Bayesian Methods, Population Mixtures

Adriano Agnello

5th May 2020

Overview

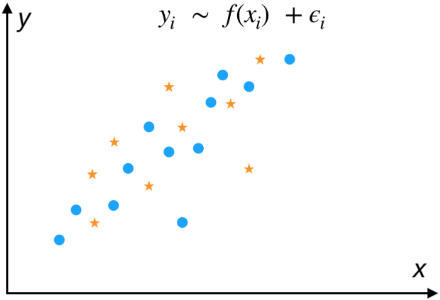
This is a chunky lecture, with some extra stuff if you want to go see it afterwards. We'll see:

- What's all this *Bayesian* stuff that people talk about?!
- Why best-fit is not always the best, and what you can do about it.
- How we can fit a distribution without binning the data (a.k.a. *binning is sinning*).
- Another *unsupervised* learning method for clustering, with some `sklearn` help of course.

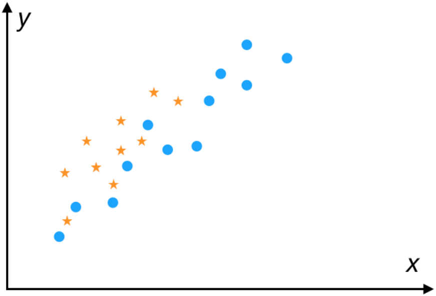
Case 0: fit a line through some points. Minimize the chi-squared

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(\mathbf{x}_i; \boldsymbol{\theta}))^2}{\epsilon_i^2} \quad (1)$$

with respect to some parameters ($\boldsymbol{\theta}$).

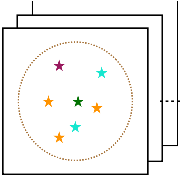


But what if the points have some intrinsic scatter around $f(\mathbf{x})$?
What if they come from different populations?
What about measurement errors on the $\mathbf{x}_i$?



Answer: go *Bayesian*.

Case 1: cases where you have *latent* variables, non-fixed dimensionality, and need to guess a complicated regression.

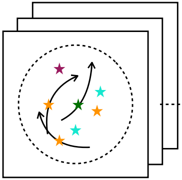


Let's consider a bag with different candies.
Each candy has an expiry date that we don't know.
The total worth of the bag is
not necessarily the sum of candy prices.

- What is the total worth of the bag?

$$\mathrm{p}(M\,|\,\mathbf{p}) \;=\; ???$$

- What about many objects of this kind?
  And ***not*** with the same amount of candies each?



- What is the ***total mass*** of this object?

$$\mathrm{p}(M\,|\,\mathbf{v}\ldots,\mathbf{x}\ldots) \;=\; ???$$

- Did different subsets of stars come in together?

- What about many objects of this kind?
  And ***not*** with the same amount of stars each?

Answer: go *Bayesian*.

What's with that P(A|B)

Probability: a measure[1] on a space of events $\mathcal{E}$, with $P(\mathcal{E}) = 1$.
Conditional probability:

$$P(A|B) \;:=\; P(A \cap B)/P(B) \tag{2}$$

Then

$$P(B)P(A|B) \;=\; P(A \cap B) = P(A)P(B|A) \tag{3}$$

This is Bayes' Thorem.

E.g.: think of a thermostat that goes on *mostly* when $T > 28°$C and is off *mostly* when $T < 28°$C... If it turns on (and I don't have a thermometer), what are the odds that it's actually hot and not a glitch?

[1] Sigma-additive, positive-definite, null on empty subset...

What's with that P(A|B)

Probability: a measure[1] on a space of events $\mathcal{E}$, with $P(\mathcal{E}) = 1$.
Conditional probability:

$$P(A|B) \ := \ P(A \cap B)/P(B) \tag{2}$$

Then

$$P(B)P(A|B) \ = \ P(A \cap B) = P(A)P(B|A) \tag{3}$$

This is Bayes' Thorem.

E.g.: think of a thermostat that goes on *mostly* when $T > 28^{\circ}$C and is off *mostly* when $T < 28^{\circ}$C... If it turns on (and I don't have a thermometer), what are the odds that it's actually hot and not a glitch?

[1] Sigma-additive, positive-definite, null on empty subset...

