

Mixture models (Ch. 16)

- Using a mixture of distributions to model a random variable provides great flexibility.
- When?
 - When the mixture, even if not directly justifiable by the problem, provides a mean to model different zones of support of the true distribution.
 - When population of sampling units consists of several sub-populations, and within each population, a simple model applies.
- Example:

$$p(y_i|\theta, \lambda) = \lambda_1 f(y_i|\theta_1) + \lambda_2 f(y_i|\theta_2) + \lambda_3 f(y_i|\theta_3)$$

is a three-component mixture. Here:

- $\lambda_m \in [0, 1]$ and $\sum_{m=1}^3 \lambda_m = 1$.

– θ_m is the (vector) of parameters of each component distribution. For example, if mixture is normal mixture, then $\theta_m = (\mu_m, \sigma_m^2)$.

Mixture models (cont'd)

- Of interest is the estimation of component probabilities λ_m and parameters of mixture components θ_m .
- The probabilities indicate the proportion of the sample that can be expected to be generated by each of the mixture components.
- Estimation within a classical context is problematic. In a two-Gaussian mixture and for any n , the MLE does not exist because there is a non-zero probability that one of the two components does not contribute any of the observations y_i and thus the sample has no information about it. (Identifiability problem.)
- An ML can be derived when the likelihood is bounded, but often, likelihood is 'flat' at the MLE.
- Bayesian estimators in mixture models are always well-defined as long as priors are proper.

Bayesian estimation

- Suppose that the mixture components $f(y_m|\theta_m)$ are all from the exponential family

$$f(y|\theta) = h(y) \exp\{\theta y - \eta(\theta)\}.$$

- A conjugate prior for θ is

$$p(\theta|\tau) \propto \exp\{\theta - \tau\eta(\theta)\}.$$

- The conjugate prior for the vector of component probabilities $\{\lambda_1, \dots, \lambda_M\}$ is the Dirichlet with parameters $(\alpha_1, \dots, \alpha_M)$ with density

$$p(\lambda_1, \dots, \lambda_M) \propto \lambda_1^{\alpha_1-1} \dots \lambda_M^{\alpha_M-1}.$$

(for a two-component mixture, the Dirichlet reduces to a Beta).

Bayesian estimation (cont'd)

- The posterior distribution of $(\lambda, \theta) = (\lambda_1, \dots, \lambda_M, \theta_1, \dots, \theta_M)$ is

$$p(\lambda, \theta) \propto p(\lambda) \prod_{m=1}^M p(\theta_m | \tau_m) \prod_{i=1}^n \left(\sum_{m=1}^M \lambda_m f(y_i | \theta_m) \right).$$

- Each θ_m has its own prior with parameters τ_m (and perhaps also depending on some constant x_m).
- The expression for the posterior involves M^n terms of the form

$$\prod_m \lambda_m^{\alpha_m + n_m - 1} p(\theta_m | n_m \bar{y}_m, \tau_m + n_m),$$

where n_m is the size of component m and \bar{y}_m is the sample mean in component m .

- For $M = 3$ and $n = 40$, direct computation with the posterior requires the evaluation of $1.2E + 19$ terms, clearly impossible.
- We now introduce additional parameters into the model to permit the implementation of the Gibbs sampler.

Mixture models - missing data

- Consider *unobserved* indicators ζ_{im} where
 - $\zeta_{im} = 1$ if y_i was generated by component m
 - $\zeta_{im} = 0$ otherwise
- Given the vector ζ_i for y_i , it is easy to estimate θ_m .
E.g., in normal mixture, MLE of μ_1 is \bar{y}_1 , where the mean is taken over the y_i for which $\zeta_{i1} = 1$.
- Indicators ζ introduce hierarchy in the model, useful for computation with EM (for posterior modes) and Gibbs (for posterior distributions).

Setting up a mixture model

- We consider an M component mixture model (finite mixture model)
- We do not know which mixture component underlies each particular observation
- Any information that permits classifying observations to components should be included in the model (see the schizophrenics example later)
- Typically assume that all mixture components are from the same parametric family (e.g., the normal) but with different parameter values
- Likelihood:

$$p(y_i|\theta, \lambda) = \lambda_1 f(y_i|\theta_1) + \lambda_2 f(y_i|\theta_2) + \dots + \lambda_M f(y_i|\theta_M)$$

Setting up a mixture model

- Introduce unobserved indicator variables ζ_{im} where $\zeta_{im} = 1$ if y_i comes from component m , and is zero otherwise
- Given λ ,

$$p(\zeta_i|\lambda) \sim \text{Multinomial}(1; \lambda_1, \dots, \lambda_M)$$

so that

$$p(\zeta_i|\lambda) \propto \prod_{m=1}^M \lambda_m^{\zeta_{im}}.$$

Then $E(\zeta_{im}) = \lambda_m$. Mixture parameters λ are viewed as hyperparameters indexing the distribution of ζ .

