Mathematical Tools
for Neural and Cognitive Science

Fall semester, 2018

Probability & Statistics: Estimation, inference, model-fitting

1

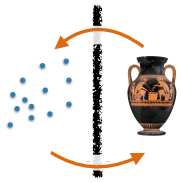
Estimation of model parameters (outline)

- How do I compute an estimate?
  (mathematics vs. numerical optimization)
- How "good" are my estimates?
  (classical stats vs. simulation vs. resampling)
- How well does my model explain the data?
  Future data (prediction/generalization)?
  (classical stats vs. resampling)
- How do I compare two (or more) models?
  (classical stats vs. resampling)

2

The sample average

$$a(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N} x_n$$



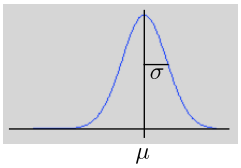
- Most common common form of estimator
- Value of *a* converges to true mean *E(x),* for all reasonable distributions
- Variance of *a* converges to zero, as σ^2/N
- Distribution *p(a)* converges to a Gaussian
  (the "Central Limit Theorem")

3

4

The Gaussian

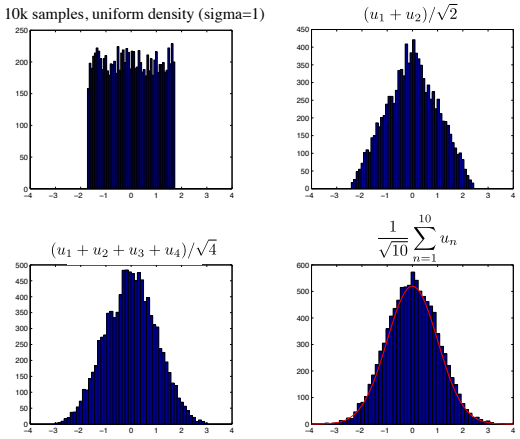
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- parameterized by mean and SD (position / width)
- product of two Gaussians is Gaussian! [easy]
- sum of Gaussian RVs is Gaussian! [moderate]
- central limit theorem: sum of many RVs is Gaussian! [hard]

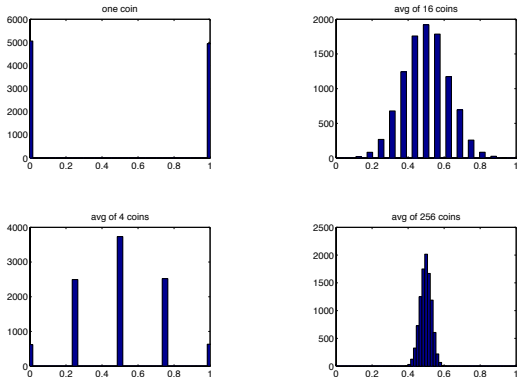
5

Central limit for a uniform distribution...

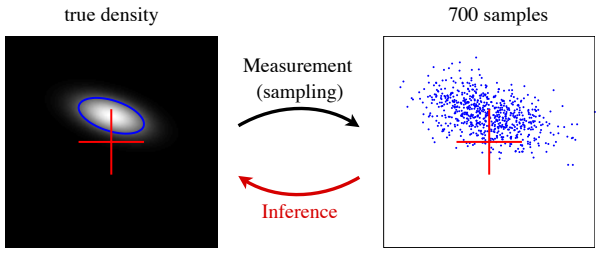


6

Central limit for a binary distribution...



7



true mean: [0 0.8]
true cov: [1.0 -0.25
-0.25 0.3]

sample mean: [-0.05 0.83]
sample cov: [0.95 -0.23
-0.23 0.29]

8

Point Estimates

- Estimator: Any function of the data, intended to provide an estimate of the true value of a parameter
- Statistically-motivated estimators:
  - Maximum likelihood (ML): $\hat{x}(\vec{d}) = \arg\max_x p(\vec{d}|x)$
  - Max a posteriori (MAP): $\hat{x}(\vec{d}) = \arg\max_x p(x|\vec{d})$
  - Bayes estimator: $\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbf{E}\left(L(x-\hat{x})|\vec{d}\right)$
  - Bayes least squares: (special case) $\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbf{E}\left((x-\hat{x})^2|\vec{d}\right) = \mathbf{E}\left(x|\vec{d}\right)$

9

Estimator quality: Bias & Variance

- Mean squared error = bias^2 + variance
- Bias is difficult to assess (requires knowing the "true" value). Variance is easier.
- Classical statistics generally aims for an unbiased estimator, with minimal variance ("MVUE").
- The MLE is *asymptotically* unbiased (under fairly general conditions), but this is only useful if
  - the likelihood model is correct
  - the optimum can be computed
  - you have lots of data
- More general view: estimation is about trading off bias and variance, through model selection, "regularization", or Bayesian priors…

ML Estimates - discrete

- Binomial: $p\left(n_{head} \mid m, p_{head}\right) = \left(\begin{array}{c} m \\ n \end{array} \right) p_{head}^{n} \left(1 - p_{head}\right)^{m-n}$

  $\hat{p}_{head} = \frac{n}{m}$

- Poisson: $p\left(k \mid \lambda\right) = \frac{\lambda^k e^{-k}}{k!}$

  $\hat{\lambda} = k$

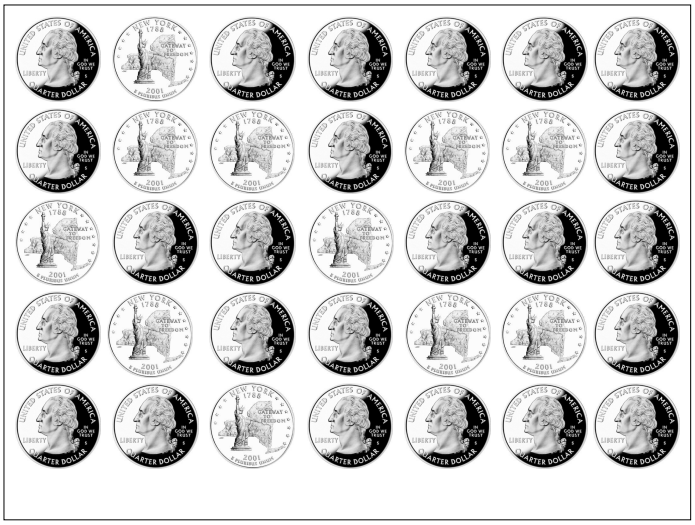
ML Estimates - continuous

The *N* independent samples are $x_1, x_2, \cdots x_N$

ML estimates are $\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N}$ biased!

Example: Estimate the bias of a coin



13

Bayes' Rule and Estimation

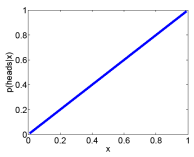
Posterior Likelihood Prior

$$p(\text{parameter value} \mid \text{data}) = \frac{p(\text{data} \mid \text{parameter value})\, p(\text{parameter value})}{p(\text{data})}$$

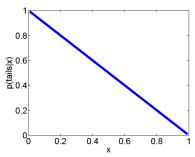
Nuisance normalizing term

14

Likelihood: 1 head



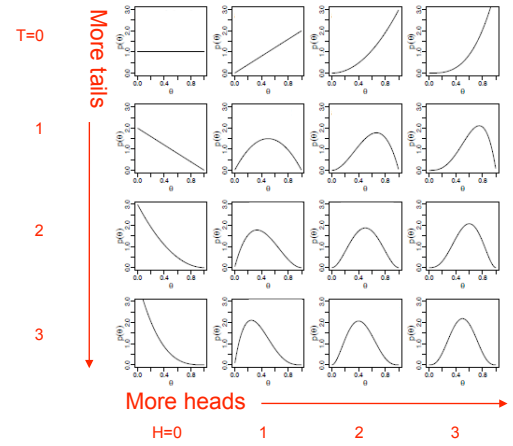
Likelihood: 1 tail



15

16

Posteriors, p(H,T|x), assuming prior p(x)=1



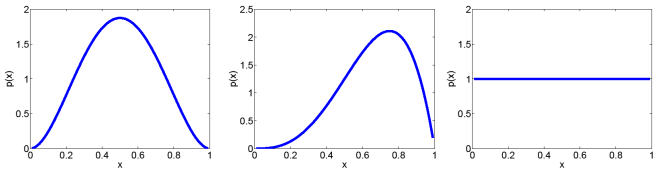
17

example

infer whether a coin is fair by flipping it repeatedly
here, x is the probability of heads (50% is fair)
$y_{1...n}$ are the outcomes of flips

