Gentle Introduction to Infinite Gaussian Mixture Modeling

… with an application in neuroscience

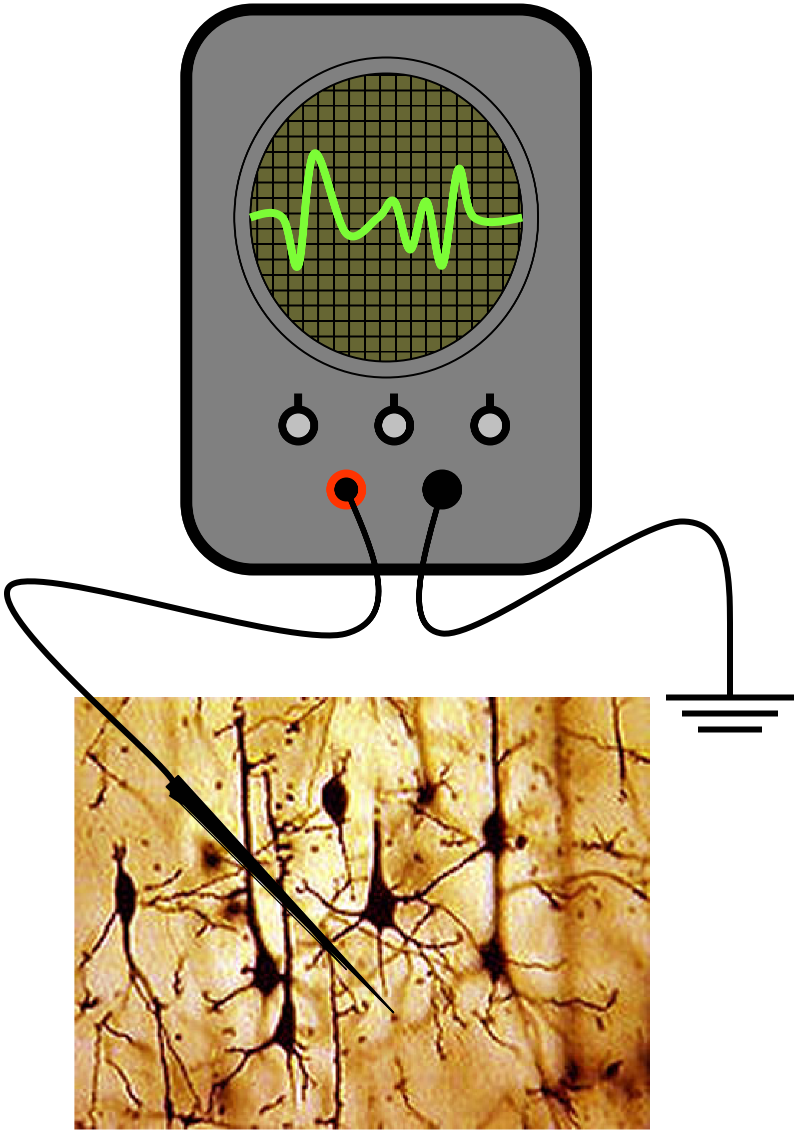
By Frank Wood

BROWN

Rasmussen, NIPS 1999

Neuroscience Application: Spike Sorting

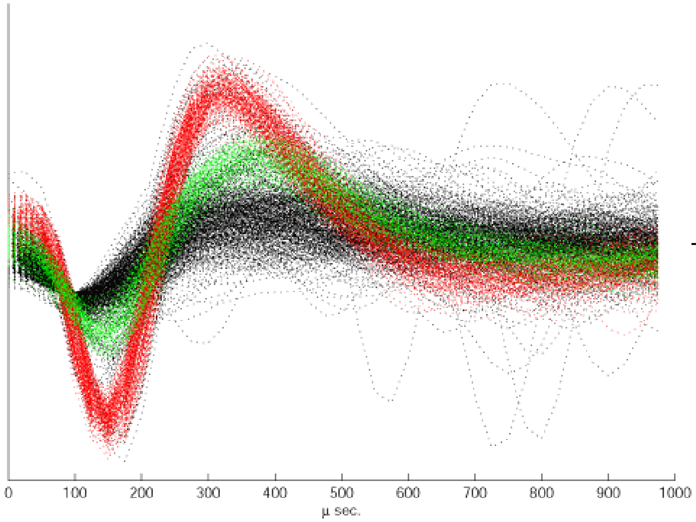
- Important in neuroscience and for medical device performance
- Neural electrical activity is recorded and "spikes" are manually detected and segmented
- "Spike sorting" is the process of deciding which waveforms are spikes and which out of an unknown number of neurons they came from



