Probability models for machine learning

Advanced topics ML4bio 2016

Alan Moses

What did we cover in this course so far?

- 4 major areas of machine learning:
  - Clustering
  - Dimensionality reduction

    } unsupervised
  - Classification
  - Regression

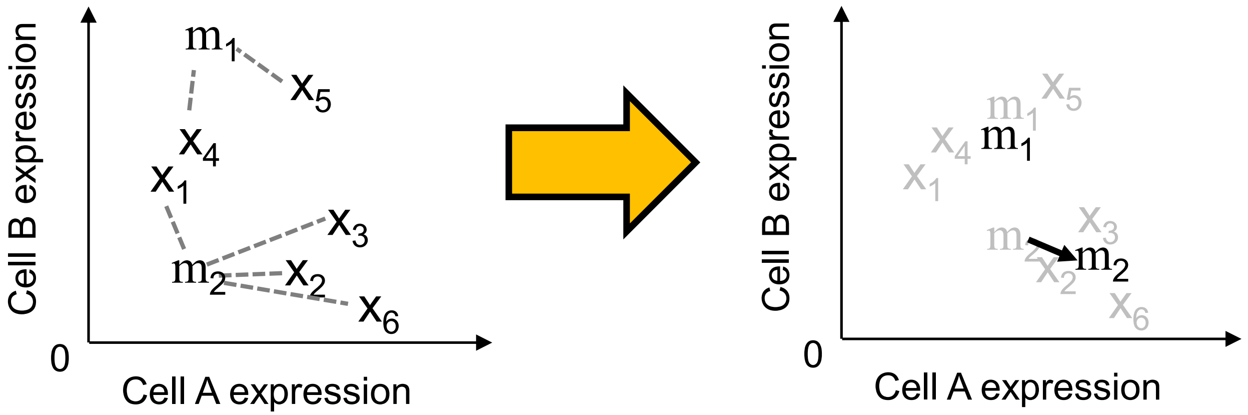
    } supervised
- Key ideas:
  - Supervised vs. unsupervised learning
  - High-dimensional data and linear algebra
  - objective functions, parameters and optimization
  - overfitting, cross-validation/held-out data, regularization

Advanced topics: the “theory” of machine learning

- What is “learning”? A taste of information theory
- Probability models for simple machine learning methods
  - What are models? Why?
  - Model-based objective functions and the connection with statistics
  - Maximum likelihood
  - Maximum a posteriori probability
  - Bayesian estimation
- Graphical models and Bayesian networks
- Derivation of E-M updates for mixture model (if time…)

What is “learning” ?

- E.g., K-means clustering



- We start “knowing nothing” and end up describing the data (pretty well in some cases)
- How much have we learned?

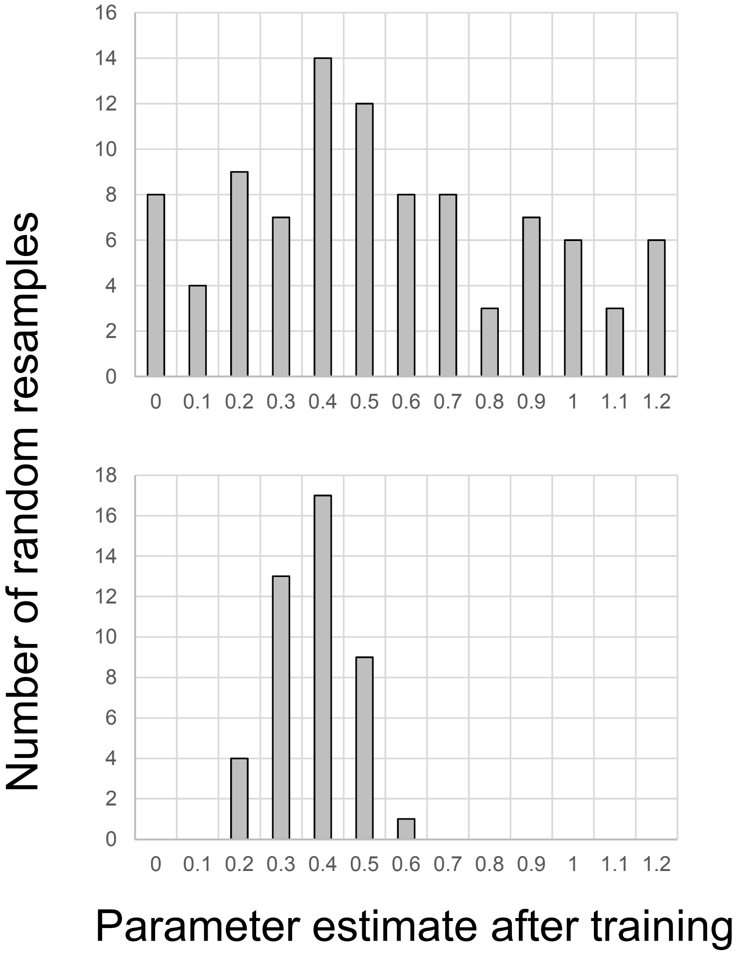
What is “learning” ?

