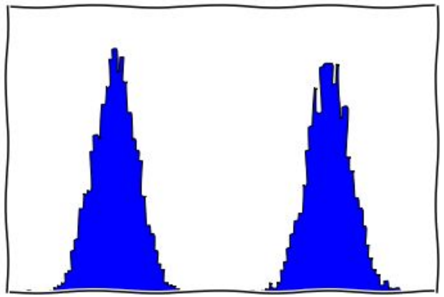
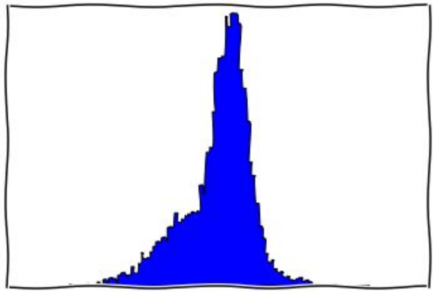
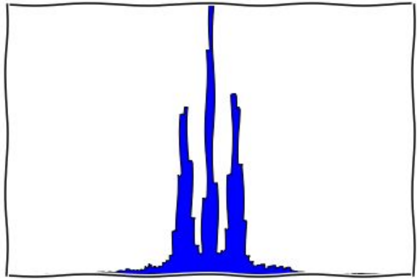






Mixture of Gaussians Models

Outline

- Inference, Learning, and Maximum Likelihood
- Why Mixtures? Why Gaussians?
- Building up to the Mixture of Gaussians
  - Single Gaussians
  - Fully-Observed Mixtures
  - Hidden Mixtures

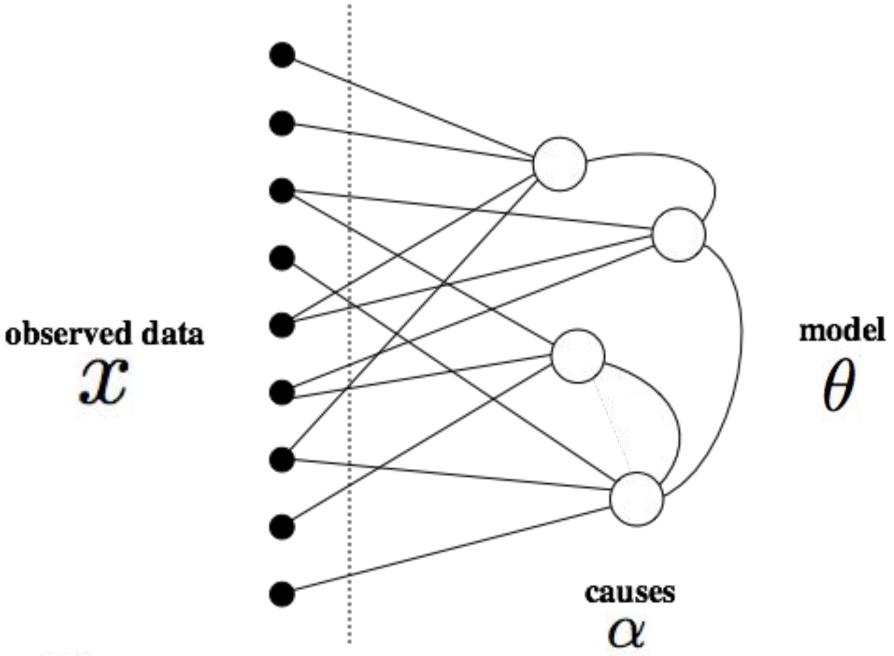
Perception Involves Inference and Learning

- Must **infer the hidden causes**, α, of sensory data, x
  - Sensory data: air pressure wave frequency composition, patterns of electromagnetic radiation
  - Hidden causes: proverbial tigers in bushes, lecture slides, sentences

- Must **learn the correct model** for the relationship between hidden causes and sensory data
  - Models will be **parameterized**, with parameters θ
  - We will use **quality of prediction** as our figure of merit

Generative models

inference $\longrightarrow$ $p(\alpha|x;\theta)$



$p(x|\alpha;\theta)$

$\longleftarrow$
explanation or prediction

Maximum Likelihood and Maximum a Posteriori

- The model parameters θ that make the data most probable are called the **maximum likelihood** parameters
- or hidden causes α or causes

INFERENCE → $$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}}\, p(\alpha|x;\theta)$$

LEARNING → $$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x;\theta)$$

$$p(x;\theta) = \sum_{\alpha} p(x,\alpha;\theta)$$

$$= \sum_{\alpha} p(x|\alpha;\theta)p(\alpha;\theta)$$

In practice, we maximize log-likelihoods

- Taking **logs doesn’t change the answer**

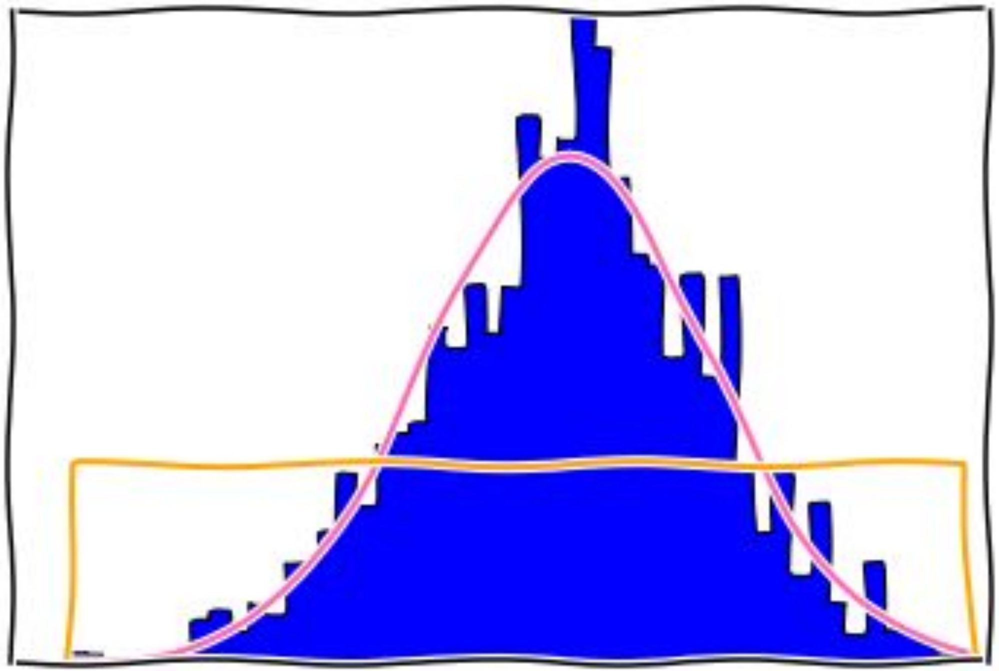
$$\hat{\alpha} = \operatorname*{argmax}_{\alpha} p(\alpha|x;\theta) = \operatorname*{argmax}_{\alpha} \log p(\alpha|x;\theta)$$

- Logs turn **multiplication into addition**

- Logs turn many natural operations on probabilities into linear algebra operations

- Negative log probabilities arise naturally in information theory

The Maximum Likelihood Answer Depends on Model Class



Outline

- Inference, Learning, and Maximum Likelihood
- Why Mixtures? Why Gaussians?
- Building up to the Mixture of Gaussians
  - Single Gaussians
  - Fully-Observed Mixtures
  - Hidden Mixtures

Why Mixtures?