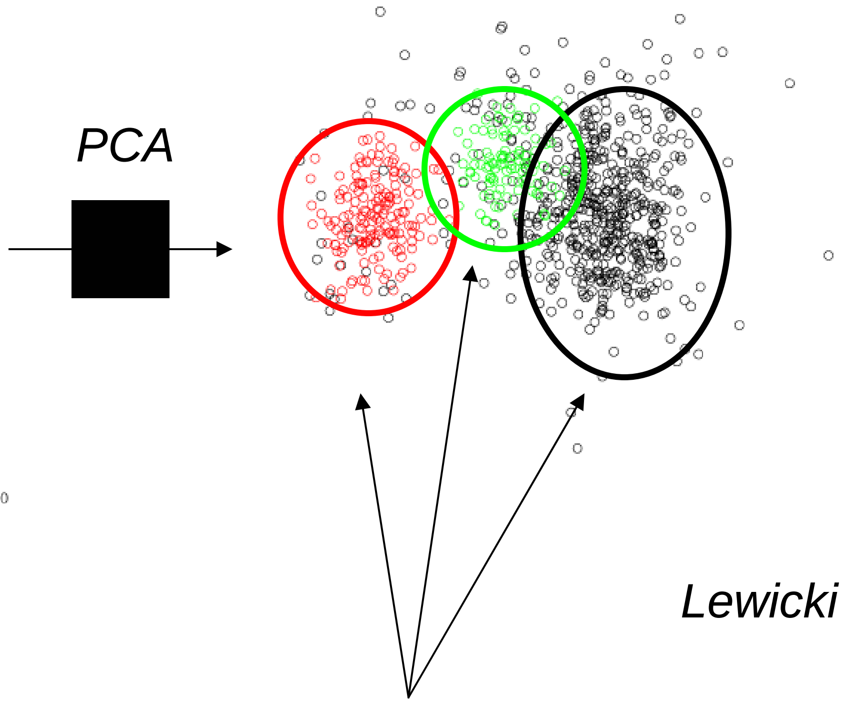
BROWN

Spike Sorting Data

Waveforms recorded on a single electrode and stacked on top of each other



PCA



Accepted neuroscience assumption: ideal mean spike, Gaussian noise

Lewicki et al 99

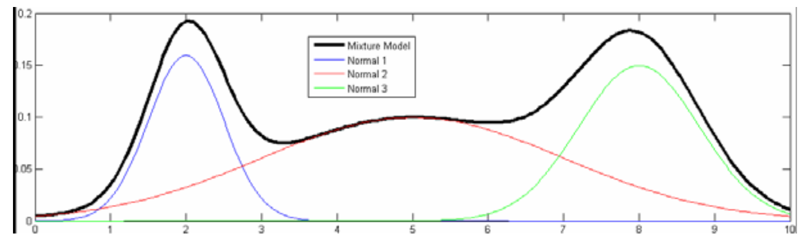
We want the number of hidden units =3

Important Questions

- Did these two spikes come from the same neuron?
  - Did these two data points come from the same hidden class?
- How many neurons are there?
  - How many hidden classes are there?
- Which spikes came from which neurons?
  - What model best explains the data?

Mixture Modeling

A formalism for modeling a probability density function as a sum of parameterized functions.