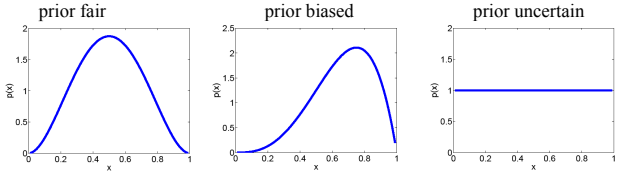
Consider three different priors:

suspect fair | suspect biased | no idea

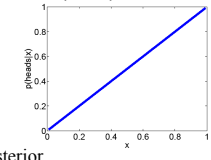


18

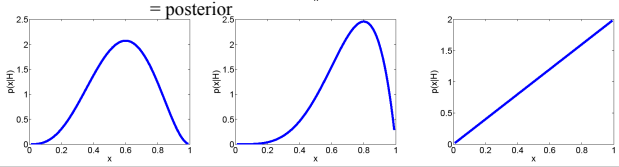
prior fair | prior biased | prior uncertain



X likelihood (heads)

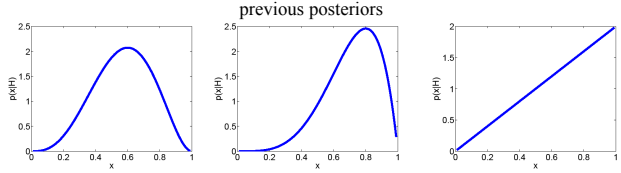


= posterior

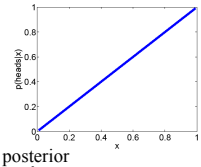


19

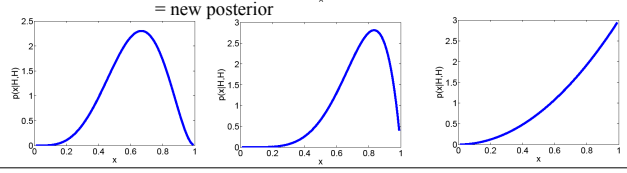
previous posteriors



X likelihood (heads)

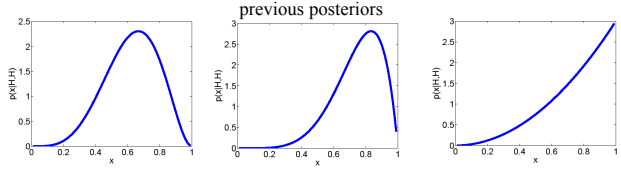


= new posterior

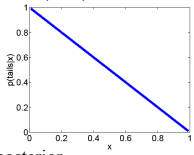


20

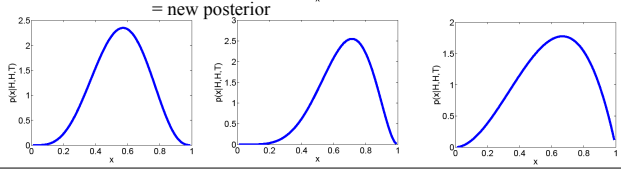
previous posteriors



X likelihood (tails)

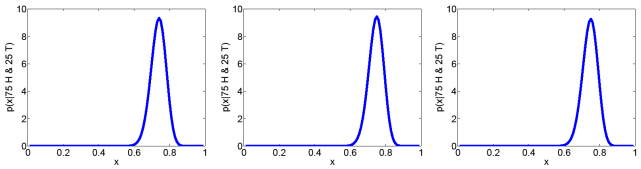


= new posterior



21

Posteriors after observing 75 heads, 25 tails

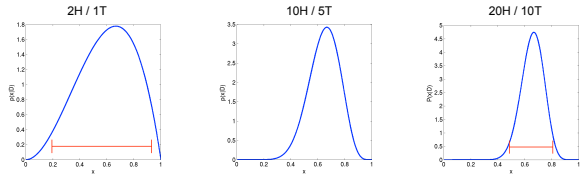


→prior differences are ultimately overwhelmed by data

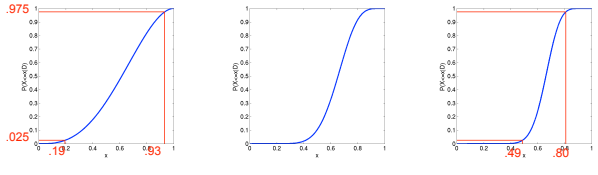
22

Confidence intervals

PDFs

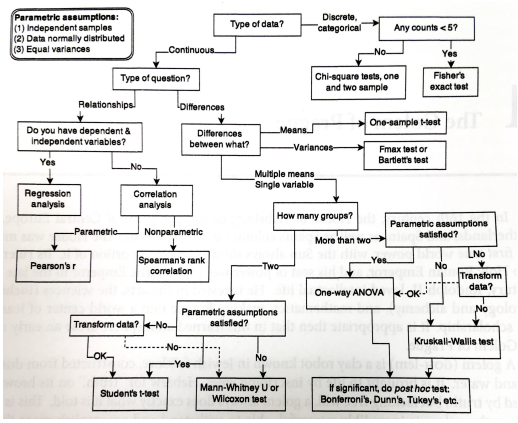


CDFs, and 95% confidence intervals



23

Classical "frequentist" statistical tests

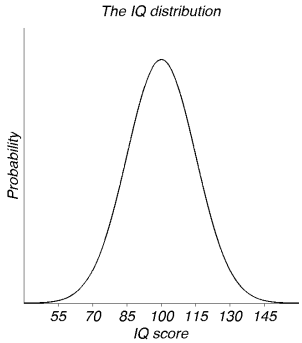


Statistical Rethinking, Richard McElreath

24

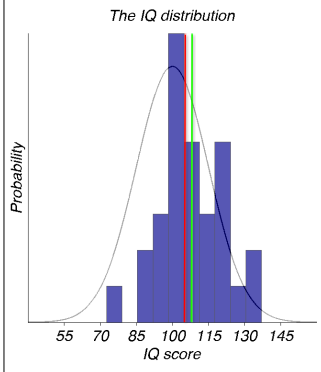
Classical/frequentist approach - *z*

- H_1: NZT improves IQ
- Null: H_0: it does nothing
- In the general population, IQ is known to be distributed normally with
- $\mu = 100$
- $\sigma = 15$
- We give the drug to 30 people and test their IQ.



25

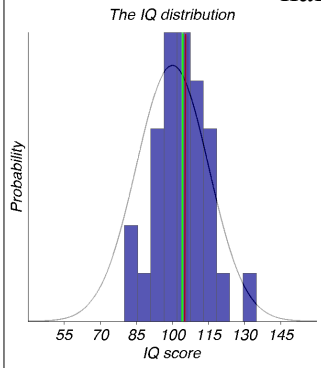
The *z*-test



- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $x = 108.3$ (Sample mean)
- $\bar{z} = (x - \mu)/\text{SE} = (108.3\text{-}100)/\text{SE}$ (Standardized score)
- $\text{SE} = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- Error bar/CI: ±2 SE
- $z = 8.3/2.74 = 3.03$
- $p = 0.0012$
- Significant?
- One- vs. two-tailed test

26

What if the measured effect of NZT had been half that?



- $\mu = 100$ (Population mean)
- $\sigma = 15$ (Population standard deviation)
- $N = 30$ (Sample contains scores from 30 participants)
- $\bar{x} = 104.2$ (Sample mean)
- $z = (\bar{x} - \mu)/\text{SE} = (104.2\text{-}100)/\text{SE}$
- $\text{SE} = \sigma / \sqrt{N} = 15/\sqrt{30} = 2.74$
- $z = 4.2/2.74 = 1.53$
- $p = 0.061$
- Significant?

27

Significance levels

- Are denoted by the Greek letter α.
- In principle, we can pick anything that we consider unlikely.
- In practice, the consensus is that a level of 0.05 or 1 in 20 is considered as unlikely enough to reject H_0 and accept the alternative.
- A level of 0.01 or 1 in 100 is considered “highly significant” or really unlikely.

28

Does NZT improve IQ scores or not?

Significant?	Reality: Yes	Reality: No
Yes	Correct	Type I error α-error False alarm
No	Type II error β-error Miss	Correct

29

Test statistic

- We calculate how far the observed value of the sample average is away from its expected value.
- In units of standard error.
- In this case, the test statistic is

$$z = \frac{\bar{x} - \mu}{SE} = \frac{\bar{x} - \mu}{\sigma / \sqrt{N}}$$

- Compare to a distribution, in this case z or $N(0,1)$

30

Common misconceptions

Is "Statistically significant" a synonym for:
- Substantial
- Important
- Big
- Real

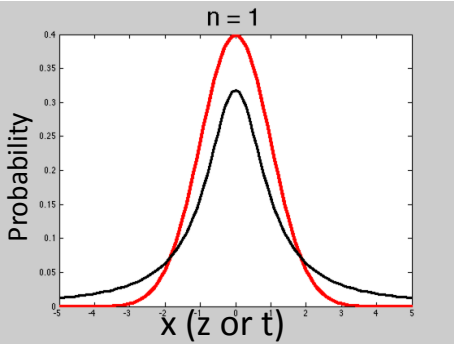
Does statistical significance gives the
- probability that the null hypothesis is true
- probability that the null hypothesis is false
- probability that the alternative hypothesis is true
- probability that the alternative hypothesis is false

Meaning of p-value. Meaning of CI.