What is a Mixture Model?

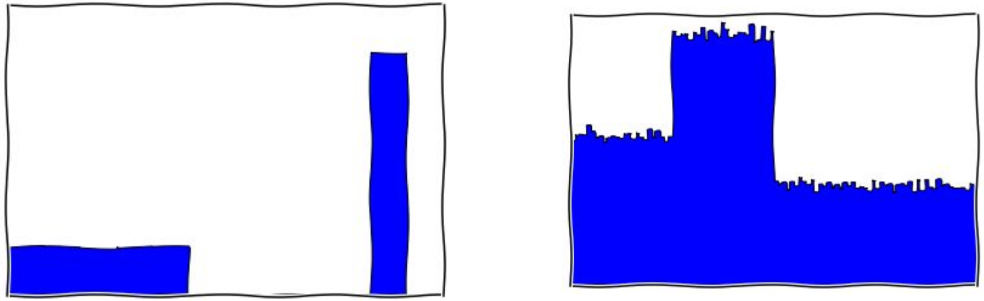
DATA $\rightarrow$ $$p(x;\theta,w) = \sum_{\alpha=1}^{K} p(x|\alpha;\theta_\alpha)\, p(\alpha;w_\alpha)$$

LIKELIHOOD $\rightarrow$ $$p(x|\alpha;\theta_\alpha)$$

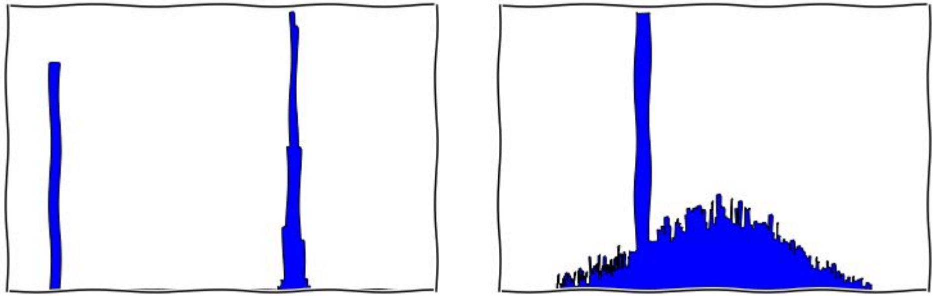
PRIOR $\rightarrow$ $$p(\alpha;w_\alpha) = w_\alpha$$

- This is precisely analogous to **using a basis to approximate a vector**

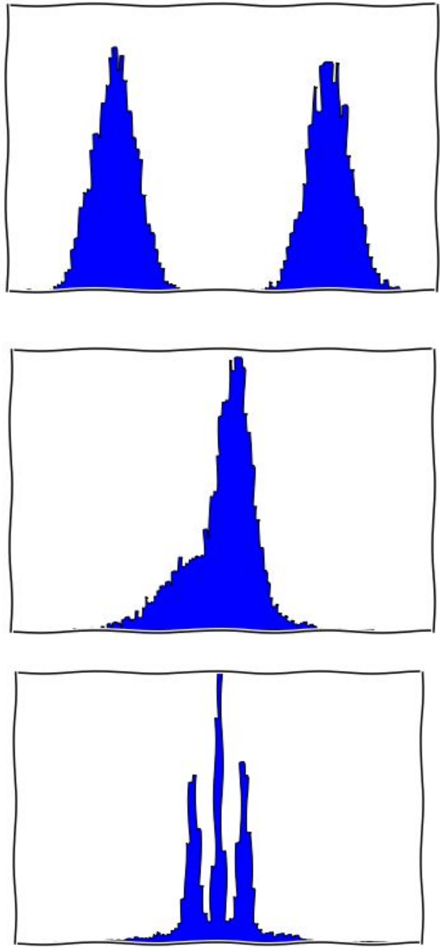
Example Mixture Datasets



Mixtures of Uniforms



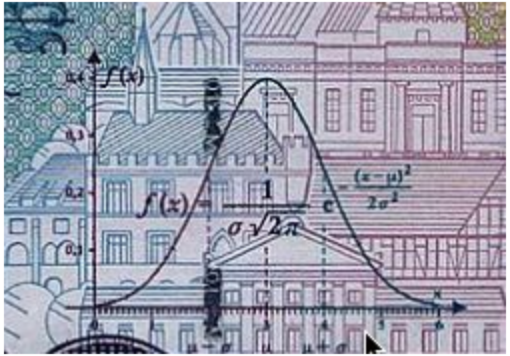
Spike-And-Gaussian Mixtures



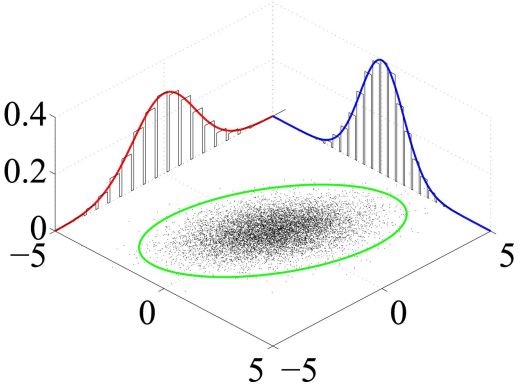
Mixtures of Gaussians

Why Gaussians?

Why Gaussians?



$$p(x; \mu, \sigma^2) \propto \mathrm{e}^{-\frac{1}{2}(x-\mu)^2\sigma^{-2}}$$



$$p(\mathbf{x}; \mu, \Sigma) \propto \mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

Wikipedia

Why Gaussians? An unhelpfully terse answer.

- Gaussians satisfy a particular differential equation:

$$\frac{d}{dx}p(x) = -xp(x)$$

- From this differential equation, all the properties of the Gaussian family can be derived *without solving for the explicit form.*
  - Gaussians are isotropic, Fourier transform of a Gaussian is a Gaussian, sum of Gaussian RVs is Gaussian, Central Limit Theorem

- See this blogpost for details: http://bit.ly/gaussian-diff-eq

Why Gaussians?

