# 4th IMPRS Astronomy Summer School

## Drawing Astrophysical Inferences from Data Sets

William H. Press
The University of Texas at Austin

Lecture 6

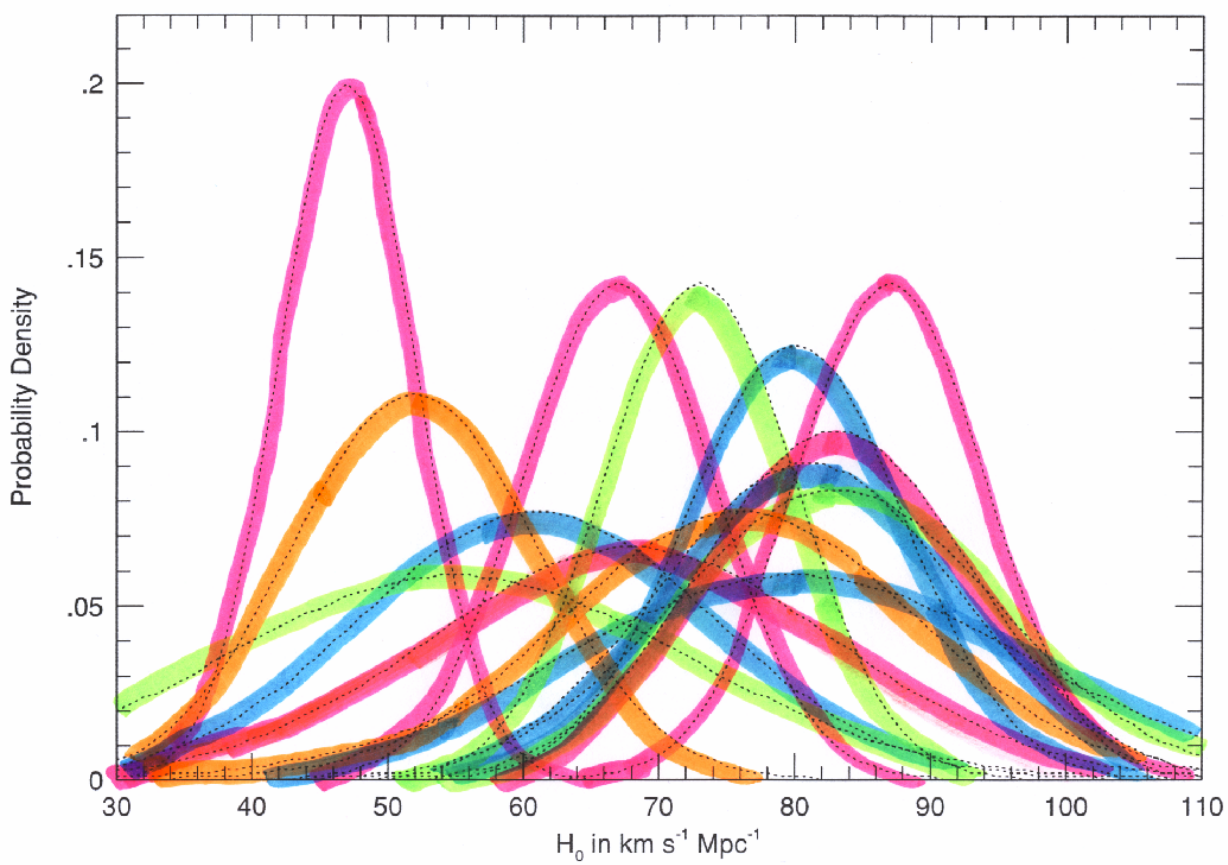
# Mixture Models

A general idea for dealing with events which can be in one of several components – and you don't know which.
Instead of doing general case, we'll illustrate an astronomical example.

this one is a 3-component mixture

- Hubble constant $H_0$ was highly contentious as late as the late 1990s.
- Measurement by classical astronomy very difficult
  - each a multi-year project
  - calibration issues
- Between 1930 and 2000 credible measurements ranged from 30 to 120 (km/s/Mpc)
  - many claimed small errors
- Consensus view was "we just don't know $H_0$ .
  - or was it just failure to apply an adequate statistical model to the existing data?

Grim observational situation: Not the range of values, but the inconsistency of the claimed errors. This forbids any kind of "just average", because goodness-of-fit rejects the possibility that these experiments are measuring the same value.



