CSE 312

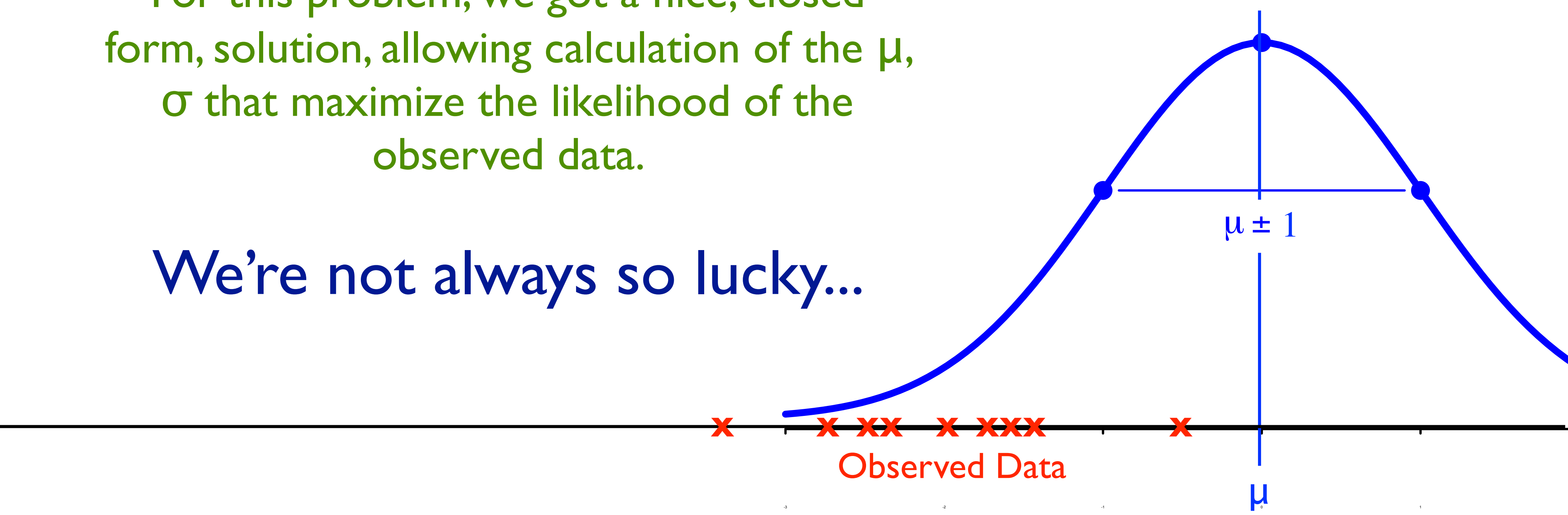
Winter 2017

EM: The Expectation-Maximization Algorithm
(for a Two-Component Gaussian Mixture)

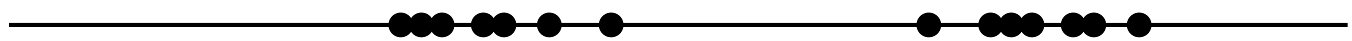
Previously:
How to estimate μ given data

For this problem, we got a nice, closed form, solution, allowing calculation of the μ, σ that maximize the likelihood of the observed data.

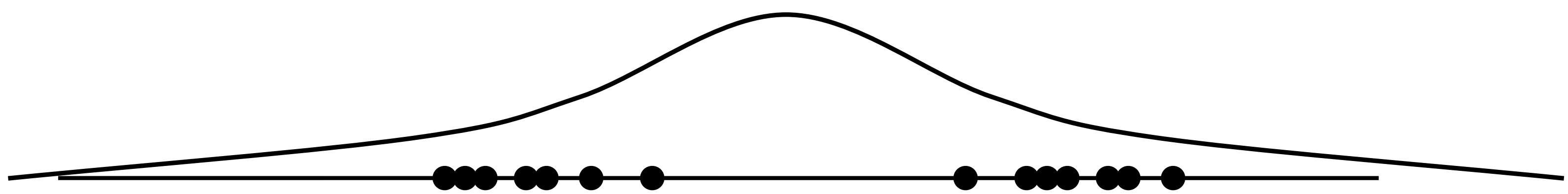
We're not always so lucky...



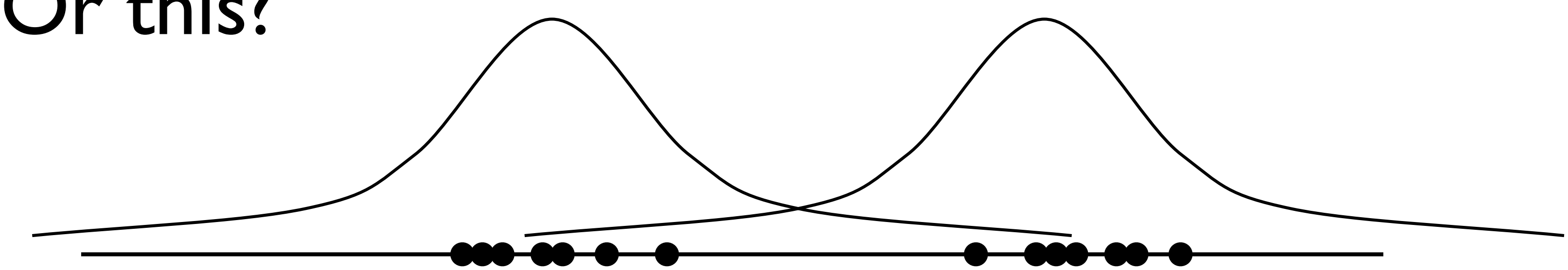
More Complex Example



This?