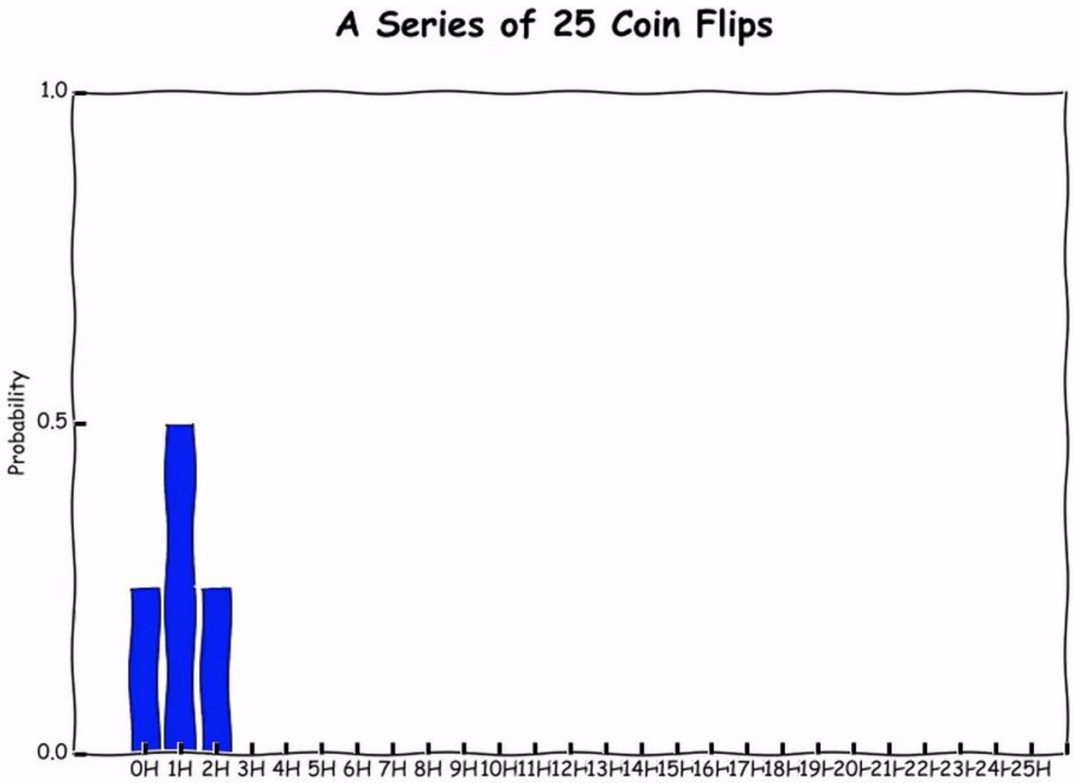
- Gaussians are everywhere, thanks to the **Central Limit Theorem**
- Gaussians are the **maximum entropy** distribution with a given center (mean) and spread (std dev)
- Inference on Gaussians is **linear algebra**

Central Limit Theorem

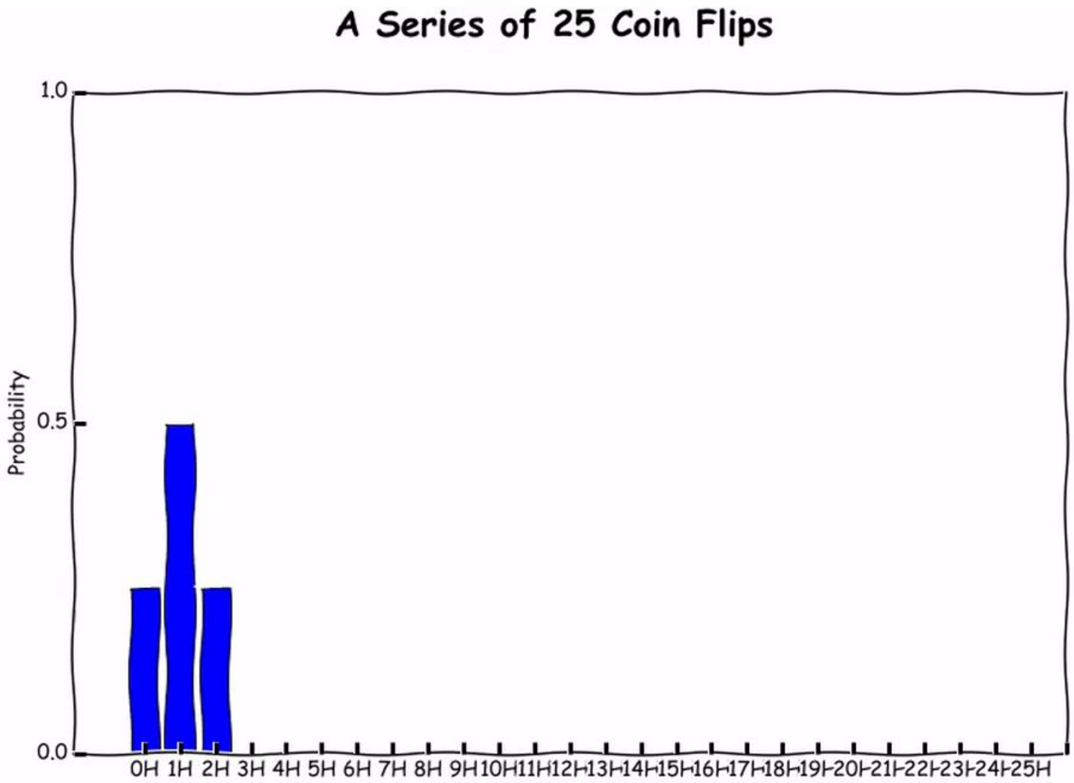
- Statistics: adding up independent random variables with finite variances results in a Gaussian distribution

- Science: if we assume that many small, independent random factors produce the noise in our results, we should see a Gaussian distribution

Central Limit Theorem in Action



Central Limit Theorem in Action



Why Gaussians?

- Gaussians are everywhere, thanks to the Central Limit Theorem
- Gaussians are the maximum entropy distribution with a given center (mean) and spread (std dev)
- Inference on Gaussians is linear algebra

Gaussians are a natural MAXENT distribution

- The principle of maximum entropy (MAXENT) will be covered in detail later
- Teaser: MAXENT maps statistics of data to probability distributions in a principled, faithful manner
- For the most common choice of statistic, mean ± s.d., the MAXENT is a Gaussian

Why Gaussians?

- Gaussians are everywhere, thanks to the Central Limit Theorem
- Gaussians are the maximum entropy distribution with a given center (mean) and spread (std dev)
- Inference on Gaussians is linear algebra

Inference with Gaussians is "just" linear algebra

- The log-probabilities of a Gaussian are a negative-definite quadratic form

$$\log p(\mathbf{x}; \mu, \Sigma) = -(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) - C$$

- Quadratic forms can be mapped onto matrices
- So solving an inference problem becomes solving a linear algebra problem
- Linear algebra is the Scottie Pippen of mathematics

Outline

- Inference, Learning, and Maximum Likelihood
- Why Mixtures? Why Gaussians?
- Building up to the Mixture of Gaussians
    - Single Gaussians
    - Fully-Observed Mixtures
    - Hidden Mixtures

What is a Gaussian Mixture Model?

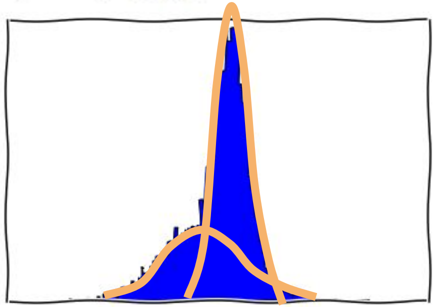
DATA $\rightarrow$ $$p(\mathbf{x}; \mu, \Sigma, w) = \sum_{\alpha=1}^{K} p(\mathbf{x}|\alpha; \mu_\alpha, \Sigma_\alpha) p(\alpha; w_\alpha)$$

LIKELIHOOD $\rightarrow$ $$p(\mathbf{x}|\alpha; \mu_\alpha, \Sigma_\alpha) = \frac{1}{Z} e^{-\frac{1}{2}((\mathbf{x}-\mu_\alpha)^T \Sigma_\alpha^{-1} (\mathbf{x}-\mu_\alpha))}$$

PRIOR $\rightarrow$ $$p(\alpha; w_\alpha) = w_\alpha$$

$$\text{Model parameters } \theta_\alpha = \{\mu_\alpha, \Sigma_\alpha, w_\alpha\}$$

Z is a normalization constant.