Student's *t*-test

- σ not assumed known
- Use $s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$
- Why *N*-1? *s* is unbiased (unlike ML version), i.e., $E(s^2) = \sigma^2$
- Test statistic is $t = \frac{\bar{x} - \mu_0}{s/\sqrt{N}}$
- Compare to *t* distribution for CIs and NHST
- "Degrees of freedom" reduced by 1 to *N*-1

The *t* distribution approaches the normal distribution for large *N*



The *z*-test for binomial data

- Is the coin fair?
- Lean on central limit theorem
- Sample is *n* heads out of *m* tosses
- Sample mean: $\hat{p} = n/m$
- H_0: $p = 0.5$
- Binomial variability (one toss): $\sigma = \sqrt{pq}$, where $q = 1 - p$
- Test statistic: $z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / m}}$
- Compare to *z* (standard normal)
- For CI, use $\pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/m}$

Many varieties of frequentist univariate tests

- χ^2 goodness of fit
- χ^2 test of independence
- test a variance using χ^2
- F to compare variances (as a ratio)
- Nonparametric tests (e.g., sign, rank-order, etc.)

Bootstrapping

- "The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps"
  [Adventures of Baron von Munchausen, by Rudolph Erich Raspe]
- A **(re)sampling** method for computing estimator distribution (incl. stdev error bars or confidence intervals)
- Idea: instead of running experiment multiple times, resample (with replacement) from the *existing* data. Compute an estimate from each of these "bootstrapped" data sets.

HEART ATTACK RISK FOUND TO BE CUT BY TAKING ASPIRIN

LIFESAVING EFFECTS SEEN

Study Finds Benefit of Tablet Every Other Day Is Much Greater Than Expected

[New York Times, 27 Jan 1987]

The summary statistics in the newspaper article are very simple:

	heart attacks (fatal plus non-fatal)	subjects
aspirin group:	104	11037
placebo group:	189	11034

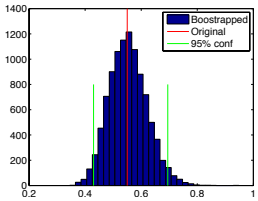
$$\widehat{\theta} = \frac{104/11037}{189/11034} = .55. \qquad (1.1)$$

If this study can be believed, and its solid design makes it very believable, the aspirin-takers only have 55% as many heart attacks as placebo-takers.

Of course we are not really interested in $\widehat{\theta}$, the estimated ratio. What we would like to know is θ, the true ratio

[Efron & Tibshirani '98]

Histogram of bootstrap estimates:



=> with 95% confidence,

$0.43 < \theta < 0.7$

37

	strokes	subjects	
aspirin group:	119	11037	
placebo group:	98	11034	(1.3)

For strokes, the ratio of rates is

$$\hat{\theta} = \frac{119/11037}{98/11034} = 1.21. \quad (1.4)$$

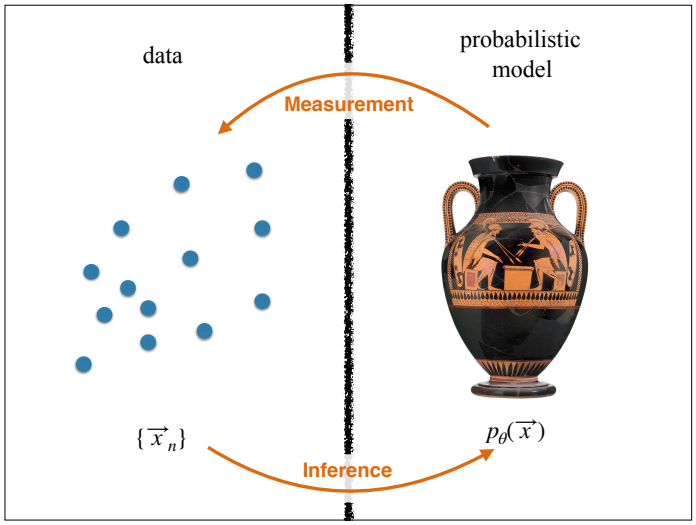
It now looks like taking aspirin is actually harmful. However the interval for the true stroke ratio θ turns out to be

$$.93 < \theta < 1.59 \quad (1.5)$$

with 95% confidence. This includes the neutral value $\theta = 1$, at which aspirin would be no better or worse than placebo vis-à-vis strokes. In the language of statistical hypothesis testing, aspirin was found to be significantly beneficial for preventing heart attacks, but not significantly harmful for causing strokes.

[Efron & Tibshirani '98]

38

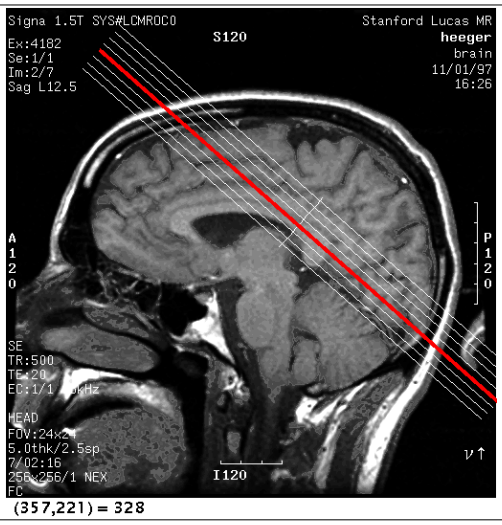


39

Point Estimates

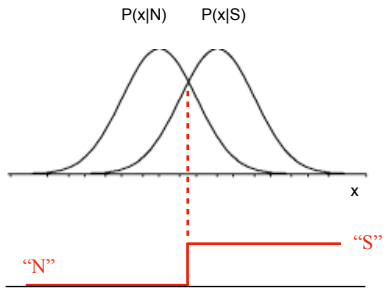
- Estimator: Any function of the data, intended to provide an estimate of the true value of a parameter
- The most common estimator is the sample average, used to estimate the true mean of a distribution.
- Statistically-motivated estimators:
  - Maximum likelihood (ML): $\hat{x}(\vec{d}) = \arg\max_x p(\vec{d}|x)$
  - Max a posteriori (MAP): $\hat{x}(\vec{d}) = \arg\max_x p(x|\vec{d})$
  - Bayes estimator: $\hat{x}(\vec{d}) = \arg\min_{\hat{x}} \mathbf{E}\left(L(x-\hat{x})|\vec{d}\right)$

40



41

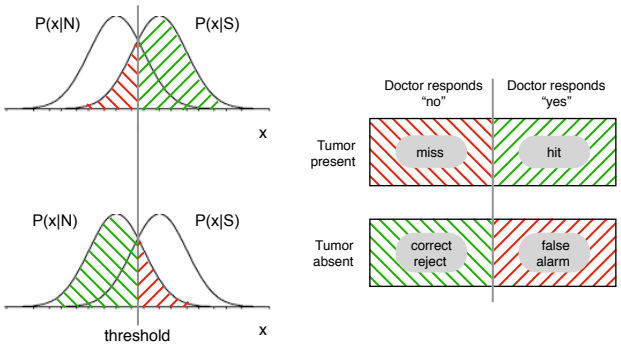
Signal Detection Theory



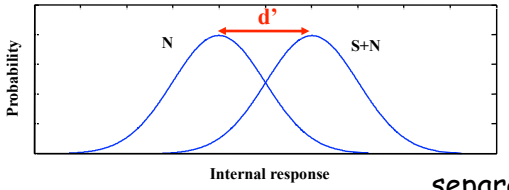
For equal, unimodal, symmetric distributions, ML decision rule is a *threshold* function.