Here, the mixture will be "some experiments are right, some are wrong, and we don't know which are which"

$$P(H_i|D) \propto P(D|H_i)P(H_i)$$

$$P(A) = \sum_i P(AB_i)$$ "law of total probability"

$$P(H_0|D) = \sum_{p,\mathbf{v}} P(H_0 p\mathbf{v}|D)$$

probability that a "random" experiment is right

bit vector of which experiments are right or wrong, e.g. (1,0,0,1,1,0,1…)

$$\propto \sum_{p,\mathbf{v}} P(D|H_0 p\mathbf{v})P(H_0 p\mathbf{v})$$ Bayes

now, expand out the prior and make reasonable assumptions about conditional independence:

$$P(H_0|D) \propto \sum_{p,\mathbf{v}} P(D|H_0 p\mathbf{v})P(H_0)P(p|H_0)P(\mathbf{v}|H_0 p)$$

$p^{\#(\mathbf{v}=1)}\ (1-p)^{\#(\mathbf{v}=0)}$

$$P(D|H_0 p\mathbf{v}) = \prod_{v_i=1} P_{Gi} \prod_{v_i=0} P_{Bi}$$

$$= \exp\Big[\sum_{v_i=1} \frac{-(H_i - H_0)^2}{2\sigma_i^2}\Big] \times \exp\Big[\sum_{v_i=0} \frac{-(H_i - H_0)^2}{2S^2}\Big]$$

any big enough value

$$P(H_0|D)$$

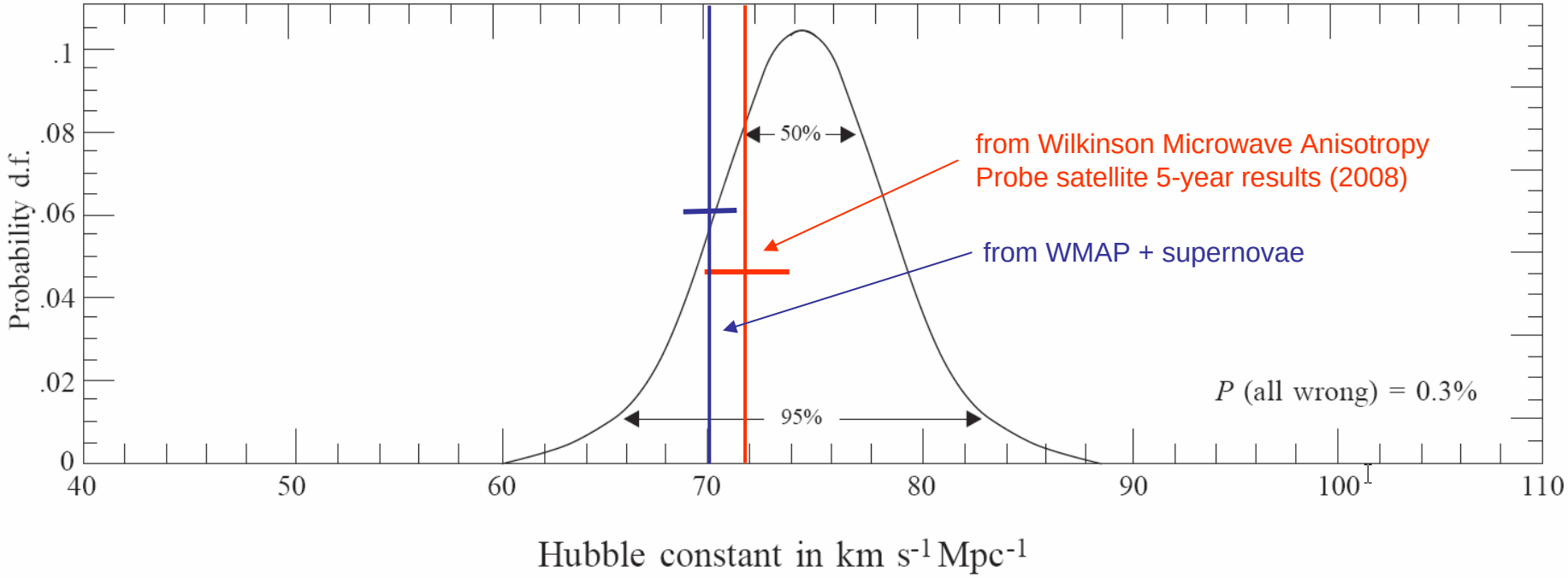
$$\propto P(H_0) \sum_p P(p) \sum_{\mathbf{v}} \Big[\prod_{v_i=1} P_{Gi} p\Big]\Big[\prod_{v_i=0} P_{Bi}(1-p)\Big]$$

now you stare at this a while and realize that the sum over v is just a multinomial expansion

$$\propto P(H_0) \sum_p P(p) \prod_i [pP_{Gi} + (1-p)P_{Bi}]$$

So it is as if each event were sampled from the linear mixture of probability distributions. We see that this is not just heuristic, but the actual, exact marginalization over all possible assignments of events to components.