Bayes, models, and data

In terms of data and models: how much would you bet on a model, given observed data?

$$\mathrm{d}\mu_{\mathrm{post.}}(\mathbf{m}|\mathbf{d}) \propto \mathcal{L}(\mathbf{d}|\mathbf{m})\mathrm{d}\mu_{\mathrm{pr.}}(\mathbf{m}) \tag{4}$$

I.e. given some model parameter values, how likely am I to obtain the observed dataset ($\mathcal{L}(\mathbf{d}|\mathbf{m})$)?

NB The model parameters need not be fixed, they may have some *prior* probability measure. This, together with the likelihood ($\mathcal{L}(\mathbf{d}|\mathbf{m})$), gives the *posterior* measure.

NB2 A posterior from a previous measurement can be a prior for a successive (independent) measurement.

Bayes, models, and data

In terms of data and models: how much would you bet on a model, given observed data?

$$\mathrm{d}\mu_{\mathrm{post.}}(\mathbf{m}|\mathbf{d}) \propto \mathcal{L}(\mathbf{d}|\mathbf{m})\mathrm{d}\mu_{\mathrm{pr.}}(\mathbf{m}) \tag{4}$$

I.e. given some model parameter values, how likely am I to obtain the observed dataset ($\mathcal{L}(\mathbf{d}|\mathbf{m})$)?

NB The model parameters need not be fixed, they may have some *prior* probability measure. This, together with the likelihood ($\mathcal{L}(\mathbf{d}|\mathbf{m})$), gives the *posterior* measure.

NB2 A posterior from a previous measurement can be a prior for a successive (independent) measurement.

Example 1: Given N independent points $x_{i=1,\ldots,N}$ on a line, and suppose that each of them is drawn from a Gaussian $\mathcal{G}(x_i - \mu, \sigma)$, where μ and σ are the unknowns that we want to work out. Our likelihood is "simply"

$$\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}) = \prod_i \mathcal{G}(x_i - \mu, \sigma) \propto \exp\left(-\frac{1}{2}\frac{\sum_i (x_i - \mu)^2}{\sigma^2}\right) \quad (5)$$

Q: what are the most likely mean μ and dispersion σ?

$\partial_\mu \mathcal{L} = 0 \Rightarrow \mu = \frac{1}{N}\sum_i x_i$

$\partial_\sigma \mathcal{L} = 0 \Rightarrow \sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$

Example 1: Given N independent points $x_{i=1,\ldots,N}$ on a line, and suppose that each of them is drawn from a Gaussian $\mathcal{G}(x_i - \mu, \sigma)$, where μ and σ are the unknowns that we want to work out. Our likelihood is "simply"

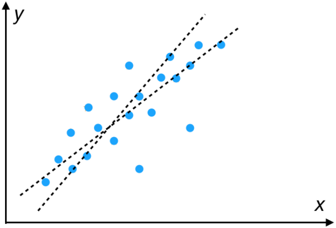
$$\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}) = \prod_i \mathcal{G}(x_i - \mu, \sigma) \propto \exp\left(-\frac{1}{2}\frac{\sum_i (x_i - \mu)^2}{\sigma^2}\right) \tag{5}$$

Q: what are the most likely mean μ and dispersion σ?
$\partial_\mu \mathcal{L} = 0 \Rightarrow \mu = \frac{1}{N}\sum_i x_i$
$\partial_\sigma \mathcal{L} = 0 \Rightarrow \sigma^2 = \frac{1}{N}\sum_i (x_i - \mu)^2$

A note about Priors

How much of the result is *driven by the prior*?

Example: fit a straight line through points. Should I write it in terms of slope or in terms of pitch angle?



$$y_i \sim q + m \cdot x_i + \epsilon_i$$

$$y_i \sim q + \tan(\theta) \cdot x_i + \epsilon_i$$

Uninformative prior on slope:

$$\mathrm{p}(\textit{slope})\mathrm{d}(\textit{slope}) = \tilde{\mathrm{p}}_{\textit{unif.}}(\textit{angle})\mathrm{d}(\textit{angle}) \propto \frac{1}{1+\textit{slope}^2}\ . \tag{6}$$

Jeffrey's prior[2]:

$$\mathrm{p}_j \propto \sqrt{\mathbb{E}[|\det(\partial^2 \log \mathcal{L})/\partial m_i \partial m_j|]} \tag{7}$$

is typically chosen as an *uninformative* prior.
NB: Jeffery's prior is independent of the parameterization.

