Are Grouped Data Robustly Fitted?

VICTORIA-FESER, Maria-Pia, RONCHETTI, Elvezio & London School of Economics

Abstract

In this paper we compute the IF of a general class of estimators for grouped data, namely the class of MPE. We find that this IF can be large although it is bounded. Therefore, we propose a more general class of estimators, the MGP-estimators, which include the class of estimators based on the power divergence statistic and permits to define robust estimators. By analogy with Hampel's theorem, we define optimal bounded IF estimators and by a simulation study, we show that under small model contaminations, they are a lot more stable than the classical estimators for grouped data. Finally, our results are applied to a particular real example.

Reference

VICTORIA-FESER, Maria-Pia, RONCHETTI, Elvezio & London School of Economics. Are Grouped Data Robustly Fitted?. London School of Economics, 1995

Available at:
http://archive-ouverte.unige.ch/unige:6507

Disclaimer: layout of this document may differ from the published version.
Are Grouped Data Robustly Fitted?

by

Maria-Pia Victoria-Feser
Department of Statistics
London School of Economics and Political Science
Houghton Street
London WC2A 2AE, UK

and

Elvezio Ronchetti
Faculty of Economic and Social Sciences
University of Geneva
CH-1211 GENEVA
Switzerland

Statistics Research Report LSERR15, LSE
February 1995

*Partially supported by the 'Fond National Suisse pour la Recherche Scientifique'.
Abstract

In this paper we compute the IF of a general class of estimators for grouped data, namely the class of MPE. We find that this IF can be large although it is bounded. Therefore, we propose a more general class of estimators, the MGP-estimators, which include the class of estimators based on the power divergence statistic and permits to define robust estimators. By analogy with Hampel’s theorem, we define optimal bounded IF estimators and by a simulation study, we show that under small model contaminations, they are a lot more stable than the classical estimators for grouped data. Finally, our results are applied to a particular real example.
1 Introduction

In many situations data are available only in a grouped form. Typically, grouped data are a simplified representation of a large number of observations. This is for instance the case of personal income data where typically only frequencies within specified classes are available. We assume that the underlying distribution of the continuous data is modelled by a parametric model \( \{F_\theta\} \). In this paper we address the robustness problems related to the estimation of \( \theta \) when the observations are available only in grouped form. In particular, we will show that we can improve the classical techniques by providing robust estimators which are more reliable in the presence of deviations from the underlying model. Whereas there exists a large body of literature on robustness for continuous data (see for instance Huber 1981 and Hampel et al. 1986), the issue of grouped data has been somewhat neglected.

Barnett (1992) in his review paper mentions outlier detection in grouped data as an open research area. He points out that intuitively an outlier is detected in a class whose corresponding probability (according to the underlying model) is small compared to the observed frequency. This can be formalized by defining the “adjusted residuals” (see also Fuchs and Kennet 1980) \( t_j = \sqrt{n\{p_j - k_j(\theta)\}} / \sqrt{\{k_j(\theta)(1 - k_j(\theta))\}}, \ j = 1,\ldots,J \), where \( k_j(\theta) \) is the expected probability of class \( j \) according to the underlying model and
$p_j$ is the observed relative frequency in class $j$. A class which yields $\max |t_j|$ is then considered outlying. Since $k_j(\theta)$ depends on the unknown parameter $\theta$, a robust estimation of $\theta$ is important to avoid the inflation of the denominator of $t_j$ and the consequent misleading effect.

We consider the following basic model. We have a parametric model $F_\theta$ with $\theta \in \mathbb{R}^p$ for the underlying continuous data but we observe only a realization of the random variable $N = (N_1, \ldots, N_J)^T$ which gives the number of observations having a characteristic corresponding to the classes $I_1, \ldots, I_J$. $N$ is distributed according to a multinomial distribution with cell probabilities $k(\theta) = (k_1(\theta), \ldots, k_J(\theta))^T$, i.e.

$$P \left[ N = (n_1, \ldots, n_J)^T \right] = \frac{n!}{n_1! \cdots n_J!} k_1^{n_1}(\theta) \cdots k_J^{n_J}(\theta),$$

where $\sum_{j=1}^{J} n_j = n$, the sample size, and $k_j(\theta) = \int_{I_j} dF_\theta(x)$, such that $\sum_{j=1}^{J} k_j(\theta) = 1$. If we denote by $x_1, \ldots, x_n$ the unknown sample of continuous data and by $F^{(n)}(x) = n^{-1} \sum_{i=1}^{n} \Delta_{x_i}(x)$ its empirical distribution function ($\Delta_{x_i}$ is the pointmass in $x_i$), then $p_j = \frac{n_j}{n} = \int_{I_j} dF^{(n)}(x)$ is the relative frequency in the class $I_j$. As $n \to \infty$, $\sqrt{n}(p - k(\theta)) \xrightarrow{p} N(0, \Sigma)$, where $p = (p_1, \ldots, p_J)^T$ and $\Sigma = \text{diag}(k(\theta)) - k(\theta)k^T(\theta)$; see Bishop et al. (1975).

