J. Japan Statist. Soc.
Vol. 38 No. 1 2008 131–143

EFFICIENT ESTIMATION AND MODEL SELECTION FOR GROUPED DATA WITH LOCAL MOMENTS

Kohtaro Hitomi*, Qing-Feng Liu**, Yoshihiko Nishiyama** and Naoya Sueishi***

This paper proposes efficient estimation methods of unknown parameters when frequencies as well as local moments are available in grouped data. Assuming the original data is an i.i.d. sample from a parametric density with unknown parameters, we obtain the joint density of frequencies and local moments, and propose a maximum likelihood (ML) estimator. We further compare it with the generalized method of moments (GMM) estimator and prove these two estimators are asymptotically equivalent in the first order. Based on the ML method, we propose to use the Akaike information criterion (AIC) for model selection. Monte Carlo experiments show that the estimators perform remarkably well, and AIC selects the right model with high frequency.

Key words and phrases: AIC, GMM, grouped data, local moments, MLE, model selection.

1. Introduction

In practice, some data are provided only in a grouped form. An example is personal income data reported by government organizations. They provide only masked data for reasons of confidentiality. Typically, an income distribution is divided into classes (by age, for instance) and only summary statistics such as frequencies and class-wise means are observable to researchers. Also, we often see insurance claim data in the same form. Researchers cannot directly observe the claim sizes of each accident, but only the summaries for each stratum are available. It is also possible to model this by some discrete distribution, however regular discrete distributions may not always be appropriate. Throughout this paper, we assume that the "original" sample $\{x_i\}$, $i = 1, \ldots, n$, which is unavailable for statisticians, is a realization of a random sample $\{X_i\}$, $i = 1, \ldots, n$ from a d-dimensional distribution with parametric density $f(x; \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ is a vector of unknown parameters. We also suppose that the bounds of each stratum are non-random.

A common and classical situation is the case when only the frequencies are available. Suppose that the support of X_1 is divided into a set of fixed disjoint classes $B_1, B_2, \ldots, B_L$, and we observe only the frequency n_j in each of the classes B_j. Since the individual data are not available, the following standard maximum

Accepted March 28, 2008.

*Kyoto Institute of Technology, Japan.
**Kyoto University, Japan.
***University of Wisconsin-Madison, U.S.A.

likelihood estimator is infeasible:

$$\hat{\boldsymbol{\theta}}_{iMLE} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}) \tag{1.1}$$

even though we know the explicit form of the density. The subscript iMLE indicates infeasible maximum likelihood estimator. In this case, however, we easily obtain the log-likelihood function with respect to n_j, $j = 1, \ldots, L$, which equals to $\sum_{j=1}^{L} n_j \log P_j(\boldsymbol{\theta})$, where $P_j(\boldsymbol{\theta}) = \int_{B_j} f(x; \boldsymbol{\theta}) dx$ is the probability that an observation falls in B_j. This gives an MLE of $\boldsymbol{\theta}$ as a solution to the normal equation:

$$\sum_{j=1}^{L} n_j \frac{\partial \log P_j(\hat{\boldsymbol{\theta}}_{nMLE})}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{1.2}$$

We call it the naive MLE (nMLE). Asymptotic properties of the nMLE have been examined in several papers (see, for example, Lindley (1949) and Tallis (1967)). It is consistent for θ_0, the true value of θ, and asymptotically normally distributed with covariance matrix $-\{\sum_{j=1}^{L} P_j(\theta_0) \frac{\partial^2 \log P_j(\theta_0)}{\partial \theta \partial \theta'}\}^{-1}$. Victoria-Feser and Ronchetti (1997) discuss the properties of the nMLE and some related estimators in terms of robustness. An estimator that is easy to compute is proposed by Brix and Pfeifer (2000) which does not require explicitly evaluating $P_j(\theta)$. See also Yanagimoto (1990) and Wooldridge (2001) which treat related topics.

In this paper, we consider the situation in which local moments in each class, namely,

$$\bar{x}_j^{(k)} = \frac{1}{n_j} \sum_{i=1}^{n} I(x_i \in B_j) x_i^k, \qquad i = 1, \ldots, n, \quad k = 1, \ldots, K$$

are also available in addition to the frequencies n_j, where $I(\cdot)$ denotes the indicator function. The purpose of this paper is to propose efficient estimation methods in this setup, and a model selection method. Though we believe we can proceed for any integer K satisfying $E(X_1^K) < \infty$, we only discuss the case when $K = 1$. It is partly because we do not know of any data sets reporting local higher order moments. We derive two forms of the joint density of frequencies and local means, then propose MLE estimators. They are proved to be consistent, asymptotically normally distributed and efficient. In the sequel, we denote $N_j = \sum_{i=1}^{n} I(X_i \in B_j)$, $n_j = \sum_{i=1}^{n} I(x_i \in B_j)$, $\bar{X}_j = \sum_{i=1}^{n} I(X_i \in B_j) X_i / N_j$, and $\bar{x}_j = \sum_{i=1}^{n} I(x_i \in B_j) x_i / n_j$.

