Published in final edited form as:
Comput Stat Data Anal. 2011 January 1; 55(1): 457–465. doi:10.1016/j.csda.2010.05.013.

Generalized weighted likelihood density estimators with application to finite mixture of exponential family distributions

Tingting Zhan[a,*], **Inna Chevoneva**[b], and **Boris Iglewicz**[a]

[a]Department of Statistics, Temple University, Philadelphia, PA 19022

[b]Department of Biostatistics, Thomas Jefferson University, Philadelphia, PA 19107

Abstract

The family of weighted likelihood estimators largely overlaps with minimum divergence estimators. They are robust to data contaminations compared to MLE. We define the class of generalized weighted likelihood estimators (GWLE), provide its influence function and discuss the efficiency requirements. We introduce a new truncated cubic-inverse weight, which is both first and second order efficient and more robust than previously reported weights. We also discuss new ways of selecting the smoothing bandwidth and weighted starting values for the iterative algorithm. The advantage of the truncated cubic-inverse weight is illustrated in a simulation study of three-components normal mixtures model with large overlaps and heavy contaminations. A real data example is also provided.

Keywords

finite normal mixture; generalized weighted likelihood estimator; influence function; smoothing bandwidth; truncated cubic-inverse weight; weighted starting value

1. Introduction

Robust density estimation through the minimum divergence estimators (MDE) dates back to 1950s. The target divergences between the empirical and model distributions could be defined through the characteristic, moment generating, distribution or density functions. Those divergences used in early research include Kolmogorov-Smirnov, Wolfowitz, Cramér-von Mises and squared L_2 norm (Parr, 1981), where the robustness of corresponding MDE is achieved through boundedness of influence function (IF) at the cost of first order efficiency. Beran (1977, 1978) started the recent line of density-based disparities by addressing both robustness and full first order efficiency of minimum Hellinger distance estimator, as well as the non-uniform convergence of ε-influence curve to the influence function. Lindsay (1994); Basu and Lindsay (1994) addressed the inaccuracy of influence function as the first order approximation to the bias response curve, thus an estimator could both have the same IF as MLE, i.e. be first order efficient, and be robust at the cost of the second order efficiency (Rao,

*Corresponding author. Tel: +1 267 324 2212 zhantt@temple.edu (Tingting Zhan), I_Chervoneva@mail.jci.tju.edu (Inna Chevoneva), borisi@temple.edu (Boris Iglewicz).

1962). Specifically, given continuous model density f_θ and empirical density $\widehat{g}_n$, define the smoothed densities $f_\theta^*(t) = \int k(x;t,h)\, f_\theta(x)\, dx$ and $\widehat{g}_n^*(t) = \int k(x;t,h)\, \widehat{g}_n(x)\, dx$, where $k(x; t, h) = k((x - t)/h)$ is the common smoothing kernel. The choice of bandwidth h has been discussed in Markatou (2000) for the finite normal mixture model. The family of δ-divergences is defined as

$$D\left(f_\theta^*, \widehat{g}_n^*\right) = \int f_\theta^*(x)\, G\left(\delta_n^*(x)\right) dx \tag{1}$$

where $\boldsymbol{G}$ is a thrice-differentiable function on $[-1, \infty)$ with $\boldsymbol{G}(0) = 0$, and the smoothed Pearson residual δ_n^* is

$$\delta_n^*(x) = \widehat{g}_n^*(x) / f_\theta^*(x) - 1 \tag{2}$$

The family (1) includes the power class (Cressie and Read, 1984), blended-χ^2 and general negative exponential (Jeong and Sarkar, 2000) through the choice of function $\boldsymbol{G}$. For example, the symmetric-χ^2 divergence (Markatou et al., 1998), which is a particular case in the blended-χ^2, corresponds to $\boldsymbol{G}(\delta) = 2\delta^2/(\delta + 2)$; and the negative exponential divergence (Bhandari et al., 2006) corresponds to $\boldsymbol{G}(\delta) = e^{-\delta} - 1$. Other recent studies involve the density power divergence (Basu et al., 1998)

$$D_\beta(f_\theta, g) = \int \left[f_\theta^{1+\beta} - \left(1+\beta^{-1}\right) g f_\theta^{\beta} + \beta^{-1} g^{1+\beta} \right] dx, \quad \beta > 0 \tag{3}$$

including Kullback-Leibler divergence ($\beta \to 0$) and L_2 distance ($\beta = 1$). Minimum density power divergence estimators are not first order efficient when $\beta > 0$ because of the bounded influence function (Jones et al., 2001).

The rest of this paper is organized as follow. In section 2 we unify previously considered minimum divergence estimators as generalized weighted likelihood estimators (GWLE), provide their influence function and discuss the efficiency requirements. In section 3 we introduce GWLE with new truncated cubic-inverse weight. In section 4 we consider the density estimation of finite mixture of exponential family distributions and describe a simple iterative algorithm for computing such GWLE. Numerical studies are carried out on simulations of three-components normal mixtures model featuring high overlaps and heavy contaminations. A real data example is also provided.

2. The generalized weighted likelihood estimator

Let $\widehat{G}_n$ be the empirical distribution of the random sample $X_1, \ldots, X_n$ and $\mathcal{F}_\Theta = \{F_\theta : \theta \in \Theta\}$ be the model family. The generalized weighted likelihood estimator (GWLE) $T\left(\widehat{G}_n\right)$, introduced here, is the solution of estimating equation $\Sigma_i\, \tilde{\psi}(x_i, \theta) = 0$ with

$$\tilde{\psi}(x, \theta) = w\left(x; F_\theta, \widehat{G}_n\right) \boldsymbol{u}_\theta(x) - \boldsymbol{a}(\theta) \tag{4}$$

where $\boldsymbol{u}_\theta = \nabla \ln f_\theta$ is the vector of score functions and $\boldsymbol{a}(\theta) = E_{F_\theta}(w(x)\boldsymbol{u}_\theta(x))$ is the bias adjustment term to ensure Fisher consistency. The weight $w(x_i; F_\theta, \widehat{G}_n)$ serves as an adjustment to potential contamination at x_i. For example, MLE is the solution of $\tilde{\psi}(x, \theta) = \boldsymbol{u}_\theta(x)$, which

assigns weight $w(x) = 1$ to all observations. However, the estimating function $\tilde{\psi}(x, \theta)$ does not have to satisfy the definition of M-estimators, for which the estimating function $\psi(x_i, \theta)$ contains only the observation x_i and parameter θ (Huber, 1981). Instead, $\tilde{\psi}(x, \theta)$ may contain the whole random sample through the weight $w(x; F_\theta, \widehat{G}_n)$. Therefore the asymptotic properties of GWLE do not immediately follow from the general theory of M-estimators. The GWLE class also includes the minimum density power divergence estimator (Basu et al., 1998), which is an M-estimator with

$$w(x, F_\theta) = f_\theta^\beta(x), \quad \boldsymbol{a}(\theta) = \int \boldsymbol{u}_\theta(x) f_\theta^{1+\beta}(x)\, dx; \tag{5}$$