And the answer is…




- This is <u>not</u> a Gaussian, it's just whatever shape it came out from the data
- It's not even necessarily unimodel (although it is for this data)
  - If you leave out some of the middle-value experiments, it splits to be bimodal
  - Thus showing that this method is <u>not</u> "tail-trimming"

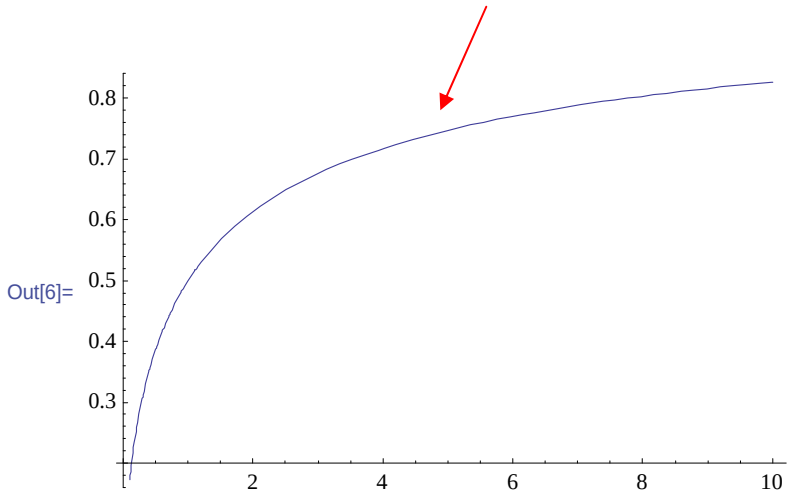
Similarly, we can get the probability that each experiment is correct (that is, the assignment of events to components):

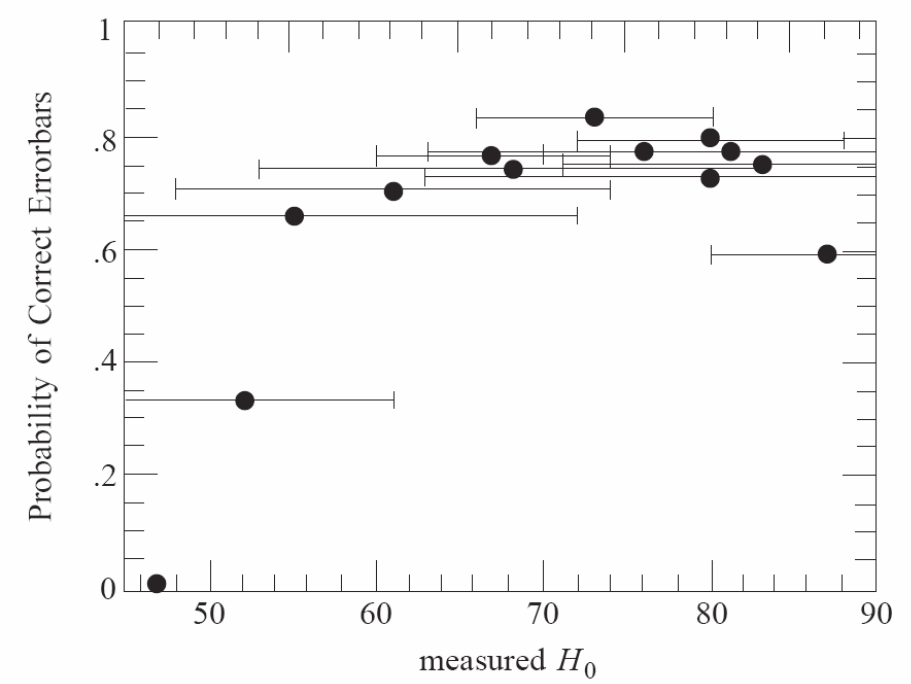
$$P(v_i = 1, p|D) \propto p\, P_{Gi}$$

$$P(v_i = 1|D) = \int_{p=0}^{1} \frac{pP_{Gi}}{pP_{Gi} + (1-p)P_{Bi}} P(p)dp$$

and if *P(p)* is uniform in (0,1)

$$= \frac{r(r - 1 - \log r)}{(r-1)^2}, \quad \text{where} \quad r = \frac{P_{Gi}}{P_{Bi}}$$

