Estimation and Testing Problems in Poisson Mixtures

Karlis Dimitrios

Abstract

Mixture models is a rapidly developing area of statistics with applications to a variety of fields. This thesis is devoted to Poisson mixtures which naturally arise as alternative models when the simple Poisson model fails to describe the data. For example, it is known that the Poisson distribution is characterised by its property of having a variance equal to its mean. This property is not usually satisfied by the data. This case is usually referred to as overdispersion. Poisson mixtures can provide flexible alternative models that can represent the inhomogeneity of the population. The idea is that the persons comprising the entire population do not have the same Poisson parameter. Instead, their parameter varies according to a distribution, termed as the mixing distribution. By the law of total probability, Poisson mixtures arise. The properties of Poisson mixtures are examined in depth. Due to their complexity, only a few have been examined to the literature. Several members of this family are presented in this thesis, emphasising their interrelations.

Among these models finite Poisson mixtures are very popular since they admit a simple and natural interpretation, as models describing a population consisting of a finite number of subpopulations. Moreover, even if the true mixing distribution is continuous, one is restricted to estimate it via a finite distribution. Estimation methods for finite Poisson mixtures are explored. Two distinct cases appear in practice. The first assumes that the number of components is fixed and tries to estimate the parameters maximising a criterion over the space of all mixing distributions with the given number of support points. The second, termed as the semiparametric case, treats the number of components as an unknown parameter which has to be estimated from the data. For the first case, the EM algorithm is critically reviewed. Initial values that can help the algorithm to converge quicker are examined via a simulation experiment. An improvement of the EM algorithm for mixtures based on a property of mixtures from the exponential family is proposed. For the semiparametric case, the algorithms proposed for obtaining maximum likelihood estimates are examined. These algorithms do not seem to be adequate for the case of Poisson mixtures since the number of support points is usually small, and the algorithms do not work properly.

The problem of determining the number of components is also examined. A new method is proposed. The method is based on sequentially applying the likelihood ratio test using bootstrap methods to determine the distribution of the test statistic. The properties of this new method are also examined.

Several other methods of estimation are also reviewed. For the moment method, the existence of the estimates is explored. The results show that the moment estimates do not exist very often. Moreover, the small sample comparison of the moment estimators to the maximum likelihood estimators discourage their use. An alternative method which

uses the zero frequency instead of the third moment is developed. this method is useful when the zero proportion is large.

A new method, which is efficient and robust at the same time is introduced. The method is based on minimising the Hellinger distance. The obtained estimators are examined and shown to be robust relative to the maximum likelihood estimators. This robustness property is used for proposing inferential procedures for Poisson mixtures. It is proposed to use the Minimum Hellinger Estimators for semiparametric estimation. Moreover, diagnostic graphs are not influenced by a few observations and can, thus, detect if a Poisson distribution is appropriate. In addition, an alternative to the likelihood ratio test is proposed. This is termed as the Hellinger Deviance Test and is based on the difference of the Hellinger distance between two hypotheses. This test statistic is powerful and robust to outlier contamination. An algorithm for the estimation of the parameters is provided which facilitates the application of minimum Hellinger methodologies.