The estimating equation based on the derivative of δ-disparity (1) takes form of the integral $\int w(\delta_n^*)\left(\nabla \ln f_\theta^*\right)\widehat{g}_n^* dx = 0$. Substitute f_θ^* and $\widehat{g}_n^*$ by f_θ and $\widehat{g}_n$, while keep $w(\delta_n^*)$ intact (Basu and Lindsay, 2004), we end up with GWLE with $w\left(x; F_\theta, \widehat{G}_n\right) = w(\delta_n^*(x))$ and $\boldsymbol{a}(\theta) = 0$. Specifically, the weight of minimum symmetric-χ^2 divergence estimator is (Markatou et al., 1998)

$$w(\delta_n^*(x)) = 1 - \delta_n^{*2}/(\delta_n^* + 2)^2 \tag{6}$$

and the weight for minimum negative exponential divergence estimator (Bhandari et al., 2006) is

$$w(\delta_n^*(x)) = \left(e - (2+\delta_n^*) e^{-\delta_n^*}\right) / (\delta_n^* + 1) \tag{7}$$

These weights discussed above either depend on f_θ as in (5) or on δ_n^* as in (6); (7). The first type down-weighs low density points, i.e. observations with small $f_\theta(x_i)$. The second type distinguishes outliers (where $\delta_n^* > 0$) and inliers (where $\delta_n^* \in (-1, 0]$) and is re-scaled so that $w(\delta_n^*)|_{0=1}$ for sample points concordant with the model, which also serves as a reference for comparison. Figure 1(a) shows weights (6); (7) together the weight of MLE and minimum Hellinger distance estimator $w(\delta^*) = \left(2\sqrt{1+\delta^*} - 1\right) / (\delta^* + 1)$. It is interesting that the negative exponential weight (7) assigns weights larger than 1 for inliers, knowing that this estimator is second order efficient and robust to both outliers and inliers (Lindsay, 1994).

Next, we develop the influence function and efficiency requirements of GWLE estimators $T(\cdot)$ with weight $w\left(x_i; F_\theta, \widehat{G}_n\right) = w(x_i; \delta_n^*, f_\theta)$, which allow down weighing simultaneously the outliers, inliers and low-density points.

Theorem 1

Let G be the true distribution and F_θ be the model distribution. $T(\cdot)$ *is the GWLE with weights* $w(x, \delta^*, f_\theta)$ *defined through smoothing kernel* $k(x; t, h)$. *Let* $\theta = T(G)$, *we have*

- **(i).** Under mild regularity conditions (Basu and Lindsay, 1994), the sequence of estimators $T(\widehat{G}_n)$ *exists.*
- **(ii).** *The influence function of T is* $IF(x_0, G, F_\theta) = T_h'(x_0, G) = -DEN_{h,g}^{-1} NUM_{h,g,x_0}$,

$$NUM_{h,g,x_0} = \int \frac{k(x; x_0, h) - g^*(x)}{f_\theta^*(x)} \cdot w_\delta' \boldsymbol{u}_\theta (g - f_\theta)\, dx + w(x_0, \delta^*, f_\theta)\, \boldsymbol{u}_\theta(x_0) - \int w(x, \delta^*, f_\theta)\, \boldsymbol{u}_\theta\, g\, dx \tag{8}$$

$$DEN_{h,g}=\int\left[(\delta^*+1)\,w'_\delta \boldsymbol{u}^*_\theta - f_\theta w'_f \boldsymbol{u}_\theta\right]\boldsymbol{u}^t_\theta\,(f_\theta - g)\,dx+\int w\,(x,\delta^*,f_\theta)\,\nabla \boldsymbol{u}^t_\theta\,g\,dx-\int w\,(x,\delta^*,f_\theta)\,\nabla^2 f_\theta\,dx \tag{9}$$

where $\boldsymbol{u}^*_\theta(x)=\partial\log f^*_n(x)/\partial\theta$. *Notation* $(\cdot)^t$ denotes vector transposition. The partial derivatives are $w'_\delta=\partial w\,(x,\delta^*,f_\theta)/\partial\delta^*$ *and* $w'_f=\partial w\,(x,\delta^*,f_\theta)/\partial f_\theta$.

(iii). *The asymptotic distribution of* $\sqrt{n}\left(T\left(\widehat{G}_n\right)-T(G)\right)$ *is multivariate normal with mean* 0 *and covariance matrix* $\mathrm{V}_G=\int T'_h(x,G)\left[T'_h(x,G)\right]^t dG(x)$, *and*

$$\widehat{V}_{\widehat{G}_n}=DEN^{-1}_{h,\widehat{g}_n}\left(\frac{1}{n}\sum_{i=1}^{n}NUM_{h,\widehat{g}_n,x_i}NUM^t_{h,\widehat{g}_n,x_i}\right)\left(DEN^{-1}_{h,\widehat{g}_n}\right)^t \tag{10}$$

Proof. See Appendix.