The paper is organized as follows. In section 2 we study the robustness properties of classical estimators of $\theta$. In particular, we show that small
deviations from the underlying model $F_{\theta}$ have the effect of introducing a large bias in the maximum likelihood estimator (MLE) and in more general minimum power divergence estimator (MPE) (Cressie and Read 1984). We use the influence function (IF) (Hampel 1968, 1974 and Hampel et al. 1986) as a basic tool to assess the robustness properties of the estimators. In section 3 we generalize the classical estimators by introducing a new class of estimators for $\theta$. In this class we find in section 4 an optimal robust estimator which is the best trade-off between efficiency at the model and robustness. Section 5 is devoted to the sensitivity of the estimators to grouping effect. Section 6 presents some numerical evidence on the performance of the new robust estimators and section 7 concludes.

2 Classical estimators and their robustness properties

In the following, we denote by $T$ an estimator of $\theta$, the parameter of interest of the underlying model. The MLE is among the most well known classical estimators of $\theta$ for the multinomial model. It belongs to a large class of asymptotically equivalent estimators. This class was proposed by Cressie and Read (1984) and Read and Cressie (1988) and is based on the power
divergence statistic. These estimators, the MPE, are defined by the solution in $\theta$ of
\[
G^\lambda(p; k(\theta)) = \sum_{j=1}^{J} \left( \frac{p_j}{k_j(\theta)} \right)^{\lambda+1} \frac{\partial}{\partial \theta} k_j(\theta) = 0 ,
\] (1)
where $-\infty < \lambda < \infty$ is a fixed parameter. When $\lambda = 0$, we have the MLE.
Cressie and Read (1984) showed that any MPE is asymptotically normal with mean $\theta$ and asymptotic covariance matrix $Q^{-1}$, where

\[
q_{ij} = \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta_i} \log k_j(\theta) \frac{\partial}{\partial \theta_l} \log k_j(\theta)
\]
so that $Q$ is indeed the Fisher information matrix for the multinomial model defined by $k(\theta)$.

One way of assessing the robustness properties of any statistic is by means of the IF; see Hampel (1968, 1974) and Hampel et al. (1986). With grouped data errors can typically occur at two levels. First there are errors or model misspecification with respect to the underlying model for continuous data. Secondly, there are grouping effects that occur at the grouping stage. We first consider deviations from the underlying model. Robustness with respect to grouping effects will be discussed in section 5. Let us introduce a perturbation to $F_\theta$ and let us write the misspecified distribution, say $G_\varepsilon$, as

\[
G_\varepsilon = (1 - \varepsilon)F_\theta + \varepsilon W ,
\] (2)
where $W$ is any distribution. The estimator $T_{MPE}$ of $\theta$ can then be written as a functional of the underlying distribution. Indeed, $T_{MPE}$ is determined by $k(\theta)$ which in turn is determined by the underlying distribution. Therefore, from (1) $T_{MPE}(G_\varepsilon)$ (for notational simplicity, we shall write $T_{MPE}(G_\varepsilon) := T(G_\varepsilon)$) is defined by

$$0 = \sum_{j=1}^{J} \left( \int_{I_j} \frac{dG_\varepsilon}{k_j(T(G_\varepsilon))} \right)^{\lambda+1} \frac{\partial}{\partial \theta} k_j(\theta) \bigg|_{\theta = T(G_\varepsilon)} =$$

$$\sum_{j=1}^{J} \left( (1 - \varepsilon)k_j(\theta) + \varepsilon \int_{I_j} dW \right)^{\lambda+1} \frac{\partial}{\partial \theta} k_j(\theta) \bigg|_{\theta = T(G_\varepsilon)} .$$

If we take the derivative with respect to $\varepsilon$ at $\varepsilon = 0$, we obtain the marginal effect of a small departure of the model $F_{\theta}$ to the distribution $W$. It is given by

$$\frac{\partial}{\partial \varepsilon} T(G_\varepsilon) \bigg|_{\varepsilon = 0} = \left[ \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \right]^{-1} \cdot$$

$$\sum_{j=1}^{J} \frac{\partial}{\partial \theta} \log k_j(\theta) \int_{I_j} dW$$