[2]Here the expectation value $\mathbb{E}[\bullet]$ is computed w.r.to $\mathcal{L}\mathrm{d}\mathbf{m}$.

Nested models, evidence tests

Model A is more general than model B, is it to be preferred?
Different *evidence tests*:

1. (from physics lab. 101) change in χ^2 vs change in d.o.f.
2. if $\mathcal{L}$ is Gaussian in the data, same phrasing as the χ^2 test
3. AIC: $2k - 2\log(\mathcal{L}_{best})$
4. BIC: $\log(N)k - 2\log(\mathcal{L}_{best})$
5. AIC for small sample sizes: $AIC + 2\frac{k^2+k}{n-k-1}$
6. significance test: how many 'sigmas' is the new parameter different from zero?
7. Jeffrey's test: compute the evidence $Z = \int \mathcal{L}\, \mathrm{d}\mu_{pr.}(\mathbf{m})$, compare $\log_{10}(Z)$.

Important to keep in mind

Important bit n.1: It's not just the goodness of fit, we want to know how good a model is and how much it can vary *given the current data*.

Important bit n.2: There are different ways of comparing models, and we typically need to integrate $\mathcal{L}\mathrm{d}\mu_{pr}$.

Sampling the likelihood and posterior

How do we explore the likelihood?

1. 'grid search': good when parameter space has few dimensions, par.range is already known, likelihood is 'easy'.
2. Monte Carlo, draw random points from prior and save their likelihood, or from uniform prior and save the posterior. Good for obtaining integrated quantities, fast convergence for high-dimensional par.space.
3. Markov-Chain Monte Carlo: build a Markov Chain s.t. the final density of points is proportional to the posterior.
4. Gibbs sampling: to sample $p(\mathbf{a}, \mathbf{b})$, alternate drawing from $p(\mathbf{a}|\mathbf{b})$ and $p(\mathbf{b}|\mathbf{a})$.
5. Optimization (e.g. amoeba, gradient descent), approx. optimization (e.g. EM in GMM).

Sampling the likelihood and posterior

How do we explore the likelihood?

1. ‘grid search’: good when parameter space has few dimensions, par.range is already known, likelihood is ‘easy’.
2. Monte Carlo, draw random points from prior and save their likelihood, or from uniform prior and save the posterior. Good for obtaining integrated quantities, fast convergence for high-dimensional par.space.
3. Markov-Chain Monte Carlo: build a Markov Chain s.t. the final density of points is proportional to the posterior.
4. Gibbs sampling: to sample $p(\mathbf{a}, \mathbf{b})$, alternate drawing from $p(\mathbf{a}|\mathbf{b})$ and $p(\mathbf{b}|\mathbf{a})$.
5. Optimization (e.g. amoeba, gradient descent), approx. optimization (e.g. EM in GMM).

Sampling the likelihood and posterior

How do we explore the likelihood?

1. 'grid search': good when parameter space has few dimensions, par.range is already known, likelihood is 'easy'.
2. Monte Carlo, draw random points from prior and save their likelihood, or from uniform prior and save the posterior. Good for obtaining integrated quantities, fast convergence for high-dimensional par.space.
3. Markov-Chain Monte Carlo: build a Markov Chain s.t. the final density of points is proportional to the posterior.
4. Gibbs sampling: to sample $p(\mathbf{a}, \mathbf{b})$, alternate drawing from $p(\mathbf{a}|\mathbf{b})$ and $p(\mathbf{b}|\mathbf{a})$.
5. Optimization (e.g. amoeba, gradient descent), approx. optimization (e.g. EM in GMM).

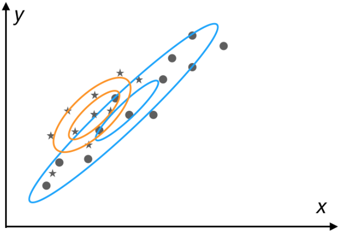
Sampling the likelihood and posterior

How do we explore the likelihood?