We can compare alternative parametric models by AIC criteria (see Akaike (1973, 1974)). It is easy to compute and a practically convenient criteria in the context of maximum likelihood, and hence it has been widely used by empirical researchers. We apply this criteria to the analysis of grouped data.

This paper is organized as follows. In Section 2 we give the MLE of θ and present its asymptotics. Section 3 discusses model selection for this problem. Section 4 shows results from Monte Carlo experiments. Section 5 concludes. All proofs are in the Appendix.

2. Efficient estimation by MLE

This section provides maximum likelihood estimators and their asymptotic properties when the frequencies and local means are available.

2.1. Estimators

We can use $(n_j, n_j\bar{x}_j)$, $j = 1, \ldots, L$, only, not the individual observations $\{x_i\}$, $i = 1, \ldots, n$. The best we can do is to obtain the likelihood function with respect to the frequencies and local means, and maximize it with respect to θ. We obtain two forms of the joint density of $(N_j, N_j\bar{X}_j)$, $j = 1, \ldots, L$. The first expression involves a distribution of sums of n_j independent random variables, and thus a n_j-fold convolution of $I(x \in B_j)f(x;\theta)$ conditional on $N_j = n_j$,

$$l(\theta) = \sum_{j=1}^{L} \log f^{(n_j)}(n_j\bar{x}_j;\theta), \tag{2.1}$$

where

$$f^{(n_j)}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} I(x - y_1 - \cdots - y_{n_j} \in B_j) f(x - y_1 - \cdots - y_{n_j};\theta) \times \prod_{k=1}^{n_j} I(y_k \in B_j) f(y_k;\theta) dy_k.$$

We rewrite it in the form of an inverse Fourier transform of a characteristic function which includes only two integrations,

$$l(\theta) = \sum_{j=1}^{L} \log \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iun_j\bar{x}_j} \left\{ \int_{-\infty}^{\infty} e^{iux} I(x \in B_j) f(x;\theta) dx \right\}^{n_j} du \right]. \tag{2.2}$$

We show later in the Theorems that it is approximated by a simpler expression,

$$\frac{1}{n}\bar{l}(\theta) = \sum_{j=1}^{L} \left[-\frac{n_j\{\bar{x}_j - \mu_j(\theta)\}^2}{2nV_j(\theta)} + \frac{n_j}{n} \log P_j(\theta) \right] \tag{2.3}$$

where Y_j is a random variable with a density $f_j(y) = I(y \epsilon B_j)f(y;\theta)/P_j(\theta)$, $\mu_j(\theta) = E(Y_j)$, $V_j(\theta) = \mathrm{Var}(Y_j)$.

Given the above likelihood functions, we can estimate parameter θ by the maximum likelihood method. There are three possibilities, namely maximizing either of (2.1), (2.2) or (2.3). The first two are of course equivalent, thus we can think of:

$$\theta_{ML} = \arg\max_{\theta} l(\theta), \qquad \bar{\theta}_{ML} = \arg\max_{\theta} \bar{l}(\theta).$$

Using the three forms of likelihood functions are shown to be asymptotically equivalent later. Among them, (2.3) will practically be the most convenient computationally in most cases.

2.2. Asymptotic properties of MLE

We first state the assumptions.

ASSUMPTION 1. (i) If $P_j(\theta) = P_j(\theta_0)$ for all $j = 1, \ldots, n$, then $\theta = \theta_0$ and no other $\theta \in \Theta$ satisfies $P_j(\theta) = P_j(\theta_0)$, where the parameter space Θ is a compact subset of R^p.

(ii) $\mu_j(\theta)$ and $V_j(\theta)$ exist for all $j = 1, \ldots, L$ and $\forall\theta \in \Theta$.

This assumption guarantees the identification.

ASSUMPTION 2. (i) $\mu_j(\theta)$ and $V_j(\theta)$ are continuous in θ for all $j = 1, \ldots, L$.

(ii) There exist $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that $\inf_{\theta\in\Theta} P_j(\theta) \geq \epsilon_1$ and $\inf_{\theta\in\Theta} V_j(\theta) \geq \epsilon_2$ for all $j = 1, \ldots, L$.