We can note that the influence on the MPE of an infinitesimal model deviation does not depend on $\lambda$. Therefore this influence is the same for all the members of the class. Moreover, $\int_{I_j} dW$ is the relative frequency $p_j$ in the class $I_j$ under the distribution $W$. Of course, at the model $\int_{I_j} dW$ is
\( k_j(\theta) \). If we choose \( W \) as the point mass 1 at any point \( x \), we obtain the IF of \( T_{MPE} \) at the model \( F_\theta \)

\[
\text{IF}(x, T_{MPE}, F_\theta) = \left[ \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \frac{\partial}{\partial \theta^T} \log k_j(\theta) \right]^{-1} - \sum_{j=1}^{J} \delta_j(x) \frac{\partial}{\partial \theta} \log k_j(\theta),
\]

where \( \delta_j(x) = \begin{cases} 1 & \text{if } x \in I_j \\ 0 & \text{otherwise} \end{cases} \). We note that the IF is bounded because

\[
\max_x \| \text{IF}(x, T_{MPE}, F_\theta) \| = \max_j \left\| \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \frac{\partial}{\partial \theta^T} \log k_j(\theta) \right\|^{-1} \frac{\partial}{\partial \theta} \log k_j(\theta) < \infty.
\]

However, this maximum value which describes the maximum bias on the MPE over the neighborhood of \( F_\theta \) defined by (2), can be very large depending on the value of \( \partial \log k_j(\theta) / \partial \theta \). Indeed

\[
\frac{\partial}{\partial \theta} \log k_j(\theta) = \frac{\int_{I_j} s(x; \theta) dF_\theta(x)}{k_j(\theta)},
\]

where \( s(x; \theta) = \partial \log f_\theta(x) / \partial \theta \) is the scores function of the underlying model, and if \( k_j(\theta) \) is small and \( s(x; \theta) \) is large in the interval \( I_j \), then the maximum bias can be very large. It should be stressed that this happens very frequently.
with commonly used distributions, since the classes with low probability correspond to the lowest or highest observations where, at the same time, the scores function is large. Typical examples are the Pareto and the Gamma distribution which are commonly used in modelling income data; see Victoria-Feser and Ronchetti (1994).

3 A general class of estimators

In the previous section we have seen that the bias of a MPE over a neighborhood of the model can be large. This is a property we would like to avoid. In this section, we generalize the class of estimators for the parameter $\theta$ in order to be able to build a more robust estimator, i.e. an estimator with a smaller IF. This is important because as we will see, the robust estimator derived from this more general class is more stable when the underlying model does not hold exactly.

For continuous data, Huber (1964) defined the class of M-estimators which are a generalization of the MLE. With grouped data, we propose the following generalization which defines MGP-estimators (i.e generalized MLE with grouped data). A MGP-estimator $T_{\psi}$ defined through a function
\[ \psi = [\psi_1, \ldots, \psi_J]^T \] is given by the solution in \( \theta \) of

\[
\sum_{j=1}^{J} \psi_j(\theta) p_j^\gamma = \sum_{j=1}^{J} \psi_j(\theta) \left[ \int_{I_j} dF^{(n)}(x) \right]^\gamma = 0 \quad (3)
\]

where \(-\infty < \gamma < \infty\) and \( \psi \) is an arbitrary function. To be able to derive some properties of MGP-estimators, we have to impose some conditions on the function \( \psi \). For simplicity of derivation, we assume differentiability for \( \psi \). However, the results we will show hold under weaker conditions like piecewise differentiability.

This class includes the MLE when \( \psi_j(\theta) = \partial \log k_j(\theta) / \partial \theta \) and \( \gamma = 1 \), and it includes the MPE when \( \psi_j(\theta) = \frac{1}{k_j(\theta)} \frac{\partial}{\partial \theta} k_j(\theta) \) with \( \gamma = \lambda + 1 \). Our goal will be to choose \( \psi_j(\theta) \) in an appropriate way in order to ensure the robustness of the resulting estimator.

Let us now look at the asymptotic properties of a MGP-estimator. First, Fisher consistency implies

\[
\sum_{j=1}^{J} \psi_j(\theta) \left[ \int_{I_j} dF_\theta(x) \right]^\gamma = \sum_{j=1}^{J} k_j(\theta)^\gamma \psi_j(\theta) = 0 \quad . \quad (4)
\]

Secondly, it can be shown (see Appendix A) that the MGP-estimator is
asymptotically normal, i.e. \( \sqrt{n} (T_{\psi} - \theta) \xrightarrow{D} N(0, \Sigma_{MGP}) \) where

\[
\Sigma_{MGP} = \left[ \sum_{j=1}^{J} k_j(\theta) \gamma_j \psi_j(\theta) \frac{\partial}{\partial \theta^T} \log k_j(\theta) \right]^{-1} 
\]

\[
\left[ \sum_{j=1}^{J} k_j(\theta) \gamma_j - 1 \psi_j(\theta) \psi_j(\theta)^T \right] . 
\]

By using the same techniques as in section 2, we obtain under (2)

\[
\frac{\partial}{\partial \varepsilon} T_{\psi}(G_\varepsilon) \bigg|_{\varepsilon = 0} = \left[ \sum_{j=1}^{J} \psi_j(\theta) k_j(\theta) \gamma \frac{\partial}{\partial \theta^T} \log k_j(\theta) \right]^{-1} 
\]

\[
\sum_{j=1}^{J} \psi_j(\theta) k_j(\theta) \gamma - 1 \int_{I_j} dW(x) . 
\]

and

\[
IF(x, T_{\psi}, F_\theta) = \left[ \sum_{j=1}^{J} k_j(\theta) \gamma \psi_j(\theta) \frac{\partial}{\partial \theta^T} \log k_j(\theta) \right]^{-1} 
\]

\[
\sum_{j=1}^{J} \delta_j(x) k_j(\theta) \gamma - 1 \psi_j(\theta) . 
\]

