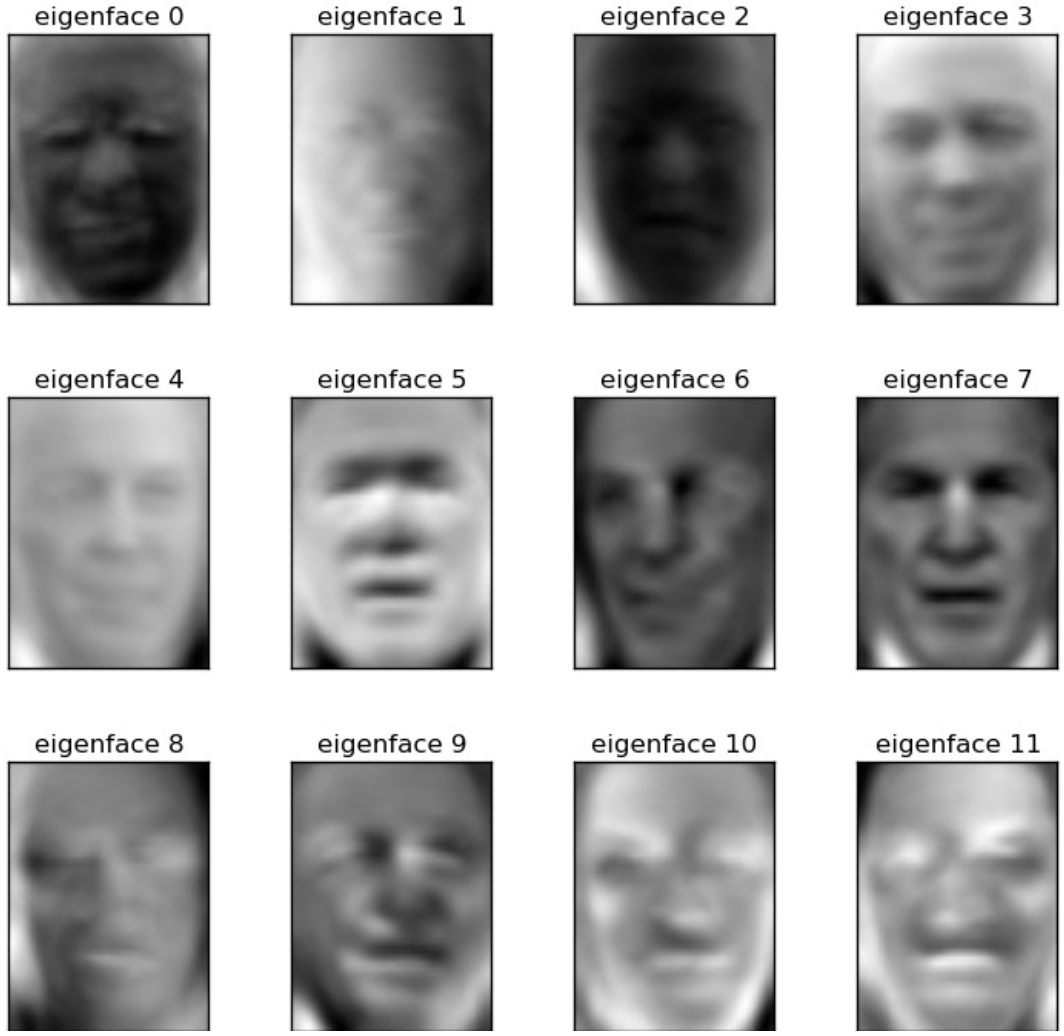




Faces


test face 0
test face 1
test face 2
test face 3
test face 4
test face 5
test face 6
test face 7
test face 8
test face 9
test face 10
test face 11

Eigenfaces


eigenface 0
eigenface 1
eigenface 2
eigenface 3
eigenface 4
eigenface 5
eigenface 6
eigenface 7
eigenface 8
eigenface 9
eigenface 10
eigenface 11