(iii) $f_j(x;\theta)$ is continuous in x in the neighborhood of $P_j(\theta_0)\mu_j(\theta_0)$ uniformly in θ for all $j = 1, \ldots, L$.

(iv) $\max_j \sup_{\theta\in\Theta} E|Y_j|^3 < \infty$.

ASSUMPTION 3. (i) θ_0 is an interior point of Θ.

(ii) $\mu_j(\theta)$, $V_j(\theta)$ and $\log P_j(\theta)$ are twice continuously differentiable in the neighborhood $\mathcal{N}$ of θ_0 for all $j = 1, \ldots, L$.

(iii) The information matrix

$$I(\theta_0) = \sum_{j=1}^{L} \left\{ \frac{P_j(\theta_0)}{V_j(\theta_0)} \frac{\partial\mu_j(\theta_0)}{\partial\theta} \frac{\partial\mu_j(\theta_0)}{\partial\theta'} - P_j(\theta_0) \frac{\partial^2 \log P_j(\theta_0)}{\partial\theta\partial\theta'} \right\}$$

is nonsingular.

The following theorems state the asymptotic properties of these estimators.

THEOREM 1. *Suppose Assumptions* 1 *and* 2 *hold, then*

(i) *The approximation error of* $\bar{l}(\theta)$ *to* $l(\theta)$ *is*

$$\frac{1}{n}\{\bar{l}(\theta) - l(\theta)\} = O_p\left(\frac{\log n}{n}\right) \tag{2.4}$$

uniformly in $\theta\epsilon\Theta$.

(ii) $\hat{\theta} \xrightarrow{p} \theta$, *where* $\hat{\theta} = \theta_{ML}, \bar{\theta}_{ML}$.

THEOREM 2. *Suppose Assumption* 3 *holds in addition to the assumptions in Theorem* 1, *then*:

(i) *The approximation error of* $\partial\bar{l}(\theta)/\partial\theta$ *to* $\partial l(\theta)/\partial\theta$ *is*

$$\frac{1}{\sqrt{n}} \left\{ \frac{\partial l(\theta)}{\partial\theta} - \frac{\partial\bar{l}(\theta)}{\partial\theta} \right\} = o_p(1). \tag{2.5}$$

(ii) *We also have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1}).$$

(2.5) guarantees that $\sqrt{n}(\theta_{ML} - \bar{\theta}_{ML}) = o_p(1)$ (see Robinson (1988), Theorem 1) so that these two estimators share the same asymptotic distribution. We note that the information matrix corresponding to the naive estimator is $-\sum_{j=1}^{L} P_j(\theta_0) \frac{\partial^2 \log P_j(\theta_0)}{\partial\theta\partial\theta'}$ which equals to the second term of the information matrix. The first term $\sum_{j=1}^{L} \frac{P_j(\theta_0)}{V_j(\theta_0)} \frac{\partial\mu_j(\theta_0)}{\partial\theta} \frac{\partial\mu_j(\theta_0)}{\partial\theta'}$ presents the efficiency gain associated with the local mean information.

Sueishi *et al.* (2006) considered a GMM estimator for the same model. It also has the same asymptotic variance.

3. Model selection

In the previous sections, we propose efficient estimation methods for the parameter θ when the underlying density is assumed to be $f(x;\theta)$. In some cases, there may exist a number of possible candidates of $f(x;\theta)$. In examining income distribution, econometricians often fit the data to log-normal, χ^2, or Pareto distributions. Researchers would like to know which is the most suitable distribution.

One criteria in choosing a suitable model is the AIC when maximum likelihood estimation is possible. Using the maximum log-likelihood $l(\hat{\theta}_{ML})$ or $\bar{l}(\hat{\theta}_{ML})$, we can construct,

$$\begin{aligned} \text{AIC} &= -2l(\hat{\theta}_{ML}) + 2p \\ &\approx -2\bar{l}(\hat{\theta}_{ML}) + 2p. \end{aligned}$$

In the current context, the meaning of AIC is slightly different from the usual one because of the unavailability of the "original" data $\{x_i\}$, $i = 1, \ldots, n$. AIC was introduced as an estimate of $-\int g(x) \log f(x;\theta) dx$ where $g(x)$ is the true density of X_1. Here, we can observe only n_j, $\bar{x}_j$, $j = 1, \ldots, L$ and thus we cannot compute

$$\text{AIC} = -2 \sum_{i=1}^{n} \log f(Y_i; \hat{\theta}_{ML}) + 2 \times (\#\text{ of parameters}).$$

Instead, the best we can do is to construct AIC in terms of the feasible log-likelihood functions (2.1)–(2.3). Therefore, we do not directly compare $f(x;\theta)$ and $g(x)$, but through the Kullback-Leibler distance between $\Pi_{j=1}^{L} f^{(n_j)}(n_j\bar{x}_j;\theta)$ and the true joint density of N_j, $N_j\bar{X}_j$, $j = 1, \ldots, L$. This seems to work satisfactory in view of the Monte Carlo results reported in the next section.