- Models with more parameters can learn “more” – they have more information “capacity”.
  - E.g., for the same data, linear regression learns only b0 and b1, while k-means learns 4 parameters.
  - Geoff Hinton: Mouse brain has 10^7 neurons, and 10^11 connections. Need to train computational models with (at least) billions of parameters
- We need a lot of data to learn all these parameters!
- Does a model with more parameters necessarily learn more?

What is “learning” ?

- Models with more parameters can learn “more” – they have more information “capacity”.
- Not all the parameters are actually measuring something useful about our data
- The confidence/certainty/variability of the parameter estimates is important

  The amount that the model has actually learned is something like the difference between filling all the parameters with random guesses vs. the parameters we estimate from data
- How do we quantify this?



Comparing two models that can only learn 1 parameter.

Which model has “learned” more?

What is “learning” ?

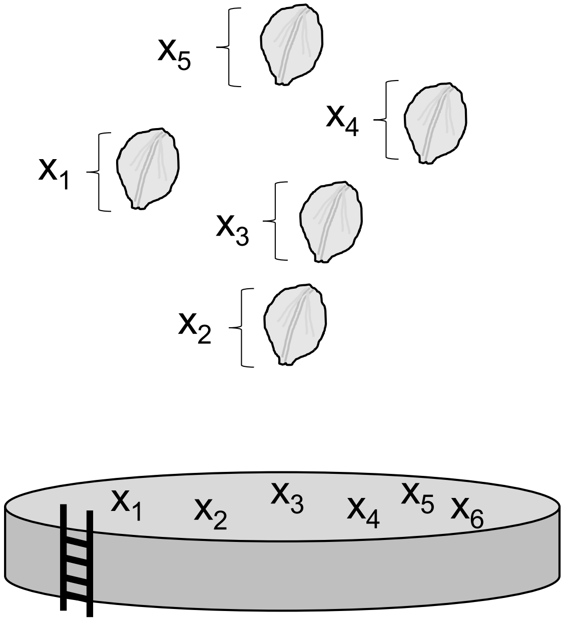
- Models with more parameters can learn “more” – they have more information “capacity”
- The amount actually learned is a comparison between the information you have before and after learning
- Information theory quantifies both the information capacity and the amount of information actually stored using mathematics
- Two ways to think about it:
  - Minimum description length/data compression
  - Analogy with statistical physics/entropy

Advanced topics: the “theory” of machine learning

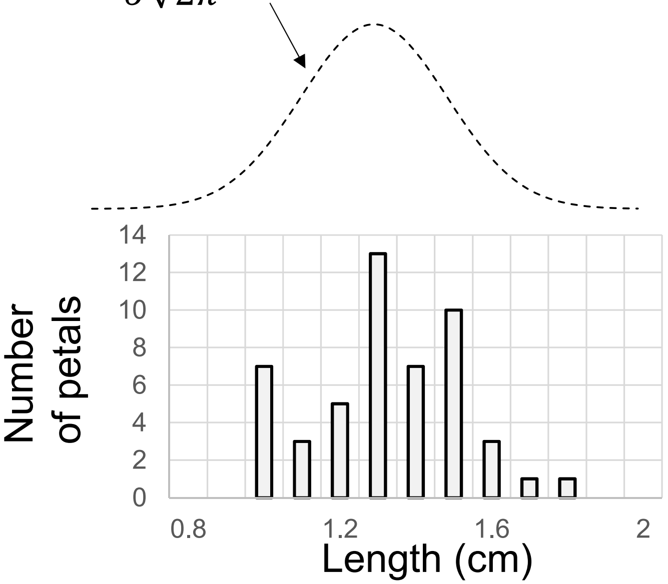
- What is “learning”? A taste of information theory
- Probability models for simple machine learning methods
  - What are models? Why?
  - Model-based objective functions and the connection with statistics
  - Maximum likelihood
  - Maximum a posteriori probability
  - Bayesian estimation
- Graphical models and Bayesian networks
- Derivation of E-M updates for mixture model (if time…)