Example

Maximum Likelihood for Gaussian Mixture Models

Plan of Attack:

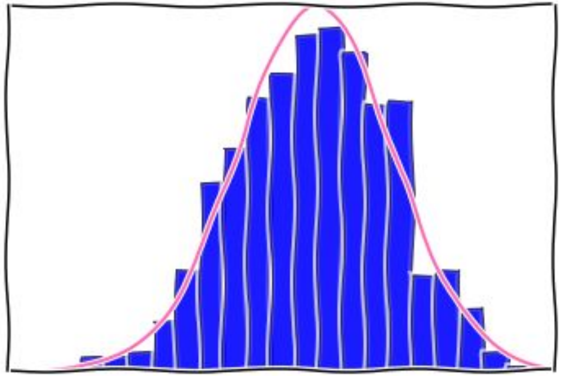
1. ML for a single Gaussian
2. ML for a fully-observed mixture
3. ML for a hidden mixture

Maximum Likelihood for a Single Gaussian

$$\mathcal{L}(\mathbf{x};\theta) := \langle \ell(\mathbf{x};\theta) \rangle := \langle \log p(\mathbf{x};\theta) \rangle$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x;\theta)$$

$$\theta = \hat{\theta} \leftrightarrow \frac{\partial \mathcal{L}}{\partial \theta} = 0$$



Maximum Likelihood for a Single Gaussian

$$\mathcal{L}(x\ ;\mu) = \langle \ell(x\ ;\mu) \rangle = \frac{1}{n} \sum_{\text{data}} \log p(x^i;\mu)$$

$$\ell(x^i;\mu) = -\frac{(x^i-\mu)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$$

$$\frac{\partial}{\partial\mu}\ell = -\frac{(x^i-\mu)}{\sigma^2} * \frac{\partial}{\partial\mu}(x^i-\mu)$$

$$\frac{\partial}{\partial\mu}\ell = \frac{(x^i-\mu)}{\sigma^2}$$

$$\Delta\mu = \frac{(x^i-\mu)}{\sigma^2}$$

$$\langle\Delta\mu\rangle = \frac{\langle x^i-\mu\rangle}{\sigma^2} = \frac{\langle x^i\rangle-\mu}{\sigma^2}$$

Maximum Likelihood for a Single Gaussian

$$\langle \Delta\mu \rangle = \frac{\langle x^i - \mu \rangle}{\sigma^2} = \frac{\langle x^i \rangle - \mu}{\sigma^2}$$

$$\frac{\partial}{\partial \mu}\ell = 0 \leftrightarrow \langle \Delta\mu \rangle = 0 \leftrightarrow \langle x^i \rangle - \mu = 0$$

$$\therefore \hat{\mu} = \langle x^i \rangle$$

By a similar argument:

$$\hat{\sigma}^2 = \langle (x^i - \mu)^2 \rangle$$

Maximum Likelihood for Gaussian Mixture Models

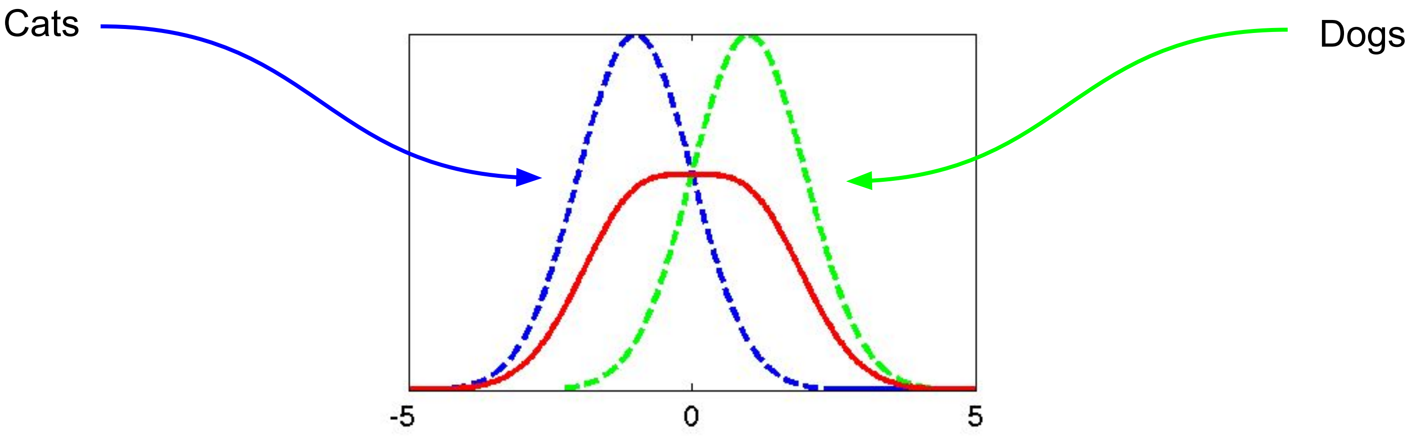
Plan of Attack:

1. ML for a single Gaussian
2. ML for a fully-observed mixture
3. ML for a hidden mixture

Maximum Likelihood for Fully-Observed Mixture

- “Observed Mixture” means we receive datapoints (x,α).

- Examples: classification (discrete), regression (continuous)



Memming Wordpress Blog

Maximum Likelihood for Fully-Observed Mixture

- **For each mixture element, the problem is exactly the same** - what are the parameters of a single Gaussian?

- Because we know which mixture each data point came from, **we can solve all these problems separately, using the same method** as for a single Gaussian.

- How do we figure out the mixture weights w?

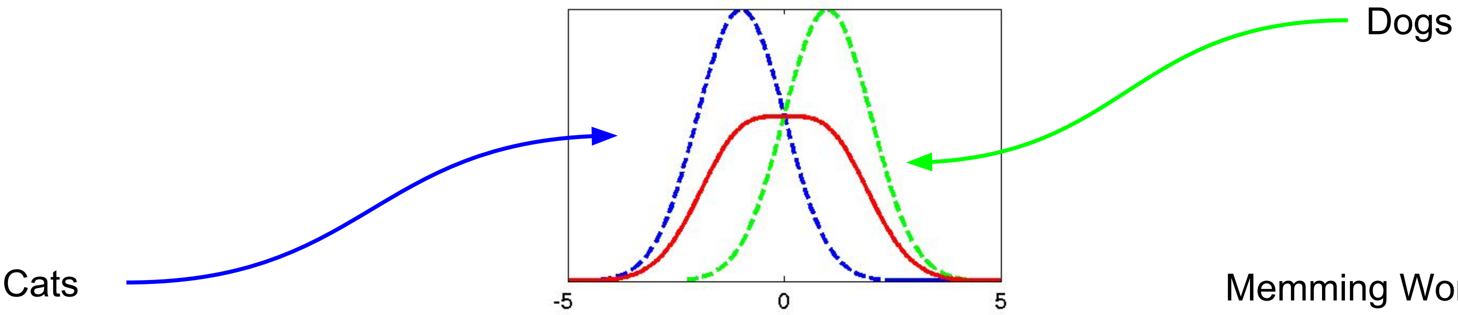
$$\hat{\mu}_\alpha = \langle x^i_\alpha \rangle$$

$$\hat{\sigma}^2_\alpha = \langle (x^i_\alpha - \mu_\alpha)^2 \rangle$$

$$p(\alpha; w_\alpha) = w_\alpha$$

Bonus: We Can Now Classify Unlabeled Datapoints

- We can **label new datapoints** x with a corresponding α using our model
- This is the key idea behind supervised learning approaches in general.
- How do we label them?
  - Max Likelihood method - find the closest mean (in z-score units), that's our label
  - Fully Bayesian method - maintain a distribution over the labels - $p(\alpha \mid x ; \theta)$