Nature. Author manuscript; available in PMC 2009 Aug 31.
Published in final edited form as:
Nature. 2008 Nov 6; 456(7218): 98–101.
Published online 2008 Aug 31. doi: 10.1038/nature07331

PMCID: PMC2735096
NIHMSID: NIHMS132060

Genes mirror geography within Europe

John Novembre,[1,2] Toby Johnson,[4,5,6] Katarzyna Bryc,[7] Zoltán Kutalik,[4,6] Adam R. Boyko,[7] Adam Auton,[7] Amit Indap,[7] Karen S. King,[8] Sven Bergmann,[4,6] Matthew R. Nelson,[8] Matthew Stephens,[2,3] and Carlos D. Bustamante[7]

Author information ▶ Copyright and License information ▶

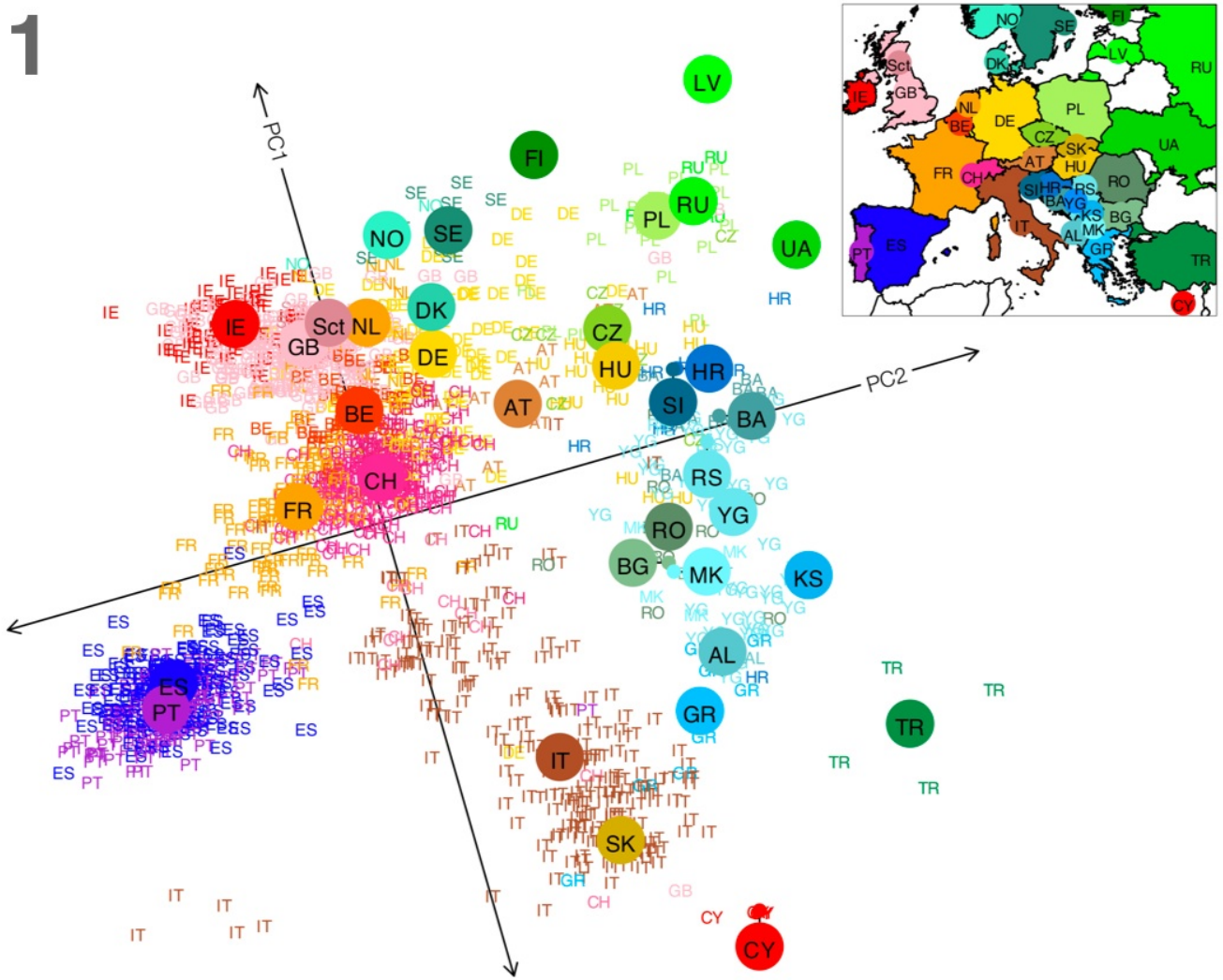
The publisher's final edited version of this article is available at Nature
This article has been corrected. See the correction in volume 456 on page 274.
See commentary "Editorial comment should accompany hot papers online." in *Nature*, volume 455 on page 861.