Probabilistic models

- E.g., Measuring the size of Iris petals

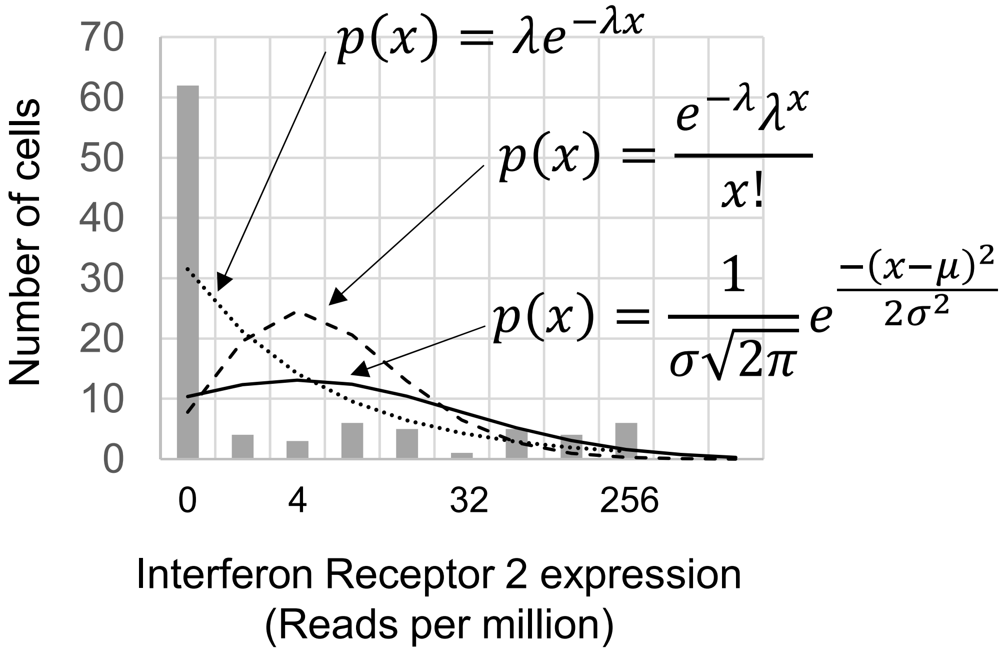


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Probabilistic models

- E.g., single cell sequence data



$$p(x) = \lambda e^{-\lambda x}$$

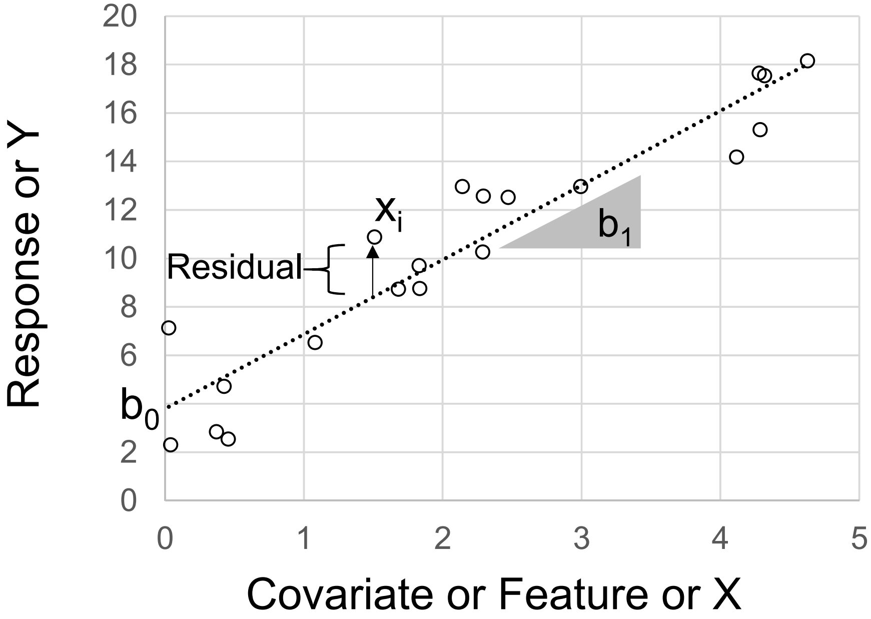
$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