1. 'grid search': good when parameter space has few dimensions, par.range is already known, likelihood is 'easy'.
2. Monte Carlo, draw random points from prior and save their likelihood, or from uniform prior and save the posterior. Good for obtaining integrated quantities, fast convergence for high-dimensional par.space.
3. Markov-Chain Monte Carlo: build a Markov Chain s.t. the final density of points is proportional to the posterior.
4. Gibbs sampling: to sample $p(\mathbf{a}, \mathbf{b})$, alternate drawing from $p(\mathbf{a}|\mathbf{b})$ and $p(\mathbf{b}|\mathbf{a})$.
5. Optimization (e.g. amoeba, gradient descent), approx. optimization (e.g. EM in GMM).

Modeling Densities

Let's consider a swarm of points $\mathbf{x}_{i=1,\ldots,N} \in \mathbb{R}^p$...



$$\rho(x_i, y_i) \sim \sum_k \rho_k(x_i, y_i)$$

...and a model $\rho(\mathbf{x})$ for how they should be distributed.

Modeling Densities

Let's consider a swarm of points $\mathbf{x}_{i=1,\ldots,N} \in \mathbb{R}^p$,, and a model $\rho(\mathbf{x})$ for how they should be distributed. Let's split $\mathbb{R}^p$ in cells of volumes δv_j, each containing n_j points. In each cell, the model predicts $mod_j \approx \rho(\mathbf{x}_j)\delta v_j$ points.
Likelihood:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{(\rho(\mathbf{x}_j)\delta v_j)^{n_j} \mathrm{e}^{-\rho(\mathbf{x}_j)\delta v_j}}{n_j!} \quad (8)$$

$$= \left(\prod_j \frac{(\delta v_j)^{n_j}}{n_j!} \right) \left(\prod_j \rho(\mathbf{x}_j)^{n_j} \right) \exp\left(-\int \rho \mathbf{d}v \right) \quad (9)$$

$\Rightarrow$ Often, what really matters is $\rho(\mathbf{x})$ and its parameters.
The bins (δv_j) factor out. ***Binning is sinning***.

Modeling Densities

Let's consider a swarm of points $\mathbf{x}_{i=1,\dots,N} \in \mathbb{R}^p$,, and a model $\rho(\mathbf{x})$ for how they should be distributed. Let's split $\mathbb{R}^p$ in cells of volumes δv_j, each containing n_j points. In each cell, the model predicts $mod_j \approx \rho(\mathbf{x}_j)\delta v_j$ points. Likelihood:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{(\rho(\mathbf{x}_j)\delta v_j)^{n_j} \mathrm{e}^{-\rho(\mathbf{x}_j)\delta v_j}}{n_j!} \tag{8}$$

$$= \left(\prod_j \frac{(\delta v_j)^{n_j}}{n_j!} \right) \left(\prod_j \rho(\mathbf{x}_j)^{n_j} \right) \exp\left(- \int \rho \mathbf{d} v \right) \tag{9}$$

$\Rightarrow$ Often, what really matters is $\rho(\mathbf{x})$ and its parameters. The bins (δv_j) factor out. ***Binning is sinning***.

Modeling Densities

Let's consider a swarm of points $\mathbf{x}_{i=1,\ldots,N} \in \mathbb{R}^p$,, and a model $\rho(\mathbf{x})$ for how they should be distributed. Let's split $\mathbb{R}^p$ in cells of volumes δv_j, each containing n_j points. In each cell, the model predicts $mod_j \approx \rho(\mathbf{x}_j)\delta v_j$ points. Likelihood:

$$\mathcal{L} = \prod_{i=1}^{N} \frac{(\rho(\mathbf{x}_j)\delta v_j)^{n_j} \mathrm{e}^{-\rho(\mathbf{x}_j)\delta v_j}}{n_j!} \quad (8)$$

$$= \left(\prod_j \frac{(\delta v_j)^{n_j}}{n_j!} \right) \left(\prod_j \rho(\mathbf{x}_j)^{n_j} \right) \exp\left(-\int \rho \mathbf{d} v \right) \quad (9)$$

$\Rightarrow$ Often, what really matters is $\rho(\mathbf{x})$ and its parameters.
The bins (δv_j) factor out. ***Binning is sinning***.