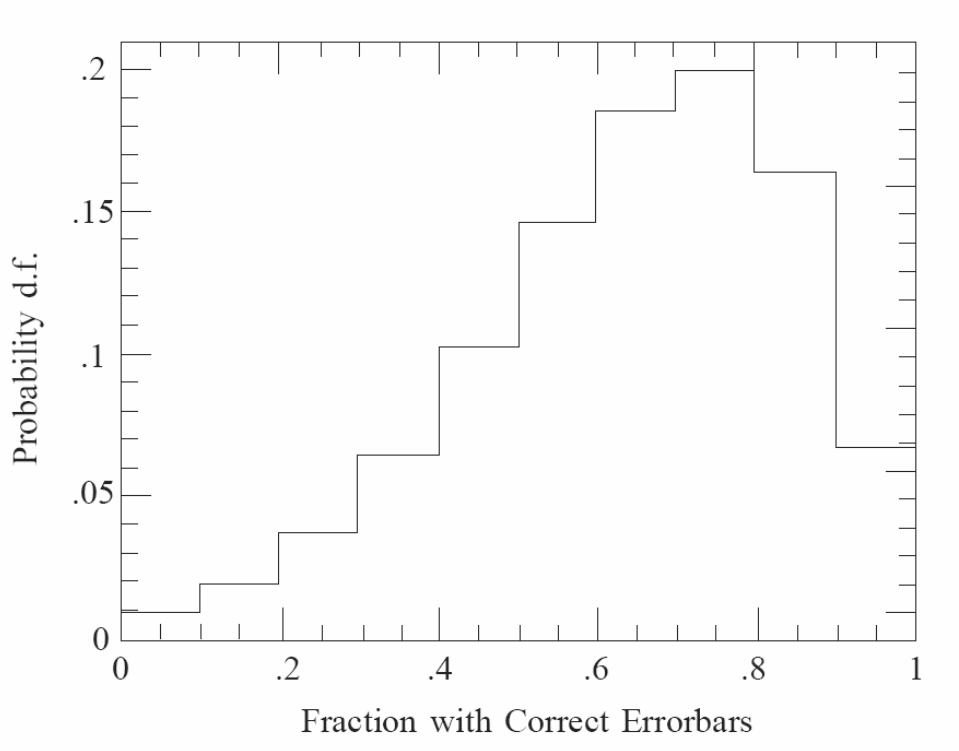


Or the probability distribution of p (probability of a measurement being correct a priori):

$$P(p|D) \propto \cancel{P(p)} \int \underbrace{\prod_i [pP_{Gi} + (1-p)P_{Bi}]}_{\text{mixture automatically marginalizes on v}} P(H_0)dH_0$$

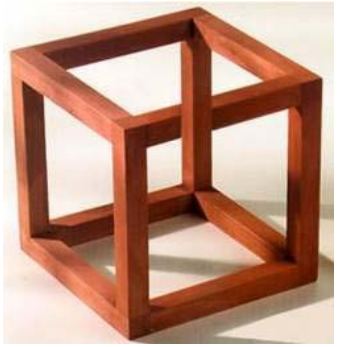
e.g., take uniform prior on $P(p)$




(this is of course not universal, but depends on the field and its current state)

# Gaussian Mixture Models (GMMs)

- What if the components have unknown parameters
  - like location and shape in N-dim space
- Can't assign the events to components until we know the components
  - but can't define the components until we know which events are assigned to them
  - the trick is to find both, self-consistently



- EM (expectation-maximization) methods are a general iterative method for dealing with problems like this
- "Gaussian Mixture Models" are the simplest example
  - the components are Gaussians defined by a mean and covariance
- Note that not all mixture models are EM methods, and not all EM methods are mixture models!
  - EM methods can also deal with missing data, e.g.

Key to the notational thicket:

$M$ dimensions
$k = 1 \ldots K$ Gaussians
$n = 1 \ldots N$ data points
$P(k)$ population fraction in $k$
$P(\mathbf{x}_n)$ model probability at $\mathbf{x}_n$

$\boldsymbol{\mu}_k$ (the $K$ means, each a vector of length $M$)
$\boldsymbol{\Sigma}_k$ (the $K$ covariance matrices, each of size $M \times M$)
$P(k|n) \equiv p_{nk}$ (the $K$ probabilities for each of $N$ data points)

"probabilistic assignment" of a data point to a component!

$$\mathcal{L} = \prod_n P(\mathbf{x}_n)$$ overall likelihood of the model

$$P(\mathbf{x}_n) = \sum_k N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k)$$ specify the model as a mixture of Gaussians

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{M/2} \det(\boldsymbol{\Sigma})^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}) \cdot \boldsymbol{\Sigma}^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu})]$$

Goal is to find *all* of the above, starting with only the $\mathbf{x}_n$

Expectation, or E-step: suppose we know the model, but not the assignment of individual points.

(so called because it's probabilistic assignment by expectation value)

$$p_{nk} \equiv P(k|n) = \frac{N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k)}{P(\mathbf{x}_n)}$$

Maximization, or M-step: suppose we know the assignment of individual points, but not the model.

$$\hat{\boldsymbol{\mu}}_k = \sum_n p_{nk} \mathbf{x}_n \Big/ \sum_n p_{nk}$$

$$\hat{\boldsymbol{\Sigma}}_k = \sum_n p_{nk} (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k) \otimes (\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k) \Big/ \sum_n p_{nk}$$

$$\hat{P}(k) = \frac{1}{N} \sum_n p_{nk}$$

(so called because [theorem!] the overall likelihood increases at each step)

- Can be proved that alternating E and M steps converges to (at least a local) maximum of overall likelihood
- Convergence is sometimes slow, with long "plateaus"
- Often start with k randomly chosen data points as starting means, and equal (usually spherical) covariance matrices
  - but then had better try multiple re-starts

Because Gaussians underflow so easily, a couple of tricks are important:

1) Use logarithms!

$$\log N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\cdot\boldsymbol{\Sigma}^{-1}\cdot(\mathbf{x}-\boldsymbol{\mu}) - \frac{M}{2}\log(2\pi) - \frac{1}{2}\log\det(\boldsymbol{\Sigma})$$

2) Do the sum $$P(\mathbf{x}_n) = \sum_k N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k)$$

by the "log-sum-exp" formula:

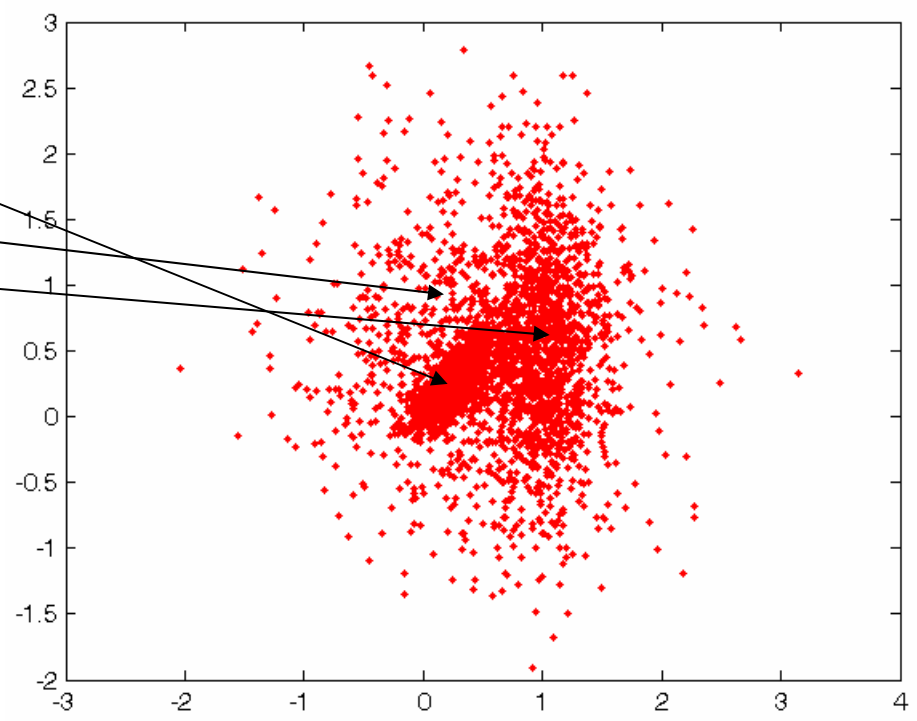
$$\log\left(\sum_i \exp(z_i)\right) = z_{\max} + \log\left(\sum_i \exp(z_i - z_{\max})\right)$$

(The code in NR3 implements these tricks.)

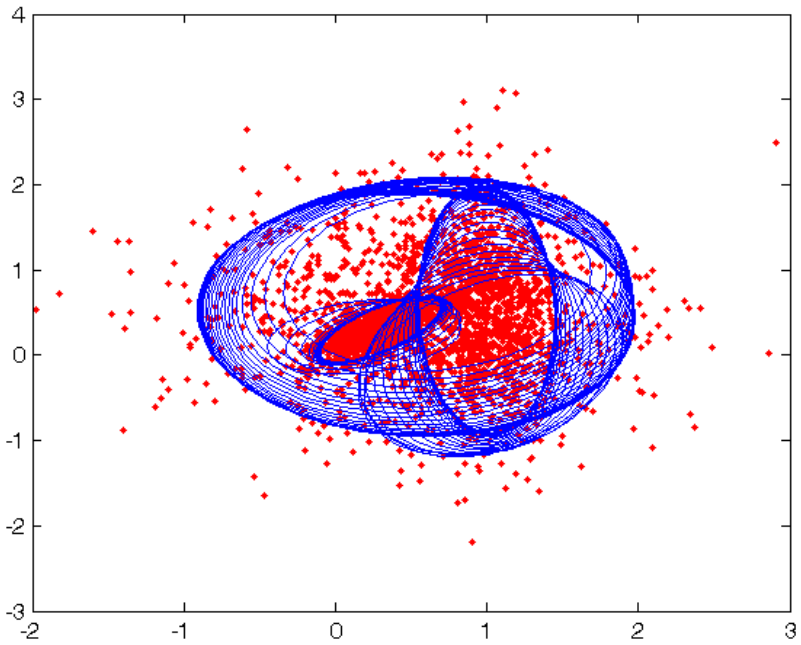
Let's look in 2 dimensions at an "ideal", and then a "non-ideal", example.

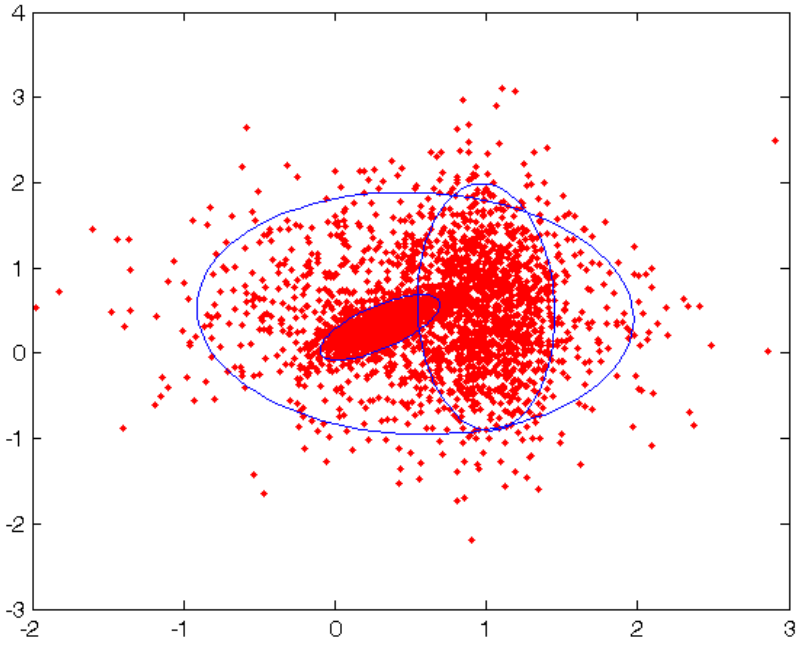
Ideal: we generate Gaussians, then, we fit to Gaussians

```
mu1 = [.3 .3];
sig1 = [.04 .03; .03 .04];
mu2 = [.5 .5];
sig2 = [.5 0; 0 .5];
mu3 = [1 .5];
sig3 = [.05 0; 0 .5];
```