Note that

\[
\Sigma_{MGP} = E \left[ IF(x, T_{\psi}, F_\theta) IF^T(x, T_{\psi}, F_\theta) \right] , 
\]

where \( \Sigma_{MGP} \) is the asymptotic covariance matrix given by (5).
We see that if we put $\gamma = \lambda + 1$ and $\psi_j(\theta) = \frac{1}{k_j(\theta)^\gamma} \frac{\partial}{\partial \theta} k_j(\theta)$ we get the expression for the MPE given by (1) and $\Sigma_{\text{MGP}}$. However, a better choice for $\psi_j(\theta)$ will lead to an optimal robust estimator, i.e. a MGP-estimator which minimizes the asymptotic variance, under a prescribed bound on its IF.

4 Optimal robust estimator with grouped data

The optimality problem is stated in terms of finding the best trade-off between efficiency and robustness. For efficiency reasons we want to minimize the asymptotic covariance of the estimator and for robustness reasons we want an IF as small as possible. A bound on $\max_x \|\text{IF}(x; T_\psi, F_\theta)\|$ controls the robustness of the estimator.

This condition is equivalent to

$$\left\| \left[ \sum_{j=1}^J k_j(\theta)^\gamma \psi_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \right]^{-1} k_j(\theta)^{\gamma-1} \psi_j(\theta) \right\| \leq c \forall j \quad (8)$$

As in the continuous case, it is difficult to minimize the asymptotic covariance matrix itself. Therefore, a similar problem is to minimize its trace. From (7)
and (8) the optimality problem is given by

\[
\min_{\psi_1, \ldots, \psi_J} \text{tr} \left\{ E \left[ \text{IF}(x, T_{\psi}, F_\theta) \text{IF}^T(x, T_{\psi}, F_\theta) \right] \right\}
\]

subject to (8), among all Fisher consistent estimators, i.e. the estimator satisfying (4). This problem is similar to Hampel’s theorem for continuous data (see Hampel et al. 1986). The solution is given by the solution in \( \theta \) of (3), where

\[
\psi_j(\theta) = \left( \frac{1}{k_j(\theta)} \right)^{\gamma} A(\theta) \left[ \frac{\partial}{\partial \theta} k_j(\theta) - a(\theta)k_j(\theta) \right] \cdot W_j(A(\theta), a(\theta)) = 0 , \quad (9)
\]

\[
W_j(A(\theta), a(\theta)) = \min \left\{ 1; \frac{c}{\left\| A(\theta) \left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - a(\theta) \right] \right\|} \right\} , \quad (10)
\]

and where the matrix \( A(\theta) \) and vector \( a(\theta) \) are determined implicitly by the equations

\[
\sum_{j=1}^J A(\theta) \left[ \frac{\partial}{\partial \theta} k_j(\theta) - a(\theta)k_j(\theta) \right] \cdot W_j(A(\theta), a(\theta)) = 0 , \quad (11)
\]

\[
\sum_{j=1}^J A(\theta) \left[ \frac{\partial}{\partial \theta} k_j(\theta) - a(\theta)k_j(\theta) \right] \frac{\partial}{\partial \theta^T} \log k_j(\theta) \cdot W_j(A(\theta), a(\theta)) = I . \quad (12)
\]

The proof is given in Appendix B.

In analogy to the continuous case we will call this estimator optimal B-
robust (OBRE). Note that we actually found a set of solutions depending on
a parameter $\gamma$. The functions $\psi_j$ in (9) are weighted linear combinations of
the ‘scores’ functions $\tilde{s}_j$ for MPE given by $\tilde{s}_j(p_j; \theta) = \left( \frac{p_j}{k_j(\theta)} \right)^\gamma \partial k_j(\theta)/\partial \theta$. Therefore, when $c = \infty$, we obtain the MPE. This is not surprising because
we know that they are asymptotically efficient.

Moreover, the functions $W_j$ in (10) can be interpreted as a weight assigned
to each class $I_j$. This can be used as diagnostic information since small
weights will reveal immediately the outlying classes.

To compute the OBRE, we use the same algorithm as in the continuous
case (see e.g. Victoria-Feser and Ronchetti 1994). We first find, for a given $
\theta$, the matrix $A(\theta)$ and the vector $a(\theta)$ by means of the equations (11) and
(12), and then compute a Newton-Raphson step for $\theta$. The algorithm is the
following.