42

Signal Detection Theory: Potential outcomes



Internal response: probability of occurrence curves



$$d' = \frac{\text{separation}}{\text{spread}}$$

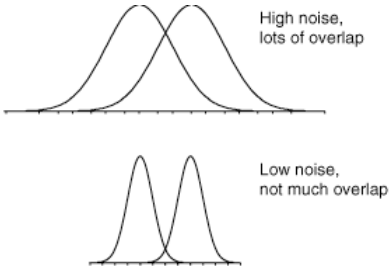
N: noise only (tumor absent)

S+N: signal plus noise (tumor present)

Discriminability ("d-prime") is the normalized separation between the two distributions

43

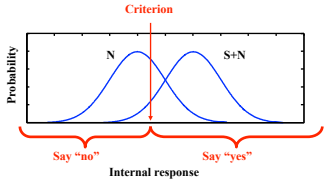
Signal Detection Theory: discriminability (d')



44

Example applications of SDT

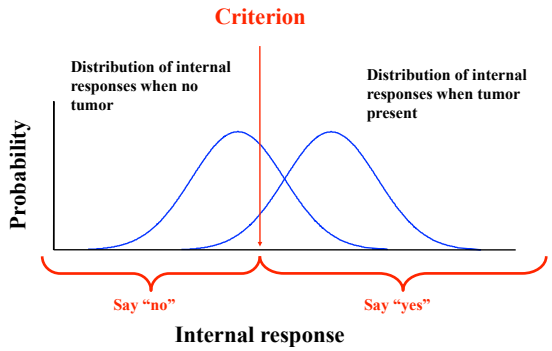
- Vision
  - Detection (something vs. nothing)
  - Discrimination (lower vs greater level of: intensity, contrast, depth, slant, size, frequency, loudness, ...
- Memory (internal response = trace strength = familiarity)
- Neurometric function/discrimination by neurons (internal response = spike count)



45

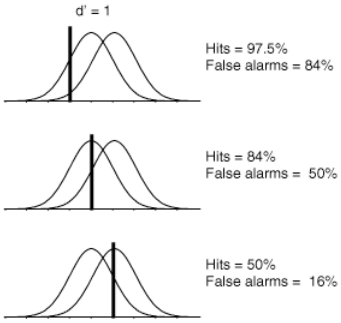
Criterion

46



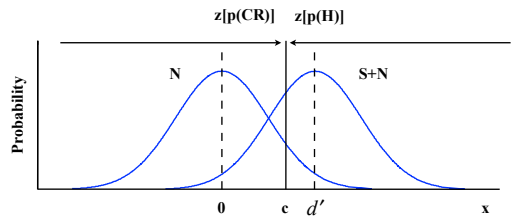
Signal Detection Theory: Criterion

47



SDT: Gaussian case

48



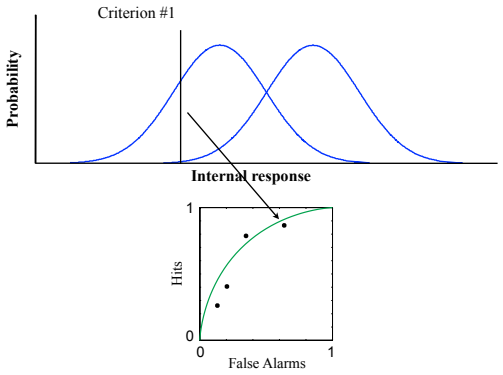
$$d' = z[p(H)] + z[p(CR)] = z[p(H)] - z[p(FA)]$$

$$c = z[p(CR)] \qquad G(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

$$\beta = \frac{p(x=c \mid S+N)}{p(x=c \mid N)} = \frac{e^{-(c-d')^2/2}}{e^{-c^2/2}}$$

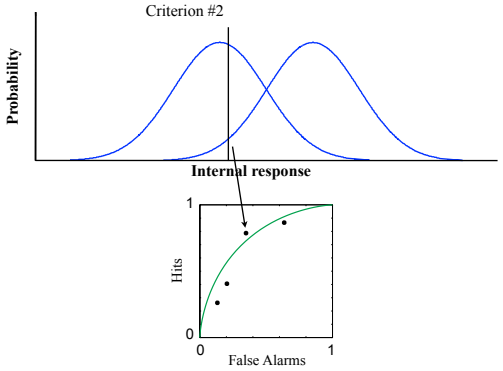
ROC (Receiver Operating Characteristic)

49



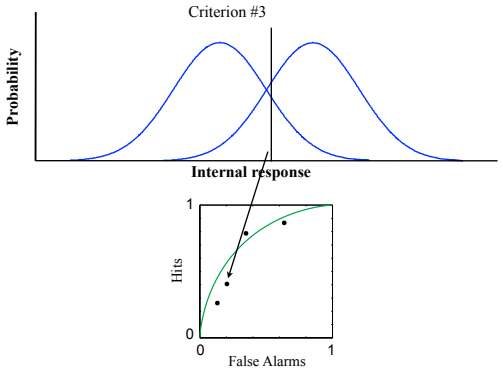
ROC (Receiver Operating Characteristic)

50



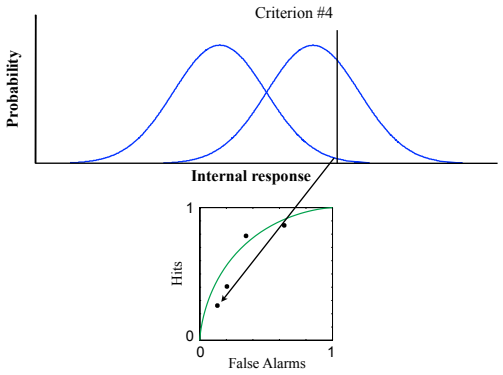
ROC (Receiver Operating Characteristic)

51



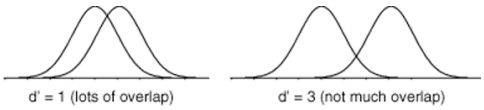
ROC (Receiver Operating Characteristic)

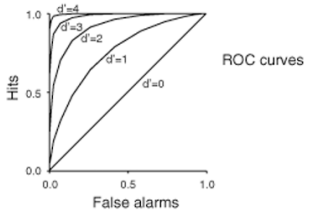
52



ROC (Receiver Operating Characteristic)

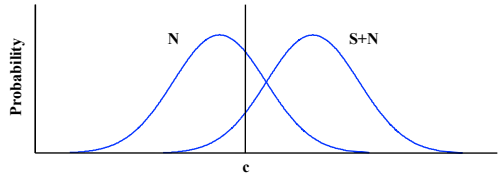
53

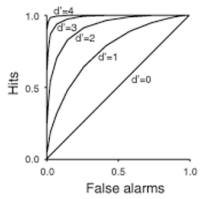


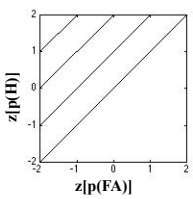


ROC: Gaussian case

54







Decision/classification in multiple dimensions

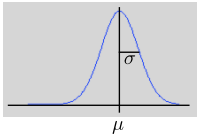
55

- Data-driven:
  - Fisher Linear Discriminant (FLD) - maximize d'
  - Support Vector Machine (SVM) - maximize margin
- Statistical:
  - ML/MAP/Bayes under a probabilistic model
  - e.g.: Gaussian, equal covariance (same as FLD)
  - e.g.: Gaussian, unequal covariance (QDA)
- Examples:
  - Visual gender identification
  - Neural population decoding

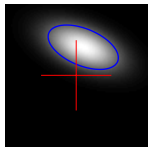
Multi-D Gaussian densities

56

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



$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}}\, e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$

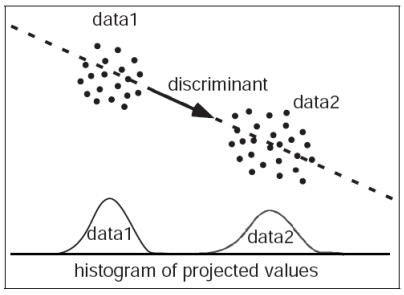


mean: [0.2, 0.8]
cov: [1.0 -0.3;
-0.3 0.4]

Linear Classifier

57

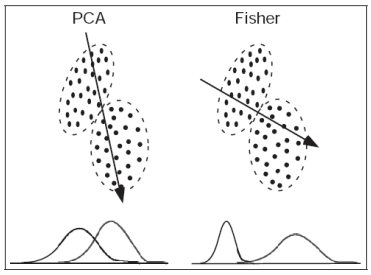
Find unit vector $\hat{w}$ ("discriminant") that best separates two distributions



Simplest choice: difference of means

Fisher Linear Discriminant

58



$$\max_{\hat{w}} \frac{\left[\hat{w}^T(\mu_A - \mu_B)\right]^2}{[\hat{w}^T C_A \hat{w} + \hat{w}^T C_B \hat{w}]}$$

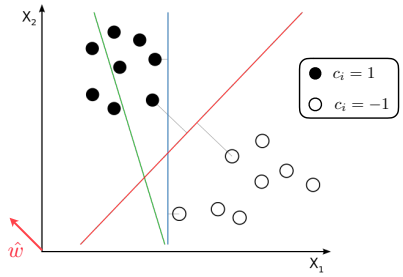
$$\hat{w} = D^{-1}V^T(\mu_A - \mu_B), \quad \text{where} \quad VD^2V^T = C_A + C_B$$

Support Vector Machine

59

Maximize the "margin" (gap between data sets)

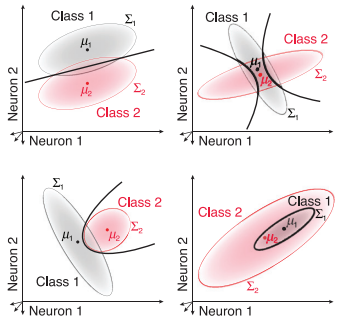
find largest m, and $\{\hat{w}, b\}$ s.t. $c_i(\hat{w}^T\vec{x}_i - b) \geq m, \quad \forall\, i$