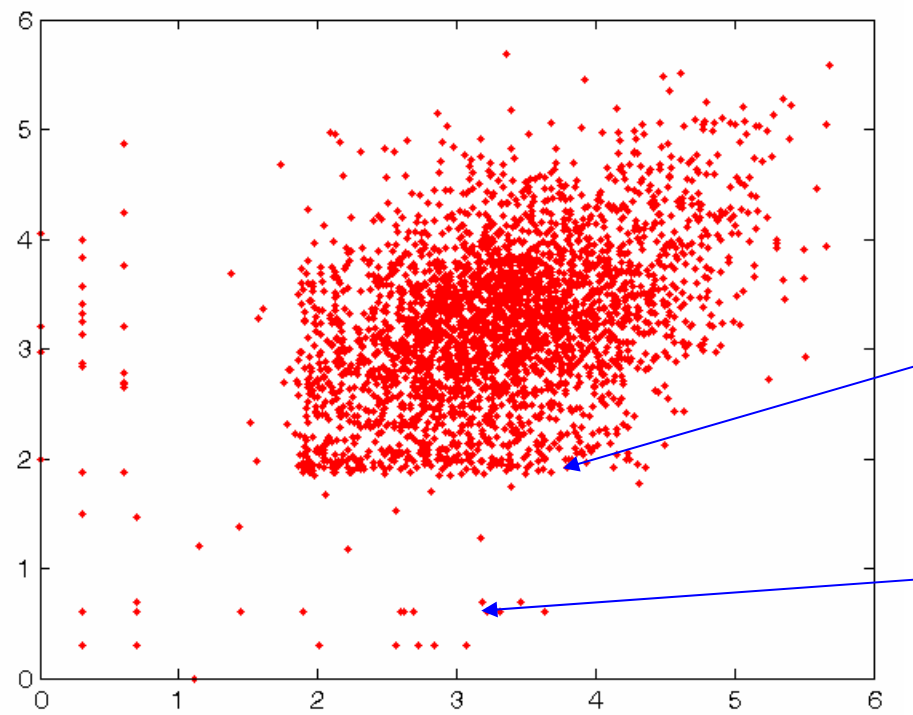
Use GMM class in NR3:




This “ideal” example converges rapidly to the right answer.

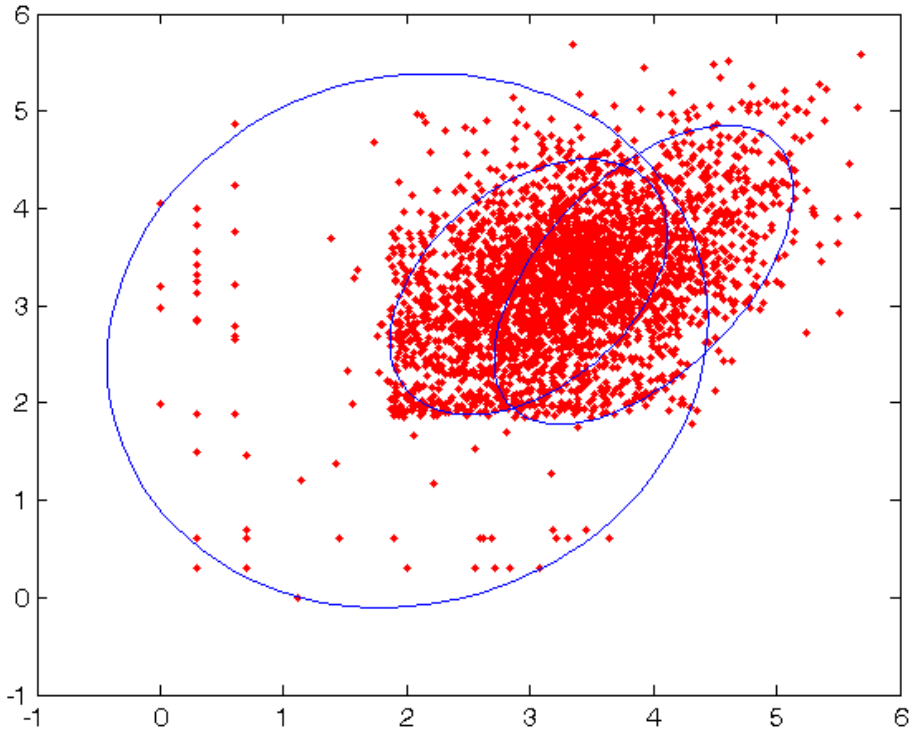
For a non-ideal example, here is some biological data on the (log10) lengths of the 1st and 2nd introns in genes. We can see that something non-GMM is going on! For general problems in >2 dimensions, it's often hard to visualize whether this is the case or not, so GMMs get used "blindly".