Algorithm for computing OBRE with grouped data

Step 1: For a given $\gamma$, fix a precision threshold $\eta$, an initial value for the pa-
parameter $\theta$ and initial values $a(\theta) = a = 0$ and

$$A(\theta) = A = \left[ \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \right]^{-1}.$$
Step 2: Solve the following equations with respect to $a$ and $A$:

$$A = \left[ \sum_{j=1}^{J} \left( \frac{\partial}{\partial \theta} k_j(\theta) - a \cdot k_j(\theta) \right) \frac{\partial}{\partial \theta} \log k_j(\theta) W_j(A, a) \right]^{-1}$$

and

$$a = \left[ \sum_{j=1}^{J} k_j(\theta) W_j(A, a) \right]^{-1} \sum_{j=1}^{J} \frac{\partial}{\partial \theta} k_j(\theta) W_j(A, a).$$

The current values of $\theta$, $a$ and $A$ are used as starting values to solve the given equations.

Step 3: Compute the Newton-Raphson step for $\theta$ given by

$$\Delta \theta = \sum_{j=1}^{J} \left( \frac{p_j}{k_j(\theta)} \right)^\gamma A \cdot \left[ \frac{\partial}{\partial \theta} k_j(\theta) - a \cdot k_j(\theta) \right] \cdot W_j(A, a).$$

Step 4: If $\Delta \theta > \eta$ then $\theta \leftarrow \theta + \Delta \theta$ and go to step 2, else stop.

It should be stressed that the initial value for $\theta$ is important since the convergence of the algorithm is guaranteed for an initial $\theta$ near the solution. As starting point one can take the MPE, and if it is too far from a very robust estimator, one can always start by computing a less robust estimator (high value for the bound $c$) and use the estimate as starting value for the computation of a very robust estimator (low value for $c$).
The efficiency of the OBRE can be expressed by the ratio between the trace of its asymptotic covariance matrix and the trace of the asymptotic covariance matrix of the MPE. From (5) and (12), we have

\[
\Sigma_{\text{MPE}} = \sum_{j=1}^{J} \frac{1}{k_j(\theta)} A(\theta) \left[ \frac{\partial}{\partial \theta} k_j(\theta) - a(\theta) k_j(\theta) \right] \cdot \\
\left[ \frac{\partial}{\partial \theta} k_j(\theta) - a(\theta) k_j(\theta) \right]^T A(\theta)^T W_j^2(A(\theta), a(\theta)) \\
= \sum_{j=1}^{J} k_j(\theta) A(\theta) \left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - a(\theta) \right] \cdot \\
\left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - a(\theta) \right]^T A(\theta)^T W_j^2(A(\theta), a(\theta)) .
\]

We can see that for \( c = \infty \), \( W_j(A(\theta), a(\theta)) = 1 \), \( \forall j \), \( a(\theta) = 0 \), \( A(\theta) = \left[ \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \right]^{-1} \), and \( \Sigma_{\text{MGP}} = \Sigma \). Therefore the asymptotic efficiency of the OBRE is given by

\[
\text{tr} \left[ \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \right]^{-1} \\
\text{tr} \left[ \sum_{j=1}^{J} k_j(\theta) A(\theta) \left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - a(\theta) \right] \left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - a(\theta) \right]^T A(\theta)^T W_j^2(A(\theta), a(\theta)) \right] .
\]

One can choose the bound \( c \) as a function of the asymptotic efficiency at the model one wants to preserve. Indeed when \( c = \infty \), we have 100% efficiency but the estimator is not robust. When we lower \( c \), we gain robustness but lose efficiency. What is the minimum value that \( c \) can achieve? From (9) and (12) we have \( I = \sum_{j=1}^{J} k_j(\theta)^{\gamma-1} \psi_j(\theta) \frac{\partial}{\partial \theta} k_j(\theta) \) and taking the traces, we get

16
\[ dim(\theta) = tr(I) = \sum_{j=1}^{J} k_j(\theta)^{\gamma-1} \psi_j(\theta)^T \frac{\partial}{\partial \theta} k_j(\theta) \leq \sum_{j=1}^{J} \left\| k_j(\theta)^{\gamma-1} \psi_j(\theta) \right\| \left\| \frac{\partial}{\partial \theta} k_j(\theta) \right\| < c \sum_{j=1}^{J} \left\| \frac{\partial}{\partial \theta} k_j(\theta) \right\|. \]  

Hence,
\[ c > \frac{dim(\theta)}{\sum_{j=1}^{J} \left\| \frac{\partial}{\partial \theta} k_j(\theta) \right\|}. \]