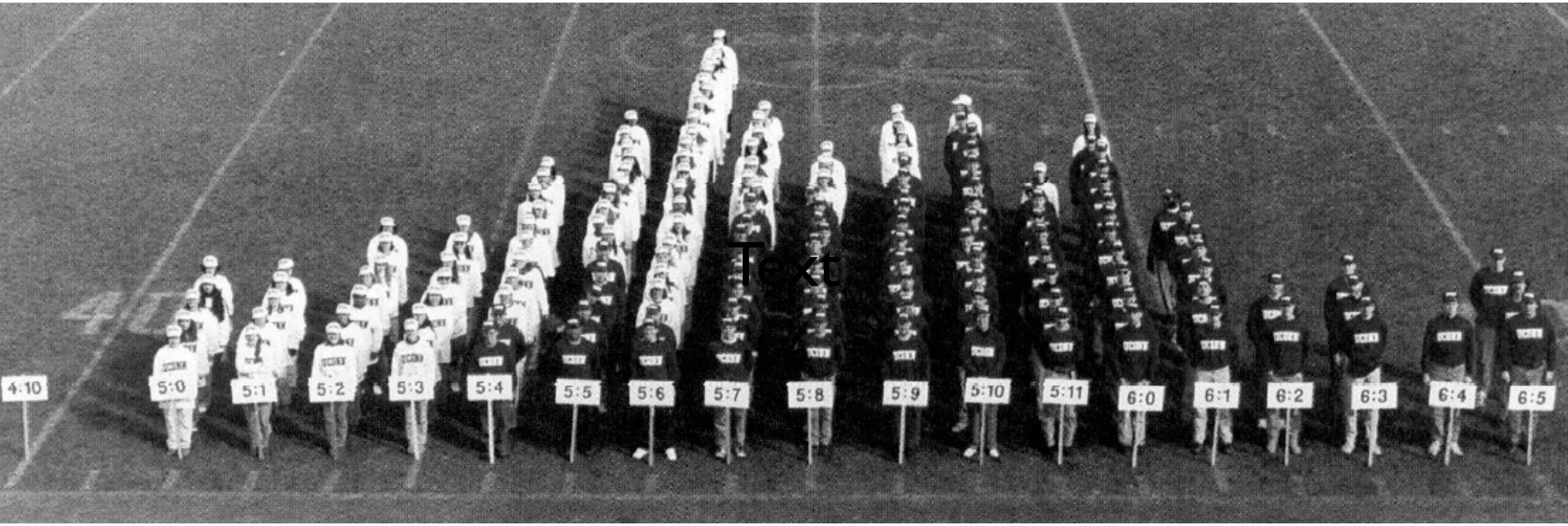


Or this?



(A modeling decision, not a math problem...,
but if the later, what math?)

A Living Histogram

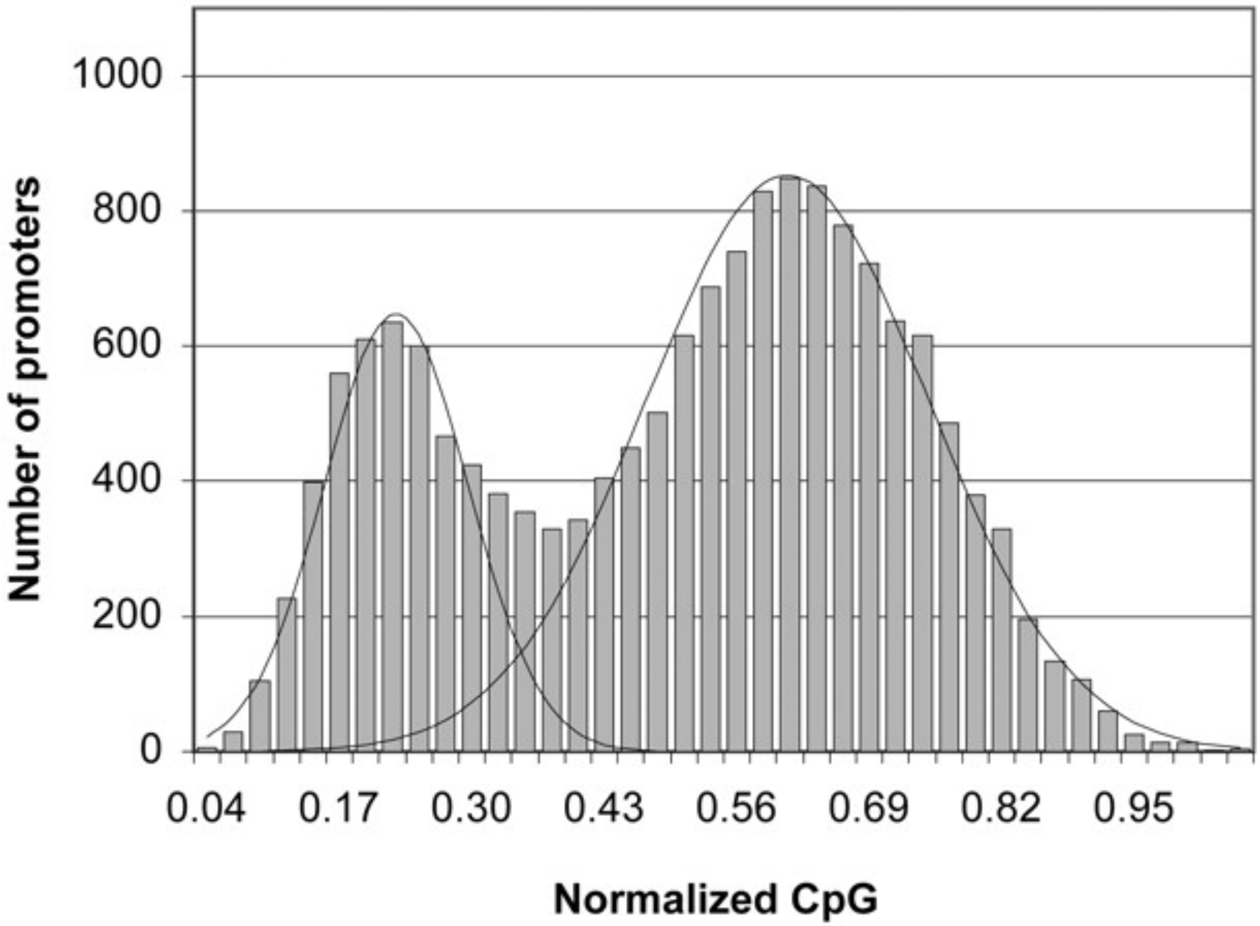


male and female genetics students, University of Connecticut in 1996

http://mindprod.com/jgloss/histogram.html

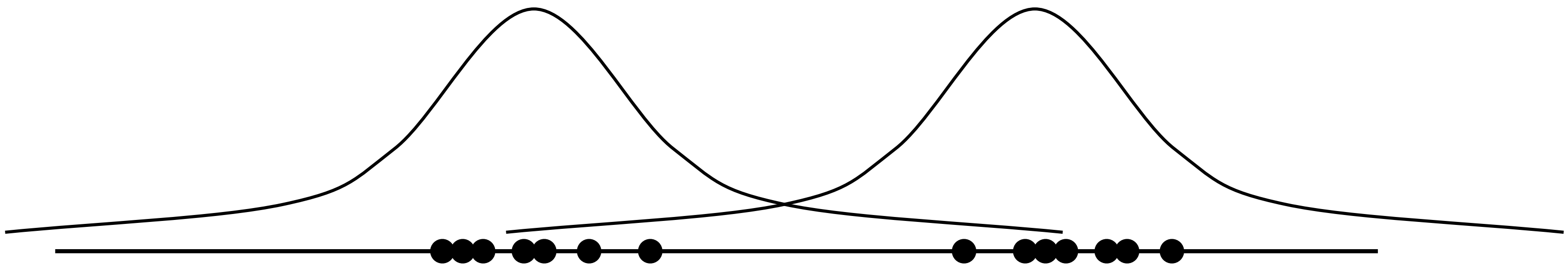
Another Real Example:

CpG content of human gene promoters



"A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters" Saxonov, Berg, and Brutlag, PNAS 2006;103:1412-1417