Abstract

Go to:

Understanding the genetic structure of human populations is of fundamental interest to medical, forensic and anthropological sciences. Advances in high-throughput genotyping technology have markedly improved our understanding of global patterns of human genetic variation and suggest the potential to use large samples to uncover variation among closely spaced populations[1–5]. Here we characterize genetic variation in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans. The results emphasize that when mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for. In addition, the results are relevant to the prospects of genetic ancestry testing[6]; an individual's DNA can be used to infer their geographic origin with surprising accuracy—often to within a few hundred kilometres.

1


PC1
PC2
LV
FI
RU
PL
NO
SE
UA
DK
IE
Sct
NL
GB
DE
CZ
HU
HR
SI
BA
AT
BE
CH
FR
RS
YG
RO
BG
MK
KS
AL
ES
PT
GR
TR
IT
SK
CY

2


PC1–PC2 space
0.03
0.02
0.01
0
UK
Germany

Where are we going?

- So far: Mostly “surface representations” & explicit features
  - Listing information for Zillow
  - Political blogs — word counts

Where are we going?

- So far: Mostly “surface representations” & explicit features
  - Listing information for Zillow
  - Political blogs — word counts
- Next: “Hidden” representations and latent variables

George Washington

State of union address excerpts

1789

Among the many interesting objects which will engage your attention that of providing for the common defense will merit particular regard. To be prepared for war is one of the most effectual means of preserving peace.

Barack Obama

The latent variable is “war and peace”

2009

PCA doesn’t pull out a semantic representation, although it does something close to it.

And for the first time, that includes the full cost of fighting in Iraq and Afghanistan. For 7 years, we have been a nation at war. No longer will we hide its price.

The elephants in the room

When we see image: it is two elephants playing with a ball. The notion of playing is a hidden variable
For a computer, it is challenging. So we need latent variables/probabilistic models