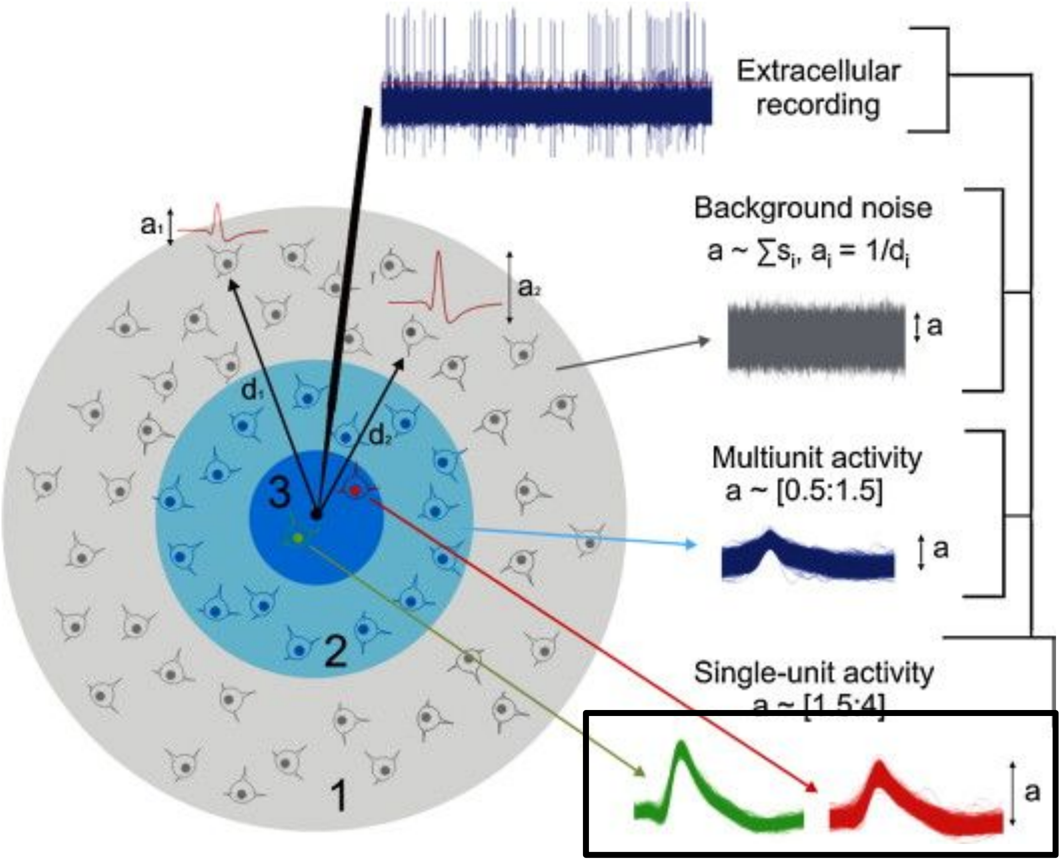
Memming Wordpress Blog

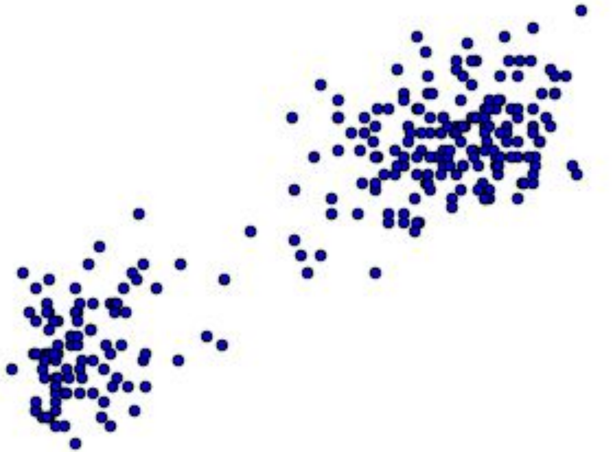
Maximum Likelihood for Gaussian Mixture Models

Plan of Attack:

1. ML for a single Gaussian
2. ML for a fully-observed mixture
3. ML for a hidden mixture

Hidden Variables Example: Spike Sorting





Martinez, Quiran Quiroga, et al., 2009
Journal of Neuroscience Methods

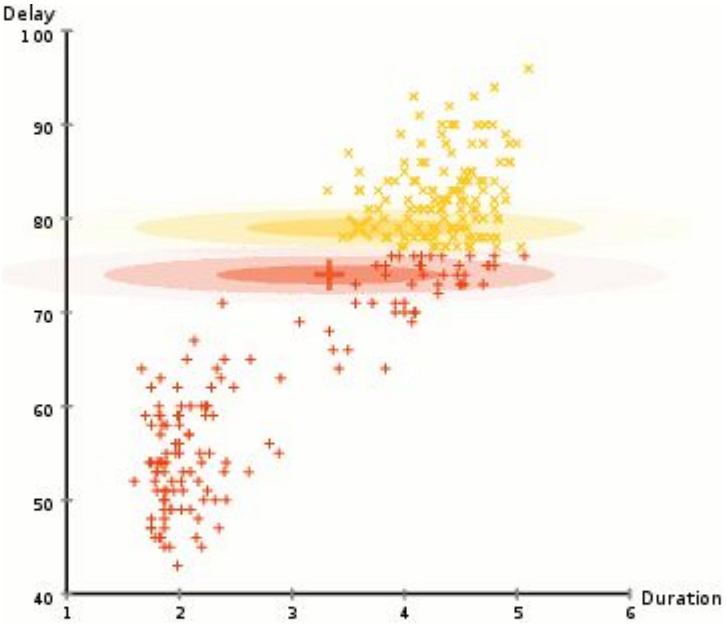
Maximum Likelihood for Models with Hidden Variables

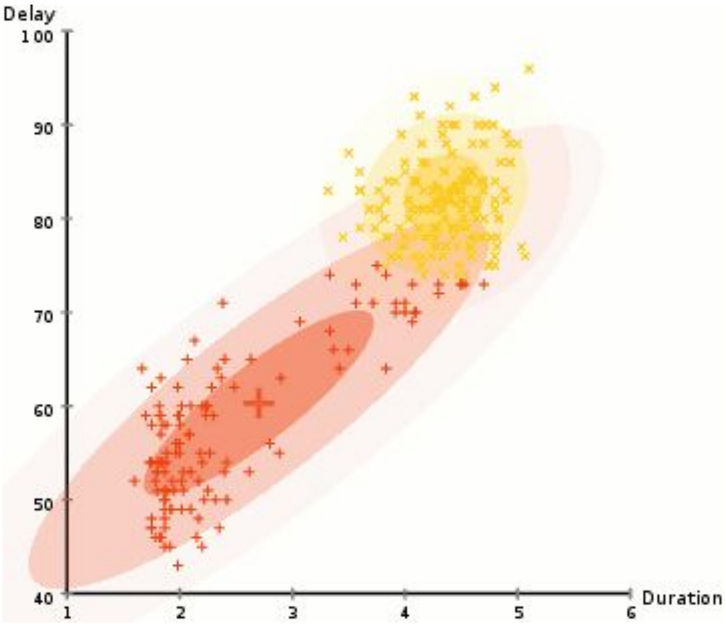
- $p(x \mid \mu, \Sigma, \alpha)$ is the same, but **now we don't have the labels** α.