Gaussian ML classifier

60

For equal covariances: Linear
For different covariances: Quadratic (three possible geometries)



[figure: Pagan et al. 2016]

61

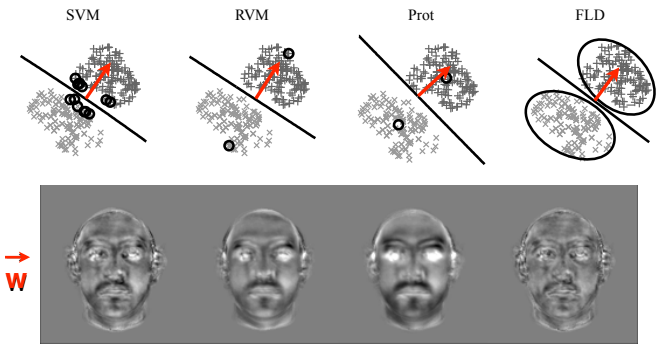
Example: Gender identification


- 200 face images (100 male, 100 female)
- Adjusted for position, size, intensity/contrast
- Labeled by 27 human subjects

[Graf & Wichmann, NIPS*03]

62

Linear classifiers



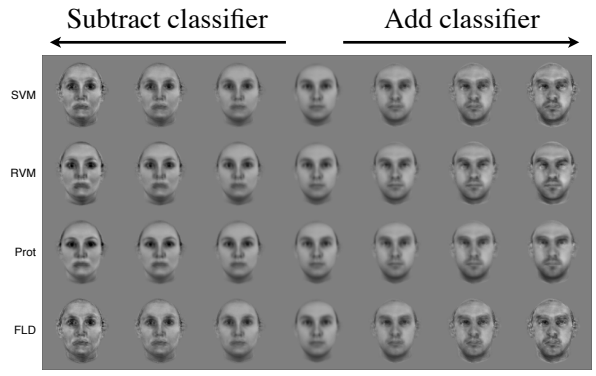
Four linear classifiers trained on subject data

63

Model validation/testing

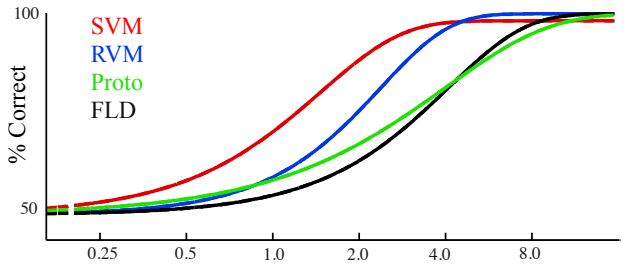
- Cross-validation: Subject responses [% correct, reaction time, confidence] are explained
  - very well by SVM
  - moderately well by RVM / FLD
  - not so well by Prot
- Curse of dimensionality strongly limits this result. A more direct test: Synthesize optimally discriminable faces...

64



[Wichmann, Graf, Simoncelli, Bülthoff, Schölkopf, NIPS*04]

65



[Wichmann, Graf, Simoncelli, Bülthoff, Schölkopf, NIPS*04]

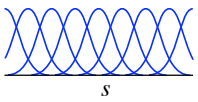
66

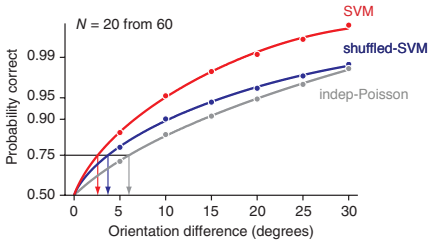
Population decoding

Independent Poisson responses
[e.g., Seung & Sompolinsky, 1993]

$$p(\vec{r}|s) = \prod_{n=1}^{N} \frac{h_n(s)^{r_n}\ e^{-h_n(s)}}{r_n!}$$

Tuning curves, $h_n(s)$





Graf, Kohn, Jazayeri, Movshon, 2011

67

Bayesian Model Comparison

- Is the coin fair? Compared to what?
- Point hypotheses: $M_1 : p = p_1 = 0.5 \quad M_2 : p = p_2 = 0.6$

$$p(M_1 \mid D) = \frac{p(D \mid M_1)P(M_1)}{p(D)} = \frac{p(D \mid M_1)P(p_1)}{p(D)}$$

Assuming equal priors over models the *Bayes factor* is

$$\frac{p(M_1 \mid D)}{p(M_2 \mid D)} = \frac{p(D \mid M_1)P(M_1)}{p(D \mid M_2)P(M_2)} = \frac{p(D \mid M_1)P(p_1)}{p(D \mid M_2)P(p_2)}$$

68

Bayesian Model Comparison

- Is the coin fair? Compared to what?
- Alternative hypothesis: $M_1 : p = p_1 = 0.5 \quad M_2 : p \neq 0.5$

$$p(M_2 \mid D) = \frac{p(D \mid M_2)p(M_2)}{p(D)}$$

$$= \int_0^1 p(p_{\text{coin}} \mid D)p(p_{\text{coin}})\,dp_{\text{coin}}$$

$$= \frac{\int_0^1 p(D \mid M_2, p_{\text{coin}})p(p_{\text{coin}})\,dp_{\text{coin}} P(M_2)}{p(D)}$$

Compute *Bayes factor* as before.

69

Continuous/Gaussian: Localization

I find that $\bar{x} \neq 0$. Is that convincing? Is the apparent bias real?

Continuous/Gaussian: Localization

Take *N* independent samples from the distribution, these act like draws from *N* independent, identically distributed (IID) RVs:

$$X_1, X_2, \cdots X_N$$

Continuous/Gaussian: Localization

The *N* independent samples are $x_1, x_2, \cdots x_N$

ML estimates are

$$\hat{\mu} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N-1}$$

Continuous/Gaussian: Localization

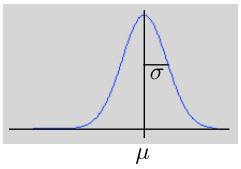
MAP estimates of the mean are based on the posterior, a product of Gaussians (assuming a Gaussian) prior. Thus there is shrinkage toward the prior.

Model comparison for hypotheses about the mean (variance assumed known) are similar to the binomial example.

73

The Gaussian

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



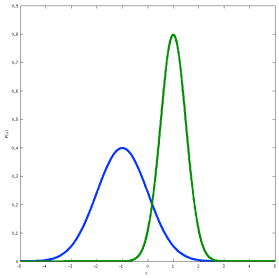
- parameterized by mean and stdev (position / width)
- joint density of two indep Gaussian RVs is circular! [easy]
- product of two Gaussian dists is Gaussian! [easy]
- conditionals of a Gaussian are Gaussian! [easy]
- sum of Gaussian RVs is Gaussian! [moderate]
- all marginals of a Gaussian are Gaussian! [moderate]
- central limit theorem: sum of many RVs is Gaussian! [hard]
- most random (max entropy) density with this variance! [moderate]

74

Product of Gaussian distributions is Gaussian

$$y = x + n, \quad x \sim N(\mu_x, \sigma_x),\ n \sim N(0, \sigma_n)$$

$$p(x|y) \propto p(y|x)p(x)$$



75

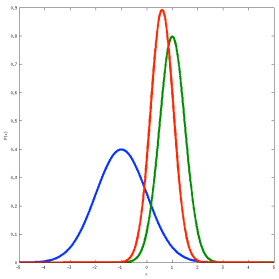
Product of Gaussian distributions is Gaussian

$$y = x + n, \quad x \sim N(\mu_x, \sigma_x),\ n \sim N(0, \sigma_n)$$

$$p(x|y) \propto p(y|x)p(x)$$

$$\propto e^{-\frac{1}{2}\left[\frac{1}{\sigma_n^2}(x-y)^2\right]} e^{-\frac{1}{2}\left[\frac{1}{\sigma_x^2}(x-\mu_x)^2\right]}$$

$$= e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma_n^2}+\frac{1}{\sigma_x^2}\right)x^2 - 2\left(\frac{y}{\sigma_n^2}+\frac{\mu_x}{\sigma_x^2}\right)x + \ldots\right]}$$



Completing the square shows that this posterior is also Gaussian, with

$$\sigma^2 = 1 \Big/ \left(\frac{1}{\sigma_n^2} + \frac{1}{\sigma_x^2}\right)$$

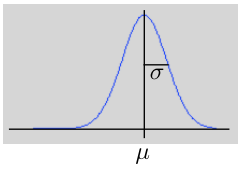
$$\mu = \left(\frac{y}{\sigma_n^2} + \frac{\mu_x}{\sigma_x^2}\right) \Big/ \left(\frac{1}{\sigma_n^2} + \frac{1}{\sigma_x^2}\right)$$

(average, weighted by *inverse* variances!)