Gaussian Mixture Models / Model-based Clustering



Parameters θ

means	μ_1	μ_2
variances	σ_1^2	σ_2^2
mixing parameters	τ_1	$\tau_2 = 1 - \tau_1$

P.D.F. —separately→ $f(x|\mu_1, \sigma_1^2)$ $f(x|\mu_2, \sigma_2^2)$

—together→

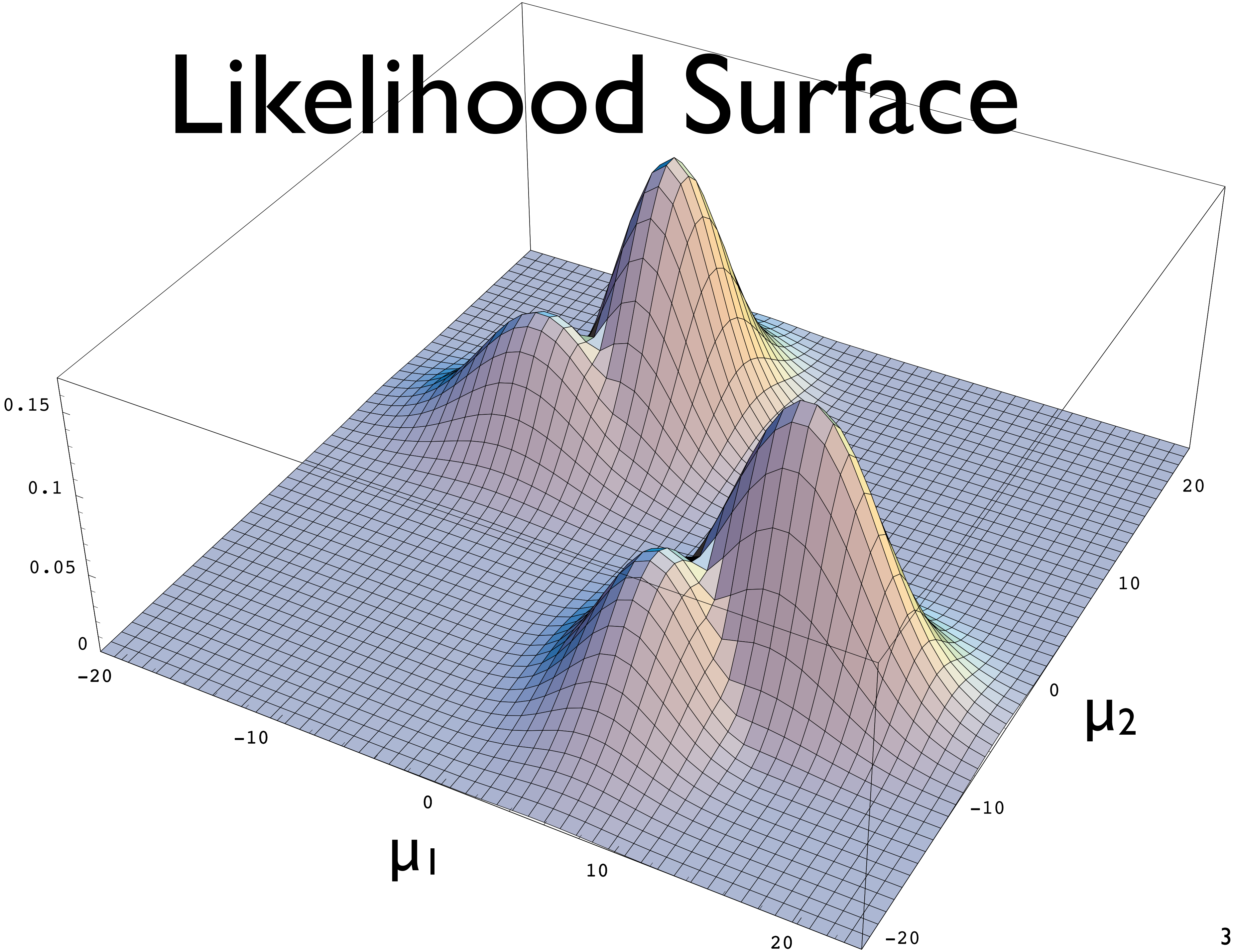
Likelihood $\tau_1 f(x|\mu_1, \sigma_1^2) + \tau_2 f(x|\mu_2, \sigma_2^2)$

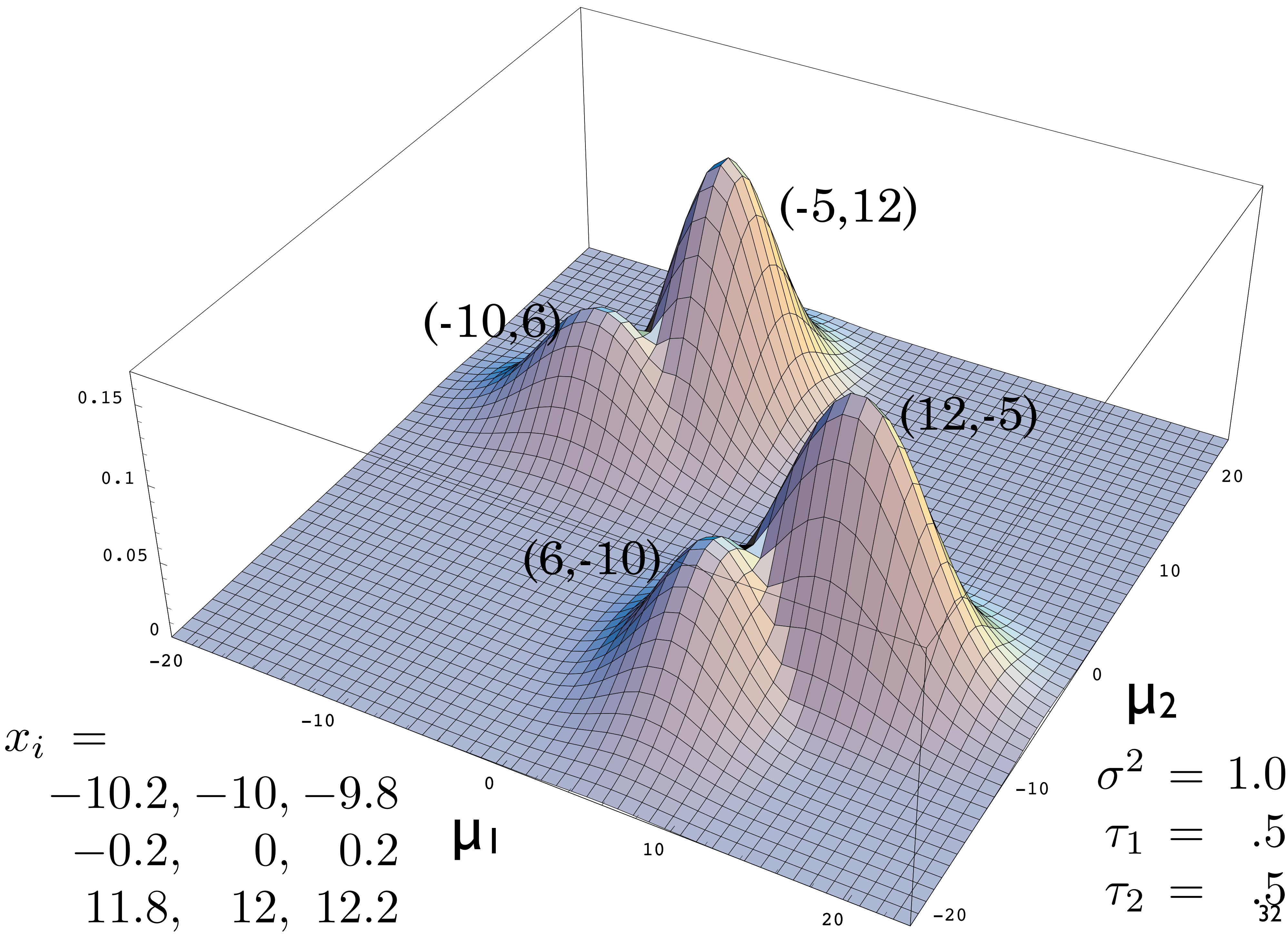
$$L(x_1, x_2, \ldots, x_n | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2)$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