Here we give the sufficient conditions for efficiency of GWLE. The GWLE $T(\cdot)$ with weight $w(x_i;\,\delta^*_n,\,f_\theta)$ is first order efficient if $w'_\delta(\delta^*,f_\theta)\,|_{\delta^*=0}=0$, and is second order efficient (Rao, 1962) if $w''_\delta(\delta^*,f_\theta)\,|_{\delta^*=0}=0$, where w''_δ is the second order derivative. The proof is a quick derivation from Remark E of Lindsay (1994), where the author listed the efficiency requirements on residual adjustment function. As there exists a one-to-one correspondence between residual adjustment function and weight function, the equivalent requirement on weight function is obtained. This implies that an estimator which is second order efficient assigns weights larger than 1 to inliers, as does the negative exponential weight (7) shown in Figure 1.

3. The truncated cubic-inverse weight

In this section we introduce the new truncated cubic-inverse weight, which both satisfies the requirements in section 2 and has better empirical robust properties than the previous weights (5), (6) and (7) as shown in section 4.2. Define the *inverse* weight

$$w_1(\delta^*_n(x))=(\delta^*_n(x)+1)^{-1}=f^*_\theta(x)/\widehat{g}^*_n(x) \tag{11}$$

A heuristic explanation of the advantage of inverse weight is given below. Consider the estimating function

$$\int_x w_1(\delta^*_n(x))\,\tilde{\boldsymbol{u}}_\theta(x)\,\widehat{g}^*_n(x)\,dx=\int\tilde{\boldsymbol{u}}_\theta(x)\,f^*_\theta(x)\,dx \tag{12}$$

where $\tilde{\boldsymbol{u}}_\theta(x)=\int k(t;x,h)\,\boldsymbol{u}_\theta(t)\,dt$. On the other hand, let estimator *MLE** be the solution of estimating equation $\Sigma_i\boldsymbol{u}^*_\theta(x_i)=\int\boldsymbol{u}^*_\theta(x)\,d\widehat{G}_n(x)=0$, where $\boldsymbol{u}^*_\theta(x)=\int k(t;x,h)\,\tilde{\boldsymbol{u}}_\theta(t)\,dt$ (Basu and Lindsay, 1994). For any random sample $\widehat{g}_n$ coming from underlying density $g=f_\theta$, the asymptotic limit of *MLE** estimating equations is (12),

$$\lim_{n\to\infty}\int_x\boldsymbol{u}^*_\theta(x)\,d\widehat{G}_n(x)=\int\boldsymbol{u}^*_\theta(x)\,dF_\theta(x)=\int\tilde{\boldsymbol{u}}_\theta(x)\,f^*_\theta(x)\,dx \tag{13}$$

Given the full efficiency of *MLE** under transparent kernels, such as a normal kernel $k(x; t, h)$ for normal model (Basu and Lindsay, 1994), the estimator as solution to (12) is also fully efficient. The estimating equation (12) is simplified by removing the smoothing kernel on score function $\boldsymbol{u}_\theta$ and $\widehat{g}_n$ (Basu and Lindsay, 2004),

$$\int_x w_1(\delta_n^*(x))\,\tilde{\boldsymbol{u}}_\theta(x)\,\widehat{g}_n^*(x)\,dx \approx \int_x w_1(\delta_n^*(x))\,\boldsymbol{u}_\theta(x)\,\widehat{g}_n(x)\,dx = 0 \tag{14}$$

which is equivalent to WLE with the inverse weight w_1 in (11).

However, the simplification (14) no longer keeps the efficiency of the estimator. We restore the efficiency by replacing the inverse weight w_1 by a cubic curve at the neighborhood of $\delta^* = 0$ in order to satisfy the efficiency requirements in section 2. Define the *cubic-inverse* weight

$$w_2(\delta_n^*(x), c, x_+, x_-) = \begin{cases} w_1(\delta_n^*(x)) & \text{if } \delta_n^*(x) \notin [x_-, x_+] \\ c[\delta_n^*(x)]^3 + 1 & \text{if } \delta_n^*(x) \in [x_-, x_+] \end{cases} \tag{15}$$

Any one of the three parameters (c, x_+, x_-) determines the other two by solving the continuity equations at the positive real root $\delta^* = x_+$. The negative real root x_- may not exist, in which case we let $w_2 = w_1$ when $\delta^* \geq x_+$ and be the cubic curve when $\delta^* < x_+$. The cubic-inverse weight w_2 satisfies the efficiency requirements in section 2. Figure 1(b) gives some examples of weight w_2 with different (c, x_+, x_-), in which the solid line represents the inverse weight w_1. The selection of cubic coefficient c is discuss in section 3.2.

3.1. Selection of smoothing bandwidth h

Selection of bandwidth h in smoothing kernel $k(x; t, h)$ plays an important role in determining the cubic coefficient c and the right truncation point. Let g be an arbitrary continuous density and $X_1, \cdots, X_n$ be a random sample with empirical density $\widehat{g}_n$. Define the nonparametric Pearson residual $\tilde{\delta}_n^*$ calculated by smoothed empirical densities at a narrow bandwidth h_1 and a wide bandwidth $h_2 = 2h_1$,

$$\tilde{\delta}_n^*(x) = \widehat{g}_1^*(x) / \widehat{g}_2^*(x) - 1 \tag{16}$$

where $\widehat{g}_i^*(x) = \int k(x; t, h_i)\,\widehat{g}_n(t)\,dt$ for $i = 1, 2$. Our criterion of choosing $h = h_1$ is that the distribution of $\tilde{\delta}_n^*(x)$ does not depend heavily on density g nor sample size n, since the subsequent choice of cubic coefficient c and truncation threshold l_2 will be decided by the approximated distribution of $\tilde{\delta}_n^*(x)$.

It has been suggested that for k-component normal mixtures one may use $h^2 = \kappa \Sigma_{i=1}^k \widehat{p}_i \widehat{\sigma}_i^2$ iteratively, where k takes values roughly in range (.001, .05) based on the criterion of average down-weight (Markatou, 2000). However, this choice of bandwidth does not satisfy our criterion; as it's difficult to generalize it to an arbitrary density g, and unrealistic to use the same bandwidth for both large and small sample size n. Therefore, we propose a new choice of bandwidth

$$h = \mathrm{MAD}(x) / \sqrt{n} \tag{17}$$

where MAD stands for median-absolute-deviation. We empirically compare the bandwidth (17) with the standard R (R Development Core Team, 2008) functions of bandwidth selection for Gaussian kernels bw.∗(), where ∗ may be replaced by nrd0 or nrd representing the "rule-of-thumb" choice (Silverman, 1986; Scott, 1992); ucv or bcv representing unbiased or biased cross-validation (Scott and Terrell, 1987); and SJ representing the method using pilot estimation of derivatives (Sheather and Jones, 1991). All these choices contain the sample size n at different scales of power.