- Problem: if we had the labels, we could find the parameters (just as before), and if we had the parameters, we could compute the labels (again, just as before). It's **a chicken-and-egg problem**!

- Solution: let's **just "make-believe"** we have the parameters.

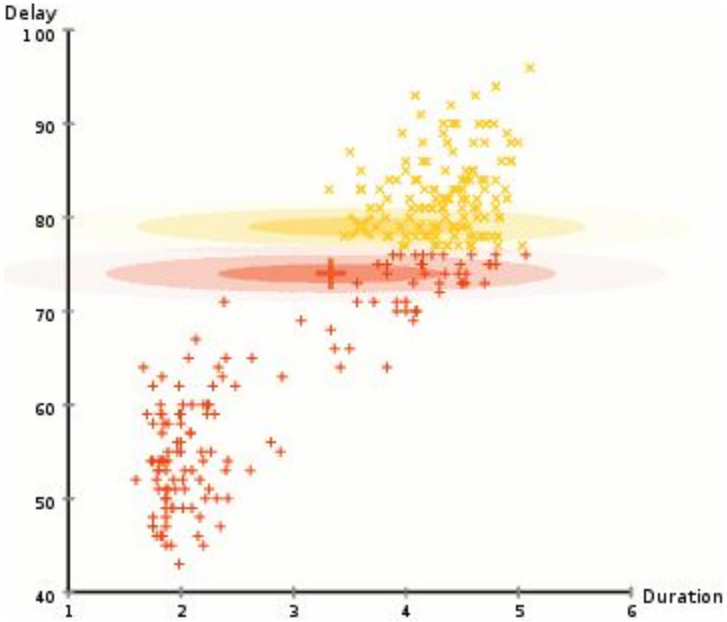
Our Clustering Algorithm on Spike Sorting





Wikipedia

Our Clustering Algorithm on Spike Sorting



Wikipedia

The K-Means Algorithm

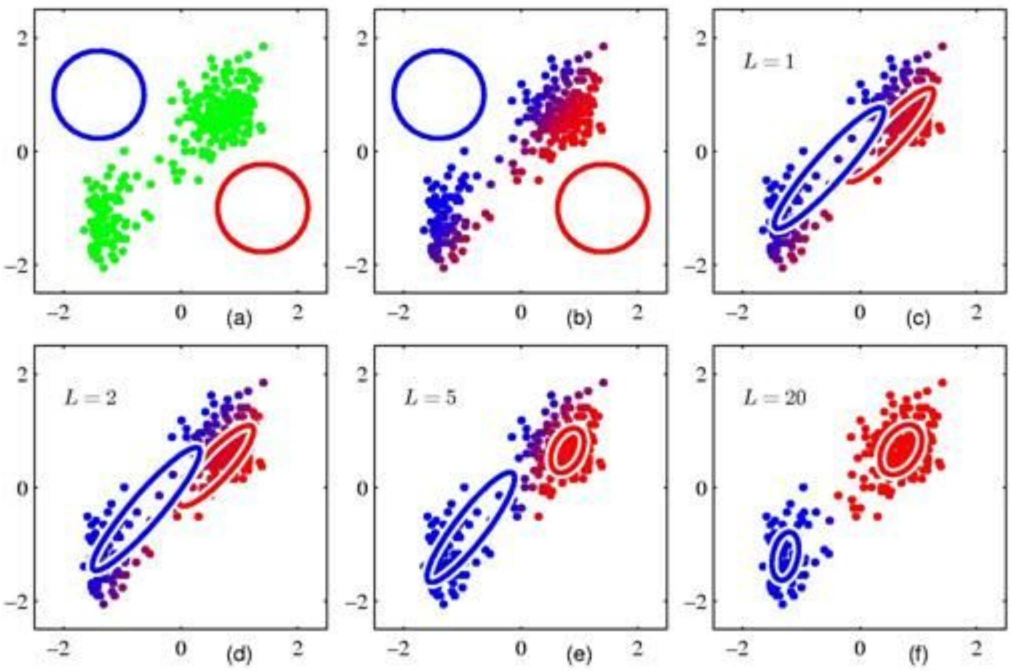
1. Make up K values for the means of the clusters
   - Usually initialized randomly
2. Assign datapoints to clusters
   - Each datapoint is assigned to the nearest cluster
3. Update the cluster means to the new empirical means
4. Repeat 2-4.

Issues with K-Means

1. Cluster assignment step (inference) is **not Bayesian**

2. Small changes in data can cause **big changes in behavior**

“Soft” Clustering?



Bishop - PRML

Expectation-Maximization for Means

1. Make up K values for the means
2. (E) Infer $p(\alpha|x)$ for each x and α
3. (M) Update the means via *weighted* average
    a. Weight the contribution of datapoint x by $p(\alpha|x)$
4. Repeat 2-4.

Full Expectation-Maximization

1. Make up K values for the means, covariances, and mixture weights
2. (E) Infer $p(\alpha|x)$ for each x and α
3. (M) Update the parameters with weighted averages
    a. Weight the contribution of datapoint x by $p(\alpha|x)$
4. Repeat 2-4.

E-Step: Bayes' Rule for Inference

$$p(\alpha|x;\theta) = \frac{p(x,\alpha;\theta)}{p(x;\theta)} = \frac{p(x|\alpha;\theta)p(\alpha;\theta)}{p(x;\theta)}$$

