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):

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

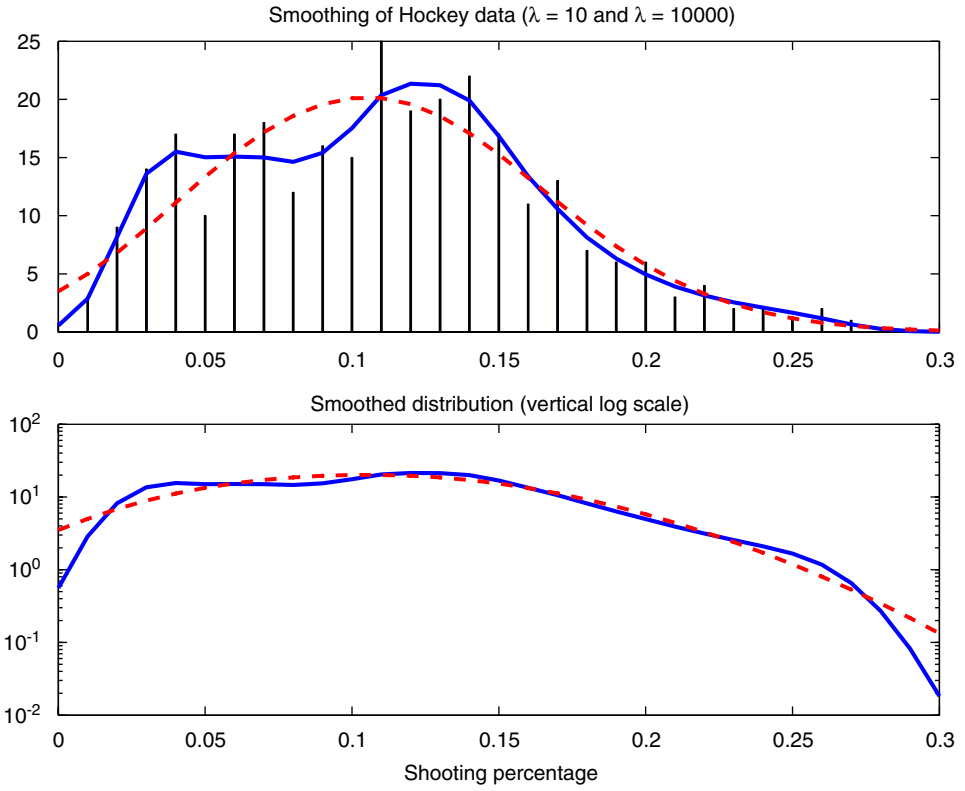

Fig. 1. Smoothing the hockey data. Top: raw and smoothed histogram, for $\lambda = 10$ and 1000 (larger λ gives a smoother curve). Bottom: logarithms of the smoothed histograms.

A simple grid search, with $\log \lambda$ running over a linear grid produces a graph of AIC, which is instructive in itself, and an approximate position of the optimum. A very precise value of λ is seldom required, but if needed one can adapt a one-dimensional minimization routine.

A distribution $f(x)$ is log-concave if $\mathrm{d}^2 \log f/\mathrm{d}x^2 < 0$ everywhere. The discrete analogon is $\Delta^2\eta < 0$ everywhere. As the penalty pushes η in the direction of a quadratic curve, we achieve this situation for large enough λ. In the (normal) limit, for very large λ, it will most likely hold. Unless we have strongly U-shaped data on a bounded domain, the tails of the distribution μ will generally slope downwards.

To illustrate the smoother, we use an example from Simonoff's book. He presents the percentage of successful shots at the goal of (ice-) hockey players. A histogram is shown in the upper panel of Fig. 1, together with the results of the smoother for several values of λ. As λ increases, the curve of μ becomes unimodal and its logarithm approaches a concave shape, as shown by the lower panel.

As convergence criterion we use the relative size of changes in $\hat{\mu}$: $\max(\delta\mu)/\max(\mu) < t$, where $\delta\mu$ is the change in μ from one iteration to the next. In our experience $t = 10^{-4}$ worked well. With this choice of criterion, convergence was reached in 6 iterations, starting from $\tilde{\eta} = \log(y + 1)$.