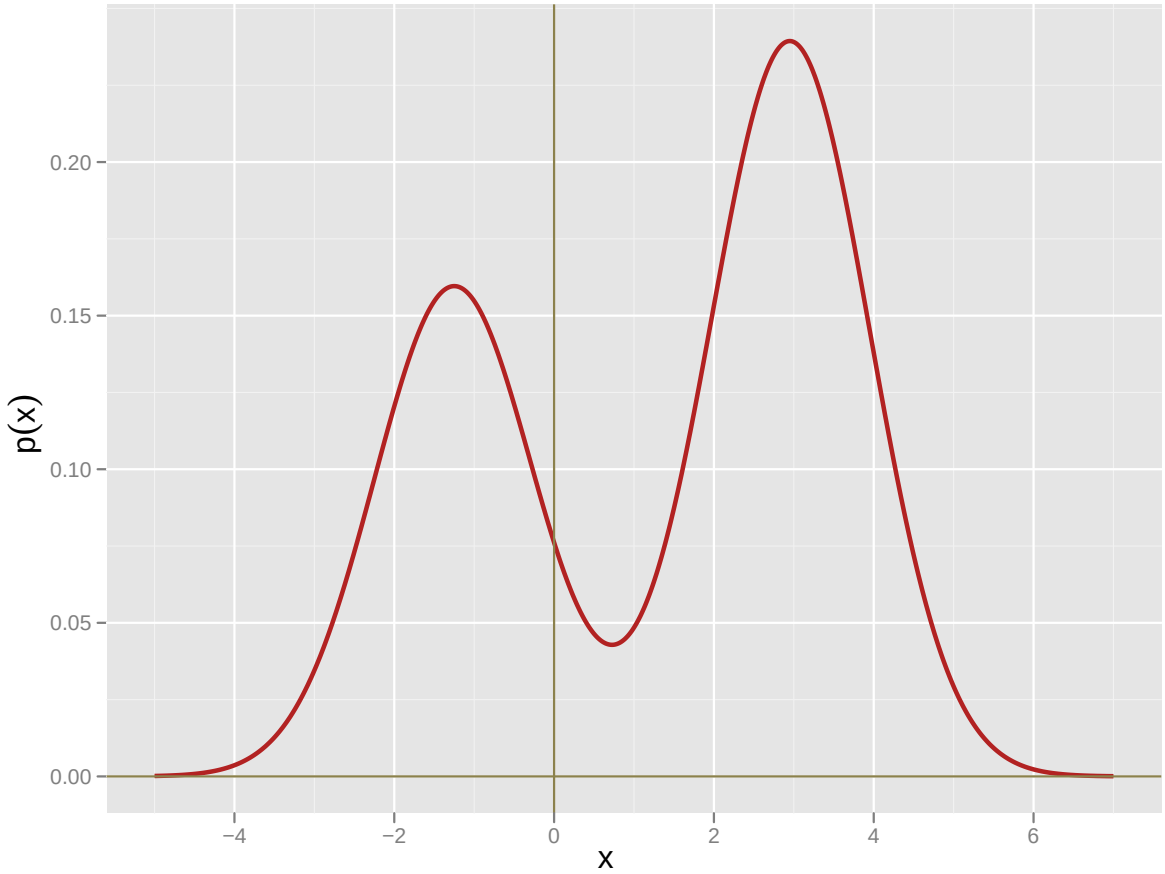
Mixtures

All latent variable models are some sort of mixture model

- Key technique: Mixture models
- Mixtures have latent variables
- Flexible tool
- Simple and difficult at the same time

Gaussian Mixture

A mixture of two gaussians



The latent variable comes from the probabilistic interpretation of this model

$$p(x) = \frac{2}{5}\phi(x; -1.25, 1) + \frac{3}{5}\phi(x; 2.95, 1)$$

The two gaussians that were mixed

Mixtures

- *Mixture of f and g*:

η is the latent variable

Or similarly, You flip a coin. If it is heads (with probability η) then sample from f(x). Else, sample from g(x). The coin flip is the latent variable, since it is unseen (you only see the final x that you got from a gaussian).