E.g., linear regression

$$Y = b_0 + b_1 X$$

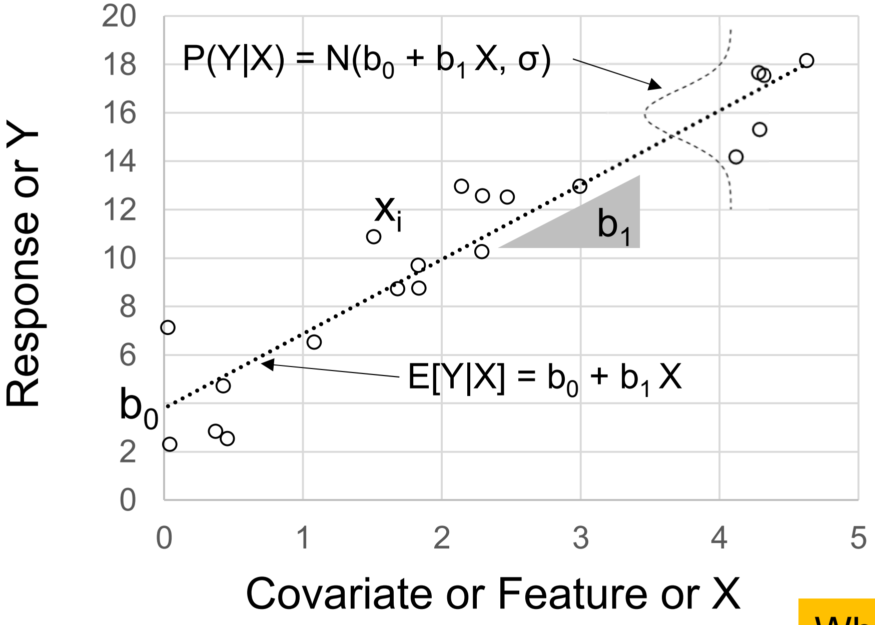
As we usually think about it:



E.g., linear regression

$Y = b_0 + b_1 X$

As a probabilistic model:

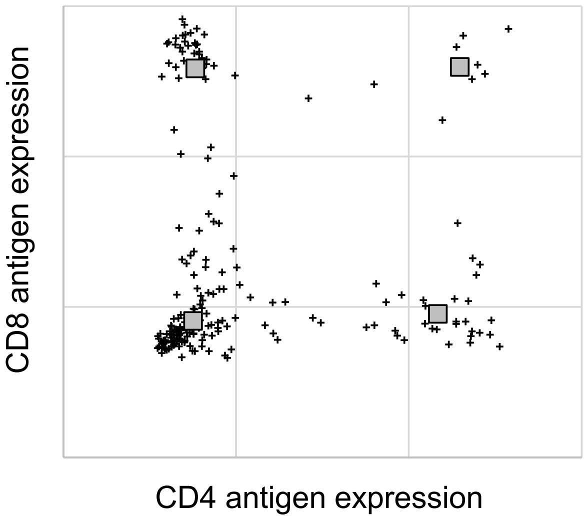


When has the regression model "learned" "a lot" ?