No closed-form max

Likelihood Surface




(-5,12)
(-10,6)
(12,-5)
(6,-10)
0.15
0.1
0.05
0
-20
-10
0
10
20
μ1
μ2
$x_i =$
$-10.2, -10, -9.8$
$-0.2, 0, 0.2$
$11.8, 12, 12.2$
$\sigma^2 = 1.0$
$\tau_1 = .5$
$\tau_2 = .5$

A What-If Puzzle

Likelihood

$$L(x_1, x_2, \ldots, x_n | \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

Messy: no closed form solution known for finding θ maximizing L

But *what if* we knew the *hidden data*?

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

A Hat Trick

Two slips of paper in a hat:

Pink: $\mu = 3$, and

Blue: $\mu = 7$.

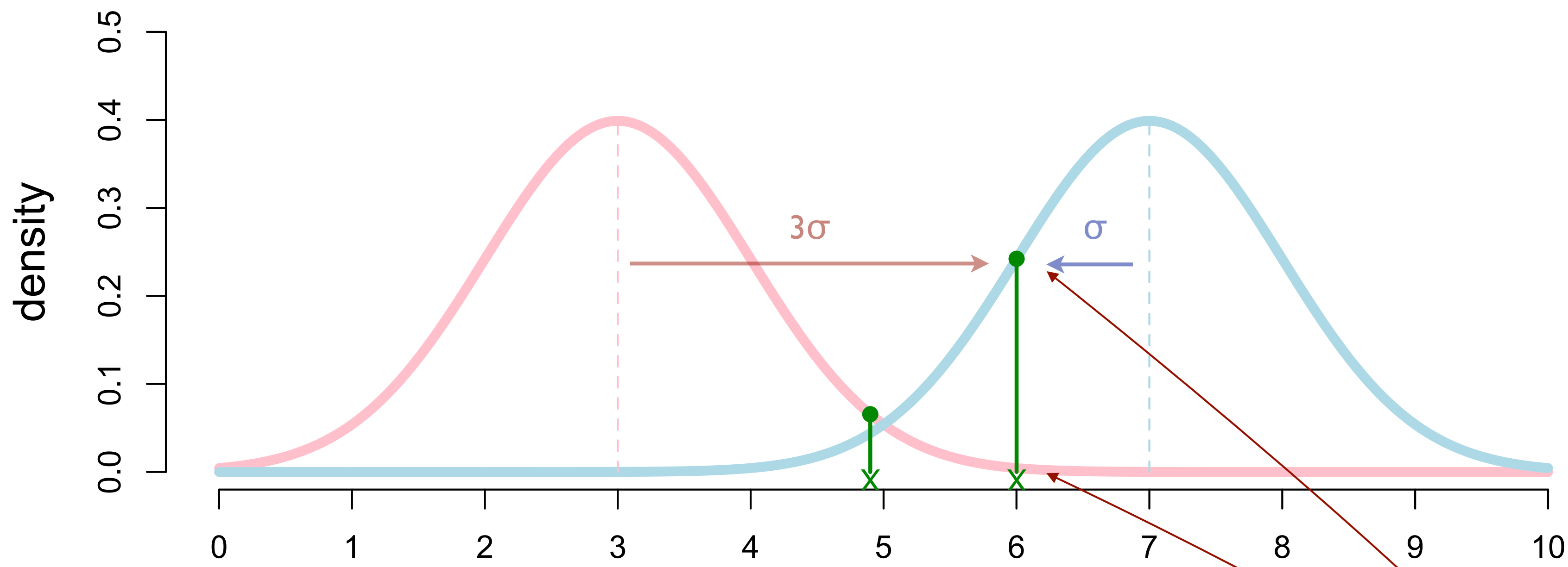
You draw one, then (without revealing color or μ) reveal a single sample $X \sim \text{Normal}(\text{mean } \mu, \sigma^2 = 1)$.

You happen to draw $X = 6.001$.

Dr. Mean says "your slip = 7." What is P(correct)?

What if X had been 4.9?

A Hat Trick



Let "$X \approx 6$" be a shorthand for $6.001 - \delta/2 < X < 6.001 + \delta/2$

$$P(\mu = 7 | X = 6) = \lim_{\delta \to 0} P(\mu = 7 | X \approx 6)$$

$$P(\mu = 7 | X \approx 6) = \frac{P(X \approx 6 | \mu = 7) P(\mu = 7)}{P(X \approx 6)}$$

Bayes rule

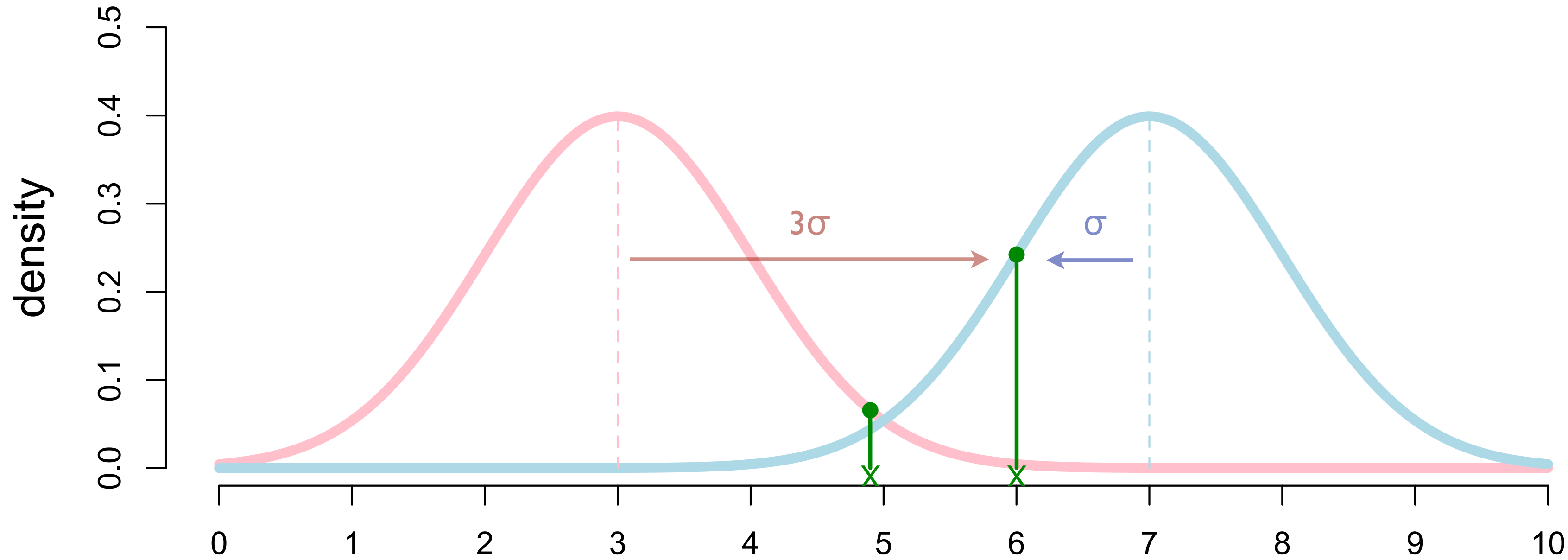
$$= \frac{0.5 P(X \approx 6 | \mu = 7)}{0.5 P(X \approx 6 | \mu = 3) + 0.5 P(X \approx 6 | \mu = 7)}$$