2.2. *Smooth mixtures*

The EM (estimation-maximization) algorithm is effective and intuitively attractive for estimating a finite mixture of parametric distributions. To simplify the presentation, we consider only two component distributions, $f_1(\theta_1)$ and $f_2(\theta_2)$. Here θ_1 and θ_2 stand for vectors of parameters. The observed histogram y will be fitted by $npf_1(\theta_1) + n(1-p)f_2(\theta_2)$, with $0 < p < 1$ the mixing parameter and $n = \sum_i y_i$. Assume that approximate solutions $\tilde{p}$, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are known. In the E-step we split each count y_i into $q_i y_i$ and $(1-q_i)y_i$, with $q_i = pf_1(\theta_1)/[pf_1(\theta_1) + (1-p)f_2(\theta_2)]$, giving two subsets of pseudo-counts. In the M-step we get an improved estimate of the parameter vector θ_1 by fitting f_1 to the

first subset and similarly for θ_2. An improved estimate of p is found from $\sum_i y_{i1}/\sum_i y_i$. Repeating this process leads to the maximum likelihood solution. The speed of convergence depends strongly on the data. Also local maxima can occur, so a prudent choice of the starting values is important.

To apply the penalized likelihood smoother, we do not use parametric models, but smooth the subsets in the M-step, giving estimates μ_{i1} and μ_{i2}. We could call this a penalized EM algorithm, because in the M-step we maximize a penalized likelihood. In AIC we use $\mu_i = \mu_{i1} + \mu_{i2}$ to compute the deviance and we take the sum of the effective dimensions of the two smoothing matrices.

We do not provide a proof of convergence to a (local) optimum. But it seems probable that this will hold. It is well known that in EM fitting of parametric distributions the likelihood increases monotonically (Everitt and Hand, 1981). It is reasonable to expect that this will be true here for the penalized likelihood. We have not encountered contradictory evidence.

With EM algorithms, convergence can be slow. In our application this can occur when mixtures are hard to separate. Our criterion is the ratio $\max(\delta\mu)/\max(\mu) < t$, where $\delta\mu$ is the change in μ from one EM iteration to the next. In our experience $t = 10^{-4}$ worked well, but when convergence is slow a more stringent value might be necessary.

3. Applications

In this section we present applications. We revisit the hockey data, we analyze frequency distributions of surveys of Tuberculin tests, and we show an application to two-dimensional data.

3.1. The hockey data

There is reason to assume a mixture for the hockey data, as teams have defenders and forwards; the latter may be assumed to be more successful in their shots at the goal. The upper panel of Fig. 2 shows a histogram, the fitted mixture