In Figure 2, the smoothed Pearson residuals $\tilde{\delta}_n^*$ for various densities g's are calculated and box-plotted under sample size n = 500 and 100, with y-axis range (−0.6, 0.6) to magnify the boxes and exclude the points outside the whiskers. The six sets of box-plots, from left to right, are from density g consisting of Unif(0, 1), Beta(5, 3), $N(0, 1)$, t_4, $F_{7,10}$ and our finite normal mixture density in Figure 3(a), respectively. Within each set of box-plots, the five parallel boxes are, from left to right, using bandwidth (17), nrd, ucv, bcv and SJ, respectively.

The bandwidth (17), which is the first box-plot of each set, best satisfies our criterion that the distribution of $\tilde{\delta}_n^*$ is closer to symmetry around 0 and stays roughly the same across different models and sample sizes. The other choices, however, have varying performances under our criterion for different sample sizes, and are overall not as good as bandwidth (17). Thus we suggest that bandwidth (17) is preferred under our criterion regardless of the underlying distribution and sample size. Nonetheless, the choice of h is still wide open for different weights, and researchers may choose other h for their own smoothing problems.

3.2. Selection of cubic coefficient c and truncation on the right tail

The cubic coefficient c is chosen such that within the inter-quartile range of the smoothed Pearson residuals (16), the cubic-inverse weight $w_2\left(\tilde{\delta}_n^*\right)$ falls in the interval $1 \pm \Delta w$. Figure 2 shows that when using $h = \mathrm{MAD}(x)/\sqrt{n}$, the inter-quartile range of nonparametric Pearson residuals $\tilde{\delta}_n^*$ is safely covered by the interval (−0.2, 0.2); thus the cubic coefficient c is determined by solving $0.2^3|c| \le \Delta w$. Let $\Delta w = 0.1$, then $|c| \le 12.5$. On the other hand, we place another restraint that $|c| \ge 8$ to avoid too much cubic modification. As shown in Figure 1(b), different c values in the range (−12.5, −8) hardly affect the shape of cubic-inverse curve, thus we may pick our choice as we like. The weight w_3 with $c = -8$ is added to Figure 1(a) for comparison.

A refinement to the cubic-inverse weight w_2 is to truncate the weight when δ* exceeds a threshold l_2, since the smoothed Pearson residuals for most continuous densities are right-skewed with a very heavy tail, which is not shown in Figure 2. In order to further reduce the influence of extreme outliers, we iteratively assign weight zero to sample points with Pearson residuals $\tilde{\delta}_n^*$ greater than a threshold l_2. Define the *truncated cubic-inverse* weight

$$w_3(x) = w_2\left(\delta_n^*(x)\right) \cdot \boldsymbol{I}_{\{\delta_n^*(x) < l_2\}} \tag{18}$$

One possible choice is to let l_2 be the 95% percentile of $\tilde{\delta}_n^*$ calculated at each iteration step.

4. Application to finite mixture of exponential family distributions

4.1. Iterative algorithm for solving GWLE

In this section we briefly outline the steps of obtaining the truncated cubic-inverse weight (18) and provide an iteration algorithm for solving GWLE (4) under the finite mixture model of exponential family distributions

$$f(x_i;\xi,\boldsymbol{p}) = \sum_{s=1}^{k} p_s \phi(x;\xi_s) \tag{19}$$

where $\boldsymbol{\xi}_s = (\xi_{s1}, \cdots, \xi_{sQ})^t$ are the canonical parameters and $\boldsymbol{\eta}_s = (\eta_{s1}, \cdots, \eta_{sQ})^t$ the mean parameters of the sth component. Let $\boldsymbol{T}_X = (\boldsymbol{T}_{X,1}, \cdots, \boldsymbol{T}_{X,Q})^t$ be the sufficient statistics of X. In this work, we do not step away to the discussion of the selection of mixture component, which itself constitutes a major topic of interest in this field (see Turner and West, 1993;Roeder, 1994;West, 1997;Richardson and Green, 1997;Stephens, 2000;Ishwaran et al., 2001;Ishwaran and James, 2002). Instead, we assume that the number of component k is fixed and known in our problem.

In this section we discuss the selection of starting value, smoothing bandwidth h, iteration steps and convergence criterion for estimating the finite mixture model (19) based on random sample $X_1, \cdots, X_n$. First of all, the choice of starting value is critical for most iterative algorithms. We obtain a tentative partition of the data into k groups through one of the existing methods such as k-means and trimmed k-means (Cuesta-Albertos et al., 1997), robust clustering (Woodward et al., 1984) or the watershed algorithm (Vincent and Soille, 1991). All of these partitions have been developed under certain robustness considerations and produce their own corresponding starting values. However, for our iterative algorithm with updating steps (20) and (21), we can greatly reduce the number of iteration if we use these partitions to produce our "weighted starting values", which is described in detail below. Within each group, we calculate the nonparametric smoothed Pearson residual (16) and the corresponding weight $\tilde{w}_i = w\left(\tilde{\delta}_n^*(x_i)\right)$. We use weight $\tilde{w}_i$ to obtain the starting values $\boldsymbol{p}^{(0)}$, as the weighted proportion of the sample size in each group $p_s^{(0)} = \Sigma_{x_i \in \mathbb{S}} \tilde{w}_i / \Sigma \tilde{w}_i$, where $\mathbb{S}$ represents the set of random sample in group s, and the weighted moments of the random sample within each group. As every exponential family distribution has a one-to-one correspondence to its first few moments, we could get the starting values $\xi_s^{(0)}$ through simple transformation of these weighted moments. Specifically, when the model (19) is a mixture of normals, the starting μ_0's are the weighted medians and σ_0's are the weighted median-absolute-deviation (MAD)'s of each group.