$$p(x) = \eta f(x) + (1 - \eta) g(x)$$

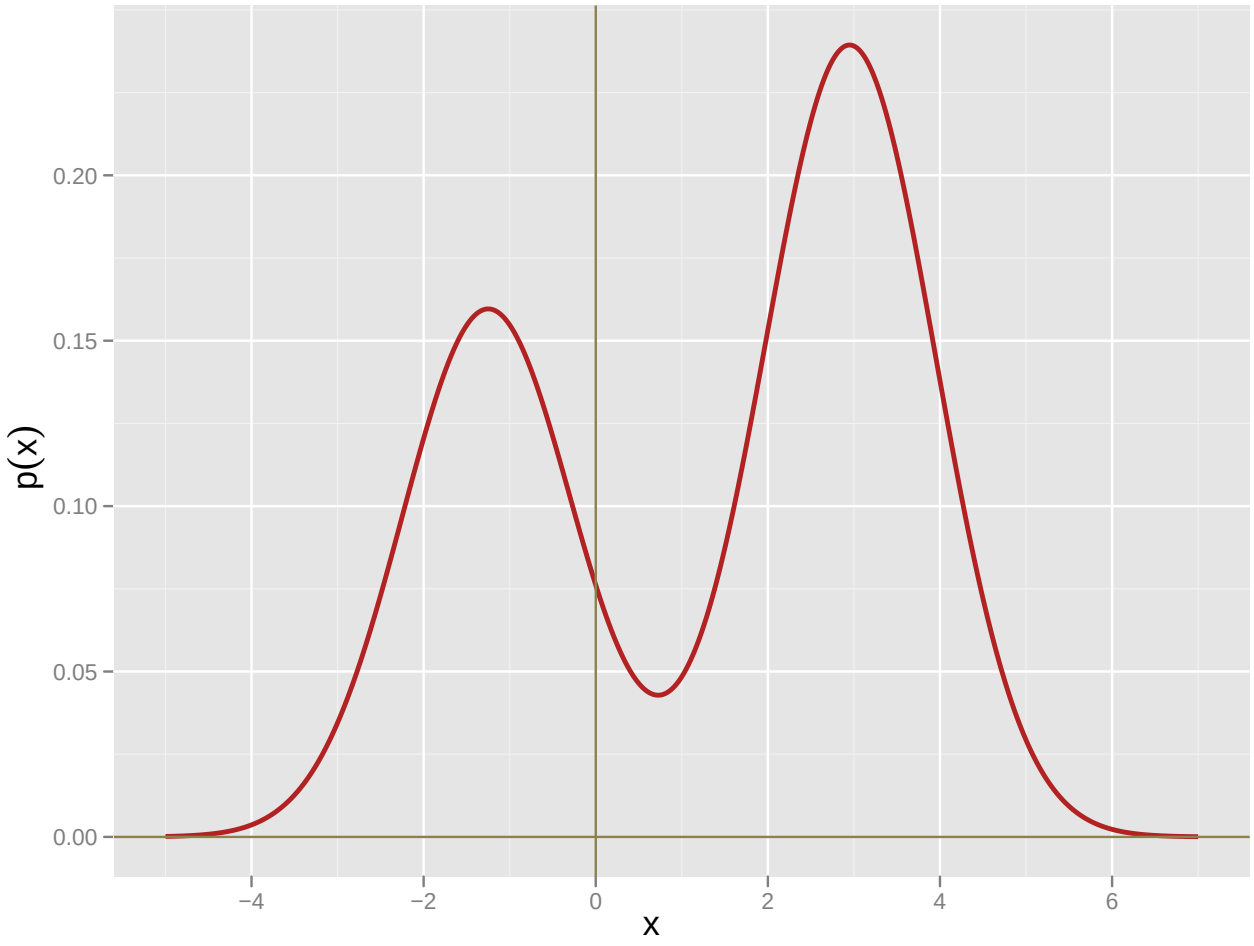
Simplest, most common kind of latent variable model

- *Hidden variable represention*: Define $Z \sim \text{Bernoulli}(\eta)$ and

$$p(x) = \sum_{z=0,1} p(x \mid z)\, p(z)$$

with $p(x \mid 0) = f(x)$, $p(x \mid 1) = g(x)$, $p(z) = \eta^z (1 - \eta)^{(1-z)}$.

Gaussian Mixture: All the Key Concepts



Bayesian Inference

Is a coin flip prob of heads / θ always 0.5? What if ridges in the coin make it change?

The parameter θ of a model is viewed as a random variable.
Inference usually carried out as follows:

Assume a distribution over your param θ i.e. no longer fixed
Flip the coin many times to gather more evidence, and update your beliefs about θ

- Choose a *generative model* $p(x \mid \theta)$ for the data.
- Choose a *prior distribution* $\pi(\theta)$ that expresses beliefs about the parameter before seeing any data.
- After observing data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, update beliefs and calculate the *posterior distribution* $p(\theta \mid \mathcal{D}_n)$.