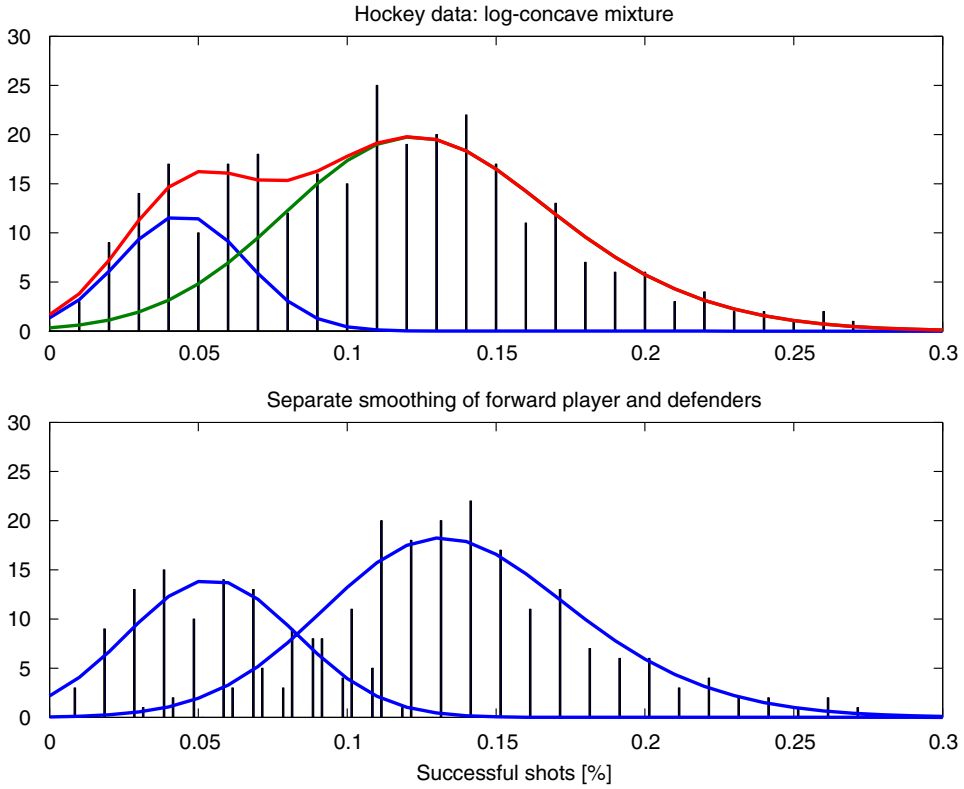

Fig. 2. Fitting a log-concave mixture to the hockey data. Top: mixed data, component distributions and their sum. Bottom: data split into defenders and forwards and smoothed separately. In both panels $\lambda = 10^4$.

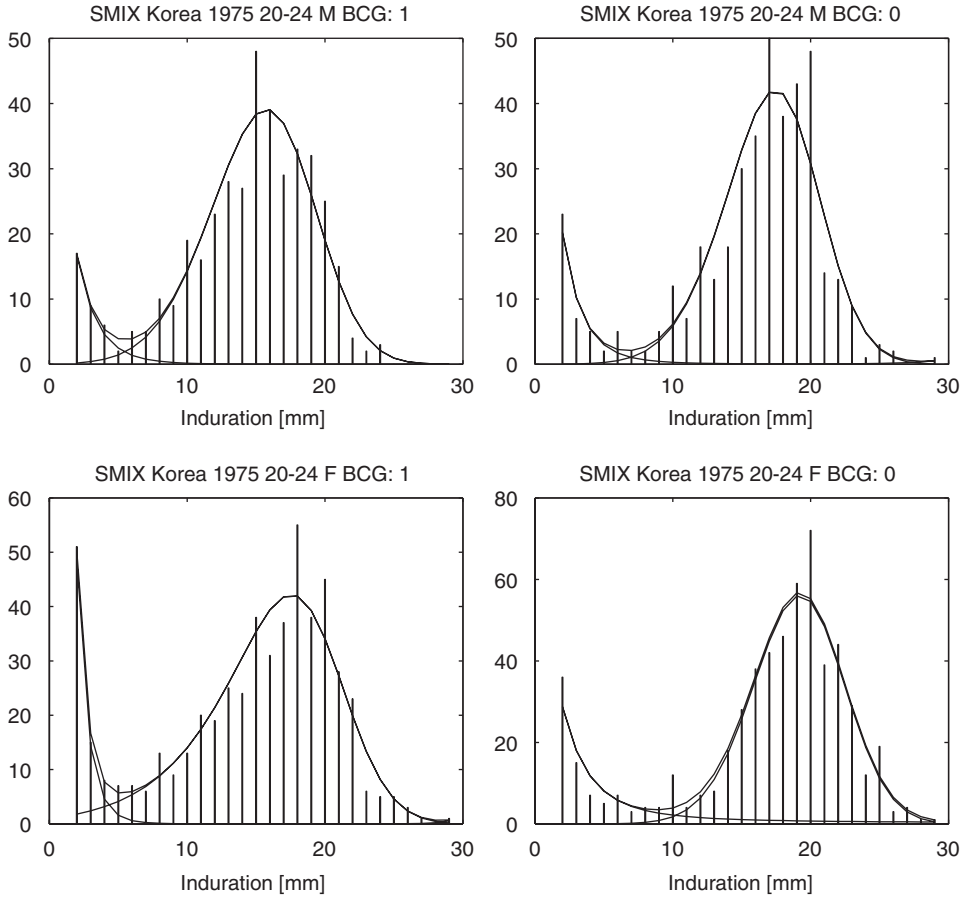

Fig. 3. Estimated mixture of two smooth induration distributions ($\lambda = 10^4$) for the age group 20–24 of the TB survey in Korea in 1975. The data (narrow bars), the smooth component distributions and their sum are shown in each panel.