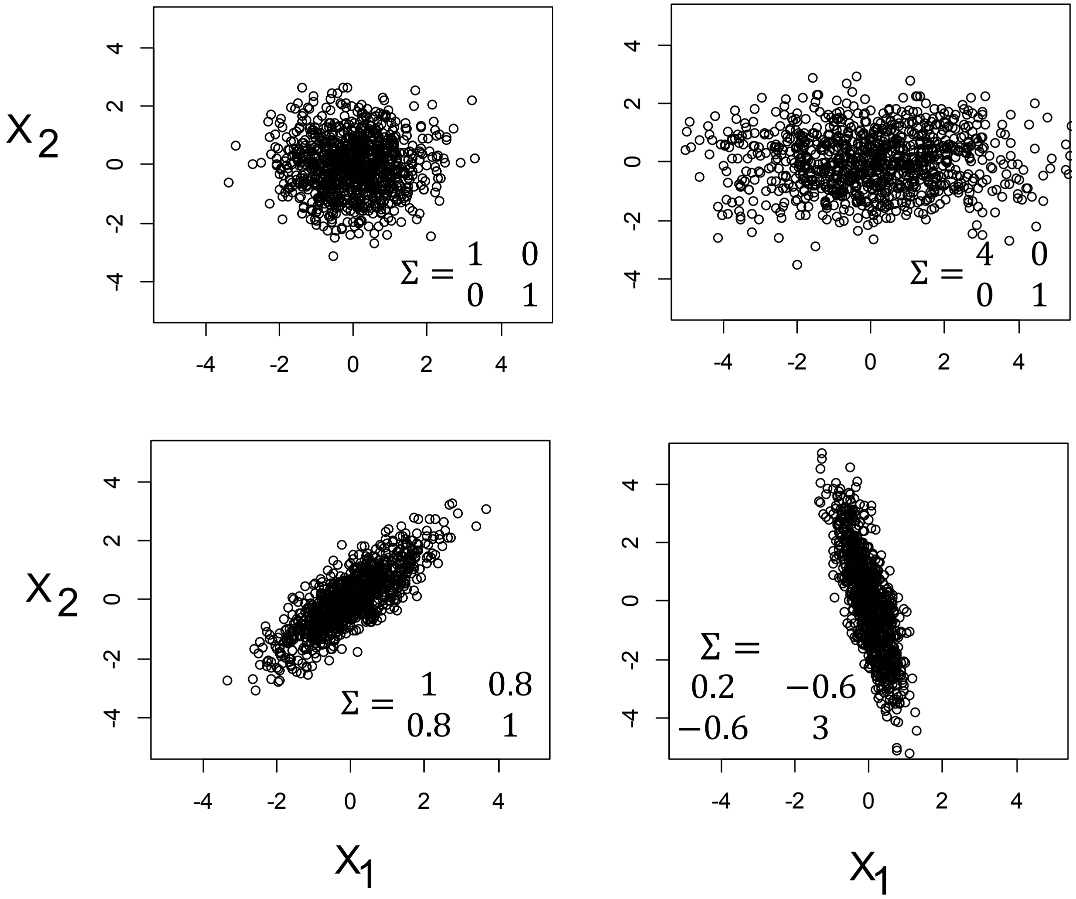
Objective functions and probabilistic models

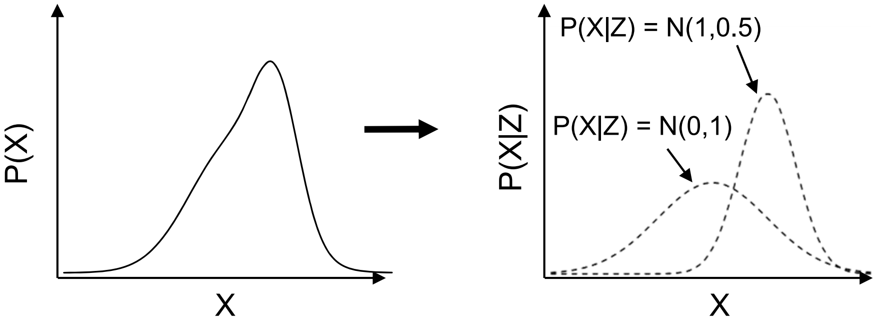
- Probabilistic models still have objective functions that depend on parameters, but…
- Parameters can now be interpreted as “mean”, “standard deviation”, etc
    - This can give us insight into the biology
    - Allow us to test hypotheses
- When we optimize an objective function based on probability theory, we are doing statistical “estimation” of the model parameters
    - In statistics jargon, the values of the parameters that optimize the objective function are called “estimators”