4. Monte Carlo results

We carry out Monte Carlo simulations to examine the small sample performances of the MLE and GMM estimators, as well as the AIC model selection proposed in the previous sections. In Subsection 4.1, we compare the iMLE (1.1), nMLE (1.2), MLE and GMM in terms of the mean squared error (MSE). iMLE is obviously the efficiency benchmark. The GMM estimator is calculated

Table 1. Efficiency comparison of estimators: relative ratio of MSE.

		iMLE	nMLE	GMM	MLE
$n = 100$					
	μ	[0.0886]	1.1109	1.0894	0.9961
	σ	[0.0431]	2.0608	1.4116	1.0221
$n = 1000$					
	μ	[0.0097]	1.0952	1.0066	1.0001
	σ	[0.0040]	1.7898	1.0233	1.0044

using the algorithm proposed by Sueishi *et al.* (2006). Subsection 4.2 provides results of sample selection by the AIC. The number of replications is $1,000$ for all experiments.

4.1. Comparison of estimators

The data are generated with samples of size $n = 100$, 1000 from $N(0, 9)$. Setting the bounds of grouping $(-30, -3, -1, 1, 3, 30)$, so that $L = 7$, we classify each observation accordingly and compute the u and the class means for each sample. The results are tabulated in Table 1. For iMLE, the level of MSE is reported, while the relative e to that of iMLE are provided for other estimators. Firstly, the efficiency is improved when the sample size increases for all estimation methods.

We immediately realize that the relative MSEs for μ are quite close to unity for all estimators, though the MLE seems to slightly outperform the others. In the case of the GMM and MLE, it may be obvious that the iMLE of μ is simply the mean of the sample, and we can compute it in fact from the local means and frequencies of each class in our hands. For the estimation of σ, we find differences in the efficiencies. nMLE is apparently less efficient than the others, meaning that the local mean information significantly increases the efficiency of estimation. Also we point out that the efficiency gain in the nMLE by increased sample size is not as large as the others. It seems MLE performs better than GMM though the asymptotic variances are the same in theory. We think this is partly because we compute the weight matrix by a bootstrap method using the nMLE as the pilot estimate, so that it is a kind of two-step procedure which may have some bias. If we use a continuous updating GMM instead, the performance may be improved. We find the MLE performs very well, almost equivalent to the iMLE where all the observations are available in this setup.

4.2. Model selection

We also study the the performance of model selection by the AIC. We consider two distributions, $N(0, 9)$ and a mixture of $N(0, 9)$ and Exp(1/2) or exponential distribution with parameter 1/2. Table 2 reports how many times each model is selected out of 1000 replications when the data is generated from the mixture for $n = 100$, 1000. Table 3 gives the same numbers when the true distribution is $N(0, 9)$. Table 2 show that the AIC chooses the correct model all

Table 2. Number of selections when mixture is true.

n	$N(0,9)$	Mixture (true)
100	144	856
1000	0	1000

Table 3. Number of selections when $N(0,3)$ is true.

n	$N(0,9)$(true)	Mixture
100	981	19
1000	980	20

the time when $n = 1000$, while not so good for smaller sample sizes. In view of Table 3, the number of wrong choices from the model is always about 20. Taking into account that the mixture nests the true distribution $N(0,9)$, it is not surprising that the mixture is selected. It may be considered outstanding that the AIC do not select unnecessarily complex models.

5. Conclusion

We applied the maximum likelihood principle to grouped data analysis when we can obtain not only counts but also local moments for each group. It provides asymptotically efficient estimates of the parameters. We carry out Monte Carlo experiments to investigate the performance of the estimator. Simulation results show that it performs remarkably well even in finite samples. Comparing with the nMLE, which does not use the local moments information, we can see they provide a significant efficiency improvement.

In the context of maximum likelihood estimation, we propose to apply the AIC model selection criterion. It shows a superb performance for the present problem. Our results suggest that local moments in grouped data could be highly informative.

Appendix A: Mathematical proofs

We give brief proofs for the results in Section 2.