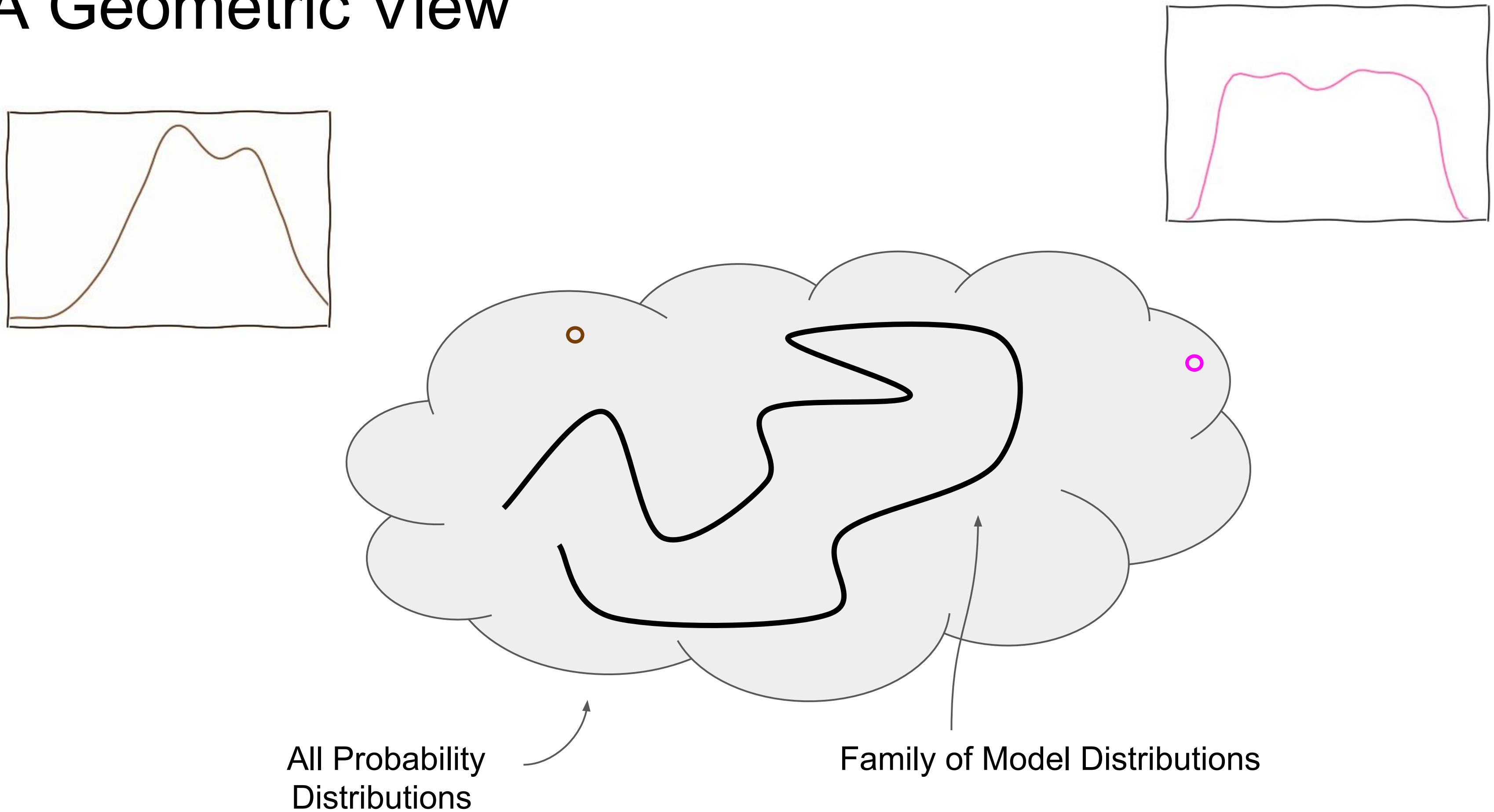
$$p(x;\theta) = \sum_{\alpha} p(x,\alpha;\theta) = \sum_{\alpha} p(x|\alpha;\theta)p(\alpha;\theta)$$

M-Step: Direct Maximization

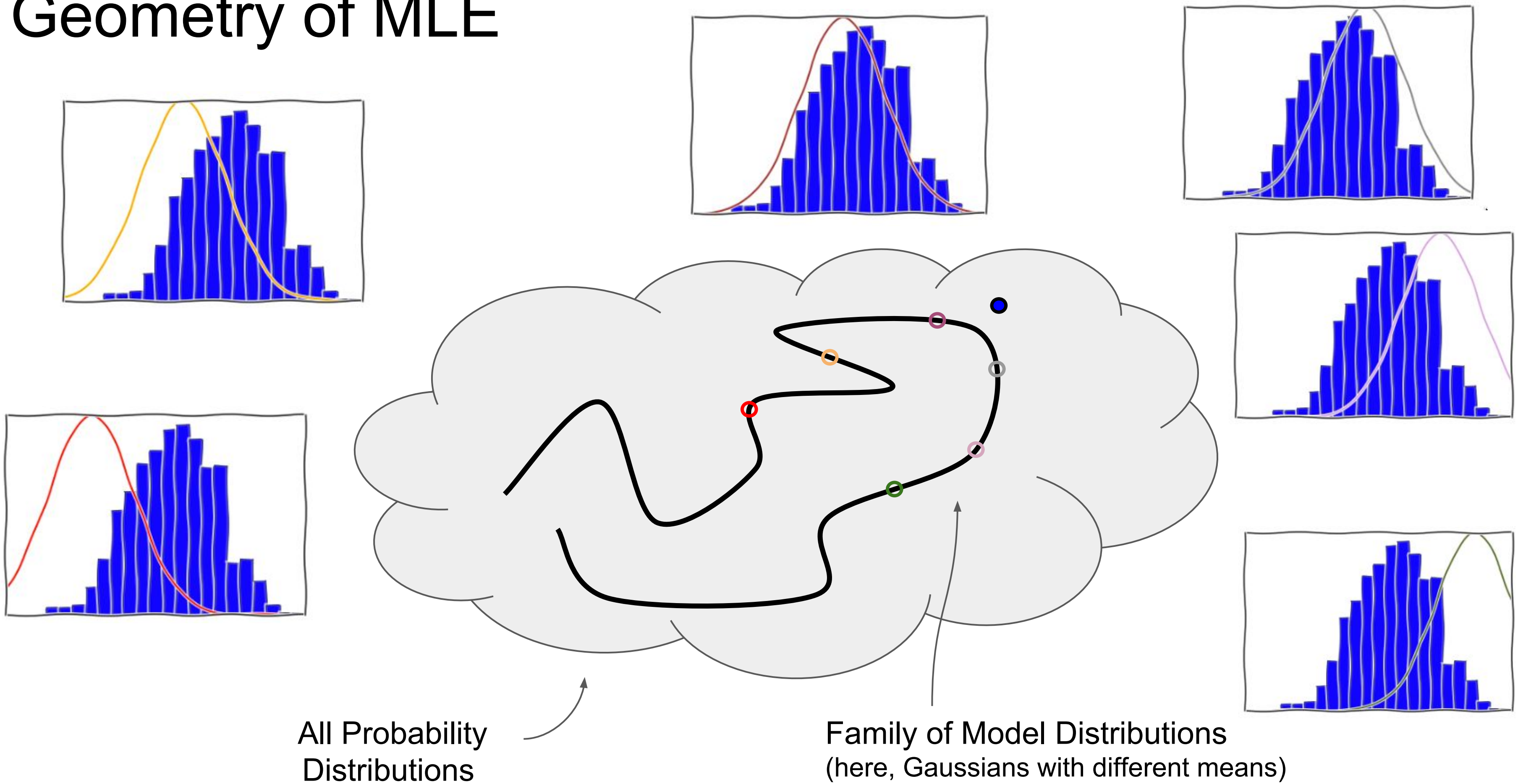
$$\begin{aligned}
\mu_\alpha &= \frac{\langle \mathbf{x}\, P(\alpha|\mathbf{x}) \rangle}{\langle P(\alpha|\mathbf{x}) \rangle} \\
\sigma_\alpha^2 &= \frac{\left\langle \frac{1}{N} |\mathbf{x} - \mu_\alpha|^2\, P(\alpha|\mathbf{x}) \right\rangle}{\langle P(\alpha|\mathbf{x}) \rangle} \\
P(\alpha) &= \langle P(\alpha|\mathbf{x}) \rangle
\end{aligned}$$