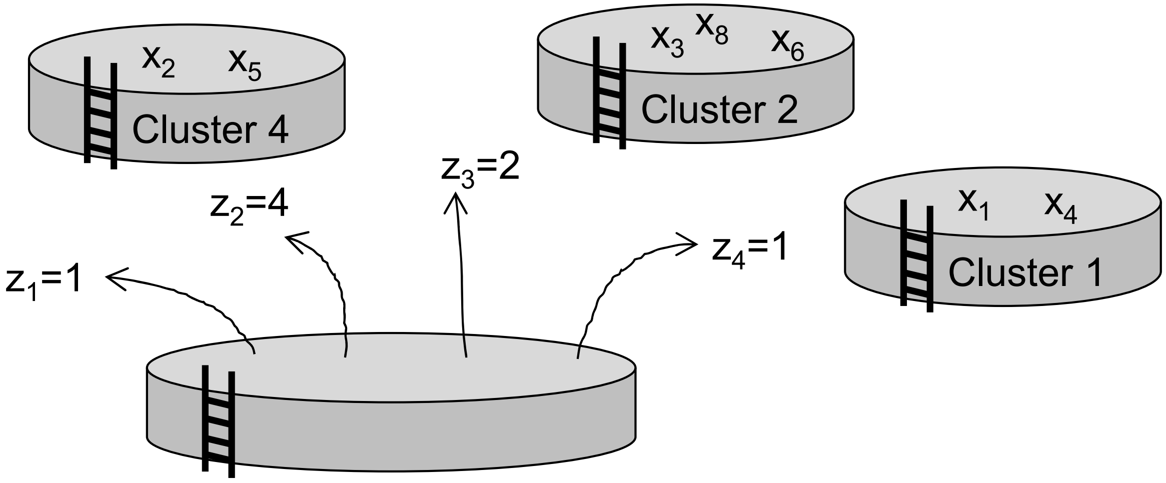
E.g., clustering

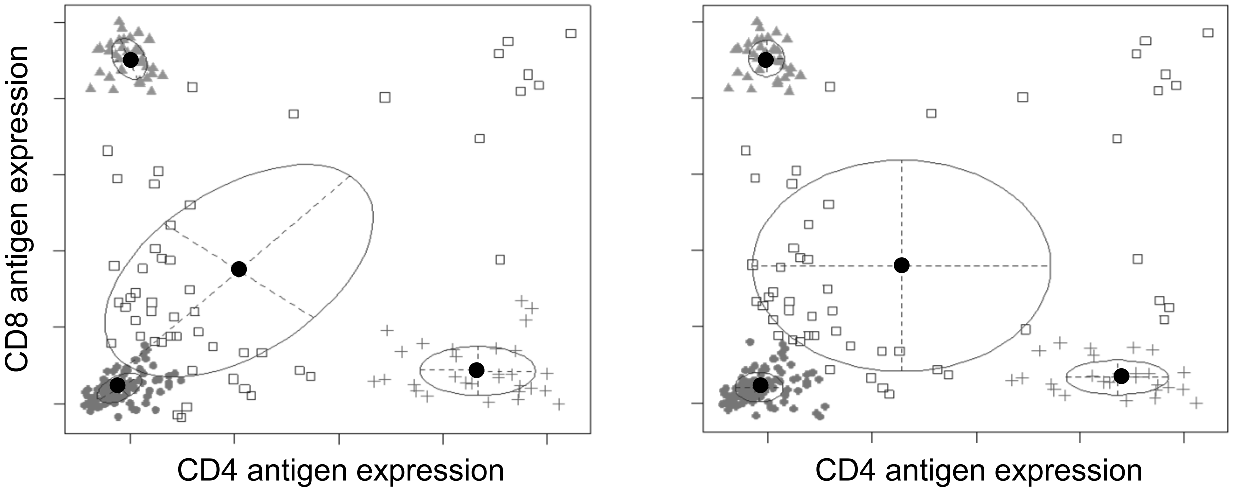


$$N\left(\vec{X}\middle|\vec{\mu},\mathbf{\Sigma}\right) = \frac{1}{\sqrt{|\mathbf{\Sigma}|(2\pi)^d}}\, e^{-\frac{1}{2}\left(\vec{X}-\vec{\mu}\right)^T\mathbf{\Sigma}^{-1}\left(\vec{X}-\vec{\mu}\right)}$$