PROOF OF THEOREM 1. (i) Let γ_j be

$$\gamma_j(u) = E(e^{iuY_j}) = P_j(\theta)^{-1} \int_{-\infty}^{\infty} e^{iux} I(x \epsilon B_j) f(x;\theta) dx. \tag{A.1}$$

Setting $u = t/n$, write

$$E(e^{i(t/n)N_j \bar{X}_j} \mid N_j = n_j) = \left\{ \gamma_j \left(\frac{t}{n} \right) \right\}^{n_j} P_j(\theta)^{n_j}. \tag{A.2}$$

Because

$$\log \left\{ \gamma_j \left(\frac{t}{n} \right) \right\}^{n_j} = n_j \log \left\{ \gamma_j \left(\frac{t}{n} \right) \right\} = n_j \log E \left\{ \exp \left(it \frac{Y_j}{n} \right) \right\}$$

$$= n_j \log\left\{1 + \frac{it}{n}E(Y_j) - \frac{t^2}{2n^2}E(Y_j^2) + O\left(\frac{|t|^3 E(Y_j^3)}{n^3}\right)\right\}$$

$$= n_j \left[\frac{it}{n}\mu_j(\theta) - \frac{t^2}{2n^2}E(Y_j^2) - \frac{1}{2}\left\{\frac{it}{n}\mu_j(\theta) - \frac{t^2}{2n^2}E(Y_j^2)\right\}^2 + O\left(\frac{|t|^3 E(Y_j^3)}{n^3}\right)\right]$$

$$= n_j \left\{\frac{it}{n}\mu_j(\theta) - \frac{t^2}{2n^2}V_j(\theta) + O\left(\frac{|t|^3 E(Y_j^3)}{n^3}\right)\right\},$$

we have

$$\text{(A.3)} \quad \left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j} = \exp\left\{\frac{itn_j}{n}\mu_j(\theta) - \frac{t^2 n_j}{2n^2}V_j(\theta) + O\left(\frac{|t|^3 n_j E(Y_j^3)}{n^3}\right)\right\}.$$

(A.2) and (A.3) yield

$$\left\{\int_{-\infty}^{\infty} e^{i(t/n)x} I(x \in B_j) f(x \mid \theta) dx\right\}^{n_j}$$

$$\text{(A.4)} \qquad = P_j(\theta)^{n_j} \exp\left\{\frac{itn_j}{n}\mu_j(\theta) - \frac{t^2 n_j}{2n^2}V_j(\theta) + O\left(\frac{|t|^3 n_j E(Y_j^3)}{n^3}\right)\right\}.$$

Plugging (A.4) into (2.2), we write, noting the transformation $u = t/n$,

$$\frac{l(\theta)}{n} = \frac{1}{n}\sum_{j=1}^{L} \log\left[\frac{P_j(\theta)^{n_j}}{2\pi n}\int_{-\infty}^{\infty} \exp\{A_j(t;\theta)\}dt\right] + R_n$$

where

$$A_j(t;\theta) = -it\frac{n_j}{n}\{\bar{x}_j - \mu_j(\theta)\} - \frac{t^2 n_j}{2n^2}V_j(\theta),$$

$$R_n = \frac{1}{n}\sum_{j=1}^{L} \log\left[\frac{\int_{-\infty}^{\infty} e^{-i(t/n)n_j\bar{x}_j}\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j} dt}{P_j(\theta)^{n_j}\int_{-\infty}^{\infty}\exp\{A_j(t;\theta)\}dt}\right].$$

Because

$$\int_{-\infty}^{\infty}\exp\{A_j(t;\theta)\}dt = \int_{-\infty}^{\infty}\exp\left\{-it\frac{n_j}{n}\{\bar{x}_j - \mu_j(\theta)\} - \frac{t^2 n_j}{2n^2}V_j(\theta)\right\}dt$$

$$= \sqrt{\frac{2\pi n^2}{n_j V_j(\theta)}}\exp\left[-\frac{n_j\{\bar{x}_j - \mu_j(\theta)\}^2}{2V_j(\theta)}\right],$$

we have

$$\frac{l(\theta)}{n} = \frac{1}{n}\sum_{j=1}^{L}\left[-\frac{n_j\{\bar{x}_j - \mu_j(\theta)\}^2}{2V_j(\theta)} - \frac{1}{2}\left\{\log\left(\frac{n_jV_j(\theta)}{n}\right)\right\} + n_j \log P_j(\theta)\right]$$

$$- \frac{L}{2n}\log(2\pi n) + R_n. \tag{A.5}$$

Now we evaluate the residual term. Write

$$R_n = \frac{1}{n}\sum_{j=1}^{L}\log\left[1 + \frac{\int_{-\infty}^{\infty}\left[e^{-i(t/n)n_j\bar{x}_j}\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j} - P_j(\theta)^{n_j}\exp A_j(t;\theta)\right]dt}{\int_{-\infty}^{\infty}P_j(\theta)^{n_j}\exp A_j(t;\theta)dt}\right].$$