$$\approx \frac{f(X = 6 | \mu = 7)\delta}{f(X = 6 | \mu = 3)\delta + f(X = 6) | \mu = 7)\delta}, \text{ so}$$

$$P(\mu = 7 | X = 6) = \frac{f(X = 6 | \mu = 7)}{f(X = 6 | \mu = 3) + f(X = 6) | \mu = 7)} \approx 0.982$$

f = normal density

A Hat Trick



Alternate View:

f = normal density

$$\text{Posterior odds} = \text{Bayes Factor} \cdot \text{Prior odds}$$

$$\frac{P(\mu = 7|X = 6)}{P(\mu = 3|X = 6)} = \frac{f(X = 6|\mu = 7)}{f(X = 6|\mu = 3)} \cdot \frac{0.50}{0.50} = \frac{0.2422}{0.0044} \cdot \frac{1}{1} = \frac{54.8}{1}$$

I.e., 50:50 prior odds become 54:1 in favor of μ=7, given X=6.001

(and would become 3:2 in favor of μ=3, given X=4.9)

Another Hat Trick

Two secret numbers, μ_{pink} and μ_{blue}

On pink slips, many samples of Normal(μ_{pink}, $\sigma^2 = 1$),

Ditto on blue slips, from Normal(μ_{blue}, $\sigma^2 = 1$).

Based on 16 of each, how would you "guess" the secrets (where "success" means your guess is within ±0.5 of each secret)?

Roughly how likely is it that you will succeed?

Another Hat Trick (cont.)

Pink/blue = red herrings; separate & independent

Given $X_1, \ldots, X_{16} \sim N(\mu, \sigma^2)$, $\sigma^2 = 1$

Calculate $Y = (X_1 + \ldots + X_{16})/16 \sim N(\ ?\ ,\ ?\ )$

$E[Y] = \mu$ →

$Var(Y) = 16\sigma^2/16^2 = \sigma^2/16 = 1/16$ →

I.e., X_i's are all $\sim N(\mu, 1)$; Y is $\sim N(\mu, 1/16)$

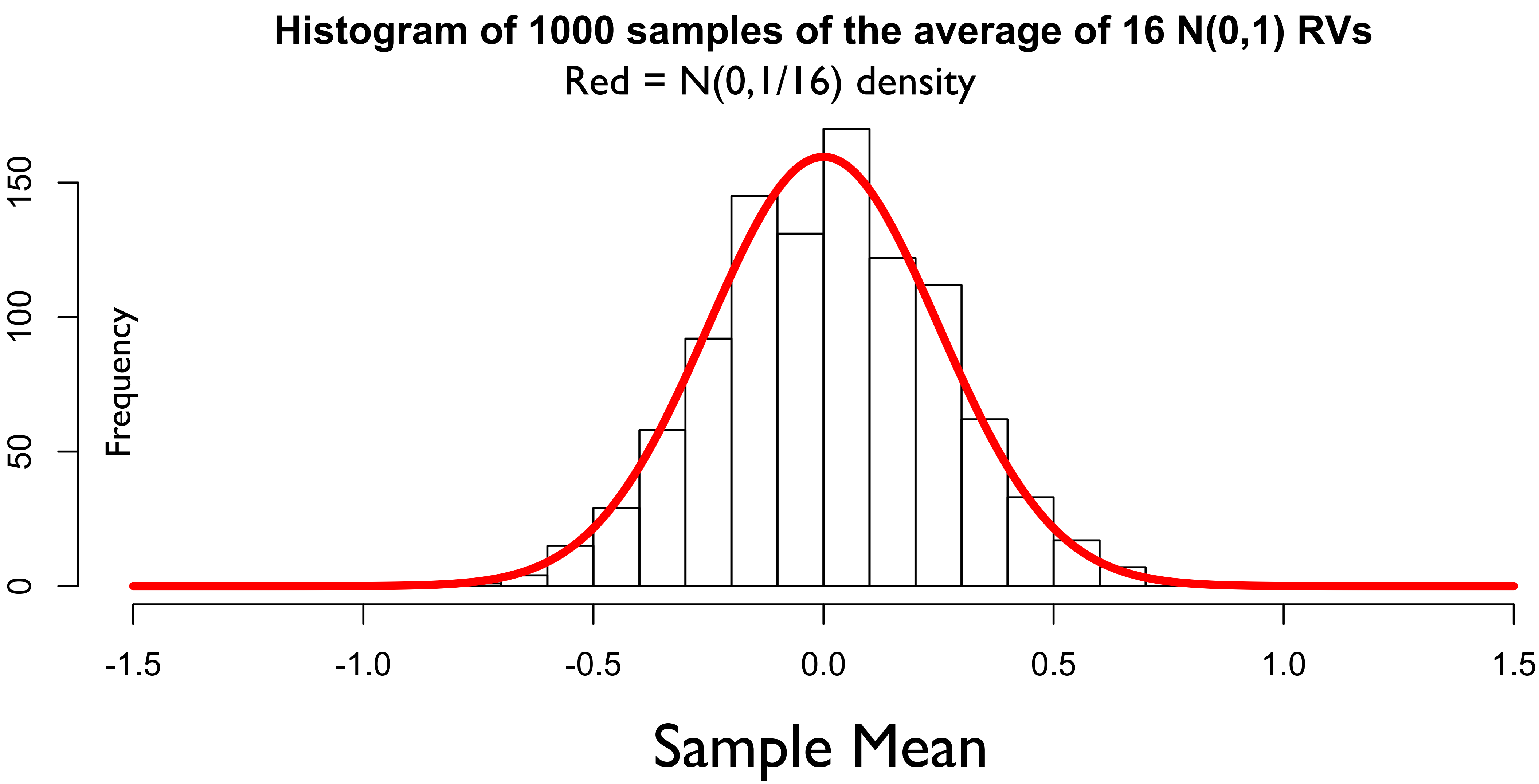
and since 0.5 = 2 sqrt(1/16), we have:

"Y within ±.5 of μ" = "Y within ±2 σ of μ" ≈ 95% prob

Note 1: Y is a *point estimate* for μ;

Y ± 2 σ is a *95% confidence interval* for μ

(More on this topic later)