Normal parameters

Class weights

$$P(y_i|\vec{\pi},\Theta) = \sum_{k=1}^{K} \pi_k P(y_i|\mu_k,\Sigma_k)$$

Number of hidden components

Observations

Class weight, class prior probability, multinomial

Multivariate Normal

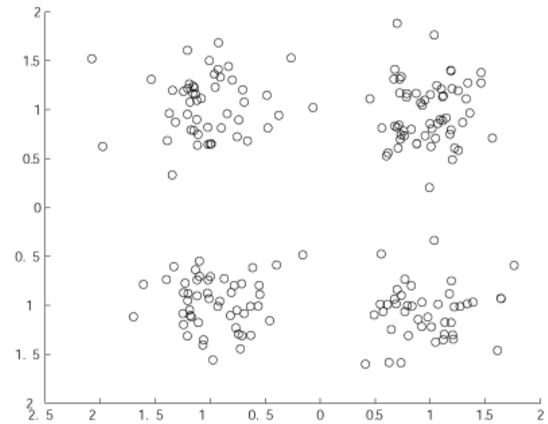
Normal = Gaussian

BROWN

Toy Data and Notation

the data, observed

$$\mathcal{Y} = \{y_i\}_{i=1}^{N}$$

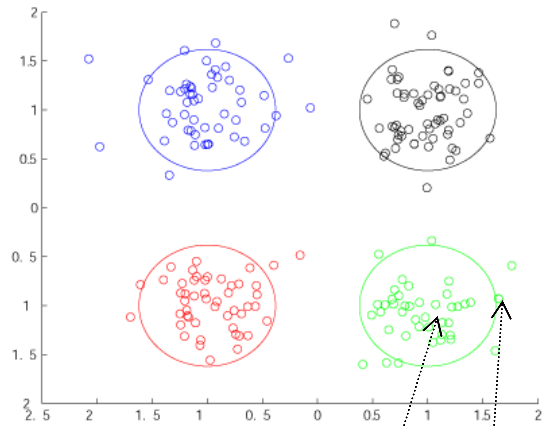


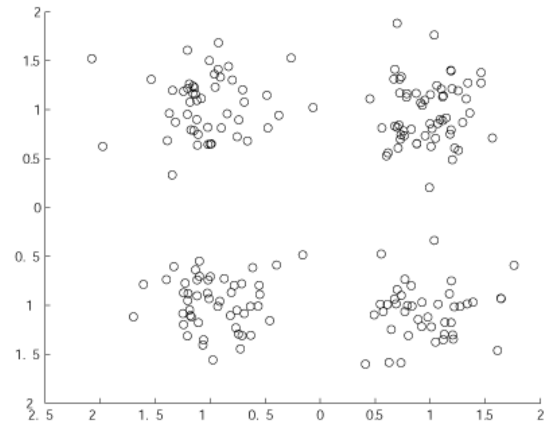
BROWN

Toy Data and Notation

the data, observed

$$\mathcal{Y} = \{y_i\}_{i=1}^{N}$$





$$\theta_k = \{\vec{\mu}_k, \Sigma_k\}, \Theta = \{\theta_k\}_{k=1}^{K}$$

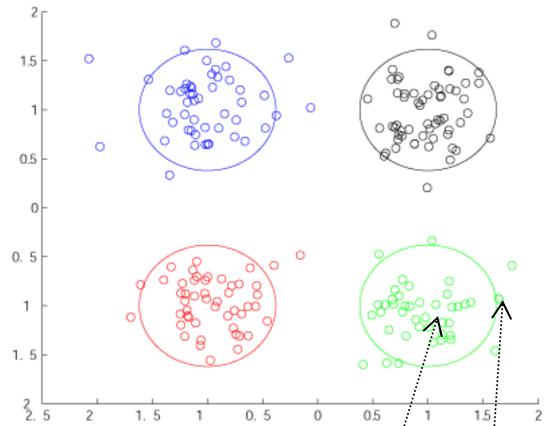
BROWN

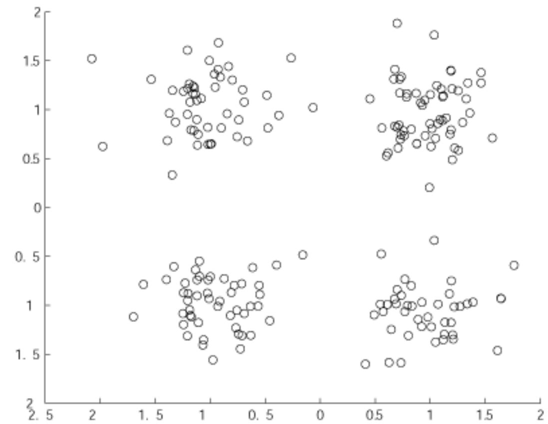
Frank Wood - fwood@cs.brown.edu

Toy Data and Notation

the data, observed

$$\mathcal{Y} = \{y_i\}_{i=1}^{N}$$





$$\theta_k = \{\vec{\mu}_k, \Sigma_k\}, \Theta = \{\theta_k\}_{k=1}^{K}$$

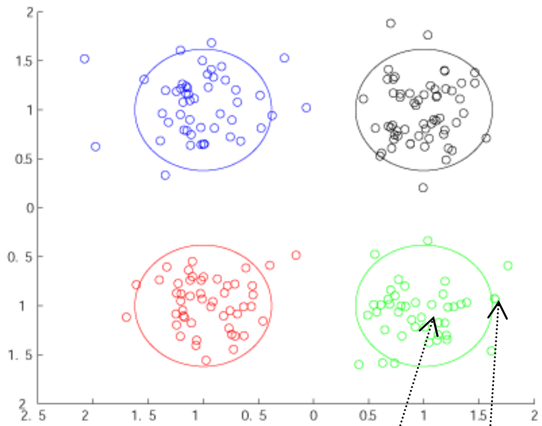
red = 1, green = 2, blue = 3, black =4

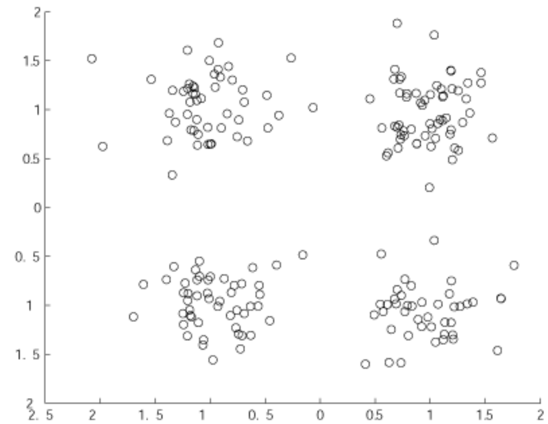
$$\mathcal{C} = \{c_i\}_{i=1}^{N}$$

Toy Data and Notation

the data, observed

$$\mathcal{Y} = \{y_i\}_{i=1}^N$$





$$\theta_k = \{\vec{\mu}_k, \Sigma_k\}, \Theta = \{\theta_k\}_{k=1}^K$$

$$\mathcal{C} = \{c_i\}_{i=1}^N$$

red = 1, *green* = 2, *blue* = 3, *black* =4

$$\pi_1 = .25, \pi_2 = .25,$$
$$\pi_3 = .25, \pi_4 = .25$$

$$\vec{\pi} = \{\pi_k\}_{k=1}^K$$

$$\pi_k = P(c_i = k)$$

Frank Wood - fwood@cs.brown.edu

Goal: learn model parameters from unlabeled data

- Learn the mixture model parameters $\mathcal{C}, \vec{\pi}, \Theta$
  - Maximum likelihood estimation
    - Good if you are certain that your generative model is correct and if all you want is a point estimate of "the right answer"
    - Fast, expectation maximization
  - Bayesian estimation
    - Better if you would like to maintain a representation of your modeling uncertainty
    - Slow, sampling
    - No 'right answer' – learn a distribution instead
    - *Can treat the number of hidden classes as a parameter to be learned*

