Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 5444–5451
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Non-parametric log-concave mixtures

Paul H.C. Eilers[a,*], M.W. Borgdorff[b]

[a]*Department of Medical Statistics, Leiden University Medical Centre, P.O. Box 9604, 2300 RC Leiden, The Netherlands*
[b]*Royal Netherlands Tuberculosis Association (KNCV), The Hague, The Netherlands*

Available online 8 September 2006

Abstract

Finite mixtures of parametric distributions are often used to model data of which it is known or suspected that there are sub-populations. Instead of a parametric model, a penalized likelihood smoothing algorithm is developed. The penalty is chosen to favor a log-concave result. The standard EM algorithm ("split and fit") can be used. Theoretical results and applications are presented.

Keywords: Finite mixtures; Penalized likelihood; Smoothing

1. Introduction

Mixture models for frequency distributions generally come in two flavors. The most general one represents situations in which an observed distribution does not fit well to any parametric theoretical distribution. Then one tries to model the data as a, discrete or continuous, mixture of simple distributions, like Poisson or normal (Böhning, 2000). By allowing enough freedom for the mixing distribution one can, in principle, model any empirical distribution. The second flavor is more specialized: the subject matter more or less dictates a discrete mixture with a few components (Everitt and Hand, 1981). The components may stand for different year classes of animals, infected and non-infected persons, types of soils, and so on. Often one can discern several modes in such a distribution. Now the choice of the component distributions becomes more critical, because there are only a few mixing parameters.

One approach is to try several parametric forms, like normal, log-normal or gamma, possibly a different one for each component. This may lead to quite complicated algorithms, especially so when the data are truncated at the boundaries. Here we present an alternative approach, using a non-parametric smoother that favors log-concave distributions. The popular and effective EM scheme is used, iterating between splitting the data into sub-groups (the E-step) and smoothing these, maximizing a penalized likelihood (a penalized M-step).

In the next section we present the penalized likelihood smoother and the EM algorithm in some detail. Some applications appear in Section 3. Section 4 discusses further work.

* Corresponding author. Tel.: +31 71 5269704; fax: +31 71 5268280.
E-mail address: p.eilers@lumc.nl (P.H.C. Eilers).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.08.027

2. Theory

2.1. Discrete penalized likelihood smoothing

Simonoff (1983) presented a simple but effective algorithm for smoothing of contingency tables with ordered categories. Let the observations be y_i, $i = 1, \ldots, m$. Assume a Poisson distribution in cell i with expected value $\mu_i = e^{\eta_i}$. Maximize the penalized log-likelihood

$$L^* = \sum_{i=1}^{m}(y_i\eta_i - \mu_i) - \lambda\sum_{i=2}^{m}(\Delta\eta_i)^2/2, \tag{1}$$

where $\Delta\eta_i = \eta_i - \eta_{i-1}$. The idea behind penalized likelihood is to strike a balance between fit to the data (the first term in L^*) and smoothness (the second term). A smooth series η will show small differences between neighboring values; this is the goal of the penalty. An efficient iterative algorithm for minimizing L^* will be presented below.

Amazingly, in his book on smoothing, Simonoff (1996) gives little attention to this smoother, favoring kernels and local likelihood instead. But it is a very useful histogram smoother. The one disadvantage is that the smooth distribution μ tends toward a constant (with value $\sum y_i/m$) as λ increases. This can easily be remedied by using third differences in the penalty, the second term of (1):

$$L^* = \sum_{i=1}^{m}(y_i\eta_i - \mu_i) - \lambda\sum_{i=4}^{m}(\Delta^3\eta_i)^2/2. \tag{2}$$

The penalized likelihood equations that follow from (2) are

$$\lambda D'D\eta = y - \mu, \tag{3}$$

where D is a matrix such that $D\eta = \Delta^3\eta$. Assume that we have an approximate solution $\tilde{\eta}$, and that $\tilde{\mu} = e^{\tilde{\eta}}$. Linearizing the system in (3), using $\mu_i - \tilde{\mu}_i \approx \tilde{\mu}_i(\eta_i - \tilde{\eta}_i)$, we get

$$(\tilde{M} + \lambda D'D)\eta = y - \tilde{\mu} + \tilde{M}\tilde{\eta}. \tag{4}$$

Starting with $\tilde{\eta} = \log(y + 0.5)$, quadratic convergence is generally reached in 5–10 iterations.

From (3) it follows that for very high λ, we will have essentially $\Delta^3\eta = 0$. As third differences are zero for any quadratic polynomial $\eta_i = a_2 i^2 + a_1 i + a_0$, this means that η will approach such a polynomial for large λ. The coefficients a will be such as to maximize the log-likelihood, the first term in (2). The result is that for large λ, μ approaches a discretized normal distribution.

Similarly we find that mean and variance of the smooth distribution are equal to those of the observed one, because $\sum_i \Delta^3 i^k = 0$ for $k = 0$, 1 or 2: the moments up to order 2 of the raw and the smoothed histogram are equal, independent of the amount of smoothing.

The system of equations in (4) generally will not be large, as most histograms have less than 50 or 100 bins. Experience has shown that the smooth μ is very insensitive to the choice of bin widths and the positions of the boundaries. Even bins that are considered far too narrow for standard histograms give very good results, thanks to the smoothing power of the penalty.

It seems that we have the perfect histogram smoother here. One question remains: how do we optimize λ, the weight of the penalty? Akaike's Information Criterion (AIC), combining the deviance $\mathrm{Dev}(y|\mu)$ and the effective model dimension (Dim) is a good choice:

$$\mathrm{AIC} = \mathrm{Dev}(y|\mu) + 2\,\mathrm{Dim} = 2\sum_{i=1}^{m} y_i \log(y_i/\mu_i) + 2\,\mathrm{Dim}. \tag{5}$$

For the effective dimension we follow the advice of Hastie and Tibshirani (1990) and take the trace of the smoothing matrix in the linearized equations (4):

$$\mathrm{Dim} = \mathrm{trace}[(M + \lambda D'D)^{-1}M]. \tag{6}$$