Mixture Models

Wilfried Seidel

Helmut-Schmidt-Universität, D-22039 Hamburg, Germany

November 4, 2015

1 Introduction

Mixture distributions are convex combinations of "component" distributions. In statistics, these are standard tools for modelling heterogeneity in the sense that different elements of a sample may belong to different components. However, they may also be used simply as flexible instruments for achieving a good fit to data when standard distributions fail. As good software for fitting mixtures is available, these play an increasingly important role in nearly every field of statistics.

It is convenient to explain finite mixtures (i.e. finite convex combinations) as theoretical models for cluster analysis, but of course the range of applicability is not at all restricted to the clustering context. Suppose that a feature vector X is observed in a heterogeneous population, which consists of k homogeneous subpopulations, the "components". It is assumed that for $i = 1, \ldots, k$, X is distributed in the i-th component according to a (discrete or continuous) density $f(x,\, \theta_i)$ (the "component density"), and all component densities belong to a common parametric family $\{f(x, \theta),\ \ \theta \in \Theta\}$, the "component model". The relative proportion of the i-th component in the whole population is p_i, $p_1 + \cdots + p_k = 1$. Now suppose that an item is drawn randomly from the population. Then it belongs to the i-th component with probability p_i, and the conditional probability that X falls in some set A is $\Pr(X \in A \mid \theta_i)$, calculated from the density $f(x,\, \theta_i)$. Consequently, the marginal probability is

$$\Pr(X \in A \mid P) = p_1 \Pr(X \in A \mid \theta_1) + \cdots + p_k \Pr(X \in A \mid \theta_k)$$

with density

$$f(x,\, P) = p_1 f(x,\, \theta_1) + \cdots + p_k f(x,\, \theta_k), \tag{1}$$

a "simple finite mixture" with parameter $P = ((p_1, \ldots, p_k), (\theta_1, \ldots, \theta_k))$. The components p_i of P are called "mixing weights", the θ_i "component paramaters". For fixed k, let $\mathcal{P}_k$ be the set of all vectors P of this type, with $\theta_i \in \Theta$ and nonnegative mixing weights summing up to one. Then $\mathcal{P}_k$ parameterizes all mixtures with not more than k components. If all mixing weights are positive and component densities are different, then k is the exact number of components. The set of all simple finite mixtures is parameterized by $\mathcal{P}_{\text{fin}}$, the union of all $\mathcal{P}_k$.

This model can be extendes in various ways. For example, all component densities may contain additional common parameters (variance parameters, say), they may depend on covariables (mixtures of regression models), and also the mixing weights may depend on covariables. Mixtures on time series models are also considered. Here I shall concentrate on simple mixtures, as all relevant concepts can be explained very easily in this setting. These need not be finite convex combinations; there is an alternative and more general definition of simple mixtures: Observe that the parameter P can be considered as a discrete probability distribution on Θ which assigns probability mass p_i to the parameter θ_i. Then equation (1) is an integral with respect to this distribution, and if ξ is an arbitrary probability distribution Θ, a mixture can be defined by

$$f(x,\, \xi) = \int_\Theta f(x,\, \theta)\, d\xi(\theta)\ . \tag{2}$$

It can be considered as the distribution of a two-stage experiment: First, choose a parameter θ according to the distribution ξ, then choose x according to $f(x,\, \theta)$. Here, ξ is called a "mixing distribution", and mixture models of this type can be parameterized over every set Ξ of probability distributions on Θ.

In statistical applications of mixture models, a nontrivial key issue is identifiability, meaning that different parameters describe different mixtures. In a trivial sense, models parametrized over vectors P are never identifiable: All vectors that correspond to the same probability distribution on Θ describe the same mixture model. For example, any permutation of the sequence of components leaves the mixing distribution unchanged, or components may be added with zero mixing weights. Therefore identifiability can only mean that parameters that correspond to different mixing distributions describe different mixture models. However, also in this sense identifiability is often violated. For example, the mixture of two uniform distributions with supports $[0,\, 0,5]$ and $[0.5,\, 1]$ and equal mixing weights is the uniform distribution with support $[0,\, 1]$. On the other hand, finite mixtures of many standard families (normal, Poisson, ...) are identifiable, see for example Titterington [8]. Identifiability of mixtures of regression models has been treated among others by Hennig [3]. A standard general reference for finite mixture models is McLachlan and Peel [6].