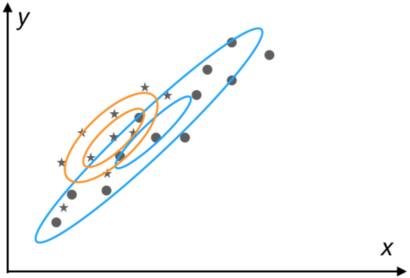
Ex.1: If $\rho = n_0\tilde{\rho}$, with $\int \tilde{\rho}\mathbf{d}v = 1$, what is the best-fitting n_0?

Ex.2: Given N points on a line, if we want to draw their histogram, what is the best bin width that we should use?

Ex.3: Can you design a way of using histograms with uneven bin-widths?

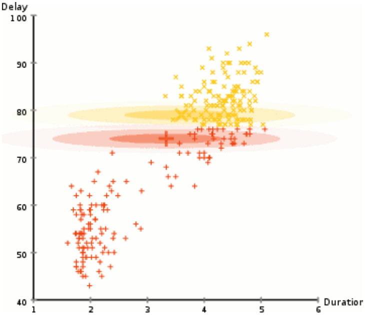
Ex.4: Given N points $x_{i=1,\ldots,N}$ on a line, ***if*** the model likelihood is $\mathcal{L}(\mathbf{x}|\boldsymbol{\theta}) = \mathcal{G}(x-\mu,\sigma)$, what are the most likely mean μ and dispersion σ?

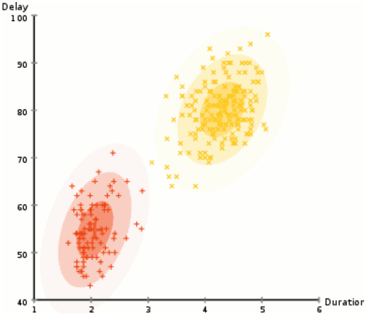
Population Mixtures



$$\rho(x_i, y_i) \sim \sum_k \rho_k(x_i, y_i)$$

Let's consider a *mixture model* $\rho(\mathbf{x}) = \sum_k w_k \tilde{\rho}_k(\mathbf{x})$, with $\int \tilde{\rho}_k \mathbf{d}v = 1$, i.e. we model the swarm of points as drawn from a mixture of blobs ($\tilde{\rho}_k$), where each blob has a weight (w_k).


Delay
100
90
80
70
60
50
40
1
2
3
4
5
6
Duration


Delay
100
90
80
70
60
50
40
1
2
3
4
5
6
Duration

Population Mixtures

Let's consider a *mixture model* $\rho(\mathbf{x}) = \sum_k w_k \tilde{\rho}_k(\mathbf{x})$, with $\int \tilde{\rho}_k \mathbf{d}v = 1$. Then:

$$\log(\mathcal{L}) \;=\; cnst. + \sum_j \log\left(\sum_k w_k \tilde{\rho}(\mathbf{x}_j)\right) - \sum_k w_k \qquad (10)$$

The most likely normalizations w_l satisfy ($\partial_{w_l} \log \mathcal{L} = 0$)

$$w_l = \sum_j \frac{w_l \tilde{\rho}_{l,j}}{\sum_k w_k \tilde{\rho}_{k,j}} \qquad (11)$$

This can be solved iteratively.
Q: What about the *internal* parameters of the components $\tilde{\rho}_k$?

NB: We are using blobs, but the points are un-labeled! Labels are *latent variables*.

A *finite mixture model* has:

- data points $\mathbf{X} = \mathbf{x}_{i=1,\ldots,N} \in \mathbb{R}^P$
- (hidden) labels $\mathbf{Z} = z_{i=1,\ldots,N}$
- mixture weights $w_{k=1,\ldots,K}$ and a prior on them
- basis functions $\tilde{\rho}_k$, with some internal parameters

Q: OK but, given the data and the mixture parameters, what are the *most likely* label values?
A: For each point x_i, consider the values $w_k\tilde{\rho}_k(x_i)$ as scores...