Setting up a mixture model

- Joint “complete data” distribution conditional on unknown parameters is:

$$\begin{aligned} p(y, \zeta | \theta, \lambda) &= p(\zeta | \lambda) p(y | \zeta, \theta) \\ &= \prod_{i=1}^n \prod_{m=1}^M [\lambda_m f(y_i | \theta_m)]^{\zeta_{im}} \end{aligned}$$

with exactly one $\zeta_{im} = 1$ for each i .

- We assume that M is known, but fit of models with different M should be tested (see later)
- When M unknown, estimation gets complicated: unknown number of parameters to estimate! Can be done using reversible jump MCMC methods (jumps are from different dimensional parameter spaces)
- If component is known for some observations, just divide $p(y, \zeta | \theta, \lambda)$ into two parts. Obs with known membership add a single factor to product with a known value for ζ_i .

Setting up a mixture model

- Identifiability: Unless restrictions in θ_m are imposed, model is un-identified: same likelihood results even if we permute group labels.
- In two component normal mixture without restriction, computation will result in 50% split of observations between two normals that are mirror images
- Unidentifiability can be resolved by
 - Better defining the parameter space. E.g. in normal mixture can require: $\mu_1 > \mu_2 > \dots > \mu_M$
 - Using informative priors on parameters

Priors for mixture models

- Typically, $p(\theta, \lambda) = p(\theta)p(\lambda)$
- If $\zeta_i \sim \text{Mult}(\lambda)$ then conjugate for λ is Dirichlet:

$$p(\lambda|\alpha) \propto \prod_{m=1}^M \lambda_m^{\alpha_m-1}$$

where

- $E(\lambda_k) = \alpha_k / \sum_m \alpha_m$ so that relative size of α_k is prior “guess” for λ_k
- “Strength” of prior belief proportional to $\sum_m \alpha_m$.
- For all other parameters, consider some $p(\theta)$
- Need to be careful with improper prior distributions:
 - Posterior improper when $\alpha_k = 0$ unless data strongly supports presence of M mixture component
 - In mixture of two normals, posterior improper for $(\log \sigma_1, \log \sigma_2) \sim 1$.

Computation in mixture models

- Exploit hierarchical structure introduced by missing data.
- Crude estimates:
 - (a) Use clustering or other graphical techniques to assign obs to groups
 - (b) Get crude estimates of component parameters using tentative grouping of obs
- Modes of posterior using EM:
 - Estimate parameters of mixture components *averaging over* indicator
 - Complete log-likelihood is
$$\log p(y, \zeta | \theta, \lambda) = \sum_i \sum_m \zeta_{im} \log[\lambda_m f(y_i | \theta_m)]$$
 - In E-step, find $E(\zeta_{im})$ conditional on $(\theta^{(\text{old})}, \lambda^{(\text{old})})$.
 - In finite mixtures, E-step is easy and can be implemented using Bayes rule (see later)

Computation in mixture models

- Posterior distributions using Gibbs alternates between two steps:
 - draws from conditional for indicators given parameters is multinomial draws
 - draws from conditional of parameters given indicators typically easy, and conjugate priors help
- Given indicators, parameters may be arranged hierarchically
- For inference about parameters, can ignore indicators
- Posterior distributions of ζ_{im} contain information about likely components from which each observation is drawn.

Gibbs sampling

- If conjugate priors are used, the simulation is rather trivial.
- Step 1: Sample the θ_m from their conditional distributions. For a normal mixture with conjugate priors, the (μ_m, σ_m^2) parameters are sampled from univariate normal and inverted gamma distributions, respectively.
- Step 2: Sample the vector of λ s from a Dirichlet with parameters $(\alpha_m + n_m)$.
- Step 3: Simulate

$$\zeta_i | y_i, \lambda_1, \dots, \lambda_M, \theta_1, \dots, \theta_M = \sum_{m=1}^M p_{im} I_{\{\zeta_i=m\}},$$

where $i = 1, \dots, n$ and

$$p_{ij} = \frac{\lambda_j f(y_i | \theta_j)}{\sum_m \lambda_m f(y_i | \theta_m)}.$$

Gibbs sampling (cont'd)

- Even though the chains are theoretically irreducible, the Gibbs sampler can get 'trapped' if one of the components in the mixture receives very few observations in an iteration.
- Non-informative priors for θ_m lead to identifiability problems. Intuitively, if each θ_m has its own prior parameters and few observations are allocated to group m , there is no information at all to estimate θ_m . The sampler gets trapped in the local mode corresponding to component m .
- Vague but proper priors do not work either.
- How to avoid this problem? 'Link' the θ_m across groups by reparameterizing. For example, for a two-component normal, Mengersen and Robert (1995) proposed

$$p(y|\lambda, \theta) \propto \lambda N(\mu, \tau^2) + (1 - \lambda) N(\mu + \tau\theta, \tau^2\sigma^2)$$

where $\sigma < 1$.

- The 'extra' parameters represent differences in mean and variance for the second component relative to the first. Notice that this reparametrization introduces a natural 'ordering' that helps resolve the unidentifiability problems.