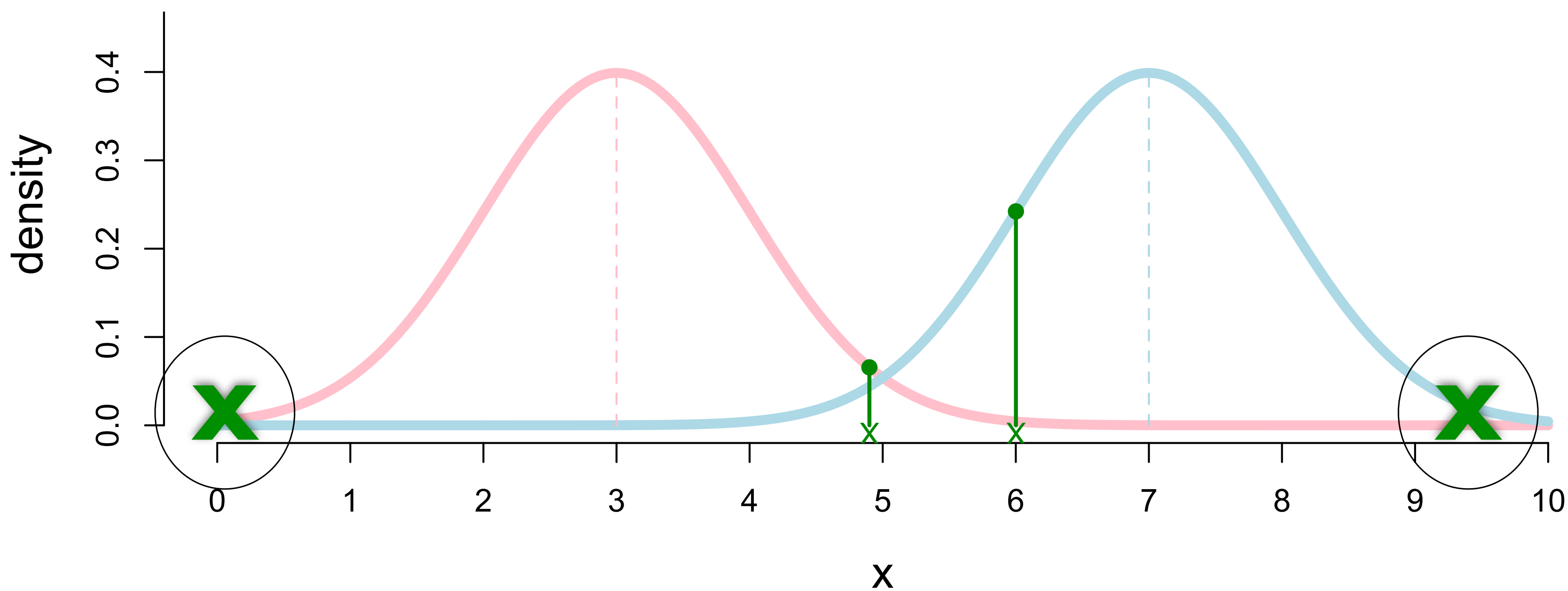

Histogram of 1000 samples of the average of 16 N(0,1) RVs
Red = N(0,1/16) density
Frequency
0
50
100
150
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Sample Mean

Hat Trick 2 (cont.)

Note 2:

What would you do if some of the slips you pulled had coffee spilled on them, obscuring color?

- If they were half way between means of the others?
- If they were on opposite sides of the means of the others



A What-If Puzzle

Likelihood

$$L(x_1, x_2, \ldots, x_n | \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta})$$

$$= \prod_{i=1}^{n} \sum_{j=1}^{2} \tau_j f(x_i | \mu_j, \sigma_j^2)$$

Messy: no closed form solution known for finding θ maximizing L

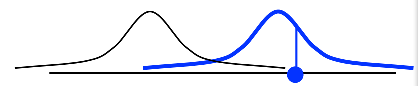
But *what if* we knew the *hidden data*?

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

EM as Egg vs Chicken

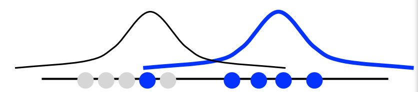
Hat Trick 1

- *IF* parameters θ known, could estimate z_{ij}
  - E.g., $|x_i - \mu_1|/\sigma_1 \gg |x_i - \mu_2|/\sigma_2 \Rightarrow P[z_{i1}=1] \ll P[z_{i2}=1]$



Hat Trick 2

- *IF* z_{ij} known, could estimate parameters θ
  - E.g., only points in cluster 2 influence μ_2, σ_2



- But we know neither; (optimistically) iterate:
  - Hat Trick 1 — E-step: calculate expected z_{ij}, given parameters
  - Hat Trick 2 — M-step: calculate "MLE" of parameters, given $E(z_{ij})$
- Overall, a clever "hill-climbing" strategy

Not what's needed for homework, but may help clarify concepts

Simple Version: "Classification EM"

If $E[z_{ij}] < .5$, pretend $z_{ij} = 0$; $E[z_{ij}] > .5$, pretend it's 1

I.e., *classify* points as component 1 or 2

Now recalc θ, assuming that partition (standard MLE)

Then recalc $E[z_{ij}]$, assuming that θ

Then re-recalc θ, assuming new $E[z_{ij}]$, etc., etc.

"K-means clustering," essentially

"Full EM" is slightly more involved, (to account for uncertainty in classification) but this is the crux.

Full EM

x_i's are known; θ unknown. Goal is to find MLE θ of:

$$L(x_1, \ldots, x_n \mid \theta) \qquad \text{(hidden data likelihood)}$$

Would be easy *if* z_{ij}'s were known, i.e., consider:

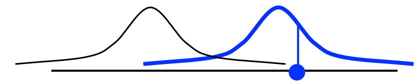
$$L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta) \qquad \text{(complete data likelihood)}$$

But z_{ij}'s aren't known.

Instead, maximize *expected* likelihood of visible data

$$E(L(x_1, \ldots, x_n, z_{11}, z_{12}, \ldots, z_{n2} \mid \theta)),$$

where expectation is over distribution of hidden data (z_{ij}'s)



The E-step:

Find $E(z_{ij})$, i.e., $P(z_{ij}=1)$

Assume θ known & fixed

A (B): the event that x_i was drawn from f_1 (f_2)