BROWN

Bayesian Modeling

- Estimate a posterior distribution
- Provides a principled way to encode prior beliefs about the form of the solution
- Posterior distribution represented by samples
- Will enable us to estimate how many hidden classes there are

$$\underbrace{P(\mathcal{M}|\mathcal{Y})}_{\text{Posterior}} \propto \underbrace{P(\mathcal{Y}|\mathcal{M})}_{\text{Likelihood}}\underbrace{P(\mathcal{M})}_{\text{Prior}}$$

$\mathcal{M}$ = *model*

$\mathcal{Y}$ = *observations / training data*

BROWN

What we need:

- Priors for the model parameters
- Sampler
  - To draw samples from the posterior distribution

BROWN

Priors for the model parameters

- Prior over class assignments
  - Class assignments are Multinomial, we will choose a conjugate Dirichlet prior. This allows us to specify a priori how likely we think each class will be.
- Prior over class distribution parameters
  - Class distributions are multivariate Normal. We will choose conjugate Normal*Inverse-Wishart priors. These let us specify a priori where and how broad we think each mixture density should be.

BROWN

Conjugate Priors

- A prior distribution is conjugate if a likelihood distribution times the prior results in a distribution with the same functional form as the prior distribution
- Examples:

Likelihood	*Conjugate Prior*
Poisson	Gamma
Binomial	Beta
Multinomial	Dirichlet
Multivariate Normal	Multivariate Normal * Inverse Wishart

BROWN

Sampling the posterior distribution

- Simulate a Markov chain whose equilibrium distribution is the Bayesian mixture model posterior distribution

Geman & Geman

$$P(\mathcal{C}, \Theta, \vec{\pi}, \alpha | \mathcal{Y})$$

Posterior: *Remember, a distribution over model parameters is what we seek.*

$$\propto P(\mathcal{Y}|\mathcal{C}, \Theta) P(\Theta|\mathcal{G}_0) \prod_{i=1}^{N} P(c_i|\vec{\pi}) P(\vec{\pi}\,|\alpha) P(\alpha).$$

Multivariate Normal

Normal Inverse-Wishart

Multinomial

Dirichlet

***Likelihood:** Multivariate Normal*

***Prior:** CRP over class assignments, normal-IW over normal parameters*

BROWN

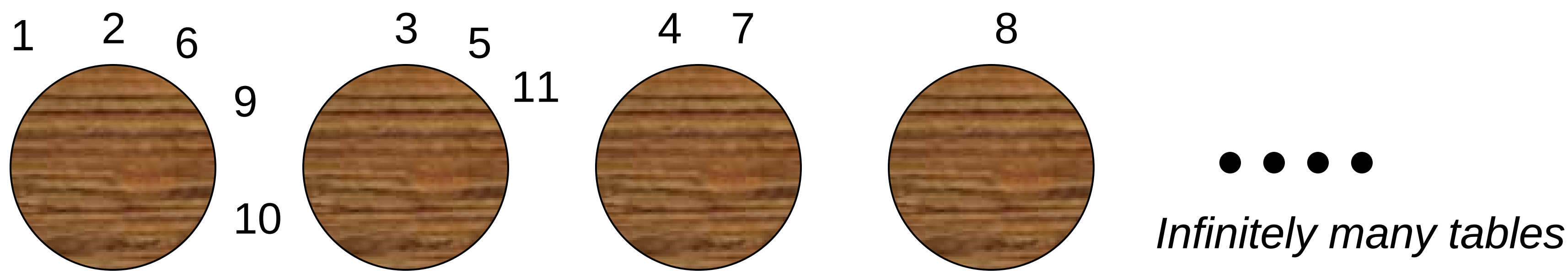
But what about the infinite part?

- Properly parameterized, a posterior formed from a Multinomial Dirichlet conjugate pair is well behaved as the number of hidden classes approaches infinity.
- This results in a model with an infinite number of hidden causes, but one that only a finite number are causal w.r.t. our finite dataset.

- The Chinese Restaurant Process is one process that generates samples from such a model.
  - A hyperparameter (prior) will remain that allows us to specify our a priori belief about how many hidden classes cause our finite data.

BROWN

Sampling class membership in an infinite mixture model: the Chinese Restaurant Process



First customer sits at the first table.