1. Form some prior beliefs, i.e. π(θ)
2. Then, consider model: p(x|θ)
3. Compute posterior distribution over θ using Bayes rule

Bayes rule: P(A | B) = P(A & B) / P(B)
P(B | A) = P(A | B) * P(B) / P(A)

Bayes' Theorem

The posterior distribution can be written as

Posterior — Generative model / likelihood? — Prior — Proportional to

$$p(\theta \mid x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{p(x_1, \ldots, x_n)} = \frac{\mathcal{L}_n(\theta)\pi(\theta)}{c_n} \propto \mathcal{L}_n(\theta)\pi(\theta)$$

evidence

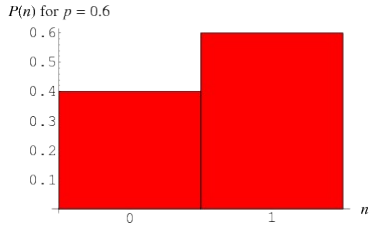
where $\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p(x_i \mid \theta)$ is the *likelihood function* and

$$c_n = p(x_1, \ldots, x_n) = \int p(x_1, \ldots, x_n \mid \theta)\pi(\theta)d\theta = \int \mathcal{L}_n(\theta)\pi(\theta)d\theta$$

is the normalizing constant, which is also called *evidence*.

Sometimes also called marginal likelihood?

Example



Assume the data comes from a Bernoulli distribution over Θ

$X \sim$ Bernoulli(θ) with data $\mathcal{D}_n = \{x_1, \ldots, x_n\}$. Prior Beta($\alpha, \beta$) distribution

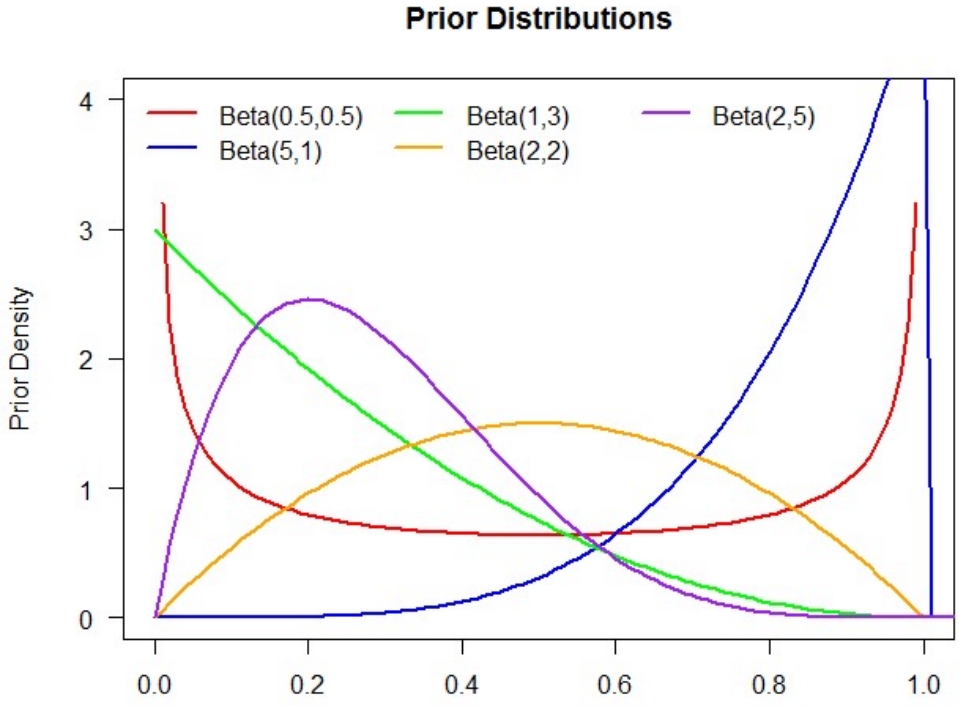
$$\pi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Let $s = \sum_{i=1}^{n} x_i$ be the number of "successes."