and its sum. To find the optimum for λ, it was varied on a "nice" grid (1, 2, 5, 10 and so on). The minimum of AIC was 30.6 (Dim $= 3.0 + 4.1 = 7.1$) at $\lambda = 10^4$. The number of EM steps was 46 (taking 1 s on a 1000 MHz Pentium III).

Judging from the effective dimensions the component distributions do not differ much from the normal distribution, which has Dim $= 3$. In this case there would be no obvious need for smooth non-parametric distributions. Notice, however, that in the mixture the left component is truncated at zero. The penalized likelihood smoother has no problem with that. When fitting parametric distributions one has to account for this, which greatly complicates the computations. This suggests that the non-parametric approach, with a very large λ, might even be useful when one plans to fit a mixture of (truncated) normals.

Simonoff also presents the actual position of each player, so we have an independent check on our result. It is shown in the lower panel of Fig. 2.

3.2. Induration frequencies

The Tuberculin test is being used on a large scale in population surveys to estimate the prevalence of tuberculosis (TB). Tuberculin is injected into the skin. If after 72 h a swelling of the skin (called induration) occurs, its size is being measured and recorded. Generally one finds that a large proportion of people do not show any reaction (zero induration). Of those who react the induration can be caused by *Mycobacterium tuberculosis*, the cause of TB, or by so-called a-specific reactions to environmental Mycobacteria. Hence the biology indicates a mixture of two distributions. Indurations of TB infected persons show a relatively high mean (15 mm or large) and a unimodal distribution.