Remaining customers seat themselves randomly.

$$P(c_i = k | \mathcal{C}_{-i}, \alpha) = \begin{cases} \frac{m_k}{i-1+\alpha} & k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & k > K_+ \end{cases}$$

Exchangeable distribution (Aldous, 1985; Pitman, 2002)

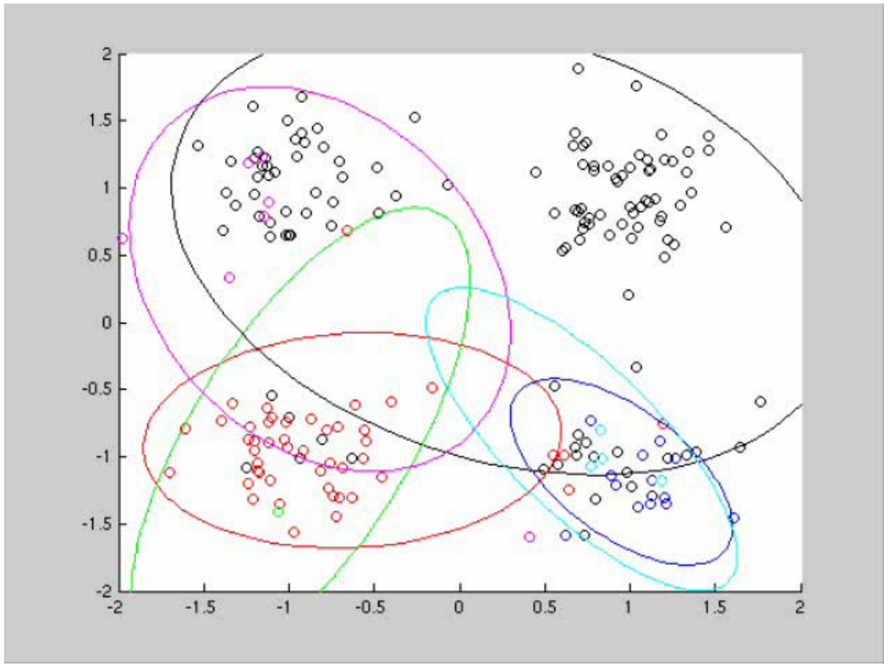
BROWN

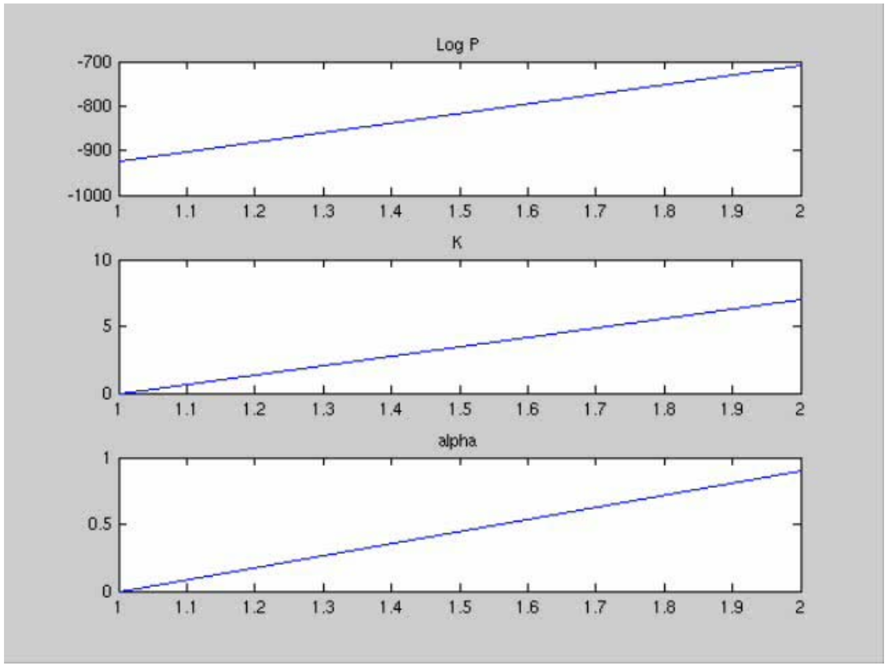
Infinite Gaussian Mixture Model Sampler

- Hard to explain – easy to implement and use
- Gibbs sampler – conjugate priors produce analytic conditional distributions for sampling
- Two step iterative sampler:
  - Sample Normal distribution means and covariances given a current assignment of data to classes
  - Sample the assignment of data to classes given current values for the means and covariances (CRP)
- After some time, sampler converges to a set of samples from the posterior, i.e. a scored set of feasible models given the training data