2 Statistical Problems

Consider a mixture model with parameter η (vector or probalility measure). In the simplest case, one has i.i.d. data $x_1, \ldots, x_n$ from $f(x,\, \eta)$, from which one wants to gain information about η. Typical questions are estimation of (parameters of) η, or mixture diagnostics: Is there strong evidence for a mixture (in contrast to homogeneity in the sense that η is concentrated at some single parameter θ)? What is the (minimum) number of mixture components?

A variety of techniques has been developed. The data provide at least implicitly an estimate of the mixture, and equations 1 and 2 show that mixture and mixing distribution are related by a linear (integral) equation. Approximate solution techniques have been applied for obtaining estimators, and moment estimators have been developed on basis of this structure. Distance estimators exhibit nice properties. Traditionally, mixture diagnostics has been handled by graphical methods. More recent approaches for estimation and diagnostics are based on Bayesian or likelihood techniques; likelihood methods will be addressed below. Although Bayesian methods have some advantages over likelihood methods, they are not straightforward (for example, usually no "natural" conjugate priors are available, therefore posteriors are simulated using MCMC. Choice of "noninformative" priors is not obvious, as improper priors usually lead to improper posteriors. Nonidentifiability of $\mathcal{P}_k$ causes the problem of "label switching"). A nice reference for Bayesian methods is Frühwirth-Schnatter [2].

Let me close this section with a short discussion of robustness. Robustness with respect to outliers is treated by Hennig [4]. Another problem is that mixture models are extremely nonrobust with respect to misspecification of the component model. Estimating the component model in a fully nonparametric way is of course not possible, but manageable alternatives are for example mixtures of log-concave distributions. Let me point out, however, that issues like nonrobustness and nonidentifiability only cause problems if the task is to interpret the model parameters somehow. If the aim is only to obtain a better data fit, one need not worry about them.

3 Likelihood Methods

In the above setting, $l(\eta) = \log(f(x_1, \eta)) + \cdots + \log(f(x_n, \eta))$ is the log likelihood function. It may have some undesirable properties: First, the log likelihood is often unbounded. For example, consider mixtures of normals. If the expectation of one component is fixed at some data point and the variance goes to zero, the likelihood goes to infinity. Singularities usually occur at the boundary of the parameter space. Second, the likelihood function is usually not unimodal, although this depends on the parametrization. For example, if the parameter is a probability distribution as in equation 2 and if the parameter space Ξ is a convex set (with respect to the usual linear combination of measures), the log likelihood function is concave. If it is bounded, there is a nice theory of "nonparametric likelihood estimation" (Lindsay [5]), and "the" "nonparametric maximum likelihood estimator" is in some sense uniquely defined and can be calculated numerically (Böhning [1], Schlattmann [7]).

Nonparametric methods, however, work only for low dimensional component models, whereas "parametric" estimation techniques like the Expectation-Maximization (EM) method work for nearly any dimension. The latter is a local maximizer for mixture likelihoods in $\mathcal{P}_k$. Here the mixture likelihood is usually multimodal; moreover, it can be very flat. Analytic expressions for likelihood maxima usually do not exist, they have to be calculated numerically. On the other hand, even for unbounded likelihoods, it is known from asymptotic theory, that the simple heuristics of searching for a large local maximum in the interior of the parameter space may lead to reasonable estimators. However, one must be aware that there exist "spurious" large local maxima that are statistically meaningless. Moreover, except from simple cases, there is no manageable asymptotics for likelihood ratio.

Some of the problems of pure likelihood approaches can be overcome by considering penalized likelihoods. However, here one has the problem of choosing a penalization parameter. Moreover, the EM algorithm is a basic tool for a number of estimation problems, and it has a very simple structure for simple finite mixtures. Therefore it will be outlined in the next sextion.

4 EM Algorithm

The EM algorithm is a local maximization technique for the log likelihood in $\mathcal{P}_k$. It starts from the complete-data log-likelihood. Suppose that for observation x_i the (fictive) component membership is known. It is defined by a vector $z_i \in \Re^k$ with $z_{ij} = 1$, if x_i belongs to j-th component, and zero elsewhere. As a random variable Z_i, it has a multinomial distribution with parameters $k, p_1, \ldots, p_k$. Then the complete data likelihood and log likelihood of P, respectively, are $L_c(P) = \prod_{i=1}^n \prod_{j=1}^k (p_j\, f(x_i,\, \theta_j))^{z_{ij}}$ and $l_c(P) = \log(L_c(P)) = \sum_{i=1}^n \sum_{j=1}^k z_{ij}\ \log p_j + \sum_{i=1}^n \sum_{j=1}^k z_{ij}\ \log f(x_i,\, \theta_j)$.