D: the observed datum x_i

Expected value of z_{i1} is P(A|D) — $E = 0 \cdot P(0) + 1 \cdot P(1)$

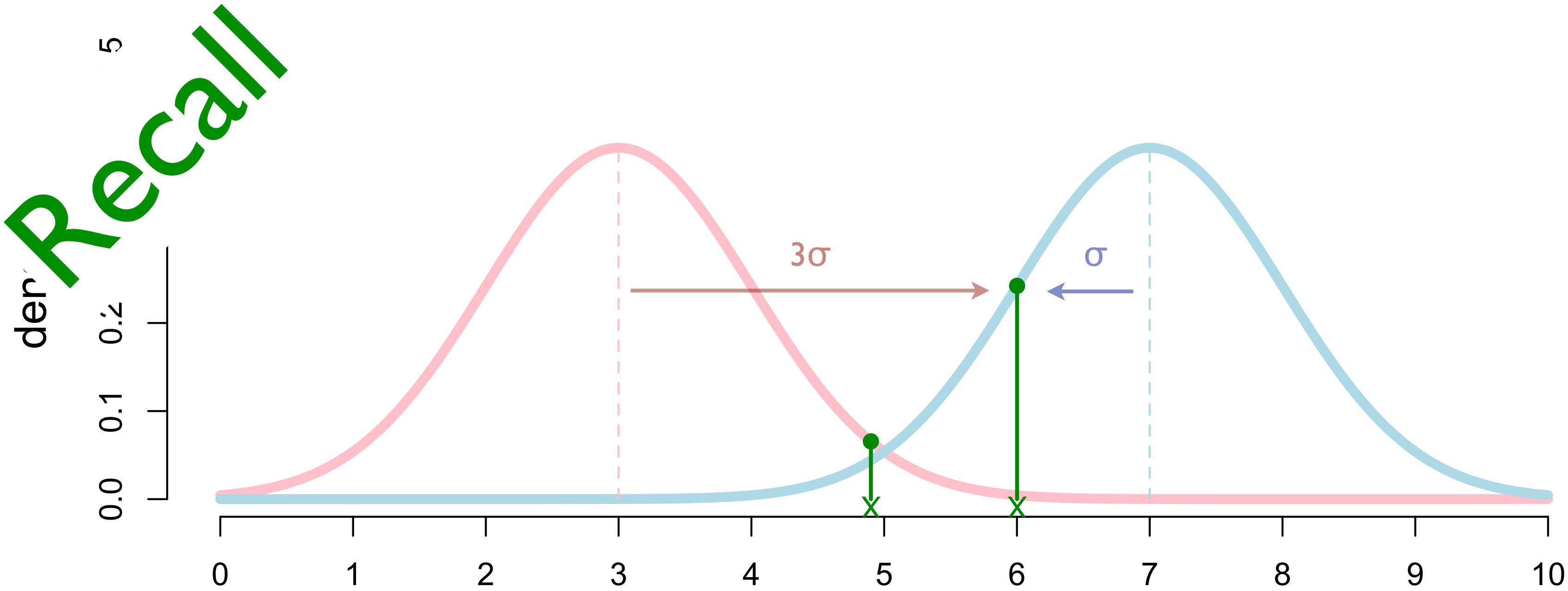
$$E[z_{i1}] = P(A|D) = \frac{P(D|A)P(A)}{P(D)}$$

$$\begin{aligned} P(D) &= P(D|A)P(A) + P(D|B)P(B) \\ &= f_1(x_i|\theta_1)\,\tau_1 + f_2(x_i|\theta_2)\,\tau_2 \end{aligned}$$

Repeat for each $z_{i,j}$

Note: denominator = sum of numerators – i.e. that which normalizes sum to 1 (typical Bayes)

A Hat Trick



Let "$X \approx 6$" be a shorthand for $6.001 - \delta/2 < X < 6.001 + \delta/2$

$$P(\mu = 7 | X = 6) = \lim_{\delta \to 0} P(\mu = 7 | X \approx 6)$$

$$P(\mu = 7 | X \approx 6) = \frac{P(X \approx 6 | \mu = 7) P(\mu = 7)}{P(X \approx 6)}$$

$$= \frac{0.5 P(X \approx 6 | \mu = 7)}{0.5 P(X \approx 6 | \mu = 3) + 0.5 P(X \approx 6 | \mu = 7)}$$

$$\approx \frac{f(X = 6 | \mu = 7)\delta}{f(X = 6 | \mu = 3)\delta + f(X = 6) | \mu = 7)\delta}, \text{ so}$$

$$P(\mu = 7 | X = 6) = \frac{f(X = 6 | \mu = 7)}{f(X = 6 | \mu = 3) + f(X = 6) | \mu = 7)} \approx 0.982$$

$E[Z_{i,pink}]=?$

$E[Z_{i,blue}]=$ (→ 0.982)

f = normal density

Complete Data Likelihood

Recall:

$$z_{1j} = \begin{cases} 1 & \text{if } x_1 \text{ drawn from } f_j \\ 0 & \text{otherwise} \end{cases}$$

so, correspondingly,

$$L(x_1, z_{1j} \mid \theta) = \begin{cases} \tau_1 f_1(x_1 \mid \theta) & \text{if } z_{11} = 1 \\ \tau_2 f_2(x_1 \mid \theta) & \text{otherwise} \end{cases}$$

equal, if z_{ij} are 0/1

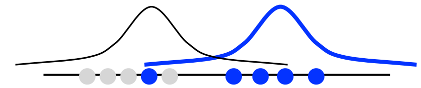
Formulas with "if's" are messy; can we blend more smoothly?
Yes, many possibilities. Idea 1:

$$L(x_1, z_{1j} \mid \theta) = z_{11} \cdot \tau_1 f_1(x_1 \mid \theta) + z_{12} \cdot \tau_2 f_2(x_1 \mid \theta)$$

Idea 2 (Better):

$$L(x_1, z_{1j} \mid \theta) = (\tau_1 f_1(x_1 \mid \theta))^{z_{11}} \cdot (\tau_2 f_2(x_1 \mid \theta))^{z_{12}}$$

M-step:



Find θ maximizing E(log(Likelihood))

(For simplicity, assume $\sigma_1 = \sigma_2 = \sigma; \tau_1 = \tau_2 = \tau = 0.5$)

$$L(\vec{x}, \vec{z} \mid \theta) = \prod_{i=1}^{n} \frac{\tau}{\sqrt{2\pi\sigma^2}} \exp\left(- \sum_{j=1}^{2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

$$E[\log L(\vec{x}, \vec{z} \mid \theta)] = E\left[\sum_{i=1}^{n} \left(\log \tau - \frac{1}{2} \log(2\pi\sigma^2) - \sum_{j=1}^{2} z_{ij} \frac{(x_i - \mu_j)^2}{2\sigma^2} \right) \right]$$

wrt dist of z_{ij}

$$= \sum_{i=1}^{n} \left(\log \tau - \frac{1}{2} \log(2\pi\sigma^2) - \sum_{j=1}^{2} E[z_{ij}] \frac{(x_i - \mu_j)^2}{2\sigma^2} \right)$$

Find θ maximizing this as before, using $E[z_{ij}]$ found in E-step. Result:

$\mu_j = \sum_{i=1}^{n} E[z_{ij}] x_i / \sum_{i=1}^{n} E[z_{ij}]$ (intuit: avg, weighted by subpop prob)