Using Assumption 2,

$$\left|\int_{-\infty}^{\infty} e^{-i(t/n)n_j\bar{x}_j}\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j}dt\right| = f_{Y_j}\left(\frac{n_j\bar{x}_j}{n}\right) < \infty,$$

$$\left|\int_{-\infty}^{\infty}\exp\{A_j(t;\theta)\}dt\right| < \infty,$$

uniformly in θ as $n \to \infty$. Therefore, $\int_{-\infty}^{\infty} R_{nj}(t;\theta)dt$ exists where

$$R_{nj}(t;\theta) = e^{-i(t/n)n_j\bar{x}_j}\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j} - P_j(\theta)^{n_j}\exp A_j(t;\theta).$$

Because of the existence, it equals to its principal value which is, for $\epsilon > 0$,

$$\int_{-\infty}^{\infty} R_{nj}(t;\theta)dt = P.V.\int_{-\infty}^{\infty} R_{nj}(t;\theta)dt = \lim_{n\to\infty}\int_{-n^\epsilon}^{n^\epsilon} R_{nj}(t;\theta)dt.$$

We show this integral converges to zero in probability uniformly in θ. Note

$$\left|\lim_{n\to\infty}\int_{-n^\epsilon}^{n^\epsilon} R_{nj}(t;\theta)dt\right| \le \lim_{n\to\infty}\int_{-n^\epsilon}^{n^\epsilon}|R_{nj}(t;\theta)|dt,$$

and

$$\left|\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j} - \left\{it\frac{n_j}{n}\mu_j(\theta) - \frac{t^2n_j}{2n^2}V_j(\theta)\right\}\right| \le C|t|^3$$

for a generic positive constant C, as $E|Y_j|^3 = \int |x|^3 I(x\epsilon B_j)f(x;\theta)dx \le \infty$ uniformly in θ by Assumption 2 (iii). Then, using the inequality $|e^{ix} - 1 - ix - (ix)^2/2| \le |t|^3/6$, and $0 < P_j(\theta)^{n_j} \le 1$ for any $\theta\epsilon\Theta$, we have,

$$\int_{-n^\epsilon}^{n^\epsilon}|R_{nj}(t;\theta)|dt \le \int_{-n^\epsilon}\frac{n_j\left|\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j} - \left\{it\frac{n_j}{n}\mu_j(\theta) - \frac{t^2n_j}{2n^2}V_j(\theta)\right\}\right|}{n^3}dt$$

$$\le \frac{Cn_j}{n}n^{4\epsilon-2}.$$

Setting $\epsilon < 1/4$, we obtain $R_n = o_p(n^{-1})$ uniformly in θ, and therefore by (A.5),

$$\frac{l(\theta)}{n} = \frac{\bar{l}(\theta)}{n} + O_p\left(\frac{\log n}{n}\right)$$

uniformly in $\theta \epsilon \Theta$, because $n_j/n \xrightarrow{p} P_j(\theta_0)$ and $\bar{x}_j \xrightarrow{p} \mu_j(\theta_0)$ by the weak law of large numbers.

(ii) Because of (i) above, it suffices to show that $\frac{\bar{l}(\theta)}{n} \xrightarrow{p} Q(\theta)$ uniformly in $\theta \epsilon \Theta$ and $\theta_0 = \max_\theta Q(\theta)$. Due to the weak law of large numbers, we have $n_j/n \xrightarrow{p} P_j(\theta_0)$ and $\bar{x}_j \xrightarrow{p} \mu_j(\theta_0)$. Therefore,

$$\begin{aligned}\frac{1}{n}\bar{l}(\theta) &= \sum_{j=1}^{L}\left[-\frac{n_j\{\bar{x}_j - \mu_j(\theta)\}^2}{2nV_j(\theta)} + \frac{n_j}{n}\log P_j(\theta)\right] \\ &\xrightarrow{p} Q(\theta) = \sum_{j=1}^{L}\left[-\frac{P_j(\theta_0)\{\mu_j(\theta_0) - \mu_j(\theta)\}^2}{2nV_j(\theta)} + P_j(\theta_0)\log P_j(\theta)\right].\end{aligned}$$

This convergence is obviously uniform in θ by Assumptions 1 and 2. It is also obvious that $Q(\theta) \leq \sum_{j=1}^{L} P_j(\theta_0)\log P_j(\theta_0)$ and the equality holds only when $\theta = \theta_0$.