P(X)
X
P(X|Z) = N(1,0.5)
P(X|Z) = N(0,1)
P(X|Z)
X


x_2 x_5
Cluster 4
x_3 x_8 x_6
Cluster 2
x_1 x_4
Cluster 1
z_3=2
z_2=4
z_4=1
z_1=1


CD8 antigen expression
CD4 antigen expression
CD4 antigen expression

likelihood

- Most famous and widely used objective function

Likelihood = P(data|model)

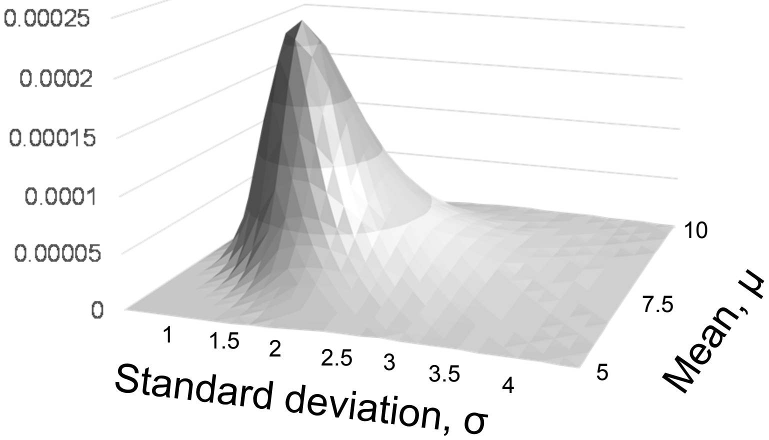
- Assuming observations are independent, the likelihood is just the product of the probabilities of the observations. Why?
- When you optimize it, you reach maximum likelihood and the parameters are “maximum likelihood estimates” or MLEs

Observation (i)	Value (Xi)
1	5.2
2	9.1
3	8.2
4	7.3
5	7.8

Likelihood

of this data under a Gaussian model

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



likelihood

- In practice we use the log likelihood. Why?
- Often, if we assume the residuals (or errors) follow a Gaussian distribution, $p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, the log likelihood is equivalent to the sum of squared residuals objective function. Why?
- Models where the objective function is the likelihood can be regularized using AIC (and compared using LRTs*)
- Models that can be formulated with likelihood as the objective function include: Mixture Models, Naïve Bayes, logistic regression, HMMs

*sometimes

MAP estimation

- Objective function is the a posteriori probability

  P(model|data)

- Which turns out to be equal to $P(data|model)\frac{P(model)}{P(data)}$
- Equivalent to log likelihood plus a term that depends only on the parameters and a term that doesn't depend on the parameters

$$\log(P(model|data)) = \log(Likelihood) + \log\big(P(model)\big) - \log\big(P(data)\big)$$

- This is a "penalty" or regularization!

MAP estimation

$$\log(P(model|data)) = \log(Likelihood) + \log\big(P(model)\big) - \log\big(P(data)\big)$$

- P(model) is known as the “prior” probability and expresses what we know about the parameters before we see the data
- Parameters in the prior probability distributions are the hyperparameters – usually there are few, but they are hard to estimate
- What kinds of distributions do we want for priors?

