

EM

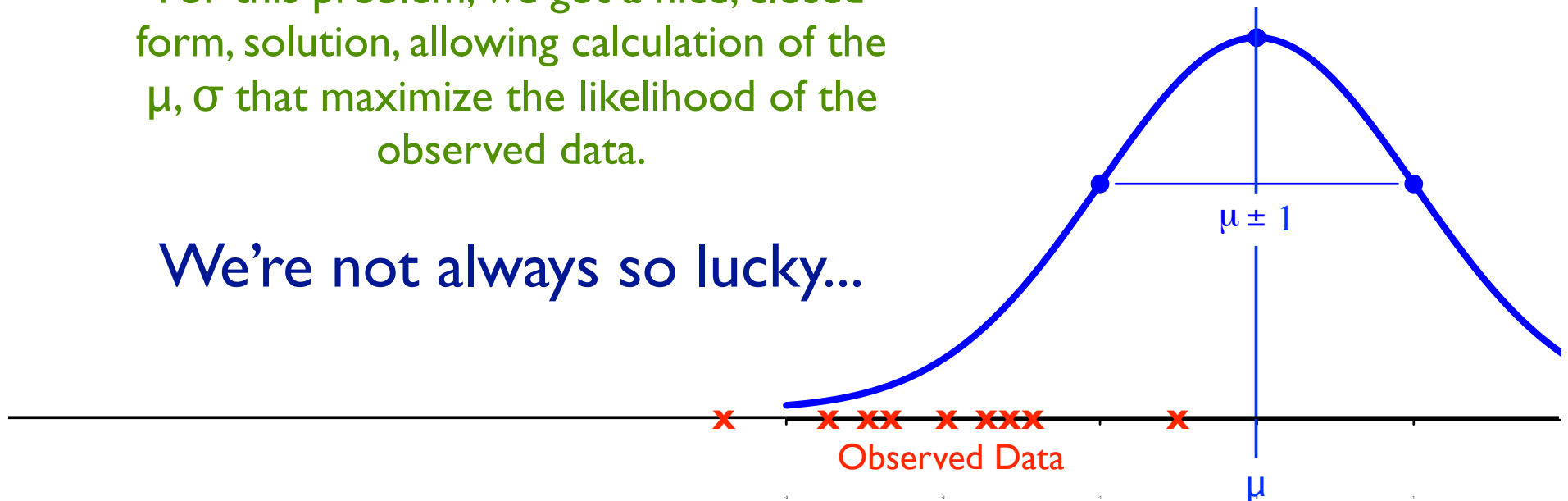
The Expectation-Maximization
Algorithm

We know:

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...

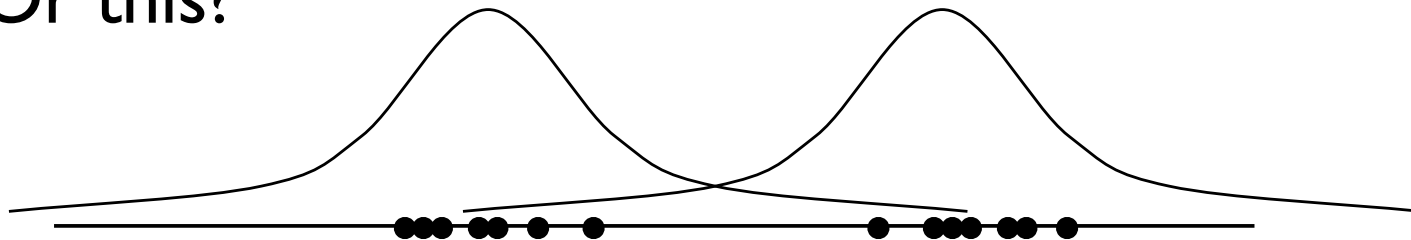
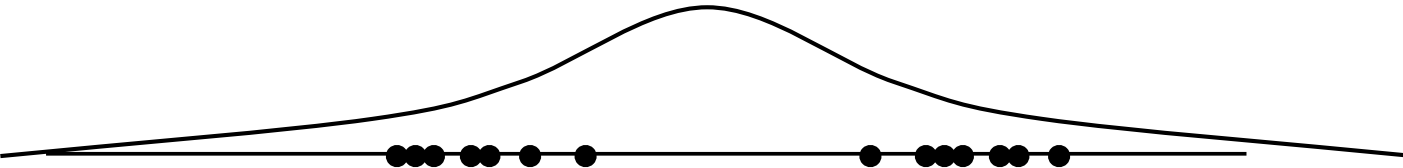


More Complex Example

This?