BROWN

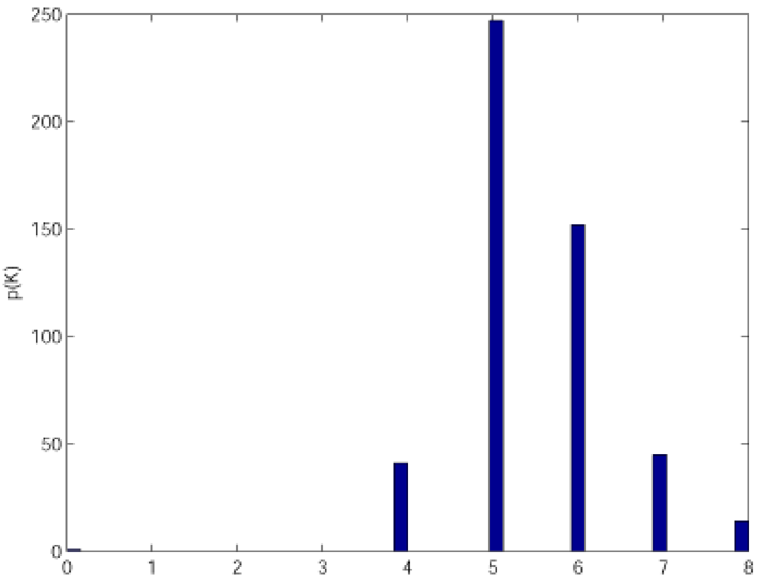
Gibbs Sampling the Posterior



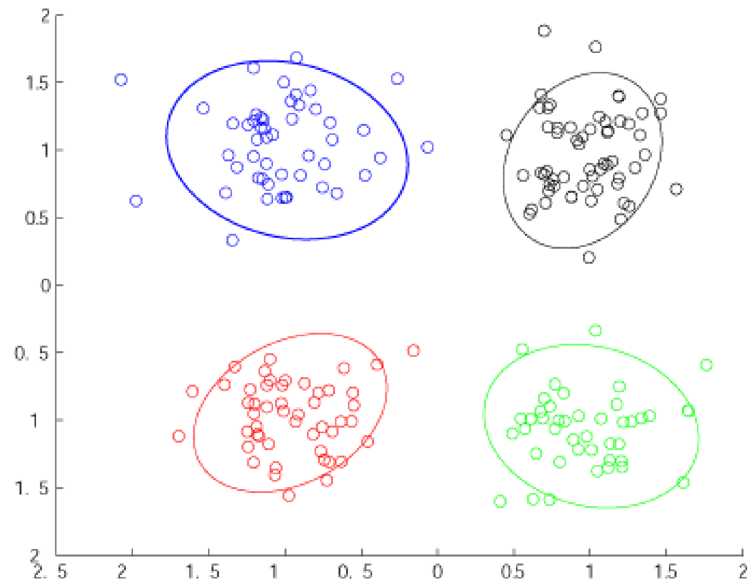


Toy Data Results

Distribution over # of classes K

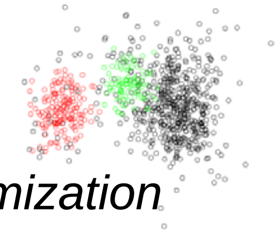


Maximum a posteriori sample

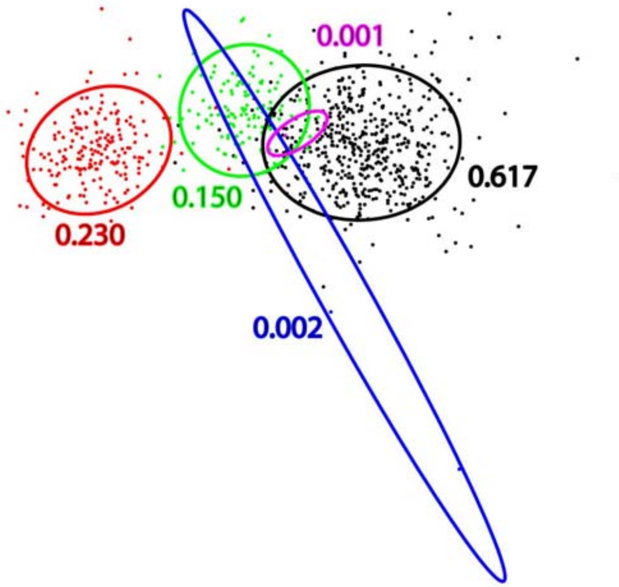


Frank Wood - fwood@cs.brown.edu

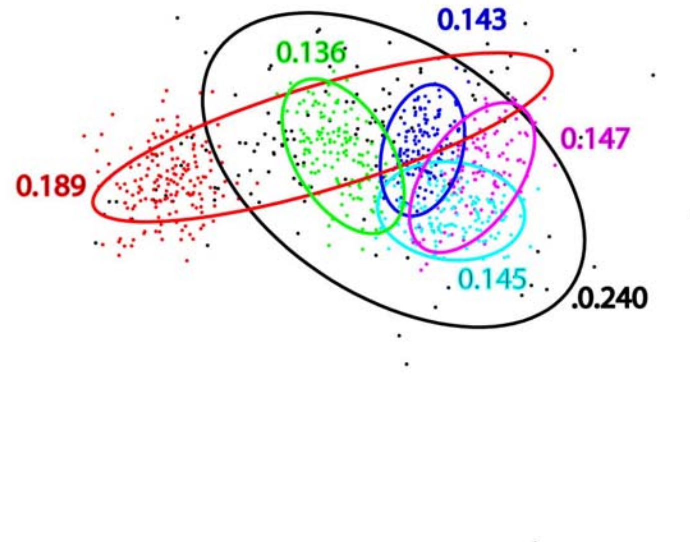
Single channel spike sorting results



Infinite Mixture Model



Expectation Maximization



- *Priors enforce preference for intuitive models*
- *CRP prior allows inference over # of hidden classes*