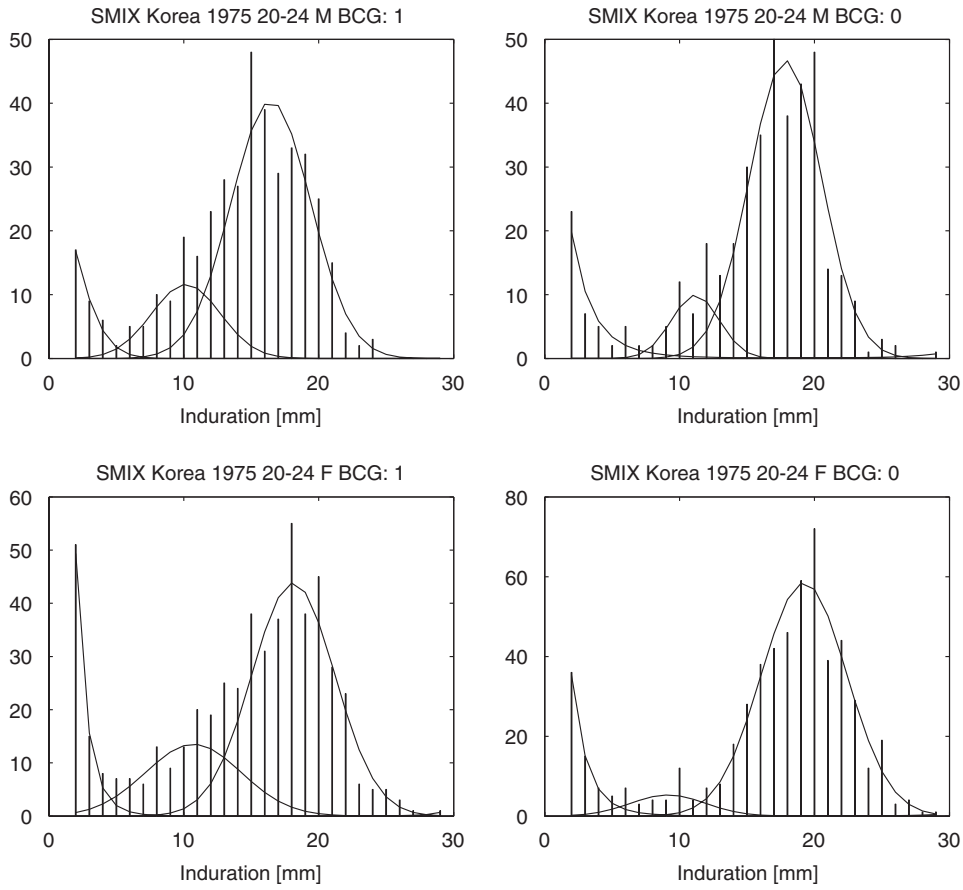

Fig. 4. Estimated mixture of three smooth induration distributions ($\lambda = 10^4$) for the age group 20–24 of the TB survey in Korea in 1975. The data (narrow bars), the smooth component distributions and their sum are shown in each panel.

The distribution of the a-specific reactors looks more or less exponential with average below 5 mm. In countries where BCG vaccination against TB is common, a third component appears in the mixture, situated more or less between the other two.

Fig. 3 shows data and estimated mixtures for a survey in Korea in 1975, covering men and women of age 20–24, split according to their BCG status (1 = vaccinated, 0 = not). In Fig. 3 the right distribution has a rather heavy left tail for BCG positive persons. This illustrates the capability of the model to recover skew unimodal components. From the biological point of view it is, however, hard to accept, as we expect this distribution to stand for TB infected people, for whom it is known that the induration seldom is less than 10 mm. Apparently a mixture of two distributions is too simple. We therefore use a model with a mixture of three distributions. Now we get a much more realistic result, as shown in Fig. 4.

From the estimated components one can easily derive important epidemiological parameters, like the prevalence of TB and means and standard deviations of the separate distributions.

3.3. Old Faithful

The data of the Old Faithful geyser (Azzalini and Bowman, 1990) are popular for illustrating smoothing and mixture estimation. Here we use them to present a simplified approach to the modeling of two-dimensional distributions by log-concave mixtures. If we have two-dimensional data and assume that the two dimensions are independent, it is sufficient to estimate the marginal distributions: the joint distribution will be their product. If we assume that a two-dimensional