Lasso regression is MAP estimation with an exponential prior

Bayesian statistics

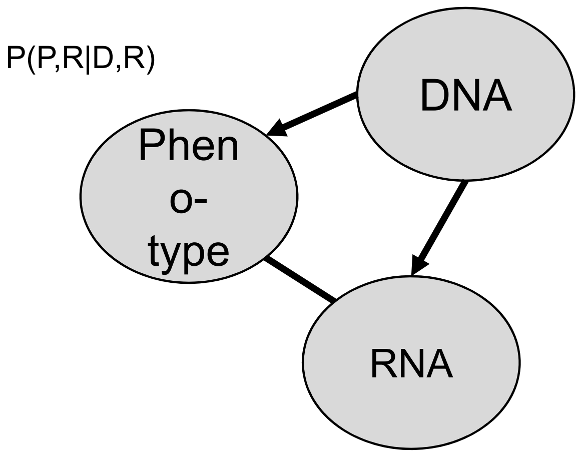
- Don’t bother trying find a single set of parameters that optimizes the objective function
    - Don’t believe in “parameters” or “estimators” – everything is random!

- Try to directly estimate the distribution of the parameters after observing the data (either analytically or by sampling)

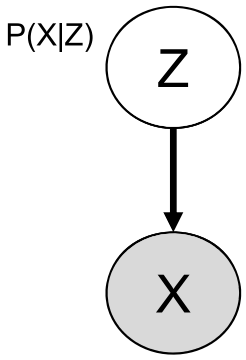
- Not used very often in molecular biology. Why?

Graphical Models and Bayesian Networks

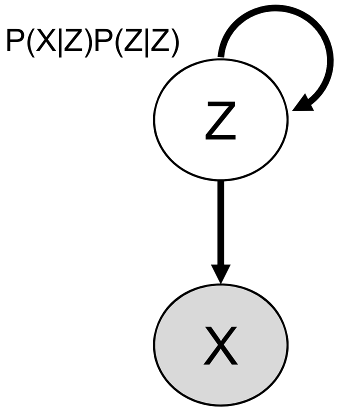
- Represent the dependence of variables as a “graph” where the edges represent statistical dependence between variables.
- Only represents the structure of the model, not the distributions



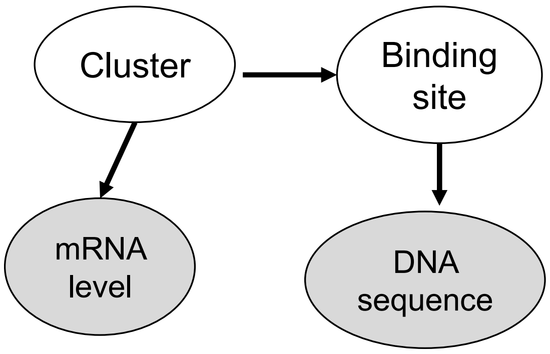
eQTL data when you know phenotypes



mixture model



HMM



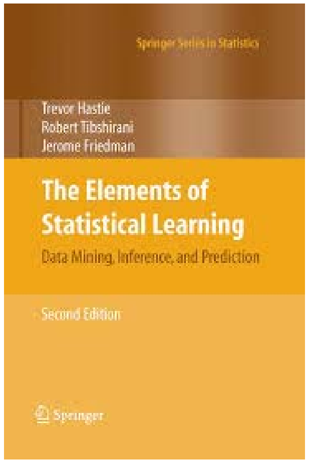
Holmes & Bruno sequence and expression clusters

Graphical Models and Bayesian Networks

- Represent the dependence of variables as a “graph” where the edges represent statistical dependence between variables.
- If you can represent your model as a “directed, acyclic graph” or DAG then you have a Bayesian network.
- Powerful algorithms can be developed if the graph has this structure, and the details of the structure of the graph determine the performance of the algorithms

Interested in statistics and probability? (and the connection to machine learning)

- I wrote an introductory textbook about this aimed at molecular biology graduate students
- I'm happy to email you a recent draft!
- More advanced books include:



Springer Series in Statistics

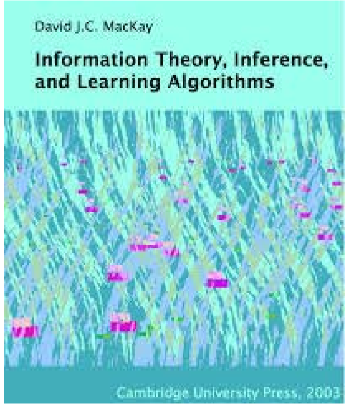
Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer



David J.C. MacKay

Information Theory, Inference, and Learning Algorithms

Cambridge University Press, 2003

Available free online!