- *Lack of priors allows non-intuitive solutions*
- *No distribution over # of hidden classes*

BROWN

Conclusions

- Bayesian mixture modeling is principled way to add prior information into the modeling process
- IMM / CRP is a way estimate the number of hidden classes
- Infinite Gaussian mixture modeling is good for automatic spike sorting

Future Work

- Particle filtering for online spike sorting

BROWN

Thank you

This work was supported by NIH-NINDS R01 NS 50967-01 as part of the NSF/NIH Collaborative Research in Computational Neuroscience Program.

IGMM Software available at http://www.cs.brown.edu/~fwood/code.html

Thanks to Michael Black, Tom Griffiths, Sharon Goldwater, and the Brown University machine learning reading group.

BROWN

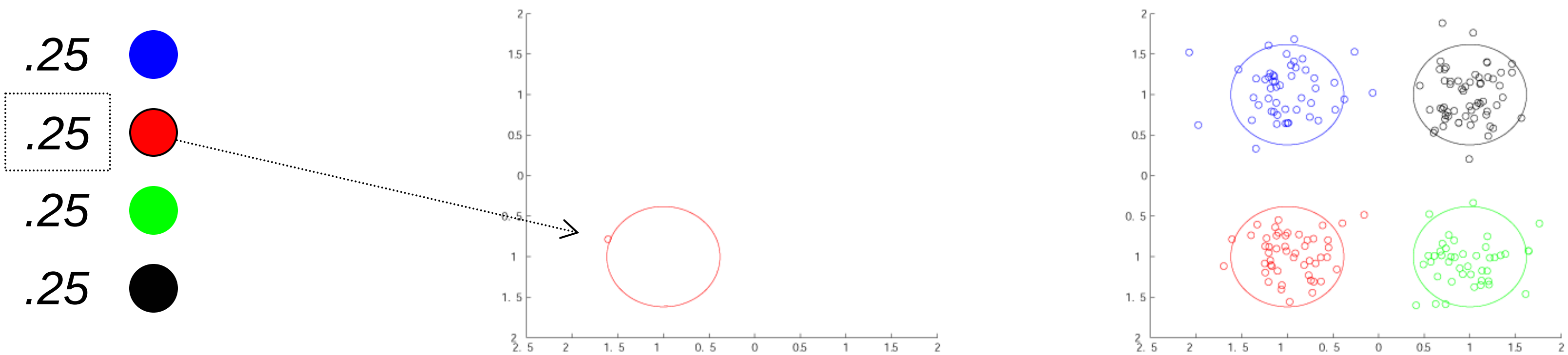
Generative Viewpoint

$$c_i|\vec{\pi} \sim \text{Multinomial}(\cdot|\vec{\pi})$$

$$\vec{y}_i|c_i = k, \Theta \sim \mathcal{N}(\cdot|\theta_k)$$

Pick class label according to multinomial

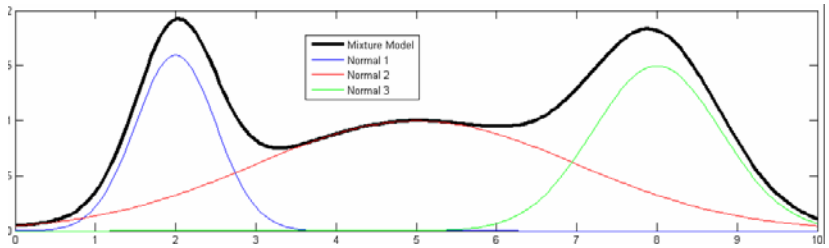
Generate observation according to class model



Frank Wood - fwood@cs.brown.edu

Mixture Modeling

- A formalism for modeling a probability density function as a sum of parameterized functions



- Observed population data is complicated – not well fit by a canonical parametric distribution
- Assume: 'Hidden' subpopulation data is simple – well fit by a canonical parametric distribution
- Hope: 1 hidden subpopulation <-> 1 simple parametric distribution
- Key questions:
  - How many hidden subpopulations are responsible for generating the data?
  - Which subpopulation did each data point come from?