### 5 Local shift sensitivity

With grouped data, grouping errors can occur easily and therefore a robust estimator should be robust to this kind of model deviation. The *local shift sensitivity* \( \lambda^* \) (see Hampel et al. 1986) measures the influence on the estimator of a shift of an observation \( x \) from one class to next. It is therefore given by the maximum difference between the value of the IF between two consecutive classes. That is, the \( \lambda^* \) is given by the maximum difference between (6) evaluated at two consecutive classes \( j \) and \( j + 1 \). For MGP-estimators, we have

\[ \lambda^* = \max_j \left\| \left[ \sum_{j=1}^{J} k_j(\theta)^{\gamma-1} \psi_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \right]^{-1} \right. \]

\[ \left. \left[ k_j(\theta)^{\gamma-1} \psi_j(\theta) - k_{j+1}(\theta)^{\gamma-1} \psi_{j+1}(\theta) \right] \right\|, \]
and in the particular case of MPE,

$$\lambda^*_{\text{MPE}} = \max_j \left| \left[ \sum_{j=1}^{J} k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \frac{\partial}{\partial \theta} \log k_j(\theta) \right]^{-1} \left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - \frac{\partial}{\partial \theta} \log k_{j+1}(\theta) \right] \right| $$

Note that

$$\frac{\partial}{\partial \theta} \log k_j(\theta) = \frac{\int_{I_j} s(x; \theta) dF(\theta)(x)}{\int_{I_j} dF(\theta)(x)} \approx \frac{s(x^{(j)}; \theta) f(\theta(x^{(j)})) (x_{j+1} - x_j)}{f(\theta(x^{(j)})) (x_{j+1} - x_j)} = s(x^{(j)}; \theta),$$

where $x_j = \inf\{I_j\}, x_{j+1} = \sup\{I_j\}$ and $x^{(j)} = \frac{x_j + x_{j+1}}{2}$. Therefore, the local shift sensitivity of the MPE is approximately proportional to

$$\max_j \left| s(x^{(j)}; \theta) - s(x^{(j+1)}; \theta) \right|. \quad (14)$$

For example, for the Pareto distribution given in (15), (14) is proportional to $\max_j |\log(x^{(j+1)}) - \log(x^{(j)})|$ which, according to the length of the classes and the positions of the classes (in high or low $x$), can be large. However, with robust estimators from the class of MGP-estimators, by construction

$$\|k_j(\theta)^{\gamma-1}\psi_j(\theta)\| < c \forall j, $$

and the $\lambda^*$ is bounded. Therefore, the OBRE derived in section 4 is robust also with respect to grouping errors.
6 Numerical Results

In order to measure the performance of our robust estimator, in a first step we computed the MLE and the OBRE of the parameter $\alpha$ of the Pareto distribution from 50 simulated samples of 1000 observations with parameter values $\alpha = 3.0$ and $x_0 = 0.5$ that we contaminated by means of

$$(1 - \varepsilon)F_{\alpha,x_0} + \varepsilon F_{\alpha,10-x_0} .$$

The Pareto distribution can be used as a model for incomes which are often reported in a grouped form. It is given by

$$F_\theta(x) = 1 - \left(\frac{x}{x_0}\right)^{-\alpha},$$

where $\alpha > 1$ and $0 \leq x_0 \leq x < \infty$, $x_0$ being a fixed parameter. The results are presented in Table 1. The standard errors for the values of the estimates in Table 1 are smaller than 0.02.

We can see that although the IF of the MLE is bounded, when the underlying model is contaminated, the MLE has a large bias. On the other hand, with the robust estimator we can see that this bias and the overall MSE are considerably smaller. For example, with 5% of contamination, the MSE of the MLE is more than 7 times greater than the MSE of the OBRE.
<table>
<thead>
<tr>
<th>ε</th>
<th>OBRE</th>
<th>MSE</th>
<th>MLE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>3.02</td>
<td>0.02</td>
<td>3.01</td>
<td>0.01</td>
</tr>
<tr>
<td>1%</td>
<td>2.97</td>
<td>0.02</td>
<td>2.81</td>
<td>0.05</td>
</tr>
<tr>
<td>3%</td>
<td>2.87</td>
<td>0.04</td>
<td>2.45</td>
<td>0.27</td>
</tr>
<tr>
<td>5%</td>
<td>2.75</td>
<td>0.08</td>
<td>2.23</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 1: **MLE and OBRE** ($\gamma = 1$, $c = 5.0$) for the Pareto model with grouped data

In a second step, we analyse a real data sample. We consider here a subsample (of size 746) of a standard data set of disposable income in the UK, 1979. The income receivers are households in receipt of social benefits (see DSS 1992). The original sample is not in a grouped form, but for our purpose we grouped the data in 58 (???) equal sized classes. The first class is extended to 0 and the last class to $\infty$.