76

The Gaussian

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

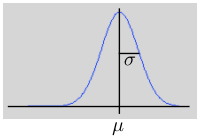


- parameterized by mean and stdev (position / width)
- joint density of two indep Gaussian RVs is circular! [easy]
- product of two Gaussian dists is Gaussian! [easy]
- conditionals of a Gaussian are Gaussian! [easy]
- sum of Gaussian RVs is Gaussian! [moderate]
- all marginals of a Gaussian are Gaussian! [moderate]
- central limit theorem: sum of many RVs is Gaussian! [hard]
- most random (max entropy) density with this variance! [moderate]

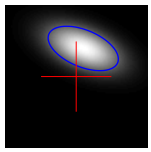
77

Multi-D Gaussian densities

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



$$p(\vec{x}) = \frac{1}{\sqrt{(2\pi)^N |C|}}\ e^{-(\vec{x}-\vec{\mu})^T C^{-1} (\vec{x}-\vec{\mu})/2}$$



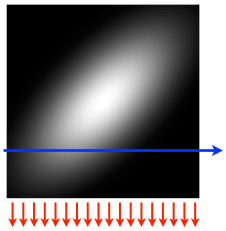
mean: [0.2, 0.8]
cov: [1.0 -0.3;
-0.3 0.4]

78

79

$\vec{x} \sim N(\vec{\mu}, C)$, let $P = C^{-1}$ (the "precision" matrix)

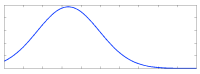
$$p(x_1|x_2 = a) \propto e^{-\frac{1}{2}\left[P_{11}(x_1-\mu_1)^2+2P_{12}(x_1-\mu_1)(a-\mu_2)+\ldots\right]}$$
$$= e^{-\frac{1}{2}\left[P_{11}x_1^2+2(P_{12}(a-\mu_2)-P_{11}\mu_1)x_1+\ldots\right]}$$
$$= e^{-\frac{1}{2}\left(x_1-\mu_1+\frac{P_{12}}{P_{11}}(a-\mu_2)\right)P_{11}\left(x_1-\mu_1+\frac{P_{12}}{P_{11}}(a-\mu_2)\right)+\ldots}$$



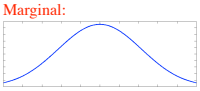
Gaussian, with:

$$\mu = \mu_1 + \frac{P_{12}}{P_{11}}(a-\mu_2)$$
$$\sigma^2 = \frac{1}{P_{11}}$$

Conditional:



Marginal:



$$p(x_1) = \int p(\vec{x})\, dx_2$$

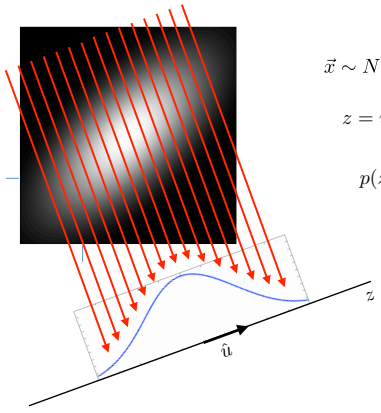
[on board]

Gaussian, with:

$$\mu = \mu_1$$
$$\sigma^2 = C_{11}$$

80

Generalized marginals of a Gaussian



$\vec{x} \sim N(\vec{\mu}_x, C_x)$

$z = \hat{u}^T\vec{x}$

$p(z)$ is Gaussian, with:

$$\mu_z = \hat{u}^T\vec{\mu}_x$$
$$\sigma_z^2 = \hat{u}^T C_x \hat{u}$$

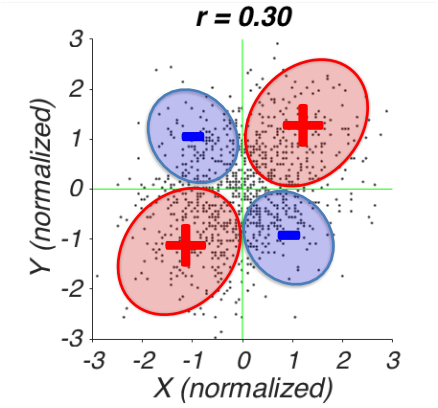
81

Bivariate statistics

- Covariance of X and Y: $Cov(X,Y) = E\left(\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right)$
- Correlation of X and Y: $\rho_{XY} = \dfrac{Cov(X,Y)}{\sigma_X\sigma_Y}$
- Unbiased estimators from samples: $s_{xy} = \dfrac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\left(N-1\right)}$

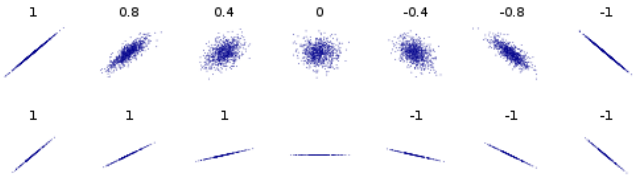
$$r_{xy} = \frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\left(N-1\right)s_x s_y} = \frac{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}}$$

Correlation: summary of data cloud shape



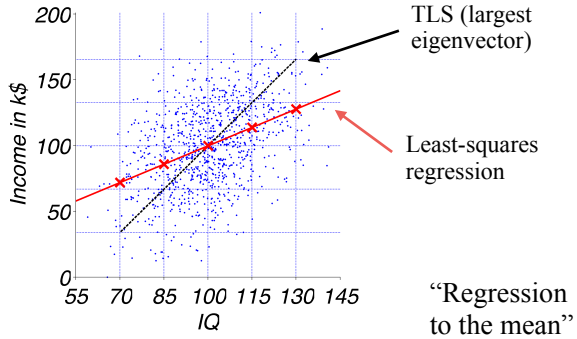
82

Correlation captures dependency, but not "shape"



83

Correlation and regression



TLS (largest eigenvector)

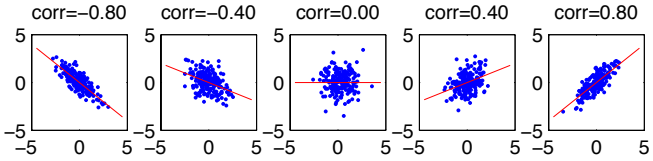
Least-squares regression

"Regression to the mean"

84

85

Correlation and regression



86

Variance partitioning and model assessment

$$SS_{\text{total}} = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2 = SS_{\text{explained}} + SS_{\text{residual}}$$

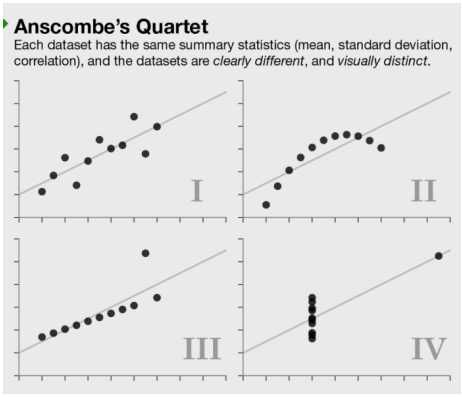
$$SS_{\text{explained}} = \sum_{i=1}^{N} \left(\hat{y}_i - \bar{y}\right)^2$$

$$SS_{\text{residual}} = \sum_{i=1}^{N} \left(\hat{y}_i - y_i\right)^2$$

$$\text{Coeff. of determination } r^2 = \frac{SS_{\text{explained}}}{SS_{\text{total}}}$$

87

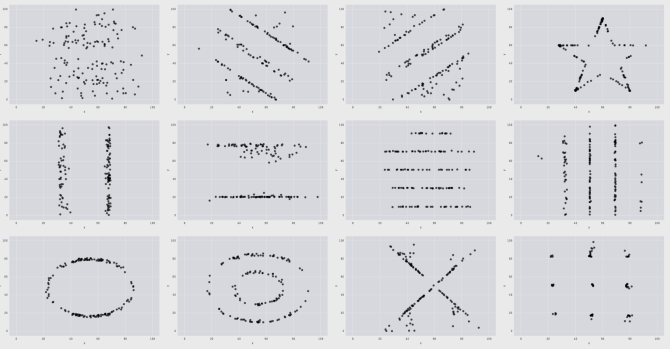
Correlation between variables does not uniquely indicate their relationship



88

More extreme examples !

X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



https://www.autodeskresearch.com/publications/samestats

89

Independence implies uncorrelated,
but uncorrelated doesn't imply independent!