PROOF OF THEOREM 2. (i) The proof proceeds similarly to the proof of Theorem 1 (i). Since evaluating all the remainder terms makes the expression too tedious, we only pick up first order terms. The detailed proof for the validation is available from the authors upon request. Write

$$f_j^{(n_j)}(n_j, n_j\bar{x}_j; \theta) = \frac{1}{2\pi n}\int_{-\infty}^{\infty} e^{-i(t/n)n_j\bar{x}_j}\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j} dt,$$

then the log-likelihood is written as

$$l(\theta) = \sum_{j=1}^{L}[\log f_j^{(n_j)}(n_j, n_j\bar{x}_j; \theta) + n_j \log P_j(\theta)]$$

and the k-th element of the score is

$$\frac{1}{\sqrt{n}}\frac{\partial l(\theta)}{\partial \theta_k} = \sum_{j=1}^{L}\left[\frac{1}{\sqrt{n} f_j^{(n_j)}(n_j, n_j\bar{x}_j; \theta)}\frac{\partial f_j^{(n_j)}(n_j, n_j\bar{x}_j; \theta)}{\partial \theta_k} + \frac{N_j}{\sqrt{n}}\frac{\partial \log P_j(\theta)}{\partial \theta_k}\right]$$

where

$$\frac{\partial f_j^{(n_j)}(n_j, n_j\bar{x}_j; \theta)}{\partial \theta_k} = \frac{n_j}{2\pi n}\int_{-\infty}^{\infty} e^{-i(t/n)n_j\bar{x}_j}\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j-1}\frac{\partial \gamma_j\left(\frac{t}{n};\theta\right)}{\partial \theta_k} dt.$$

We can show after some tedious but straightforward algebra that,

$$\frac{\partial \gamma_j\left(\frac{t}{n};\theta\right)}{\partial\theta_k} \approx -\gamma_j\left(\frac{t}{n}\right) + \gamma_j\left(\frac{t}{n}\right)\left\{1+\frac{it}{n}\frac{\partial\mu_j(\theta)}{\partial\theta_k}\right\},$$

so that we have

$$\frac{\partial f_j^{(n_j)}(n_j, n_j\bar{x}_j;\theta)}{\partial\theta_k} = -n_j f_j^{(n_j)}(n_j, n_j\bar{x}_j;\theta) + \frac{n_j}{2\pi n}\int_{-\infty}^{\infty} e^{-i(t/n)n_j\bar{x}_j}\left\{\gamma_j\left(\frac{t}{n}\right)\right\}^{n_j}\left\{1+\frac{it}{n}\frac{\partial\mu_j(\theta)}{\partial\theta_k}\right\}dt.$$

Using approximation $\exp\{\frac{it}{n}\frac{\partial\mu_j(\theta)}{\partial\theta_k}\} \approx 1+\frac{it}{n}\frac{\partial\mu_j(\theta)}{\partial\theta_k}$, the integral above is shown to be approximated by

$$\sqrt{\frac{2\pi n^2}{n_j V_j(\theta)}}\exp\left[-\frac{\left\{n_j(\bar{x}_j-\mu_j(\theta))+\frac{\partial\mu_j(\theta)}{\partial\theta_k}\right\}^2}{2n_jV_j(\theta)}\right].$$

This gives,

$$\begin{aligned}
&\frac{1}{\sqrt{n}f_j^{(n_j)}(n_j, n_j\bar{x}_j;\theta)}\frac{\partial f_j^{(n_j)}(n_j, n_j\bar{x}_j;\theta)}{\partial\theta_k} \\
&\quad\approx \frac{n_j\exp\left[-\frac{\left\{n_j(\bar{x}_j-\mu_j(\theta))+\frac{\partial\mu_j(\theta)}{\partial\theta_k}\right\}^2}{2n_jV_j(\theta)}\right]}{\sqrt{n}\exp\left[-\frac{n_j\{\bar{x}_j-\mu_j(\theta)\}^2}{2V_j(\theta)}\right]} - n_j \\
&\quad= \frac{n_j\{\bar{x}_j-\mu_j(\theta)\}}{\sqrt{n}V_j(\theta)}\frac{\partial\mu_j(\theta)}{\partial\theta_k} - \frac{1}{2\sqrt{n}V_j(\theta)}\left\{\frac{\partial\mu_j(\theta)}{\partial\theta_k}\right\}^2 + o_p(1).
\end{aligned}$$

Therefore, we have

$$\frac{1}{\sqrt{n}}\left\{\frac{\partial l(\theta)}{\partial\theta}-\frac{\partial\bar{l}(\theta)}{\partial\theta}\right\} = o_p(1)$$

uniformly in θ.