The iteration steps of solving model (19) is given below,

1. Let $\boldsymbol{\theta}^{(d-1)} = (p^{(d-1)}, \boldsymbol{\eta}^{(d-1)})$ or $\boldsymbol{\theta}^{(d-1)} = (p^{(d-1)}, \boldsymbol{\xi}^{(d-1)})$ be the estimates from $(d-1)$th iteration.
2. Calculate Pearson residual δ_n^* with $\boldsymbol{\theta}^{(d-1)}$ as in (1).
3. Let $w^{(d-1)}(x_i) = w(x_i, \delta_n^*, f_{\theta^{(d-1)}})$. Specifically for truncated cubic-inverse weight (18), choose c and l_2 as suggested in section 3.2. Let $p_w^{(d-1)}(s;x_i) = w^{(d-1)}(x_i)\, p_s \phi\left(x_i;\xi_s^{(d-1)}\right) / f\left(x_i;\theta^{(d-1)}\right)$, $s = 1, \cdots, k$.
4. Obtain the bias adjustment term $\boldsymbol{a}(\boldsymbol{\xi}^{(d-1)}, \boldsymbol{p}^{(d-1)})$ by numeric integration. Let $\boldsymbol{a}_{:\xi sq}$ and $\boldsymbol{a}_{:ps}$ be the elements of a corresponding to ξ_{sq} and p_s.

5. Update p_s and η_{sq}, $s = 1, \cdots, k$, $q = 1, \cdots, Q$, by

$$p_s^{(d)} = \frac{1}{n}\sum_{i=1}^{n} p_w^{(d-1)}(s; x_i) - p_s \boldsymbol{a}\left(\xi^{(d-1)}, \boldsymbol{p}^{(d-1)}\right)_{:p_s} \tag{20}$$

$$\eta_{sq}^{(d)} = \left[\sum_{i=1}^{n} p_w^{(d-1)}(s; x_i)\right]^{-1} \left(\sum_{i=1}^{n} \boldsymbol{T}_{x_i,q} p_w^{(d-1)}(s; x_i) - n\boldsymbol{a}\left(\xi^{(d-1)}, \boldsymbol{p}^{(d-1)}\right)_{:\xi_{sq}}\right) \tag{21}$$

The updating steps for finite normal mixtures model are included in this algorithm with $\eta_s = \left(\mu_s, \mu_s^2 + \sigma_s^2\right)^t$ and $\boldsymbol{T}_x = (x, x^2)^t$; this special case together with density power weight (5) is discussed in Fujisawa and Eguchi (2006).

The convergence of the series of estimates $\boldsymbol{\theta}$'s is equivalent to the convergence of the series of weights w, since they depend on each other iteratively, i.e. $\boldsymbol{\theta}^{(n-1)} \Rightarrow w^{(n)} \Rightarrow \boldsymbol{\theta}^{(n)} \Rightarrow w^{(n+1)} \Rightarrow \boldsymbol{\theta}^{(n+1)}$. The convergence criterion based on weights w rather than estimates $\boldsymbol{\theta}$ is more reliable and not specific to particular model. Let $w^{\text{new}}(x) = w\left(x, \delta_n^*, f_\theta | \theta^{\text{new}}\right)$. Similar to R^2 in regression, the criterion is set to be

$$\frac{n^{-1} \| w^{\text{new}} - w \|^2}{\text{var}(w^{\text{new}})} < \varepsilon \tag{22}$$

where function var() calculates the sample variance. We use $\varepsilon = 1\%$ in the simulation studies.

4.2. Simulation results

The simulation study is carried out on a scenario of three components normal mixtures sketched in Figure 3 with sample size $n = 400$. We generate 1000 data sets from each scenarios: a clean density (a) .2N(−10, 3) + .5N(0, 5) + .3N(15, 4); and a contaminated density (a1) where outliers of 5% are added at N(−18, 4) and the third component is replaced by a shifted and re-scaled F-distribution with the same mode as the original normal density. The overlaps (Woodward et al., 1984) between adjacent components in scenario (a), the areas highlighted, are 7.2% and 3.7%, respectively.

Estimators compared are MLE and GWLEs with weights of density power (5), symmetric χ^2 (6), negative exponential (7), cubic-inverse (15) and truncated cubic-inverse (18). The starting partitions are obtained from robust clustering (Woodward et al., 1984). The tuning parameter for density power weight (5) is selected from $\beta = .15, .20, .25, .30$ by minimizing Cramer-von Mises divergence through cross-validation (Fujisawa and Eguchi, 2006). The kernel smoothing bandwidth is $h = \text{MAD}(X) / \sqrt{n}$. The truncated cubic-inverse weight (18) has the cubic coefficient $c = -8$. The convergence criterion for MLE is a set of pre-specified thresholds on the L_2 norms of $\|p^{\text{new}} - p\|^2$, $\|\mu^{\text{new}} - \mu\|^2$ and $\|\sigma^{\text{new}} - \sigma\|^2$; while for all GWLE estimators we use criterion (22). Figure 4 shows boxplots of errors of different estimators for μ's, σ's and p's, where the scale of p's are multiplied by a factor of 10 in order to enlarge the boxplots. The truncated cubic-inverse weight (18) generally retains efficiency under the uncontaminated scenario (lower panel of Figure 4) and provides the overall best estimation under the contaminated scenario (upper panel of Figure 4), in terms of smaller bias and mean squared error. This advantage is especially obvious when estimating the scale parameters σ_1 and σ_3, where either inliers or outliers are present.

4.3. Tendon fibrilogenesis data

The case study described in this section comes from the tendon collagen fibrilogenesis experiment (Zhang et al., 2006). From the two mice strains, decorin deficient (DD) and wild type (WT), the fibril cross-sections are made, photographed under the microscope and the fibril diameters are measured. Typically, slices of mature fibril (2 month or older) have around 50–150 diameter measurements on each field. The top row of Figure 5 show histograms of fibril diameters on selected microscopic fields, which are modeled by finite normal mixtures which provide insight into the mechanisms of collagen fibrillogenesis. It was suggested (Zhang et al., 2006) that three components were appropriate for 2 month or older fibrils. The contaminations, such as heavy tails shown in histograms, are either due to genetic alterations (e.g. abnormally large fused fibrils) or cross sections through the tapered ends of fibrils. Robust GWLE estimates with all weights are calculated, among which the negative exponential (7) and truncated cubic-inverse (18) are plotted in the bottom rows of Figure 5. The first selected field shows a scenario with similar estimates from all weights, while the second and third fields show that truncated cubic-inverse (and cubic-inverse) weights are less affected by the small cluster outliers appearing to the left of the data.

