

MINIMUM L2 AND ROBUST KULLBACK–LEIBLER ESTIMATION*

Nils Lid Hjort, University of Oslo

Department of Mathematics, N-0316 Oslo, Norway

Abstract. *This paper introduces two new robust methods for estimation of parameters in a given parametric family. The first method is that of ‘minimum weighted L2’, effectively minimising an estimate of the integrated (and possibly weighted) squared error. The second is ‘robust Kullback–Leibler’, consisting of minimising a robust version of the empirical Kullback–Leibler distance, and can be viewed as a general robust modification of the maximum likelihood procedure. This second method is also related to recent local likelihood ideas for semiparametric density estimation. The methods are described, influence functions are found, as are formulae for asymptotic variances. In particular large-sample efficiencies are computed under the home turf conditions of the underlying parametric model. The methods and formulae are illustrated for the normal model.*

1. Minimum weighted L2 estimation.

Let $X_1, \dots, X_n$ be independent data points from an unknown density f , and suppose that the data are to be fitted to some given regular parametric family of densities $f_\theta(x)$. A simple and natural estimation idea is to minimise an estimate of $\int w(f_\theta - f)^2 dx$, where $w(\cdot)$ is a suitable weight function, perhaps the constant 1. Disregarding the one term that does not depend on the parameter, this leads to the following strategy: minimise

$$Q_n(\theta) = \int w f_\theta^2 dx - 2 \frac{1}{n} \sum_{i=1}^n w(x_i) f_\theta(x_i). \quad (1.1)$$

Taking the derivative this is also the same as solving

$$V_n(\theta) = \int w f_\theta u_\theta (dF_n - f_\theta dx) = 0, \quad (1.2)$$

where $u_\theta(x) = \partial \log f_\theta(x) / \partial \theta$ is the score function of the model, and where F_n is the empirical distribution of the data.

We derived (1.2) as a consequence of the natural (1.1), but forming an estimator by solving this second equation can also be motivated separately. It forces a weighted integral of the nonparametric $dF_n(x)$ to be equal to the corresponding weighted integral of the parametric $f_\theta(x) dx$. In spite of much work in the literature on various minimum distance strategies, the particular estimator (1.1)–(1.2) does not seem to have been studied earlier. It has also been proposed independently by M.C. Jones (personal communication). A method recently considered in Brown and Hwang (1993) has intentions similar to that of (1.1), but is unnecessarily hampered with an intermediate histogram approximation. This is a case of ‘don’t smooth if you don’t have to’.

2. Influence function. Let $\hat{\theta}$ be the estimator and let θ_0 minimise $\int w(f_\theta - f)^2 dx$. There is typically a unique parameter achieving this, and we interpret this θ_0 as the ‘least false’ or ‘most appropriate’ parameter value. As n grows $\hat{\theta}$ converges almost surely to θ_0 . Standard Taylor arguments show that

$$\hat{\theta} - \theta_0 \doteq -V_n^*(\theta_0)^{-1} V_n(\theta_0), \quad (2.1)$$

where $V_n^*(\theta_0)$ is the matrix of derivatives of $V_n(\theta)$. Letting u_θ^* be the matrix of second order derivatives of the log density we have

$$\begin{aligned} V_n^*(\theta) &= \int w(f_\theta^2 u_\theta u_\theta' + f_\theta u_\theta^*) (dF_n - f_\theta dx) \\ &\quad - \int w f_\theta^2 u_\theta u_\theta' dx, \end{aligned}$$

so that $-V_n^*(\theta_0) \rightarrow_p J = J(\theta_0)$, where

$$\begin{aligned} J(\theta) &= \int w f_\theta^2 u_\theta u_\theta' dx \\ &\quad - \int w(f_\theta^2 u_\theta u_\theta' + f_\theta u_\theta^*) (f - f_\theta) dx. \end{aligned}$$

The influence function of the estimator can now be established, via (2.1); see for example Huber

*From Proceedings of the 12th Prague Conference, 1994

(1981) for definition of and important uses of influence functions. Here it becomes

$$I(f, x) = J^{-1}\{w(x)f(x, \theta_0)u(x, \theta_0) - \xi_0\}, \quad (2.2)$$

where

$$\begin{aligned} \xi_0 &= E_f w(X)f(X, \theta_0)u(X, \theta_0) \\ &= \int w(x)f(x, \theta_0)f(x)u(x, \theta_0) dx \\ &= \int w(x)f(x, \theta_0)^2 u(x, \theta_0) dx. \end{aligned}$$

Where notationally convenient we write $f(x, \theta)$ for $f_\theta(x)$, and so on. The (2.2) function is typically bounded, which means robustness. The influence function is in fact also redescending in most cases, going to zero for x -values outside mainstream. This is often considered an attractive robustness feature of an estimation method.