The EM needs a starting value P_0, and then proceeds as an iteration between an "E-step" and an "M-step" until "convergence". The first E-step consists in calculating the conditional expectation $E_{P_0}(l_c(P)\,|\,x_1,, \ldots, x_n)$ of $l_c(P)$ for arbitrary P, given the data, under P_0. However, as the only randomness is in the z_{ij}, we obtain

$$E_{P_0}(l_c(P)\,|\,x_1,, \ldots, x_n) = \sum_{i=1}^n \sum_{j=1}^k \tau_j(x_i|P_0) \log p_j + \sum_{i=1}^n \sum_{j=1}^k \tau_j(x_i|P_0) \log f(x_i,\, \theta_j)\,,$$

where

$$\tau_j(x_i|P_0) = \mathrm{Pr}_{P_0}(Z_{ij} = 1\,|\,x_i) = \frac{p_j f(x_i,\, \theta_j)}{f(x_i,\, P_0)}$$

is the conditional probability that the i-th observation belongs to component j, given the data, with respect to P_0.

In the following M-step, $E_{P_0}(l_c(P)\,|\,x_1,, \ldots, x_n)$ is maximized with respect to P. As it is the sum of terms depending on the mixing weights and on the parameters only, respectively, both parts can be maximized separately. It is easily shown that the maximum in the p_j is achieved for $p_j^{(1)} = (1/n) \sum_{i=1}^n \tau_j(x_i|P_0), j = 1, \ldots, n$. For component densities from exponential families, similar simple solutions exist for the θ_j, therefore both the E-step and te M-step can be carried out here analytically. It can be shown that (i) the log-likelihood is not decreasing during the iteration of the EM, and (ii) tat under some regularity conditions it converges to a stationary point of the likelihood function. However, this may also be a saddle point.

It remains to define the stopping rule and the starting point(s). Both are crucial, and the reader is referred to the literature. There are also techniques that prevent from convergence to singularities or spurious maxima. A final nice issue of the EM is that it yields a simple tool for classification of data points: If $\hat{P}$ is an estimator, then $\tau_j(x_i|\hat{P})$ is the posterior probability that x_i belongs to class j with respect to the "prior" $\hat{P}$. The Bayesian classification rule assigns observation i to the class j that maximizes $\tau_j(x_i|\hat{P})$, and the $\tau_j(x_i|\hat{P})$ measure the plausibility of such a clustering.

Let me finally remark that the Bayesian analysis of a mixture model via the Gibbs sampler takes advantage of exactly the same structure by simulating both the parameters and the missing data (see, e.g., Frühwirth-Schnatter, [2]).

5 Number of Components, Testing and Asymptotics

Even if one has an estimator in each $\mathcal{P}_k$ from the EM, the question is how to assess the number of components (i.e. how to choose k). Usually information criteria like AIC and BIC are recommended. An alternative is to perform a sequence of tests of k against $k+1$ components, for $k = 1, 2 \ldots$.

There are several tests for homogeneity, i.e. for the "component model", as for example goodness of fit or dispersion score tests. For testing k_0 against k_1 components, a likelihood ratio test may be performed. However, the usual χ^2-asymptotics fails, so critical values have to be simulated. Moreover, the distribution of the test statistic usually depeds on the specific parameter under the null hypothesis. Therefore some sort of bootstrap is needed, and as estimators have to be calculated numerically, likelihood ratio tests are computationally intensive.

Let me close with some remarks on asymptotics. Whereas asymptotic normality of estimators is guaranteed under some conditions, the usual asymptotics for the likelihood ratio test fails. The reason is that under the null hypothesis, the parameter P_0 is on the boundary of the parameter space, it is not identifiable and the Fisher information matrix in P_0 is singular. There is an asymptotic theory under certain restrictive assumptions, but it is usually hard to calculate critical values from it.

References

[1] Böhning, D. (2000), *Finite Mixture Models*, Chapman & Hall, Boca Raton.

[2] Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer, New York.

[3] Hennig, C. (2000), *Identifiability of Models for Clusterwise Linear Regression*, Journal of Classification 17, 273-296.

[4] Hennig, C. (2004), *Breakdown points for ML estimators of location-scale mixtures*, Annals of Statistics 32, 1313-1340.

[5] Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*, NSC-CBMS Regional Conference Series in Probability and Statistics, Volume 5.

[6] McLachlan, G. J., Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

[7] Schlattmann, P. (2009), Medical Applications of Finite Mixture Models . Spinger, Berlin, Heidelberg.

[8] Titterington, D. M.; Smith, A. F. M, Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.