5. Concluding remarks

In this paper, we proposed the truncated cubic-inver weight for the class of generalized weighted likelihood density estimators, which retains the first and second order efficiency and is more robust than previously reported weights (Fujisawa and Eguchi, 2006; Markatou et al., 1998; Bhandari et al., 2006) under heavy data contamination. We also proposed an iterative algorithm for solving GWLE under the model family of finite exponential-family distribution mixtures, with the new weighted starting values. This approach can be generated to multivariate density estimation.

Acknowledgments

The research reported in this paper was partially supported by NIH/NIAMS grant NO.AR054596.

6. Appendix

Proof of Theorem 1

Given a fixed distribution G, contaminated $G_{\varepsilon,x_0} = (1-\varepsilon)G+\varepsilon\Delta(x_0)$ and model family $\mathcal{F}_\Theta$, the estimating equation (4) for GWLE $T(\cdot)$ with weight $w(x, \delta^*_\varepsilon, f_\theta)$ is equivalent to

$$\int w(x)\,\boldsymbol{u}_\theta(x)\cdot\left(g_{\varepsilon,x_0}(x)-f_\theta(x)\right)dx=0 \tag{23}$$

where $\theta = T(G_{\varepsilon,x_0})$ and $k(x; t, h)$-smoothed Pearson residual $\delta^*_\varepsilon(x)=g^*_\varepsilon(x)/f^*_\theta(x)-1$. The general theory about influence function of M-estimator (Hampel et al., 1985) is not applicable (section 2). Here let $(\cdot)^t$ denote transposition of vectors and take the derivative of equation (23) with respect to ε,

$$\int\left(\frac{\partial\delta^*_\varepsilon}{\partial\varepsilon}\frac{\partial w(\delta^*,f_\theta)}{\partial\delta^*_\varepsilon}+\frac{\partial\theta^t}{\partial\varepsilon}\frac{\partial f_\theta}{\partial\theta}\frac{\partial w(\delta^*,f_\theta)}{\partial f_\theta}\right)\boldsymbol{u}^t_\theta\left(g_{\varepsilon,x_0}-f_\theta\right)dx+\int w(\delta^*,f_\theta)\frac{\partial\theta^t}{\partial\theta}\frac{\partial\boldsymbol{u}^t_\theta}{\partial\theta}\left(g_{\varepsilon,x_0}-f_\theta\right)dx+\int w(\delta^*,f_\theta)\,\boldsymbol{u}^t_\theta\left((\chi_{x_0}-g)-\frac{\partial\theta^t}{\partial\varepsilon}\frac{\partial f_\theta}{\partial\theta}\right)dx=0 \tag{24}$$

where $\partial\theta/\partial\varepsilon|_{\varepsilon=0} = \mathrm{IF}(x_0)$, $\partial g^*_{\varepsilon,x_0}(x)/\partial\varepsilon = k(x;x_0,h) - g^*(x)$ and $\boldsymbol{u}^*_\theta(x) = \left[f^*_\theta(x)\right]^{-1}\partial f^*_\theta(x)/\partial\theta$.

$$\left.\frac{\partial\delta^*}{\partial\varepsilon}\right|_{\varepsilon=0} = \frac{k(x;x_0,h) - g^*(x)}{f^*_\theta(x)} - \frac{g^*(x)}{f^*_\theta(x)} \cdot \mathbf{IF}^t \cdot \boldsymbol{u}^*_\theta(x) \tag{25}$$

Substitute (25) into (24) and evaluate at ε = 0,

$$\begin{aligned}
&\int\left(\frac{k(x;x_0,h)-g^*(x)}{f^*_\theta(x)} - (\delta^*+1)\,\mathbf{IF}^t\boldsymbol{u}^*_\theta\right)w'_\delta\boldsymbol{u}^t_\theta\cdot(g-f_\theta)\,dx\\
&\quad+\int\mathrm{IF}^t f_\theta w'_f\boldsymbol{u}_\theta\boldsymbol{u}^t_\theta\cdot(g-f_\theta)\,dx\\
&\quad+\int\mathrm{IF}^t w\frac{\partial\boldsymbol{u}^t_\theta}{\partial\theta}\cdot(g-f_\theta)\,dx\\
&\quad+w(x_0)\,\boldsymbol{u}^t_\theta(x_0)\\
&\quad-\int w\boldsymbol{u}^t_\theta g\,dx\\
&\quad-\int\mathrm{IF}^t f_\theta w\boldsymbol{u}_\theta\boldsymbol{u}^t_\theta dx = 0
\end{aligned} \tag{26}$$

and substitute $f_\theta\cdot\left(\partial\boldsymbol{u}^t_\theta/\partial\theta + \boldsymbol{u}_\theta\boldsymbol{u}^t_\theta\right) = \partial(f_\theta\boldsymbol{u}_\theta)^t/\partial\theta = \partial^2 f_\theta/\partial\theta^2$ into (26), one could solve for influence function (8); (9). Note this is influence function of the WLE from Eq (4), which is an approximation of minimum disparity estimator of Lindsay (1994); for influence function of the latter, refer to Basu and Lindsay (1994). The proof is applicable to empirical distribution $\widehat{G}_n$ and contaminated $\widehat{G}_{n,\varepsilon,x}$.

References

Basu A, Harris IR, Hjort NL, Jones MC. Robust and efficient estimation by minimising a density power divergence. Biometrika 1998;85(3):549–559.

Basu A, Lindsay BG. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. Annals of the Institute of Statistical Mathematics 1994;46(4):683–705.

Basu A, Lindsay BG. The iteratively reweighted estimating equation in minimum distance problems. Comput. Statist. Data Anal 2004;45(2):105–124.