We chose to model the data with the Gamma distribution with shape parameter $\alpha$ and scale parameter $\delta$. We computed the OBRE ($\gamma = 1$) for different bound $c$. The results are presented in Table 2. We can see that the MLE ($c = \infty$) and a robust estimator (e.g. $c = 100$) give considerably different results. In a graph of the histogram of the empirical distribution and the classically and robustly estimated densities, the difference becomes clearer (see Figure 1). While the MLE tries to accommodate the tails of the


<table>
<thead>
<tr>
<th>$c$</th>
<th>$\infty$</th>
<th>150</th>
<th>120</th>
<th>115</th>
<th>110</th>
<th>105</th>
<th>100</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>16.59</td>
<td>20.198</td>
<td>22.536</td>
<td>23.593</td>
<td>25.939</td>
<td>30.125</td>
<td>34.41</td>
<td>37.792</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.1535</td>
<td>0.1888</td>
<td>0.2121</td>
<td>0.2228</td>
<td>0.2467</td>
<td>0.2892</td>
<td>0.3329</td>
<td>0.3658</td>
</tr>
</tbody>
</table>

Table 2: MLE and OBRE ($\gamma = 1$) for the Gamma parametrization of UK data

distribution, it misses the description of the bulk of the data, i.e. the data in the center. On the other hand, the OBRE concentrates on the majority of the data and therefore gives a better fit.

If one looks at the weights given by the OBRE to the data, one would see that they are low for classes corresponding to low and high incomes, thus confirming what we stated above.

It is difficult to advise a choice for the bound $c$. It depends on the problem and the type of data. As we have done, one should try different bounds and see how the OBRE behaves. For our example, we also computed the relative efficiency of the OBRE for the values of $c$ we have used. As given in (13), the asymptotic efficiency depends on the parameters. We therefore computed the asymptotic efficiency of the OBRE as a function of the shape parameter $\alpha$ of the Gamma distribution ($\delta = 0.2$) for different bound $c$. The results are presented in Figure 2.
Figure 1: Histogram of the empirical distribution and the classically and robustly estimated densities for the UK data.
7 Conclusion

In this paper we have studied robustness properties of MPE for grouped data. We have shown that although the IF of such estimators is bounded, it can be very large, meaning that the bias on the estimator under small model contaminations can be indeed large. We have then proposed a more general class of estimators that include more robust estimators, namely the class of MGP estimators. Within this class we have minimized the asymptotic covariance matrix under the constraint of bounded IF and have in this way defined optimal B-robust estimators for grouped data. We have also proposed an algorithm to find them. We have studied the local shift sensitivity of MPE and OBRE and have seen that in the latter case it is controlled. Finally, we have applied our results first in a simulation study and then to a real data set. The conclusion is that the OBRE is a lot more stable than the classical MPE under small departures from the assumed model.
Figure 2: Asymptotic efficiency of the OBRE as a function of the shape parameter $\alpha$ of the Gamma distribution ($\delta = 0.2$) for different bound $c$. 
Appendix

A Asymptotic distribution of MGP estimators

In this appendix we give a sketch of the proof of the asymptotic normality of MGP estimators. The techniques used are similar to those in Cressie and Read (1984) and Birch (1964). By a Taylor series expansion at $T_\psi = \theta$, we get

$$0 = \sum_{j=1}^{J} \psi_j(T_\psi) \left[ \int_{I_j} dF^{(n)}(x) \right]^\gamma = \sum_{j=1}^{J} \psi_j(\theta) \left[ \int_{I_j} dF^{(n)}(x) \right]^\gamma$$

$$+ (T_\psi - \theta) \sum_{j=1}^{J} \left[ \int_{I_j} dF^{(n)}(x) \right]^\gamma \frac{\partial}{\partial \theta} \psi_j(\theta) + O(\|T_\psi - \theta\|^2).$$

Hence,

$$T_\psi = \theta - \left[ \sum_{j=1}^{J} \frac{\partial}{\partial \theta} \psi_j(\theta)p_j^\gamma \right]^{-1} \sum_{j=1}^{J} \psi_j(\theta)p_j^\gamma + O(\|T_\psi - \theta\|^2).$$

(16)
Expanding $\psi_j(\theta)p_j^\gamma$ in a Taylor series around $k_j(\theta)$, we get

$$
\psi_j(\theta)p_j^\gamma = \psi_j(\theta)k_j(\theta)^\gamma + (p_j - k_j(\theta))\gamma \psi_j(\theta)k_j(\theta)^{\gamma-1} + O((p_j - k_j(\theta))^2)
$$

and summation over $j$ gives

$$
\sum_{j=1}^J \psi_j(\theta)p_j^\gamma = \gamma \sum_{j=1}^J (p_j - k_j(\theta))\psi_j(\theta)k_j(\theta)^{\gamma-1} + O(\|p - k(\theta)\|^2) \quad (17)
$$

On the other hand, by (4)

$$
\sum_{j=1}^J \frac{\partial}{\partial \theta} \psi_j(\theta)p_j^\gamma = \sum_{j=1}^J \frac{\partial}{\partial \theta} \psi_j(\theta)k_j(\theta)^\gamma + O(\|p - k(\theta)\|) \\
= -\gamma \sum_{j=1}^J \psi_j(\theta)k_j(\theta)^{\gamma-1} \frac{\partial}{\partial \theta} k_j(\theta) + O(\|p - k(\theta)\|) \quad (18)
$$

Since $p = k(\theta) + O_p(n^{-1/2})$ and

$$
\sqrt{n}(p - k(\theta)) \overset{D}{\rightarrow} N(0, diag(k_j(\theta)) - k(\theta)k^T(\theta))
$$

from (16), (17) and (18), we deduce that

$$
\sqrt{n}(T_{\psi\theta} - \theta) \overset{D}{\rightarrow} N(0, \Sigma_{MGP})
$$
where $\Sigma_{\text{MGP}}$ is given by (5).