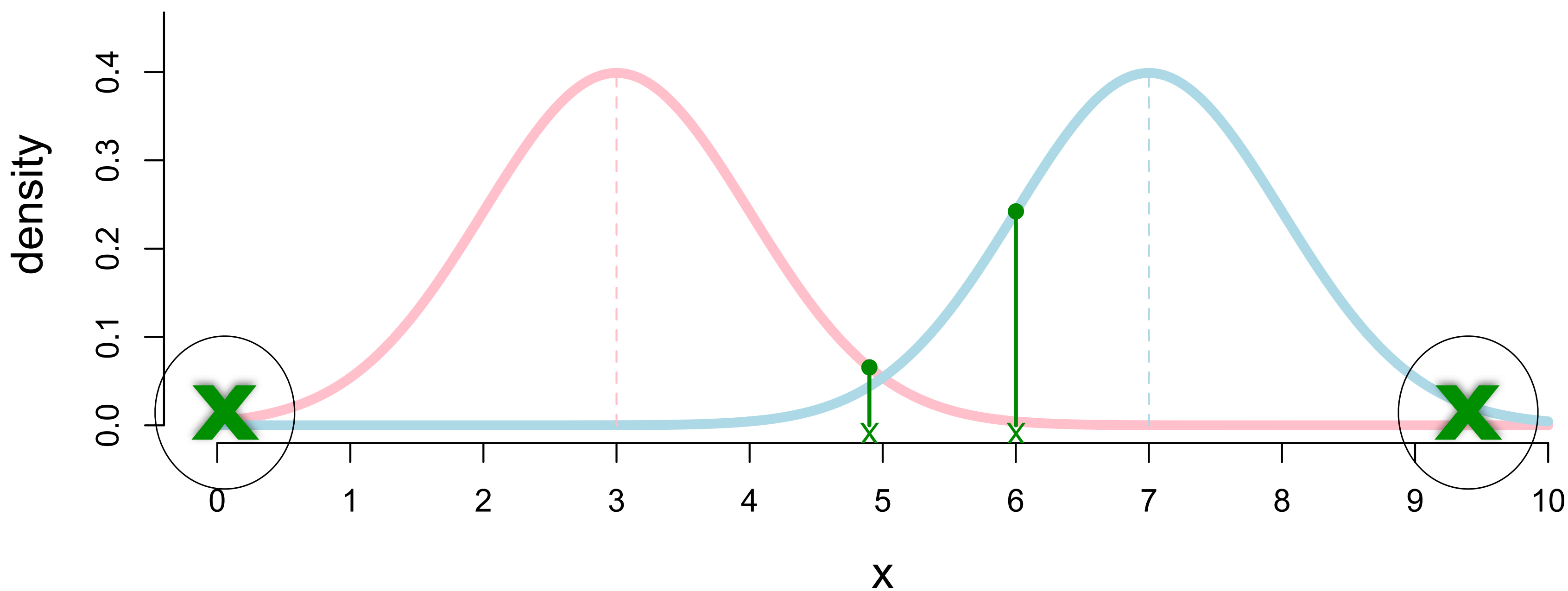
Recall

Hat Trick 2 (cont.)

Note 2: red/blue separation is just like the M-step of EM *if values of the hidden variables (z_{ij}) were known.*

What if they're not? E.g., what would you do if some of the slips you pulled had coffee spilled on them, obscuring color?

> If they were half way between means of the others?
> If they were on opposite sides of the means of the others



M-step:calculating mu's

$$\mu_j = \sum_{i=1}^{n} E[z_{ij}]x_i / \sum_{i=1}^{n} E[z_{ij}]$$

In words: μ_j is the average of the observed x_i's, weighted by the probability that x_i was sampled from component j.

							row sum	avg
old E's: $E[z_{i1}]$	0.99	0.98	0.7	0.2	0.03	0.01	2.91	
old E's: $E[z_{i2}]$	0.01	0.02	0.3	0.8	0.97	0.99	3.09	
x_i	9	10	11	19	20	21	90	15
$E[z_{i1}]x_i$	8.9	9.8	7.7	3.8	0.6	0.2	31.0	10.66 (new μ's)
$E[z_{i2}]x_i$	0.1	0.2	3.3	15.2	19.4	20.8	58.98	19.09 (new μ's)

2 Component Mixture

$$\sigma_1 = \sigma_2 = 1; \;\; \tau = 0.5$$

		mu1	-20.00		-6.00		-5.00		-4.99
		mu2	6.00		0.00		3.75		3.75
x1	-6	z11		5.11E-12		1.00E+00		1.00E+00	
x2	-5	z21		2.61E-23		1.00E+00		1.00E+00	
x3	-4	z31		1.33E-34		9.98E-01		1.00E+00	
x4	0	z41		9.09E-80		1.52E-08		4.11E-03	
x5	4	z51		6.19E-125		5.75E-19		2.64E-18	
x6	5	z61		3.16E-136		1.43E-21		4.20E-22	
x7	6	z71		1.62E-147		3.53E-24		6.69E-26	

Essentially converged in 2 iterations

(Excel spreadsheet on course web)

EM Summary

Fundamentally a maximum likelihood parameter estimation problem; broader than just Gaussian

Useful if 0/1 hidden data, and if analysis would be more tractable if hidden data z were known

Iterate:

E-step: estimate E(z) for each z, given θ

M-step: estimate θ maximizing E[log likelihood] given E[z] [where "E[logL]" is wrt random z ~ E[z] = p(z=1)]

Bayes

MLE

EM Issues

Under mild assumptions, EM is guaranteed to increase likelihood with every E-M iteration, hence will *converge*.

But it may converge to a *local*, not global, max. (Recall the 4-bump surface...)

Issue is intrinsic (probably), since EM is often applied to problems (including clustering, above) that are *NP-hard* (so fast alg is unlikely)

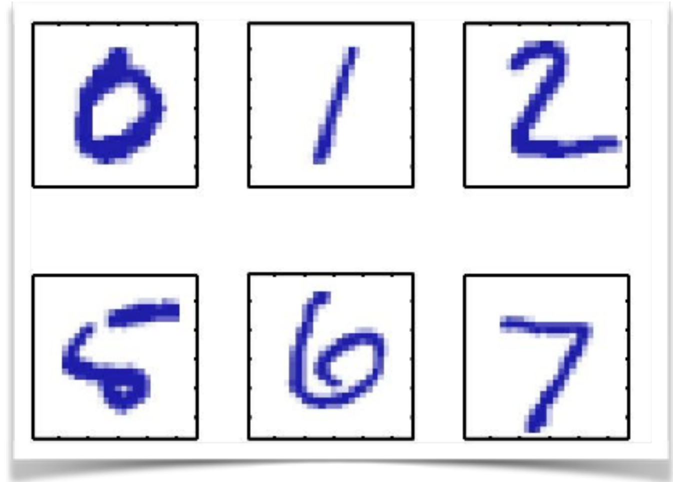
Nevertheless, widely used, often effective

Applications

- Clustering is a remarkably successful exploratory data analysis tool
  - Web-search, information retrieval, gene-expression, ...
  - Model-based approach above is one of the leading ways to do it
- Gaussian mixture models widely used
  - With many components, empirically match arbitrary distribution
  - Often well-justified, due to "hidden parameters" driving the visible data
- EM is extremely widely used for "hidden-data" problems
  - Hidden Markov Models – speech recognition, DNA analysis, ...

A “Machine Learning” Example

Handwritten Digit Recognition



Given: 10^4 unlabeled, scanned images of handwritten digits, say 25 x 25 pixels,

Goal: automatically classify new examples

Possible Method:

Each image is a point in $\mathbb{R}^{625}$; the “ideal” 7, say, is one such point; model other 7’s as a Gaussian cloud around it

Do EM, as above, but 10 components in 625 dimensions instead of 2 components in 1 dimension

“Recognize” a new digit by best fit to those 10 models, i.e., basically max E-step probability