(ii) We show that $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\bar{l}(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$ and $-\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta'}\bar{l}(\theta_0) \xrightarrow{p} I(\theta_0)^{-1}$. Firstly, replacing n_j and $\bar{x}_j$ by N_j and $\bar{X}_j$,

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\bar{l}(\theta_0) = \sum_{j=1}^{L}\left[\frac{N_j\sqrt{n}\{\bar{X}_j-\mu_j(\theta_0)\}}{nV_j(\theta_0)}\frac{\partial\mu_j(\theta_0)}{\partial\theta} + \frac{N_j}{\sqrt{n}}\frac{\partial\log P_j(\theta_0)}{\partial\theta}\right].$$

Using $N_j = \sum_{i=1}^{n} I(X_i \epsilon B_j)$ and $N_j \bar{X}_j = \sum_{i=1}^{n} I(X_i \epsilon B_j) X_i$, we can rewrite it as

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \bar{l}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\sum_{j=1}^{L} W_{ji} \right),$$

where

$$W_{ji} = \frac{\{X_i - \mu_j(\theta_0)\} I(X_i \epsilon B_j)}{V_j(\theta_0)} \frac{\partial \mu_j(\theta_0)}{\partial \theta} + I(X_i \epsilon B_j) \frac{\partial \log P_j(\theta_0)}{\partial \theta}.$$

It is straightforward to show that $\sum_{j=1}^{L} W_{ji}$ are i.i.d. with mean zero and variance $I(\theta_0)$. Therefore, due to a central limit theorem,

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \bar{l}(\theta_0) \xrightarrow{d} N(0, I(\theta_0)).$$

Secondly, we obtain the information matrix as follows using the weak law of large numbers:

$$\begin{aligned} &-\frac{1}{n} \frac{\partial^2}{\partial \theta \partial \theta'} \bar{l}(\theta_0) \\ &\quad = -\sum_{j=1}^{L} \left[\frac{n_j \{\bar{x}_j - \mu_j(\theta_0)\}}{n V_j(\theta_0)} \frac{\partial^2 \mu_j(\theta_0)}{\partial \theta \partial \theta'} - \frac{n_j}{n V_j(\theta_0)} \frac{\partial \mu_j(\theta_0)}{\partial \theta} \frac{\partial \mu_j(\theta_0)}{\partial \theta'} \right. \\ &\qquad \left. + \frac{n_j}{n} \frac{\partial^2 \log P_j(\theta_0)}{\partial \theta \partial \theta'} \right] \\ &\quad \xrightarrow{p} \sum_{j=1}^{L} \left\{ \frac{P_j(\theta_0)}{V_j(\theta_0)} \frac{\partial \mu_j(\theta_0)}{\partial \theta} \frac{\partial \mu_j(\theta_0)}{\partial \theta'} - P_j(\theta_0) \frac{\partial^2 \log P_j(\theta_0)}{\partial \theta \partial \theta'} \right\} = I(\theta_0). \end{aligned}$$

Acknowledgements

We would like to thank an anonymous referee, Yoshinori Kawasaki, Atsushi Yoshida, Nakahiro Yoshida, seminar participants at Kyoto University and participants for the JEA annual meeting at Kyoto Sangyo University and Workshop on Stochastic Analysis and Statistical Inference II at Hitotsubashi University for helpful comments. This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grand-in-Aid for 21st Century COE Program "Interfaces for Advanced Economic Analysis".

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium of in Information Theory* (eds. B. N. Petrov and F. Csaki), 267–281, Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Brix, J. and Pfeifer, D. (2000). A simple method to estimate parametric claim size distributions from grouped data, *Blätter der Deutschen Gesellscaft für Versicherungsmatematik*, **XXIV**, 495–505.

Lindley, D. V. (1949). Grouping dorrections and maximum likelihood equations, *Proc. Camb. Philos. Soc.*, **46**, 106–110.

Robinson, P. M. (1988). The stochastic differences between econometric statistics, *Econometrica*, **56**(3), 531–548.

Sueishi, N., Liu, Q. F., Hitomi, K. and Nishiyama, Y. (2006). Efficiency improvement by local moments in grouped data analysis, CAEA Discussion Paper Series No. 107, Kyoto University.

Tallis, G. M. (1967). Approximate maximum likelihood estimates from grouped data, *Technometrics*, **9**, 599–606.

Victoria-Feser, M. and Ronchetti, E. (1997). Robust estimation for grouped data, *Journal of the American Statistical Association*, **92**, 333–340.

Wooldridge, J. M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples, *Econometric Theory*, **17**, 451–470.

Yanagimoto, T. (1990). Combining moment estimates of a parameter common through strata, *Journal of Statistical Planning and Inferences*, **25**, 187–198.