B Optimal B-robust Estimators for Grouped Data

To prove the result on the OBRE for grouped data we need to write the problem in the following way

$$
\min_{\psi_1, \ldots, \psi_J} \text{tr} \left\{ \sum_{j=1}^{J} k_j(\theta)^{2\gamma-1} \psi_j(\theta) \psi_j(\theta)^T \right\}, \tag{19}
$$

subject to

$$
\forall j, \|k_j(\theta)^{\gamma-1} \psi_j(\theta)\| \leq c \tag{20}
$$

$$
\sum_{j=1}^{J} k_j(\theta)\gamma \psi_j(\theta) = 0 \tag{21}
$$

$$
\sum_{j=1}^{J} k_j(\theta)\gamma \psi_j(\theta) \frac{\partial}{\partial \theta^T} \log k_j(\theta) = I. \tag{22}
$$

There is no loss of generality in setting (22) because $\psi_j(\theta)$ is determined only up to a multiplicative matrix; cf (3).

Solving the problem (19) under the constraints (21) and (22) is equivalent
to minimize the Lagrange-function given by

\[
\sum_{j=1}^{J} \int_{I_j} k_j(\theta)^{\gamma-1} \psi_j(\theta)^T \psi_j(\theta) k_j(\theta)^{\gamma-1} dF_\theta(x) \\
- \mathbf{a}^*^T \sum_{j=1}^{J} \int_{I_j} k_j(\theta)^{\gamma-1} \psi_j(\theta) dF_\theta(x) \\
- \mathbf{a}^* \sum_{j=1}^{J} \int_{I_j} k_j(\theta)^{\gamma-1} \psi_j(\theta)^T dF_\theta(x) \\
- \sum_{j=1}^{J} \int_{I_j} k_j(\theta)^{\gamma-1} \psi_j(\theta) \frac{\partial}{\partial \theta^T} \log k_j(\theta) dF_\theta(x) \cdot \mathbf{A}^*^T + \mathbf{A}^*^T \\
- \mathbf{A}^* \sum_{j=1}^{J} \int_{I_j} k_j(\theta)^{\gamma-1} \frac{\partial}{\partial \theta} \log k_j(\theta) \psi_j(\theta)^T dF_\theta(x) + \mathbf{A}^* \\
= \sum_{j=1}^{J} \int_{I_j} \left[ k_j(\theta)^{\gamma-1} \psi_j(\theta) - \mathbf{a}^* - \mathbf{A}^* \frac{\partial}{\partial \theta^T} \log k_j(\theta) \right]^T . \\
\left[ k_j(\theta)^{\gamma-1} \psi_j(\theta) - \mathbf{a}^* - \mathbf{A}^* \frac{\partial}{\partial \theta} \log k_j(\theta) \right] dF_\theta(x) \\
+ B \left[ \mathbf{A}^*, \mathbf{a}^*, k_j(\theta), \frac{\partial}{\partial \theta} k_j(\theta) \right] ,
\]

where the vector \(\mathbf{a}^*\) and the matrix \(\mathbf{A}^*\) are the Lagrange multipliers and \(B\) is a function that does not depend on \(\psi_j\). We see then that the optimal functions \(\psi_j\) are of the form

\[
\frac{1}{k_j(\theta)^{\gamma}} \mathbf{A}(\theta) \left[ \frac{\partial}{\partial \theta^T} k_j(\theta) - \mathbf{a}(\theta) k_j(\theta) \right] ,
\]

where \(\mathbf{A}(\theta)\) and \(\mathbf{a}(\theta)\) are determined implicitly by (21) and (22). If we add
the constraint (20) on the IF, we can see that we have to multiply these functions by the weights given by

$$
\min \left\{ 1 ; \frac{c}{\| A(\theta) \left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - a(\theta) \right] \|} \right\} .
$$

The optimal functions $\psi_j$’s are then given by

$$
\psi_j(\theta) = \frac{1}{k_j(\theta)} A(\theta) \left[ \frac{\partial}{\partial \theta} k_j(\theta) - a(\theta) k_j(\theta) \right] \cdot \\
\min \left\{ 1 ; \frac{c}{\| A(\theta) \left[ \frac{\partial}{\partial \theta} \log k_j(\theta) - a(\theta) \right] \|} \right\} .
$$

Replacing (23) in (19), (21) and (22) leads to the solution given in (9), (11) and (12).
References