3. Limit distribution. By the central limit theorem and the definition of θ_0 , $\sqrt{n}V_n(\theta_0)$ tends to $\mathcal{N}\{0, M\}$, where

$$\begin{aligned} M &= \text{VAR}_f\{w(X)f(X, \theta_0)u(X, \theta_0)\} \\ &= \int w^2 f_\theta^2 f u_\theta u'_\theta dx - \xi_0 \xi'_0. \end{aligned}$$

From (2.1) follows

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d \mathcal{N}\{0, J^{-1}MJ^{-1}\}, \quad (3.1)$$

with J as given above. Note that this result has been reached without having to assume that the true f belongs to the parametric model.

The expressions for J and M simplify under model conditions. Of course there is some loss of efficiency, that is, the limiting covariance matrix $J^{-1}MJ^{-1}$ is larger than the best possible one under the model, namely $(\int f_\theta u_\theta u'_\theta dx)^{-1}$, achieved by the maximum likelihood method.

4. Local and weighted L2 fitting. The size of the limiting variances depend on the weight function $w(\cdot)$. Choosing a local weight function can be contemplated, say of the kernel type $K_h(x_0 - t)$ around a given x_0 . Here $K_h(u) = h^{-1}K(h^{-1}u)$ and K is a given kernel

function. This gives a locally estimated normal, for example, in a spirit similar to local likelihood methods discussed in Hjort and Jones (1994). The apparatus above can be used to investigate influence functions and large-sample properties.

It is sometimes desirable to let the weight function be data driven too, perhaps to increase precision under close to the model circumstances. One example would be to use $w_n(x) = w_0((x - \tilde{\mu})/\tilde{\sigma})$, for a suitable $w_0(\cdot)$ function, with preliminary robust estimates of location and scale. Result (3.1) is still true under appropriate conditions, with J and M being defined in terms of the limit function version of $w_n(\cdot)$.

5. Local and robust Kullback–Leibler fitting. The local kernel smoothed likelihood function, around a given x_0 , is

$$\begin{aligned} L_n(x_0, \theta) &= \sum_{i=1}^n K_h(x_i - x_0) \log f(x_i, \theta) \\ &\quad - n \int K_h(t - x_0) f(t, \theta) dt, \end{aligned} \quad (5.1)$$

see Hjort and Jones (1994). As shown in Hjort and Jones (1994), maximising (5.1) aims at minimising the localised Kullback–Leibler distance

$$\begin{aligned} d(f, f_\theta) &= \int K_h(t - x_0) \left[f(t) \log \frac{f(t)}{f_\theta(t)} \right. \\ &\quad \left. - \{f(t) - f_\theta(t)\} \right] dt \end{aligned}$$

from true density to parametric density. In other words, the maximiser of (5.1) aims at a ‘least false’ parameter value θ_0 that in general is different from the one associated with the minimum weighted L2 method. Note that a large h gives a flat $K_h(t - x_0)$ function, and brings back the ordinary Kullback–Leibler distance and the traditional full likelihood method.

The aim of Hjort and Jones (1994) is primarily the complete semiparametric estimation of the full density curve, as partly opposed to concentrating on the locally estimated parameters themselves. But this is also automatically one way of obtaining robust parameter estimates for a given parametric family: Apply the above

for a suitable centrally placed x_0 , for a reasonably sized h . The resulting maximiser $\hat{\theta}$ is a robust estimate of θ , and $f(x, \hat{\theta})$ a robust estimate of the underlying density curve.

Hjort and Jones (1994) demonstrate that

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d \mathcal{N}\{0, J_h^{-1} M_h J_h^{-1}\},$$

with certain generally valid expressions available there for J_h and M_h . At the moment it will suffice to give these under model conditions:

$$\begin{aligned} J_h &= \int K_h(t - x_0) u_\theta u'_\theta f_\theta dt, \\ M_h &= \int K_h(t - x_0)^2 u_\theta u'_\theta f_\theta dt - \xi_0 \xi'_0, \end{aligned} \quad (5.2)$$

where $\xi_0 = \int K_h(t - x_0) u_\theta f_\theta dt$. The influence function of this robustified maximum likelihood is also derived in Hjort and Jones (1994), and is of the form

$$I(f, x) = J_h^{-1} \{K_h(x - x_0) u(x, \theta) - \xi_0\}. \quad (5.3)$$

This is reasonably similar to the influence function (2.2) for the minimum L2 method. In many cases the present method, with a suitably chosen h , is more efficient at the model than at least the unweighted version of the minimum L2 method.

