Preliminaries: mixtures and Markov chains

Yun-Hsiang Chan

May 2021

I. Independent mixture models

1.1 Definition and properties

In general, independent mixture distribution consists of a finite number, say m, of component distributions and a 'mixing distribution' which selects from these components.

To specify the component, one needs a discrete random variable C which performs the mixing:

$$C = \begin{cases} 1 & \text{with probability } \delta_1 \\ 2 & \text{with probability } \delta_2 = 1 - \delta_1 \end{cases}$$

The extension to m components is staightforward. Let $\delta_1, ..., \delta_m$ denote the probabilities assigned to the different components, and let $p_1, ..., p_m$ denote their probability or density functions. Let X denote the random variable which has the mixture distribution.

In the discrete case, the probability function of X is given by:

$$p(x) = \sum_{i=1}^{m} Pr(X = x|C = i)Pr(C = i) \tag{1}$$

$$= \sum_{i=1}^{m} \delta_i p_i(x) \tag{2}$$

In the continous case, the expectation of the mixture can be given in terms of the expectations of the component distributions. Letting Y_i denote the random variable with probability function p_i, we have

$$E(X) = \sum_{i=1}^{m} Pr(C = i)E(X|C = i) = \sum_{i=1}^{m} \delta_i E(Y_i)$$

More generally, for a mixture the kth moment about the origin is simply a linear combination of the kth momentions of its components Y_i:

$$E(X^k) = \sum_{i=1}^{m} \delta_i E(Y_i^k),\ k = 1, 2, ...$$

In particular, the variance of X is not a linear combination of the variances of its components Y_i.

1.2 Parameter Estimation

The estimation of parameters of mixture distribution is often performed by maximum likelihood. The likelihood of a mixture model with m components is given, for both discrete and continuous cases, by

$$L(\theta_1, ..., \theta_m, \delta_1, ..., \delta_m | x_1, ..., x_n) = \Pi_{j=1}^{n} \sum_{i=1}^{m} \delta_i p_i(x_j, \theta_i)$$

Here, $\theta_1, ..., \theta_m$ are the parameter vectors of the component distributions, $\delta_1, ..., \delta_m$ are the mixing parameters, totalling 1, and $x_1, ..., x_n$ are observations. Thus, $2m - 1$ independent parameters have to be estimated.

Parameter estimation is more conveniently carried out by direct numerical maximization of the likelihood, although the EM algorithm is a commonly used alternative.

1.3 Unbounded likelihood in mixtures

There is one aspect of mixtures of continuous distributions that differs from the discrete case and is worth highlighting. It is this: **in the vicinity of certain parameter combinations, the likelihood is unbounded**.

For instance, in the case of a mixture of normal distributions, the likelihood becomes arbitarily large if one sets of component mean equal to one of the observations and allows the corresponding vairance to tend to zero.

The source of problem: **the use of densities rather than probabilities in the likelihood**.

Solution:
Replace each density value in a likelihood by the probability of the inter-value corresponding to the recorded value.

Example:

$$\Pi_{j=1}^{n} \sum_{i=1}^{m} \delta_i p_i(x_j, \theta_i) \Rightarrow L = \Pi_{j=1}^{n} \sum_{i=1}^{m} \delta_i \int_{a_j}^{b_j} p_i(x, \theta_i) dx$$

where the interval (a_j, b_j) consists of those values would be recorded as x_j if observed.

Another Solution: Impose a lower bound to the variances and search for the best local maximum subject to that bound.

II. Markov chains

2.1 Definition and example

A sequence of discrete random variables $\{C_t : t \in N\}$ is said to be a **Markov chain** (MC) if, for all $t \in N$, it satisfies the **Markov property**

$$Pr(C_{t+1}|C_t, ..., C_1) = Pr(C_{t+1}|C_t)$$

i.e. The conditioning on the history of the process up to time t is equivalent to conditioning only on the most recent value C_t.

For compactness, we define $C^{(t)}$ as the history $(C_1, C_2, ..., C_t)$, in which case the Markov property can be written as

$$Pr(C_{t+1}|C^{(t)}) = Pr(C_{t+1}|C_t)$$

The markov property can be regarded as a first relaxation of the assumption of independence.