Beran R. Minimum Hellinger distance estimates for parametric models. The Annals of Statistics 1977;5 (3):445–463.

Beran R. An efficient and robust adaptive estimator of location. The Annals of Statistics 1978;6(2):292–313.

Bhandari SK, Basu A, Sarkar S. Robust inference in parametric models using the family of generalized negative exponential dispatches. Australian & New Zealand Journal of Statistics 2006;48(1):95–114.

Cressie N, Read TRC. Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society 1984;46 (3):440–464.Series B. Methodological

Cuesta-Albertos JA, Gordaliza A, Matrán C. Trimmed k-means: an attempt to robustify quantizers. Ann. Statist 1997;25(2):553–576.

Fujisawa H, Eguchi S. Robust estimation in the normal mixture model. J. Statist. Plann. Inference 2006;136(11):3989–4011.

Hampel, FR.; Ronchetti, EM.; Rousseeuw, PJ.; Stahel, WA. Robust Statistics: The Approach Based on Influence Functions. 1985. Wiley Series in Probability and Statistics

Huber, PJ. Robust statistics. John Wiley & Sons Inc.; New York: 1981. Wiley Series in Probability and Mathematical Statistics

Ishwaran H, James LF. Approximate dirichlet process computing in finite normal mixtures: Smoothing and prior information. Journal of Computational and Graphical Statistics 2002;11(3):508–532.

Ishwaran H, James LF, Sun J. Bayesian model selection in finite mixtures by marginal density decompositions. Journal of the American Statistical Association 2001;96(456):1316–1332.

Jeong D-B, Sarkar S. Negative exponential disparity based family of goodness-of-fit tests for multinomial models. J. Statist. Comput. Simulation 2000;65(1):43–61.

Jones MC, Hjort NL, Harris IR, Basu A. A comparison of related density-based minimum divergence estimators. Biometrika 2001;88(3):865–873.

Lindsay BG. Efficiency versus robustness: the case for minimum Hellinger distance and related methods. The Annals of Statistics 1994;22(2):1081–1114.

Markatou M. Mixture models, robustness, and the weighted likelihood methodology. Biometrics 2000;56 (2):483–486. [PubMed: 10877307]

Markatou M, Basu A, Lindsay BG. Weighted likelihood equations with bootstrap root search. J. Amer. Statist. Assoc 1998;93(442):740–750.

Parr WC. Minimum distance estimation: A bibliography. Communications in Staitstics - Theory and Methods 1981;10(12):1205–1224.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2008. ISBN 3-900051-07-0

Rao CR. Efficient estimates and optimum inference procedures in large samples. Journal of the Royal Statistical Society. Series B (Methodological) 1962;24(1):46–72.

Richardson S, Green PJ. On bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society. Series B (Methodological) 1997;59(4):731–792.

Roeder K. A graphical technique for determining the number of components in a mixture of normals. Journal of the American Statistical Association 1994;89(426):487–495.

Scott, DW. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley; 1992.

Scott DW, Terrell GR. Biased and unbiased cross-validation in density estimation. Journal of the American Statistical Association 1987;82(400):1131–1146.

Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological) 1991;53(3):683–690.

Silverman, BW. Density Estimation. Chapman and Hall; London: 1986.

Stephens M. Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. The Annals of Statistics 2000;28(1):40–74.

Turner DA, West M. Bayesian analysis of mixtures applied to post-synaptic potential fluctuations. Journal of Neuroscience Methods 1993;47(1–2):1–21. [PubMed: 8321009]

Vincent L, Soille P. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. IEEE PAMI 1991;13(6):583–598.

West M. Hierarchical mixture models in neurological transmission analysis. Journal of the American Statistical Association 1997;92(438):587–606.

Woodward WA, Parr WC, Schucany WR, Lindsey H. A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. Journal of the American Statistical Association 1984;79(387):590–598.

Zhang G, Ezura Y, Chervoneva I, Robinson PS, Beason DP, Carine ET, Soslowsky LJ, Iozzo RV, Birk1 DE. Decorin regulates assembly of collagen fibrils and acquisition of biomechanical properties during tendon development. Journal of Cellular Biochemistry 2006;98:1436–1449. [PubMed: 16518859]