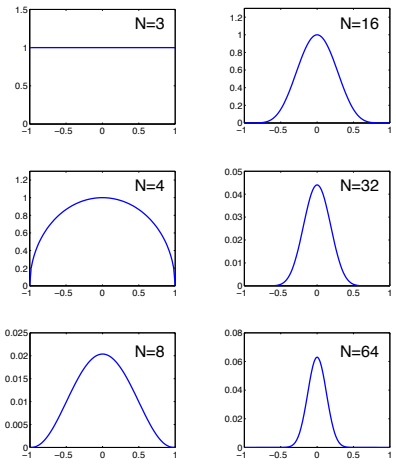
r = 0.0 0.0 0.0 0.0 0.0 0.0

90

Lack of correlation is favored in N>3 dimensions

Null Hypothesis: Distribution of normalized dot product of pairs of Gaussian vectors in N dimensions:

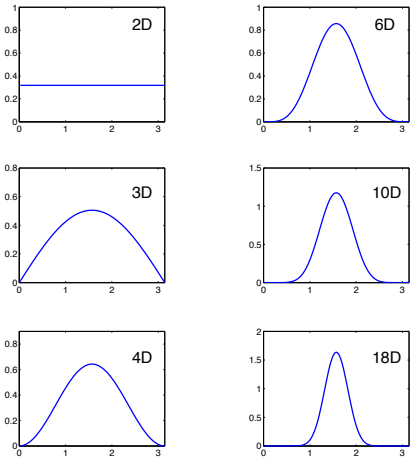
$$(1-d^2)^{\frac{N-3}{2}}$$



91

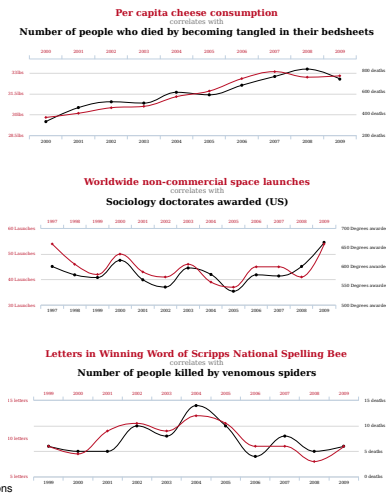
Distribution of angles of pairs of Gaussian vectors

sin(theta)^(N-2)



92

Nevertheless, one can find correlation if one looks for it!



http://www.tylervigen.com/spurious-correlations

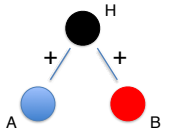
93

Correlation does not imply causation

- Beware selection bias
- Correlation does not provide a *direction* for causality. For that, you need additional (temporal) information.
- More generally, correlations are often a result of hidden (unmeasured, uncontrolled) variables…

Example: conditional independence:

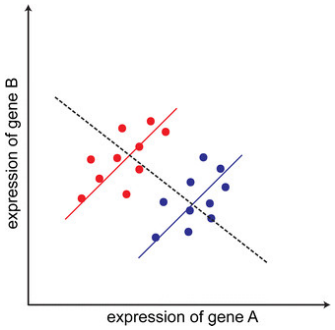
$p(A,B \mid H) = p(A \mid H)\, p(B \mid H)$

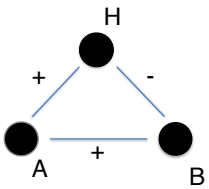


[on board: In Gaussian case, connections are explicit in the Precision Matrix]

94

Another example: Simpson's paradox



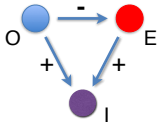


95

Milton Friedman's Thermostat

O = outside temperature (assumed cold)
I = inside temperature (ideally, constant)
E = energy used for heating

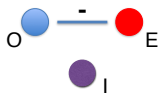
True interactions:



Statistical observations:

- O and I uncorrelated
- I and E uncorrelated
- O and E anti-correlated

Statistical interactions, $P=C^{-1}$:



Some nonsensical conclusions:

- O and E have no effect on I, so shut off heater to save money!
- I is irrelevant, and can be ignored. Increases in E cause decreases in O.

Statistical summary cannot replace scientific reasoning/experiments!

96

Summary: Correlation misinterpretations

- Correlation does not imply data lie on a line (subspace), with noise perturbations
- Correlation => dependency, but lack of correlation does *not* imply independence
- Correlation does not imply causation (temporally, or by direct influence/connection)
- Correlation is a descriptive statistic, and cannot replace the need for scientific reasoning/experiment!

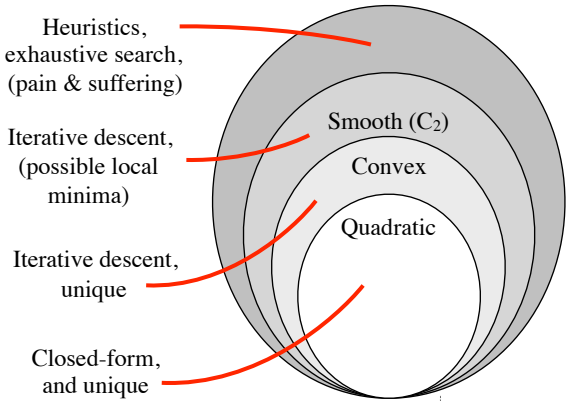
97

Taxonomy of model-fitting errors

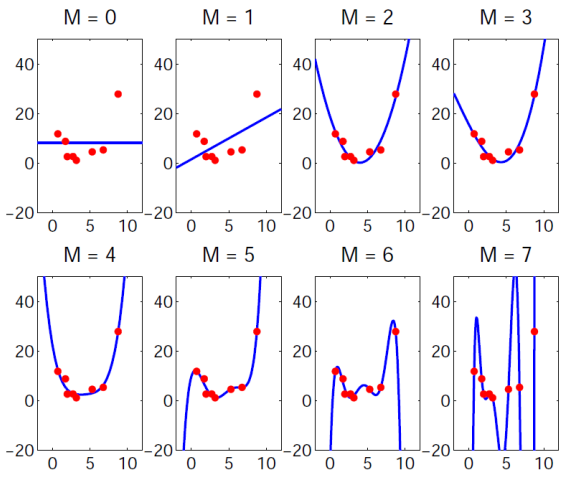
- Optimization failures (e.g., local minima) [convex relaxation, test with simulations]
- Overfitting (too many params, not enough data) [use cross-validation to select complexity, or to control regularization]
- Experimental variability (due to finite/noisy measurements) [use math/distributional assumptions, or simulations, or bootstrapping]
- Model failures

98

Optimization…



99



100

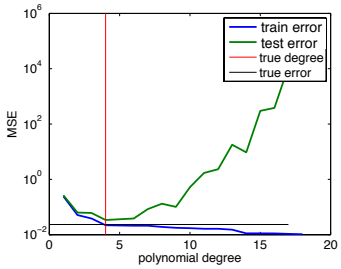
Cross-validation

A resampling method for constraining a model. Widely used to identify/avoid over-fitting.

(1) Randomly partition data into a "training" set, and a "test" set.
(2) Fit model to training set. Measure error on test set.
(3) Repeat (many times)

(4) Choose model that minimizes the cross-validated (test) error

Using cross-validation to select the degree of a polynomial model:



101

Ridge regression

(a.k.a. Tikhonov regularization)

Ordinary least squares regression:

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2$$

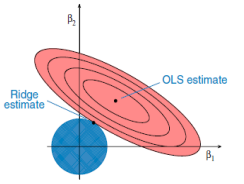
"Regularized" least squares regression:

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2 + \lambda||\vec{\beta}||^2$$

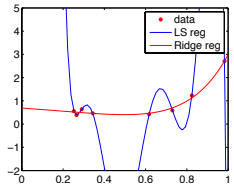
Equivalent formulation: MAP estimate, assuming Gaussian likelihood & prior!

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T \vec{y}$$

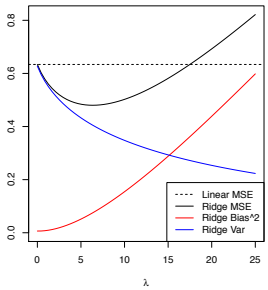
Choose lambda by cross-validation:



7th-order polynomial regression:



102



Linear regression:
Squared bias ≈ 0.006
Variance ≈ 0.627
Pred. error $\approx 1 + 0.006 + 0.627$
≈ 1.633

Ridge regression, at its best:
Squared bias ≈ 0.077
Variance ≈ 0.403
Pred. error $\approx 1 + 0.077 + 0.403$
≈ 1.48

from http://www.stat.cmu.edu/~ryantibs/datamining/

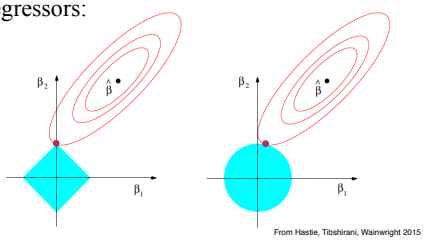
L_1 regularization

(a.k.a. least absolute shrinkage and selection operator - LASSO)

$$\arg\min_{\vec{\beta}} ||\vec{y} - X\vec{\beta}||^2 + \lambda \sum_k |\beta_k|$$

L_1 norm (still convex)

Using an absolute error regularization term promotes binary *selection* of regressors:



From Hastie, Tibshirani, Wainwright 2015

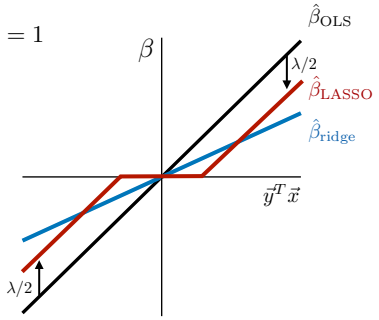
103

Solving for LASSO in 1D: soft-thresholding

$$\arg\min_{\beta} ||\vec{y} - \vec{x}\beta||^2 + \lambda|\beta|$$

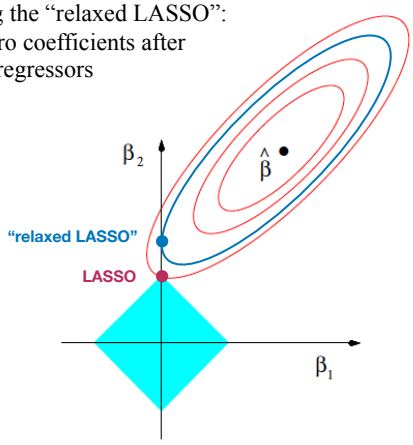
assume $||\vec{x}||^2 = 1$

[solution on board]



104

Bias reduction using the "relaxed LASSO": Re-solve for non-zero coefficients after eliminating unused regressors



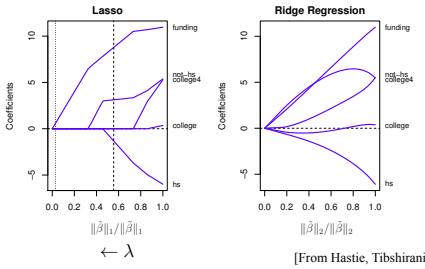
105

LASSO vs. Ridge regression

106

Table 2.1 *Crime data: Crime rate and five predictors, for $N = 50$ U.S. cities.*

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮	⋮	
50	66	67	26	18	16	940



[From Hastie, Tibshirani, Wainwright 2015]

Clustering

107

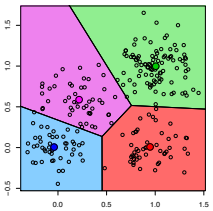
- K-Means (Lloyd, 1957)
- "Soft-assignment" version of K-means
  (a form of Expectation-Maximization - EM)
- In general, alternate between:
  1) Estimating cluster assignments
  2) Estimating cluster parameters
- Coordinate descent: converges to (possibly local) minimum
- Need to choose K (number of clusters)

K-Means

108

- Estimating cluster assignments: given class centers, assign each point to closest one.

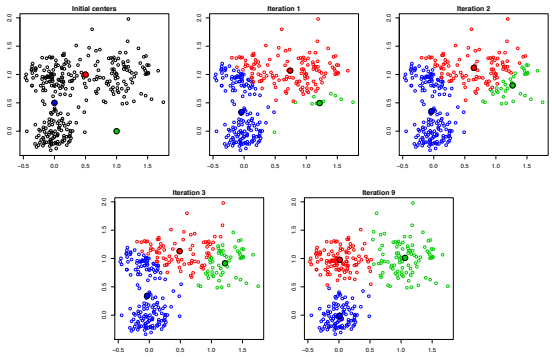
Soap bubbles:



- Estimating cluster parameters: given assignments, re-estimate the centroid of each cluster.

K-means example

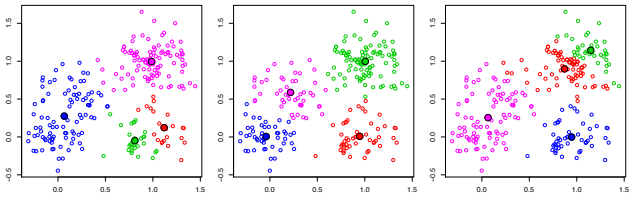
Here $X_i \in \mathbb{R}^2$, $n = 300$, and $K = 3$



[from R. Tibshirani, 2013]

Warning: Initialization matters…

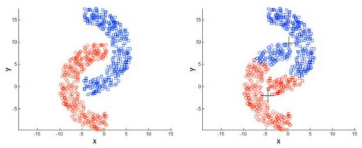
Three solutions obtained with different random starting points:



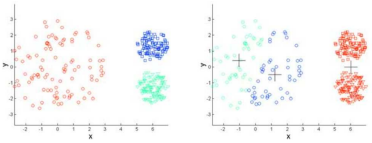
[from R. Tibshirani, 2013]

K-means failures

Non-convex/non-round-shaped clusters



Clusters with different densities



Picture courtesy: Christof Monz (Queen Mary, Univ. of London)

112

ML for discrete mixture of Gaussians

$$p(\vec{x}_n|a_{nk},\vec{\mu}_k,\Lambda_k) \propto \sum_k \frac{a_{nk}}{\sqrt{|\Lambda_k|}} e^{-(\vec{x}_n-\vec{\mu}_k)^T\Lambda_k^{-1}(\vec{x}_n-\vec{\mu}_k)/2}$$

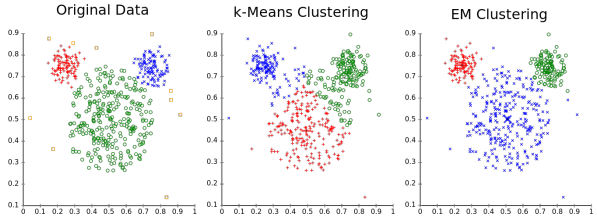
a_{nk} = assignment *probability*

$\{\vec{\mu}_k, \Lambda_k\}$ = mean/covariance of class *n*

Intuition: alternate between maximizing these two sets of variables ("coordinate descent")

113

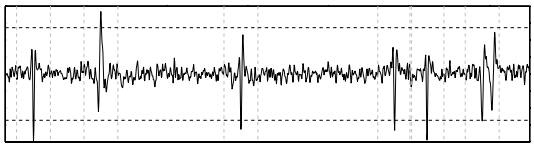
Different cluster analysis results on "mouse" data set:



[wikipedia]

114

Application to neural "spike sorting"



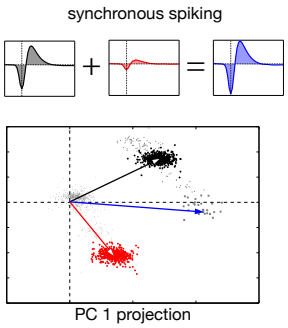
Standard solution:

1. Threshold to find segments containing spikes
2. Reduce dimensionality of segments using PCA
3. Identify spikes using clustering (e.g., K-means)

Note: Fails for overlapping spikes!

115

Failures of clustering for near-synchronous spikes

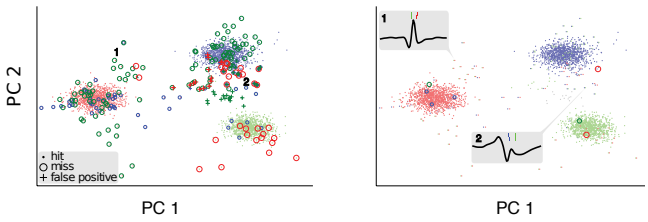


[Pillow et. al. 2013]

116

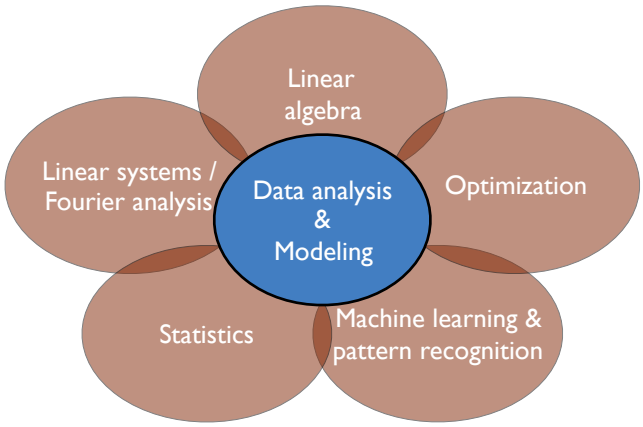
Simulated data [Quiroga et. al. 2004]

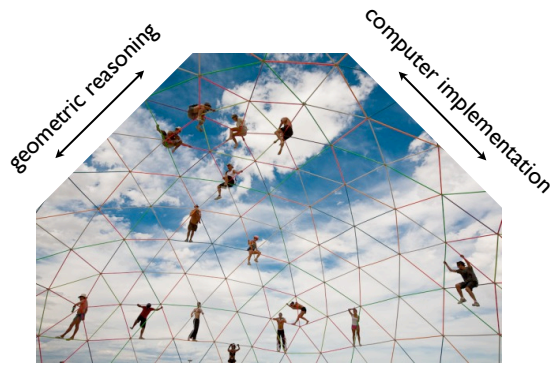
clustering (K-means) CBP



[Ekanadham et al, 2014]

117




geometric reasoning
computer implementation
mathematical manipulation