- Another important quantities: **transition probabilities**

$$Pr(C_{s+t} = j|C_s = i)$$

Term: homogeneous
- If the transition probabilities do not depend on s, the Markov chain is called homogeneous.
- Generally, we assume that the Markov chain under discussion is homogeneous.
- The transition probabilities will be denoted by

$$\gamma_{ij}(t) = Pr(C_{s+t} = j|C_s = i)$$

The matrix $\Gamma(t)$ is defined as the matrix with (i, j) element $\gamma_{ij}(t)$.

- Chapman-Kolmogorov equations

$$\Gamma(t + u) = \Gamma(t)\Gamma(u)$$

- It implies that

$$\Gamma(t) = \Gamma(1)^t$$

that is, the t-step transition probabilities is the t-th power of $\Gamma(1)$, the matrix of one-step transition probabilities.

- The matrix $\Gamma(1)$, which will be abbreviated as Γ, is a square matrix of probabilties with row sums equal to 1.

$$\Gamma = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ & \cdots & \\ \gamma_{m1} & \cdots & \gamma_{mm} \end{pmatrix}$$

where m denotes the number of stats of the Markov chain, and the row sums are equal to 1 can be written as $\Gamma 1' = 1'$. That is, the column vector $1'$ is a right eigenvector of Γ and corresponds to eigenvalue 1.

Term: Γ, transitional probability matrix
We shall refer to Γ as the (one-step) transition probability matrix (transition matrix).

Term: Unconditional probabilities
$Pr(C_t = j)$ is a Markov chain being in a given state at a given time t are often of interest. We denote tthese by the row vector

$$u(t) = (Pr(C_t = 1), ..., Pr(C_t = m)),\ t \in N$$

Term: initial distribution
$u(1)$ is refered as the initial distribution of Markov chain.

Result: distribution at time t+1

$$u(t+1) = u(t)\Gamma$$

2.2 Stationary distributions

A Markov chain with transition probablity matrix Γ is said to have **stationary distribution** δ (a row vector with non-negative elements), if $\delta\Gamma = \Gamma$ and $\delta 1' = 1$, where $1'$ is the right eigenvector of Γ corresponds to eigenvalue 1.

Since $u(t+1) = u(t)\Gamma$, a Markov chain started from its staionary distribution will continue to have that distribution at all subsequent time points, and we shall refer to such a process as a **stationary Markov chain**.

A general result that can conveniently be used to compute a stationary distribution is as follows: the vector Γ with non-negative elements is a stationary

distribution of the Markov chain if

$$\delta(I_m - \Gamma + U) = 1$$

i.e. a stationary distribution can be found by deleting one of the equations in the system $\delta\Gamma = \delta$ and replacing it by $\sum_{i=1} \delta_i = 1$.

2.3 Autocorrelation function

The ACF of $\{C_t\}$, assumed stationary and irreducible, may be obtained as follows.

$$Cov(C_t, C_{t+k}) = \delta V \Gamma^k v' - (\delta v')^2$$

If Γ is diagonalizable, and its eigenvalues are denoted by $\omega_2, \omega_3, ..., \omega_m$, then Γ can be written as

$$\Gamma = U \Omega U^{-1}$$

Then,

$$\begin{aligned} Cov(C_t, C_{t+k}) &= \delta V U \Omega^k U^{-1} v' - (\delta v')^2 && (3) \\ &= a \Omega^k b' - a_1 b_1 && (4) \\ &= \sum_{i=1}^{m} a_i b_i \omega_i^k && (5) \end{aligned}$$

where $a = \delta V U$ and $b' = U^{-1} v'$

$$\rho(k) = Corr(C_t, C_{t+k}) = \sum_{i=2}^{m} a_i b_i w_i^k / \sum_{i=2}^{m} a_i b_i$$

2.4 Estimating transition probabilites

By using ML estimator, we will obtain the equivalent estimator as empirical estimator:

$$\hat{\gamma}_i j = \frac{f_{ij}}{\sum_{k=1}^{m} f_{ik}},\ i, j = 1, ..., m$$

The assumption of stationary of the Markov chain was not used in the above derivation. If we wish to assume stationarity, we may use the uncondtional likelihood.

2.5 Higher-order Markov chains

Omit it before we enter Section 10.3 or Sections 17.3.2 and 19.2.2