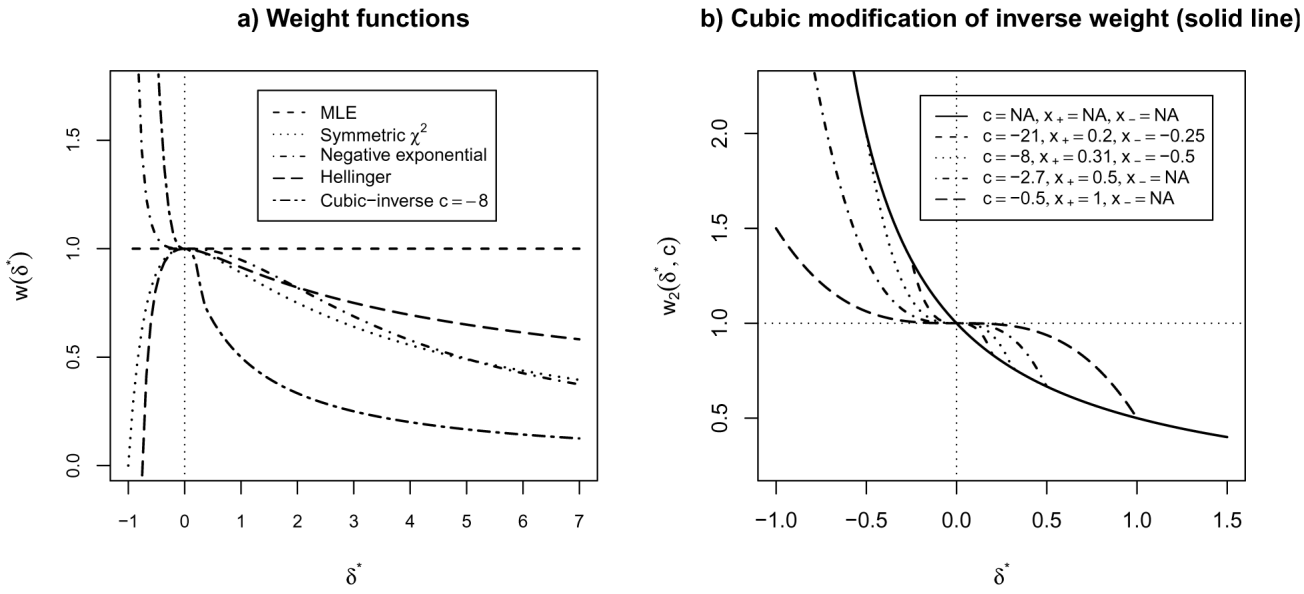

Figure 1.
Weight functions depending on smoothed Pearson residual δ_n^*

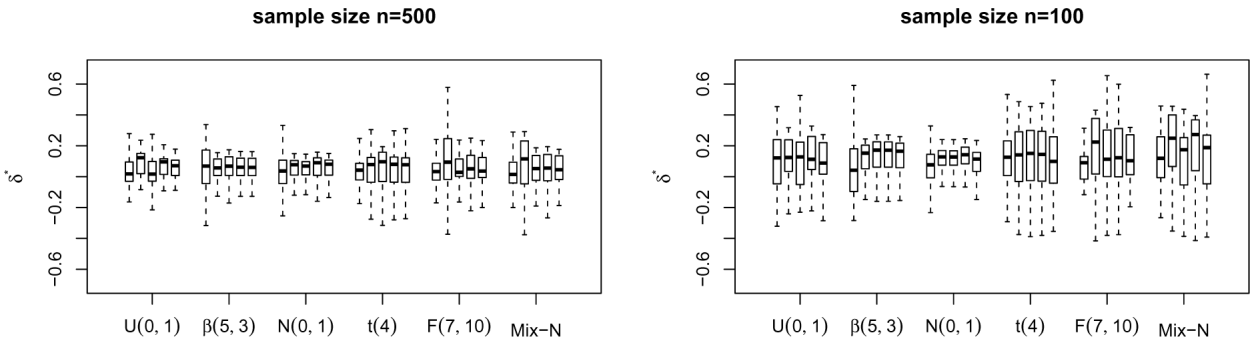

Figure 2.
Distribution of δ* for 6 generating distributions. For each distribution the 5 boxplots represent smoothing bandwidths *h* for (17), nrd, ucv, bcv, and SJ.

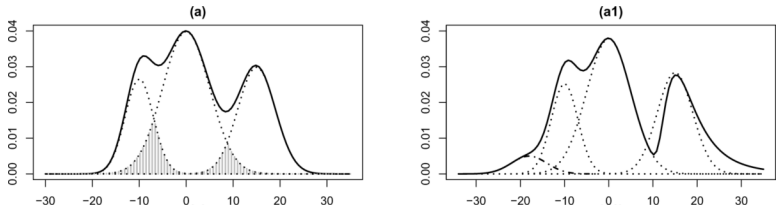

Figure 3.
(a) clean data; (a1) contaminated data.

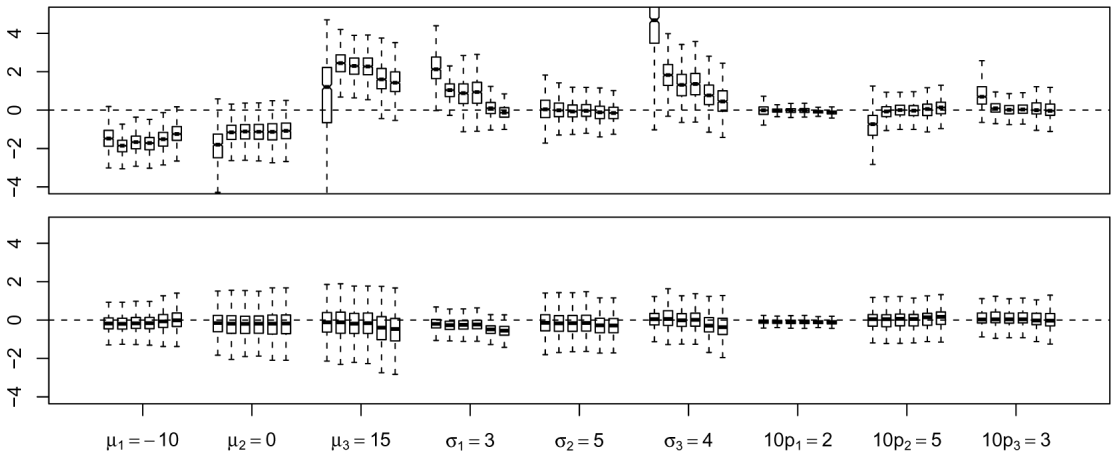

Figure 4.

Box-plot of estimation errors of contaminated (a1) at top vs. clean (a) at bottom, with sample size 400. From left to right: MLE, density power, symmetric χ^2, negative exponential, cubic-inverse, truncated cubic-inverse.

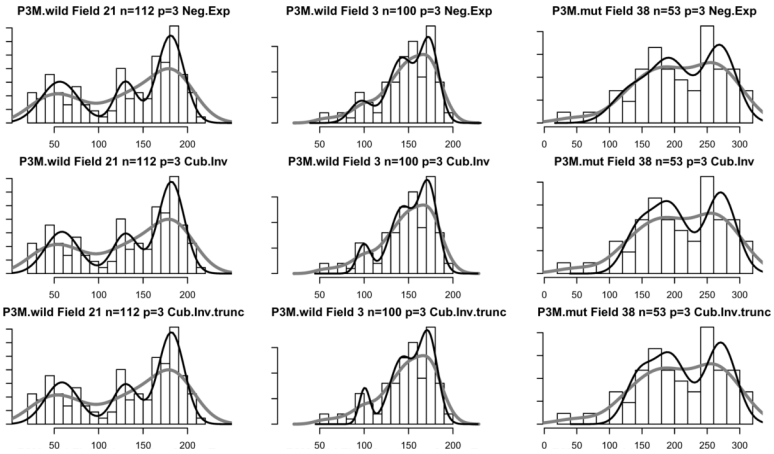

Figure 5.
Selected microscopic fields of the fibril diameter measures from 3 months old mice.