Mixture Models in Statistics:

Given a sample X_i for $1 \leq i \leq n$, can it be described as coming from a *mixture* of m different subpopulations?

For example, each X_i is independently of type a with probability p_a, with then X_i normal $N(\mu_a, \Sigma_a)$, where $\sum_{a=1}^{m} p_a = 1$,?

The problem is to estimate the parameters $\theta = (p_a, \mu_a, \Sigma_a : 1 \leq a \leq m)$ and try to determine the type $a = a(i)$ for each X_i

This could be done either because we are interested in detecting population subgroups from the values X_i alone or as a device to isolate outliers.

The *likelihood* of θ for the data X_i is

$$\prod_{i=1}^{n} \left(\sum_{a=1}^{m} p_a \phi(X_i, \mu_a, \Sigma_a) \right)$$

If n is large, this is very awkward to work with directly.

See the accompanying graphics for two examples.

An example from genetics: Chromosomes can be thought of as long strings of DNA, which in turn can be thought of as long strings of nucleotides A, C, G, T.

A *gene* or *genetic locus* is a segment of a chromosome that controls some trait, usually by generating a *protein* or *enzyme*.

Parts of each genetic locus directly code for *amino acids* by consecutive triples of bases that are called *codons*, for example:

ATG GCA GAA GGC TTT AAC TTC ATT GGT ACC $\cdots$

Met Ala Glu Gly Phe Asn Phe Ile Gly Thr $\cdots$

---

where Met etc are amino acids. *Proteins* are built from strings of amino acids.

There are 20 amino acids and 64 codons. Base changes that change the amino acid are called *replacement*. Changes that do not are called *silent*.

Most silent variation is in the 3^{rd} nucleotide position, and is mostly either one of

(GAT,GAC) = Asp (GAA,GAG) = Glu or
(GCT, GCC, GCA, GCG) = Ala

In most cases, any change in either of the first two nucleotides changes is *replacement* (that is, it changes the amino acid). For the four codons for Alanine above, all 3^{rd} position changes are *silent*.

Suppose that we have samples of DNA sequences from two closely-related "sibling" species:

Species 1: ...T...A.....A...C...C...
...T...T.....A...G...C...
...T...A.....A...C...C...
Species 2: ...G...A.....C...T...T...
...G...A.....C...T...A...
...G...A.....C...T...T...

We then collect counts at each of $n = 35$ or 91 or 112 loci of the form (*McDonald-Kreitman tables*):

	mono. at diff. bases	poly. in either sp.	Sum
Replacement	M_{ri}	P_{ri}	T_{ri}
Silent	M_{si}	P_{si}	T_{si}
(Sum)	T_{Mi}	T_{Pi}	T_i

where there is a different table at each locus.

An overall excess of fixed replacements ($M_r > \frac{T_r T_M}{T}$, so that $\frac{M_r}{T} > \frac{T_r}{T}\frac{T_M}{T}$) suggests favorable mutation. Conversely, an overall deficit ($M_r < \frac{T_r T_M}{T}$) suggests unfavorable mutation.

DNA changes at individual genetic loci are likely to be too sparse for individual tables to be significant. Nevertheless, the *Mantel-Haenszel strata test* based on the differences $Z_i = M_r - \frac{T_r T_M}{T}$ can often be significant.

Random changes due to random choices of mates and who survives happen on a time scale of $N = N_e$ generations, where N_e is the effective population size, so that it is natural to scale time in this way.

It is useful to consider five different kinds of mutations, where s is the amount of selection (relative advantage) per generation:

(i)	$s < 0,\ \ \lvert sN\rvert \gg 1$	Evolutionary lethal
(ii)	$s < 0,\ \ \lvert sN\rvert = O(1)$	Weakly deleterious
(iii)	$s = 0$	Neutral
(iv)	$s > 0,\ \ \lvert sN\rvert = O(1)$	Weakly advantageous
(v)	$s > 0,\ \ \lvert sN\rvert \gg 1$	Hopeful monsters(?)

Evolutionary lethal mutations can be ignored since they rapidly disappear in time scaled by N generations, and hopeful monsters are never polymorphic in this time scale.

We will restrict ourselves to weakly selected mutations, (ii,iii,iv). This ignores the more interesting "hopeful monsters" (v), but these are relatively rare.

THE MODEL: A probability model leads to 3 parameters at each locus, specifically one parameter γ_i for selection and two mutation rates θ_{si} and θ_{ri}.

Mutations at the i^{th} locus have selection coefficients s that are normal $N(\gamma_i, \sigma_w)$. Mutants with s on the upper tail of this distribution control evolution, except for "hopeful monsters".

We also assume the γ_i are $N(\mu_\gamma, \sigma_b)$. This is then a *random effects* model with parameters $(\gamma_i, \sigma_w, \sigma_b)$ for the distribution of s within and between loci.