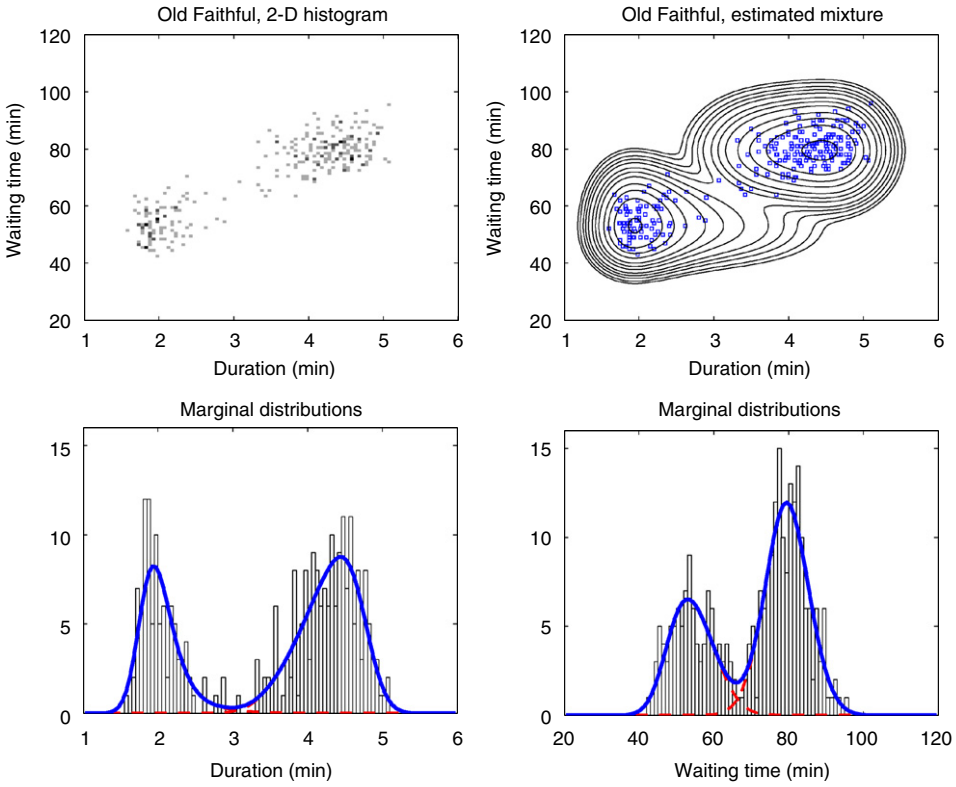

Fig. 5. Old Faithful data, modeled as a sum of two independent two-dimensional distributions with log-concave marginal densities.

distribution can be approximated well by a sum of independent log-concave components, our algorithm can be adapted with little extra work. We sketch the procedure for a mixture of two components.

Assume that approximations to the component distributions have been obtained. Then we can split the data, a two-dimensional histogram, proportionally into two sub-histograms (the E-step). The row sum and column sum of each sub-histogram are smoothed as one-dimensional histograms (the M-step). Their product gives an estimate of the mixture component. These steps are repeated until convergence.

Data and results for Old Faithful are shown in Fig. 5. The histogram consists of 100 by 100 bins. One value of λ (10^4) was used. In principle, a separate λ can be used for each mixture component for each marginal distribution, but optimization of AIC will be non-trivial. It is also probably too simple to take simply the sum of the four effective dimensions as the effective dimension. We propose this approach only as an explorative tool.

4. Discussion

Histogram smoothing with penalized likelihood is a generally useful tool. With a third-order penalty, the smoothed histogram is gently pushed toward a log-concave gaussian shape. This is of great value when fitting mixtures of non-parametric smooth distributions to data. The applications illustrate this claim.

We are investigating several extensions. In the case of the induration data, it might be useful to assume identical shapes of the components for men and women, perhaps independent of the BCG status, but with mixing proportions that change with BCG status, age and calendar time. This can be implemented by means of a two-phase EM algorithm, one phase for splitting and smoothing, the other for fitting mixing proportions.

In our applications the information in the data was strong enough to push the estimated mixture components in the right (log-concave) direction. One might encounter less favorable situations, where this is not the case and some outside help may be needed. Bollaerts et al. (2006) and Eilers (2005) describe the use of asymmetric penalties to enforce shape

constraints. In the present application this approach would allow to set an upper bound to second differences, effectively setting an upper limit to the standard deviations of the mixture components.

In our approach (approximate) log-concavity is achieved as a consequence to a third-order roughness penalty, so smoothness is an essential assumption. We think this is reasonable for most applications. It also gives pleasing curves. Walther (2002) has a more general aim, taking only log-concavity as a goal, without considering smoothness. He also proposes a test for determining whether a data distribution might be a mixture. We have a simpler objective: not testing, but data exploration.

Acknowledgment

We thank Sang Jae Kim of the Korean Institute of Tuberculosis in Seoul for providing us with the data.

References

Azzalini, A., Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. Appl. Statist. 39, 357–365.
Böhning, D., 2000. Computer-assisted Analysis of Mixtures and Applications. Chapman & Hall, CRC Press, London, Boca Raton.
Bollaerts, K., Eilers, P.H.C., van Mechelen, I.M., 2006. Simple and multiple P-splines regression with shape constraints. British J. Math. Statist. Psych., in press.
Eilers, P.H.C., 2005. Unimodal smoothing. J. Chemometrics 19, 317–328.
Everitt, B., Hand, D.J., 1981. Finite Mixture Distributions. Chapman & Hall, London.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.
Simonoff, J.S., 1983. A penalty function approach to smoothing large sparse contingency tables. Ann. Statist. 11, 208–218.
Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, Berlin.
Walther, G., 2002. Detecting the presence of mixing with multiscale maximum likelihood. J. Amer. Statist. Assoc. 97, 508–513.