Posterior distribution $\theta \mid \mathcal{D}_n$ is Beta($\alpha + s, \beta + n - s$). Posterior mean is a mixture:

$$\bar{\theta} = \frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{n}{\alpha+\beta+n}\right)\widehat{\theta} + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\theta_0$$

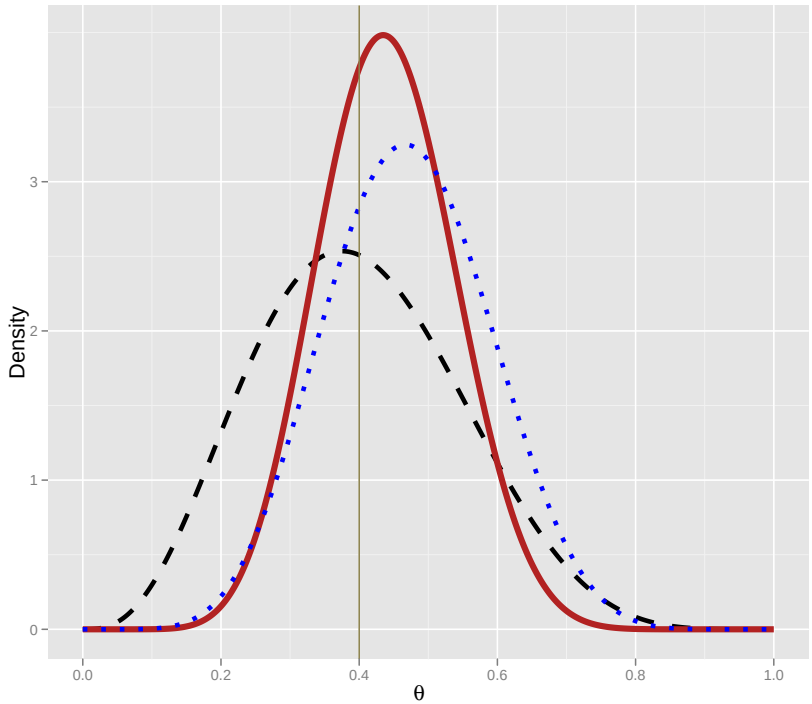
where $\widehat{\theta} = s/n$ is the MLE and $\theta_0 = \alpha/(\alpha+\beta)$ is the prior mean.



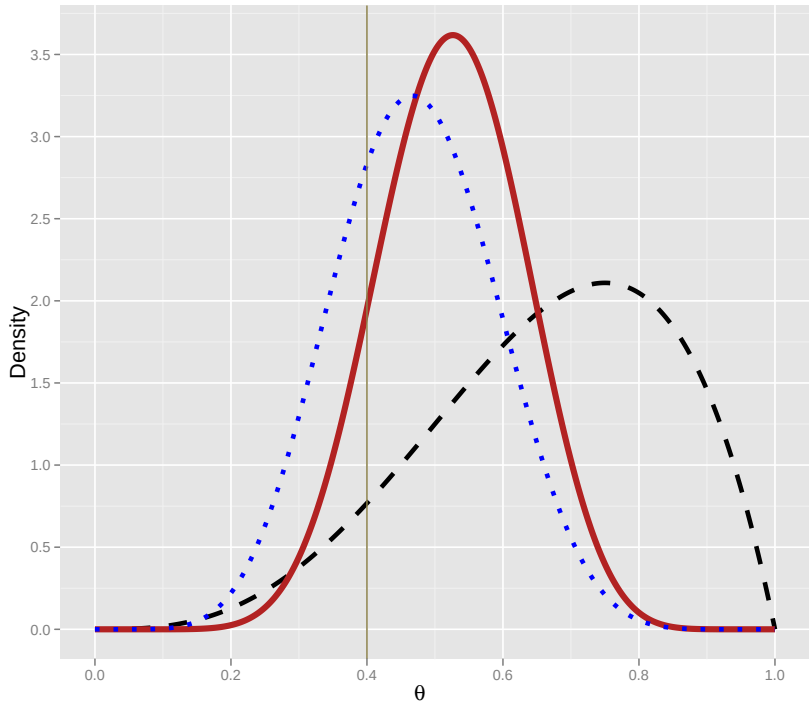
Example of beta priors

Example

$n = 15$ points sampled as $X \sim \text{Bernoulli}(\theta = 0.4)$, with $s = 7$ heads.