Under these assumptions, given data $M_{ri}, P_{ri}, M_{si}, P_{si}$, we can write down a *likelihood*

$$L = L(M_{ri}, P_{ri}, M_{si}, P_{si} \mid \mu_\gamma, \sigma_w, \sigma_b, \theta_{si}, \theta_{ri}, \gamma_i, t_{\text{div}})$$

for $1 \leq i \leq n$, if we have data for n loci. Here t_{div} is the divergence time between the two species. If $n = 56$, there are 172 parameters, of which we are primarily interested in $\mu_\gamma, \sigma_w, \sigma_b, \gamma_i$.

See the accompanying graphics for more examples.

The *maximum likelihood method* says that we should guess those parameter values that maximize L given our data, but we have far too many parameters for numerical maximization methods to work well.

Instead, we will use a *Bayesian technique* called MCMC, for *Markov chain Monte Carlo*. The first step is to assume a "prior distribution" $\pi_0(\theta)$ for $\theta = (\mu_\gamma, \sigma_w, \sigma_b, \theta_{si}, \theta_{ri}, \gamma_i, t_{\rm div})$ that is a probability distribution in those parameters.

The expression $\pi_0(\theta)L(M_{ri}, P_{ri}, M_{si}, P_{si}, \theta)$ is then a joint probability distribution for both our parameters θ as well as our data $M_{ri}, P_{ri}, M_{si}, P_{si}$. We now consider the *conditional* or *posterior distribution*

$$\pi_1(\theta) = C(M,P)\,\pi_0(\theta)L(M_{ri}, P_{ri}, M_{si}, P_{si} \mid \theta)$$

where

$$C(M,P) = 1 \Big/ \int \pi_0(z)L(M_{ri}, P_{ri}, M_{si}, P_{si} \mid z)\,dz$$

If we have enough data, this distribution should be concentrated near the true value of θ, and the center of the distribution should not depend on $\pi_0(\theta)$.

Thus we want to find means or median values of various components of $\pi_1(\theta)$. This is a reasonably tractable expression of θ except for the hideously complicated normalizing constant $C(M, P)$, which is here a 172-dimensional integral that does not simplify.

Two approaches to this problem — of finding integrals or median values of a moderately complicated expression $\pi_0(\theta)L(M_{ri}, P_{ri}, M_{si}, P_{si}, \theta)$ times an impossibly complicated normalizing constant — were found by Metropolis *et al.* (1953) and Geman and Geman (1984), and are based on three ideas.

The first idea is to look for a Markov chain Z_n on θ-space (here part of R^{172}) that has $\pi_1(\theta)$ as a stationary measure. If the Markov chain is ergodic, we can estimate the conditional distribution of the parameters θ given our data by considering the sample distribution of a single very long sample path of Z_n. That is, we can estimate

$$\theta_k \approx \frac{1}{L}\sum_{b=1}^{L}(Z_b)_k$$

with corresponding expressions for the median, exact confidence intervals, etc.

The second idea is, for parameters $\theta \in R^{172}$, is break up each step of the Markov chain $Z_n \in R^{172}$ into a sequence of 172 single steps of one-dimensional Markov chains for each component of θ. If each of the one-dimensional Markov chains is ergodic in one dimension, then one can usually show that the resulting 172-dimensional Markov chain is ergodic in R^{172}.

The third idea is due to Metropolis (1953): Given any Markov-chain transition function $q(\theta_1, \theta_2)$ on θ-space that is symmetric in θ_1 and θ_2, they show how to modify $q(\theta_1, \theta_2)$ in a simple way to form a second Markov-chain transition function $q_M(\theta_1, \theta_2)$ such that $q_M(\theta_1, \theta_2)$ has $\pi_1(\theta)$ as a stationary measure. Hastings (1970) removed the condition of symmetry on $q(\theta_1, \theta_2)$: The resulting slightly more-complicated procedure is called the Metropolis-Hasting algorithm.

Alternatively, Geman and Geman (1984) introduced the idea, for updating the i^{th} component of θ or Z_n, of sampling from the conditional distribution of that component given the current value of all of the other components as well as the data.

This idea carries the colorful name of *Gibbs Sampler*. In practice, many MCMC algorithms use Gibbs' sampler steps for some components of $\theta = Z_n$ and Metropolis random-walk steps for other components. In some cases, groups of key components are highly correlated and can be updated together.

BAYESIAN MIXTURE MODELS: As before, we have data X_i that come from m unknown subpopulations or "mixture components". If X_i comes from the a^{th} subpopulation, it is normal $N(\mu_a, \Sigma_a)$. The likelihood of the data is

$$L(X, \theta) = \prod_{i=1}^{n} \Big(\sum_{a=1}^{m} p_a \phi(X_i, \mu_a, \Sigma_a) \Big)$$

for $\theta = (p_a, \mu_a, \Sigma_a : 1 \le a \le m)$.