6. The normal model. The most important special case is that of fitting data to a normal (μ, σ^2) .

6.1. THE MINIMUM L2 METHOD. From (1.2) two equations are easily put up to define minimum L2 estimators $\hat{\mu}$ and $\hat{\sigma}$. These are solved for example by the iterative Newton–Raphson technique. Regarding performance, under Gaussian circumstances, and using a constant weight function, we find

$$\begin{aligned} J &= (\sigma^3 \sqrt{2\pi})^{-1} \text{diag}(1/2^{3/2}, 3/2^{5/2}), \\ M &= (\sigma^4 2\pi)^{-1} \text{diag}(1/3^{3/2}, 2/3^{3/2} - 1/8). \end{aligned}$$

This gives an asymptotic variance for $\hat{\mu}$ of size $1.5396 \sigma^2/n$ and an asymptotic variance for $\hat{\sigma}$ of size $0.9241 \sigma^2/n$. These should be compared to the minimum possible values, under model

conditions. These optimal figures are achieved by the ML method, and are σ^2/n and $\frac{1}{2}\sigma^2/n$, respectively. This makes the direct minimum L2 method qualify as a ‘quite robust but perhaps too inefficient method’. Increased efficiency at the model is achieved through appropriate choices of weight function $w(\cdot)$, cf. comments at the end of Section 4. One possibility here is $w_n(x) = \exp\{\frac{1}{2}\delta(x - \tilde{\mu})^2/\tilde{\sigma}^2\}$, defined in terms of preliminary robust estimates of location and scale, and with an extra tuning parameter $\delta \in (0, 1)$. Choosing e.g. $\delta = 0.8$ leads to quite good efficiency at the model, while still retaining a reasonable robustness.

6.2. THE ROBUSTIFIED ML METHOD. The robust Kullback–Leibler fitting method of Section 5 can easily be made better than the unweighted minimum L2 method. For the present normal model, let us use a normal kernel. The method is then to minimise, for given x_0 , the function

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \phi\left(\frac{x_i - x_0}{h}\right) \frac{1}{h} \{\log \sigma + \frac{1}{2}(x_i - \mu)^2/\sigma^2\} \\ &+ \phi\left(\frac{x_0 - \mu}{\sqrt{\sigma^2 + h^2}}\right) \frac{1}{\sqrt{\sigma^2 + h^2}} \end{aligned} \quad (6.1)$$

over all (μ, σ) . We may compute the J_h and M_h matrices of (5.2) without serious difficulties. But in the present context the interest lies more in getting hold of a single, robust (μ, σ) -estimate, than in obtaining a full function of local estimates. Therefore we suggest using $x_0 = \tilde{\mu}$, a robust preliminary estimate of the mean, say the simple median. Minimising (6.1) with this x_0 defines the proposed $\hat{\mu}$ and $\hat{\sigma}$.

Again it is of interest to see how well the method fares under Gaussian home-turf conditions. Somewhat arduous calculations give two diagonal matrices (J_μ, J_σ) and (M_μ, M_σ) for J_h and M_h of (5.2). Here

$$J_\mu = \frac{1}{\sigma^2} \frac{1}{\sqrt{2\pi}} \frac{1}{h} \frac{1}{R^3} \quad \text{and} \quad M_\mu = \frac{1}{\sigma^2} \frac{1}{2\pi} \frac{1}{h^2} \frac{1}{S^3},$$

in which

$$R = (1 + \sigma^2/h^2)^{1/2} \quad \text{and} \quad S = (1 + 2\sigma^2/h^2)^{1/2}.$$

Similarly,

$$J_\sigma = \frac{1}{\sigma^2} \frac{1}{\sqrt{2\pi}} \frac{1}{h} \left(1 - \frac{2}{R^2} + \frac{3}{R^4}\right),$$

$$M_\sigma = \frac{1}{\sigma^2} \frac{1}{2\pi} \frac{1}{h^2} \left\{ \frac{1}{S} \left(1 - \frac{2}{S^2} + \frac{3}{S^4}\right) - \frac{1}{R^2} \left(1 - \frac{1}{R^2}\right)^2 \right\}.$$