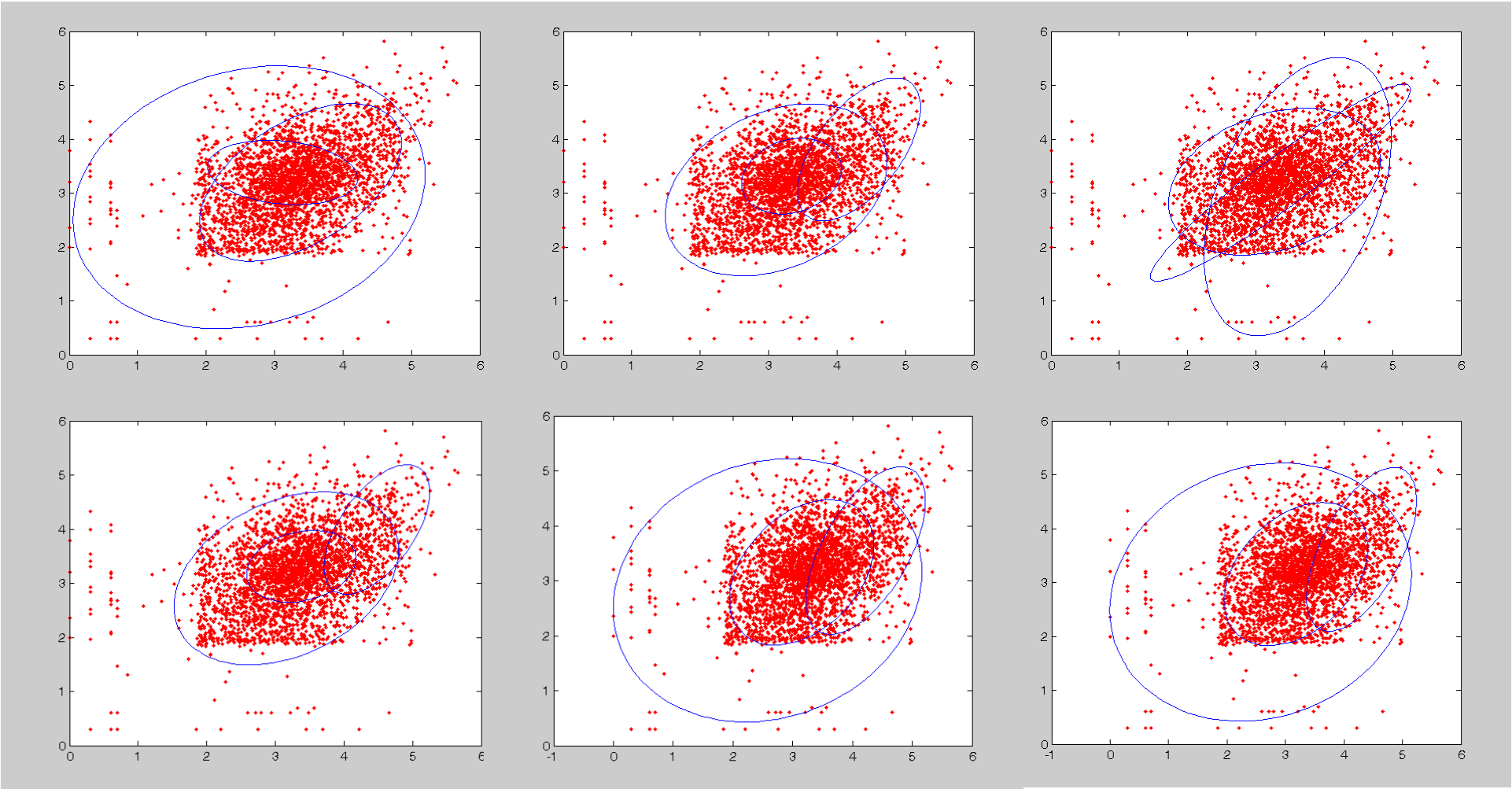

spliceosome can't deal with introns <100 in length, so strong evolutionary constrant

except that, like everything else in biology, there are exceptions (these are not "experimental error" in the physics sense!)