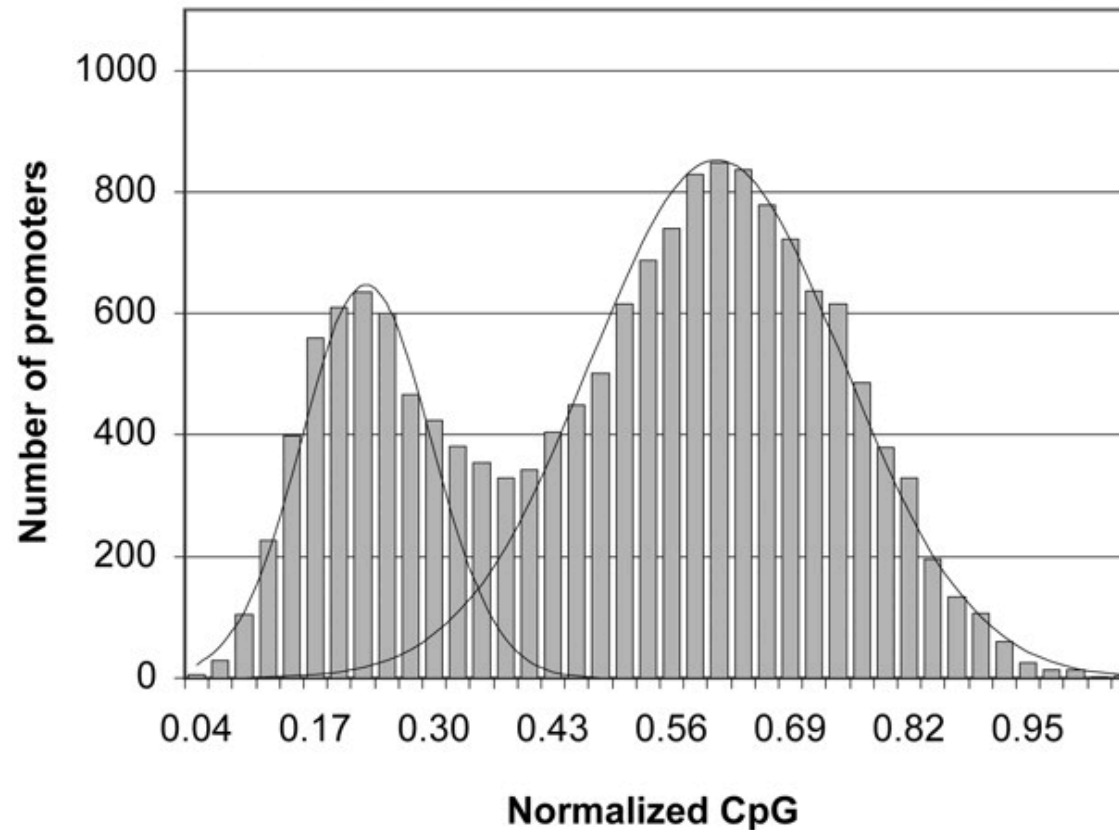
Or this?



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

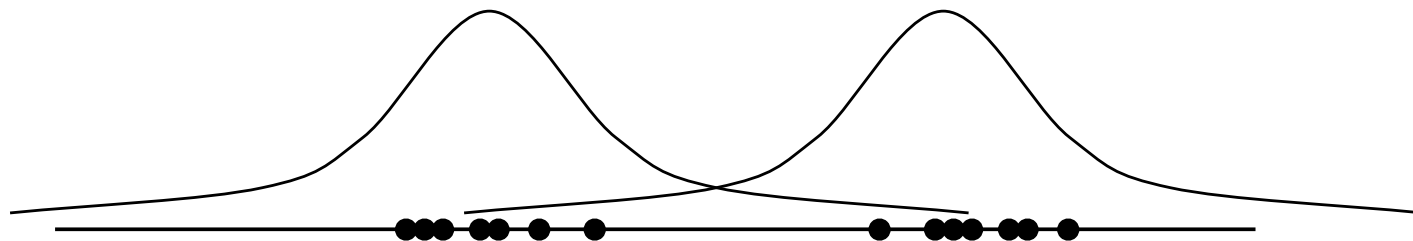
A Real Example:

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. $f(x|\mu_1, \sigma_1^2)$ $f(x|\mu_2, \sigma_2^2)$

Likelihood

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

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

No
closed-
form
max

A What-If Puzzle

Likelihood

$$L(x_1, x_2, \dots, x_n \mid \overbrace{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \tau_1, \tau_2}^{\theta})$$
$$= \prod_{i=1}^n \sum_{j=1}^2 \tau_j f(x_i \mid \mu_j, \sigma_j^2)$$

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

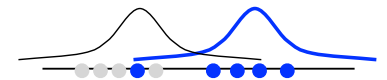
But *what if* we knew the *hidden data*?

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

EM as Egg vs Chicken

IF z_{ij} known, could estimate parameters θ

E.g., only points in cluster 2 influence μ_2, σ_2



IF parameters θ known, could estimate z_{ij}

E.g., if $|x_i - \mu_1|/\sigma_1 \ll |x_i - \mu_2|/\sigma_2$, then $z_{i1} \gg z_{i2}$



But we know neither; (optimistically) iterate:

E: calculate expected z_{ij} , given parameters

M: calc “MLE” of parameters, given $E(z_{ij})$

Overall, a clever “hill-climbing” strategy

Simple Version: “Classification EM”

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

I.e., *classify* points as component 0 or 1

Now recalc θ , assuming that partition

Then recalc z_{ij} , assuming that θ

Then re-recalc θ , assuming new z_{ij} , etc., etc.

“Full EM” is a bit more involved, but this is the crux.

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

EM Summary

Fundamentally a maximum likelihood parameter estimation problem

Useful if hidden data, and if analysis is more tractable when 0/1 hidden data z known

Iterate:

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

M-step: estimate θ maximizing $E(\log \text{likelihood})$
given $E(z)$ [where “ $E(\log L)$ ” is wrt random $z \sim E(z) = p(z=1)$]

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.

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

Nevertheless, widely used, often effective