Thus $\sqrt{n}(\hat{\mu} - \mu) \rightarrow_d \mathcal{N}\{0, \kappa_\mu^2\}$ and $\sqrt{n}(\hat{\sigma} - \sigma) \rightarrow_d \mathcal{N}\{0, \kappa_\sigma^2\}$, where the asymptotic variances are found as M_μ/J_μ^2 and M_σ/J_σ^2 . Some calculations give

$$\kappa_\mu^2 = \sigma^2 \frac{R^6}{S^3} = \sigma^2 \frac{(1 + 1/k^2)^3}{(1 + 2/k^2)^{3/2}},$$

writing $h = k\sigma$, and similarly

$$\kappa_\sigma^2 = \sigma^2 \frac{(1 + 1/k^2)^2}{(1 + 2/k^2)^{5/2}}$$

$$\frac{(1 + 1/k^2)^3 (2 + 4/k^4) - (1 + 2/k^2)^{5/2} 1/k^4}{(2 + 1/k^4)^2}.$$

How large should h be chosen? We think of h as $k\tilde{\sigma}$, where $\tilde{\sigma}$ is a robust preliminary estimate of standard deviation, and need to choose the factor k . As a mild surprise the value $k = 1$ gives precisely the same large-sample variances under the model as the straightforward minimum L2 method of Section 2, respectively $1.5396\sigma^2$ and $0.9241\sigma^2$. A more efficient but still quite robust value would be $k = 2$, ‘place a normal with two standard deviations around the median and maximise the local kernel smoothed likelihood’. Then the values are $1.063\sigma^2$ and $0.563\sigma^2$, only a few percent above the values that are optimal under the model, viz. σ^2 and $\sigma^2/2$. Increasing the value to three estimated standard deviations brings the large-sample variances further down to $1.015\sigma^2$ and $0.5152\sigma^2$. One should not go much further if robustness is aimed for, but of course a large h gives back these optimal values.

Comparing the performance of the weighted L2 method, say with the data driven weight function indicated above, with that of the robust Kullback–Leibler estimator, is an interesting problem for further research. One should also devise criteria for choosing the necessary fine-tuning parameters.

7. Robust estimation of location and covariance matrix.

The ideas and results above generalise easily to the multi-dimensional case. In particular the localised Kullback–Leibler method seems to constitute a fruitful way of obtaining robust estimates of μ and Σ , the mean vector and covariance matrix of the underlying distribution. The estimates can be viewed as robust estimates of these parameters under normality assumptions but also outside normality.

One concrete version of this scheme, in the p -dimensional case, is as follows: Start out with preliminary and robust estimates $\tilde{\mu}$ and $\tilde{\Sigma}$ for mean and covariance matrix. Then carry out local likelihood estimation with a Gaussian kernel function centred at $\tilde{\mu}$ and with covariance matrix of size $h^2\tilde{\Sigma}$. This is seen to be the same as minimising the criterion function

$$\left[\frac{1}{n} \sum_{i=1}^n \frac{\exp\{-\frac{1}{2}(x_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (x_i - \tilde{\mu})/h^2\}}{h^p |\tilde{\Sigma}|^{1/2}} \right]$$

$$\left\{ \frac{1}{2} \log |\Sigma| + \frac{1}{2} (x_i - \mu)' \Sigma^{-1} (x_i - \mu) \right\}$$

$$+ \frac{\exp\{-\frac{1}{2}(\mu - \tilde{\mu})' (h^2 \tilde{\Sigma} + \Sigma)^{-1} (\mu - \tilde{\mu})\}}{|h^2 \tilde{\Sigma} + \Sigma|^{1/2}}$$

over all possible (μ, Σ) . Note that this method properly generalises that of (6.1). For h larger than say 5 the procedure is practically the same as ordinary maximum likelihood estimation. A value of perhaps $h = 2$ constitutes a modified maximum likelihood procedure with quite good robustness qualities without sacrificing much in efficiency under multinormal conditions.

Acknowledgements. This paper has benefited from ongoing joint work on related matters with M.C. Jones, I.R. Harris and A. Basu.

References

- Brown, L.D. and Hwang, J.T.G. (1993). How to approximate a histogram by a normal density. *American Statistician* **47**, 251–255.
- Hjort, N.L. and Jones, M.C. (1994). Locally parametric nonparametric density estimation. Submitted for publication.
- Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.