”good prior”



”bad prior”

Prior distribution (black-dashed), likelihood function (blue-dotted), posterior distribution (red-solid).

Dirichlet

Fancier coin flipping models i.e. rolling a die

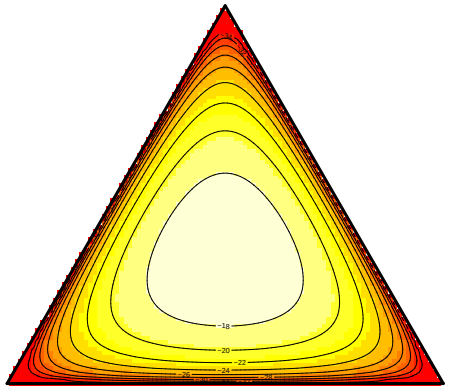
Multinomial model with Dirichlet prior is generalization of the Bernoulli/Beta model.

$$\text{Dirichlet}_\alpha(\theta) = \frac{\Gamma(\sum_{j=1}^K \alpha_j)}{\prod_{j=1}^K \Gamma(\alpha_j)} \prod_{j=1}^K \theta_j^{\alpha_j - 1}$$

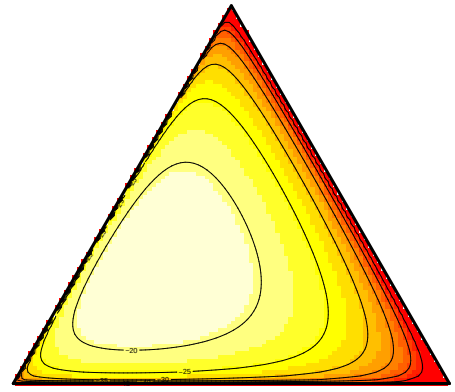
where $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}_+^K$ is a non-negative vector.

Example

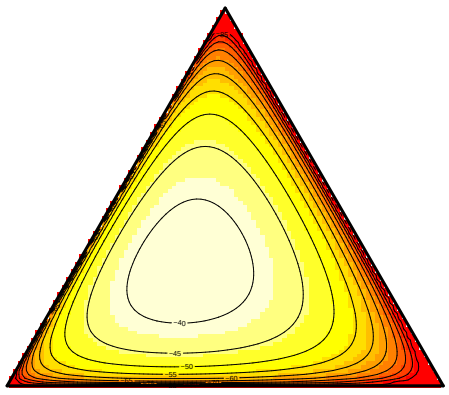
Likelihood function is just p(x | Θ)



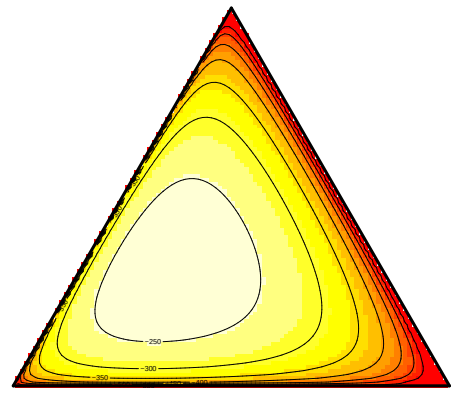
prior with Dirichlet(6,6,6)



likelihood function with $n = 20$



posterior distribution with $n = 20$



posterior distribution with $n = 200$

Reading

- Please be sure to read the detailed notes on Bayesian inference that have been posted to Canvas
- Not necessary to understand everything—but you should be able to do some of the basic “coin flipping” calculations
- We’ll need this when discussing topic models

Summary

- PCA is classical and modern—key tool for ML
- Mixtures are latent variable models
- The mixing weight encodes a hidden variable
- Computing with mixtures uses basic probabilistic reasoning