Three component model converges rapidly to something reasonable:

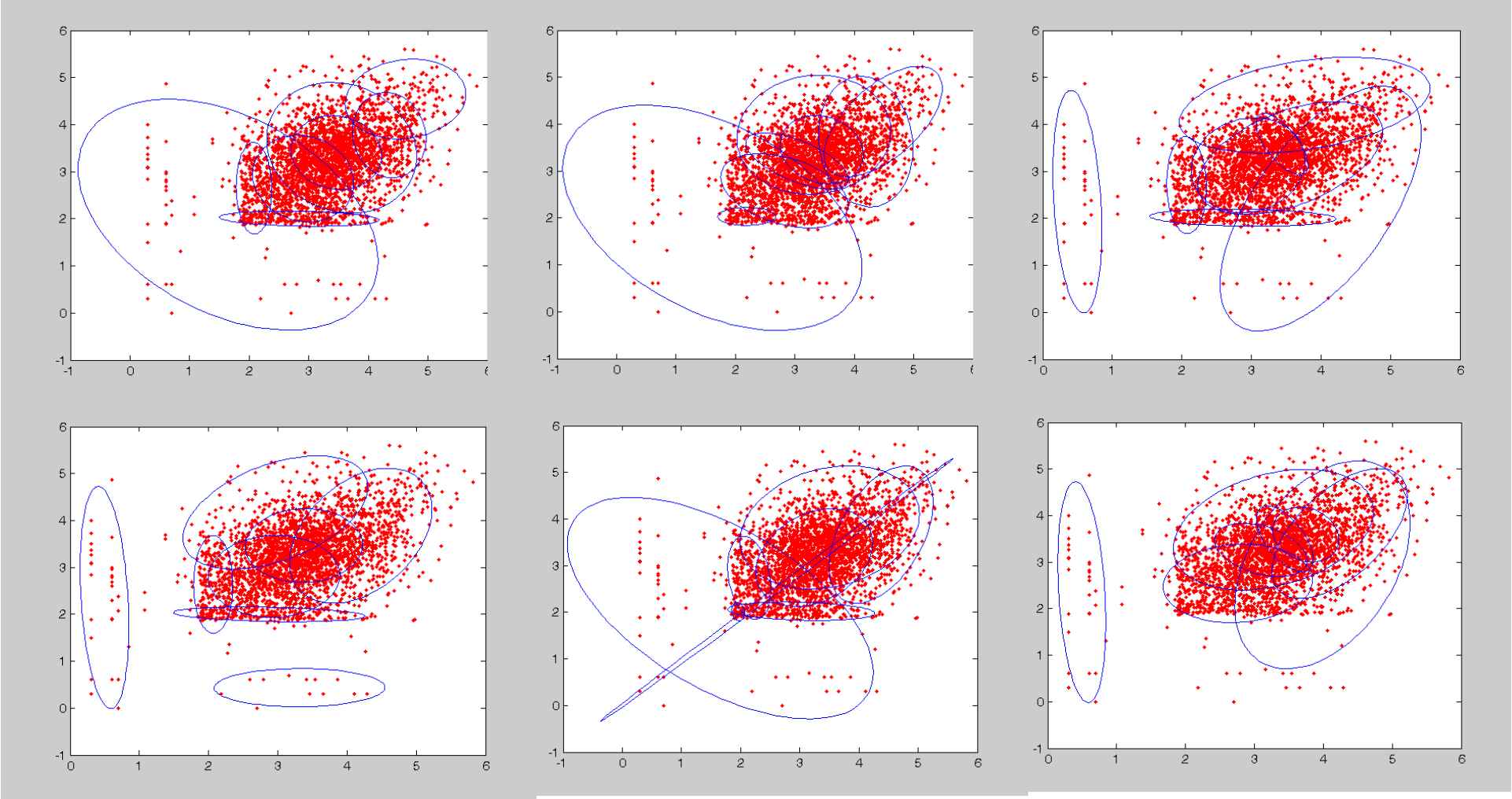


But, we don't always land on the same local maximum, although there seem to be just a handfull.



(One of these presumably has the higher likelihood.)

Eight components:



The ones with higher likelihood are pretty good as summaries of the data distribution (absent a predictive model). But the individual components are unstable and have little or no meaning. "Fit a lot of Gaussians for interpolation, but don't believe them."

**Variations on the theme of GMMs:**

- You can constrain the $\Sigma$ matrices to be diagonal
  - when you have reason to believe that the components individually have no cross-correlations (align with the axes)

$$(\hat{\boldsymbol{\Sigma}}_k)_{mm} = \sum_n p_{nk}[(\mathbf{x}_n)_m - (\hat{\boldsymbol{\mu}}_k)_m]^2 \Big/ \sum_n p_{nk}$$

- Or constrain them to be multiples of the unit matrix
  - make all components spherical

$$(\hat{\boldsymbol{\Sigma}}_k) = \mathbf{1} \times \left( \sum_n p_{nk} |\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k|^2 \Big/ \sum_n p_{nk} \right)$$

- Or fix $\Sigma = \varepsilon \mathbf{1}$ (infinitesimal times unit matrix)
  - don't re-estimate $\Sigma$, only re-estimate $\mu$
  - this assigns points 100% to the closest cluster (so don't actually need to compute any Gaussians, just compute distances)
  - it is called "K-means clustering"
    - kind of GMM for dummies
    - widely used (there are a lot of dummies!)
    - probably always better to use spherical GMM (middle bullet above)