To make life simpler, we introduce *hidden variables* $h_i = 1, 2, \ldots, m$ that given the assumed state of the i^{th} observation. The likelihood in terms of h_i is considerably simpler:

$$\begin{aligned} L(X, \theta) &= \prod_{i=1}^{n} p_{h_i} \phi(X_i, \mu_{h_i}, \Sigma_{h_i}) \\ &= \prod_{a=1}^{m} p_a^{N_a} \prod_{[h_i = a]} \phi(X_i, \mu_a, \Sigma_a) \end{aligned}$$

where $N_a = \text{num}\{a : h_i = a\}$ and now

$$\theta = (h_i : 1 \le i \le n,\ (p_a, \mu_a, \Sigma_a) : 1 \le a \le m)$$

What makes Bayesian mixture models work so well is that this likelihood is amenable to Gibbs samplers in $\{h_i\}$, $\{p_a\}$, and $\{\mu_a, \Sigma_a\}$, in any order, providing that we choose the prior $\pi_0(\theta)$ appropriately:

$$\pi_0(\theta) = \prod_{a=1}^{m} p_a^{\alpha_a - 1} \,\mathsf{IVGN}(\mu_\epsilon, \Sigma_\epsilon)$$

where IVG is "inverse-gamma-normal", so that

$$\begin{aligned} \pi_1(\theta \mid X) &= \prod_{a=1}^{m} p_a^{N_a + \alpha_a - 1} \prod_{[h_i = a]} \phi(X_i, \mu_a, \Sigma_a) \,\mathsf{IVGN} \\ &= \prod_{a=1}^{m} p_a^{N_a + \alpha_a - 1} \,\mathsf{IVGN}(X_a, \mu_a, \Sigma_a, \epsilon) \end{aligned}$$

where X_a means all observations X_i with $h_i = a$.

The product form of $\pi_1(\theta \mid X)$ means that individual component updates will be much simpler, because under conditioning with respect to X and all other parameters, most factors in the likelihood will go away.

Gibbs sampler updates for likelihood

$$\pi_1(\theta \mid X) \;=\; \prod_{a=1}^{m} p_a^{N_a+\alpha_a-1}\,\mathsf{IVGN}(X_a, \mu_a, \Sigma_a, \epsilon)$$

Variables p_a: The conditional density in the p_a is

$$C(p, N) \prod_{a=1}^{m} p_a^{N_a+\alpha_a-1}$$

or if $m = 2$, so that $p_2 = 1 - p_1$:

$$C(p, N) p_1^{N_1} (1 - p_1)^{m-N_1}$$

This is a *beta density* for $m = 2$ or a *Dirichlet density* for $m \geq 2$. It is known how to sample efficiently from either.

Hidden variables h_i: For each i:

$$\Pr(h_i = a \mid X \text{ etc}) = p_a\,\mathsf{IVGN}(X_i, \mu_a, \Sigma_a, \epsilon)$$

For each i, this is a *Bernoulli sample* (biased coin toss) for $m = 2$ and the multinomial analog (biased multi-sided coin toss). This is also easy to do, but must be done each each hidden variable h_i.

Within-subpopulation parameters $(\boldsymbol{\mu_a, \Sigma_a})$:

The conditional density for each a is

$$\prod_{[h_i=a]} \phi(X_i, \mu_a, \Sigma_a)\, \mathsf{IVGN}(a, \epsilon) \;=\; \mathsf{IVGN}(X_a, \mu_a, \Sigma_a, \epsilon)$$

For $d = \dim(X_i) = 1$, this can be carried out by sampling γ_a from a gamma distribution with $\Sigma_a = 1/\gamma_a$ and then sampling μ_a from a normal distribution with mean approximately $\overline{X}_a$ and variance approximately Σ_a/N_a.

For $d > 2$, these steps are replaced by sampling from an *inverse Wishart* distribution and then from a multivariate normal.

Thus we have Gibbs-sampler updates of all variables, which is the principal reason why MCMC converges rapidly for Bayesian mixture models.

MCMC output for first two examples:

Example 1: Analysis based on 5000 records.
1000 burnins then 1000 samples.
Median and 95% credible interval, true values, and Gelman-Rubin statistics:

pp	(0.0567,	0.0715,	0.0871)	0.08	OK	1.2479
mu1	(-0.0800,	-0.0253,	0.0352)	0.00	OK	1.1139
sig1	(0.2231,	0.2810,	0.3391)	0.30	OK	1.2179
mu2	(1.9443,	1.9840,	2.0257)	2.00	OK	1.1077
sig2	(0.9706,	1.0004,	1.0287)	1.00	OK	1.1140

Example 2: Analysis based on 5000 records.
Analysis based on 5000 records.
1000 burnins then 20,000 samples.

pp	(0.6488,	0.8358,	0.8893)	0.88	OK	1.4948
mu1	(-0.1898,	-0.0730,	0.0097)	0.00	OK	1.3587
sig1	(0.8972,	0.9685,	1.0156)	1.00	OK	1.3023
mu2	(1.1478,	2.1527,	2.6047)	2.50	OK	1.4609
sig2	(0.9272,	1.1205,	1.4645)	1.00	OK	1.3931

Thank you for coming.