Bonus Slides: Information Geometry

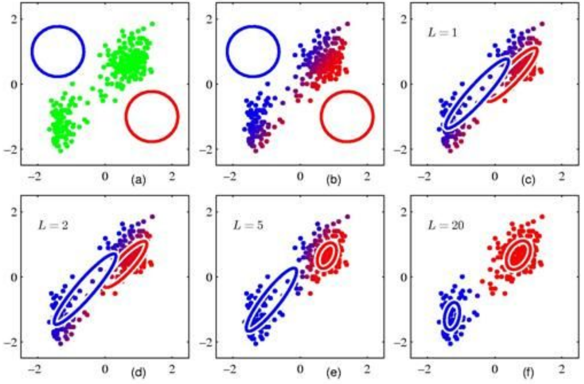
A Geometric View



Geometry of MLE

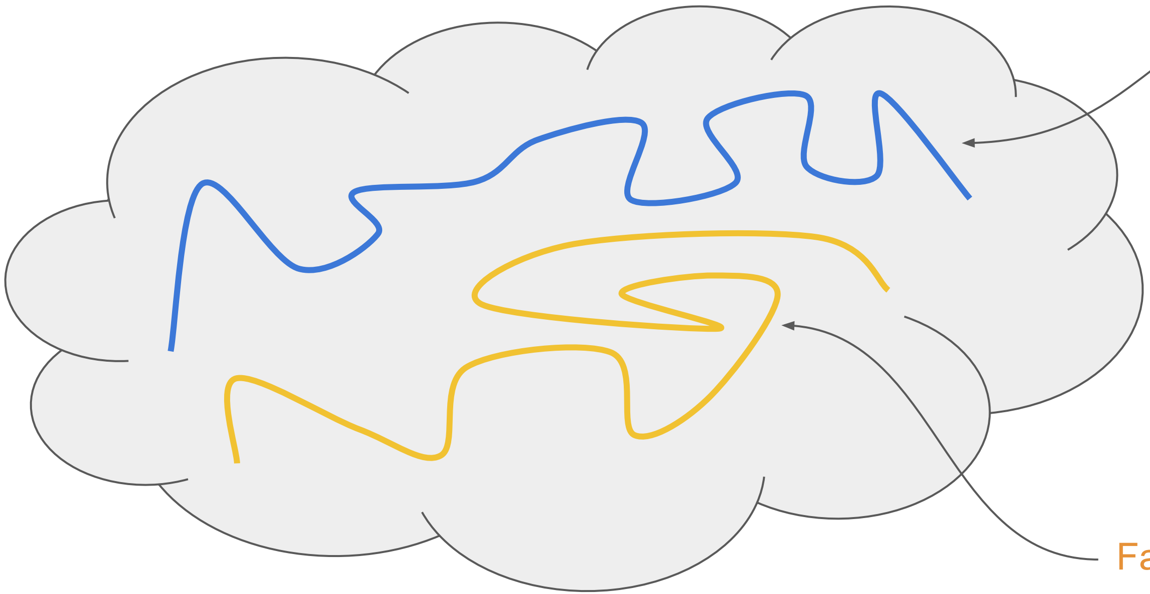


Geometry of EM



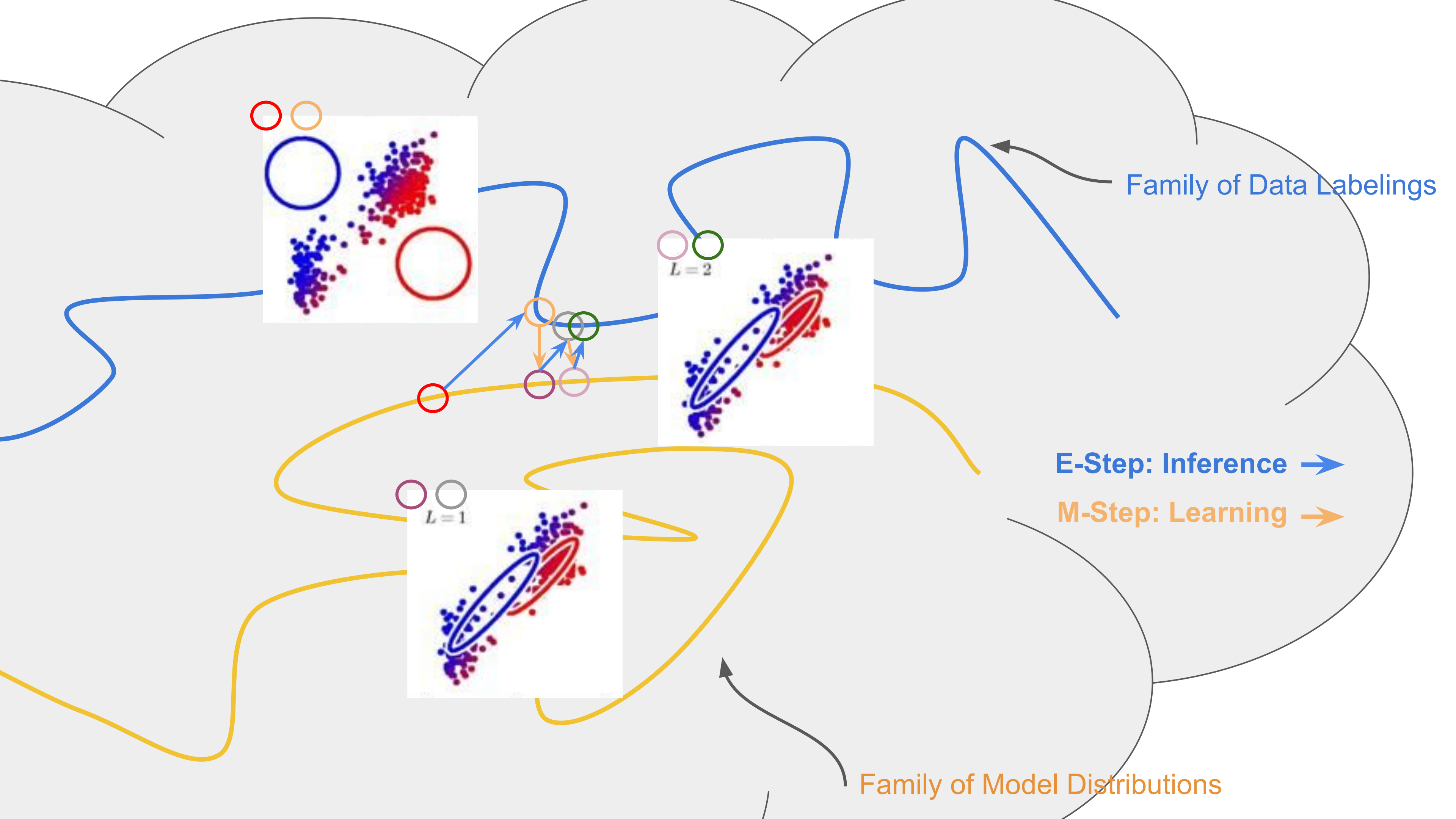
Family of Data Labelings

(same p(x), different p(a|x))



All Probability Distributions

Family of Model Distributions



Family of Data Labelings
$L=2$
$L=1$
E-Step: Inference
M-Step: Learning
Family of Model Distributions