A simple Gaussian Mixture Model

Use Gaussian blobs $w_k \tilde{\rho}_k = w_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

Everything is simple to write in terms of Gaussians!
There is a simple *Expectation-Maximization* algorithm :

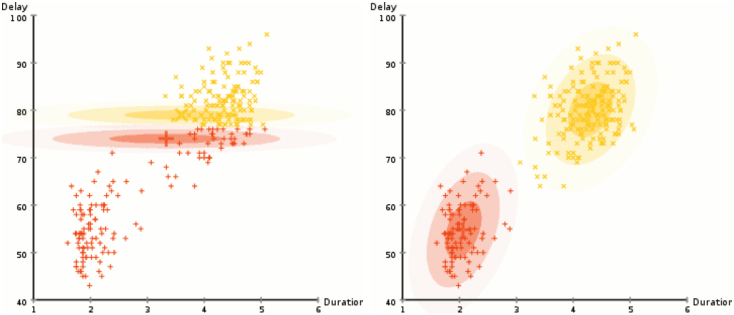
$$w_l \mapsto \sum_j \frac{w_l \mathcal{N}_{l,j}}{\sum_k w_k \mathcal{N}_{k,j}} \tag{12}$$

$$\boldsymbol{\mu}_l \mapsto \sum_j \frac{w_l \mathbf{x}_j \mathcal{N}_{l,j}}{\sum_k w_k \mathcal{N}_{k,j}} \tag{13}$$

$$\Sigma_{l,\alpha\beta} \mapsto \sum_j \frac{w_l (x_{j,\alpha} - \mu_{k,\alpha})(x_{j,\beta} - \mu_{k,\beta}) \mathcal{N}_{l,j}}{\sum_k w_k \mathcal{N}_{k,j}} \tag{14}$$

First line: *expectation* step. Second and third: *maximization* step.

The EM is often good enough, if data are well separated and well populated.

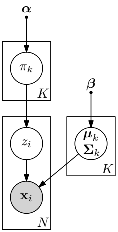


Let's go fully Bayesian...

Bayesian GMM[3]: priors on the weights and on the averages and covariances. Convenient choices:

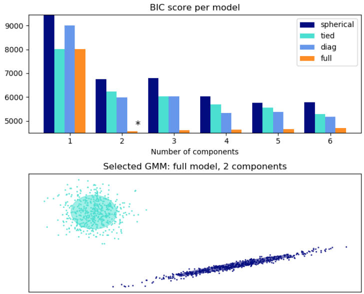
- ***Dirichlet prior*** on the weights $p(\mathbf{w}) \propto \prod_{k=1}^{K} w_k^{\alpha_k - 1}$, with hyperparameters α_k.
- Gaussian prior on averages; (inverse) Wishart prior on covariances.

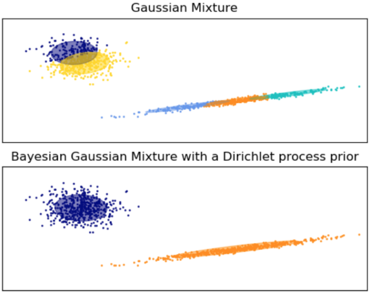
These are useful in order to integrate the posterior over the latent variables.



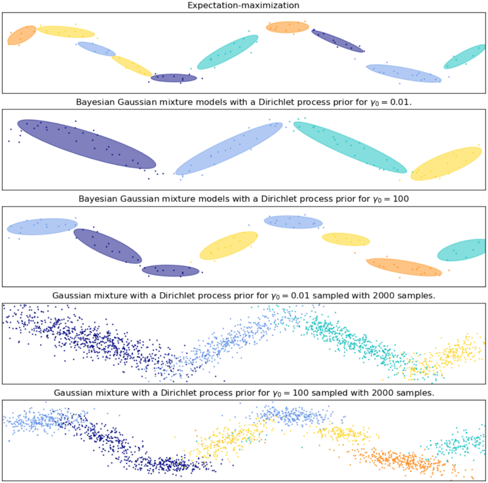
[3] For the brave:
https://www.kamperh.com/notes/kamper_bayesgmm13.pdf

As for many tools, `scikit-learn` has a module for it!
`https://scikit-learn.org/stable/modules/mixture.html`
Q: How should I choose how many blobs I need for my dataset?
(**A:** try with BIC score... or use Infinite GMM!)





Infinite GMM (Dirichlet *process* prior).



Summing up:

1. When we have noisy data or multiple populations, we can go Bayesian.
2. Bayesian: how would you bet on a model, given the data at hand?
3. posterior = likelihood $\times$ prior
4. Quantitative ways to control model complexity.
5. Binning is sinning! We can often do without bins (eq.8-9).
6. Mixture models, EM, Bayesian GMM $\rightarrow$ Infinite GMM.