BROWN

Limiting Behavior of Uniform Dirichlet Prior

$$
\begin{aligned}
P(\mathcal{C}|\alpha) &= \int \prod_{i=1}^{N} P(c_i|\vec{\pi})P(\vec{\pi}\,|\alpha)d\vec{\pi} \\
&= \frac{\prod_{k=1}^{K} \Gamma(m_k + \frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K} \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}
\end{aligned}
$$

$$
\lim_{K \to \infty} P(\mathcal{C}|\alpha) = \alpha^{K_+} \left(\prod_{k=1}^{K_+} (m_k - 1)! \right) \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}
$$

BROWN

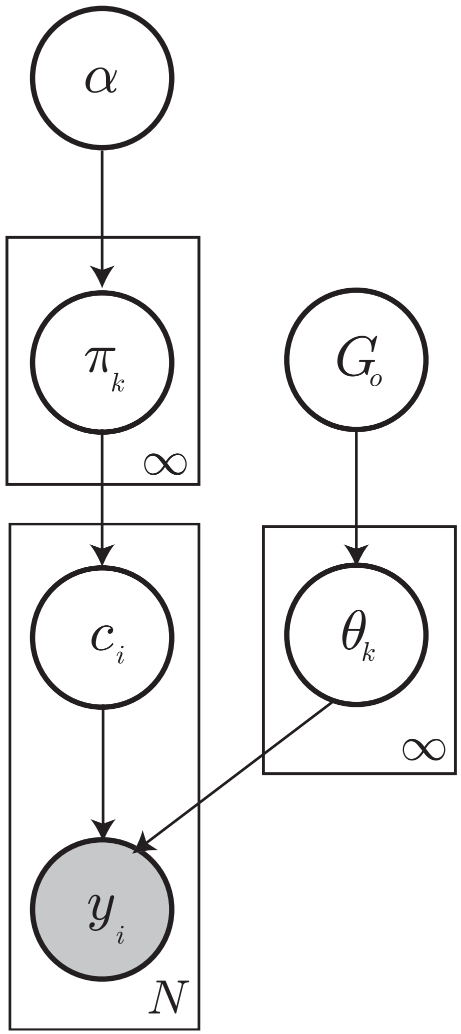
Bayesian Mixture Model Priors

- Prior over class assignments

$$
\begin{aligned}
\vec{\pi}|\alpha &\sim \text{Dirichlet}(\cdot|\frac{\alpha}{K},\ldots,\frac{\alpha}{K}) \\
\Theta &\sim \mathcal{G}_0
\end{aligned}
$$

- Prior over class distribution parameters

$$
\begin{aligned}
\Sigma_k &\sim \text{Inverse-Wishart}_{v_0}(\Lambda_0^{-1}) \\
\vec{\mu}_k &\sim \mathcal{N}(\vec{\mu}_0, \Sigma_k/\kappa_0).
\end{aligned}
$$



BROWN

Frank Wood - fwood@cs.brown.edu

Conjugacy – our friend

- If you choose a conjugate prior then the posterior will be in the same family as the prior.
  - Normal <-> Normal * Inverse-Wishart
  - Dirichlet <-> Multinomial

$$
\begin{aligned}
P(\mathcal{C}|\alpha) &= \int \prod_{i=1}^{N} P(c_i|\vec{\pi})P(\vec{\pi}\,|\alpha)d\vec{\pi} \\
&= \frac{\prod_{k=1}^{K}\Gamma(m_k+\frac{\alpha}{K})}{\Gamma(\frac{\alpha}{K})^K}\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}
\end{aligned}
$$

  - Analytic posteriors allow Gibbs sampling

BROWN

Sampler

State of the sampler $\{\mathcal{C}, \Theta\}$

$$P(\theta_k|\mathcal{C}, \mathcal{Y}, \Theta_{-k}, \vec{\pi}, \alpha) \propto \prod_{i \text{ s.t. } c_i = k} P(\vec{y}_i|c_i, \theta_k) P_{\mathcal{G}_0}(\theta_k).$$

$$P(c_i = k|\mathcal{C}_{-i}, \mathcal{Y}, \Theta, \vec{\pi}, \alpha) \propto P(\vec{y}_i|c_i, \Theta) P(c_i|\mathcal{C}_{-i})$$

$$P(c_i = k|\mathcal{C}_{-i}) = \begin{cases} \frac{m_k}{i-1+\alpha} & k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & k > K_+ \end{cases}$$

$$\Theta_{-k} = \{\theta_k, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_N\}$$

BROWN

Maximum likelihood techniques

- Expectation maximization

$$P(\mathcal{Y}, \mathcal{C} | \vec{\pi}, \Theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k P(\vec{y}_i | c_i = k, \Theta).$$

$$\hat{\vec{\pi}}, \hat{\Theta} = \arg\max_{\vec{\pi}, \Theta} \log(P(\mathcal{Y}, \mathcal{C} | \vec{\pi}, \Theta))$$

- Bayesian Information Criteria

$$\text{BIC} = -2\log(P(\mathcal{Y}, \mathcal{C} | \vec{\pi}, \Theta)) + \nu_K \log(N)$$

-- but not Bayesian; no distribution over

Example applications

- Modeling network packet traffic
  - Network applications' performance dependent on distribution of incoming packets
  - Want a population model to build a fancy scheduler
  - Potentially multiple heterogeneous applications generating packet traffic
  - How many types of applications are generating packets?
- Clustering sensor data (robotics, sensor networks)
  - Robot encounters multiple types of physical environments (doors, walls, hallways, etc.)
  - How many types of environments are there?
  - How do we tell what type of space we are in?

BROWN